This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Wavelet Knowledge Distillation: Towards Efficient Image-to-Image Translation

Linfeng Zhang[1] Xin Chen[2] Xiaobing Tu[3] Pengfei Wan[3] Ning Xu[3] Kaisheng Ma[1]*
Tsinghua Univeristy[1] Intel Corporation[2] Kuaishou Technology[3†]
zhang-lf19@mails.tsinghua.edu.cn, xin.chen@intel.com, {tuxiaobing, wanpengfei}@kuaishou.com
ningxu01@gmail.com, kaisheng@mail.tsinghua.edu.cn

Abstract

Remarkable achievements have been attained with Generative Adversarial Networks (GANs) in image-to-image translation. However, due to a tremendous amount of parameters, state-of-the-art GANs usually suffer from low efficiency and bulky memory usage. To tackle this challenge, firstly, this paper investigates GANs performance from a frequency perspective. The results show that GANs, especially small GANs lack the ability to generate high-quality high frequency information. To address this problem, we propose a novel knowledge distillation method referred to as wavelet knowledge distillation. Instead of directly distilling the generated images of teachers, wavelet knowledge distillation first decomposes the images into different frequency bands with discrete wavelet transformation and then only distills the high frequency bands. As a result, the student GAN can pay more attention to its learning on high frequency bands. Experiments demonstrate that our method leads to 7.08× compression and 6.80× acceleration on CycleGAN with almost no performance drop. Additionally, we have studied the relation between discriminators and generators which shows that the compression of discriminators can promote the performance of compressed generators.

1. Introduction

Tremendous progress has been achieved with Generative adversarial networks (GANs) in generating high-fidelity, high-resolution, and photo-realistic images and videos with both paired and unpaired datasets [4,13,17,18,25,41,43,59]. The excellent performance of GANs has promoted its application in various image-to-image translation tasks, such as image style transfer [20, 21] and super-resolution [22]. Compared with other tasks such as image classification and object detection, image-to-image generation is more complex since it has a much larger output space. As a consequence, existing GANs always have high computational demands and a huge amount of parameters, which lead to inefficient inference and intolerant memory footprint, and limit their usage in resource-constrained platforms.

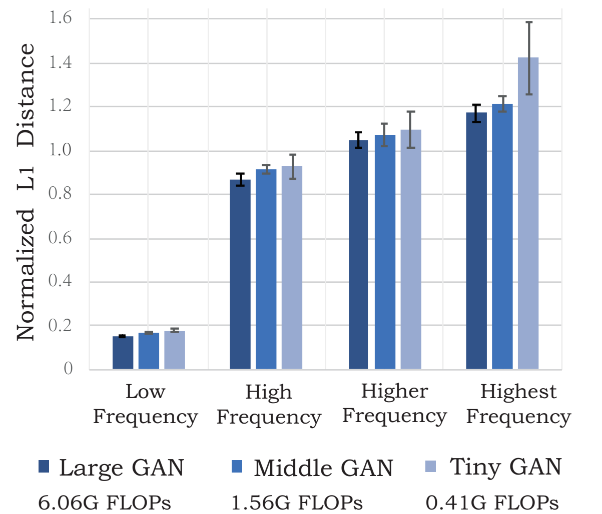


Figure 1. Normalized L_1 distance between the images generated by GANs and the ground truth images on different frequency bands. Different colors indicate GANs with different FLOPs. Results are averaged over 8 trials on Edge→Shoe dataset.

Knowledge distillation (KD) has been an effective tool for improving the performance of small models [5, 14]. By imitating the prediction results and the intermediate features from a cumbersome teacher model, the performance of a lightweight student model can be improved significantly. Following previous knowledge distillation methods on classification [44], object detection [53], semantic segmentation [31] and action recognition [28], some recent research has tried to directly apply knowledge distillation to GANs. Unfortunately, most of them obtain very limited and even negative effects [23, 26].

Why does KD not work well on GAN? In this paper, we first study this question from a frequency perspective with the following experiment. Firstly, discrete wavelet

* Corresponding author. † This project is funded by Kuaishou Research Program. This work was done during internship of Linfeng Zhang and Xin Chen with Y-tech Kuaishou Technology. Codes are relased on Github.

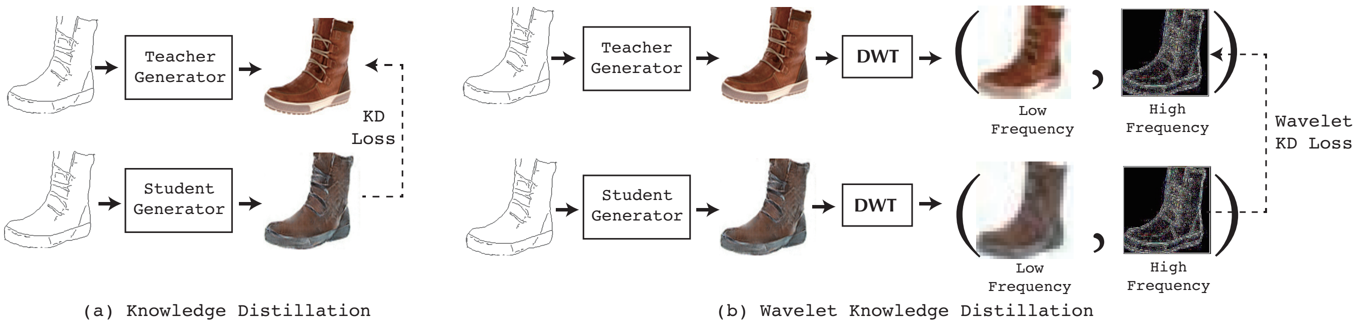


Figure 2. Comparison between knowledge distillation [14] (subfigure a) and the proposed wavelet knowledge distillation (subfigure b) on Edges→Shoes. Wavelet knowledge distillation first applies discrete wavelet transformation (DWT) to the generated images and then only minimizes the difference on high frequency bands.

transformation (DWT) is utilized to decompose the generated images and the ground truth images into different frequency bands. Then, we compute the normalized L_1-norm distance on each frequency band respectively[1]. As shown in Figure 1, all the GANs achieve very low error on the low frequency band but fail in the generation on high frequency bands, which is in line with the observation that images generated by GANs do not have good details. Besides, it is observed that compared with the large GAN, the tiny GAN achieves comparable performance on the low frequency band but much worse performance on high frequency bands. These two observations demonstrate that more attention should be paid to the high frequency during GAN compression.

