Deceive D: Adaptive Pseudo Augmentation for GAN Training with Limited Data

Supplementary Material

Liming Jiang[1] **Bo Dai**[1] **Wayne Wu**[2] **Chen Change Loy**[1*]

[1]S-Lab, Nanyang Technological University [2]SenseTime Research

{liming002, bo.dai, ccloy}@ntu.edu.sg wuwenyan@sensetime.com

Abstract

This document provides supplementary information that is not elaborated in our main paper due to the space constraints: Section A describes some preliminary concepts and definitions of our methodology. Section B provides additional discussions on previous techniques for regularizing GANs. Section C details our used datasets. Section D presents the implementation details of our experiments. Section E shows more results and analysis.

A Preliminaries of Methodology

To further facilitate readers' understanding of the theoretical analysis and the proposed Adaptive Pseudo Augmentation (APA), this section will provide some preliminary concepts and definitions for the methodology section in our main paper.

Generative adversarial networks (GANs) [7] aim at capturing the real data distribution to synthesize new data. Two networks are trained alternately via an adversarial process: a generator G learns to produce new samples, and a discriminator D (*i.e.*, a binary classifier) predicts the probability that a sample comes from the real data rather than from G. Following [7, 13], our main paper focuses on the fundamental problem of GANs, *i.e.*, unconditional image synthesis, which is generating random samples from a noise input in the latent space. The noise is sampled from a Gaussian distribution.

The goal of GANs is to learn an ideal generated distribution p_g from the real data distribution p_{data}. Let $p_z(z)$ be the prior on the input noise variable. The mapping from the latent space to the image space can be denoted as $G(z)$. For sample x, $D(x)$ represents the estimated probability of x coming from the real data. Here, both G and D should be differentiable functions that are defined by the network parameters. To quantify the adversarial process, G and D play a minimax two-player game with the value function $V(G, D)$:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{\mathrm{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] . \tag{1}$$

Let the virtual training criterion [7] for the generator G be $C(G)$. The global minimum of $C(G)$ is achieved if and only if $p_g = p_{\mathrm{data}}$, and the minimum value is $-\log 4$, as proved by [7]. This indicates that GANs can perfectly model the real data distribution if given sufficient capacity and time. In practice, we usually use a non-saturated form for G and train it to maximize $\log D(G(z))$ instead of minimizing $\log(1 - D(G(z)))$ to ensure a healthy gradient at the early training stage.

In all the figures of this work, we show raw output logits of D before the last Sigmoid activation to better visualize its prediction confidence. Let logit denotes the logit function, we define:

$$D_{\mathrm{real}} = \mathrm{logit}(D(x)), \quad D_{\mathrm{fake}} = \mathrm{logit}(D(G(z))) . \tag{2}$$

[*]Corresponding author.

B Additional Discussions on Previous Techniques for Regularizing GANs

As mentioned in our main paper, previous techniques for regularizing GANs include adding noise to the inputs of the discriminator [2, 22, 9], gradient penalty [18, 8, 20], one-sided label smoothing [21], spectral normalization [19], label noise [5], *etc*. These approaches are designed for stabilizing training or preventing mode collapse [21]. The essence of their objectives could be considered similar to our method since training GANs in the low-data regime exhibits similar behaviors as previously observed in early GANs with sufficient data. However, several differences are worth highlighting.

First of all, the actual goals of previous strategies and the proposed APA may not be completely the same. Specifically, APA is specialized for training GANs in the low-data regime, which was not carefully considered by prior studies. This difference in the data setting is quite important since it is the core problem we wish to address. Even the performance of state-of-the-art StyleGAN2 [13] deteriorates when trained with a limited amount of data, although it has exploited many advanced techniques for stabilizing training or preventing mode collapse, such as R1 regularization [18]. The main challenge that lies within the low-data regime is the overfitting of the discriminator. Although this issue might also appear on early GANs, it becomes more severe when data is limited.

Besides, we have presented the comparative studies on APA and conventional techniques in the main paper. Empirically, we showed that previous methods could not boost the performance of StyleGAN2 much under limited data. Some of them showed a huge performance gap in comparison to the proposed APA. The experiments indicate that previous relevant strategies cannot handle the low-data regime well, further suggesting the effectiveness and usefulness of APA.

