IMITATE YOUR OWN REFINEMENT: KNOWLEDGE DISTILLATION SHEDS LIGHT ON EFFICIENT IMAGE-TO-IMAGE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The excellent performance of the state-of-the-art Generative Adversarial Networks (GANs) is always accompanied by enormous parameters and computations, making them unaffordable on resource-limited mobile devices. As an effective model compression technique, knowledge distillation (KD) has been proposed to transfer the knowledge from a cumbersome teacher to a lightweight student. Following its success on classification, some recent works have applied KD to GAN-based image-to-image translation but lead to unsatisfactory performance. In this paper, to tackle this challenge, we propose a novel knowledge distillation framework named **IYOR** (*Imitate Your Own Refinement*), which consists of the following two techniques. Firstly, since image-to-image translation is an ill-posed problem, knowledge distillation on image-to-image translation may force the student to learn the average results between multiple correct answers and thus harm student performance. To address this problem, we propose to replace the teacher network in knowledge distillation with a refining network, which is trained to refine the images generated by the student to make them more realistic. During the training period, the refining network and the student are trained simultaneously, and the student is trained to imitate the refined results in a knowledge distillation manner. Secondly, instead of only distilling the knowledge in the generated images, we propose SIFT KD, which firstly extracts the distinctive and scale-invariant features of the generated images with Scale-invariant feature transform (SIFT), and then distills them from the refining network to the student. Extensive experimental results demonstrate the effectiveness of our method on five datasets with nine previous knowledge distillation methods. Our codes are available in the supplementary material and will be released on Github.

1 INTRODUCTION

In the last decade, Generative Adversarial Networks (GANs) have evolved to one of the most dominated methods for content generation of images (Isola et al., 2017; Zhu et al., 2017a), videos (Vondrick et al., 2016), text (Zhang et al., 2016), audios (Kong et al., 2020), graphs (Wang et al., 2018a), point clouds (Li et al., 2019) and multi-modal systems (Zhu et al., 2017b). Their remarkable ability of representation and generation has significantly boosted the performance of image-to-image translation and further promoted their usage in real-world applications. Despite their impressive performance, GANs models usually suffer from massive parameters and computation, which have limited them to deploy on resource-restricted platforms such as mobile phones. This problem further raises the research trend in model compression such as network pruning (Buciluǎ et al., 2006; He et al., 2018a; 2017), weights quantization (Lee et al., 2019; Nagel et al., 2019), lightweight model design (Ma et al., 2018; Sandler et al., 2018; Howard et al., 2017), neural network architecture search (Howard et al., 2019; He et al., 2018b), and knowledge distillation (Hinton et al., 2014).

Knowledge distillation (KD), which aims to improve the performance of lightweight students by transferring knowledge from an over-parameterized teacher model, has become a popular technique for model compression. By imitating the prediction results and the intermediate features of teachers, students can achieve significant performance improvements. Following its success in image classification (Hinton et al., 2014; Zhang et al., 2020), object detection (Zhang & Ma, 2021) and semantic