However, naive knowledge distillation in GANs application directly minimizes the difference between the images generated by students and teachers and ignores the priority of high frequency. Motivated by these observations, we propose wavelet knowledge distillation, which highlights students learning on the high frequency in knowledge distillation. As shown in Figure 2, we first apply discrete wavelet transformation to decompose the images generated by teachers and students into different frequency bands and then only minimize the L_1 loss on the high frequency bands.

Abundant experiments on both paired and unpaired image-to-image translation demonstrate the effectiveness of our method both quantitatively and qualitatively. On Horse→Zebra and Zebra→Horse datasets, our method leads to 7.08× compression and 6.80× acceleration on CycleGAN with almost no performance drop. In the discussion section, we have further studied the effectiveness of different frequency bands and the influence of knowledge distillation schemes. Additionally, studies on the relation between discriminators and generators in model compression have also been introduced, showing that the compression of discriminators can significantly promote the performance of compressed generators.

[1] Details of this experiment can be found in the supplementary material.

Our main contributions can be summarized as follows.

- We have analyzed the performance of GANs from a frequency perspective, which quantitatively shows that GAN, especially small GAN lacks the ability to generate high-quality high frequency information in images.
- Based on the above observation, wavelet knowledge distillation is proposed to address this issue by only distilling the high frequency information, instead of all the information from images generated by the teacher.
- Quantitative and qualitative results on three models and eight datasets with six comparison methods have demonstrated the effectiveness of our method.
- We have studied the relation between discriminators and generators during model compression. It shows that compression on discriminators is necessary for maintaining its competition with compressed generators in adversarial learning. which further benefits the performance of generators.

2. Related Work

2.1. Image-to-Image Translation

Generative adversarial networks have shown powerful ability in formulating and generating high-fidelity, high-resolution, and photo-realistic images and videos, and thus become the dominant models in image-to-image translation [2, 7, 10, 13, 30, 46, 47]. Pix2Pix is first proposed to apply conditional generative adversarial networks for paired image-to-image translation [17]. Then, Pix2PixHD is proposed to generate images with higher resolution with coarse to fine generators and multi-scale discriminators [48]. A more challenging task is to perform image-to-image translation with unpaired datasets. CycleGAN tackles this challenge by introducing the cycle-consistency loss, which reconstructs the generated image to the input domain [59]. Then, attention CycleGAN is proposed to find the crucial pixels in the images with an attention module [9]. Spade module is introduced in GauGAN to avoid the loss of semantic information in batch normalization layers [39]. Recently, researchers find that the perfect reconstruction in CycleGAN may be too difficult to achieve [36]. To address this issue, Park *et al.* introduce the patch-wise contrastive learning which improves the generation quality, stabilizes the training process, and reduces the training time, simulanteously [38]. The high-resolution and photo-realistic generated images come at the expense of intensive computation and massive parameters. To tackle this challenge, fruitful compression methods such as network pruning and network architecture search have been proposed recently. Li *et al.* propose the GAN Compression, which applies once-for-all search to find the best tiny GAN architecture [23]. Jin *et al.* introduce an inception-based residual block into generators and further compress them with channel pruning [19]. Liu *et al.* propose the Content-Aware GAN Compression, which enables GANs to maintain the content of crucial regions during compression [32]. Li *et al.* propose to revisit the role of discriminator in GAN compression with a selective activation discriminator [24].

2.2. Knowledge Distillation

Knowledge distillation, which aims to facilitate the training of a lightweight student model under the supervision of a cumbersome teacher model, has been considered as an effective approach for both model compression and model accuracy boosting. The idea of employing a teacher model to train a student model is first proposed by Buciluǎ *et al.* for ensemble model compression [5]. Then Hinton *et al.* propose the concept of knowledge distillation, which introduces a hyper-parameter named temperature in softmax to soften the distribution of teacher logits [14]. Recently, abundant methods have been proposed to distill the knowledge in the intermediate features [52, 54] and their relation [40, 45]. Besides image classification, recent works have also successfully applied knowledge distillation to more challenging tasks such as object detection [3, 53], semantic segmentation [31], pre-trained language models [42, 50], machine translation [27], distributed training [37], multi-exit models [56] and so on.

However, the effect of knowledge distillation on GANs for image-to-image translation has not been well-studied. Existing research shows that directly minimizing the distance between the generated images of students and teachers does not improve but sometimes harms the performance of students [26]. A few previous methods have tried to apply the classification-based knowledge distillation to image-to-image translation but earned very limited improvements. For instance, Li *et al.* propose to minimize the distance between intermediate features of teacher and student GANs [23] and Li *et al.* have tried to distill the semantic relation between the features of different patches [26]. Recently, Chen *et al*. propose an overall knowledge distillation framework on GANs by distilling both the generator and the discriminator [6]. Jin *et al*. propose to distill generators with global kernel alignment on intermediate features [19], which boosts student performance without introducing additional layers. The main difference between our method and the previous GAN knowledge distillation methods is that our method distills the generated images instead of the intermediate features. As a result, our method is orthogonal to the previous methods, and it can be utilized with previous methods to achieve better performance.

