NoisyTwins: Class-Consistent and Diverse Image Generation through StyleGANs Supplementary Document (Appendix)

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

Organization of Appendix

A. Notations and Code

We summarize the notations used throughout the paper in Table A.1. We provide PyTorch-style pseudo code for NoisyTwins in `noisy_twins.py` in the supplementary material. We will open-source our code to promote reproducible research.

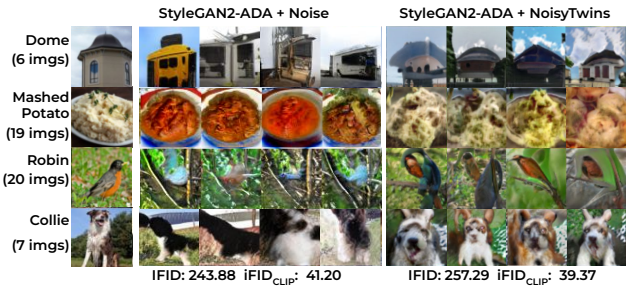

Figure A.1. **Qualitative Results and iFID.** We observe that the noise-only baseline suffers from the mode collapse and class confusion for tail categories as shown on (*left*). Despite this, it is found that the mean iFID based on Inception V3 shows a smaller value for StyleGAN2ADA+Noise, whereas a higher value for diverse and class-consistent NoisyTwins. Hence, this metric does not align with qualitative results. On the other hand, the proposed mean iFID$_{\text{CLIP}}$ is lower for NoisyTwins, demonstrating its reliability for evaluating GAN models.

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

Table A.1. **Notation Table**

Symbol	Space	Meaning
$\mathbf{c}$	$\mathbb{R}^d$	Class Embedding
$\mathbf{z}$	$\mathbb{R}^d$	Noise vector
$\mathbf{w}$	$\mathbb{R}^d$	Vector in $\mathcal{W}$ latent Space
$\mathcal{D}$		Discriminator
$\mathcal{G}$		Generator
BS	$\mathbb{R}^+$	Batch Size
$\mathbf{x_i}$	$\mathbb{R}^{3\times H\times W}$	Image
$\tilde{\mathbf{c}}$	$\mathbb{R}^d$	Noise Augmented Class Embedding
n_c	$\mathbb{R}^+$	Frequency of training samples in class $\mathbf{c}$
σ_c	$\mathbb{R}^+$	Effective number of samples based noise standard deviation
σ	$\mathbb{R}^+$	Hyperparameter for scaling noise
$\mu_{\mathbf{c}}$	$\mathbb{R}^d$	Mean embedding parameters of class $\mathbf{c}$
$\tilde{\mathbf{W}}_A$ $\tilde{\mathbf{W}}_B$	$\mathbb{R}^{BS\times d}$	Batches of augmented latents
$\mathbf{C}_{j,k}$	$\mathbb{R}$	Cross-correlation between jth and kth latent variables
λ	$\mathbb{R}^+$	Strength of NoisyTwins regularization
γ	$\mathbb{R}^+$	Relative importance of the two terms of NoisyTwins loss
ρ	$\mathbb{R}^+$	Imbalance ratio of dataset: Ratio between the most and the least frequent classes

B. Comparison of iFID and iFID$_{\text{CLIP}}$

In this section, we present failure cases of InceptionV3-based iFID in the detection of mode collapse, and show how CLIP-based iFID can detect these cases. InceptionV3-based iFID assigns a lower value to a generator with mode collapse, compared to another generator which creates diverse and class-consistent images. In addition to the example given in the main text (Fig. 5), we provide examples from three different classes (Fig. B.2). In all the four cases, the InceptionV3-based iFID is better for mode collapsed classes. *Whereas iFID$_{\text{CLIP}}$ follows the correct behavior, where the class consistent and diverse model is ranked better.* Due to this inconsistent behavior, mean iFID (mean across classes) which is a commonly used as a metric for quantifying class confusion [1] can be incorrect.

