Assessing Sample Quality via the Latent Space of Generative Models

Jingyi Xu[1], Hieu Le[2], and Dimitris Samaras[1]

[1] Stony Brook University, New York, USA
[2] EPFL, Lausanne, Switzerland

Abstract. Advances in generative models increase the need for sample quality assessment. To do so, previous methods rely on a pre-trained feature extractor to embed the generated samples and real samples into a common space for comparison. However, different feature extractors might lead to inconsistent assessment outcomes. Moreover, these methods are not applicable for domains where a robust, universal feature extractor does not yet exist, such as medical images or 3D assets. In this paper, we propose to directly examine the latent space of the trained generative model to infer generated sample quality. This is feasible because the quality a generated sample directly relates to the amount of training data resembling it, and we can infer this information by examining the density of the latent space. Accordingly, we use a latent density score function to quantify sample quality. We show that the proposed score correlates highly with the sample quality for various generative models including VAEs, GANs and Latent Diffusion Models. Compared with previous quality assessment methods, our method has the following advantages: 1) pre-generation quality estimation with reduced computational cost, 2) generalizability to various domains and modalities, and 3) applicability to latent-based image editing and generation methods. Extensive experiments demonstrate that our proposed methods can benefit downstream tasks such as few-shot image classification and latent face image editing. Code is available at https://github.com/cvlab-stonybrook/LS-sample-quality.

Keywords: Generative Model, Quality Assessment, VAE, GAN, Diffusion

1 Introduction

Generative models have emerged as powerful modeling tools that can capture diverse and complex distribution from a large training dataset to synthesize new data [48–51]. A single pre-trained diffusion model [38] can generate thousands of images of "Yorkshire Terrier" or "Notre-Dame de Paris". In this paper, we aim to answer the question: among the samples generated from the model, how to measure the quality of each individual one? Such an instance-wise quality assessment metric is essential for users and consumers to select samples among the ones provided by those recently released text-to-image models, *e.g.*, DALL-E

2 [36] and Stable Diffusion [38], rather than model-wise metrics such as Frechèt Inception Distance (FID) [16].

For the most part, previous instance-wise evaluation methods [15, 24] rely on a pre-trained feature extractor (*e.g.*, VGG16 [43]) to embed the generated samples and real samples into a common feature space. K-nearest neighbor (k-NN) based approaches are then applied under the assumption that close samples in this feature space correspond to semantically similar images. The realism score [24], for example, measures the maximum of the inverse relative distance of a fake sample in a real k-NN latent sphere. The rarity score [15], on the other hand, measures the minimum radius of a real k-NN sphere that contains the fake latent representation. However, relying on a pre-trained feature extractor suffers from two shortcomings. First, different feature extractors might lead to inconsistent assessment outcomes: the rarity score shows a negative correlation with Frechèt Inception Distance (FID) [16] when using VGG16 as backbone, while the correlation becomes positive under DINO [6] or CLIP [35] backbones. Moreover, these methods are not applicable for domains where a robust, universal feature extractor is not yet available, *e.g.*, 3D shapes, human-drawn art or medical images.

In this paper, we propose to assess sample quality from another perspective: instead of using a pre-trained feature space, we directly use *the latent space of the generative models themselves*. The intuition is that the quality of a generated sample directly relates to the amount of the training samples that closely resemble it, and we can infer this information solely by examining the density of the latent space. Specifically, the samples lying in the latent area with dense latent codes are likely to have sufficient training data resembling them while low-density latent areas would correspond to the rare cases in the data manifold. This is because generative models typically map similar data points to similar latent embeddings. The latent embeddings in those low-density areas are less exposed in model training, consequently receiving less supervision, and leading to potentially inferior reconstruction quality.

To this end, we propose a latent density score function to measure the quality of generated samples. Given a pre-trained generative model, our proposed function quantitatively measures the density of a randomly sampled latent code w.r.t. a set of latent codes extracted from the training data. We show that the proposed latent density score highly correlates with the sample quality for various generative models including Variational Autoencoders (VAEs) [22], Generative Adversarial Networks (GANs) [14] and Latent Diffusion Models (LDMs) [38]. Compared with previous quality assessments that require an additional embedding network for feature extraction, our method estimates the sample quality by directly examining the latent space of the generative models, which brings several key advantages: 1) **efficiency**: our method enables quality assessment without generating image pixels, which significantly reduces the computational cost; 2) **generalizability**: our method eliminates the reliance on external feature extractors, which allows for generalization to the domains where a universal pre-trained feature extractor might not exist; 3) **applicability**: our method can be

seamlessly incorporated into latent-based image editing and generation methods, which can benefit various downstream tasks.

In short, our main contributions can be summarized as follows:

- We demonstrate that we can directly assess sample quality via the latent space of generative models themselves, while previous quality assessment methods rely on a pre-trained feature extractor to embed real and generated samples to a common space.
- We propose a score function to quantify sample quality by measuring the density in latent space. The proposed function is applicable to various generative models trained on a variety of datasets.
- We show the clear advantages of our proposed method over previous instance-wise evaluation methods, including significantly saving computational cost, generalizing across different domains and facilitating various downstream tasks.

2 Related Work

Previous metrics for quality assessment can be grouped into two main categories: model-wise evaluation metrics and instance-wise evaluation metrics. Model-wise evaluation metrics measure the performance of different generative models, while instance-wise evaluation metrics aim to compare the quality of each individual generated sample.

2.1 Model-wise Evaluation Metrics

Various model-wise evaluation metrics have been proposed to quantify the performances of generative models. Prevalent model-wise metrics include Inception Score (IS) [40], Kernel Inception Distance (KID) [3] and Frechèt Inception Distance (FID) [16]. They quantify the performance of a generative model by measuring the distribution discrepancy between the generated samples and real samples in a high-dimensional feature space. Sajjadi *et al.* [39] propose to further disentangle this discrepancy between distributions into two components: precision and recall. Precision represents the quality of generated samples while recall corresponds to the coverage of the real target distribution. Naeem *et al.* [32] improve upon precision and recall by introducing density and coverage: density improves upon precision by being more robust to outliers and coverage improves upon recall by preventing the overestimation of the latent manifold. Although the above metrics have demonstrated their effectiveness in assessing generative models, they are not suitable to measure individual sample quality since they work on a set of generations.

2.2 Instance-wise Evaluation Metrics

Unlike model-wise metrics, instance-wise metrics are applied on individual generated samples for performance evaluation. They are helpful for users to select

samples from generative models, which might produce noisy, unrealistic samples with artifacts, especially for underrepresented cases [29] such as rare categories or extreme object poses. The realism score [24] measures the perceptual quality of individual samples by estimating how close a given fake sample is to the latent manifold of real samples. Recently, Han *et al.* have proposed the rarity score [15], which measures how rare a synthesized sample is based on the real data distribution. Our proposed method and rarity score share the spirit of estimating the density around the target fake sample on the real manifold. Nevertheless, rarity score defines this manifold using a pre-trained classification network, *e.g.*, VGG16, while our method directly leverages the latent manifold of the generative models themselves. We show that in this latent manifold, the density correlates with the perceptual quality of the generated samples.

