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

Abstract

Generative Adversarial Networks (GANs) rely heavily on large-scale training data for training high-quality image generation models. With limited training data, the GAN discriminator often suffers from severe overfitting which directly leads to degraded generation especially in generation diversity. Inspired by the recent advances in knowledge distillation (KD), we propose KD-DLGAN, a knowledge-distillation based generation framework that introduces pre-trained vision-language models for training effective data-limited generation models. KD-DLGAN consists of two innovative designs. The first is aggregated generative KD that mitigates the discriminator overfitting by challenging the discriminator with harder learning tasks and distilling more generalizable knowledge from the pre-trained models. The second is correlated generative KD that improves the generation diversity by distilling and preserving the diverse image-text correlation within the pre-trained models. Extensive experiments over multiple benchmarks show that KD-DLGAN achieves superior image generation with limited training data. In addition, KD-DLGAN complements the state-of-the-art with consistent and substantial performance gains. Note that codes will be released.

1. Introduction

Generative Adversarial Networks (GANs) [12] have become the cornerstone technique in various image generation tasks. On the other hand, effective training of GANs relies heavily on large-scale training images that are usually laborious and expensive to collect. With limited training data, the discriminator in GANs often suffers from severe overfitting [40, 53], leading to degraded generation as shown in Fig. 1. Recent works attempt to address this issue from two major perspectives: i) massive data augmentation that aims to expand the distribution of the limited training data [53]; ii) model regularization that introduces regularizers to modulate the discriminator learning [40]. We intend to mitigate the discriminator overfitting from a new perspective.

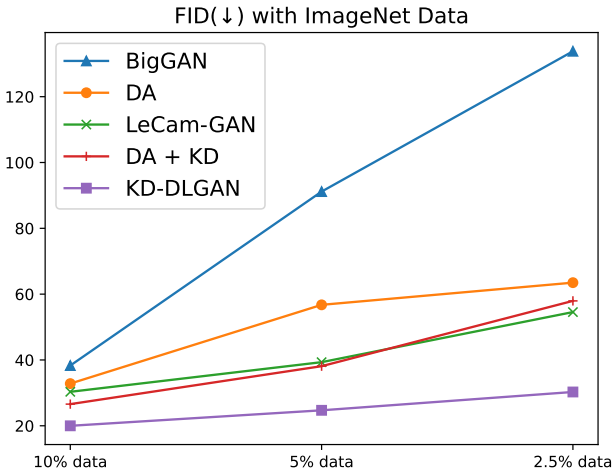

Figure 1. With limited training samples, state-of-the-art GANs such as BigGAN suffer from clear discriminator overfitting which directly leads to degraded generation. The recent work attempts to mitigate the overfitting via mass data augmentation in DA [53] or regularization in LeCam-GAN [40]. The proposed KD-DLGAN distills the rich and diverse text-image knowledge from the powerful visual-language model to the discriminator which greatly mitigates the discriminator overfitting. Additionally, KD-DLGAN is designed specifically for image generation tasks, which also outperforms the vanilla knowledge distillation (DA+KD) greatly.

Recent studies show that knowledge distillation (KD) from powerful vision-language models such as CLIP [36] can effectively relieve network overfitting in visual recognition tasks [2, 9, 29, 42]. Inspired by these prior studies, we explore KD for data-limited image generation, aiming to mitigate the discriminator overfitting by distilling the rich

*corresponding author.

image-text knowledge from vision-language models. One intuitive approach is to adopt existing KD methods [17, 37] for training data-limited GANs, e.g., by forcing the discriminator to mimic the representation space of vision-language models. However, such approach does not work well as most existing KD methods are designed for visual recognition instead of GANs as illustrated in *DA+KD* in Fig. 1.

We propose KD-DLGAN, a knowledge-distillation based image generation framework that introduces the idea of generative KD for training effective data-limited GANs. KD-DLGAN is designed based on two observations in data-limited image generation: 1) the overfitting is largely attributed to the simplicity of the discriminator task, i.e., the discriminator can easily memorize the limited training samples and distinguish them with little efforts; 2) the degradation in data-limited generation is largely attributed to poor generation diversity, i.e., the trained data-limited GAN models tend to generate similar images.

Inspired by the two observations, we design two generative KD techniques that jointly distill knowledge from CLIP [36] to the GAN discriminator for effective training of data-limited GANs. The first is aggregated generative KD (AGKD) that challenges the discriminator by forcing fake samples to be similar to real samples while mimicking CLIP's visual feature space. It mitigates the discriminator overfitting by aggregating features of real and fake samples and distilling generalizable CLIP knowledge concurrently. The second is correlated generative KD (CGKD) that strives to distill CLIP image-text correlations to the GAN discriminator. It improves the generation diversity by enforcing the diverse correlations between images and texts, ultimately improving the generation performance. The two designs distill the rich yet diverse CLIP knowledge which effectively mitigates the discriminator overfitting and improve the generation as illustrated in *KD-DLGAN* in Fig. 1.