2.3. Wavelet Analysis in Deep Learning

Compared with the other frequency analysis methods such as Fourier analysis, wavelet transformation can capture both the spatial and the frequency information in the sign and is thus considered as a more effective method in image processing [33]. Along with the success of deep learning, fruitful methods have been proposed to apply wavelet methods into neural networks for different targets. Williams *et al*. propose the wavelet pooling which replaces the max and average pooling with discrete wavelet transformation to preserve the global information of images during down sampling [49]. Chen *et al.* propose wavelet-like auto-encoder, which compresses the original image into two low-resolution images to accelerate the inference computation [8]. Liu *et al.* introduce wavelet transformation to the convolutional neural networks to leverage the spectral information in texture classification [12].

Recent works have also applied wavelet analysis to the image-to-image translation tasks. Huang *et al.* first propose the wavelet-SRNet, which performs single image super-resolution by predicting the wavelet coefficients of high-resolution images [15]. Inspired by the architecture of U-Net, Liu *et al.* apply the wavelet package in convolutional neural networks to obtain a large receptive field efficiently [29]. To the best of our knowledge, this paper is the first work which applies wavelet analysis to knowledge distillation and GANs compression.

3. Methodology

3.1. Wavelet Analysis

Given a function ψ, let $\mathcal{X}(\psi)$ be the collection of the dilations and shift of ψ:

$$\mathcal{X}(\psi) = \{\psi_{jk} = 2^{-j/2}\psi(2^{-j}x - k)|\, j, k \in Z\}, \qquad (1)$$

where ψ is the orthogonal wavelet if $\mathcal{X}(\psi)$ forms a basis in $\mathcal{L}_2$ spaces. Discrete wavelet transformation (DWT) is a mathematical tool for pyramidal image decomposition. With DWT, each image can be decomposed into

Table 1. Experiment results on paired image-to-image translation on Edges→Shoes with Pix2Pix and Pix2PixHD. A lower FID is better performance. Δ indicates the performance improvements compared with the origin student. Each result is averaged over 8 trials.

Pix2PixHD

#Params (M)	FLOPs (G)	Method	Metric FID↓	Metric Δ↑
45.59	48.36	Teacher	41.59±0.42	–
		Origin Student	44.64$_{\pm 0.54}$	–
		Hinton *et al.* [14]	45.31$_{\pm 0.63}$	-0.67
1.61	1.89	Zagoruyko *et al.* [52]	44.21$_{\pm 0.72}$	0.43
28.32×	25.59×	Li and Lin *et al.* [23]	44.03$_{\pm 0.41}$	0.61
		Li and Jiang *et al.* [26]	43.90$_{\pm 0.36}$	0.74
		Jin *et al.* [19]	43.97$_{\pm 0.17}$	0.67
		Ahn *et al.* [1]	44.53$_{\pm 0.48}$	0.11
		Ours	**42.53**$_{\pm 0.29}$	**2.11**

Pix2Pix

#Params (M)	FLOPs (G)	Method	Metric FID↓	Metric Δ↑
54.41	6.06	Teacher	59.70$_{\pm 0.91}$	–
		Origin Student	85.06$_{\pm 0.98}$	–
		Hinton *et al.* [14]	86.97$_{\pm 3.49}$	-1.91
13.61	1.56	Zagoruyko *et al.* [52]	84.25$_{\pm 2.08}$	0.81
4.00×	3.88×	Li and Lin *et al.* [23]	83.63$_{\pm 3.12}$	1.43
		Li and Jiang *et al.* [26]	84.01$_{\pm 2.31}$	1.05
		Jin *et al.* [19]	84.39$_{\pm 3.62}$	0.67
		Ahn *et al.* [1]	84.92$_{\pm 0.78}$	0.14
		Ours	**80.13**$_{\pm 2.18}$	**4.93**

Table 2. Experiment results on unpaired image-to-image translation on Horse→Zebra and Zebra→Horse with CycleGAN. A lower FID is better. Δ indicates the performance improvements compared with the origin student. Each result is averaged over 8 trials.

Horse→Zebra

#Params (M)	FLOPs (G)	Method	Metric FID↓	Metric Δ↑
11.38	49.64	Teacher	61.34$_{\pm 4.35}$	–
		Origin Student	85.04$_{\pm 6.88}$	–
		Hinton *et al.* [14]	84.08$_{\pm 3.78}$	0.96
0.72	3.35	Zagoruyko *et al.* [52]	81.24$_{\pm 2.01}$	3.80
15.81×	14.82×	Li and Lin *et al.* [23]	83.97$_{\pm 5.01}$	1.07
		Li and Jiang *et al.* [26]	81.74$_{\pm 4.65}$	3.30
		Jin *et al.* [19]	82.37$_{\pm 8.56}$	2.67
		Ahn *et al.* [1]	82.91$_{\pm 2.41}$	2.13
		Ours	**77.04**$_{\pm 3.52}$	**8.00**
		Ours + Li and Lin *et al.*	**76.40**$_{\pm 3.17}$	**8.64**
		Origin Student	70.54$_{\pm 9.63}$	–
		Hinton *et al.* [14]	70.35$_{\pm 3.27}$	0.19
1.61	7.29	Zagoruyko *et al.* [52]	67.51$_{\pm 4.57}$	3.03
7.08×	6.80×	Li and Lin *et al.* [23]	68.58$_{\pm 4.31}$	1.96
		Li and Jiang *et al.* [26]	68.94$_{\pm 2.98}$	1.60
		Jin *et al.* [19]	67.31$_{\pm 3.01}$	3.23
		Ahn *et al.* [1]	69.32$_{\pm 5.89}$	1.22
		Ours	**61.65**$_{\pm 4.73}$	**8.89**
		Ours + Li and Lin *et al.*	**60.13**$_{\pm 4.08}$	**10.41**