3 Latent Density Score

Given a well-trained generative model, *e.g.*, GAN, VAE or latent diffusion model, we aim to estimate the quality of the generated samples by examining the latent space of the model. Let $\mathcal{Z} = \{z_1, z_2, ...z_i\}$ denote a set of latent codes extracted from the training samples, and z_g denote a latent code randomly sampled from the latent space, we measure the latent density of z_g quantitatively by calculating the latent density score as:

$$D(z_g, \mathcal{Z}) = \frac{1}{|\mathcal{Z}|} * \sum_{z_i \in \mathcal{Z}} e^{-\frac{\|z_g - z_i\|^2}{2\sigma^2}}, \tag{1}$$

where σ is a hyper-parameter of this score function. Latent density score measures the average Gaussian kernelized Euclidean distance [10] between z_g and each latent code in $\mathcal{Z}$. The score is high when z_g resides in an area where the trained codes are densely distributed. σ controls the relative contribution of each latent code in $\mathcal{Z}$ to the final density value, *i.e.*, using a small σ places more emphasis on the local area surrounding z_g, while applying a large σ places more focus on the global density. In the case where there are multiple local clusters in the latent manifold, different values of σ will lead to different assessment results (see Section 6.2).

In GAN-based generative models, truncation trick [5, 20, 23] is a widely used technique to increase the sample fidelity at the cost of lowering the diversity. It works by shifting a randomly sampled code towards the mean latent code. The mean code typically resides in a high-density latent area. In fact, we observe that the proposed latent density score well correlates with the degree of truncation. We analyze this correlation further in Section 6.1 and provide more qualitative results in the Supplementary Material. Another highly relevant quality assessment metric is the realism score [24]. The realism score measures the relative distance of a fake sample in a real latent sphere, which is defined by a pre-trained feature extractor. We show that the latent density score behaves similarly with the realism score for images from the domains previously seen by the feature extractor (see Section 6.1). However, for images from non-ImageNet-like domains

(*e.g.*, medical images and anime-style images) or domains other than 2D images (*e.g.*, 3D shapes), quality assessment with realism score will be infeasible (see Section 4.2).

(a) Latent Diffusion - CelebA-HQ

(b) Latent Diffusion - LSUN-Bedrooms

(c) Latent Diffusion - LSUN-Churches

Fig. 1: Top 6, Middle 6 and Bottom 6 generated images in terms of the proposed latent density score on CelebA-HQ, LSUN-Bedrooms and LSUN-Churches for unconditional latent diffusion models. (Zoom-in for best view). The proposed latent density scores highly correlate with the quality of generated images.

4 Experimental Results

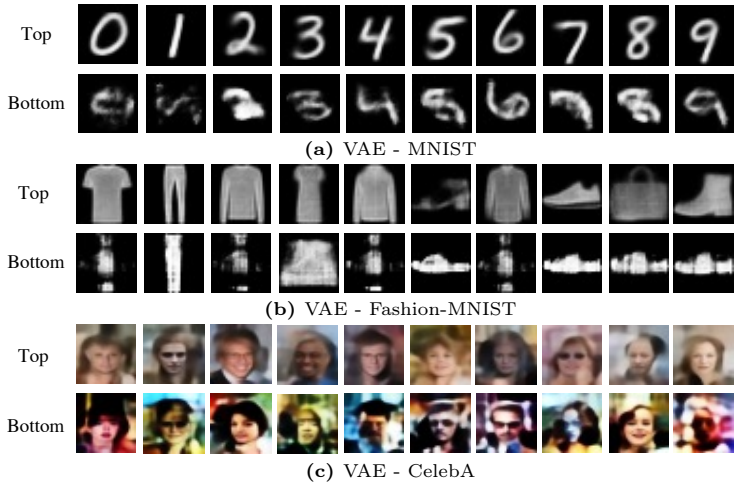
4.1 Results on Different Generative Models

In this section, we provide the experimental results of the proposed metric for various generative models and datasets. We experiment with three types of generative models, *i.e.*, GANs, VAEs and LDMs. For each trained model, we extract latent codes from 60k training samples and calculate the latent density scores for 20k randomly sampled latent codes. In particular, for VAEs and LDMs, we

Fig. 2: Top 2 and bottom 2 Stable Diffusion generated samples for eight classes in terms of the proposed latent density score. Images in the 'top 2' rows are high-resolution images with natural, realistic backgrounds, whereas images in the 'bottom 2' rows contain visual noise and artifacts. The only difference in model configuration for images of top / bottom rows is the initial noise.

take the output of the image encoder as the latent representation of each real input image. For LDMs, we further flatten the 2D representations (before the denoising process) for computing the latent density score. We use the pre-trained Stable Diffusion v1.5 model [38] as the text-to-image diffusion model. For the unconditional diffusion models, we choose the LDMs pre-trained on CelebA-HQ [9], LSUN-Bedrooms and LSUN-Churches [53] released by [38]. For GANs, we experiment with StyleGAN2 [21] and StyleGAN2-ADA [19]. To obtain the latent representations, we input vectors sampled from a normal distribution to their mapping networks and extract latent features from the $\mathcal{W}$-space. We use $\sigma = 20$ for computing latent density scores. We analyze the choice of σ and how it affects the results in Section 6.2.

Latent Diffusion Models Latent diffusion models [38] use pre-trained autoencoders to construct a low-dimensional latent space, from which the original data can be reconstructed at high fidelity with reduced computational costs. In Figure 1, we show images synthesized by unconditional latent diffusion models trained on CelebA-HQ, LSUN-Bedrooms and LSUN-Churches. For each dataset, we show samples using latent codes with the top 6 highest, top 6 lowest and 6 middle latent density scores. As shown in the figure, the proposed latent density scores highly correlate with the quality of generated images. For example, on the CelebA-HQ dataset, we can see human faces generated from codes with high latent density scores are visually realistic with clear hair, eye and eyebrow details, whereas those with low latent density scores are of degraded quality due to blur, artifacts or distorted facial structures. Similarly, on LSUN-Bedrooms



(a) VAE - MNIST

(b) VAE - Fashion-MNIST

(c) VAE - CelebA

Fig. 3: Top 10 and Bottom 10 generated images in terms of the proposed latent density score on MNIST, Fashion-MNIST and CelebA for VAE. The samples with high latent density scores display clear instances, whereas those with low latent density scores are often distorted / blurred.

and LSUN-Churches, we observe unrealistic artifacts (*i.e.*, distorted textures or inharmonious colors) from images with low latent density scores.

