Discriminator-Cooperated Feature Map Distillation for GAN Compression

Tie Hu[1], Mingbao Lin[3], Lizhou You[1], Fei Chao[1], Rongrong Ji[1,2,4]*
[1]MAC Lab, School of Informatics, Xiamen University
[2]Institute of Artificial Intelligence, Xiamen University
[3]Tencent Youtu Lab
[4]Shenzhen Research Institute of Xiamen University
{hutie, lmbxmu, youlizhou}@stu.xmu.edu.cn, {fchao, rrji}@xmu.edu.cn

Abstract

Despite excellent performance in image generation, Generative Adversarial Networks (GANs) are notorious for its requirements of enormous storage and intensive computation. As an awesome "performance maker", knowledge distillation is demonstrated to be particularly efficacious in exploring low-priced GANs. In this paper, we investigate the irreplaceability of teacher discriminator and present an inventive discriminator-cooperated distillation, abbreviated as DCD, towards refining better feature maps from the generator. In contrast to conventional pixel-to-pixel match methods in feature map distillation, our DCD utilizes teacher discriminator as a transformation to drive intermediate results of the student generator to be perceptually close to corresponding outputs of the teacher generator. Furthermore, in order to mitigate mode collapse in GAN compression, we construct a collaborative adversarial training paradigm where the teacher discriminator is from scratch established to co-train with student generator in company with our DCD. Our DCD shows superior results compared with existing GAN compression methods. For instance, after reducing over 40× MACs and 80× parameters of CycleGAN, we well decrease FID metric from 61.53 to 48.24 while the current SoTA method merely has 51.92. This work's source code has been made accessible at https://github.com/poopit/DCD-official.

1. Introduction

Image generation transforms random noise or source-domain images to other images in user-required domains. Recent years have witnessed the burgeoning of generative adversarial networks (GANs) that lead to substantial progress in image-to-image translation [8, 9, 18, 49], style transfer [11, 12, 42], image synthesis [3, 22, 23, 32, 46], *etc.*

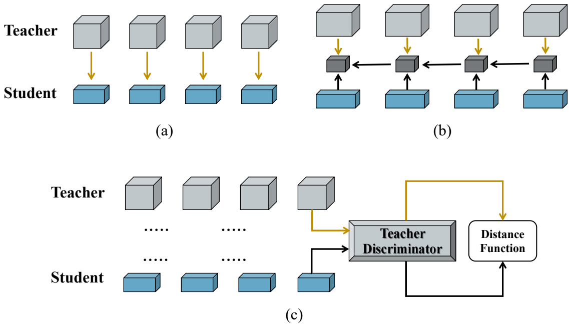

Figure 1. (a) Layer-by-layer feature map distillation [34]. (b) Cross-layer feature map distillation [6]. (c) Our discriminator-cooperated feature map distillation.

Image generation has a wide application in daily entertainment such as TikTok AI image generator, Dream by WOMBO, Google Imagen, and so on. Running platforms performing these applications are typically featured with poor memory storage and limited computational power. However, GANs are also ill-famed for the growing spurt of learnable parameters and multiply-accumulate operations (MACs), raising a huge challenge to the storage requirement and computing ability of deployment infrastructure.

To address the above dilemma for better usability of GANs in serving human life, methods such as pruning [7, 27, 28, 33], neural network architecture search (NAS) [10, 19, 26] and quantization [39, 40], have been broadly explored to obtain a smaller generator. On the premise of these compression researches, knowledge distillation, in particular to distilling feature maps, has been accepted as a supplementary means to enhance the performance of compressed generators [1, 4, 17, 26, 29, 41]. Originated from image classification, as illustrated in Fig. 1(a), feature map based distillation, which extracts information of intermediate activations and transfers the knowledge from the teacher model to the student one, has been extensively explored and demonstrated to well improve the capability of lightweight mod-

[*]Corresponding Author

els [5, 25, 34, 43, 45]. Distinctive from passing on common feature maps from teacher to student, AT [45] calculates feature attentions as the delivered knowledge; MGD [43] randomly masks feature maps to indirectly guide the student to learn from the teacher; KRD [6] uses a cross-layer distillation method to allow the "new knowledge" of the student to learn from the "old knowledge" in teacher, as shown in Fig. 1(b). Whatever, most methods execute pixel-to-pixel feature maps matching between teacher and student.

Alike to the implementations on image classification, feature map based distillation is also considered in GAN compression. For example, GCC [28] considers a well pre-trained discriminator to absorb high-level information from the teacher-generated image, and fuses it with intermediate activations from the teacher generator, results of which are passed to the corresponding position of the student generator. OMGD [33] utilizes an online multi-granularity strategy to allow a deeper teacher generator and a wider one to simultaneously deliver output image knowledge of different granularities to the student generator.These two methods follow the pipeline of image classification to tune the intermediate outputs of student generator with those of teacher generator in a fashion of per-pixel matching. Although the sustainable progress on multiple benchmark datasets demonstrates the efficacy of intermediate activation outputs, the feature-based distillation, as we reveal in this paper, is not well compatible with the very nature of generating perceptually similar images and adversarial training paradigm.

Concretely speaking, conversely to image classification that relies on feature vector representations, the essence of image generation is to improve perceptually alike between the real images and generated images. Two important facts cause it is eventually difficult to use per-pixel match to analyze a pair of images: First, two similar images can contain many different pixel values; Second, two dissimilar images can still comprise the same pixel values. Thus, it is not suitable to simply use the per-pixel match. Regarding adversarial training in GANs, a generator learns to synthesize samples that best resemble the dataset, meanwhile a discriminator differentiate samples in the dataset from the generator generated samples. The adversarial results finally lead the generator to creating images of out-of-the-ordinary visual quality, indicating that the discriminator is also empowered with informative capacity and can be exploited to enrich the distillation of feature maps. Therefore, it might be inappropriate to directly extend feature map distillation in image classification to image generation. And GAN compression oriented feature map distillation with discriminator included remains to be well explored.