Zebra→Horse

#Params (M)	FLOPs (G)	Method	Metric FID↓	Metric Δ↑
11.38	49.64	Teacher	138.07$_{\pm 4.01}$	–
		Origin Student	152.67$_{\pm 9.63}$	–
		Hinton *et al.* [14]	148.64$_{\pm 1.62}$	4.03
0.72	3.35	Zagoruyko *et al.* [52]	148.92$_{\pm 1.20}$	3.75
15.81×	14.82×	Li and Lin *et al.* [23]	151.32$_{\pm 2.31}$	1.35
		Li and Jiang *et al.* [26]	151.09$_{\pm 3.67}$	1.58
		Jin *et al.* [19]	149.73$_{\pm 3.94}$	2.94
		Ahn *et al.* [1]	150.31$_{\pm 3.55}$	2.36
		Ours	**146.01**$_{\pm 1.86}$	**6.66**
		Ours + Li and Lin *et al.*	**145.96**$_{\pm 1.92}$	**6.71**
		Origin Student	141.86$_{\pm 1.57}$	–
		Hinton *et al.* [14]	142.03$_{\pm 1.61}$	-0.17
1.61	7.29	Zagoruyko *et al.* [52]	141.23$_{\pm 1.88}$	0.63
7.08×	6.80×	Li and Lin *et al.* [23]	141.32$_{\pm 1.27}$	0.54
		Li and Jiang *et al.* [26]	151.09$_{\pm 3.67}$	1.58
		Jin *et al.* [19]	140.98$_{\pm 1.41}$	0.88
		Ahn *et al.* [1]	141.50$_{\pm 2.51}$	0.36
		Ours	**138.84**$_{\pm 1.47}$	**3.02**
		Ours + Li and Lin *et al.*	**138.52**$_{\pm 0.95}$	**3.34**

four bands, including LL, LH, HL and HH, where LL indicates the low frequency band and the others are high frequency bands. The LL band can be further decomposed by DWT into LL2, LH2, HL2 HH2 and so on. Denote DWT as $\Psi(\cdot)$, then the high frequency and the low frequency bands of an image x can be written as $\Psi^{\mathrm{H}}(x)$ and $\Psi^{\mathrm{L}}(x)$, respectively. More specifically, in this paper, we apply 3-level discrete wavelet transformation in all the experiments. $\Psi^{\mathrm{L}}(x)$ indicates LL3 band. $\Psi^{\mathrm{H}}(x) = \{\mathrm{HL3, LH3, HH3, HL2, LH2, HH2, HL1, LH1, HH1}\}$.

3.2. Knowledge Distillation

Revisit Knowledge Distillation for Classification At the beginning of this subsection, we revisit the formulation of knowledge distillation on classification [14]. Given a set of training samples $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ and their labels $\mathcal{Y} = \{y_1, y_2, ..., y_n\}$, denoting the networks of the student and the teacher as f_s and f_t, the loss function of the student can be formulated as $\mathcal{L}_{\text{Student}} = \alpha \cdot \mathcal{L}_{\text{CE}} + (1-\alpha) \cdot \mathcal{L}_{\text{KD}}$, where $\mathcal{L}_{\text{CE}}$ indicates the cross-entropy loss between the prediction $f(x)$ and its label y. $\alpha \in (0, 1]$ is a hyper-parameter to balance two loss items, and $\mathcal{L}_{\text{KD}}$ indicates the knowledge distillation loss.

On classification tasks, $\mathcal{L}_{\text{KD}}$ can be formulated as

$$\mathcal{L}_{\text{KD}} = \frac{1}{n}\sum_{i}^{n} \mathcal{KL}\left(\text{softmax}\left(\frac{f_t(x_i)}{\tau}\right), \text{softmax}\left(\frac{f_s(x_i)}{\tau}\right)\right), \quad (2)$$

where $\mathcal{KL}$ indicates the Kullback-Leibler divergence,

Figure 3. Qualitative results on Horse→Zebra with CycleGAN (a-d) and Edges→Shoes with Pix2Pix (e-h). Numbers in the brackets indicate the acceleration ratio compared with their teachers. "Baseline" indicates the students trained without knowledge distillation.

which measures the distance between the categorical probability distribution of students and teachers. τ is the temperature hyper-parameter in softmax function.

Knowledge Distillation for Image-to-Image Translation On the task of image-to-image translation, since the prediction result $f(x_i)$ is the value of pixels instead of categorical probability distribution, KL divergence can not be utilized to measure the difference between students and teachers. A naive alternative is to replace KL divergence with the L_1-norm distance between the generated images from students and teachers. Then, we can extend Hinton knowledge distillation for image-to-image translation, whose loss function can be formulated as

$$\mathcal{L}_{\text{KD}} = \frac{1}{n}\sum_{i}^{n} \|(f_t(x_i) - f_s(x_i)\|_1. \quad (3)$$

Besides Hinton knowledge distillation, there are also abundant feature knowledge distillation methods which can be applied to image-to-image translation directly. Since our method is not feature-based, we do not introduce them here.