Last but not least, APA and these techniques themselves are not exactly the same. The proposed APA is an effective practice and improvement of these ideas on modern GANs, whose implementations are very different from the early ones. Compared to previous techniques, APA is more adaptive to fit different settings and the overfitting status in training. Although using adaptive heuristics was also explored in the past, it had been found unpractical at the time [5]. APA makes the adaptive control scheme possible in practice.

We believe that the proposed APA could contribute to the community for its effectiveness, simplicity, and adaptability for training state-of-the-art GANs in the low-data regime. Hopefully, our approach could extend the breadth and potential of solutions to GAN training with limited data.

C Dataset Details

This section will detail our explored datasets in the main paper. We randomly select subsets to confine the size of large datasets and directly use small datasets for GAN training with limited data under different settings. Our main paper focuses on the fundamental unconditional image synthesis task with powerful contemporary GANs. Thus, there is no need to split a separate test set. We exploit a high-quality Lanczos filter [17] for image resizing and save images in the uncompressed PNG format. We will detail each dataset separately as follows.

- **AFHQ-Cat.** We use the AFHQ-Cat dataset released by [6], which is the cat category of the high-quality Animal Faces-HQ (AFHQ) dataset [6]. The original authors mentioned that they collected images with permissive licenses from the Flickr and Pixabay websites. All images were vertically and horizontally aligned at the center. The dataset was released under Creative Commons BY-NC 4.0 license by NAVER Corporation. As a small dataset, we exploit $5,153$ training images of cat faces with various breeds. The original resolution is 512×512, which is scaled to 256×256 for training.
- **FFHQ.** The full Flickr-Faces-HQ (FFHQ) dataset [12] consists of $70,000$ high-quality (1024×1024) human face images. The faces contain considerable variation in terms of age, ethnicity, background, and accessories. The original authors mentioned that the images were crawled from the Flickr website, and only the ones under permissive licenses were collected. The images were automatically aligned [14] and cropped. The individual images were published in Flickr by their respective authors under either Creative Commons BY 2.0, Creative Commons BY-NC 2.0, Public Domain Mark 1.0, Public Domain CC0 1.0, or U.S. Government Works license. The dataset itself was made available under Creative Commons BY-NC-SA 4.0 license by NVIDIA Corporation. We randomly select different

subsets of FFHQ, *i.e.*, FFHQ-1k ($1,000$ images, $\sim 1.4\%$ data), FFHQ-5k ($5,000$ images, $\sim 7\%$ data), FFHQ-7k ($7,000$ images, 10% data), and FFHQ-70k ($70,000$ images, 100% data) to perform GAN training given different data amounts. The original images are resized to 256×256 for training in the experiments of our main paper. We will also show some synthesized examples with the original resolution in this document.

- **Danbooru2019 Portraits (Anime).** The Danbooru2019 Portraits (Anime) [3] is a dataset consisting of 512×512 anime faces cropped from solo SFW Danbooru2019 images [1]. The full dataset contains a total of $302,652$ images in a broad portrait style encompassing ears, necklines, hats, *etc*., rather than the tightly cropped faces. The dataset was released under the Creative Commons public domain (CC-0) license. We artificially confine the dataset into a subset for GAN training, *i.e.*, Anime-5k ($5,000$ images, $\sim 2\%$ data). The training image size is 256×256.
- **Caltech-UCSD Birds-200-2011 (CUB).** The Caltech-UCSD Birds-200-2011 (CUB) [24] is an extended version of CUB-200 [25], a challenging dataset containing 200 bird species. The dataset consists of $11,788$ bird images at diverse locations with heavy background clutter. The images were harvested using the Flickr image search. We were unable to find the license information about this dataset. We employ all the available images for GAN training since the dataset is already small and difficult. The resolutions of original images vary, and we uniformly resize them to 256×256.

D Implementation Details

In the main paper, we choose the state-of-the-art StyleGAN2 [13] as the backbone to verify the effectiveness of the proposed APA on limited data. Besides, we compare our method with representative approaches designed for the low-data regime, including the adaptive discriminator augmentation (ADA) [11] and LC-regularization (LC-Reg) [23], which perform standard data augmentations and model regularization, respectively. In addition, we compare APA with representative conventional techniques for regularizing GANs, *i.e.*, instance noise [22] and one-sided label smoothing [21]. For a fair and controllable comparison, we reimplement all baselines and run the experiments from scratch using official code. The qualitative and quantitative results of each method are reported using the best model throughout training.