Figure 2 shows images synthesized by a pre-trained text-to-image diffusion model, *i.e.*, Stable Diffusion, using latent codes with the top 2 highest and lowest latent density scores from eight classes. As shown in the figure, samples with high latent density scores have superior visual quality while latent codes with low scores often lead to erroneous samples. The most obvious failures are the unrealistic backgrounds. For example, the boat images in the 'top 2' rows are high-resolution images with natural, realistic backgrounds, whereas the backgrounds of the boats in the 'bottom 2' rows contain visual noise and artifacts. In some other failure cases, the generated objects exhibit structural integrity artifacts, *i.e.*, the spoons and the clocks. We note that all images here are generated with the same model configuration and the only difference is the initial noise. Previous works [11, 17, 33] have shown that the guidance methods used in the denoising process are essential for improving the quality of generated images. Here we observe that the randomly initialized noise can also affect the sample quality significantly.

VAEs and GANs Figure 3 presents images generated by VAEs using codes with the top 10 highest and top 10 lowest latent density scores. As shown in the figure, our proposed score function is applicable to VAEs as well. For MNIST [28] and Fashion-MNIST [47], for example, we observe the samples with high latent

density scores display clear instances from the given class, whereas those with low latent density scores are often distorted / blurred to unrecognizable.

Figure 4 shows images generated by StyleGAN2 [21] trained on FFHQ [20], StyleGAN2-ADA [19] trained on AFHQ Dog [9] and StyleGAN2 trained on AFHQ Cat [9] using codes with the top 6 highest, top 6 lowest and 6 middle latent density scores. We observe clear generation quality differences between samples with different scores. For example, on the AFHQ Dog dataset, the samples with high scores show clear, frontal dog faces. On the other hand, the dog faces in samples with low density scores are highly distorted in various ways.

(a) StyleGAN2 - FFHQ

(b) StyleGAN2-ADA - AFHQ Dog

(c) StyleGAN2 - AFHQ Cat

Fig. 4: Top 6, Middle 6 and Bottom 6 generated images in terms of the proposed latent density score on FFHQ for StyleGAN2, on AFHQ Dog for StyleGAN2-ADA and on AFHQ Cat for StyleGAN2. (Zoom-in for best view). Samples with high scores are of better quality while samples with low scores are often highly distorted.

4.2 Results on Other Domains and Modalities

Our proposed metric does not rely on any additional feature extractor, which enables quality assessment in the domains where robust pre-trained models might

not be available. In this section, we show the applications of our method on quality assessment for generated 3D shapes and non-ImageNet-like images.

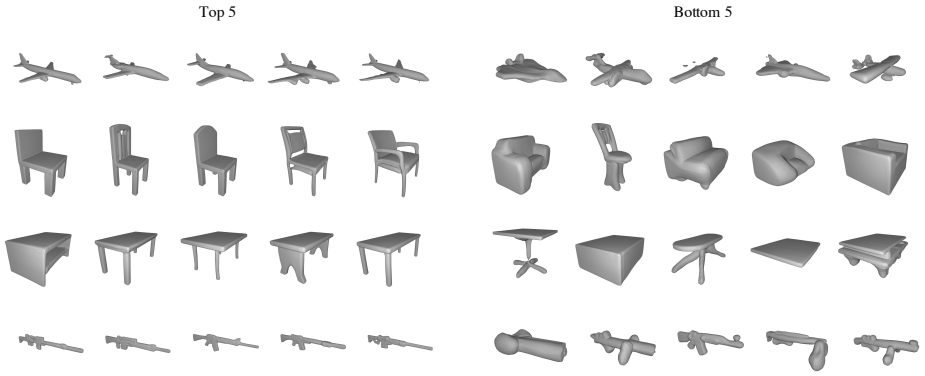
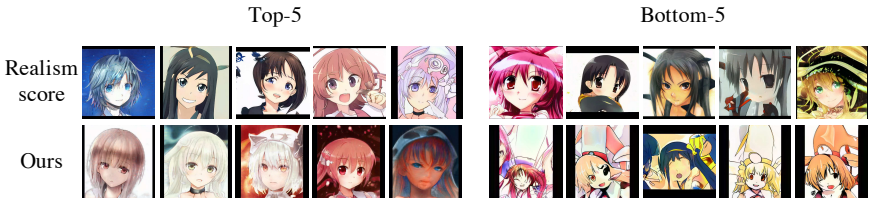


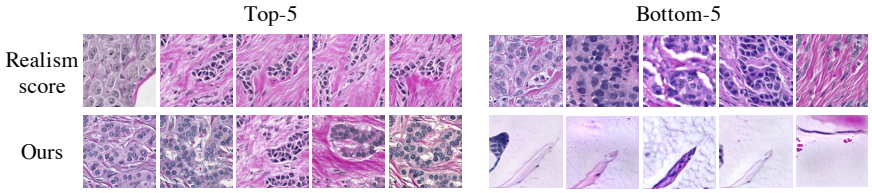
Fig. 5: Top 5 and Bottom 5 generated 3D shapes for four categories (*i.e.*, airplane, chair, table and rifle) in terms of the proposed latent density score on ShapeNet Core V1 for SDF-StyleGAN. The generated samples with high scores have plausible 3D shapes and complete geometry structures, while samples with low scores exhibit unrealistic shapes and severe geometry distortion.

Quality Assessment for 3D Shapes We first show the application of our method on generated 3D shapes. Specifically, we generate shapes for four categories, *i.e.*, airplane, chair, table and rifle, using a StyleGAN2-based 3D shape generation framework, SDF-StyleGAN [54] trained on ShapeNet Core V1 [7]. For each shape category, we extract the latent embeddings in the $\mathcal{W}$ space of SDF-StyleGAN for 30k randomly sampled vectors and compute the corresponding latent density scores. Figure 15 visualizes the generated shapes with the top 5 highest and lowest scores. We observe that the generated 3D shapes with high scores have better visual quality with plausible 3D shapes and complete geometry structures. The generated 3D shapes with low scores, in contrast, exhibit unrealistic shapes and severe geometry distortion.

Quality Assessment on Non-ImageNet-like Images Existing quality assessment methods operate under the assumption that semantically similar images are mapped to points close to each other in the embedding space of a pre-trained feature extractor. However, this assumption might not hold true across different data domains. In this section, we conduct quality assessment for images from two non-ImageNet-like domains, *i.e.*, the medical domain and anime-style domain. Figure 6 shows the samples with the highest and lowest latent density scores / realism scores among 5k candidate samples on each domain. We can see



(a) StyleGAN2 - Danbooru



(b) StyleGAN2 - BreCaHAD

Fig. 6: Top and bottom generated images in terms of latent density score and realism score using StyleGAN2 pre-trained on BreCaHAD [1] and Danbooru [4] datasets. There is clear visual difference between images with the highest and lowest latent density scores. In comparison, we do not observe visually distinguishable difference between samples with the highest and lowest realism scores.

that the images with the highest and lowest latent density scores exhibit clear visual difference. On the BreCaHAD dataset [1], for example, the high-density images contain representative human cells while low-density images are mostly blank. We note that the low-density samples do not show degraded perceptual quality. This is probably because although these images are underrepresented cases in the training set, reconstructing them is relatively easy due to their simple layouts.

On the other hand, we do not observe visually distinguishable differences between samples with the highest and lowest realism scores. This suggests that the pre-trained VGG space used by the realism score is not semantically meaningful for non-ImageNet-like domain images. In addition, our method is more computationally efficient, since we directly operate on the latent codes instead of actually generating all the 5k candidate images.