In order to achieve this objective, in this paper, we propose a discriminator-cooperated distillation (DCD) method to involve the teacher discriminator in distilling feature maps for student generator. A simple illustration is given in Fig. 1(c), in contrast to the vanilla pixel-wise distance constraint, our DCD measures the distance at the end of teacher discriminator with the intermediate generator outputs as its inputs. Our DCD is perspicacious in multiple benchmark datasets with a simple implementation. Akin to perceptual loss [20] which employs a pre-trained neural network such as a VGG model [36] to extract features upon which the ℓ_1 distance is calculated from activations of hidden layers, the teacher discriminator in DCD also acts as a feature extractor. Due to pooling operations in the hidden layers, feature maps from different sources (student generator and teacher generator) as inputs to the discriminator may lead to identical latent representations, therefore encouraging natural and perceptually pleasing results. In addition, the proposed DCD is used in conjunction with collaborative adversarial training, which is also simple but perspicacious to allow the student generator to fool the discriminator for generating better images. In contrast to discriminator-free paradigm training [28], we find our DCD empowers the compressed student generator with a better capability to compete against teacher discriminator. Thus, we also employ the teacher discriminator to collaboratively determine whether inputs from the student generator are real or not.

This work intends to raise the level of feature map distillation to strengthen the compressed student generator to generate high-quality images. The major contributions we have made across the entire paper are listed as follows: (1) An incentive GAN-oriented discriminator-cooperated feature map distillation method to produce images with high fidelity; (2) One novel collaborative adversarial training paradigm to better reach a global equilibrium point in compressing GANs; (3) Remarkable reduction on the generator complexity and significant performance increase.

2. Related Work

2.1. GANs and GAN Compression

Generative adversarial networks (GANs) [13] have attracted the attention of substantive researchers due to their outstanding performance in image generation tasks [3, 8, 9, 18, 22, 23, 32, 46, 49]. Since the infancy of GAN, many variants have emerged, from DCGAN [32], which embraces convolutional neural networks for the first time, to CycleGAN [49] and Pix2Pix [18] which implement image-to-image translation, to StyleGAN [22] to enable controllable manipulation of various attributes of image synthesis. CycleGAN [49] transfers a source-domain image into a target style in an unpaired configuration, such as a horse image to a zebra pattern or a summer image to a winter style. On the other hand, Pix2Pix [18] is given a ground-truth image and converts a semantic segmentation or contour map into a photo-realistic picture. Albeit the performance, GANs suffer heavy burden on storage and computation.

Therefore, recent years have witnessed increasing attention on compressing GANs [7, 10, 19, 21, 24, 26, 27, 30, 35, 39, 40, 47]. Co-evolution [35] prunes filters in the generator under the constraint of consistent output distributions. GAN Slimming [39] suggests a compression framework that can integrate pruning, knowledge distillation and quantization. GAN Compression [26] views the teacher output as the pseudo label for student generator and unifies the compression framework for GANs trained on paired data and unpaired data. GCC [28] combines information from discriminator in knowledge distillation, an idea closer to our approach, but the use of discriminator is still under-explored. Although the above approaches have compressed parameters and computation costs, both the generated image quality and the model size are still far from practical applications on mobile devices. OMGD [33] bridges this huge gap in a way by constructing discriminator-free distillation.

Of the above methods, GCC is the most similar to our approach with major differences in: (1) GCC involves summation of discriminator activations in distilling intermediate layers of student generator. Our DCD calculates distance between teacher and student directly upon feature maps of the discriminator. (2) GCC focuses on a more suitable student discriminator for student generator. Our DCD discards student discriminator by adopting that of teacher. (3) Results show merits of our DCD over GCC in both quantitative and visual quality with a higher compression rate.

2.2. Knowledge Distillation

Pioneered by FitNet as of 2014 [34], knowledge distillation (KD) [16] has become a regular approach in model compression especially to GAN compression, where a larger model with better performance imparts knowledge to a smaller model. Since then, great efforts have been made to dig out opulent knowledge hints, such as output logits [21, 47, 48], intermediate feature maps [6, 14, 34], instance relation [31, 38] and so on. In this paper, we are mainly inspired by the intermediate feature map based distillation that has been extensively mined to efficiently guide the training of the student network. Compared to other knowledge hints, feature maps often accommodate a richer level of information and provide more detailed guidance for the student network. AT [45] extracts the attention map from the feature maps and trains the student's attention maps to be as close as possible to the one of teacher. MGD [43] transforms direct learning from the teacher into a generative intermediate goal. KRD [6] uses the idea of cross-layer distillation to allow old knowledge from the teacher to guide new knowledge learning from the student network. Nevertheless, conventional methods require per-pixel match between feature maps of teacher and student, which does not fit well in GAN compression because the goal of GANs is to generate perceptually similar images. In this paper, we dig deeper into the distinctiveness of generative adversarial networks.

3. Methodology

3.1. Preliminaries

Generative adversarial networks [13], or GANs for short, are an interesting manner to train a generative model by modelling the problem with two sub-models including a generator model $\mathcal{G}$ and a discriminator model $\mathcal{D}$. The two models are trained together in a zero-sum game as:

$$\begin{aligned}\min_{\mathcal{G}}\max_{\mathcal{D}}\mathcal{L}_{gan} = \mathbb{E}_{y\sim p_{real}}\log\mathcal{D}(y)\\ +\mathbb{E}_{x\sim p(x)}\Big[\log\Big(1-\mathcal{D}\big(\mathcal{G}(x)\big)\Big)\Big].\end{aligned} \tag{1}$$

Herein, the generator model $\mathcal{G}$ is trained to generate new examples, and the discriminator model $\mathcal{D}$ tries to classify examples as either real (from the domain) or fake (generated). The two models are trained adversarially until the discriminator model is fooled at most times, which indicates that the generator model is producing plausible examples. Then, the generator $\mathcal{G}$ is deployed online to complete service for reality.