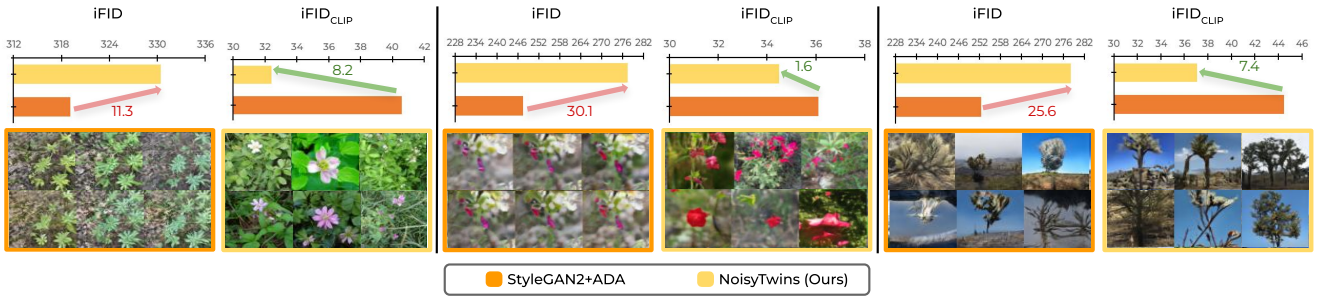

Figure B.2. **iFID Comparison on iNaturalist 2019 dataset.** We provide examples of classes where the quality of images generated by StyleGAN2-ADA is worse, which either suffers from mode collapse or artifacts in generation. Yet iFID based on Inception V3 ranks it higher in terms of quality, which doesn't align with human judgement. On the other hand the proposed iFID$_{\text{CLIP}}$ is able to rank the models correctly and gives a lower value to diverse generations from NoisyTwins.

For example, we observe that the StyleGAN2-ADA baseline with proposed noise augmentation achieves mean iFID (243.88) on ImageNet-LT, compared to 257.29 for the NoisyTwins model (Table 1 in main text). However, while examining the tail class samples (Fig. A.1), we find that noise augmented baseline suffers from mode collapse and class confusion, whereas NoisyTwins generates diverse and class-consistent images. Hence, the mean iFID based on Inception-V3 does not align well with qualitative results. On the contrary, the iFID$_{\text{CLIP}}$ value is 41.20 for the noise-augmented model compared to 39.37 for NoisyTwins, which correlates with the human observation that the NoisyTwins model should have a lower FID as it is diverse and class-consistent. Hence, the proposed metric iFID$_{\text{CLIP}}$ can be used to to evaluate models for class-conditional image generation reliably.

C. Experimental Details

We run our experiments using PyTorchStudioGAN [2] as the base framework. For most baseline experiments, we use the standard StyleGAN configurations present in the framework. We use a learning rate of 0.0025 for the discriminator ($\mathcal{D}$) and the generator ($\mathcal{G}$) network. We use a batch size of 128 for all our experiments. In addition, following the observations of previous work [6], we apply a delayed Path Length Regularization (PLR) starting at 60k iterations for all our experiments on ImageNet-LT. For NoisyTwins, the most important hyperparameters are λ (regularization strength) and σ (noise variance). We perform a grid search on λ values of $\{0, 0.001, 0.01, 0.1\}$ and σ values of $\{0.10, 0.25, 0.50, 0.75\}$. We provide a detailed list of optimal hyperparameters used in Table C.2. All the models trained on a particular dataset use the same hyperparameters, to maintain fairness in the comparison of models. We summarize all the hyperparameters used for respective datasets in Table C.2.

For our experiments on few-shot datasets with SotA transitional-cGAN, we use the author's official code implementation available on GitHub [1]. We use the same configuration specified to first evaluate on ImageNet Carnivores and AnimalFaces datasets. To integrate NoisyTwins, we generate the noise augmentations by augmenting the class embeddings and then apply NoisyTwins regularization in $\mathcal{W}$ space. We use the same hyperparameter setting used by the authors and NoisyTwins with $\lambda = 0.001$ and $\gamma = 0.05$.

