Supplementary Material for KD-DLGAN: Data Limited Image Generation via Knowledge Distillation

Kaiwen Cui[1], Yingchen Yu[1], Fangneng Zhan[2], Shengcai Liao[3], Shijian Lu[1]*, Eric Xing[4]
[1] Nanyang Technological University, [2] Max Planck Institute for Informatics
[3] Inception Institute of Artificial Intelligence
[4] Mohamed bin Zayed University of Artificial Intelligence
{Kaiwen.Cui, Yingchen.Yu, Shijian.Lu}@ntu.edu.sg, fzhan@mpi-inf.mpg.de
shengcai.liao@inceptioniai.org, Eric.Xing@mbzuai.ac.ae

1. Supplementary Material

1.1. Overview

We provide this supplementary material due to the space limitation of the main manuscript. The information in this supplementary material includes: 1) dataset details; 2) implementation details; 3) text label generation; 4) parameter study; 5) more quantitative and qualitative comparisons with the state-of-the-art.

1.2. Dataset

We conduct experiments over multiple widely adopted datasets including: 100-shot, AFHQ, CIFAR-10, CIFAR-100 and ImageNet.

100-shot: 100-shot contains three datasets each of which has 100 samples of resolution 256×256. The three datasets are 100-shot Obama, 100-shot Grumpy Cat and 100-shot Panda.

AFHQ: AFHQ consists of face images of three types of animals including Cat, Dog and Wildlife, each of which has 5k training images. We follow DA [9] and use 160 AFHQ-Cat images and 389 AFHQ-Dog images (at a resolution of 256×256) for training.

CIFAR-10: CIFAR-10 contains 50k training images and 10k validation images with 10 classes. The image resolution is 32×32. In our experiments, three networks are trained with 100%, 20% or 10% training images, respectively, and the trained models are valuated over all the validation images.

CIFAR-100: CIFAR-100 contains 50k training images and 10k validation images of 100 classes. The image resolution is 32×32. In our experiments, three networks are trained with 100%, 20% or 10% training images, respectively, and the trained models are evaluated over all the validation data.

ImageNet: ImageNet contains $1,281,167$ training images of 1000 classes. We employ the resolution 64×64 in our experiments. We trained three networks with different amounts of training data including ~10%, ~5% and ~2.5% data. We perform evaluations over all the training images ($i.e.$, $1,281,167$ training images).

1.3. Implementation Details

StyleGAN-v2 on 100-shot and AFHQ: For experiments with 100-shot and AFHQ, our KD-DLGAN is built on top of StyleGAN-v2 as implemented with PyTorch [9]. The learning rate for G and D is $2e-3$. The batch size is set at 8 and we employ Adam optimizer with $\beta_1 = 0$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$. The FID is evaluated on the whole training set. All models are trained with 1 NVIDIA V100 GPU.

BigGAN on CIFAR-10 and CIFAR-100: In experiments with CIFAR-10 and CIFAR-100, our KD-DLGAN is built on top of BigGAN as implemented with PyTorch [9]. Following [9], the learning rate for G and D is set at $2e-4$. The batch size is set at 50. In addition, we use Adam optimizer with $\beta_1 = 0$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We evaluate FID over the whole validation set. All the models are run on 2 NVIDIA V100 GPUs.

BigGAN on ImageNet: For experiments on ImageNet, our KD-DLGAN is built on top of BigGAN as implemented with PyTorch [9]. We use a learning rate of $1e-4$ for G and $4e-4$ for D for 100% data setting and decrease the learning rate of D to $2e-4$ for the 5% and 2.5% data settings. The batch size is set at 512. In addition, we use Adam optimizer with $\beta_1 = 0$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We evaluate FID over the whole training set. All the models are run on 2 NVIDIA V100 GPUs.

*corresponding author.

Dataset	Probability of applying the designed aggregated loss						
	0.4	0.5	0.6	0.7	0.8	0.9	1.0
CIFAR-10 (10% data)	15.63	14.72	14.61	14.20	19.48	20.56	21.41

Table 1. Experiments on the probability of applying the designed aggregated loss in aggregated generative knowledge distillation. We report FID (↓) scores.

Methods	CIFAR-10			CIFAR-100		
	100% data	20% data	10% data	100% data	20% data	10% data
BigGAN [2]	9.07 ± 0.03	8.52 ± 0.10	7.09 ± 0.03	10.71 ± 0.14	08.58 ± 0.04	06.74 ± 0.04
DA [9]	9.16 ± 0.13	8.65 ± 0.14	8.09 ± 0.08	10.66 ± 0.08	09.47 ± 0.14	08.38 ± 0.12
KD-DLGAN	**9.38** ± 0.28	**9.20** ± 0.24	**9.08** ± 0.21	**10.99** ± 012.	**10.65** ± 0.22	**10.26** ± 0.29

Table 2. Comparing KD-DLGAN with the state-of-the-art over CIFAR-10 and CIFAR-100: KD-DLGAN outperforms the state-of-the-art clearly and consistently by mitigating the discriminator over-fitting. We report IS (↑) scores.