The main contributions of this work can be summarized in three aspects. *First,* we propose KD-DLGAN, a novel image generation framework that introduces knowledge distillation for effective GAN training with limited training data. To the best of our knowledge, this is the first work that exploits the idea of knowledge distillation in data-limited image generation. *Second,* we design two generative KD techniques including aggregated generative KD and correlated generative KD that mitigate the discriminator overfitting and improves the generation performance effectively. *Third,* extensive experiments over multiple widely adopted benchmarks show that KD-DLGAN achieves superior image generation and it also complements the state-of-the-art with consistent and substantial performance gains.

2. Related Works

Generative Adversarial Network: Generative adversarial network [12] (GAN) has achieved significant progress in automated image generation and editing [23, 34, 47, 49]. Following the idea in [12], quite a few generation applications have been developed in the past few years. They intend to generate more realistic images from different aspects by adopting novel training objectives [3, 13, 30], designing more advanced networks [31, 32, 50], introducing elaborately designed training strategies [20, 27, 51], etc. On the contrary, most existing GANs rely heavily on large-scale training samples. With only limited samples, they often suffer from clear discriminator overfitting and severe generation degradation.

We target data-limited image generation, which intends to train effective GAN models with limited number of samples yet without sacrificing much generation performance.

Data-Limited Image Generation: Data-limited image generation is a challenging yet meaningful task for circumventing the laborious image collection process. Prior studies [14, 43] suggest that one of the main obstacles of training effective data-limited GAN lies with the overfitting of GAN discriminator. Recent studies [10, 18, 19, 21, 40, 44, 53] attempt to mitigate the overfitting issue mainly through massive data augmentation or model regularization. For example, [53] introduces different types of differentiable augmentation to improve the generation performance. [21] presents an adaptive augmentation mechanism that prevents undesirable leaking of augmentation to the generated images. [40] introduces a regularization scheme to modulate the discriminator. Several studies instead employ external knowledge for data-limited image generation. For example, [33] pretrains the GAN model on a larger dataset. [25] employs off-the-shelf models as additional discriminators to improve the data-limited GAN performance.

We target to tackle the discriminator overfitting issue from a new perspective of knowledge distillation and design two generative knowledge distillation techniques to effectively distill knowledge from a powerful vision-language model to the GAN discriminator.

Knowledge Distillation: Knowledge distillation is a general technique for supervising the training of student networks by transferring the knowledge of trained teacher networks. Knowledge distillation is initially designed for model compression [6] via mimicking the output of an ensemble of models. [4] further compresses deep networks into shallower but wider ones via mimicking the logits. [17] presents a more general knowledge distillation technique by applying the prediction of the teacher model as a soft label. [41] measures the similarity between pairs of instances in the teacher's feature space and encourages the student to mimic the pairwise similarity. Leveraging on these ideas, Knowledge distillation has recently been widely explored and adopted in various applications such as image classification [48, 54], domain adaptation [1, 15] , object detection [7, 8], semantic segmentation [28, 45], etc.

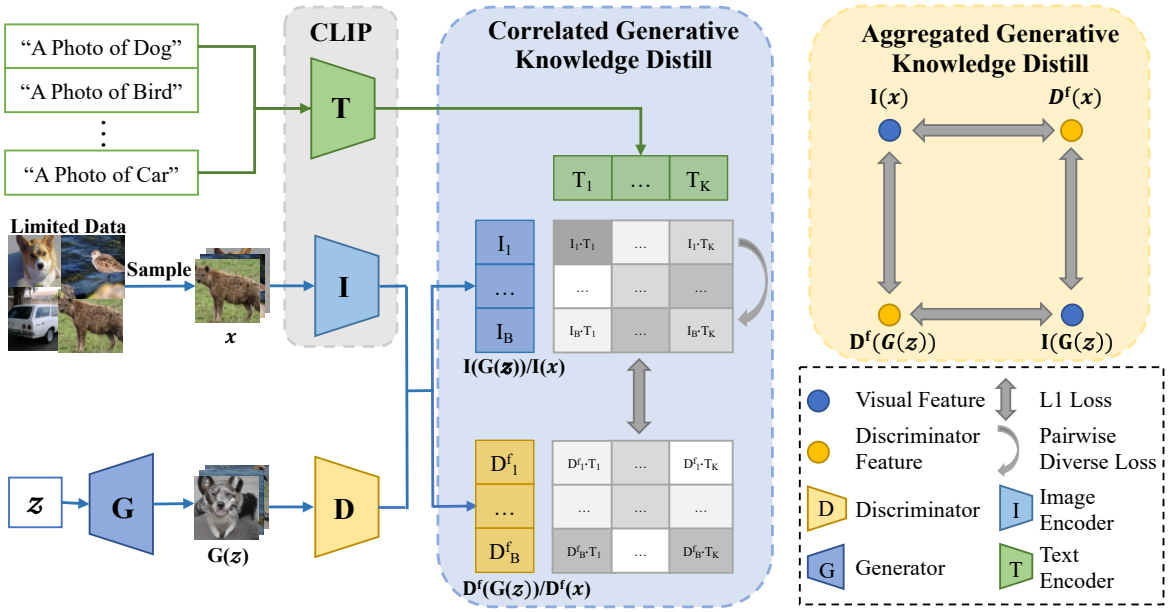

Figure 2. Architecture of the proposed KD-DLGAN: KD-DLGAN distills knowledge from the powerful vision-language model CLIP [36] to the discriminator for effective GAN training with limited training data. We design two generative knowledge distillation techniques that work orthogonally, including aggregated generative knowledge distillation and correlated generative knowledge distillation. Aggregated generative knowledge distillation mitigates the discriminator overfitting with harder learning tasks and distills general knowledge from the pretrained model. Correlated generative knowledge distillation improves the generation diversity by distilling and preserving the diverse image-text correlation within the pretrained model.