Wavelet Knowledge Distillation Based on the above notations, now we can introduce the proposed wavelet knowledge distillation, which only minimizes the difference on the high frequency between students and teachers. Its loss function $\mathcal{L}_{\text{WKD}}$ can be formulated as

$$\mathcal{L}_{\text{WKD}} = \frac{1}{n}\sum_{i}^{n} \|(\Psi^H \circ f_t)(x_i) - (\Psi^H \circ f_s)(x_i)\|_1. \quad (4)$$

On unpaired image-to-image translation models such as CycleGAN, there are sometimes two generators for the two translation directions. In this circumstance, the proposed wavelet knowledge distillation loss can be applied to the two directions simultaneously.

The overall training loss can be formulated as $\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{origin}} + \alpha \cdot \mathcal{L}_{\text{WKD}}$, where $\mathcal{L}_{\text{origin}}$ indicates the original training loss of different models, such as the adversarial learning loss and recycling loss. α is the hyper-parameter to balance the two loss functions. Hyper-parameter sensitivity studies have been given in the supplementary material.

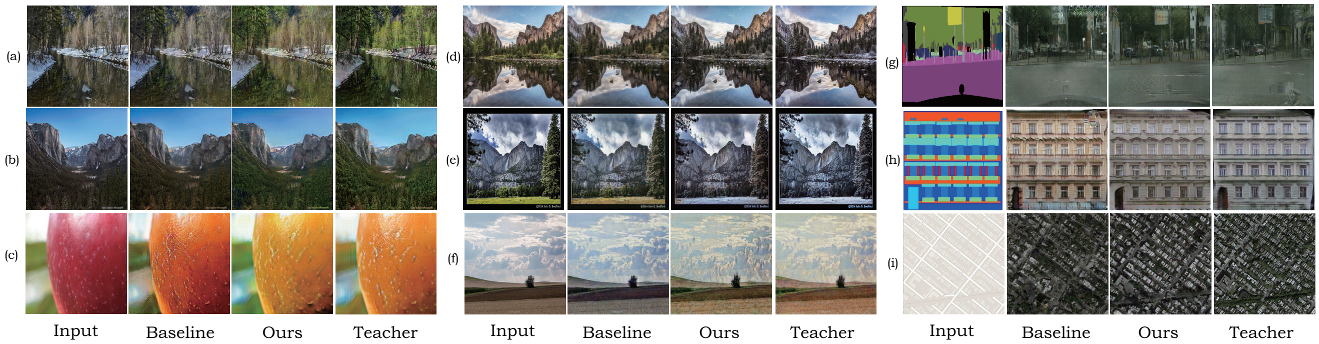


Figure 4. Qualitative experiments on the other datasets: Winter→Summer (subfig. a-b), Summer→Winter (subfig. d-e), Apple→Orange (subfig. c), Photo→Monet (subfig. f), Cityscapes (subfig. g), Facades (subfig. h) and Maps (subfig. i).

Table 3. Paired image-to-image translation experiment results on Cityscapes with Pix2Pix. A higher mIoU is better. Δ indicates the performance improvements compared with the origin student. Each experiment is averaged over 8 trials.

#Params (M)	FLOPs (G)	Method	Metric	
			mIoU↑	Δ ↑
54.41	96.97	Teacher	46.51±0.32	–
13.61 $_{4.00\times}$	24.90 $_{3.88\times}$	Origin Student	41.35$_{\pm 0.22}$	–
		Hinton *et al.* [14]	40.49$_{\pm 0.41}$	-0.86
		Zagoruyko *et al.* [52]	40.17$_{\pm 0.36}$	-1.18
		Li and Lin *et al.* [23]	41.52$_{\pm 0.34}$	0.17
		Li and Jiang *et al.* [26]	41.77$_{\pm 0.30}$	0.42
		Jin *et al.* [19]	41.29$_{\pm 0.51}$	-0.06
		Ahn *et al.* [1]	41.88$_{\pm 0.45}$	0.53
		Ours	**42.93**$_{\pm 0.25}$	**1.58**

4. Experiment

4.1. Experiment Settings

Models and Datasets Experiments are mainly conducted with three models including Pix2Pix [17], CycleGAN [59] and Pix2PixHD [48]. Our teacher is the original model with the setting from their released codes (*ngf=64*)[2]. The student model has the same architecture and depth but fewer channels (*ngf=32/24/16*) compared with its teacher. Quantitative experiments have been conducted on Horse →Zebra, Edge→Shoe and Cityscapes [11]. Besides, we also conduct qualitative experiments on Winter→Summer, Summer→Winter, Apple→Orange, Photo→Monet, Facades and Maps [16, 58].

Evaluation Settings Following previous works [19], we adopt *Fréchet Inception Distance (FID)* and mIoU as the performance metrics on Cityscapes and the other datasets. A lower FID and a higher mIoU indicate that the generated images have better quality. Please refer to the codes in the supplementary material for more details.

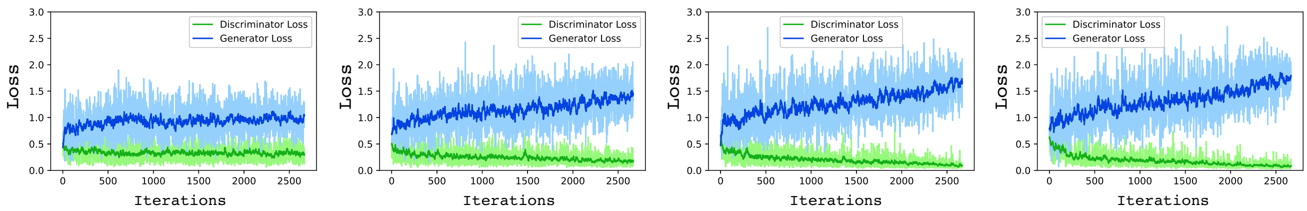
[2]"ngf" indicates "number of generator filters"