5 Applications

5.1 Latent Face Editing

Our proposed method operates directly on the latent space of the generator. Thus, it can be seamlessly incorporated into latent-based image editing methods. Previous work [42] has shown that by moving a latent code along certain directions in the latent space of a well-trained face synthesis model, one could control facial attributes of the generated images. However, if the code is moved

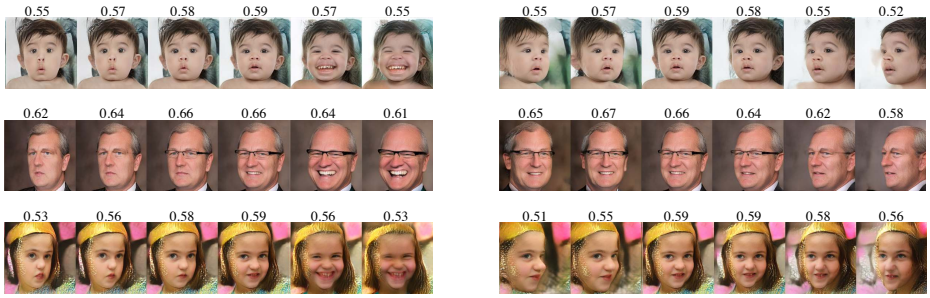


Fig. 7: Latent-based face editing results and the corresponding latent density scores. The latent density scores highly correlate with the quality of the images.

too far from the well-behaved regions [55] of the latent space, the generated samples will suffer from severe changes [42] as well as degradation in image quality. Here we use our method to estimate the perceptual quality of the edited samples. Specifically, we take a latent code and move it along the direction for the attribute "pose" in the latent space $\mathcal{W}$ of StyleGAN2 following [42]. We compute the latent density score of the moved latent code based on Equation 1. Figure 7 shows the generated edited images and the corresponding scores.

As shown in the figure, the latent density scores well correlate with the quality of the manipulated images: images with low scores contain artifacts while images with high scores are of better quality. Our method provides a reliable way to assess the quality of edited images even before generating them, which helps to avoid image corruption during latent space traversal and facilitates meaningful image manipulation.

5.2 Few-Shot Image Classification

Our method enables selecting strictly high-quality generated images with clear, high-resolution objects. These images are particularly useful for augmenting the training set in low-shot scenarios [48–51]. Here we show these samples can be used in the task of few-shot image classification and greatly boost performance. Specifically, we synthesize images using a pre-trained text-to-image model, *e.g.*, Stable Diffusion, with the class name as the text condition. The synthesized images are then used as support samples for the corresponding class. We generate k images for k-shot learning ($k = 1$ or 5). For the feature extractor, we use ResNet12 [34] trained following previous work [8]. Table 1 compares the performance of using different sets of latent codes during image generation including: 1) k randomly sampled codes, 2) top-k codes with the highest, and 3) top-k codes with the lowest latent density scores. Using samples with high latent density scores as support data leads to better few-shot performance on both the *mini*ImageNet [45] and *tiered*ImageNet [37] datasets for the 1-shot and 5-shot settings. In particular, results on 1-shot *mini*ImageNet show the largest margin,

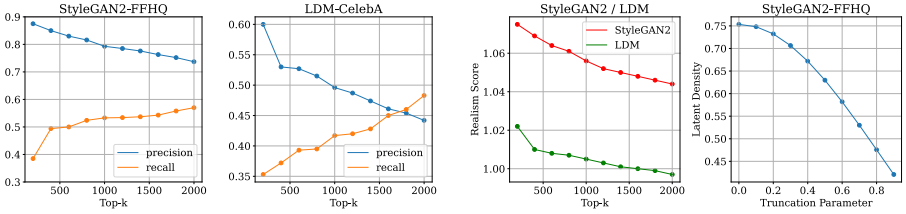
Table 1: Few-shot image classification accuracy on *mini*ImageNet and *tiered*ImageNet using images generated from different sets of latent codes. Using images from the latent codes with highest latent density scores achieves better classification performance, which validates the superior quality of images with high latent density scores.

	Support Samples	*mini*ImageNet		*tiered*ImageNet	
		1-shot	5-shot	1-shot	5-shot
Real	-	63.17 ± 0.23	79.26 ± 0.17	68.62 ± 0.27	83.74 ± 0.18
SD-generated	bottom-k	63.43 ± 0.45	71.97 ± 0.37	64.62 ± 0.55	75.08 ± 0.45
	random-k	63.87 ± 0.43	72.76 ± 0.38	66.04 ± 0.52	77.10 ± 0.44
	top-k	**67.15 ± 0.44**	**73.60 ± 0.37**	**68.39 ± 0.54**	**77.42 ± 0.43**

with a 3.28% improvement over using random codes. This validates the superior quality of images generated from codes with high latent density scores.

6 Analysis

6.1 Relationship with Existing Metrics



(a) Precision / Recall (b) Realism Score (c) Truncation

Fig. 8: Relationship between our proposed latent density score and other metrics. Top-k samples are ranked according to our latent density score. The results of our proposed metric are aligned with existing evaluation metrics on images of common domains.

In this section, we investigate the relationship of our proposed metric with other existing metrics. In particular, we generate 2000 fake samples and rank them based on the latent density score. Each time we select top-k samples and calculate the corresponding precision / recall / realism scores of the selected samples. The scores of these metrics under different k values are shown in Figure 8 (a) and Figure 8 (b). In addition, we show in Figure 8 (c) how the latent density score changes when we increase the value of truncation parameter used in truncation trick. We conduct this experiment using StyleGAN2 trained on FFHQ [20] and LDM trained on CelebA-HQ [9].

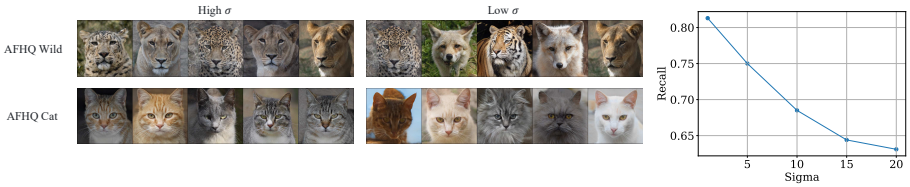
Precision and Recall. Precision and recall are commonly used evaluation metrics in many tasks, such as image classification or natural language processing. In particular, precision measures the fraction of the generated samples that are realistic. Recall, on the other hand, measures the fraction of the real data distribution which can be covered by the distribution of fake data. As shown in Figure 8 (a), a small value of k leads to high precision and low recall. This suggests that the samples with high latent density scores are of high quality. As k increases, more diverse samples are selected, which improves recall, while the decrease in precision indicates the newly selected samples are of inferior quality. The correlation between precision / recall and latent density score validates that our proposed metric reliably indicates sample quality.

