NoisyTwins: Class-Consistent and Diverse Image Generation through StyleGANs

Harsh Rangwani[1*] Lavish Bansal[1,3*] Kartik Sharma[1,4] Tejan Karmali[2]
Varun Jampani[2] R. Venkatesh Babu[1]
[1] Vision and AI Lab, IISc Bangalore [2] Google Research [3] IIT BHU Varanasi [4] BITS Pilani

Abstract

StyleGANs are at the forefront of controllable image generation as they produce a latent space that is semantically disentangled, making it suitable for image editing and manipulation. However, the performance of StyleGANs severely degrades when trained via class-conditioning on large-scale long-tailed datasets. We find that one reason for degradation is the collapse of latents for each class in the $\mathcal{W}$ latent space. With NoisyTwins, we first introduce an effective and inexpensive augmentation strategy for class embeddings, which then decorrelates the latents based on self-supervision in the $\mathcal{W}$ space. This decorrelation mitigates collapse, ensuring that our method preserves intra-class diversity with class-consistency in image generation. We show the effectiveness of our approach on large-scale real-world long-tailed datasets of ImageNet-LT and iNaturalist 2019, where our method outperforms other methods by $\sim 19\%$ on FID, establishing a new state-of-the-art.

1. Introduction

StyleGANs [21, 22] have shown unprecedented success in image generation, particularly on well-curated and articulated datasets (eg. FFHQ for face images, etc.). In addition to generating high fidelity and diverse images, StyleGANs also produce a disentangled latent space, which is extensively used for image editing and manipulation tasks [49]. As a result, StyleGANs are being extensively used in various applications like face-editing [11, 41], video generation [46, 52], face reenactment [3], etc., which are a testament to their usability and generality. However, despite being successful on well-curated datasets, training StyleGANs on in-the-wild and multi-category datasets is still challenging. A large-scale conditional StyleGAN (i.e. StyleGAN-XL) on ImageNet was recently trained successfully by Sauer *et al*. [39] using the ImageNet pre-trained model through the idea of a projection discriminator [38]. While the StyleGAN-XL uses additional pre-trained models, obtaining such models for distinctive image domains like medical, forensics, and fine-grained data may not be feasible, which limits its generalization across domains.

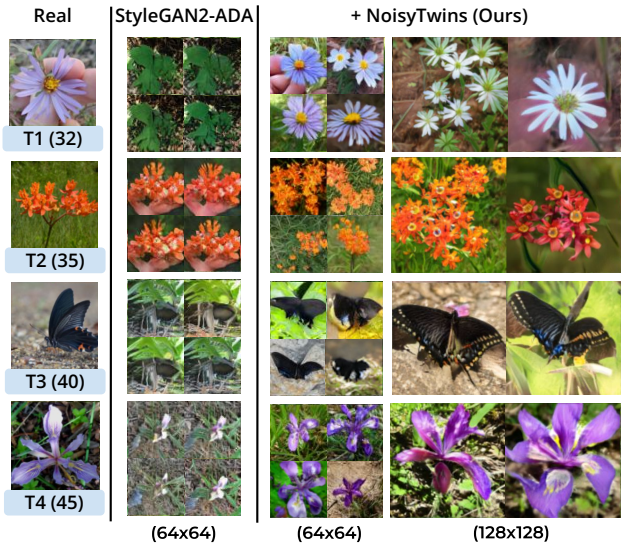

Figure 1. **Qualitative Comparison on tail classes (T1-T4) for iNaturalist 2019.** We provide sample(s) from real class (with class frequency), generated by StyleGAN2-ADA and after adding proposed NoisyTwins. NoisyTwins achieves remarkable diversity, class-consistency and quality by just using 38 samples on average.

In this work, we aim to train vanilla class-conditional StyleGAN without any pre-trained models on challenging real-world long-tailed data distributions. As training StyleGAN with augmentations [20, 54] leads to low recall [24] (which measures diversity in the generated images) and mode collapse, particularly for minority (i.e. tail) classes. For investigating this phenomenon further, we take a closer look at the latent $\mathcal{W}$ space of StyleGAN that is produced by a fully-connected mapping network that takes the conditioning variables $\mathbf{z}$ (i.e. random noise) and class embedding $\mathbf{c}$ as inputs. The vectors $\mathbf{w}$ in $\mathcal{W}$ space are used for conditioning various layers of the generator (Fig. 2). We find that output vectors $\mathbf{w}$ from the mapping network hinge on the conditioning variable $\mathbf{c}$ and become invariant to random conditioning vector $\mathbf{z}$. This collapse of latents leads to un-

*Equal Contribution. Link: rangwani-harsh.github.io/NoisyTwins

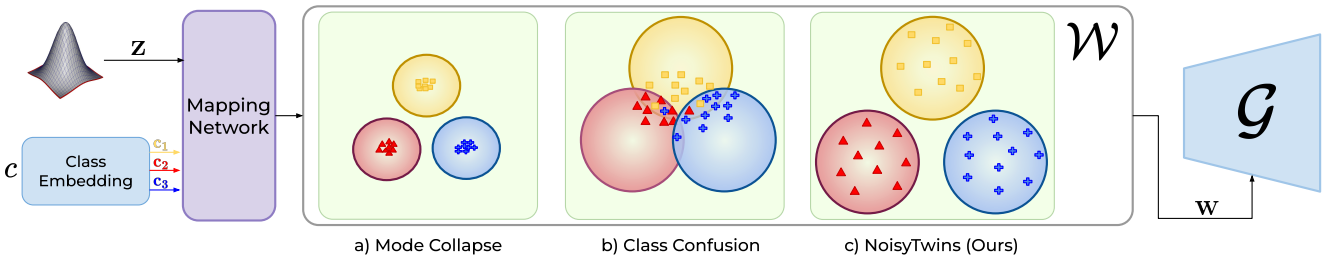

Figure 2. **Schematic illustration of $\mathcal{W}$ space for different GANs.** Existing conditioning methods either suffer from mode collapse [20] or lead to class confusion [35] in $\mathcal{W}$ space. With proposed NoisyTwins, we achieve intra class diversity while avoiding class confusion.

stable training and is one of the causes of poor recall (a.k.a. mode collapse) for minority classes. Further, on augmenting StyleGAN with recent conditioning and regularization techniques [17, 35], we find that they either lead to a poor recall for minority classes or lead to class confusion (Fig. 2) instead of mitigating the collapse.