We introduce knowledge distillation into the data-limited image generation task for mitigating its overfitting issue. To the best of our knowledge, this is the first work that explores knowledge distillation in data-limited image generation.

3. Method

3.1. Overview

We describes the detailed methodology of the proposed KD-DLGAN in this section. As shown in Fig. 2, we introduce knowledge distillation for data-limited image generation. Specifically, leveraging CLIP [36] as the teacher model, we design two generative knowledge distillation techniques, including aggregated generative knowledge distillation that leads to less distinguishable real-fake samples for the discriminator while distilling more generalizable knowledge from the pretrained model, and correlated generative knowledge distillation that encourages the discriminator to mimic the diverse vision-language correlation in CLIP. The ensuing subsections will describe the problem definition of data-limited image generation, details of the proposed aggregated generative knowledge distillation and correlated generative knowledge distillation, and the overall training objective, respectively.

3.2. Problem Definition

GANs greatly change the paradigm of image generation. Each GAN consists of a discriminator D and a generator G. The general loss function for discriminator and generator is formulated as:

$$\mathcal{L}_d(D; x, G(z)) = \mathbb{E}[\log(D(x))] + \mathbb{E}[\log(1 - D(G(z))] \quad (1)$$

$$\mathcal{L}_g(D; G(z)) = \mathbb{E}[\log(1 - D(G(z))] \quad (2)$$

where $\mathcal{L}_d$ and $\mathcal{L}_g$ denote the general discriminator loss and generator loss, respectively. x denotes a training sample and z is randomly sampled from Gaussian Distribution.

In data-limited image generation, the discriminator in GANs tends to memorize the exact training data and is prone to suffer from overfitting, leading to sub-optimal image generation. Recent studies show that knowledge distillation from powerful and generalizable models can relieve overfitting [2, 9, 29, 42] effectively. However, these state-of-the-art knowledge distillation methods are mainly designed for visual recognition tasks, which cannot be naively adapted to the image generation tasks. We design two novel knowledge distillation techniques that can greatly improve data-limited image generation, more details to be presented in the following subsections.

3.3. Aggregated Generative Knowledge Distillation

We design aggregated generative knowledge distillation (AGKD) to mitigate the discriminator overfitting by distilling more generalizable knowledge from the pretrained CLIP [36] model. Thus, we force the discriminator feature space to mimic CLIP visual feature space. Specifically, for real samples x, we distill knowledge from CLIP visual feature of x (denoted by $I(x)$) to the discriminator feature of x (last layer feature, which is denoted by $D^f(x)$) with L1 loss. Similarly, for generated samples $G(z)$, we also distill knowledge from CLIP visual feature $I(G(z))$ to the discriminator feature $D^f(G(z))$. The knowledge distillation loss $\mathcal{L}^{KD}_{AGKD}$ in AKGD can thus be formulated by:

$$\mathcal{L}^{KD}_{AGKD} = |I(x) - D^f(x)| + |I(G(z)) - D^f(G(z))|$$

AGKD also mitigates the discriminator overfitting by aggregating features of real and fake samples to challenge the discriminator learning. Specifically, we match the CLIP visual feature of real samples $I(x)$ and discriminator features of fake samples $D^f(G(z))$, as well as CLIP visual feature of fake samples $I(G(z))$ and discriminator features of real samples $D^f(x)$ by the L1 loss. Such design lowers the real-fake discriminability and makes it harder to distinguish real-fake samples for the discriminator. The aggregated loss $\mathcal{L}^{AGG}_{AGKD}$ can be formulated by:

$$\mathcal{L}^{AGG}_{AGKD} = |I(x) - D^f(G(z))| + |I(G(z)) - D^f(x)|$$

For effective GAN training, the designed aggregated loss $\mathcal{L}^{AGG}_{AGKD}$ is controlled by a hyper-parameter p, where the loss is applied with probability p or skipped with probability 1-p. We empirically set p at 0.7 in our trained networks. Thus, the aggregated loss $\mathcal{L}'^{AGG}_{AGKD}$ can be re-formulated by:

$$\mathcal{L}'^{AGG}_{AGKD} = \begin{cases} \mathcal{L}^{AGG}_{AGKD}, & \text{if } q \leq p, \\ 0, & \text{if } q > p, \end{cases}$$

where q is a random number sampled between 0 and 1.