Realism Score. The realism score is a highly relevant metric that measures the fidelity of an individual generated sample. As shown in Figure 8 (b), as k increases, the average realism score of the top-k selected samples decreases for both StyleGAN2 and LDM. This suggests that our proposed metric is aligned with realism score, *i.e.*, samples with low latent density scores also have low realism scores, and vice versa. However, the realism score relies on another feature extractor to project the generated samples to another space. Thus, it is not able to generalize to other domains or modalities (as shown in Section 4.2). Moreover, computing realism score requires generating image data, which is time-consuming and not easily scalable, while our method directly operates on the latent space without the need for generating images.

Truncation Trick. Truncation trick [20] is used to increase the fidelity of the generated images in GAN-based generative models by moving the latent code towards the mean latent code. The degree of truncation is controlled by the truncation parameter ψ, *i.e.*, $\psi = 0$ indicates full truncation using the mean code and $\psi = 1$ indicates no truncation. We see from Figure 8 (c) that the latent density score decreases as the truncation parameter increases. A higher degree of truncation typically leads to high-fidelity image generation, which, as we show, corresponds to a higher value of the latent density score. This suggests that the latent density score is a valid measure for generated sample quality.

6.2 Effect of Hyper-Parameter σ

In this section, we analyze how the choice of σ in Equation 1 affects quality assessment. Equation 1 measures the average Gaussian kernelized Euclidean distance between a given code and the latent codes extracted from training data, with σ being the standard deviation of the kernel function. When applying a small σ, the final density value will rely relatively more on the area surrounding the given code. This will increase the chance that points residing in local clusters are selected as high-density points. As a result, the selected samples are likely to be more diverse. Figure 19 (a) shows the images with high latent density scores on the AFHQ Wild and AFHQ Cat datasets [9] under large and small σ values respectively. We observe that the selected high-density samples when using a small σ are more diverse compared to using a large σ. Correspondingly, we observe a higher recall under a smaller σ (as shown in Figure 19 (b)), indicating



(a) (b)

Fig. 9: Images with high latent density scores using high and low σ values on the AFHQ Wild and AFHQ Cat datasets (a) and different recall rates under different σ values (b). We observe that under a lower σ value, the images selected as high-density ones exhibit more diversity, which corresponds to a higher recall rate.

a more complete coverage of the latent manifold. σ enables us to control the relative contribution of local density and global density w.r.t to the final density score. In this case, applying a small σ allows us to select more diverse samples.

7 Discussion and Conclusions

In this paper, we have proposed a novel approach to estimate sample quality via the latent space of generative models. Our method can be particularly useful in many scenarios. When training generative models, our proposed score points out the underrepresented cases that would possibly require collecting additional data [12,13,26,27,30,48–52]. It also allows us to select high-quality samples that best benefit downstream tasks. For large-scale generative models, pre-generation quality assessment can greatly reduce computational costs. However, only sampling data with high scores, might result in an incomplete coverage of the data manifold. This is because the scores are likely to be higher for large clusters of data representing common cases, as opposed to minority groups such as rare animal species [25] or uncommon medical conditions. One way to alleviate this issue is by considering only small neighborhood areas when measuring the density, which can be achieved by applying a small value of σ. Further, previously proposed sampling techniques such as accept-reject sampling [2,30] can be used together with our method to increase sample diversity. Combining our score with diversity-related scores also allows us to select diverse samples with high quality. In future work, we intend to extend our method to generative models with high latent dimensions, such as deep hierarchical VAEs [44], or video generative models [46] with an additional temporal dimension in latent space. For these models, latent dimension reduction techniques might be a potential solution.

Acknowledgement. This research was partially supported by NSF grants IIS-2123920 and IIS-2212046.

A Appendix-Overview

In the appendix, we provide additional experiments and analyses. In particular:

- Section B provides additional visualizations of the selected samples with the highest/lowest latent density scores across different generative models and datasets.
- Section C provides additional results of applying our quality assessment method on other domains and modalities.
- Section D shows the latent density scores for images at different truncation levels.
- Section F provides visualizations of the latent space of different generative models.
- Section E provides additional analysis on the choice of hyper-parameter σ in our score function.
- Section G justifies the choice of our score function.
- Section H provides quantitative results of the efficiency gain of our method.
- Section I includes a user study of our method.

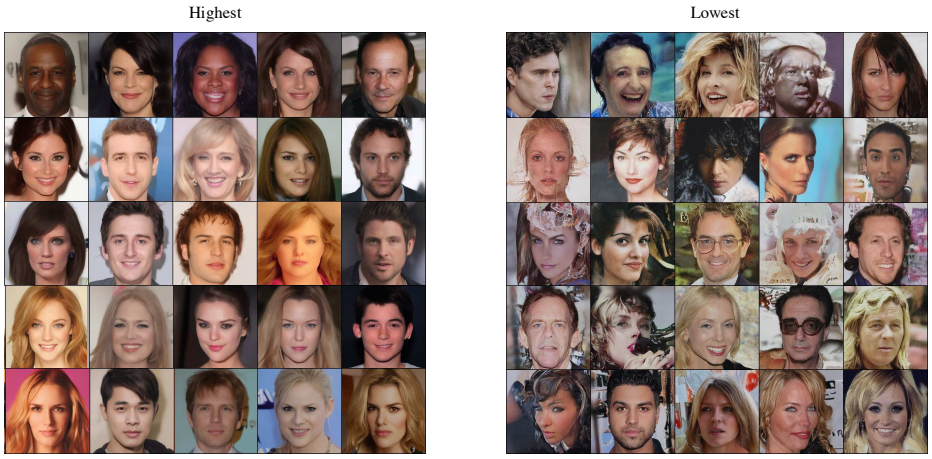


Fig. 10: Samples generated from Latent Diffusion [38] trained on Celeba-HQ [18] with the lowest/highest latent density scores.

B Additional Qualitative Results

In this section, we show additional selected samples with the highest/lowest latent density scores across different generative models and datasets.

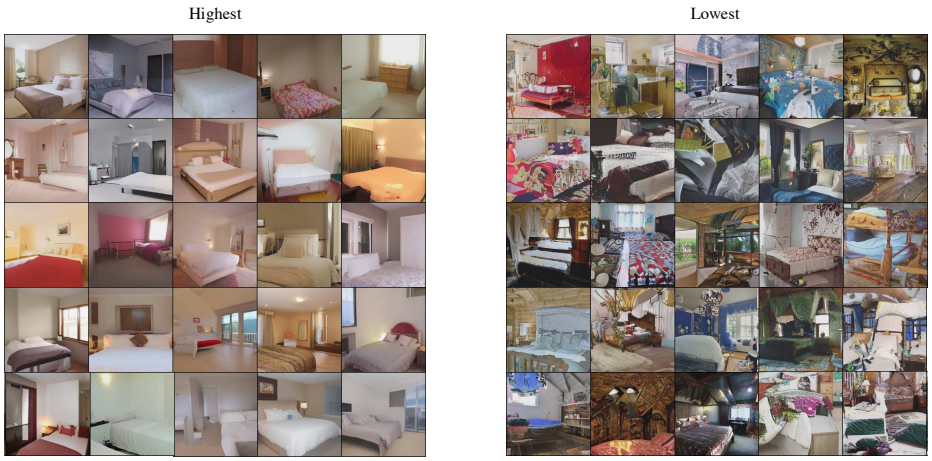


Fig. 11: Samples generated from Latent Diffusion trained on LSUN-Bedrooms [53] with the lowest/highest latent density scores.