The serviceability of generator $\mathcal{G}$ rests with not only the performance, but also the hardware capability that makes a greater demand on generator complexity. Therefore, a lighter student generator, $\mathcal{G}^S$, can be developed by various methods. The original generator and discriminator, respectively denoted as $\mathcal{G}^T$ and $\mathcal{D}^T$ in this situation to differentiate, play as a teacher to enhance the ability of student generator $\mathcal{G}^S$.

Giving an input variable $x \sim p(x) \in \mathbb{R}^{H\times W\times C}$, we denote the i-th layer output of generator as $\mathcal{G}_i(x)$, and $\mathcal{I}_{\mathcal{G}}(x)$ as layer index set of extracted intermediate outputs. Thus, the feature maps based distillation is formulated as:

$$\mathcal{L}_{fea-dis} = \sum_{i\in I_{\mathcal{G}}}\ell\Big(\mathcal{G}_i^T(x), f\big(\mathcal{G}_i^S(x)\big)\Big), \tag{2}$$

where $f(\cdot)$ is the affine transformation function to align the channel dimensions between the teacher and student, such as 1×1 convolution operators in existing studies [26, 28, 33]. Also, $\ell(\cdot,\cdot)$ refers to the distance measure function, such as Euclidean distance.

Additionally, perceptual loss [20], $\mathcal{L}_{per}$, is also widely-adopted in existing studies to encourage natural and perceptually pleasing restored images. $\mathcal{L}_{per}$ comprises a feature reconstruction loss $\mathcal{L}_{fea}$ and a style reconstruction loss $\mathcal{L}_{sty}$ as:

$$\mathcal{L}_{per} = \lambda_{fea}\cdot\mathcal{L}_{fea} + \lambda_{sty}\cdot\mathcal{L}_{sty}. \tag{3}$$

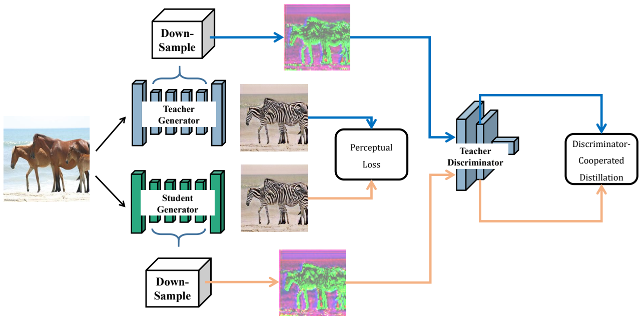

Figure 2. Framework of our DCD. Intermediate feature maps from student and teacher are downsampled first to align the dimension, results of which are then fed to the teacher discriminator to minimize distance for a perceptually vivid generated image.

Herein, $\mathcal{L}_{fea}$ propels the output representation of the teacher generator to approach to that of the student generator. This is achieved by a pre-trained VGG network $\Phi(\cdot)$ [36] and formalized as:

$$\mathcal{L}_{fea} = \sum_{j\in I_\Phi} \frac{1}{H_j W_j C_j} \left\| \Phi_j\left(\mathcal{G}^T(x)\right) - \Phi_j\left(\mathcal{G}^S(x)\right) \right\|_1, \quad (4)$$

where $\Phi_j(\cdot)$ returns the j-th activation output of VGG network and $H_j \times W_j \times C_j$ is its shape. Alike to $I_\mathcal{G}$, I_Φ contains layer index of extracted intermediate outputs.

As for $\mathcal{L}_{sty}$, it minimizes the difference between Gram matrices of the output and target images in order to preserve style characteristics such as color, textures and common pattern [33]. The $\mathcal{L}_{sty}$ is calculated as:

$$\mathcal{L}_{sty} = \sum_{j\in I_\Phi} \left\| G\Big(\Phi_j\left(\mathcal{G}^T(x)\right)\Big) - G\Big(\Phi_j\left(\mathcal{G}^S(x)\right)\Big) \right\|_1, \quad (5)$$

where $G(\cdot)$ represents abbreviation of Gram matrices.

3.2. Discriminator-Cooperated Distillation

Stepping back and reflecting on the feature map based distillation in Eq. (2), we realize that a simple utilization of generator capacity is not intact in earlier methods. The central principle of a GAN is contingent on an "indirect" training route through the discriminator updated dynamically to discern how "realistic" its input (*i.e.*, generator output) seems. This means that the generator is not trained to minimize the distance from a generated image to a target image, but rather to deceive the discriminator. It is the coopetition pattern between the generator and discriminator that even brings about superficially authentic generated images. Thus, the discriminator is also empowered with informative capacity and must be utilized to enrich the distillation of feature maps.

As shown in Fig. 2, we rethink Eq. (2) and integrate teacher discriminator to cooperate with the distillation process. Alike to feature reconstruction loss defined in Eq. (4), while taking the generator outputs as its inputs, we accomplish our distillation by aligning the intermediate outputs of the discriminator. We formulate this learning process as:

$$\mathcal{L}_{dcd} = \sum_{k\in I_\mathcal{D}} \sum_{i\in I_\mathcal{G}} \ell\bigg(\mathcal{D}_k^T\Big(f\big(\mathcal{G}_i^T(x)\big)\Big), \mathcal{D}_k^T\Big(f\big(\mathcal{G}_i^S(x)\big)\Big)\bigg), \quad (6)$$

where $\mathcal{D}_k^T$ stands for the output of its k-th layer. Similar to $I_\mathcal{G}$, $I_\mathcal{D}$ is a layer index set of the teacher discriminator; herein, $f(\cdot)$ denotes 1×1 convolution operations to downsample the channel dimensions of both the teacher and student feature maps to those of discriminator input. Usually, the channel number of discriminator input is set to three for an RGB-encoded image. Here, $f(\cdot)$ for teacher generator becomes constant once initialized while that for student generator continues updating for a better fit with teacher.