To mitigate the collapse of $\mathbf{w}$ in $\mathcal{W}$ space, we need to ensure that the change in conditioning variable $\mathbf{z}$ leads to the corresponding change in $\mathbf{w}$. Recently in self-supervised learning, several techniques [2, 53] have been introduced to prevent the collapse of learned representations by maximizing the information content in the feature dimensions. Inspired by them we propose NoisyTwins, in which we first generate inexpensive twin augmentations for class embeddings and then use them to decorrelate the $\mathbf{w}$ variables through self-supervision. The decorrelation ensures that $\mathbf{w}$ vectors are diverse for each class and the GAN is able to produce intra-class diversity among the generated images.

We evaluate NoisyTwins on challenging benchmarks of large-scale long-tailed datasets of ImageNet-LT [30] and iNaturalist 2019 [48]. These benchmarks are particularly challenging due to a large number of classes present, which makes GANs prone to class confusion. On the other hand, as these datasets are long-tailed with only a few images per class in tail classes, generating diverse images for those classes is challenging. We observe that existing metrics used in GAN evaluations are not able to capture both class confusion and mode collapse. As a remedy, we propose to use intra-class Frechet Inception Distance (FID) [12] based on features obtained from pre-trained CLIP [34] embeddings as an effective metric to measure the performance of class-conditional GANs in long-tailed data setups. Using NoisyTwins enables StyleGAN to generate diverse and class-consistent images across classes, mitigating the mode collapse and class confusion issues in existing state-of-the-art (SotA) (Fig. 1). Further, with NoisyTwins, we obtain diverse generations for tail classes even with ≤ 30 images, which can be attributed to the transfer of knowledge from head classes through shared parameters (Fig. 1 and 6). In summary, we make the following contributions:

1. We evaluate various recent SotA GAN conditioning and regularization techniques on the challenging task of long-tailed image generation. We find that all existing methods either suffer from mode collapse or lead to class confusion in generations.
2. To mitigate mode collapse and class confusion, we introduce NoisyTwins, an effective and inexpensive augmentation strategy for class embeddings that decorrelates latents in the $\mathcal{W}$ latent space (Sec. 4).
3. We evaluate NoisyTwins on large-scale long-tailed datasets of ImageNet-LT and iNaturalist-2019, where it consistently improves the StyleGAN2 performance ($\sim 19\%$), achieving a new SotA. Further, our approach can also prevent mode collapse and enhance the performance of few-shot GANs (Sec. 5.3).

2. Related Works

StyleGANs. Karras *et al*. introduced StyleGAN [21] and subsequently improved its image quality in StyleGAN2. StyleGAN could produce high-resolution photorealistic images as demonstrated on a wide variety of category-specific datasets. It introduced a mapping network, which mapped the sampled noise into another latent space, which is more disentangled and semantically coherent, as demonstrated by its downstream usage for image editing and manipulation [1,32,33,42,43]. Further, StyleGAN has been extended to get novel views from images [27, 28, 44], thus making it possible to get 3D information from it. These downstream advances are possible due to the impressive performance of StyleGANs on class-specific datasets (such as faces). However, similar photorealism levels are yet uncommon on multi-class long-tailed datasets (such as ImageNet).

GANs for Data Efficiency and Imbalance. Failure of GANs on less data was concurrently reported by Karras *et al*. [20] and Zhao *et al*. [54]. The problem is rooted in the overfitting of the discriminator due to less real data. Since then, the proposed solutions for this problem have relied on a) augmenting the data, b) introducing regularizers, and c) architectural modifications. Karras *et al*. [20] and Zhao *et al*. [54] relied on differentiable data augmentation before passing images into the discriminator to solve this problem. DeceiveD [15] proposed to introduce label-noise for

discriminator training. LeCamGAN [47] finds that enforcing LeCam divergence as a regularization trick in the discriminator can robustify GAN training under a limited data setting. DynamicD [50] tunes the capacity of the discriminator on-the-fly during training. While these methods can handle the data inefficiency, they are ineffective on class imbalanced long-tailed data distribution [35].

CBGAN [36] proposed a solution to train the unconditional GAN model on long-tailed data distribution by introducing a signal from the classifier to balance the classes generated by GAN. In a long-tailed class-conditional setting, gSR [35] proposes to regularize the exploding spectral norms of the class-specific parameters of the GAN. Collapse-by-conditioning [40] addresses the limited data in classes by introducing a training regime that transitions from an unconditional to a class-conditioned setting, thus exploiting the shared information across classes during the early stages of the training. However, these methods suffer from either class confusion or poor generated image quality on large datasets, which is resolved by NoisyTwins.

Self-Supervised Learning for GANs. Ideas from Self-supervised learning have shown their benefits in GAN training. IC-GAN [6] trains GAN conditioned on embeddings on SwAV [5], which led to remarkable improvement in performance on the long-tailed version of ImageNet. InsGen [51] and ReACGAN [16, 17] introduce the auxiliary task of instance discrimination for the discriminator, thereby making the discriminator focus on multiple tasks and thus alleviating discriminator overfitting. While InsGen relies on both noise space and image space augmentations, ReACGAN and ContraGAN follow only image space augmentations. Contrary to these, NoisyTwins performs augmentations in the class-embedding space and contrasts them in the $\mathcal{W}$-space of the generator instead of the discriminator.

3. Preliminaries

3.1. StyleGAN

StyleGAN [21] is a Generative Adversarial Network comprising of its unique Style Conditioning Based Generator ($\mathcal{G}$) and discriminator network ($\mathcal{D}$) trained jointly. We will focus on the architecture of StyleGAN2 [23] as we use it in our experiments, although our work is generally applicable to all StyleGAN architectures. The StyleGAN2 generator is composed of blocks that progressively upsample the features and resolution inspired by Progressive GAN [19], starting from a single root image. The diversity in the images comes from conditioning each block of image generation through conditioning on the latent coming from the mapping network (Fig. 2). The mapping network is a fully connected network that takes in the conditioning variables, the $\mathbf{z} \in \mathbb{R}^d$ coming from a random distribution (e.g., Gaussian, etc.) and class conditioning label c which is converted to an embedding $\mathbf{c} \in \mathbb{R}^d$. The mapping network takes these and outputs vectors $\mathbf{w}$ in the $\mathcal{W}$ latent space of StyleGAN, which is found to be semantically disentangled to a high extent [49]. The $\mathbf{w}$ is then processed through an affine transform and passed to each generator layer for conditioning the image generation process through Adaptive Instance Normalization (AdaIN) [13]. The images from generator $\mathcal{G}$, along with real images, are passed to discriminator $\mathcal{D}$ for training. The training utilizes the non-saturating adversarial losses [9] for $\mathcal{G}$ and $\mathcal{D}$ given as:

$$\min_{\mathcal{D}} \mathcal{L}_{\mathcal{D}} = \sum_{i=1}^{m} \log(\mathcal{D}(\mathbf{x}_i)) + \log(1 - \mathcal{D}(\mathcal{G}(\mathbf{z}_i, \mathbf{c}_i))) \quad (1)$$

$$\min_{\mathcal{G}} \mathcal{L}_{\mathcal{G}} = \sum_{i=1}^{m} -\log(\mathcal{D}(\mathcal{G}(\mathbf{z}_i, \mathbf{c}_i))) \quad (2)$$