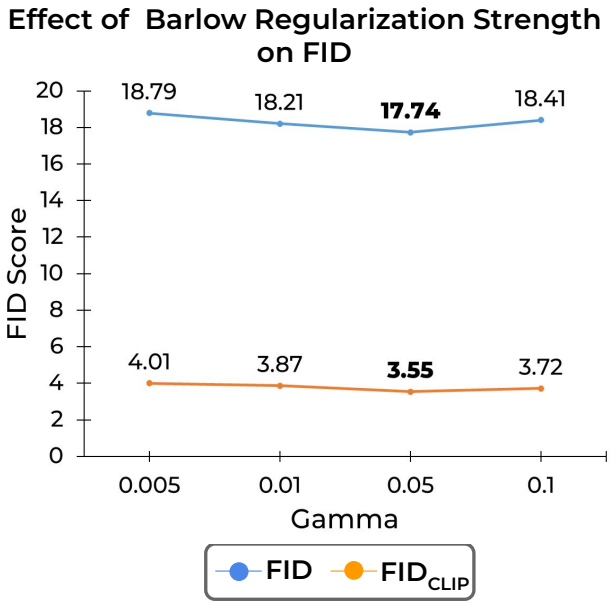

Figure C.3. **Ablation on γ:** Quantitative comparison on CIFAR10-LT for the strength of hyperparameter (γ) in NoisyTwins loss function.

C.1. Statistical Significance of the Experiments

We report mean and standard deviation over three evaluation runs for all baselines on the CIFAR10-LT (Table C.3).

[1] https://github.com/mshahbazi72/transitional-cGAN

Table C.2. **HyperParameter Configurations used for experiments.** We provide a detailed list of hyperparameters used for the experiments across datasets for NoisyTwins on StyleGANs.

	Long-Tail Datasets			Few-Shot Datasets	
	iNaturalist-2019	ImageNet-LT	CIFAR10-LT (ρ=100)	ImageNet Carnivores	AnimalFaces
Resolution	64	64	32	64	64
Augmentation	ADA	ADA	DiffAug	ADA	ADA
Regularizers					
Effective Samples α	0	0	0.99	0	0
Noise Scaling σ	0.1	0.25	0.75	0.5	0.5
NoisyTwins Start Iter.	25k	60k	0	0	0
NoisyTwins Weights (λ, γ)	0.001, 0.005	0.001, 0.005	0.01, 0.05	0.001, 0.05	0.001, 0.05
LeCam Reg Weight	0.01	0	0	0	0
R1 Regularization γ_{R1}	0.2048	0.2048	0.01	0.01	0.01
PLR Start Iter.	0	60k	No PLR	0	0
StyleGAN					
Mapping Net Layers	2	8	8	2	2
$\mathcal{D}$ Backbone	ResNet	ResNet	Orig	ResNet	ResNet
Style Mixing	0.9	0.9	0	0	0
$\mathcal{G}$ EMA Rampup	None	None	0.05	0.05	0.05
$\mathcal{G}$ EMA Kimg	20	20	500	500	500
MiniBatch Group	8	8	32	32	32

Table C.3. **Statistical Analysis for CIFAR10-LT.** This table provides the mean and one standard deviation of metrics for all methods on CIFAR10-LT performed on three independent evaluation runs by generating 50k samples across random seeds.