As an analogy to the vanilla feature map based distillation in Eq. (2) that compels the per-pixel match between intermediate outputs of both the teacher generator and student generator, the role of the teacher discriminator resembles the pre-trained VGG in Eq. (4), which acts as a transformed network to enable the intermediate results of student generator to be perceptually alike to these of teacher generator. In contrast, it does not require them to do pixel-by-pixel match, since two images might look similar in perspective, but they often have different pixel values, thus we cannot depend on per-pixel match. Notice that our discriminator-cooperated distillation is complementary to perceptual loss. The former concentrates on object localization while the latter pays attention to style discrepancies. The analysis of this phenomenon is in Sec. 4.4.

3.3. Collaborative Adversarial Training

GANs perform an alternating training paradigm for a unique global equilibrium point: 1) Training the discriminator $\mathcal{D}$ to identify real and generated data while keeping the generator constant; 2) Training the generator $\mathcal{G}$ to generate vivid data that fools the discriminator while keeping the discriminator constant; 3) Repeating steps 1) and 2) till the discriminator model is fooled at most time. However, the equilibrium is no longer guaranteed when $\mathcal{D}$ and $\mathcal{G}$ are empowered with inconsistent abilities, in which case unstable convergence frequently occurs. Generally, the instability stems from two ill-famed issues including vanishing gradient where loss for the generator is zero when the discriminator is perfect, and mode collapse where the stronger generator produces a small set of outputs for any input and the weaker discriminator traps in a local minimum [2].

Especially, the mode collapse issue is even widespread in GAN compression because the compressed student generator $\mathcal{G}^S$ is powerless to compete with the original full discriminator [27], in particular a pre-trained one [4, 19, 30]. The community has excavated various approaches to weaken the student discriminator. For example, GCC [28] selectively activates discriminator neurons. Nevertheless, a rule-of-thumb selection has to be carefully designed. Also, training a student discriminator is computationally redundant since it is unwanted in the testing stage. OMGD [33] co-trains teacher generator with the teacher discriminator while the student generator is discriminator-free. The missing adversarial training of student generator somehow barricades the further performance increase.

In this paper, we also reject real student discriminator, and to ensure the teacher's optimization, the teacher discriminator, $\mathcal{D}^T$, is online trained from scratch to be well-matched with the teacher generator $\mathcal{G}^T$. Comparing to [33], teacher discriminator $\mathcal{D}^T$ also appears in the form of a collaborative discriminator to determine if the inputs from student generator $\mathcal{G}^T$ are real or fake. The major concern originates from the actuality that the teacher discriminator is much powerful than the compressed student generator, which causes mode collapse in adversarial training. Luckily, our discriminator-cooperated distillation in Eq. (6) furnishes student generator with increasing capability to battle against teacher discriminator and leads to better performance than [33] as demonstrated in the experiment.

Based on Eq. (1), our adversarial training is rewritten as:

$$\begin{aligned}\min_{\mathcal{G}^T,\mathcal{G}^S}\max_{\mathcal{D}^T}\mathcal{L}_{col} &= \mathbb{E}_{y\sim p_{real}}\log\mathcal{D}^T(y)\\ &+\mathbb{E}_{x\sim p(x)}\Big[\log\Big(1-\mathcal{D}^T\big(\mathcal{G}^T(x)\big)\Big)\Big]\\ &+\lambda_{stu}\cdot\mathbb{E}_{x\sim p(x)}\Big[\log\Big(1-\mathcal{D}^T\big(\mathcal{G}^S(x)\big)\Big)\Big],\end{aligned} \tag{7}$$

where λ_{stu} refers to a trade-off parameter; and $\lambda_{stu} = 0$ degenerates to discarding student discriminator when co-trained with the student generator [33].

3.4. Training Objective

Looking back to Sec. 3.1, the loss terms in most conventional feature map based distillation methods include $\mathcal{L}_{gan}$ in Eq. (1), $\mathcal{L}_{fea-dis}$ in Eq. (2) and $\mathcal{L}_{per}$ in Eq. (3). In this paper, we improve $\mathcal{L}_{fea-dis}$ through a discriminator-cooperated feature map distillation loss $\mathcal{L}_{dcd}$ in Eq. (6), and $\mathcal{L}_{gan}$ through collaborative adversarial training loss $\mathcal{L}_{col}$ in Eq. (7). Therefore, the overall training objective in this paper is given in the following:

$$\min_{\mathcal{G}^T,\mathcal{G}^S}\max_{\mathcal{D}^T}\left(\mathcal{L}_{gan}+\mathcal{L}_{per}+\lambda_{dcd}\cdot\mathcal{L}_{dcd}\right), \tag{8}$$

where λ_{dcd} balances the loss term. Four hyper-parameters: λ_{fea}, λ_{sty}, λ_{stu} and λ_{dcd} are used in this paper, influence of each parameter is ablated in the appendix.

4. Experimentation

4.1. Setups

GAN Models and Benchmarks. We present the performance of compressed CycleGAN [49] and Pix2Pix [18] to follow and compare with existing methods [7, 19, 21, 26, 27, 35, 39, 47]. For a fair comparison with the current SoTA OMGD [33], the compressed generators (students) consist of 1/4 channels of the original full ResNet generators (teachers) [26]. CycleGAN translates images from one domain to another without a one-to-one mapping between the source and target domain. Therefore, we verify the performance upon unpaired image translation benchmarks including horse2zebra [49] and summer2winter [49]. As for Pix2Pix, we perform distillation upon paired edges2shoes [44] because it requires learning a mapping from input images to output images.