We ran our experiments on an internal computing cluster with Slurm Workload Manager. All the models are trained on 8 NVIDIA Tesla V100 GPUs with 32 GB memory capacity. We follow [11] and employ its mixed-precision FP16/FP32 training scheme in all our experiments. The actual memory consumption for each model is around 11 GB per GPU. We perform $25,000$ kimg (*i.e.*, thousands of images shown to the discriminator, measuring the training progress [12, 13, 11]) of training for each model. For the average training time cost of different models, please refer to the Training cost of Section 4.2 in the main paper.

We implemented the proposed APA on top of the official implementation of StyleGAN2 [13]. The network architecture is kept unchanged. The mapping network contains 8 fully connected layers, and the dimensionality of the input and intermediate latent space is 512. We use the combination of a generator with output skips and a residual discriminator. The detailed structures of the generator and the discriminator are the same as [13], *e.g.*, using weight demodulation [13] in the generator. The activation function is Leaky ReLU with a negative slope of 0.2. We apply several other standard techniques in [10, 12, 13], including the mini-batch standard deviation layer at the end of the discriminator [10], equalized learning rate for all the trainable parameters [10], pixel-wise feature vector normalization [10], the exponential moving average of generator weights [10], style mixing regularization [12], path length regularization [13], and lazy regularization [13]. The training loss is the non-saturating logistic loss [7, 13] with R_1 regularization [18]. The batch size is 64 for our experiments trained with the resolution of 256×256. The Adam [15] optimizer is applied with $\beta_1 = 0, \beta_2 = 0.99$. For other network and training details, we follow the original paper and official code of StyleGAN2 [13].

For APA, we set the overfitting heuristic $\lambda = \lambda_r$ in our main experiments and study other variants (*i.e.*, $\lambda = \lambda_f$ and $\lambda = \lambda_{rf}$) through the ablation study. Aside from the ablation study of a fixed deception probability $p = 0.5$, p is adaptively adjusted according to λ. The adaptive adjustment of p is as follows. We first initialize p to be zero and set a threshold value t ($t = 0.6$ in most cases

Table 1: The FID (lower is better) and IS (higher is better) scores (256×256) of **transfer learning** on MetFaces [11] with limited data amounts from the pre-trained StyleGAN2 model on FFHQ-70k [12].

Method	MetFaces-1336 (full data)		MetFaces-500 ($\sim$ 37% data)	
	FID $\downarrow$	IS $\uparrow$	FID $\downarrow$	IS $\uparrow$
StyleGAN2 [13]	30.988	3.719	54.691	3.218
APA (Ours)	**21.050**	**4.103**	**29.508**	**3.986**

Table 2: The FID (lower is better) and IS (higher is better) scores (256×256) **under a lower amount of training data** on FFHQ-500 [12] (a subset of 500 images, $\sim$ 0.7% of full data).

Method	FFHQ-500 ($\sim$ 0.7% data)	
	FID $\downarrow$	IS $\uparrow$
StyleGAN2 [13]	119.815	2.446
APA (Ours)	**50.989**	**4.099**

unless specified otherwise). If λ signifies too much/little overfitting regarding t (*i.e.*, larger/smaller than t), the probability p will be increased/decreased by one fixed step. Using this step size, p can increase quickly from zero to one, *i.e.*, in 500k images shown to D. We adjust p once every four iterations. We clamp p from below to zero after each adjustment so that p can always be larger than zero. We do not set an upper bound for p while it can be naturally restricted under a safe limit. Then, the pseudo augmentation of each instance will be applied with the probability p or be skipped with the probability $1 - p$. In this way, the strength of pseudo augmentation can be adaptively controlled based on the degree of overfitting throughout training.

As for other methods used for comparison, we strictly follow all the details in their papers and released official code using recommended setups.

E More Results and Analysis

Effectiveness of APA for transfer learning. Table 1 reports the additional transfer learning results on MetFaces [11] from the pre-trained StyleGAN2 model on FFHQ-70k [12]. All the metrics can be boosted by APA, further verifying its effectiveness for transfer learning besides training from scratch.

The performance of APA under a lower amount of data (*i.e.*, fewer than 1k). We have reported transfer learning results on MetFaces-500 [11] (a subset of 500 images, $\sim$ 37% of full data) in Table 1. Table 2 shows more results on FFHQ-500 [12] (a subset of 500 images, $\sim$ 0.7% of full data). Under fewer data, APA can still boost StyleGAN2 performance by a large margin. However, the quality itself of synthesized images by APA can still be improved, consistent with our discussion on limitations in the main paper.