4.2. Quantitative Results

The quantitative experiment results on paired and unpaired image-to-image translation have been shown in Table 1 and Table 2, respectively. Experiments on Cityscapes have been shown in Table 3. It is observed that: **(a)** Directly applying the native knowledge distillation (Equation 3) to GANs sometimes leads to performance degradation. For instance, there are 1.91 and 0.67 FID increments (performance degradation) on Edges→Shoes with Pix2pix and Pix2PixHD, respectively. **(b)** In contrast, our method achieves consistent and significant performance improvements on all the datasets and models, which outperforms the other GAN knowledge distillation methods by a clear margin, *e.g.* 3.78 FID lower than the second-best method on CycleGAN, on average. **(c)** On Horse→Zebra and Zebra→Horse, the student models trained with our method achieve almost the same FID with the teacher model, which indicates 7.08× compression and 6.80× acceleration with almost no performance degradation. **(d)** Compared with the students trained without knowledge distillation, the distilled students usually not only achieve lower FID, but also tend to have lower FID standard deviation, which shows that knowledge distillation may stabilize the training of GANs. **(e)** Our method and previous feature knowledge distillation method can be utilized together, which further leads to 0.63 FID reduction on CycleGAN on average.

4.3. Qualitative Results

The qualitative results of CycleGAN on Horse→Zebra (a-d) and Pix2Pix on Edges→Shoes (e-h) have been shown in Figure 3. It is observed that:**(a)** On Horse→Zebra, the baseline model can not transform the whole body of horses to zebras (*e.g.* subfigures a, b and c). Besides, the generated stripes of zebras are chaotic and unnatural (*e.g.* subfigure d). This problem also exists in the other knowledge distillation methods (*e.g.* subfigure c). In contrast, the images generated by the distilled students don't have these issues. **(b)** On Edges→Shoes, the generated images from distilled students



(a)D-Params=0.18M (b)D-Params=0.69M (c)D-Params=1.56M (d)D-Params=2.77M

Figure 5. The discriminator loss and generator loss during the training period. In all the subfigures, the generators are 15.81× compressed. In subfigure (d), the discriminator has its origin size. In subfigure (a-c), the discriminators are compressed by 15.39×, 4.01× and 1.78×. The FID of these four experiments have been shown in Figure 6.

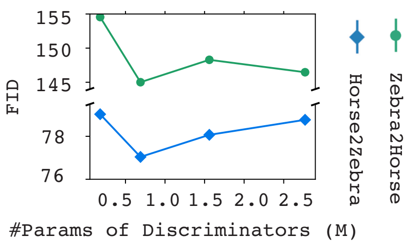


Figure 6. Experiments on the distilled CycleGAN with discriminators in different sizes from Figure 5. Lower FID is better.

have much better color and details (*e.g.* shoestrings in subfigure f and g). In subfigure (f), the distilled students have successfully generated the highlights on the shoes, which makes the images more realistic.

5. Discussion

5.1. Ablation Study

In this subsection, we give a detailed study on the individual influence of different frequency bands in knowledge distillation. Experimental results on Horse→Zebra and Zebra→Horse with CycleGAN are shown in Table 4. It is observed that: **(a)** The performance of student models has been severely harmed by only distilling the low frequency band (11.07/10.39 FID increments). **(b)** The best performance can be achieved by only distilling the high frequency bands (*i.e.* the proposed wavelet knowledge distillation). **(c)** Distilling both the high and the low frequency bands achieves slight FID reduction, but its performance is still worse than only distilling the high frequency band. These observations have clearly demonstrated the benefits of distilling the high frequency band and the negative influence of distilling the low frequency band, which also conform to the conclusion in Figure 1 – more attention should be paid to the high frequency band during GAN compression.

5.2. A Small Discriminator Makes the Compressed Generator Better

Usually, in real-world GAN applications, only the generators are required to be deployed in devices, while the discriminators are always discarded at this time. As a result, most of the previous works only perform compression on generators but ignore what should be done on discriminators. However, since the discriminator directly influences the training loss of of generators, it has a crucial impact on the performance of generators. In this subsection, we study how the capacity of discriminators influences generators. Figure 5 has shown the training loss of generators and discriminators for four CycleGANs with discriminators in different sizes. In all the subfigures, the generators are 15.81× compressed and trained with wavelet knowledge distillation. In subfigure (d), the discriminator has its origin size. In subfigure (a-c), the discriminators are compressed by 15.39×, 4.01×, and 1.78×, respectively. Besides, their corresponding FIDs have been shown in Figure 6.

Observation & Analysis It is observed that: **(i)** When the generator is compressed but the discriminator is not compressed (subfigure d), the loss of generator is much higher and the loss of discriminator is much lower. This observation indicates that when the discriminator has a much larger size than the generator, it achieves overwhelming success in its competition with generators. Thus, the balance between discriminators and generators is broken, which makes generators hard to learn useful information from the adversarial loss. **(ii)** The distilled generator achieves the best performance when the discriminator is 4.01× compressed (0.69M). Both a too small and a too large discriminator lead to performance degradation on generators, indicating that the unbalance between discriminators and generators in adversarial learning harms the training of generators.

Based on these observations, we can conclude that although discriminators are not utilized in application, they are still required to be properly compressed to maintain the balance between them and generators in adversarial learning, which further benefits the training of generators.

5.3. Knowledge Distillation Paradigm

Knowledge distillation is first proposed in a teacher-student (TSKD) paradigm, where the teacher model is first trained and then distilled to a student model. Recently, abundant knowledge distillation paradigms are proposed to achieve better performance, such as deep mutual learning

Table 4. Ablation study on different frequency bands. Each experiment is averaged over 8 trials. Lower FID is better.