Evaluations. Fréchet Inception Distance (FID), or FID for short [15], is particularly developed to evaluate the performance of GANs. It accesses the quality of generated images by an Inception-V3 network [37] to separately embed synthetic and real images to feature space, and then calculate the Wasserstein distance of their distributions. A lower FID score indicate better quality of generated images.

Implementations. We train CycleGAN and Pix2Pix for a total of 100 epochs. The initial learning rate is given as $2\times e^{-4}$ and then linearly decayed to 0 as training goes. The batch size is set to 4 on edges2shoes, and 1 on horse2zebra [49] and summer2winter [49]. We have λ_{dcd}=1, λ_{fea}=1×e^1, λ_{sty}=1×e^4 and λ_{stu}=1 across all experiments. In the appendix, specific ablations in regard to these hyper-parameters have been provided.

4.2. Quantitative Comparison

GAN Compression. We first compare with existing implementations on GAN compression in Table 1 where the blue number indicates performance drop, while the red number denotes performance increase compared with the original GAN models. We conclude the following observations from Table 1: First, on summer2winter, all methods lead to a performance increase of more or less. On the contrary, most methods cause FID drops on the challenging horse2zebra and edges2shoes while methods such as OMGD [33] and our DCD consistently enhance the performance. Second, our DCD leads to the most complexity reduction *w.r.t.* MACs and parameters, meanwhile it gains the best performance increase on all the three benchmark datasets. Third, with the same reduction of MACs (40.3×) and parameters (82.5×) on unpaired CycleGAN, our DCD well outperforms the recent SoTA method, *i.e.*, OMGD. In particular, we drastically increase the performance of 51.92 for OMGD to 48.24, which is also 13.29 increase compared to the original CycleGAN. This increase is very challenging

Table 1. Performance comparison when compressing CycleGAN on horse2zebra, and Pix2Pix on edges2shoes.

Model	Dataset	Method	MACs	#Parameters	FID(↓)
CycleGAN [49]	horse2zebra [49]	Original [49]	56.80G(1.0×)	11.30M(1.0×)	61.53 (-)
		Co-Evolution [35]	13.40G(4.2×)	-	96.15(-34.62)
		DMAD [27]	2.41G(23.6×)	0.28M(40.0×)	62.96(-1.43)
		Wavelet KD [47]	1.68G(33.8×)	0.72M(15.81×)	77.04(-15.51)
		GAN-Compression [26]	2.67G(21.3×)	0.34M(33.2×)	64.95(-3.42)
		GCC [28]	2.40G(23.6×)	-	59.31(+2.22)
		OMGD [33]	1.408G(40.3×)	0.137M(82.5×)	51.92(+9.61)
		DCD (Ours)	**1.408G(40.3×)**	**0.137M(82.5×)**	**48.24(+13.29)**
	summer2winter [49]	Original [49]	56.80G(1.0×)	11.30M(1.0×)	79.12(-)
		Co-Evolution [35]	11.10G(5.1×)	-	78.58(+0.54)
		AutoGAN-Distiller [10]	4.34G(13.1×)	-	78.33(+0.79)
		DMAD [27]	3.18G(17.9×)	0.30M(37.7×)	78.24(+0.88)
		OMGD [33]	1.408G(40.3×)	0.137M(82.5×)	73.79(+5.33)
		DCD (Ours)	**1.408G(40.3×)**	**0.137M(82.5×)**	**73.63(+5.49)**
Pix2Pix [18]	edges2shoes	Original [18]	18.60G(1.0×)	54.40M(1.0×)	34.31(-)
		DMAD [27]	2.99G(6.2×)	2.13M(25.5×)	46.95(-12.64)
		Wavelet KD [47]	1.56G(11.92×)	13.61M(4.00×)	80.13(-45.82)
		OMGD [33]	1.219G(15.3×)	3.404M(16.0×)	25.00(+9.41)
		DCD (Ours)	**1.219G(15.3×)**	**3.404M(16.0×)**	**23.43(+10.98)**

Table 2. Comparison between different feature map distillation methods on horse2zebra.

Dataset	Method	FID(↓)
horse2zebra [49]	Baseline	65.13
	FitNet [34]	65.84
	MGD [43]	67.57
	KRD [6]	61.53
	DCD (Ours)	**48.24**

since 51.92 already is a strong performance. Nevertheless, our increase is very evident. Lastly, similar to CycleGAN, the results on paired Pix2Pix also show the superiority of our DCD over OMGD when compressing 15.3× MACs and 16.0× parameters. In this case, OMGD leads to 9.41 FID gains while our DCD has a better increase of 10.98.

Therefore, our discriminator-cooperated feature map distillation has well demonstrated its great capability to boost the performance of a light-weighted generator.

Feature Map Distillation. As introduced in Sec. 1, current feature map distillation methods such as FitNet [34], MGD [43] and KRD [6] construct pixel-to-pixel matching between the student generator and teacher generator. In contrast, our DCD is constrained to generate perceptually alike images. In Table 2, we replace our DCD with the aforementioned distillation scenarios for performance comparison. All experiments are performed on CycleGAN of 40.3× MACs reduction as shown in Table 1. We can see that FitNet and MGD incur performance degradation of 0.71 and 2.44 FID. Though KRD allows cross-layer connections and rises the FID by 4.40, the performance increase is very limited if compared with our 13.29 performance gains. To find out the root cause, these methods were initially invented to perform image classification that focuses more on extracting robust image feature vectors. However, GANs often pay great attention to the image contents, which do not call for identical pixel values between two images but urge more for perceptual discrepancies. Therefore, off-the-shelf per-pixel matching studies fail to fit well when directly extended to GAN compression.