More examples of the effectiveness of APA on various datasets. We show more comparative results of StyleGAN2 [13] and the proposed APA to demonstrate the effectiveness of our method to improve the state-of-the-art baseline on various datasets with limited data amounts, *i.e.*, AFHQ-Cat-5k [6] (Figure 1), FFHQ-5k [12] (Figure 2), Anime-5k [3] (Figure 3), and CUB-12k [24] (Figure 4). Regardless of applying the truncation trick [4, 12, 13] or not, the proposed APA can significantly ameliorate the degraded synthesis quality of StyleGAN2 with limited training data in all cases. The generated images by our method are highly photorealistic using only limited training data, being closer to the real data distributions.

Qualitative results of comparison with other solutions for GAN training with limited data. In the main paper, we have quantitatively shown the effectiveness of the proposed APA over other state-of-the-art approaches designed for the low-data regime, including ADA [11] and LC-regularization (LC-Reg) [23], which perform standard data augmentations and model regularization, respectively. In Figure 5, we show the qualitative results for further illustration. The synthesis quality of the StyleGAN2 [13] baseline deteriorates on the limited amount of training data. Ripple artifacts and substantial distortions appear on the generated images by StyleGAN2. LC-Reg [23] slightly improves the quality of synthesis by reducing the distortions on the images. The amelioration of visual quality is more evident by applying ADA [11], where the ripple artifacts are clearly subsided while some minor artifacts exist on the hair and beard. The proposed APA achieves comparable or even better visual quality than LC-Reg [23] and ADA [11], effectively improving the StyleGAN2 synthesized results on limited data. Notably, APA is also complementary to ADA [11] for gaining a further

Figure 1: More examples of the effectiveness of our method to improve state-of-the-art StyleGAN2 [13] synthesized results (256×256) on **AFHQ-Cat-5k** [6] ($5,153$ images, which is small by itself).

Figure 2: More examples of the effectiveness of our method to improve state-of-the-art StyleGAN2 [13] synthesized results (256×256) on **FFHQ-5k** [12] (a subset of $5,000$ images, $\sim 7\%$ of full data).

Figure 3: More examples of the effectiveness of our method to improve state-of-the-art StyleGAN2 [13] synthesized results (256×256) on **Anime-5k** [3] (a subset of $5,000$ images, $\sim 2\%$ of full data).