Frequency		FID↓	
Low	High	Horse→Zebra	Zebra→Horse
×	×	$85.04_{\pm 6.88}$	$152.67_{\pm 5.07}$
✓	×	$96.11_{\pm 14.39}$	$163.06_{\pm 3.91}$
×	✓	$\mathbf{77.04}_{\pm \mathbf{3.52}}$	$\mathbf{146.01}_{\pm \mathbf{1.86}}$
✓	✓	$81.81_{\pm 4.52}$	$148.09_{\pm 2.18}$

Table 5. Comparison on knowledge distillation paradigms. Wavelet knowledge distillation is utilized in all these experiments. Each experiment is averaged over 8 trials. Lower FID is better.

Dataset	KD Scheme	Metric	
		FID↓	Δ
Horse→Zebra	Origin Student	$85.04_{\pm 6.88}$	–
	TSKD	$\mathbf{77.04}_{\pm \mathbf{3.52}}$	**8.00**
	TAKD[1]	$78.53_{\pm 2.98}$	6.51
	TAKD[2]	$78.69_{\pm 3.26}$	6.35
	SD	$83.51_{\pm 2.00}$	1.53
	DML[1]	$81.06_{\pm 3.56}$	3.98
	DML[2]	$84.72_{\pm 5.17}$	0.32
Zebra→Horse	Origin Student	$152.67_{\pm 5.07}$	–
	TSKD	$\mathbf{146.01}_{\pm \mathbf{1.86}}$	**6.66**
	TAKD[1]	$148.03_{\pm 1.40}$	4.56
	TAKD[2]	$147.75_{\pm 1.75}$	4.92
	SD	$151.74_{\pm 3.46}$	0.93
	DML[1]	$150.37_{\pm 2.18}$	2.30
	DML[2]	$152.03_{\pm 1.93}$	0.64

TAKD[1]: The FIDs of teacher assistants on Horse → Zebra and Zebra → Horse are 55.00 and 140.49, respectively.
TAKD[2]: The FIDs of teacher assistants on Horse → Zebra and Zebra → Horse are 51.34 and 133.29, respectively.
DML[1] and DML[2]: There are 2 and 3 peers , respectively.

(DML) [57], self-distillation (SD) [55] and teacher-assistant knowledge distillation (TAKD) [34]. Many of these methods lead to higher effectiveness than the traditional (TSKD) paradigm. Unfortunately, these KD paradigms are usually only evaluated on classification tasks and their performance in more challenging tasks has not been well-studied. In this subsection, we have given a comparison of the following KD paradigms on Image-to-Image translation with GANs.

- *TSKD* is the most common KD diagram which trains a large teacher first and then distills it to a small student.
- *TAKD* is proposed to bridge the gap between students and teachers with a teacher-assistant. It first distills knowledge from teachers to teacher-assistants and then distills knowledge from teacher assistants to the students [34].
- *SD* is a special case in TSKD when the student and the teacher have the identical architecture. Experimental and theoretical results have proven its success [35].
- *DML* (*a.k.a.* online knowledge distillation, collaborative learning) trains several students (*a.k.a.* peers) to learn from each other [57].

Observation Experimental results of different knowledge distillation paradigms have been shown in Table 5. It is observed that: **(a)** All the knowledge distillation schemes lead to performance gain compared with the baseline. Besides, the most common TSKD achieves better performance than the other KD schemes. **(b)** The performance improvements in DML and SD are much lower than that in TSKD and TAKD, which indicates a pre-trained and high-quality teacher is very crucial on image-to-image translation. **(c)** There is no significant performance difference between TSKD and TAKD, which means that the teacher assistants can not facilitate the training of a tiny student in knowledge distillation on image-to-image translation.

Analysis These observations show that there is a huge difference between knowledge distillation on image classification and image-to-image translation. We believe this difference is caused by the following reasons: **(a)** Compared with image classification, image-to-image translation is more challenging and thus a high performance teacher is more necessary to provide better guidance. **(b)** Besides, on classification, one of the benefits of the novel knowledge distillation schemes comes from their effectiveness as label smoothing [51]. However, label smoothing is effective in classification but can not be utilized in image-to-image translation, which is a pixel-level regression problem.

6. Conclusion

This paper has proposed to analyze and distill GANs on image-to-image translation tasks from a frequency perspective. To the best of our knowledge, we first quantitatively show the difference of GANs performance on different frequency bands and propose to highlight its learning on the high frequency bands during knowledge distillation. Abundant experiments on both paired and unpaired image-to-image translation have demonstrated its significant performance in terms of both quantitative and qualitative results. For instance, 7.08× compression and 6.80× acceleration can be achieved on CycleGAN with almost no performance drop. Experimental results in the ablation study have further shown the merits of distilling the high frequency bands. Besides, studies on the relation between discriminators and generators in model compression have been introduced, showing that a small discriminator is beneficial during the compression of the generator by maintaining their balance in adversarial learning. Moreover, we have also analyzed the influence of different knowledge distillation paradigms on GANs for image-to-image translation. Surprisingly, different from the results in classification, most of novel KD paradigms do not work well on GANs. We expect this observation may encourage studies of knowledge distillation in tasks beyond classification. Our limitations and future work are discussed in supplementary materials.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019. 4, 6

[2] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 862–871. Computer Vision Foundation / IEEE, 2021. 2

[3] Mohammad Farhadi Bajestani and Yezhou Yang. Tkd: Temporal knowledge distillation for active perception. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 953–962, 2020. 3

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

[5] Cristian Buciluǎ, 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. ACM, 2006. 1, 3