4.3. Visualization

We further present visualization of generated images from original CycleGAN generator as well as its compressed versions in Fig. 3. Evidently, our DCD results in not only more vivid objects (zebras), but also well retains background information. Regarding the objects, our DCD produces sharper and brighter stripes compared to other methods. In particular, existing methods are more or less influenced by the input. For example, the brown fur on horse is retained on the generated zebras (third row) while DCD well overcomes this drawback. As for the background, we find that DCD sometimes presents better visual perception even than the inputs, such as greener meadows and trees (last row). The better visual results are in accordance with better FID in Table 1. More examples of summer2winter and edges2shoes can be referred to the appendix.

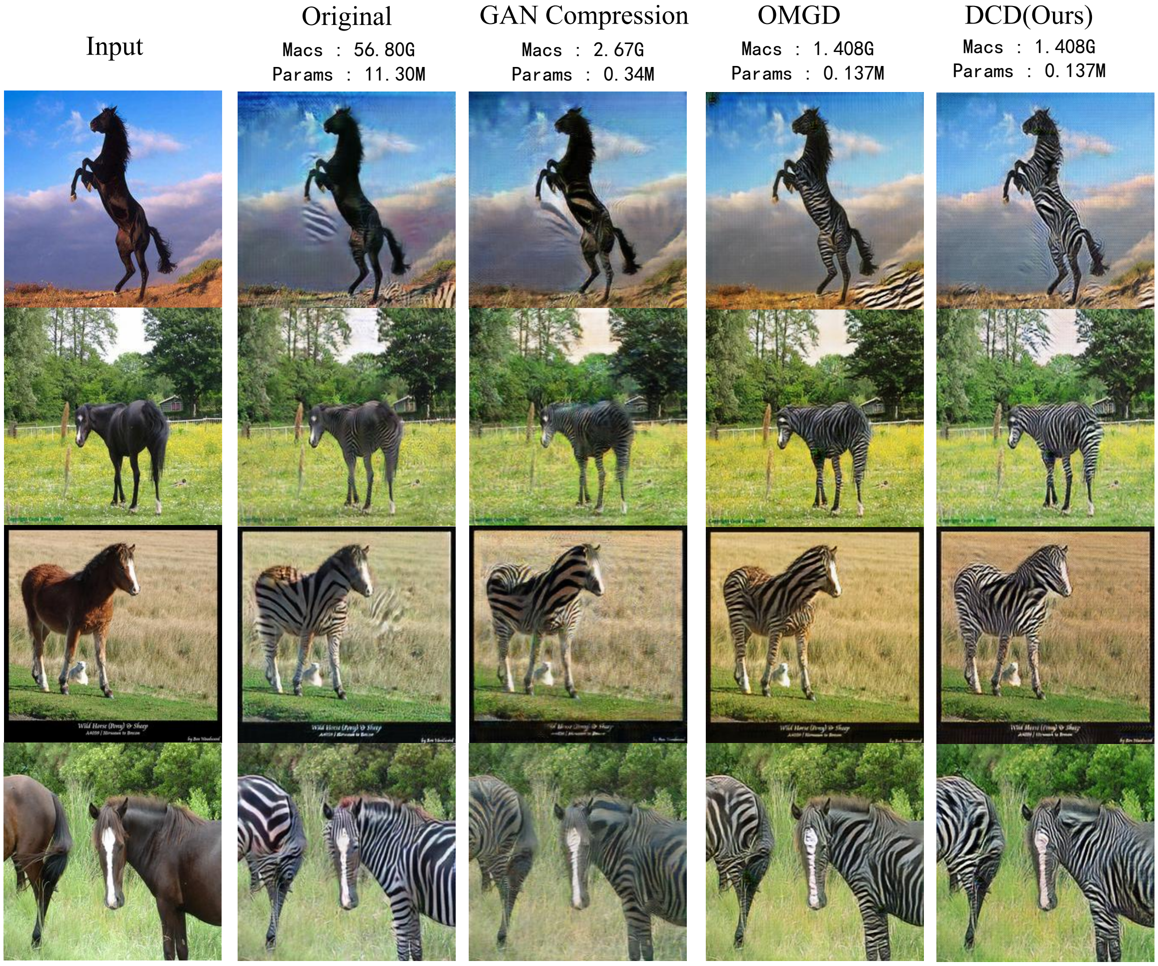

Figure 3. Visualization comparison on horse2zebra with CycleGAN.

Table 3. Training loss influence to CycleGAN on horse2zebra.

$\mathcal{L}_{per}$	$\mathcal{L}_{dcd}$	$\mathcal{L}_{gan}$	**FID(↓)**
✓			67.20
	✓		324.27
		✓	389.77
✓	✓		55.24
✓		✓	65.13
	✓	✓	323.63
✓	✓	✓	**48.24**

4.4. Analysis

We continue to conduct ablation studies to analyze influences of different training losses defined in Eq. (8), as well as our downsampling strategies defined in Eq. (6), trying to reveal why our DCD performs well. All experiments are constructed by using CycleGAN on horse2zebra.

Training Loss. Table 3 manifests the performance of different loss combinations. We observe the significance of the perceptual loss, without which the FID drastically increases to hundreds. Also, both our DCD loss and adversarial training loss are complementary to perceptual loss where $\mathcal{L}_{per}$+$\mathcal{L}_{dcd}$ increases performance to 55.24 and it is 65.13 for $\mathcal{L}_{per}$+$\mathcal{L}_{gan}$. Combining all the training losses results in the optimal FID of 48.24.