	CIFAR10-LT (ρ=100)				
Method	FID(↓)	FID_{CLIP}(↓)	$\text{iFID}_{\text{CLIP}}$(↓)	Precision(↑)	Recall(↑)
SG2+DiffAug [3]	$31.72_{\pm 0.16}$	$6.24_{\pm 0.02}$	$11.63_{\pm 0.03}$	$0.63_{\pm 0.00}$	$0.35_{\pm 0.00}$
SG2+D2D-CE [1]	$20.08_{\pm 0.15}$	$4.75_{\pm 0.04}$	$11.35_{\pm 0.01}$	$\mathbf{0.73}_{\pm 0.00}$	$0.43_{\pm 0.00}$
gSR [5]	$22.50_{\pm 0.29}$	$5.55_{\pm 0.01}$	$9.94_{\pm 0.00}$	$0.70_{\pm 0.00}$	$0.28_{\pm 0.01}$
SG2+DiffAug+Noise (Ours)	$28.85_{\pm 0.18}$	$5.29_{\pm 0.02}$	$10.64_{\pm 0.01}$	$0.71_{\pm 0.00}$	$0.38_{\pm 0.00}$
+ NoisyTwins (Ours)	$\mathbf{17.72}_{\pm 0.08}$	$\mathbf{3.56}_{\pm 0.01}$	$\mathbf{7.27}_{\pm 0.02}$	$0.69_{\pm 0.01}$	$\mathbf{0.52}_{\pm 0.01}$

It can be observed that most metrics that we have reported have a low standard deviation, and metrics are close to the mean value across runs. As we find standard deviation to be low across the metrics evaluated and the process of evaluating iFID to be expensive, we do not explicitly report them on large multi-class datasets.

D. Additional Details of Analysis

We perform our ablation experiments on CIFAR10-LT using the same configuration as mentioned in Table C.2. We provide ablation experiments on the standard deviation of noise (σ) and the strength of regularization loss (λ) (Sec. 6), as we observe that they influence the performance of the system most. We further provide ablation on the parameter γ in Fig. C.3, which controls the relative importance

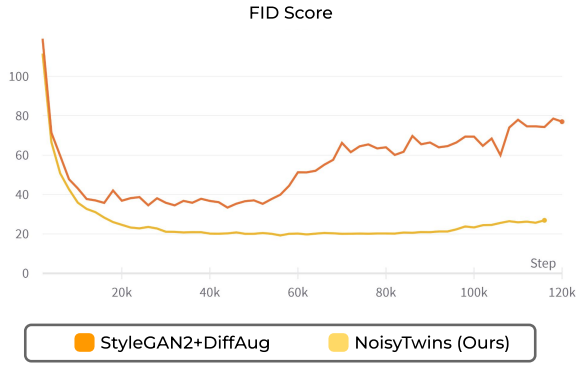

Figure C.4. **Comparison of FID curves for CIFAR10-LT (ρ=100).** NoisyTwins leads to stable training with decreasing FID with iterations.

Table C.4. **Evaluation of NoisyTwins by varying degree of imbalance.** NoisyTwins can produce diverse and class-consistent results across imbalance ratios.

		CIFAR10-LT				
Method	ρ	FID(↓)	FID_{CLIP}(↓)	$\text{iFID}_{\text{CLIP}}$(↓)	Precision(↑)	Recall(↑)
SG2+DiffAug [3]	50	26.79	5.83	9.61	0.65	0.38
+NoisyTwins (Ours)		**14.92**	**2.99**	**6.38**	**0.71**	**0.57**
SG2+DiffAug [3]	100	31.73	6.27	11.59	0.63	0.35
+NoisyTwins (Ours)		**17.74**	**3.55**	**7.24**	**0.70**	**0.51**
SG2+DiffAug [3]	200	55.48	10.59	19.49	0.65	0.36
+NoisyTwins (Ours)		**23.57**	**4.91**	**9.17**	**0.68**	**0.46**

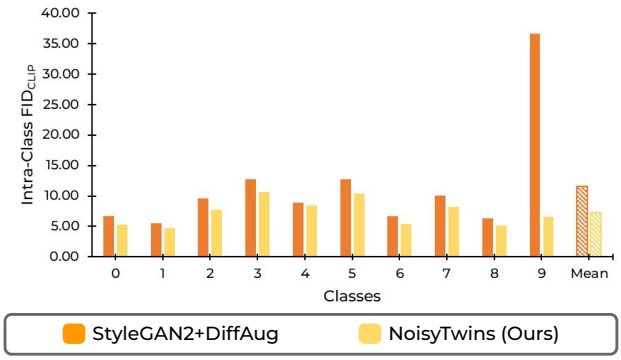

Figure E.5. **Class-wise $\text{iFID}_{\text{CLIP}}$** comparison of models on CIFAR10-LT (ρ=100) dataset.

between the invariance enforcement and decorrelation enhancement terms in Eq. 6 of the main text. We find that performance remains almost the same while varying γ from 0.005 to 0.1, with optimal value occurring around 0.05 for CIFAR10-LT. Hence, the model is robust to γ.