[6] 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*, volume 34, pages 3585–3592, 2020. 3

[7] Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Dualast: Dual style-learning networks for artistic style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 872–881. Computer Vision Foundation / IEEE, 2021. 2

[8] Tianshui Chen, Liang Lin, Wangmeng Zuo, Xiaonan Luo, and Lei Zhang. Learning a wavelet-like auto-encoder to accelerate deep neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 3

[9] Xinyuan Chen, Chang Xu, Xiaokang Yang, and Dacheng Tao. Attention-gan for object transfiguration in wild images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 164–180, 2018. 3

[10] Jiaxin Cheng, Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Prem Natarajan. Style-aware normalized loss for improving arbitrary style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 134–143. Computer Vision Foundation / IEEE, 2021. 2

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[12] Shin Fujieda, Kohei Takayama, and Toshiya Hachisuka. Wavelet convolutional neural networks for texture classification. *arXiv preprint arXiv:1707.07394*, 2017. 3

[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*, 27, 2014. 1, 2

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS*, 2014. 1, 2, 3, 4, 6

[15] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1697, 2017. 3

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Pix2pix datasets, `http://efrosgans.eecs.berkeley.edu/pix2pix/datasets/`. 6

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 2, 6

[18] Somi Jeong, Youngjung Kim, Eungbean Lee, and Kwanghoon Sohn. Memory-guided unsupervised image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 6558–6567. Computer Vision Foundation / IEEE, 2021. 1

[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*, pages 13600–13611, 2021. 3, 4, 6

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1

[21] 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. 1

[22] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1

[23] 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*, pages 5284–5294, 2020. 1, 3, 4, 6

[24] Shaojie Li, Jie Wu, Xuefeng Xiao, Fei Chao, Xudong Mao, and Rongrong Ji. Revisiting discriminator in GAN compression: A generator-discriminator cooperative compression scheme. *CoRR*, abs/2110.14439, 2021. 3

[25] Xinyang Li, Shengchuan Zhang, Jie Hu, Liujuan Cao, Xiaopeng Hong, Xudong Mao, Feiyue Huang, Yongjian Wu,

and Rongrong Ji. Image-to-image translation via hierarchical style disentanglement. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8639–8648. Computer Vision Foundation / IEEE, 2021. 1

[26] Zeqi Li, Ruowei Jiang, and Parham Aarabi. Semantic relation preserving knowledge distillation for image-to-image translation. In *European Conference on Computer Vision*, pages 648–663. Springer, 2020. 1, 3, 4, 6

[27] Ye Lin, Yanyang Li, Ziyang Wang, Bei Li, Quan Du, Tong Xiao, and Jingbo Zhu. Weight distillation: Transferring the knowledge in neural network parameters. *arXiv preprint arXiv:2009.09152*, 2020. 3

[28] Miao Liu, Xin Chen, Yun Zhang, Yin Li, and James M Rehg. Attention distillation for learning video representations. In *BMVC*, 2020. 1

[29] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 773–782, 2018. 3

[30] Xiao-Chang Liu, Yong-Liang Yang, and Peter Hall. Learning to warp for style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3702–3711. Computer Vision Foundation / IEEE, 2021. 2

[31] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. 1, 3

[32] Yuchen Liu, Zhixin Shu, Yijun Li, Zhe Lin, Federico Perazzi, and Sun-Yuan Kung. Content-aware gan compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12156–12166, 2021. 3

[33] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999. 3

[34] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. *arXiv preprint arXiv:1902.03393*, 2019. 8

[35] Hossein Mobahi, Mehrdad Farajtabar, and Peter L Bartlett. Self-distillation amplifies regularization in hilbert space. *arXiv preprint arXiv:2002.05715*, 2020. 8

[36] Ori Nizan and Ayellet Tal. Breaking the cycle-colleagues are all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7860–7869, 2020. 3

[37] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy*, pages 582–597, 2016. 3

[38] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020. 3

[39] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Gaugan: semantic image synthesis with spatially adaptive normalization. In *ACM SIGGRAPH 2019 Real-Time Live!*, pages 1–1. 2019. 3

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

[41] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: A stylegan encoder for image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2287–2296. Computer Vision Foundation / IEEE, 2021. 1

[42] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 3

[43] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580, 2019. 1

[44] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 1

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

[46] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29:613–621, 2016. 2

[47] Pei Wang, Yijun Li, and Nuno Vasconcelos. Rethinking and improving the robustness of image style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 124–133. Computer Vision Foundation / IEEE, 2021. 2

[48] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2, 6

[49] Travis Williams and Robert Li. Wavelet pooling for convolutional neural networks. In *International Conference on Learning Representations*, 2018. 3

[50] Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. Bert-of-theseus: Compressing bert by progressive module replacing. *arXiv preprint arXiv:2002.02925*, 2020. 3

[51] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng. Revisiting knowledge distillation via label smoothing regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[52] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 3, 4, 6

[53] Linfeng Zhang and Ma Kaisheng. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *ICLR*, 2021. 1, 3

[54] Linfeng Zhang, Yukang Shi, Zuoqiang Shi, Kaisheng Ma, and Chenglong Bao. Task-oriented feature distillation. In *NeurIPS*, 2020. 3

[55] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *arXiv preprint:1905.08094*, 2019. 8

[56] Linfeng Zhang, Zhanhong Tan, Jiebo Song, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Scan: A scalable neural networks framework towards compact and efficient models. *ArXiv*, abs/1906.03951, 2019. 3

[57] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018. 8

[58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Pix2pix datasets, `http://efrosgans.eecs.berkeley.edu/cyclegan/datasets`. 6

[59] 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*, pages 2223–2232, 2017. 1, 2, 6