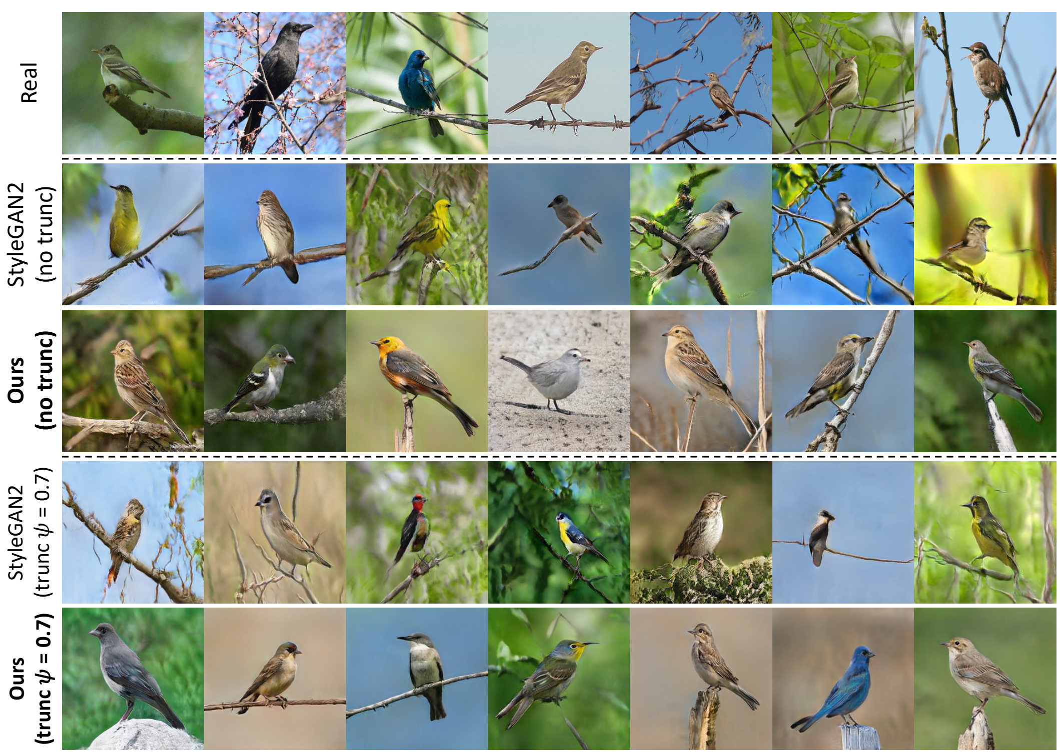

Figure 4: More examples of the effectiveness of our method to improve state-of-the-art StyleGAN2 [13] synthesized results (256×256) on **CUB-12k** [24] ($11,788$ images, which is small by itself).

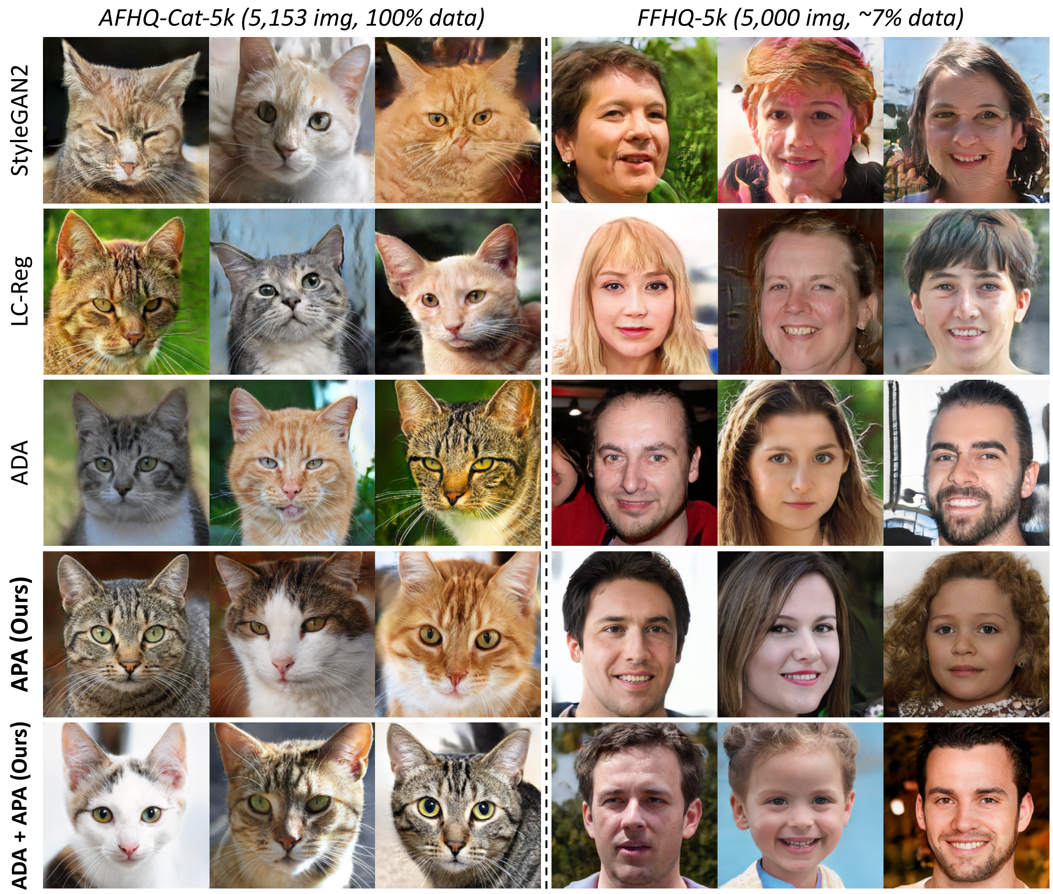

Figure 5: The synthesized results (256 × 256, no truncation) of our method **compared to other state-of-the-art solutions designed for GAN training with limited data** on StyleGAN2 [13]. We confine the data amount of FFHQ [12] and directly use AFHQ-Cat [6] that is small by itself.

Table 3: The FID (lower is better) and IS (higher is better) scores (256 × 256) of **additional comparison with the state-of-the-art ADA [11]** tailored for GAN training with limited data on StyleGAN2 trained with FFHQ [12] and MetFaces [11]. The bold number indicates the best value, and the underline marks the second best.