We now describe the issues present in the StyleGANs trained on long-tailed data and their analysis in $\mathcal{W}$ space.

3.2. Class Confusion and Class-Specific Mode Collapse in Conditional StyleGANs

To finely analyze the performance of StyleGAN and its variants on long-tailed datasets, we train them on the CIFAR10-LT dataset. In Fig. 3, we plot the qualitative results of generated images and create a t-SNE plot for latents in $\mathcal{W}$ space for each class. We first train the StyleGAN2 baseline with augmentations (DiffAug) [20, 54]. We find that it leads to mode collapse, specifically for tail classes (Fig. 3). In conjunction with images, we also observe that corresponding t-SNE embeddings are also collapsed near each class's mean in $\mathcal{W}$ space. Further, recent methods which have proposed the usage of contrastive learning for GANs, improve their data efficiency and prevent discriminator overfitting [14, 16]. We also evaluate them by adding the contrastive conditioning method, which is D2D-CE loss-based on ReACGAN [17], to the baseline, where in results, we observe that the network omits to learn tail classes and produces head class images at their place (i.e., class confusion). In Fig. 3, it can be seen that the network confuses semantically similar classes, that is, generating cars (head or majority class) in place of trucks and airplanes (head class) instead of ships. In the $\mathcal{W}$ space, we find the same number of clusters as the number of classes in the dataset. However, the tail label cluster images also belong to the head classes of cars and airplanes. In a very recent work gSR [35], it has been shown that constraining the spectral norm of $\mathcal{G}$ embedding parameters can help reduce the mode collapse and lead to stable training. However, we find that constraining the embeddings leads to class confusion, as seen in t-SNE visualization in Fig. 3. We find that this class confusion gets further aggravated when StyleGAN is trained on datasets like ImageNet-LT, which contain a large number of classes, along with a bunch of semantically similar classes (Sec. 5.2). Based on our $\mathcal{W}$

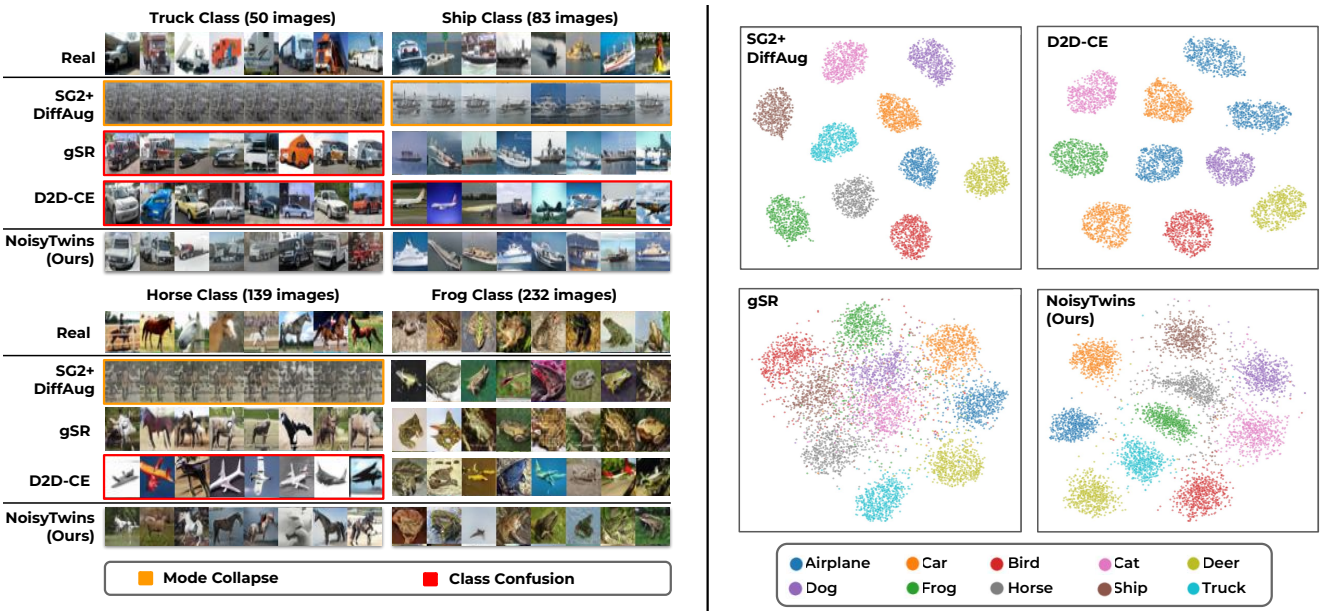

Figure 3. **Comparison of GANs and their $\mathcal{W}$ space for CIFAR10-LT.** We plot the generated images on (*left*) and generate a t-SNE plot of $\mathbf{w}$ latents for generated images in $\mathcal{W}$ space (*right*). We find that mode collapse and class confusion in images is linked to the corresponding collapse and confusion in latent $\mathcal{W}$ space. Our proposed NoisyTwins mitigates both collapse (*left*) and confusion (*right*) simultaneously.

space analysis and qualitative results above, we observe that the class confusion and mode collapse of images is tightly coupled with the structure of $\mathcal{W}$ space. Further, the recent SotA methods are either unable to prevent collapse or suffer from class confusion. Hence, this work aims to develop a technique that mitigates both confusion and collapse.

4. Approach

In this section, we present our method NoisyTwins, which introduces noise-based augmentation twins in the conditional embedding space (Sec. 4.1), and then combines it with the Barlow-Twins-based regularizer from the self-supervised learning (SSL) paradigm to resolve the issue of class confusion and mode collapse (Sec. 4.2).