The overall AGKD loss $\mathcal{L}_{AGKD}$ can be formulated by:

$$\mathcal{L}_{AGKD} = \mathcal{L}^{KD}_{AGKD} + \mathcal{L}'^{AGG}_{AGKD} \tag{3}$$

3.4. Correlated Generative Knowledge Distillation

Correlated generative knowledge distillation (CGKD) aims to improve the generation diversity by two steps: 1) it enforces the diverse correlations between generated images and texts in CLIP with a pairwise diversity loss ($\mathcal{L}^{PD}_{CGKD}$); 2) it distills the diverse correlations from CLIP to the GAN discriminator with a distillation loss ($\mathcal{L}^{KD}_{CGKD}$).

To achieve diverse image-text correlations, it first builds the correlations (indicated by inner products) between CLIP visual features of generated images $I(G(z)) \in \mathbb{R}^{B\times M}$ and CLIP texts features $T \in \mathbb{R}^{K\times M}$. Here, B is the batch size of generated images, K is the number of texts and M is the features dimension for each text feature or each image feature. For conditional datasets and unconditional datasets, we employ the corresponding image labels as input texts and predefine a set of relevant text labels as input texts, respectively. Details of text selection for our datasets are introduced in the supplementary material. Thus, the correlation $C_T \in \mathbb{R}^{B\times K}$ between $I(G(z))$ and T (i.e., their L2-normalized inner products) can be defined as follows:

$$C_T = \frac{I(G(z)) \cdot T'}{||I(G(z)) \cdot T'||_2},$$

where T' is the transpose of T.

With the defined correlation, the diverse CLIP image-text correlations can be extracted in a pairwise manner. Specifically, for each image-text correlation $C_T[i,:] \in \mathbb{R}^K$, we diversify it with another image-text correlation $C_T[j,:] \in \mathbb{R}^K$ by minimizing the cosine similarity between them. Note [i,:] or [j,:] denotes the i-th or j-th row in C_T and $j \neq i$. The pairwise diversity loss $\mathcal{L}^{PD}_{CGKD}$ can thus be defined as the sum of the cosine similarity of all pairs:

$$\mathcal{L}^{PD}_{CGKD} = \sum_{i=1}^{K} \sum_{j=1, j\neq i}^{K} Cos(C_T[i,:], C_T[j,:]),$$

where $Cos(\overrightarrow{a}, \overrightarrow{b})$ indicates the cosine similarity between the two vectors $\overrightarrow{a}$ and $\overrightarrow{b}$.

Then, the obtained diverse correlations are distilled from CLIP to the GAN discriminator, aiming to improve the generation diversity. We build the correlations $C_S \in \mathbb{R}^{B\times K}$ between discriminator features of generated samples $D^f(G(z)) \in \mathbb{R}^{B\times M}$ and CLIP text features $T \in \mathbb{R}^{K\times M}$ as follows:

$$C_S = \frac{D^f(G(z)) \cdot T'}{||D^f(G(z)) \cdot T'||_2}$$

The correlation distillation from C_T to C_S is defined by the L1 loss between them:

$$\mathcal{L}^{KD}_{CGKD} = |C_T - C_S|$$

The overall CGKD loss $\mathcal{L}_{CGKD}$ can thus be defined by:

$$\mathcal{L}_{CGKD} = \mathcal{L}^{PD}_{CGKD} + \mathcal{L}^{KD}_{CGKD} \tag{4}$$

3.5. Overall Training Objective

The overall training objective of the proposed KD-DLGAN can thus be formulated by:

$$\min_G \max_D \mathcal{L}_G + \mathcal{L}_D \tag{5}$$

where $\mathcal{L}_G = \mathcal{L}_g$ as introduced in Eq. 2 and $\mathcal{L}_D = \mathcal{L}_d + \mathcal{L}_{AGKD} + \mathcal{L}_{CGKD}$ as introduced in Eqs. 1, 3 and 4.

Methods	100-shot			AFHQ	
	Obama	Grumpy Cat	Panda	Cat	Dog
DA [53] + KD (CLIP [36])	45.22	25.62	11.24	38.31	55.13
DA [53] (Baseline)	46.87	27.08	12.06	42.44	58.85
+ **KD-DLGAN (Ours)**	**31.54** ± 0.27	**20.13** ± 0.13	**8.93** ± 0.06	**32.99** ± 0.10	**51.63** ± 0.17
LeCam-GAN [40]	33.16	24.93	10.16	34.18	54.88
+ **KD-DLGAN (Ours)**	**29.38** ± 0.15	**19.65** ± 0.17	**8.41** ± 0.05	**31.89** ± 0.09	**50.22** ± 0.16
InsGen [44]	45.85	27.48	12.13	41.33	55.12
+ **KD-DLGAN (Ours)**	**38.28** ± 0.25	**22.16** ± 0.12	**9.51** ± 0.07	**32.39** ± 0.08	**50.13** ± 0.12
APA [19]	43.75	28.49	12.34	39.13	54.15
+ **KD-DLGAN (Ours)**	**34.68** ± 0.21	**23.14** ± 0.14	**8.70** ± 0.05	**31.77** ± 0.09	**51.23** ± 0.13
ADA [21]	45.69	26.62	12.90	40.77	56.83
+ **KD-DLGAN (Ours)**	**31.78** ± 0.22	**19.76** ± 0.11	**8.85** ± 0.05	**32.81** ± 0.10	**51.12** ± 0.15

Table 1. Comparison with the state-of-the-art over 100-shot and AFHQ: KD-DLGAN outperforms and complements the state-of-the-art data-limited image generation approaches consistently. In addition, KD-DLGAN outperforms vanilla knowledge distillation in DA + KD (CLIP [36]) consistently. All the compared methods employ StyleGAN-v2 [22] as backbone. We report FID(↓) averaged over three runs.