We further analyze our method for a range of imbalance ratios (i.e., ρ, ratio of the most frequent to least frequent class) in the class distribution. We present results for CIFAR10-LT with imbalance factors (ρ) values of 50, 100, and 200 in Table C.4. Our method can prevent mode collapse and improves the baseline FID significantly in all cases. Also note that the baseline gets more unstable (high FID) as the imbalance ratio increases, which shows the necessity of using NoisyTwins as it stabilizes the training even when large imbalances are present in the dataset (Fig. C.4).

E. Additional Results

Fig. E.5 provides the class-wise comparison of the proposed $\text{iFID}_{\text{CLIP}}$ for the baseline and after adding NoisyTwins. NoisyTwins produces better $\text{iFID}_{\text{CLIP}}$ for all classes, hence does not lead to performance degradation for head classes while improving performance on tail classes.

We now provide additional qualitative results for models. Similar to ImageNet-LT, we also provide a full-scale comparison of images from different methods in Fig. E.10 for iNaturalist-2019. In addition to the images from the tail classes, we also show generations from the head and middle classes. In Fig. E.10, it is clearly shown that NoisyTwins can obtain high-quality and diverse samples compared to the baseline. We find that the StyleGAN2-ADA baseline produces similar images across a class for tail classes, which confirms the occurrence of class-wise mode collapse even in large datasets. Further, it can be seen that the regularizer-based method (gSR) is unable to capture the identity of the real class and suffers from the issue of class confusion (as also seen in t-SNE of Fig. 2 of the main text). Our method NoisyTwins, can produce realistic-looking diverse images even for tail classes, which shows the successful transfer of knowledge from head classes. Training a class-conditioned GAN on long-tailed datasets leads to class confusion when the extent of knowledge transfer is not controlled. NoisyTwins strikes the right balance between knowledge transfer from the head classes to benefit the quality of generation in the tail classes, thus not allowing class confusion. This would not be possible if we train GAN independently on tail classes ($\sim$ 30 images), which shows the practical usefulness of joint training on complete long-tailed data (i.e., our setup).

We showcase qualitative results of generations from few-shot datasets (i.e., ImageNet Carnivore and AnimalFaces). Fig. E.6 and E.7 show the results of the SotA few-shot baseline of Transitional-cGAN (*left*) and after augmenting it with our proposed NoisyTwins (*right*). Our proposed method, NoisyTwins, can further stabilize the training of Transitional-cGAN and improve the quality and diversity of the generated samples on both datasets of ImageNet Carnivores and AnimalFaces.