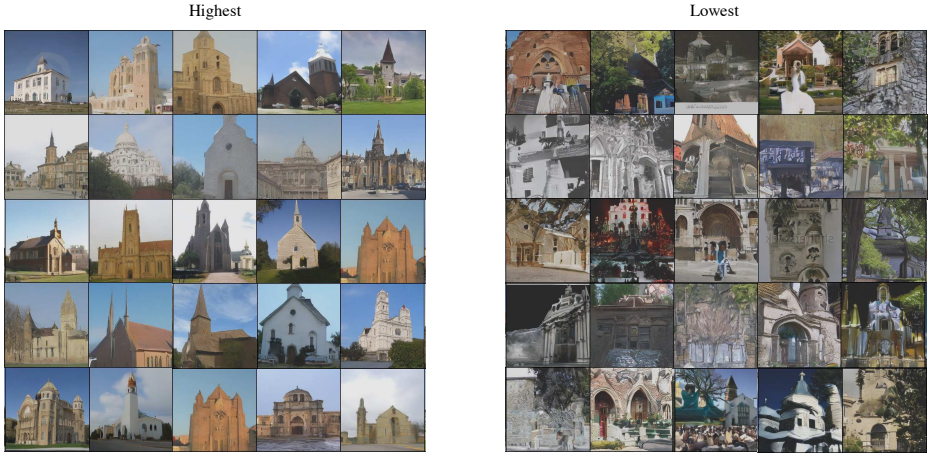


Fig. 12: Samples generated from Latent Diffusion trained on LSUN-Churches [53] with the lowest/highest latent density scores.

C Additional Results on Other Domains and Modalities

C.1 Quality Assessment on 3D Shapes

In Figure 15, we show additional quality assessment results on generated 3D shapes using our proposed latent density score. The 3D shapes are generated via a StyleGAN2-based [21] 3D shape generation framework, SDF-StyleGAN [54]

Fig. 13: Samples generated from StyleGAN2-ADA [19] trained on AFHQ Dog [9] with the lowest/highest latent density scores.

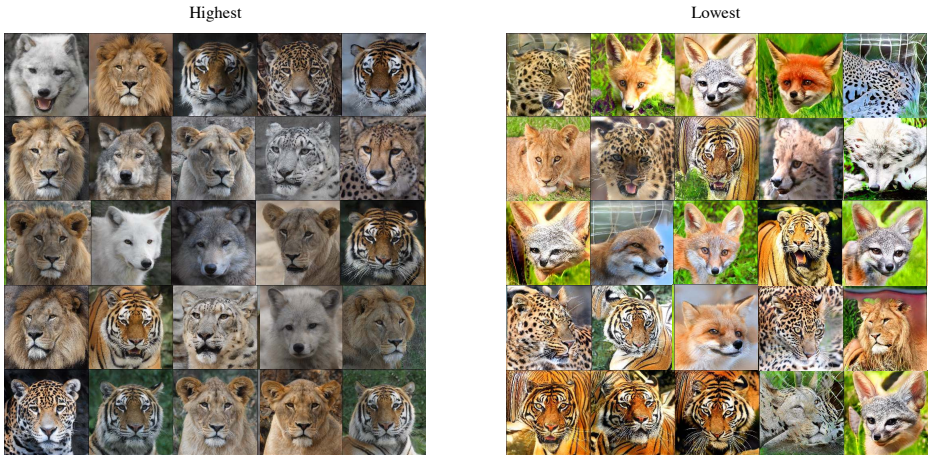
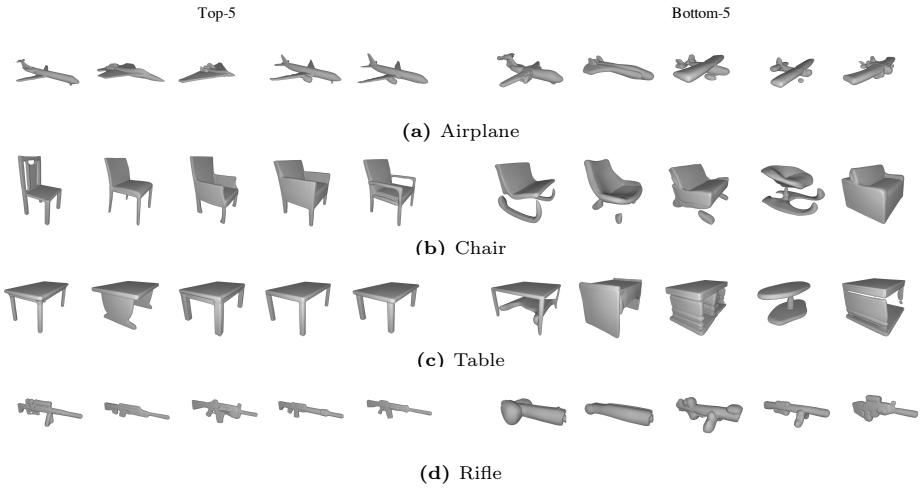


Fig. 14: Samples generated from StyleGAN2-ADA trained on AFHQ Wild [9] with the lowest/highest latent density scores.

trained on ShapeNet Core V1 [7]. As shown in the figure, the generated samples with high scores have better visual quality with meaningful object structures. The generated samples with low scores, in contrast, exhibit irregular objects shapes and severe geometry distortion.



(a) Airplane

(b) Chair

(c) Table

(d) Rifle

Fig. 15: Top 5 and Bottom 5 generated 3D shapes for four categories (i.e., airplane, chair, table and rifle) in terms of the proposed latent density score on ShapeNet Core V1 for SDF-StyleGAN.

C.2 Quality Assessment on Non-ImageNet-like Domain Images

In Figure 16 and Figure 17, we show additional selected samples with the highest and lowest latent density scores / realism scores on the BreCaHAD [1] and Danbooru [4] datasets. We see there exhibit clear visual differences between images with the highest and lowest latent density scores on both datasets. However, we do not observe visually distinguishable differences between samples with the highest and lowest realism scores. This is because the pre-trained feature space used by the realism score is not semantically meaningful for non-ImageNet-like domain images. Our proposed latent density score, on the other hand, directly leverages the latent space of generative models and generalizes well across different domains.