Figure 3. Qualitative comparison with the state-of-the-art over 100-shot: Samples generated by KD-DLGAN are clearly more realistic than those generated by DA [53], the state-of-the-art data-limited generation approach.

4. Experiments

In this section, we conduct extensive experiments to evaluate our KD-DLGAN. We first introduce the datasets and the evaluation metrics used in our experiments. We then benchmark KD-DLGAN with StyleGAN-v2 [22] and BigGAN [5]. Moreover, we conduct extensive ablation studies and discussions to support our designs.

4.1. Datasets and Evaluation Metrics

We conduct experiments over the following datasets: CIFAR [24], ImageNet [11], 100-shot [53] and AFHQ [39]. Datasets details are provided in the supplementary material. We perform evaluations with Frechet Inception Distance (FID) [16] and inception score (IS) [38].

4.2. Experiments with StyleGAN-v2

Table 1 shows unconditional image generation results over 100-shot and AFHQ datasets, where we employ StyleGAN-v2 [22] as the backbone. Following the data settings in DA [53], the models are trained with 100 samples (100-shot Obama, Grumpy Cat, Panda), 160 samples (AFHQ Cat) and 389 samples (AFHQ Dog), respectively.

As Row 3 of Table 1 shows, including the proposed KD-DLGAN into DA [53] achieves superior performance across all data settings consistently as compared with DA alone (in Row 2), demonstrating the complementary relation between KD-DLGAN and DA [53]. In addition, the vanilla knowledge distillation in DA + KD (CLIP [36]) (Row 1) trains the GAN discriminator to mimic the visual feature representation of CLIP. We can observe that KD-

Method	CIFAR-10			CIFAR-100		
	100% Data	20% Data	10% Data	100% Data	20% Data	10% Data
DA [53] + KD (CLIP [36])	8.70 ± 0.02	13.70 ± 0.08	22.03 ± 0.07	11.74 ± 0.02	21.76 ± 0.06	33.93 ± 0.09
DA [53] (Baseline)	8.75 ± 0.03	14.53 ± 0.10	23.34 ± 0.09	11.99 ± 0.02	22.55 ± 0.06	35.39 ± 0.08
+ **KD-DLGAN (Ours)**	**8.42** ± 0.01	**11.01** ± 0.07	**14.20** ± 0.06	**10.28** ± 0.03	**15.60** ± 0.08	**18.03** ± 0.11
APA [19]	8.28 ± 0.02	15.31 ± 0.04	25.98 ± 0.06	11.42 ± 0.04	23.50 ± 0.06	45.79 ± 0.15
+ **KD-DLGAN (Ours)**	**8.26** ± 0.02	**11.15** ± 0.06	**13.86** ± 0.07	**10.23** ± 0.02	**19.22** ± 0.07	**27.11** ± 0.10
LeCam-GAN [40]	8.46 ± 0.06	14.55 ± 0.08	16.69 ± 0.02	11.20 ± 0.09	22.45 ± 0.09	27.28± 0.05
+ **KD-DLGAN (Ours)**	**8.19** ± 0.01	**11.45** ± 0.07	**13.22** ± 0.03	**10.12** ± 0.03	**18.70** ± 0.05	**22.40** ± 0.06
ADA [21]	8.99 ± 0.03	19.87 ± 0.09	30.58 ± 0.11	12.22 ± 0.02	22.65 ± 0.10	27.08 ± 0.15
+ **KD-DLGAN (Ours)**	**8.46** ± 0.02	**14.12** ± 0.10	**16.88** ± 0.08	**10.48** ± 0.04	**19.26** ± 0.06	**20.62** ± 0.09

Table 2. Comparison with the state-of-the-art over CIFAR-10 and CIFAR 100: KD-DLGAN outperforms and complements the state-of-the-art clearly. KD-DLGAN also performs better than vanilla knowledge distillation in DA + KD (CLIP [36]) consistently as well. All the compared methods employ BigGAN [5] as backbone. And we report FID(↓) averaged over three runs.