Downsampling. In Eq. (6), we adopt 1×1 convolution operations to downsample feature maps to a three-channel RGB-encoded image before being fed to the teacher discriminator. We fix downsampling modules for teacher generator while updating those for student generator, which is very crucial to the performance of the student generator. We visualize RGB features from layer 3, 6, 9, and 12 of the teacher and student generators at different training stages, referred to as stage I$\sim$IV in order. Results are displayed in Fig. 4 where two variants are introduced for comparison including updating both the teacher and student downsampling modules, and updating the teacher downsampling modules while fixing those of student. We can observe that the student generator keeps pace with teacher generator, so that they generate similar feature maps. This indicates the teacher knowledge has been well transferred to learn

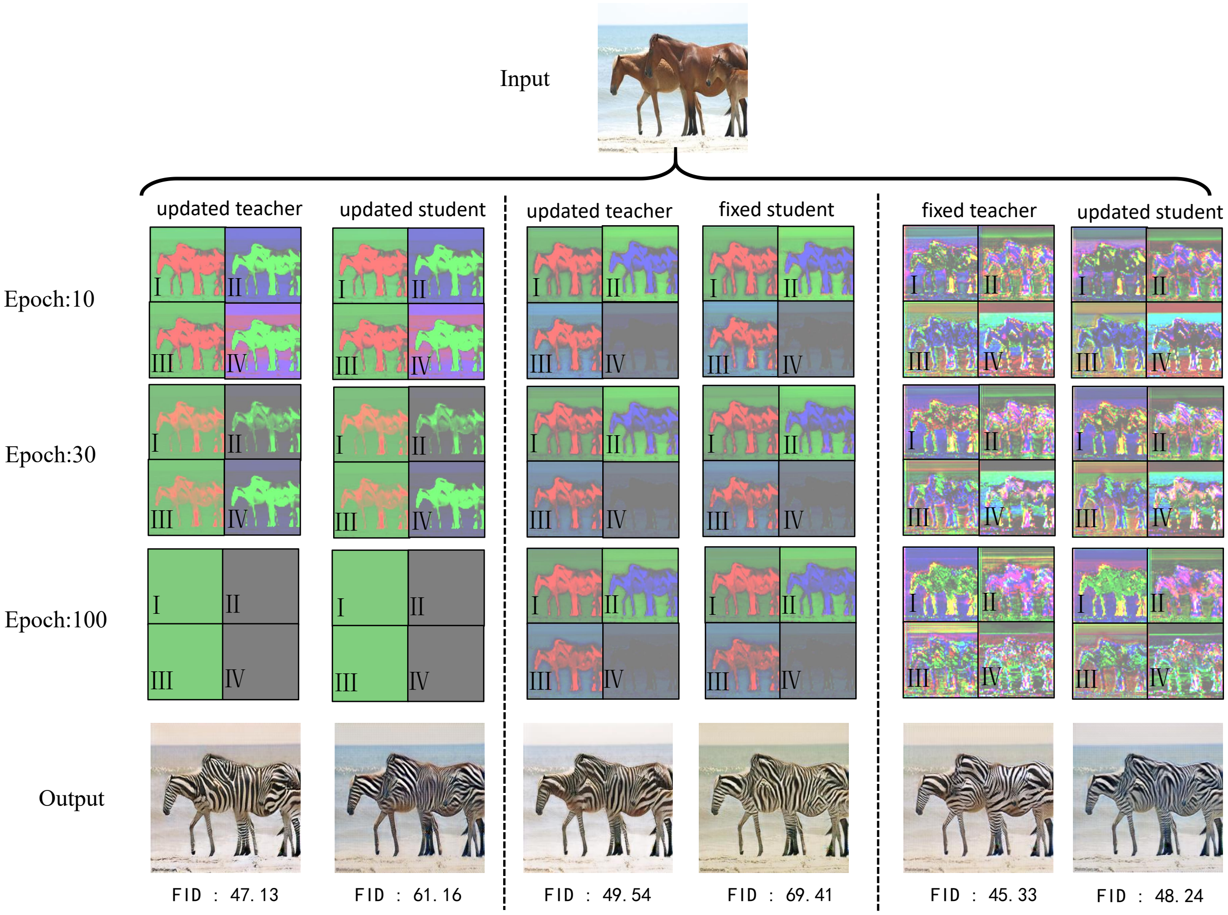

Figure 4. Feature map visualization of teacher and student generators at different training stages.

student model. Unfortunately, updating the teacher downsampling modules severely damages the knowledge from the teacher generator where only 61.16 FID is obtained if the student downsampling modules are updated as well and 69.41 otherwise. The poor performance mainly stems from the broken feature maps of teacher, where horse objects become invisible. Therefore, the student generator collapses as well. As for our downsampling strategy where downsampling modules are fixed for teacher generator while updated for student generator, the attention is gradually paid to the target horse objects as network training. Therefore, our discriminator-cooperated feature map distillation benefits more to localizing target objects. This is complementary to the perceptual loss that is engaged in style transferring. Their combination leads the lightweight student generator to finally synthesise vivid images with intricate textures.

5. Conclusion

In this paper, we proposed a novel discriminator-cooperated distillation (DCD) that reshapes the conventional pixel-to-pixel feature map match by skillfully utilizing the teacher discriminator as a transformation to pursue better visual perception in generated images. Our methods show that the teacher discriminator can also be utilized to co-train with the compressed student generator and accordingly invent a collaborative adversarial training paradigm. Our experimental results demonstrated the significant improvement of DCD in both quantitative and qualitative performance, meanwhile, the complexity of student generator is reduced by a large volume.

Acknowledgement

This work was supported by National Key R&D Program of China (No.2022ZD0118201) , the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No.U21B2037, No.U22B2051, No.62176222, No.62176223, No.62176226, No.62072386, No.62072387, No.62072389, No.62002305 and No.62272401), and the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001), Guangdong Basic and Applied Basic Research Foundation (No.2019B1515120049).