4.1. Noise Augmentation in Embedding Space

As we observed in the previous section that $\mathbf{w}$ vectors for each sample become insensitive to changes in $\mathbf{z}$. This collapse in $\mathbf{w}$ vectors for each class leads to mode collapse for baselines (Fig. 3). One reason for this could be the fact that $\mathbf{z}$ is composed of continuous variables, whereas the embedding vectors $\mathbf{c}$ for each class are discrete. Due to this, the GAN converges to the easy degenerate solution where it generates a single sample for each class, becoming insensitive to changes in $\mathbf{z}$. For inducing some continuity in $\mathbf{c}$ embeddings vectors, we introduce an augmentation strategy where we add i.i.d. noise of small magnitude in each of the variables in $\mathbf{c}$. Based on our observation (Fig. 3) and existing works [35], there is a high tendency for mode collapse in tail classes. Hence we add noise in embedding space that is proportional to the inverse of the frequency of samples . We provide the mathematical expression of noise augmentation $\tilde{\mathbf{c}}$ below:

$$\tilde{\mathbf{c}} \sim \mathbf{c} + \mathcal{N}(\mathbf{0}, \sigma_c \mathbb{I}_d) \text{ where } \sigma_c = \sigma \frac{(1-\alpha)}{1-\alpha^{n_c}} \tag{3}$$

Here n_c is the frequency of training samples in class c, $\mathbb{I}$ is the identity matrix of size $d \times d$, and α, σ are hyper-parameters. The expression of σ_c is from the effective number of samples [8], which is a softer version of inverse frequency proportionality. In contrast to the image space augmentation, these *noise augmentations come for free* as there is no significant additional computation overhead. This noise is added in the embedding $\mathbf{c}$ before passing it to the generator and the discriminator, which ensures that the class embeddings occupy a continuous region in latent space.

Insight: The augmentation equation above (Eq. 3) can be interpreted as approximating the discrete random variable $\mathbf{c}$ with a Gaussian with finite variance and the embedding parameters $\mathbf{c}$ being the mean $\mu_{\mathbf{c}}$.

$$\tilde{\mathbf{c}} \sim \mathcal{N}(\mu_{\mathbf{c}}, \sigma_c \mathbb{I}_d) \tag{4}$$

This leads to the class-embedding input $\tilde{\mathbf{c}}$ to mapping network to have a Gaussian distribution, similar in nature to $\mathbf{z}$. This noise augmentation strategy alone mitigates the degenerate solution of class-wise mode collapse to a great extent (Table 1) and helps generate diverse latent $\mathbf{w}$ for each class. Due to the diverse $\mathbf{w}$ conditioning of the GAN, it leads to diverse image generation.

4.2. Invariance in $\mathcal{W}$-Space with NoisyTwins

The augmentation strategy introduced in the previous section expands the region for each class in $\mathcal{W}$ latent space.

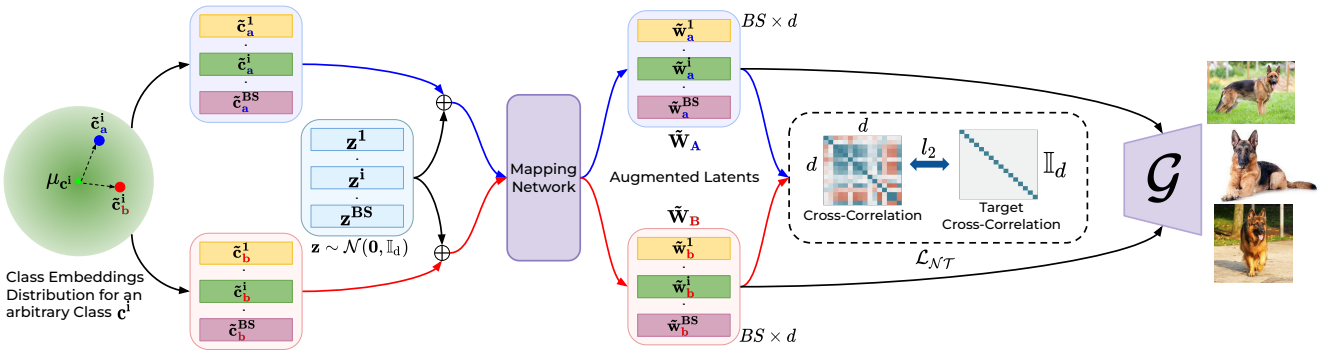

Figure 4. **Overview of NoisyTwins.** For the i^{th} sample of class c^i, we create twin augmentations ($\tilde{\mathbf{c}}_a^i$, $\tilde{\mathbf{c}}_b^i$), by sampling from a Gaussian centered at class embedding ($\mu_{\mathbf{c}^i}$). After this, we concatenate them with the same $\mathbf{z}^i$ and obtain ($\tilde{\mathbf{w}}_a^i, \tilde{\mathbf{w}}_b^i$) from the mapping network, which we stack in batches of augmented latents ($\tilde{\mathbf{W}}_\mathbf{A}$ and $\tilde{\mathbf{W}}_\mathbf{B}$). The twin ($\tilde{\mathbf{w}}_a^i, \tilde{\mathbf{w}}_b^i$) vectors are then made invariant to augmentations (similar) in the latent space by minimizing cross-correlation [2, 53] between the latents of two augmented batches ($\tilde{\mathbf{W}}_\mathbf{A}$ and $\tilde{\mathbf{W}}_\mathbf{B}$).

Although that does lead to diverse image generation as $\mathbf{w}$ are diverse; however, this does not ensure that these $\mathbf{w}$ will generate class-consistent outputs for augmentations in embedding ($\tilde{\mathbf{c}}$). To ensure class consistent predictions, we need to ensure invariance in $\mathbf{w}$ to noise augmentation.