	FFHQ-7k		FFHQ-1k		MetFaces-1336		MetFaces-500	
Method	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑
StyleGAN2 [13]	27.738	4.264	86.407	2.806	30.988	3.719	54.691	3.218
ADA [11]	10.275	4.813	22.590	4.239	20.834	4.005	30.368	3.974
APA (Ours)	10.800	4.860	45.192	4.130	21.050	4.103	29.508	3.986
ADA + APA (Ours)	**7.333**	**4.994**	**18.892**	**4.316**	**18.865**	**4.207**	**28.408**	**4.044**

performance boost, suggesting the compatibility of our method with standard data augmentations. The visual results we present here are in line with the quantitative performance in our main paper.

More comparative results with ADA [11]. Table 3 reports additional comparative results with the state-of-the-art ADA [11] tailored for GAN training with limited data to further highlight the advantages of the proposed APA. We include additional comparisons on FFHQ-7k [12] and FFHQ-1k (see our main paper for 5k and 70k). To make the comparison settings more comprehensive, we also provide the transfer learning results on MetFaces [11] from the pre-trained StyleGAN2 model on FFHQ-70k. We also make a comparison with a fewer data amount (*i.e.*, 500 images). Combined with the results we presented in the main paper, the proposed APA achieves comparable or even better performance than ADA [11] while with less computational cost. Both methods outperform the StyleGAN2 baseline under limited data. Although applying APA solely may be inferior to ADA [11] on FFHQ-1k (in line with our discussion in the main paper), it is worth mentioning that APA is also complementary to ADA [11], which is very important to boost the performance further.

Figure 6: The effectiveness of our method to improve state-of-the-art **StyleGAN2** [13] higher-resolution synthesized results (1024×1024, no truncation) **trained with the limited data**, *i.e.*, FFHQ-5k [12] (a subset of $5,000$ images, $\sim 7\%$ of full data). Our method achieves an FID score of $\mathbf{9.545}$, outperforming the original StyleGAN2 of 18.296.

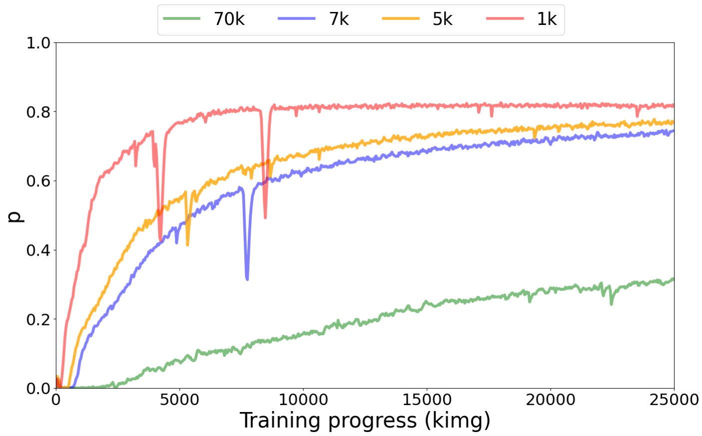

Figure 7: The **evolution of the deception probability** $\boldsymbol{p}$ during training on FFHQ [12] (256×256) with different data amounts. The “kimg” denotes thousands of real images shown to the discriminator.

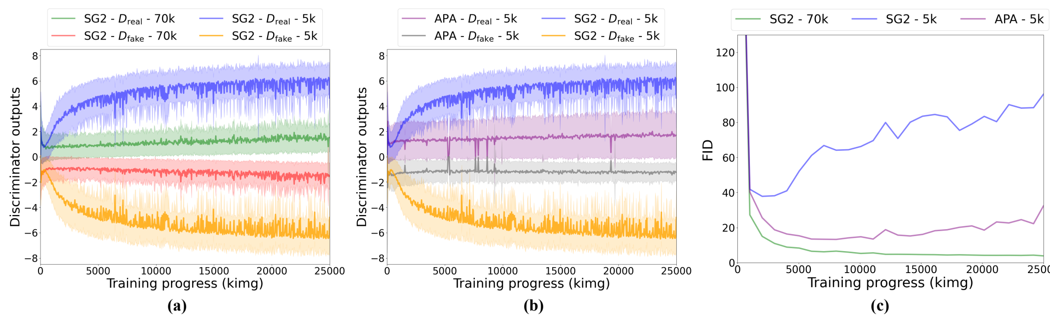

Figure 8: The **overfitting and convergence status** of APA compared to StyleGAN2 (SG2) on FFHQ [12] (256×256). (a) The discriminator raw output logits of StyleGAN2 on the full (70k) or limited (**5k**) datasets. (b) The discriminator raw output logits of StyleGAN2 and APA on the limited (**5k**) dataset. (c) The training convergence shown by FID.