Method	10% training data		5% training data		2.5% training data	
	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓
DA [53] + KD (CLIP [36])	13.29 ± 0.50	26.58 ± 0.21	11.63 ± 0.29	38.11 ± 0.33	9.43 ± 0.25	57.95 ± 0.41
DA [53] (Baseline)	12.76 ± 0.34	32.82 ± 0.18	9.63 ± 0.21	56.75 ± 0.35	8.17 ± 0.28	63.49 ± 0.51
+ **KD-DLGAN (Ours)**	**14.25** ± 0.66	**19.99** ± 0.11	**12.71** ± 0.34	**24.70** ± 0.14	**13.45** ± 0.51	**30.27** ± 0.16
LeCam-GAN [40]	11.59 ± 0.44	30.32 ± 0.24	10.53 ± 0.22	39.33 ± 0.27	9.99 ± 0.26	54.55 ± 0.46
+ **KD-DLGAN (Ours)**	**13.98** ± 0.23	**22.12** ± 0.12	**13.86** ± 0.45	**23.85** ± 0.21	**13.22** ± 0.44	**31.33** ± 0.15
ADA	12.67 ± 0.31	31.89 ± 0.17	9.44 ±0.25	43.21 ± 0.37	8.54 ± 0.26	56.83 ± 0.48
+ **KD-DLGAN (Ours)**	**14.14** ± 0.32	**20.32** ± 0.10	**14.06** ± 0.39	**22.35** ± 0.11	**14.65** ± 0.47	**28.79** ± 0.14

Table 3. Comparison with the state-of-the-art over ImageNet [11]: KD-DLGAN achieves the best performance consistently and complements the state-of-the-art. Besides, KD-DLGAN outperforms vanilla knowledge distillation in DA + KD (CLIP [36]) consistently as well. All the compared methods employ BigGAN [5] as backbone. We report IS(↑) and FID(↓) averaged over three runs.

DLGAN outperforms DA + KD (CLIP [36]) consistently as well, indicating that the performance gain in KD-DLGAN is largely attributed to our generative knowledge distillation designs instead of solely from the powerful vision-language model. Table 1 also tabulates the results of KD-DLGAN when implementing over four state-of-the-art data-limited generation approaches including LeCam-GAN [40], InsGen [44], APA [19] and ADA [21]. We can see that KD-DLGAN complement all the state-of-the-art consistently, demonstrating the superior generalization and complementary property of our proposed KD-DLGAN.

Fig. 3 shows qualitative comparison with DA [53]. It can be observed that KD-DLGAN clearly outperforms the state-of-the-art in the data-limited generation, especially in term of the generated shapes and textures.

4.3. Experiments with BigGAN

Table 2 and Table 3 show the conditional image generation results on CIFAR-10, CIFAR-100 and ImageNet, respectively. All models employ BigGAN [5] as the backbone. CIFAR-10 and and CIFAR-100 are trained with 100% (50K images), 20% (10K images) or 10% (5K images) training data where the FIDs are evaluated over the validation sets (10K images). ImageNet is trained with 10% (~100K images), 5% (~50K images) and 2.5% (~25K images), where the evaluations are performed over the whole training set (~1.2M images).

The experiments show that our KD-DLGAN outperforms the state-of-the-art substantially. The superior 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. We also show the results of vanilla knowledge distillation from the powerful vision-language model CLIP [36] in Row 1 of Table 2 and Table 3. We can see that KD-DLGAN outperforms the vanilla knowledge distillation method by a large margin, indicating that the performance gain is largely attributed to our generative knowledge distillation design instead of the powerful vision-language model. In addition, Table 2 and

Method	AGKD	CGKD	CIFAR-10		100-shot	
			20% data	10% data	Obama	Grumpy Cat
DA [53] (Baseline)			14.53	23.34	46.87	27.08
	✓		12.97 ± 0.08	15.85 ± 0.06	35.51± 0.25	23.24 ± 0.16
		✓	12.77 ± 0.08	18.66 ± 0.09	36.18 ± 0.22	23.17 ± 0.11
Ours	✓	✓	**11.01** ± 0.07	**14.20** ± 0.06	**31.54** ± 0.27	**20.13** ± 0.13

Table 4. Quantitative ablation study of KD-DLGAN: AGKD and CGKD in KD-DLGAN both improves the generation performance over the baseline DA [53]. KD-DLGAN performs the best as AGKD and CGKD complement each other. The FIDs (↓) are averaged over three runs.

Figure 4. Qualitative ablation study over 100-shot Obama: AGKD (Row 2) and CGKD (Row 3) can generate more realistic images than the baseline DA [53] (Row 1), the state-of-the-art in data-limited image generation. KD-DLGAN combining AGKD and CGKD generates the most realistic images.

Table 3 also show the results of KD-DLGAN when implementing over the state-of-the-art data-limited image generation approaches. KD-DLGAN complements the state-of-the-art and improves the generation performance greatly.

4.4. Ablation study

The proposed KD-DLGAN consists of two generative knowledge distillation (KD) techniques, namely, AGKD and CGKD. The two techniques are separately evaluated to demonstrate their contributions to the overall generation performance. As Table 4 shows, including either AGKD or CGKD clearly outperforms DA [53], the state-of-the-art in data-limited image generation, demonstrating the effectiveness of the proposed AGKD and CGKD in mitigating the discriminator overfitting and improving the generation performance. In addition, combining AGKD and CGKD leads to the best generation performance which shows that the two KD techniques complement each other.