For enforcing invariance to augmentations, a set of recent works [2, 10, 53] in self-supervised learning make the representations of augmentations similar through regularization. Among them, we focus on Barlow Twins as it does not require a large batch size of samples. Inspired by Barlow twins, we introduce NoisyTwins (Fig. 4), where we generate twin augmentations $\tilde{\mathbf{c}}_a$ and $\tilde{\mathbf{c}}_b$ of the same class embedding (μ_c) and concatenate them to same $\mathbf{z}$. After creating a batch of such inputs, they are passed to the mapping network to get batches of augmented latents ($\tilde{\mathbf{W}}_A$ and $\tilde{\mathbf{W}}_B$). These batches are then used to calculate the cross-correlation matrix of latent variables given as:

$$\mathbf{C}_{j,k} = \frac{\sum\limits_{(\tilde{\mathbf{w}}_a, \tilde{\mathbf{w}}_b) \in (\tilde{\mathbf{W}}_A, \tilde{\mathbf{W}}_B)} \tilde{\mathbf{w}}_a^j \tilde{\mathbf{w}}_b^k}{\sum\limits_{\tilde{\mathbf{w}}_a \in \tilde{\mathbf{W}}_A} \tilde{\mathbf{w}}_a^j \tilde{\mathbf{w}}_a^j \sum\limits_{\tilde{\mathbf{w}}_b \in \tilde{\mathbf{W}}_B} \tilde{\mathbf{w}}_b^k \tilde{\mathbf{w}}_b^k} \tag{5}$$

The matrix $\mathbf{C}$ is a square matrix of size same as of latents $\mathbf{w}$. The final loss based on confusion matrix is given as:

$$\mathcal{L}_{\mathcal{NT}} = \sum_j (1 - \mathbf{C}_{jj}^2) + \gamma \sum_{j \neq k} \mathbf{C}_{j,k}^2 \tag{6}$$

The first term tries to make the two latents ($\tilde{\mathbf{w}}_a$ and $\tilde{\mathbf{w}}_b$) invariant to the noise augmentation applied (i.e. similar), whereas the second term tries to de-correlate the different variables, thus maximizing the information content of the embedding vector [53]. The γ is the hyper-parameter that determines the relative importance of the two terms. This loss is then added to the generator loss term ($\mathcal{L}_\mathcal{G} + \lambda L_{\mathcal{NT}}$) and optimized through backpropagation. The above procedure comprises our proposed method, NoisyTwins (Fig. 4), which we empirically evaluate in the subsequent sections.

5. Experimental Evaluation

5.1. Setup

Datasets: We primarily apply all methods on long-tailed datasets, as GANs trained on them are more prone to class confusion and mode collapse. We first report on the commonly used CIFAR10-LT dataset with an imbalance factor (*i.e.* ratio of most to least frequent class) of 100. To show our approach's scalability and real-world application, we test our method on the challenging ImageNet-LT and iNaturalist 2019 datasets. The ImageNet-LT [30] is a long-tailed variant of the 1000 class ImageNet dataset, with a plethora of semantically similar classes (e.g., Dogs, Birds etc.), making it challenging to avoid class confusion. iNaturalist-2019 [48] is a real-world long-tailed dataset composed of 1010 different variants of species, some of which have fine-grained differences in their appearance. For such fine-grained datasets, ImageNet pre-trained discriminators [38, 39] may not be useful, as augmentations used to train the model makes it invariant to fine-grained changes.

Training Configuration: We use StyleGAN2 architecture for all our experiments. All our experiments are performed using PyTorch-StudioGAN implemented by Kang *et al.* [18], which serves as a base for our framework. We use Path Length regularization (PLR) as it ensures changes in $\mathcal{W}$ space lead to changes in images, using a delayed PLR on ImageNet-LT following [39]. We use a batch size of 128 for all our experiments, with one G step per D step. Unless stated explicitly, we used the general training setup for StyleGANs from [18], including methods like R_1 regularization [31], etc. More details on the exact training configuration for each dataset are provided in Appendix.

Metrics: In this work we use the following metrics for evaluation of our methods:
a) FID: Fréchet Inception Distance [12] is the Wasserstein-2 Distance between the real and validation data. We use a

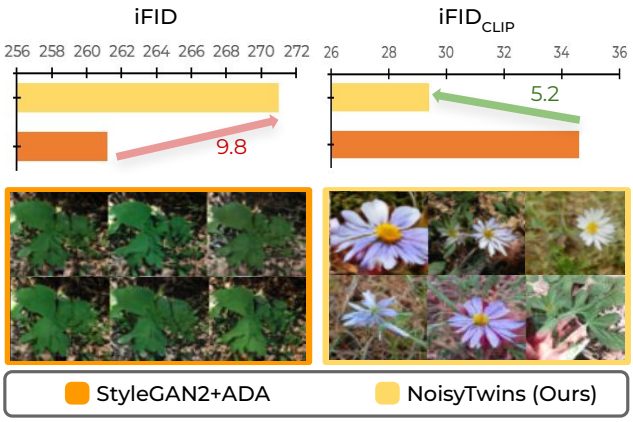

Figure 5. **Choice of Eval. backbone:** intra-FID (iFID) of a class based on InceptionV3 backbone (left plot) is not able to capture the mode collapse (increase in iFID in the absence of mode collapse). This is well-captured by iFID$_{\text{CLIP}}$ based on CLIP [34] backbone (right plot, decrease in iFID in the absence of mode collapse).

held-out validation set and 50k generated samples to evaluate FID in each case. As FID is biased towards ImageNet and can be arbitrarily manipulated, we also report FID$_{\text{CLIP}}$.

b) Precision & Recall: As we aim to mitigate the mode collapse and achieve diverse generations across classes, we use improved Precision & Recall [25] metrics, as poor recall indicates mode collapse [39].

d) Intra-Class FID$_{\text{CLIP}}$ (iFID$_{\text{CLIP}}$): The usage of only FID based on Inception-V3 Networks for evaluation of Generative Models has severe limitations, as it has been found that FID can be reduced easily by some fringe features [26]. iFID is computed by taking FID between 5k generated and real samples for the same class. As we want to evaluate both class consistency and diversity, we find that similar limitations exist for intra-class FID (iFID), which has been used to evaluate class-conditional GANs [18]. In Fig. 5, we show the existence of generated images for a particular class (more in Appendix) from models trained on iNaturalist 2019, where iFID is better for the mode collapsed model than the other model generating diverse images. Whereas the iFID$_{\text{CLIP}}$, based on CLIP backbone can rank the models correctly with the model having mode collapse having high iFID$_{\text{CLIP}}$. Further, we find that the mean iFID can be deceptive in detecting class confusion and collapse cases, as it sometimes ranks models with high realism better than models generating diversity (See Appendix). Hence, mean iFID$_{\text{CLIP}}$ (ref. to as iFID$_{\text{CLIP}}$ in result section for brevity) can be reliably used to evaluate models for class consistency and diversity.