D Latent Density Scores for Images at Different Truncation Levels

The Truncation trick [20] is used to increase the fidelity of the generated images in GAN-based generative models by moving the latent code towards the mean latent code. The degree of truncation is controlled by the truncation parameter ψ, *i.e.*, $\psi = 0$ indicates full truncation using the mean code and $\psi = 1$ indicates no truncation. Figure 18 shows the images truncated using different values of ψ and the corresponding latent density scores. As shown in the figure, as ψ goes from 1 to 0.1, the perceptual quality of the truncated images gradually increases. For example, the face image in the third row without any truncation shows clear

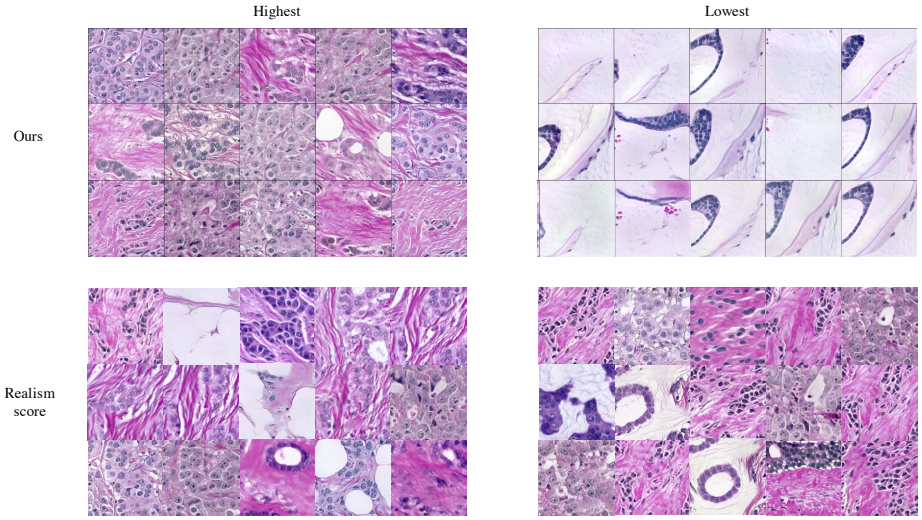


Fig. 16: Images with highest and lowest latent density scores / realism scores on BreCaHAD dataset

Fig. 17: Images with highest and lowest latent density scores / realism scores on Danbooru dataset

artifacts, while the artifacts are hardly noticeable when ψ is 0.1. Correspondingly, the latent density score gradually increases as the image quality becomes better. This suggests that the latent density score is a valid measure for generated sample quality.

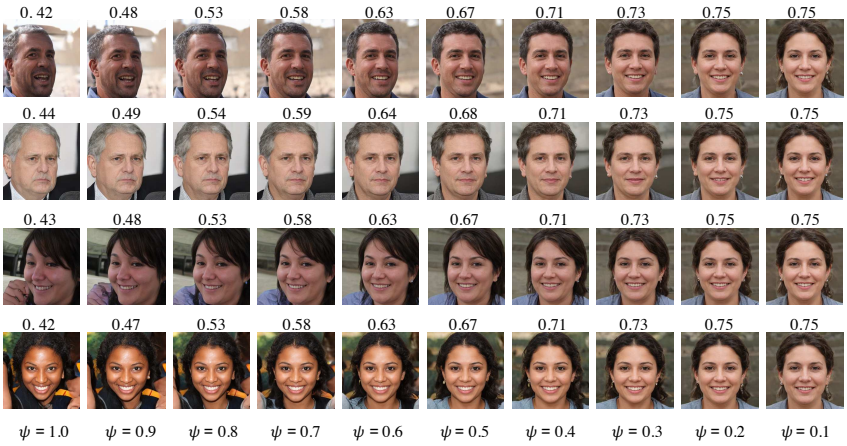


Fig. 18: Latent density scores of images at different truncation levels.

E Effect of Hyper-Parameter σ

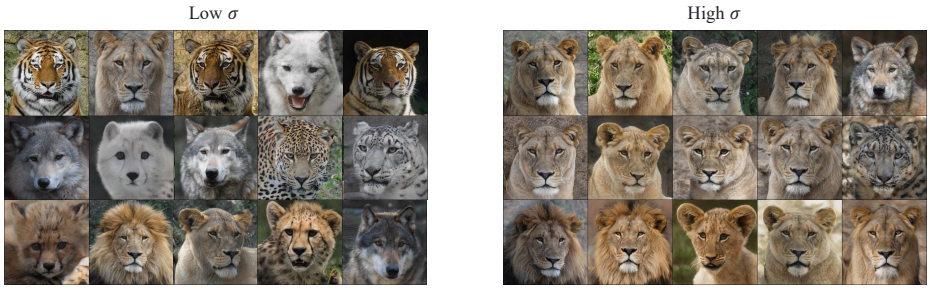
As discussed in the main paper, the hyper-parameter σ in our score function controls the relative contribution of local density and global density w.r.t. to the final density score. Applying a small σ will increase the chance of selecting the points residing in local clusters as high-density points. Thus, the selected samples are likely to be more diverse. Figure 19 shows additional selected high-score samples when using small and large σ values. As shown in the figure, the selected high-density samples when using a small σ are more diverse compared to using a large σ.

F T-SNE Visualization of Latent Space

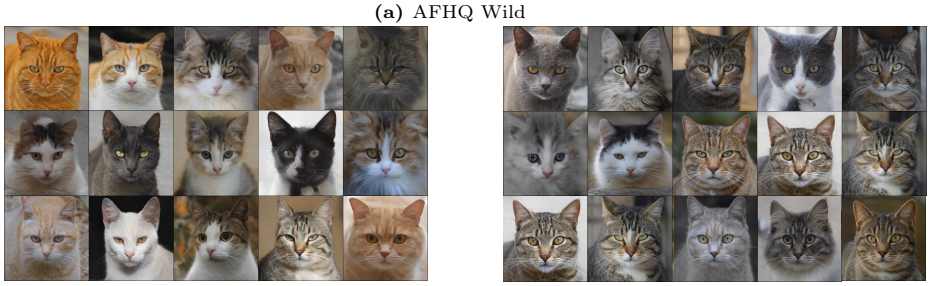
In Figure 20, we show the t-SNE [31] visualization of the latent space of three generative models, *i.e.*, StyleGAN2-ADA [19] trained on AFHQ Wild [9], Latent Diffusion [38] trained on CelebA-HQ [18] and Stable Diffusion [38] trained on LAION-5B [41]. We randomly select 4k images from the training set of each model and extract their latent embeddings. As shown in the figure, different regions in the latent space show different density patterns: latent codes are densely distributed in some areas, while sparsely distributed in other areas. In fact, we show via extensive experiments that this latent density directly correlates with the quality of generated samples.

G Justification of the Score Function

In our main experiments, we choose Gaussian score function because it is simple, intuitive and effective. In addition, we explore alternative formulations, includ-

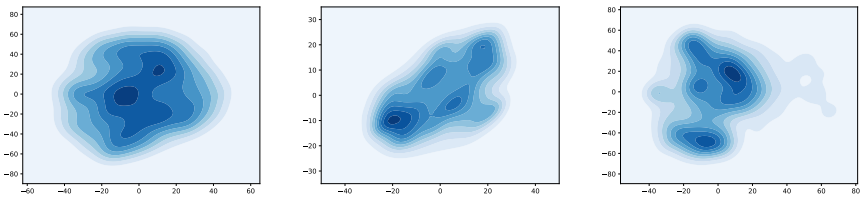


(a) AFHQ Wild



(b) AFHQ Cat

Fig. 19: Selected high-density samples when using low and high values of σ on the AFHQ Wild and AFHQ Cat datasets.