Qualitative ablation studies in Fig. 4 show that the proposed AGKD and CGKD can produce clearly more realistic generation than the baseline, demonstrating the effectiveness of these two generative knowledge distillation techniques. In addition, KD-DLGAN combining AGKD and CGKD performs the best, which further verifies that AGKD and CGKD are complementary to each other.

4.5. Discussion

In this subsection, we analyze our KD-DLGAN from several perspectives. All the experiments are based on the CIFAR-10 and CIFAR-100 dataset with 10% data.

Generation Diversity: The proposed KD-DLGAN improves the generation diversity by enforcing the diverse correlation between images and texts, which eventually improves the generation performance. In this subsection, we evaluate the generation diversity with LPIPS [52]. Higher LPIPS means better diversity of generated images. As Table 5 shows, the proposed KD-DLGAN outperforms the baseline DA [53], the state-of-the-art in data-limited image generation, demonstrating the effectiveness of KD-DLGAN in improving generation diversity. We also show the results of KD-DLGAN without CGKD; it further verifies that the improved generation diversity is largely attributed to the CGKD, which distills diverse image-text correlation from CLIP [36] to the discriminator, ultimately improving the generation performance. Note we choose Alexnet model with linear configuration for LPIPS evaluation.

Generalization of KD-DLGAN: We study the generalization of our KD-DLGAN by performing experiments with different GAN architectures, generation tasks and the number of training samples. Specifically, as shown in Table 1-3, we perform extensive evaluations over BigGAN and StyleGAN-v2. Meanwhile, we benckmark KD-DLGAN on object generation tasks with CIFAR and ImageNet and face generation tasks with 100-shot and AFHQ. Besides, we perform extensive evaluations on 100-shot and AFHQ with few hundred samples, CIFAR with 100%, 20% and 10% data, ImageNet with 10%, 5% and 2.5% data.

Method	CIFAR-10 10% data	CIFAR-100 10% data
DA [53] (Baseline)	0.202	0.236
KD-DLGAN w/o CKGD	0.204	0.237
KD-DLGAN	**0.221**	**0.264**

Table 5. KD-DLGAN improves the generation diversity clearly. And the improvement is largely attributed to CGKD, which distills diverse image-text correlations from CLIP [36] to the discriminator. We report LPIP (↑) averaged over three runs.

Method	CIFAR-10 10% data	CIFAR-100 10% data
DA [53] (Baseline)	23.34 ± 0.09	35.39 ± 0.08
Fitnets [37]	22.03 ± 0.07	33.93 ± 0.09
Label Distillation [17]	20.46 ± 0.10	34.14 ± 0.11
PKD [35]	21.34 ± 0.08	32.15 ± 0.13
SPKD [41]	19.11 ± 0.07	31.97 ± 0.10
KD-DLGAN (Ours)	**14.20** ± 0.06	**18.03** ± 0.11

Table 6. KD-DLGAN outperforms the state-of-the-art knowledge distillation methods by large margins, demonstrating the effectiveness of the two generative knowledge distillation techniques designed specifically for data-limited image generation. We report FID(↓) averaged over three runs.

Comparison with state-of-the-art knowledge distillation methods: KD-DLGAN is the first to explore the idea of knowledge distillation in data-limited image generation. To validate the superiority of our designs, we compare KD-DLGAN with state-of-the-art knowledge distillation methods designed for other tasks in Table 6. It shows that our KD-DLGAN outperforms the state-of-the-art knowledge distillation approaches consistently by large margins. The superior generation performance demonstrates the effectiveness of our designed generative knowledge distillation techniques for data-limited image generation.

Comparison with other Vision-Language teacher models: KD-DLGAN adopted CLIP [36] as the teacher model for knowledge distillation. We perform experiments to study how different vision-language models affect the generation performance. As shown in Table 7, different vision-language models produce quite similar FIDs. We conjecture that these pertrained models provide sufficient vision-language information for distillation and the performance gain mainly comes from our designed generative knowledge distillation techniques instead of the selected teacher model.

Comparison with other CLIP-based methods: KD-DLGAN distills knowledge from CLIP [36] to the discriminator with two novelly designed generative knowledge distillation techniques while Vision-aided GAN [25] employs off-the-shelf models as additional discriminators for data-limited generation. Table 8 compares KD-DLGAN with Vision-aided GAN. We can observe that KD-DLGAN outperforms CLIP-based Vision-aided GAN consistently, demonstrating its effectiveness in mitigating discriminator overfitting and improving generation performance.

Method	CIFAR-10 10% data	CIFAR-100 10% data
DA [53] (Baseline)	23.34 ± 0.09	35.39 ± 0.08
+ TCL [46]	14.98 ± 0.09	18.43 ± 0.12
+ BLIP [26]	15.74 ± 0.10	18.88 ± 0.11
+ CLIP [36] (Ours)	**14.20** ± 0.06	**18.03** ± 0.11

Table 7. Employing different pretrained vision-language models as teacher models, the results are similar. We report FID(↓) averaged over three runs.