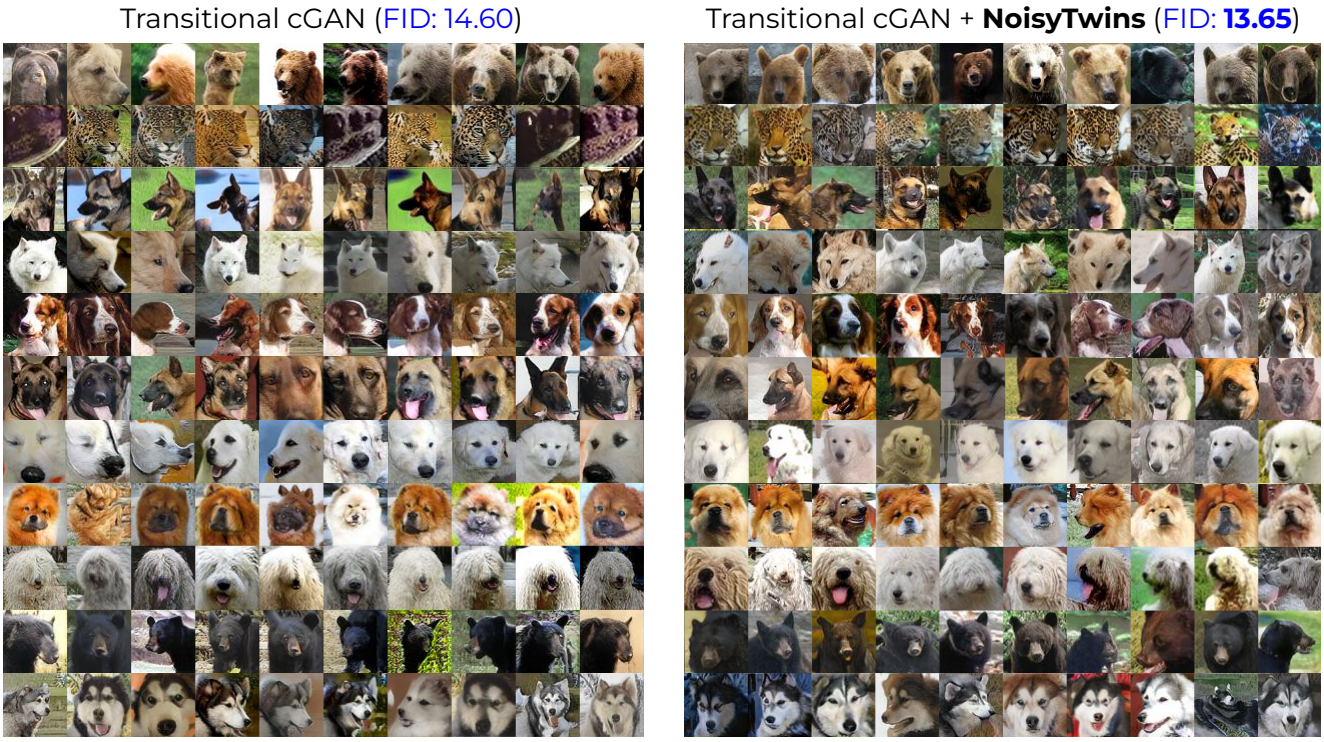

Figure E.6. **Qualitative comparison on few-shot ImageNet Carnivores dataset.**

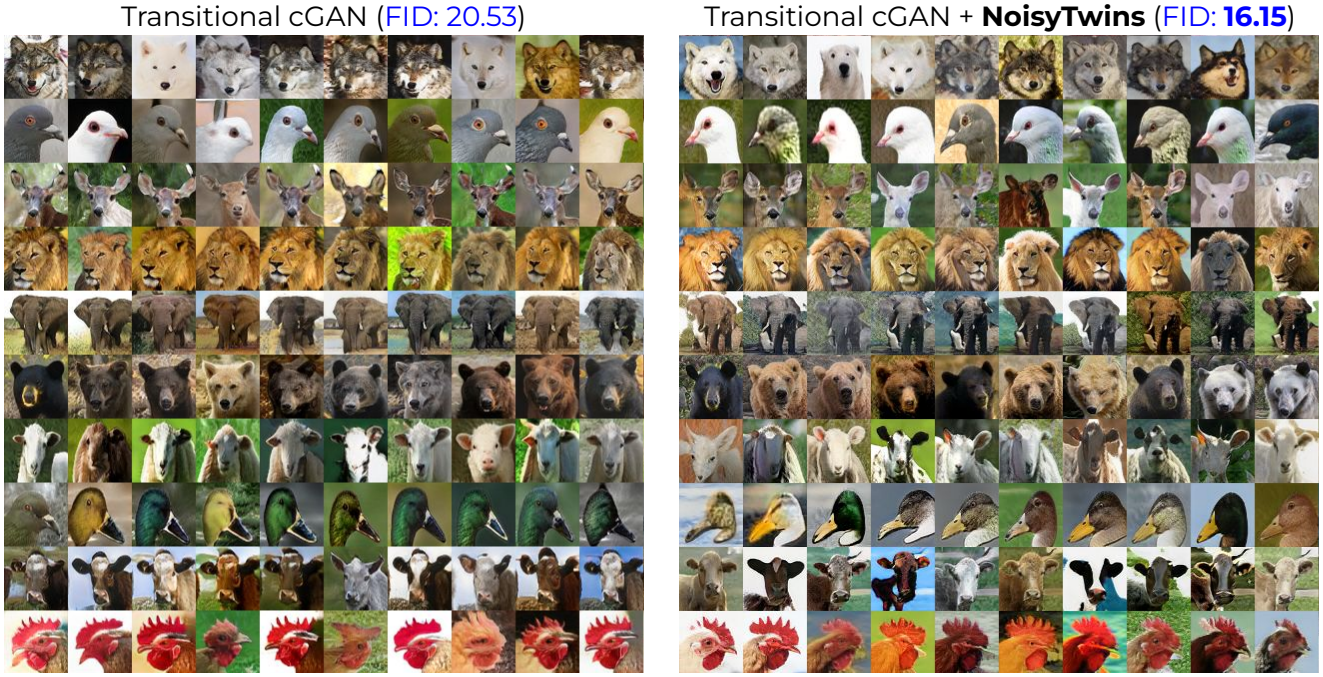

Figure E.7. **Qualitative comparison on few-shot AnimalFaces dataset.**

Results across other Resolutions: NoisyTwins scales well on larger resolutions as demonstrated on few-shot AnimalFaces (AF) dataset using Transitional-cGAN [7] in Table E.5, where we observe a significant improvement if FID

Table E.5. Results for Large Resolutions on Animal Faces dataset

FID (↓)	AF (128 × 128)	AF (256 × 256)
Transitional-cGAN [7]	22.59	22.28
+NoisyTwins (Ours)	**16.79**	**19.14**

Table E.6. Results for large iNaturalist 2019 dataset (128 × 128)

	FID (80k) (↓)	FID (↓)	FID_{CLIP} (↓)
StyleGAN2-ADA [3]	16.58	12.31	2.18
+NoisyTwins (Ours)	**15.29**	**12.01**	**1.93**