Figure 1. Qualitative results over 100-shot and AFHQ datasets: KD-DLGAN generates more realistic images than DA [9], the state-of-the-art data-limited generation method.

1.4. Text Labels

Conditional Datasets: For conditional datasets, we employ the corresponding image labels as input texts. Hence, CIFAR-10, CIFAR-100, and ImageNet have 10, 100, and 1000 input texts, respectively. Following [6], we expand a single word to a sentence with the prompt template "A photo of a {label}".

Unconditional Datasets: For unconditional datasets, we pre-define a set of relevant text labels as input texts. In our experiments, all unconditional datasets ($i.e.$ 100-shot Obama, 100-shot Grumpy Cat, 100-shot Panda, AFHQ-Cat, AFHQ-Dog) are face-related. We define text labels for these datasets based on face expressions, $i.e.$, neutral, happy, sad, surprise, disgust, anger, fear. Similarly for the conditional

Figure 2. Qualitative results over 100-shot and AFHQ datasets: The generation by KD-DLGAN is clearly more realistic than that by ADA [3], the state-of-the-art data-limited generation method.

datasets, we expand a single word to a sentence with the prompt template "A photo of a {label}".

1.5. Parameter Study

For effective GAN training, the designed aggregated loss in aggregated generative knowledge distillation is controlled by a hyper-parameter p, where the loss is applied with probability p or skipped with probability $1-p$. We perform experiments to study how different p affect the generation performance. Table 1 shows the experimental results while applying the designed aggregated loss with different p. We can see that the image generation performs best when p is 0.7 and is tolerant to p when it lies between 0.5 and 0.7. However, the performance deteriorates clearly while p is too large (higher than 0.7) or small (lower than 0.5). We conjecture that the knowledge distillation performance would be overwhelmed by the aggregated loss (can be regarded as regularization term) when p is too large. On the contrary, the regularization performance of feature aggregation will be poor when p is too small. In our study, we fix this hyper-parameter at 0.7 for all conducted experiments.

1.6. Additional Results

We present more experimental results to demonstrate that our KD-DLGAN can mitigate the discriminator overfitting and achieve superior image generation effectively. Table 2 shows the comparison over CIFAR-10 and CIFAR-100 datasets. Evaluating with inception score (IS) [7], we can observe that KD-DLGAN outperforms the state-of-the-art consistently, especially when training samples are limited. The superior generation performance is largely attributed to our designed generative knowledge distillation techniques in KD-DLGAN, which mitigates the discriminator overfitting and improves the generation performance effectively.

Fig 1 and 2 qualitatively show that KD-DLGAN outperforms the state-of-the-art ($i.e.$, DA [9] and ADA [3])

Methods	FFHQ-5k	Anime-5k	CUB-12k
StyleGAN2	37.88	24.55	24.13
DA	18.76	14.39	13.14
ADA	18.42	14.15	13.60
APA	14.33	13.31	12.99
KD-DLGAN	**10.44**	**9.01**	**7.21**

Table 3. Comparison with the state-of-the-art over FFHQ, CUB and Anime: All the compared methods employ StyleGAN-v2 [5] as backbone. We report FID(↓) averaged over three runs.

Method	AGKD	CGKD	Imagenet-10%	CUB-12k
DA			32.82	13.14
	✓		23.98	8.98
		✓	24.55	9.52
Ours	✓	✓	**19.99**	**7.21**

Table 4. Quantitative ablation study of KD-DLGAN: AGKD and CGKD in KD-DLGAN both improves the generation performance over the baseline DA [9]. KD-DLGAN performs the best as AGKD and CGKD complement each other. The FIDs (↓) are averaged over three runs.

in data-limited image generation, especially in terms of the generated shapes and textures.

We conducted additional experiments over FFHQ [4] (5k samples), CUB [8] (12k samples) and Anime [1] (5k samples), where all the image resolutions are 256 × 256. As Table 3 shows, the proposed KD-DLGAN achieves superior FID for all the suggested datasets. Note all our experiments are performed with one NVIDIA Tesla V100 GPU.

We conducted additional quantitative ablation studies over the Imagenet with 10% data where the image resolution is 64 × 64 and CUB dataset with 12k samples where the image resolution is 256 × 256. As the Table 4 shows, the results on the two new datasets are consistent with that in Table 4 in our manuscript. Both AGKD and CGKD improves generation performance clearly and combining the two complementary designs leads to further improvement consistently.

References

[1] Gwern Branwen, Anonymous, and Danbooru Community. Danbooru2019 portraits: A large-scale anime head illustration dataset. https://www.gwern.net/Crops#danbooru2019-portraits, March 2019. Accessed: DATE. 4

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[3] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020. 3

[4] 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. 4

[5] 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. 3

[6] 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

[7] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 3

[8] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 4

[9] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020. 1, 2, 3, 4