Baselines: For evaluating NoisyTwins performance in comparison to other methods, we use the implementations present in StudioGAN [16]. For fairness, we re-run all the baselines on StyleGAN2 in the same hyperparameter setting. We compare our method to the StyleGAN2 (SG2) and StyleGAN2 with augmentation (DiffAug [54] and ADA [20]) baselines. We further tried to improve the baseline by incorporating the recent LeCam regularization method; however, it resulted in gains only for the iNaturalist 2019 dataset, where we use LeCam for all experiments. Further on StyleGAN2, we also use contrastive D2D-CE loss conditioning (*i.e.* ReACGAN) as a baseline. However, the D2D-CE baseline completely ignores learning of tail classes (Fig. 3) for CIFAR10-LT and is expensive to train; hence we do not report results for it for large-scale long-tailed datasets. We also compare our method against the recent SotA group Spectral Normalization (gSR) [35] method, which we implement for StyleGAN2 by constraining the spectral norms of embedding parameters of the generator ($\mathcal{G}$) as suggested by authors. As a sanity check, we reproduce their results on CIFAR10-LT and find that our implementation matches the reported results correctly. We provide results on all datasets, for both the proposed Noise Augmentation (+ Noise) and the overall proposed NoisyTwins (+NoisyTwins) method.

5.2. Results on Long-Tailed Data Distributions

CIFAR10-LT. We applied DiffAug [54] on all baselines, except on gSR, where we found that DiffAug provides inferior results compared to ADA (as also used by authors [35]). It can be observed in Table 2 that the addition of NoisyTwins regularization significantly improves over baseline by ($\sim$ 14 FID) along with providing superior class consistency as shown by improved iFID$_{\text{CLIP}}$. NoisyTwins is also able to outperform the recent gSR regularization method and achieves improved results for all metrics. Further, NoisyTwins improves FID for StyleGAN2-ADA baseline used by gSR too from 32.08 to 23.02, however the final results are inferior than reported DiffAug baseline results. Further, we observed that despite not producing any tail class images (Fig. 3), the D2D-CE baseline has much superior FID in comparison to baselines. Whereas the proposed iFID$_{\text{CLIP}}$ value is similar for the baseline and D2D-CE model. This clearly demonstrates the superiority of proposed iFID$_{\text{CLIP}}$ in detecting class confusion.

Real-world, Large-scale Long-Tailed Datasets. We experiment with iNaturalist 2019 and ImageNet-LT. These datasets are particularly challenging as they contain long-tailed imbalances and semantically similar classes, making GANs prone to mode collapse and class confusion. The baselines StyleGAN2 and StyleGAN2-ADA both suffer from mode collapse (Fig. 6), particularly for the tail classes. Whereas for the recent SotA gSR method, we find that although it undergoes less collapse in comparison to baselines, it suffers from class confusion as seen from similar Intra-FID$_{\text{CLIP}}$ in comparison to baselines (Table 1). Compared to that, our method NoisyTwins improves when used with StyleGAN2-ADA significantly, leading to

Table 1. **Quantitative results on ImageNet-LT and iNaturalist 2019 Datasets.** We compare FID(↓), FID_{CLIP}(↓), $\text{iFID}_{\text{CLIP}}$(↓), Precision(↑) and Recall(↑) with other existing approaches on StyleGAN2 (SG2). We obtain an average $\sim$ 19% relative improvement on FID, $\sim$ 33% on FID_{CLIP}, and $\sim$ 11% on $\text{iFID}_{\text{CLIP}}$ metrics over the previous SotA on ImageNet-LT and iNaturalist 2019 datasets.

	ImageNet-LT					iNaturalist 2019				
Method	FID(↓)	FID_{CLIP}(↓)	$\text{iFID}_{\text{CLIP}}$(↓)	Precision(↑)	Recall(↑)	FID(↓)	FID_{CLIP}(↓)	$\text{iFID}_{\text{CLIP}}$(↓)	Precision(↑)	Recall(↑)
SG2 [23]	41.25	11.64	46.93	0.50	0.48	19.34	3.33	38.24	0.74	0.17
SG2+ADA [20]	37.20	11.04	47.41	0.54	0.38	14.92	2.30	35.19	0.75	0.57
SG2+ADA+gSR [35]	24.78	8.21	44.42	0.63	0.35	15.17	2.06	36.22	0.74	0.46
SG2+ADA+Noise (Ours)	22.17	7.11	41.20	**0.72**	0.33	12.87	1.37	**31.43**	**0.81**	0.63
+ NoisyTwins (Ours)	**21.29**	**6.41**	**39.74**	0.67	**0.49**	**11.46**	**1.14**	31.50	0.79	**0.67**

Table 2. **Quantitative results on CIFAR10-LT Dataset.** We compare with other existing approaches. We obtain $\sim$ 26% relative improvement over the existing methods on FID_{CLIP} and $\text{iFID}_{\text{CLIP}}$ metrics.

Method	FID(↓)	FID_{CLIP}(↓)	$\text{iFID}_{\text{CLIP}}$(↓)	Precision(↑)	Recall(↑)
SG2+DiffAug [54]	31.73	6.27	11.59	0.63	0.35
SG2+D2D-CE [17]	19.97	4.77	11.35	**0.73**	0.42
gSR [35]	22.10	5.54	9.94	0.70	0.29
SG2+DiffAug+Noise (Ours)	28.90	5.26	10.65	0.71	0.38
+ NoisyTwins(Ours)	**17.74**	**3.55**	**7.24**	0.70	**0.51**

Table 3. **Comparison with SotA approaches on BigGAN.** We compare FID(↓) with other existing models on ImageNet-LT (IN-LT) and iNaturalist 2019 (iNat-19).

Method	iNat-19	IN-LT
BigGAN [4]	14.85	28.10
+ gSR [35]	13.95	-
ICGAN [6]	-	23.40
StyleGAN2-ADA [20]	14.92	37.20
+ NoisyTwins (Ours)	**11.46**	**21.29**

a relative improvement of 42.7% in FID for ImageNet-LT and 23.19% on the iNaturalist 2019 dataset when added to StyleGAN2-ADA baseline. Further with Noise Augmentation (+Noise), we observe generations of high-quality class-consistent images, but it also suffers from mode collapse. This can be observed by the high-precision values in comparison to low-recall values. However, adding NoisyTwins regularization over the noise augmentation improves diversity by improving recall (Table 1).