(a) StyleGAN2 - ADA - AFHQ Wild (b) Latent Diffusion – CelebA-HQ (c) Stable Diffusion - LAION-5B

Fig. 20: T-SNE visualization of the latent space of three generative models.

ing two k-NN based approaches in [24] and [15]. We show the average realism scores of the top-k samples based on different formulations in Tab. 2. While our formulation better aligns with the realism score in this case, other functions can also be feasible choices.

H Quantifying Efficiency Improvements

The efficiency gain of our method mainly comes from bypassing pixel-level image generation. For example, generating a single image with an LDM takes 29.378

top-k	200	400	600	800	1000
[15]	1.058	1.056	1.055	1.054	1.053
[24]	1.061	1.056	1.054	1.053	1.053
Ours	1.065	1.063	1.059	1.056	1.053

Table 2: Avg. realism scores of top-k samples.

secs (with 50 timesteps on a TITAN RTX GPU). In comparison, our method only requires 0.009 secs to extract one latent embedding from the LDM's encoder and 0.4 ms to compute the density score.

I User Study

We conduct a small-scale user study with 12 participants to further verify our method. We ask them to choose the more realistic images between two sets with different scores. Each set contains 9 images. 11 answers (91.7%) align with the density score.

References

1. Aksac, A., Demetrick, D.J., an Reda Alhajj, T.O.: Brecahad: A dataset for breast cancer histopatholog- ical annotation and diagnosis. BMC Research Notes (2019)
2. Azadi, S., Olsson, C., Darrell, T., Goodfellow, I., Odena, A.: Discriminator rejection sampling. In: ArXiv (2019)
3. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton., A.: Demystifying mmd gans. In: ICLR (2018)
4. Branwen, G., Anonymous, Community., D.: Danbooru2019 portraits: A large-scale anime head illus- tration dataset. https://www.gwern.net/crops#danbooru2019-portraits (2019)
5. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. In: ICLR (2019)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin., A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
7. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
8. Chen, Y., Liu, Z., Xu, H., Darrell, T., Wang, X.: Meta-baseline: Exploring simple meta-learning for few-shot learning. In: ICCV (2021)
9. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: CVPR (2020)
10. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. In: TPAMI (2002)
11. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
12. Durasov, N., Dorndorf, N., Le, H., Fua, P.: Zigzag: Universal sampling-free uncertainty estimation through two-step inference. Transactions on Machine Learning Research (2024)

13. Durasov, N., Oner, D., Donier, J., Le, H., Fua, P.: Enabling uncertainty estimation in iterative neural networks. In: Forty-first International Conference on Machine Learning (2024)
14. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio., Y.: Generative adversarial networks. In: NeurIPS (2014)
15. Han, J., Choi, H., Choi, Y., Kim, J., Ha, J.W., Choi, J.: Rarity score : A new metric to evaluate the uncommonness of synthesized images. In: ICLR (2023)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter., S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
17. Ho, J., Salimans., T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
18. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, sta- bility, and variation. CoRR (2017)
19. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: NeurIPS (2020)
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
21. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR (2020)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
23. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. In: NeurIPS (2018)
24. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Mimproved precision and recall metric for assessing generative models. In: NeurIPS (2019)
25. Le, H., Goncalves, B., Samaras, D., Lynch, H.: Weakly labeling the antarctic: The penguin colony case. In: CVPR Workshops (June 2019)
26. Le, H., Samaras, D.: Physics-based shadow image decomposition for shadow removal. IEEE Computer Society, Los Alamitos, CA, USA. https://doi.org/10.1109/TPAMI.2021.3124934
27. Le, H., Samaras, D.: From shadow segmentation to shadow removal. In: European Conference on Computer Vision(ECCV) (2020)
28. LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit databases. Technical Report (2014)
29. Lee, J., Kim, H., Hong, Y., Chung, H.W.: Self-diagnosing gan: Diagnosing underrepresented samples in generative adversarial networks. In: NeurIPS (2021)
30. Li, S., Liu, C., Zhang, T., Le, H., Süsstrunk, S., Salzmann, M.: Controlling the fidelity and diversity of deep generative models via pseudo density (2024), https://arxiv.org/abs/2407.08659
31. Maaten, L.V.D., Hinton, G.E.: Visualizing data using t-sne. In: Journal of Machine Learning Research (2008)
32. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo., J.: Reliable fidelity and diversity metrics for generative models. In: ICML (2020)
33. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen., M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv: abs/2112.10741 (2021)
34. Oreshkin, B., Lopez, P.R., Lacoste., A.: Tadam: Task dependent adaptive metric for improved few-shot learning. In: NeurIPS (2018)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)

36. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen., M.: Hierarchical text- conditional image generation with clip latents. ArXiv **abs/2204.06125** (2022)
37. Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. arXiv preprint arXiv:1803.00676 (2018)
38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
39. Sajjadi, M.S.M., Bachem, O., Lucic, M., Bousquet, O., Gelly, S.: Assessing generative models via precision and recall. In: NeurIPS (2018)
40. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen., X.: Improved techniques for training gans. In: NeurIPS (2016)
41. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models. ArXiv **abs/2210.08402** (2022), `https://api.semanticscholar.org/CorpusID:252917726`
42. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: CVPR (2020)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
44. Vahdat, A., Kautz, J.: Nvae: A deep hierarchical variational autoencoder. In: NeurIPS (2020)
45. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: NeurIPS (2016)
46. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: ICCV (2023)
47. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
48. Xu, J., Le, H.: Generating representative samples for few-shot classification. In: CVPR (2022)
49. Xu, J., Le, H., Huang, M., Athar, S., Samaras, D.: Variational feature disentangling for fine-grained few-shot classification. In: ICCV (2021)
50. Xu, J., Le, H., Nguyen, V., Ranjan, V., Samaras, D.: Zero-shot object counting. CVPR (2023)
51. Xu, J., Le, H., Samaras, D.: Generating features with increased crop-related diversity for few-shot object detection. In: CVPR (2023)
52. Xu, J., Le, H., Samaras, D.: Zero-shot object counting with language-vision models (2023), `https://arxiv.org/abs/2309.13097`
53. Yu, F., Zhang, Y., Song, S., Seff, A., Xiao., J.: Lsun: construction of a large-scale image dataset using deep learning with humans in the loop. CoRR (2015)
54. Zheng, X.Y., Liu, Y., Wang, P.S., Tong, X.: Sdf-stylegan: Implicit sdf-based stylegan for 3d shape generation. In: SGP (2022)
55. Zhu, P., Abdal, R., Qin, Y., Femiani, J., Wonka, P.: Improved stylegan embedding: Where are the good latents? arXiv preprint arXiv:2012.09036 (2020)