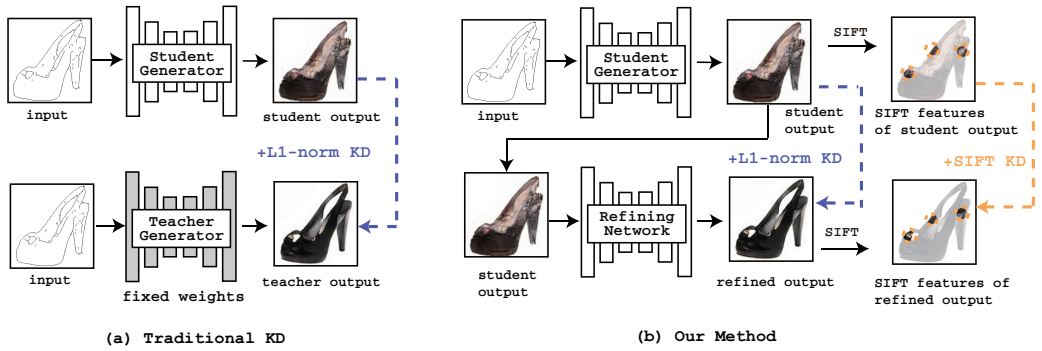

Figure 1: Comparison between traditional KD and **IYOR** on Edge→Shoe. **(a)** The teacher network in traditional KD is trained to translate images of edges to shoes. **(b)** In contrast, the refining network in **IYOR** is trained to refine the shoe images generated by students to make them more realistic. Since the refining network is conditioned on the student outputs, the refined results tend to be more consistent with student outputs than teacher outputs, and thus they are better learning targets for the student. **(c)** In traditional KD, the teacher network is firstly pretrained and then used for knowledge distillation with its weights frozen. In our method, the refining network is trained with the student simultaneously. **(d)** Besides, we propose SIFT KD, which firstly extracts the scale-invariant and highly distinctive features in the generated images with Scale-invariant feature transform (SIFT), and then distills them to the students. Note that the student and refining network are also trained with the original loss in GANs, which are not shown in the figure for simplicity.

segmentation (Yang et al., 2022), Recently, some researchers have tried to apply knowledge distillation to image-to-image translation by training students to mimic the images generated by the teachers. Unfortunately, these trials usually lead to limited and even sometimes negative performance (Li et al., 2020c; Zhang et al., 2022). Some works have been proposed to distill teacher knowledge in their features and lead to positive effectiveness (Ren et al., 2021; Li et al., 2020c). However, there is still no analysis on the reason that why traditional image-based knowledge distillation fails.

In this paper, we mainly impute the unsatisfactory performance of naive knowledge distillation to the *ill-posed property of image-to-image translation*. Unlike image classification, where each image always has a unique categorical label, an image can have multiple different but correct post-translation answers in image-to-image translation. For example, in Edge→Shoe translation (*i.e.,* translating edges of shoes to photos), given an input image of edges, there are multiple corresponding images of shoes with different colors, styles, and contents. All of these images can be correct answers while the average of them may have low quality. Unfortunately, in traditional KD, the student and teacher are likely to give two *different* but correct predictions for the same input image. In this case, the knowledge distillation loss forces the students to learn the average between the student outputs and the teacher outputs, which can harm student performance acutely. In contrast, the ideal case to avoid this problem is to guarantee that the student and teacher output the *consistent* answers for the input image. However, this assumption does not always hold since the student and the teacher in traditional KD are two independent image-to-image translation models.

To address this problem, we propose **IYOR** (*Imitate Your Own Refinement*), a generalized knowledge framework which introduces a different manner to build the "teacher network" in knowledge distillation. Taking Edge→Shoe translation as an example, as shown in Figure 1, instead of building a teacher network which translates edges into shoes, **IYOR** introduces a refining network, which takes the shoe images generated by the student as inputs, refines them, and outputs the images of the shoe which have much better quality. Note that the refining network is trained with the student simultaneously and can be discarded during inference to avoid additional parameters and computations. Since the refining network has much more parameters than the student, this refining process can significantly improve the quality of images generated by students. Hence, the refined results can be considered as the "teacher outputs" in traditional knowledge distillation, and utilized as the learning targets of the students. The major advantage of **IYOR** is that *the refining network is conditioned on the outputs of students, instead of the original inputted images*. Hence, the refined results are more likely to be consistent with the student outputs than the teacher outputs in traditional knowledge distillation. As a result, it can alleviate the problem of ineffective knowledge distillation caussed by

the ill-posed property. Extensive experiments show that dramatic performance gain of five datasets can be observed by simply replacing the traditional teacher network with the refining network.