Fig. 6 presents the generated images of tail classes for various methods on ImageNet-LT , where NoisyTwins generations show remarkable diversity in comparison to others. The presence of diversity for classes with just 5-6 training images demonstrates successful transfer of knowledge from head classes to tail classes, due to shared parameters. Further, to compare with existing SotA reported results, we compare FID of BigGAN models from gSR [35] and Instance Conditioned GAN (ICGAN) [6]. For fairness, we compare FID on the validation set for which we obtained gSR models from authors and re-evaluate them, as they reported FID on a balanced training set. As BigGAN models are more common for class-conditioned generation [18], their baseline performs superior to StyleGAN2-ADA baselines (Table 3). However, the addition of NoisyTwins to the StyleGAN2-ADA method improves it significantly, even outperforming the existing methods of gSR (by 18.44%) and ICGAN (by 9.44%) based on BigGAN architecture. This shows that NoisyTwins allows the StyleGAN2 baseline to scale to large and diverse long-tailed datasets.

5.3. NoisyTwins on Few-Shot Datasets

We now demonstrate the potential of NoisyTwins in another challenging scenario of class-conditional few-shot image generation from GANs. We perform our experiments using a conditional StyleGAN2-ADA baseline, for which we tune hyper-parameters to obtain a strong baseline. We then apply our method of Noise Augmentation and NoisyTwins over the strong baseline for reporting our results. We use the few-shot dataset of LHI-AnimalFaces [45] and a subset of ImageNet Carnivores [29, 40] to report our results. Table 4 shows the results of these experiments, where we find that our method, NoisyTwins, significantly improves the FID of StyleGAN2 ADA baseline by (22.2%) on average for both datasets. Further, combining Noisy Twins with SotA Transitional-cGAN [40] through official code, also leads to effective improvement in FID. These results clearly demonstrate the diverse potential and applicability of our proposed method NoisyTwins.

6. Analysis

We perform analysis of NoisyTwins w.r.t. to its hyperparameters, standard deviation (σ) of noise augmentation and regularization strength (λ). We also compare NoisyTwins objective (ref. Eq. 6) with contrastive objective. Finally, we compare NoisyTwins over Latent Diffusion Models for long-tailed class conditional generation task. We perform ablation experiments on CIFAR10-LT, for which additional details and results are present in Appendix. We also present comparison of NoisyTwins for GAN fine-tuning.

How much noise and regularization strength is optimal? In Fig. 7, we ablate over the noise variance parameter σ for CIFAR10-LT. We find that a moderate value of noise strength 0.75 leads to optimal results. For the strength of NoisyTwins loss (λ), we find that the algorithm performs similarly on values near 0.01 and is robust to it (Fig. 7).

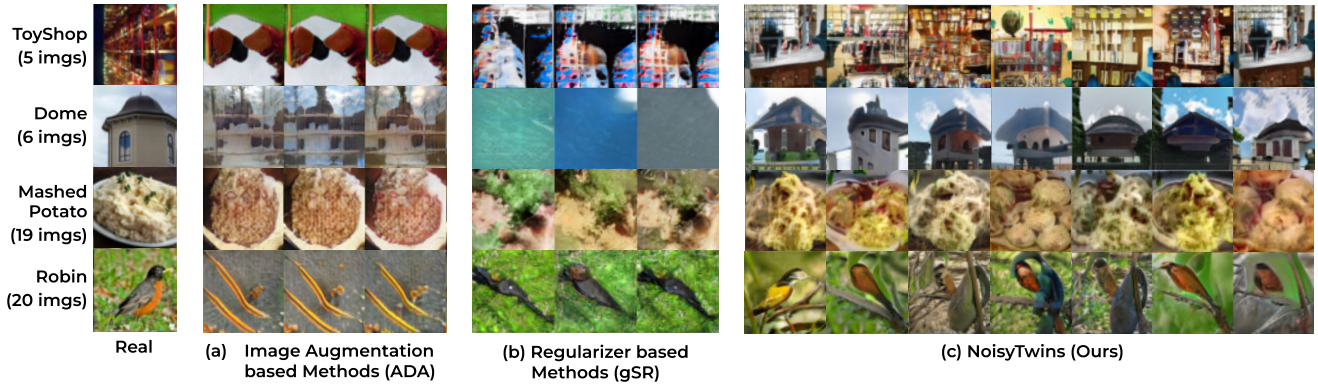

Figure 6. **Qualitative results on ImageNet-LT for tail classes.** We find that existing SotA methods for tail classes show collapsed (a) or arbitrary image generation (b). With NoisyTwins, we observe diverse and class-consistent image generation, even for classes having 5-6 images. The tail classes get enhanced diversity by transferring the knowledge from head classes, as they share parameters.

Table 4. **Quantitative results on ImageNet Carnivore and AnimalFace Datasets.** Our method improves over both StyleGAN2-ADA (SG2-ADA) baseline and SotA Transitional-cGAN .

	ImageNet Carnivore		AnimalFace	
Method	FID(↓)	$\text{iFID}_{\text{CLIP}}$(↓)	FID(↓)	$\text{iFID}_{\text{CLIP}}$(↓)
SG2 [23]	111.83	36.34	94.09	29.94
SG2+ADA [20]	22.77	12.85	20.25	11.12
SG2+ADA+Noise (Ours)	19.25	12.51	18.78	10.42
+ NoisyTwins (Ours)	**16.01**	**12.41**	**17.27**	**10.03**
	FID(↓)		FID(↓)	
Transitional-cGAN [40]	14.60		20.53	
+ NoisyTwins (Ours)	**13.65**		**16.15**	

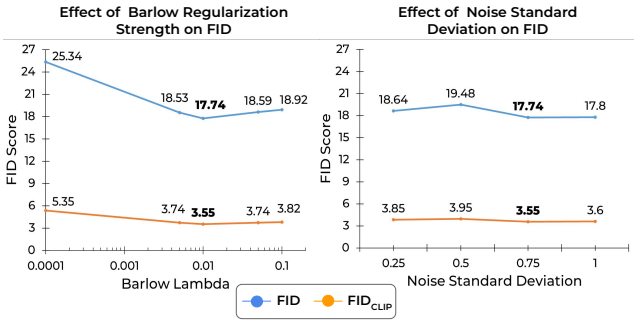

Figure 7. **Ablation of Hyperparameters.** Quantitative comparison on CIFAR10-LT for standard deviation of Noise Augmentation (σ) and strength (λ) of NoisyTwins loss.