Higher-resolution examples on StyleGAN2. In Figure 6, we show some higher-resolution (1024×1024) synthesized images on FFHQ-5k [12] (a subset of $5,000$ images, $\sim 7\%$ of full data) to further illustrate the effectiveness of our approach in improving StyleGAN2 with limited training data. The truncation trick [4, 12, 13] is not applied. The proposed APA evidently improves StyleGAN2 in both perceptual quality and the FID score, indicating its effectiveness for GAN training with limited data.

Evolution of the deception probability. In Figure 7, we visualize the evolution of the deception probability p in training on FFHQ [12] (256×256) with different data amounts. The evolution of p may be relatively more unstable when the data amount is very limited (*i.e.*, 1k). Notably, the proposed APA possesses a desired property that the deception probability p can be naturally restricted under a safe limit (~ 0.8) regardless of the training data amounts. Hence, the fundamental capability of discriminator in adversarial training may be better preserved thanks to the adaptive control scheme.

More overfitting and convergence analysis. Aside from the overfitting and convergence analysis on FFHQ-7k (10% of full data) in our main paper, we present the additional analysis on FFHQ-5k ($\sim 7\%$ of full data) and FFHQ-1k ($\sim 1.4\%$ of full data) in Figure 8 and Figure 9, respectively. The trend of model overfitting and convergence on FFHQ-5k and FFHQ-1k is similar to FFHQ-7k. The divergence of StyleGAN2 discriminator predictions can be effectively restricted by applying the proposed APA, indicating its effectiveness in mitigating the discriminator overfitting. Meanwhile, APA improves the training convergence measured by FID.

Furthermore, we show the comparative overfitting analysis on APA and previous techniques for regularizing GANs in Figure 10. We observe that the effectiveness to counteract overfitting is in line with the generation performance of these methods in the main paper. Specifically, StyleGAN2 experiences diverged predictions most rapidly, and APA obtains the most effective restriction on the divergence of discriminator outputs. Applying instance noise (ISN) [22] produces curves that are very

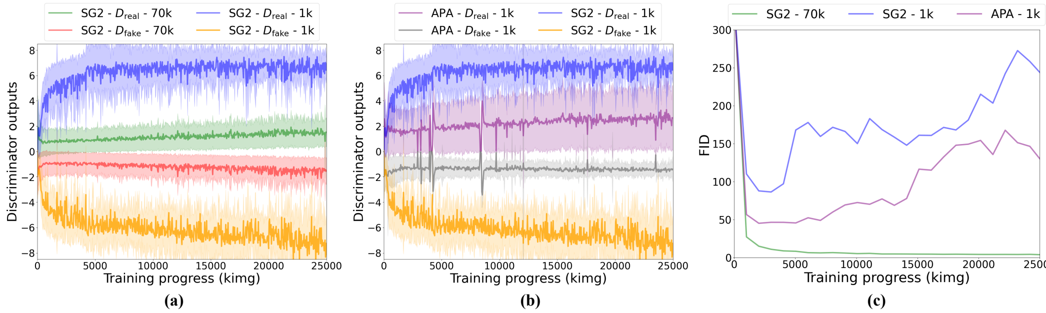

Figure 9: The **overfitting and convergence status** of APA compared to StyleGAN2 (SG2) on FFHQ [12] (256×256). (a) The discriminator raw output logits of StyleGAN2 on the full (70k) or limited (**1k**) datasets. (b) The discriminator raw output logits of StyleGAN2 and APA on the limited (**1k**) dataset. (c) The training convergence shown by FID.