References

[1] Angeline Aguinaldo, Ping-Yeh Chiang, Alex Gain, Ameya Patil, Kolten Pearson, and Soheil Feizi. Compressing gans using knowledge distillation. *arXiv preprint arXiv:1902.00159*, 2019. 1

[2] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2017. 4

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1, 2

[4] Hanting Chen, Yunhe Wang, Han Shu, Changyuan Wen, Chunjing Xu, Boxin Shi, Chao Xu, and Chang Xu. Distilling portable generative adversarial networks for image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3585–3592, 2020. 1, 5

[5] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16296–16305, 2021. 2

[6] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5008–5017, 2021. 1, 2, 3, 6

[7] Xuxi Chen, Zhenyu Zhang, Yongduo Sui, and Tianlong Chen. Gans can play lottery tickets too. *arXiv preprint arXiv:2106.00134*, 2021. 1, 3, 5

[8] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. Cartoongan: Generative adversarial networks for photo cartoonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9465–9474, 2018. 1, 2

[9] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797, 2018. 1, 2

[10] Yonggan Fu, Wuyang Chen, Haotao Wang, Haoran Li, Yingyan Lin, and Zhangyang Wang. Autogan-distiller: Searching to compress generative adversarial networks. *arXiv preprint arXiv:2006.08198*, 2020. 1, 3, 6

[11] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 1

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. 1

[13] 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 (NeurIPS)*, 2014. 2, 3

[14] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1921–1930, 2019. 3

[15] 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 I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc., 2017. 5

[16] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[17] Liang Hou, Zehuan Yuan, Lei Huang, Huawei Shen, Xueqi Cheng, and Changhu Wang. Slimmable generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 7746–7753, 2021. 1

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017. 1, 2, 5, 6

[19] Qing Jin, Jian Ren, Oliver J Woodford, Jiazhuo Wang, Geng Yuan, Yanzhi Wang, and Sergey Tulyakov. Teachers do more than teach: Compressing image-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13600–13611, 2021. 1, 3, 5

[20] Justin Johnson, Alahi Alexandre, and Fei-Fei Li. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711, 2016. 2, 3

[21] Chanyong Jung, Gihyun Kwon, and Jong Chul Ye. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18260–18269, 2022. 3, 5

[22] 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 (CVPR)*, pages 4401–4410, 2019. 1, 2

[23] 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 (CVPR)*, pages 8110–8119, 2020. 1, 2

[24] Bo-Kyeong Kim, Shinkook Choi, and Hancheol Park. Cut inner layers: A structured pruning strategy for efficient u-net gans. *arXiv preprint arXiv:2206.14658*, 2022. 3

[25] Chenxin Li, Mingbao Lin, Zhiyuan Ding, Nie Lin, Yihong Zhuang, Yue Huang, Xinghao Ding, and Liujuan Cao. Knowledge condensation distillation. *arXiv preprint arXiv:2207.05409*, 2022. 2

[26] Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, and Song Han. Gan compression: Efficient architectures for interactive conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5284–5294, 2020. 1, 3, 5, 6

[27] Shaojie Li, Mingbao Lin, Yan Wang, Chao Fei, Ling Shao, and Rongrong Ji. Learning efficient gans for image translation via differentiable masks and co-attention distillation. *IEEE Transactions on Multimedia (TMM)*, 2022. 1, 3, 5, 6

[28] Shaojie Li, Jie Wu, Xuefeng Xiao, Fei Chao, Xudong Mao, and Rongrong Ji. Revisiting discriminator in gan compression: A generator-discriminator cooperative compression scheme. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 28560–28572, 2021. 1, 2, 3, 5, 6

[29] Zeqi Li, Ruowei Jiang, and Parham Aarabi. Semantic relation preserving knowledge distillation for image-to-image translation. In *European Conference on Computer Vision (ECCV)*, pages 648–663, 2020. 1

[30] Ji Lin, Richard Zhang, Frieder Ganz, Song Han, and Jun-Yan Zhu. Anycost gans for interactive image synthesis and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14986–14996, 2021. 3, 5

[31] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3967–3976, 2019. 3

[32] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1, 2

[33] Yuxi Ren, Jie Wu, Xuefeng Xiao, and Jianchao Yang. Online multi-granularity distillation for gan compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6793–6803, 2021. 1, 2, 3, 4, 5, 6

[34] 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. 1, 2, 3, 6

[35] Han Shu, Yunhe Wang, Xu Jia, Kai Han, Hanting Chen, Chunjing Xu, Qi Tian, and Chang Xu. Co-evolutionary compression for unpaired image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3235–3244, 2019. 3, 5, 6

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 4

[37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 5

[38] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1365–1374, 2019. 3

[39] Haotao Wang, Shupeng Gui, Haichuan Yang, Ji Liu, and Zhangyang Wang. Gan slimming: All-in-one gan compression by a unified optimization framework. In *European Conference on Computer Vision (ECCV)*, pages 54–73, 2020. 1, 3, 5

[40] Peiqi Wang, Dongsheng Wang, Yu Ji, Xinfeng Xie, Haoxuan Song, XuXin Liu, Yongqiang Lyu, and Yuan Xie. Qgan: Quantized generative adversarial networks. *arXiv preprint arXiv:1901.08263*, 2019. 1, 3

[41] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Kdgan: Knowledge distillation with generative adversarial networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1

[42] Wenju Xu, Chengjiang Long, Ruisheng Wang, and Guanghui Wang. Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6383–6392, 2021. 1

[43] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. *arXiv preprint arXiv:2205.01529*, 2022. 2, 3, 6

[44] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 5

[45] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations (ICLR)*, 2016. 2, 3

[46] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 7354–7363, 2019. 1, 2

[47] Linfeng Zhang, Xin Chen, Xiaobing Tu, Pengfei Wan, Ning Xu, and Kaisheng Ma. Wavelet knowledge distillation: Towards efficient image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12464–12474, 2022. 3, 5, 6

[48] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11953–11962, 2022. 3

[49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 1, 2, 5, 6