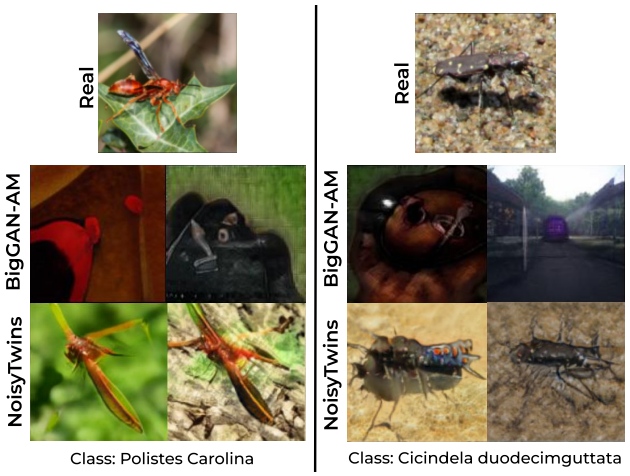

Figure E.8. **BigGAN-AM results on iNaturalist Dataset.**

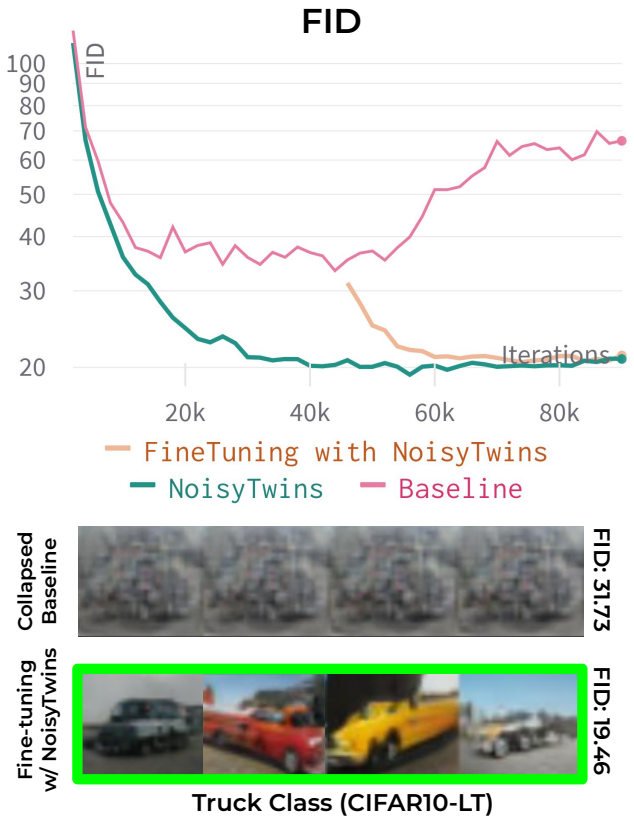

Figure E.9. **Fine-tuning Results.** (Top) FID Curve during fine-tuning with NoisyTwins for CIFAR10-LT dataset. (Below) Diverse images of the truck class generated after fine-tuning baseline with NoisyTwins.

for both 128 × 128 and 64 × 64 resolution data. Further, on large-scale iNat-19 StyleGAN2-ADA baseline in Tab. E.6, we also find that NoisyTwins is able to improve performance. The NoisyTwins method also converges faster as at intermediate stage of 80k iterations in full run of 150k iterations, the FID for NoisyTwins is lower than baseline. As NoisyTwins method is based on the information maximization principle [8] and generalizes on datasets, we expect it benefits other large resolutions of StyleGAN too, similar to what is observed in Sauer *et al*. [6].

Comparison to Fine-Tuning Approaches: We tested NoisyTwins in fine-tuning setting to investigate if it is able to overcome mode collapse. For this we first train StyleGAN-2 DiffAug baseline (Table 2) and then obtain the checkpoint which has collapse, we then resume training of baseline after adding the NoisyTwins regularizer. As seen in Fig. E.9, NoisyTwins is able to reconstruct the collapsed class of baseline on fine-tuning, improving the FID to 19.46 from 31.73 on the CIFAR10-LT dataset.

We also compare our method to other fine-tuning approaches like BigGAN-AM [4], which tries to adapt the embeddings for new classes or repair collapsed classes using knowledge transfer from a pre-trained classifier trained on the target dataset. However, we see in Fig. E.8, when fine-tuned for fine-grained datasets like iNaturalist, these approaches fail completely due to the significant domain shift of these datasets compared to ImageNet. We hypothesize that this is because the activation maximization(AM) [4] using a classifier trained on iNaturalist is unable to produce meaningful images as there is presence of distribution shift between the datasets.

Figure E.10. **Qualitative Analysis on iNaturalist2019 (1010 classes).** Examples of generations from various classes for evaluated baselines (Table 1). The baseline ADA suffers from mode collapse, whereas gSR suffers from class confusion particularly for tail classes, particularly for tail classes as seen above on the left. NoisyTwins generates diverse and class-consistent images across all categories.

References

[1] 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. 1, 3

[2] MinGuk Kang, Joonghyuk Shin, and Jaesik Park. StudioGAN: A Taxonomy and Benchmark of GANs for Image Synthesis. *2206.09479 (arXiv)*, 2022. 2

[3] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020. 3, 4, 6

[4] Qi Li, Long Mai, Michael A Alcorn, and Anh Nguyen. A cost-effective method for improving and re-purposing large, pre-trained gans by fine-tuning their class-embeddings. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 6

[5] 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. 3

[6] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. volume abs/2201.00273, 2022. 2, 6

[7] 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. 5, 6

[8] 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. 6