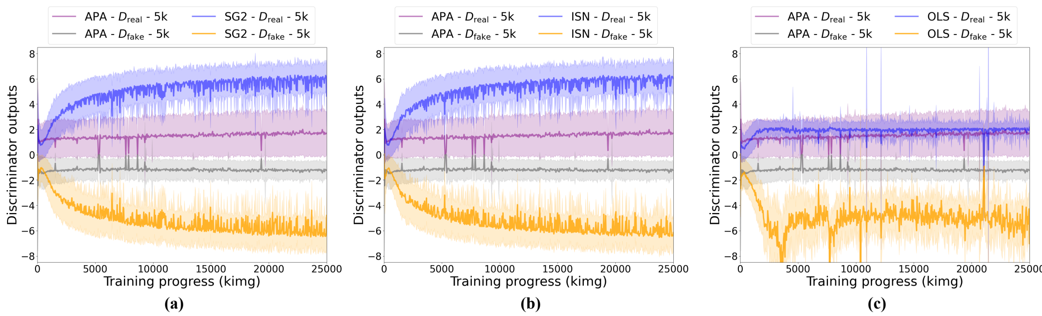

Figure 10: The **overfitting and convergence status** of APA compared to previous techniques for regularizing GANs on FFHQ-5k [12] (256×256, $\sim 7\%$ data). (a) The discriminator raw output logits of StyleGAN2 (SG2) and APA. (b) The discriminator raw output logits of instance noise (ISN) [22] and APA. (c) The discriminator raw output logits of one-sided label smoothing (OLS) [21] and APA.

Table 4: The FID (lower is better) and IS (higher is better) scores (32×32) of **class-conditional image synthesis on BigGAN** trained with the subsets of **CIFAR-10** [16] with limited data amounts.

	full data		20% data		10% data	
Method	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑
BigGAN [4]	9.531	9.078	22.024	8.343	44.061	7.589
APA (Ours)	**8.283**	**9.362**	**15.316**	**8.822**	**25.987**	**8.410**

Table 5: The FID (lower is better) and IS (higher is better) scores (32×32) of **class-conditional image synthesis on BigGAN** trained with the subsets of **CIFAR-100** [16] with limited data amounts.

	full data		20% data		10% data	
Method	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑
BigGAN [4]	13.281	10.525	35.590	8.706	64.828	6.635
APA (Ours)	**11.429**	**11.243**	**23.506**	**9.811**	**45.794**	**8.114**

close to StyleGAN2, and one-sided label smoothing (OLS) [21] only restricts the real predictions. This further suggests: 1) the importance of addressing the discriminator overfitting for training GANs with limited data; 2) the effectiveness of APA in alleviating the discriminator overfitting, outperforming previous strategies.

Additional training convergence visualizations. Please refer to our supplementary video for additional training convergence visualizations on FFHQ-7k [12] ($7,000$ images, 10% of full data). The truncation trick [4, 12, 13] with $\psi = 0.7$ is applied to the synthesized images. The proposed APA effectively improves the training convergence of StyleGAN2 on limited data.

Class-conditional image synthesis with BigGAN [4]. To further enrich our benchmark and demonstrate the effectiveness of the proposed APA under diverse settings, we show additional class-conditional image synthesis results on the state-of-the-art BigGAN [4] trained with CIFAR-10 [16] and CIFAR-100 [16] in Table 4 and Table 5, respectively. It can be observed that APA outperforms BigGAN under limited training data in all cases, further suggesting its adaptability and effectiveness for class-conditional image synthesis with other powerful contemporary GANs, such as BigGAN.

References

[1] Anonymous, Danbooru community, and Gwern Branwen. Danbooru2019: A large-scale crowdsourced and tagged anime illustration dataset. `https://www.gwern.net/Danbooru2020#danbooru2019`. Accessed: 2021-04-10.

[2] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017.

[3] Gwern Branwen, Anonymous, and Danbooru Community. Danbooru2019 Portraits: A large-scale anime head illustration dataset. `https://www.gwern.net/Crops#danbooru2019-portraits`. Accessed: 2021-04-10.

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2018.

[5] Soumith Chintala. How to train a GAN? NeurIPS workshop on adversarial training. `https://sites.google.com/site/nips2016adversarial/`, 2016. Accessed: 2021-08-03.

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[8] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017.

[9] Simon Jenni and Paolo Favaro. On stabilizing generative adversarial training with noise. In *CVPR*, 2019.

[10] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

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

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

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

[14] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*, arXiv:1412.6980, 2014.

[16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[17] Cornelius Lanczos. *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Governm. Press Office Los Angeles, CA, 1950.

[18] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, 2018.

[19] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

[20] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *NeurIPS*, 2017.

[21] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016.

[22] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. In *ICLR*, 2017.

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

[24] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. 2011.

[25] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD Birds 200. 2010.