Which type of self-supervision to use with noise augmentation? The goal of our method is to achieve invariance to Noise Augmentation in the $\mathcal{W}$ latent space. This can be achieved using either contrastive learning-based methods like SimCLR [7] or negative-free method like Barlow Twins [53]. Contrastive loss (SimCLR based) produces FID of 26.23 vs 17.74 by NoisyTwins (BarlowTwins based). We find that contrastive baseline improves over the noise augmentation baseline (28.90) however falls significantly below the NoisyTwins, as the former requires a large batch size to be effective which is expensive for GANs.

How does NoisyTwins compare with modern Vision and Language models? For evaluating the effectiveness of modern vision language-based diffusion models, we test the generation of the iNaturalist 2019 dataset by creating the prompt "a photo of `S`" where we replace the class name in place of `S`. We use the LDM [37] model trained on LAION-400M to perform inference, generating 50 images per class. We obtained an FID of 57.04 in comparison to best FID of 11.46 achieved by NoisyTwins. This clearly demonstrates that for specific use cases like fine-grained generation, GANs are still ahead of general-purpose LDM.

7. Conclusion

In this work, we analyze the performance of StyleGAN2 models on the real-world long-tailed datasets including iNaturalist 2019 and ImageNet-LT. We find that existing works lead to either class confusion or mode collapse in the image space. This phenomenon is rooted in collapse and confusion in the latent $\mathcal{W}$ space of StyleGAN2. Through our analysis, we deduce that this collapse occurs when the latents become invariant to random conditioning vectors $\mathbf{z}$, and collapse for each class. To mitigate this, we introduce inexpensive noise based augmentation for discrete class embeddings. Further, to ensure class consistency, we couple this augmentation technique with BarlowTwins' objective in the latent $\mathcal{W}$ space which imparts intra-class diversity to latent $\mathbf{w}$ vectors. The noise augmentation and regularization comprises our proposed NoisyTwins technique, which improves the performance of StyleGAN2 establishing a new SotA on iNaturalist 2019 and ImageNet-LT. The extension of NoisyTwins for conditioning on more-complex attributes for StyleGANs is a good direction for future work.

Acknowledgements: This work was supported in part by SERB-STAR Project (STR/2020/000128). Harsh Rangwani is supported by PMRF fellowship.

References

[1] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18511–18521, June 2022. 2

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. 2, 5

[3] Stella Bounareli, Vasileios Argyriou, and Georgios Tzimiropoulos. Finding directions in gan's latent space for neural face reenactment. *arXiv preprint arXiv:2202.00046*, 2022. 1

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 7

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. 2020. 3

[6] Arantxa Casanova, Marlène Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero-Soriano. Instance-conditioned gan. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3, 7

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 8

[8] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 4

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 5

[11] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020. 1

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2, 5

[13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 3

[14] Jongheon Jeong and Jinwoo Shin. Training GANs with stronger augmentations via contrastive discriminator. In *International Conference on Learning Representations*, 2021. 3

[15] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Deceive D: Adaptive Pseudo Augmentation for GAN training with limited data. In *NeurIPS*, 2021. 2

[16] Minguk Kang and Jaesik Park. ContraGAN: Contrastive Learning for Conditional Image Generation. 2020. 3, 6

[17] Minguk Kang, Woohyeon Shim, Minsu Cho, and Jaesik Park. Rebooting acgan: Auxiliary classifier gans with stable training. *Advances in neural information processing systems*, 34:23505–23518, 2021. 2, 3, 7

[18] MinGuk Kang, Joonghyuk Shin, and Jaesik Park. StudioGAN: A Taxonomy and Benchmark of GANs for Image Synthesis. *2206.09479 (arXiv)*, 2022. 5, 6, 7

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 3

[20] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020. 1, 2, 3, 6, 7, 8

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 2, 3

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 3, 7, 8

[24] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. *Improved Precision and Recall Metric for Assessing Generative Models*. 2019. 1

[25] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[26] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. *CoRR*, abs/2203.06026, 2022. 6

[27] Thomas Leimkühler and George Drettakis. Freestylegan: Free-view editable portrait rendering with the camera manifold. 40(6), 2021. 2

[28] Feng Liu and Xiaoming Liu. 2d gans meet unsupervised single-view 3d reconstruction. In *Computer Vision – ECCV 2022*, pages 497–514, 2022. 2

[29] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of*

the IEEE/CVF international conference on computer vision, pages 10551–10560, 2019. 7

[30] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 2, 5

[31] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018. 5

[32] Rishubh Parihar, Ankit Dhiman, Tejan Karmali, and R. Venkatesh Babu. Everything is there in latent space: Attribute editing and attribute style manipulation by stylegan latent space exploration. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1828–1836, 2022. 2

[33] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021. 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 6

[35] Harsh Rangwani, Naman Jaswani, Tejan Karmali, Varun Jampani, and R. Venkatesh Babu. Improving gans for long-tailed data through group spectral regularization. In *European Conference on Computer Vision*, 2022. 2, 3, 4, 6, 7

[36] Harsh Rangwani, Konda Reddy Mopuri, and R Venkatesh Babu. Class balancing gan with a classifier in the loop. In *Uncertainty in Artificial Intelligence*, pages 1618–1627. PMLR, 2021. 3

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 8

[38] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 5

[39] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. volume abs/2201.00273, 2022. 1, 5, 6

[40] Mohamad Shahbazi, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Collapse by conditioning: Training class-conditional GANs with limited data. In *International Conference on Learning Representations*, 2022. 3, 7, 8

[41] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 1

[42] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE TPAMI*, 2020. 2

[43] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 2

[44] Yichun Shi, Divyansh Aggarwal, and Anil K. Jain. Lifting 2d stylegan for 3d-aware face generation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6254–6262, 2021. 2

[45] Zhangzhang Si and Song-Chun Zhu. Learning hybrid image templates (hit) by information projection. *IEEE Transactions on pattern analysis and machine intelligence*, 34(7):1354–1367, 2011. 7

[46] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2, 2021. 1

[47] Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularing generative adversarial networks under limited data. In *CVPR*, 2021. 3

[48] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 2, 5

[49] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 1, 3

[50] Ceyuan Yang, Yujun Shen, Yinghao Xu, Deli Zhao, Bo Dai, and Bolei Zhou. Improving gans with a dynamic discriminator. 2022. 3

[51] Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. Data-efficient instance generation from instance discrimination. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 3

[52] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations*, 2022. 1

[53] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 2, 5, 8

[54] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 3, 6, 7