Method	CIFAR-10 10% data	CIFAR-100 10% data
DA [53] (Baseline)	23.34 ± 0.09	35.39 ± 0.08
Vision-aided GAN [25]	16.24 ± 0.08	19.11 ± 0.10
KD-DLGAN (Ours)	**14.20** ± 0.06	**18.03** ± 0.11

Table 8. KD-DLGAN outperforms CLIP-based Vision-add GAN [25], demonstrating the effectiveness of our KD-DLGAN in mitigating discriminator overfitting and improving the generation performance. We report FID(↓) averaged over three runs.

5. Conclusion

In this paper, we present KD-DLGAN, a novel data-limited image generation framework that introduces knowledge distillation for effective GAN training with limited data. We design two novel generative knowledge distillation techniques, including aggregated generative knowledge distillation (AGKD) and correlated generative knowledge distillation (CGKD). AGKD mitigates the discriminator overfitting by forcing harder learning tasks and distilling more general knowledge from CLIP. CGKD improves the generation diversity by distilling and preserving the diverse image-text correlation within CLIP. Extensive experiments show that both AGKD and CGKD can improve the generation performance and combining them leads to the best performance. We also show that KD-DLGAN complements the state-of-the-art data-limited generation methods consistently. Moving forward, we will explore KD-DLGAN in more tasks such as image translation and editing.

Acknowledgement

This study is funded by the Ministry of Education Singapore, under the Tier-1 scheme with a project number RG94/20, as well as cash and in-kind contribution from Singapore Telecommunications Limited (Singtel), through Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU).

References

[1] Unsupervised multi-target domain adaptation through knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1339–1347, 2021. 2

[2] Alex Andonian, Shixing Chen, and Raffay Hamid. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16430–16441, June 2022. 1, 3

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 2

[4] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27, 2014. 2

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 5, 6

[6] Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 2

[7] Akshay Chawla, Hongxu Yin, Pavlo Molchanov, and Jose Alvarez. Data-free knowledge distillation for object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3289–3298, 2021. 2

[8] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017. 2

[9] Ruizhe Cheng, Bichen Wu, Peizhao Zhang, Peter Vajda, and Joseph E. Gonzalez. Data-efficient language-supervised zero-shot learning with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3119–3124, June 2021. 1, 3

[10] Kaiwen Cui, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, Fangneng Zhan, and Shijian Lu. Genco: generative co-training for generative adversarial networks with limited data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 499–507, 2022. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5, 6

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1, 2

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017. 2

[14] Ishaan Gulrajani, Colin Raffel, and Luke Metz. Towards gan benchmarks which require generalization. *arXiv preprint arXiv:2001.03653*, 2020. 2

[15] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016. 2

[16] 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. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 5

[17] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2, 8

[18] Jiaxing Huang, Kaiwen Cui, Dayan Guan, Aoran Xiao, Fangneng Zhan, Shijian Lu, Shengcai Liao, and Eric Xing. Masked generative adversarial networks are data-efficient generation learners. In *Advances in Neural Information Processing Systems*. 2

[19] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Deceive d: Adaptive pseudo augmentation for gan training with limited data. *Advances in Neural Information Processing Systems*, 34:21655–21667, 2021. 2, 5, 6

[20] 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. 2

[21] 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. 2, 5, 6

[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. 5

[23] Ali Koksal and Shijian Lu. Rf-gan: A light and reconfigurable network for unpaired image-to-image translation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[25] Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10651–10662, June 2022. 2, 8

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 8

[27] Steven Liu, Tongzhou Wang, David Bau, Jun-Yan Zhu, and Antonio Torralba. Diverse image generation via self-conditioned gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14286–14295, 2020. 2

[28] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for

semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. 2

[29] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14074–14083, June 2022. 1, 3

[30] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 2

[31] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 2

[32] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018. 2

[33] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze discriminator: A simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020. 2

[34] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019. 2

[35] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–284, 2018. 8

[36] 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. 1, 2, 3, 4, 5, 6, 7, 8

[37] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2, 8

[38] 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. 5

[39] 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. 5

[40] Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularizing generative adversarial networks under limited data. 2021. 1, 2, 5, 6

[41] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019. 2, 8

[42] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 939–948, June 2022. 1, 3

[43] Ryan Webster, Julien Rabin, Loic Simon, and Frédéric Jurie. Detecting overfitting of deep generative networks via latent recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11273–11282, 2019. 2

[44] Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. Data-efficient instance generation from instance discrimination. *Advances in Neural Information Processing Systems*, 34:9378–9390, 2021. 2, 5, 6

[45] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12319–12328, 2022. 2

[46] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 8

[47] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jiahui Zhang, Shijian Lu, Miaomiao Cui, Xuansong Xie, Xian-Sheng Hua, and Chunyan Miao. Towards counterfactual image manipulation via clip. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3637–3645, 2022. 2

[48] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13876–13885, 2020. 2

[49] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15028–15038, 2021. 2

[50] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 2

[51] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[53] 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, 5, 6, 7, 8

[54] Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. *Advances in neural information processing systems*, 31, 2018. 2