Moreover, instead of directly training the student to imitate the images generated by the refining network pixel by pixel, we further propose SIFT distillation which adopts Scale Invariant Feature Transform (SIFT) (Lowe, 1999), a typical image feature extraction method in traditional image processing to extract the scale-invariant and highly distinctive features of the generated images and then distills them from the refining network to the students. As pointed out by abundant previous research (Lowe, 1999; 2004; Yuan et al., 2008), the features extracted by SIFT are invariant to image scaling, rotation and illumination, and highly distinctive for downstream tasks such as detection and tracking. Hence, these features carry more semantic information of the images, and they are more beneficial in knowledge distillation than traditional pixel-wise imitating. Another advantage of SIFT KD is that SIFT does not contain any trainable parameters, which makes SIFT KD generalize well on different image-to-image translation tasks as a plug-and-play knowledge distillation technique.

Experimental results on five image-to-image translation tasks have demonstrated the performance of **IYOR** for both paired and unpaired image-to-image translation in terms of both quantitative and qualitative analysis. Despite its simplicity, **IYOR** outperforms the previous nine knowledge distillation methods by a clear margin. Besides, experimental results also demonstrate that **IYOR** can be combined with the previous feature-based knowledge distillation methods to achieve better performance. To sum up, our main contributions can be summarized as follows.

- We propose **IYOR**, a knowledge distillation method for efficient image-to-image translation. To the best of our knowledge, **IYOR** firstly shows that the most naive image-based knowledge distillation can be effective by replacing the teacher with a refining network.
- We propose SIFT distillation, which adopts SIFT to extract the distinctive and scale-invariant features of images and distill them from the refining network to the student.
- Extensive experiments on both paired and unpaired translation tasks have demonstrated the performance of **IYOR** over nine previous methods and five datasets in terms of both quantitative and qualitative results. Our codes have been released for future research.

2 Related Work

2.1 Image-to-Image Translation with GANs

Remarkable progress has been achieved in image-to-image translation with the rapid development of generative adversarial networks (GANs) (Goodfellow et al., 2014; Brock et al., 2018). Pix2Pix is first proposed to perform paired image-to-image translation with conditional GANs (Isola et al., 2017). Then, Pix2PixHD is proposed to improve the generation quality with multi-scale generators and discriminators (Wang et al., 2018b). The similar idea has also been extended in text-to-image translation (Zhang et al., 2017), multi-modal image-to-image translation (Huang et al., 2018; Zhu et al., 2017c) and applications such as super-resolution and image dehazing (Wang et al., 2018d; Ledig et al., 2017; Zhang et al., 2017). In the real-world applications, the paired image-to-image translation dataset is usually not available. To address this problem, abundant methods have been proposed to perform image-to-image translation on unpaired datasets with cycle-consistency regularization (Zhu et al., 2017a; Yi et al., 2017; Kim et al., 2017). StarGAN is proposed to perform image-to-image translation for multiple domains with a single model (Choi et al., 2018), and StarGAN v2 is proposed to increase the scalability and the diversity of image-to-image translation models at the same time (Choi et al., 2020). Attention based GANs have been widely utilized to improve the performance of image-to-image translation by localizing the to-be-translated regions with attention modules (Tang et al., 2021; Chen et al., 2018; Emami et al., 2020; Alami Mejjati et al., 2018). Recently, some researchers have proposed to replace the convolutional layers in GAN with MLP-mixers and vision transformers, which leads to better high-fidelity translation (Wan et al., 2021; Cazenavette & De Guevara, 2021).

2.2 Knowledge Distillation

The idea that employing a large model to improve the performance of a small model is firstly proposed by Buciluǎ (Buciluǎ et al., 2006) for the compression of neural network ensemble. Then, Hinton *et al.* propose the concept of knowledge distillation, which introduces a temperature hyper-parameter in the softmax layer to flatter teacher prediction (Hinton et al., 2014). Following their

success, many researchers have proposed to not only distill the teacher knowledge in its predicted categorical probability distribution, but also the dark knowledge in features (Romero et al., 2015; Tian et al., 2019), spatial attention (Zagoruyko & Komodakis, 2017), channel-wise attention (Liu et al., 2021a; Shu et al., 2021; Li et al., 2021a), pixel-wise relation (Zhang & Ma, 2021; Li et al., 2020c; Yoon et al., 2020), instance-wise relation (Park et al., 2019b; Tung & Mori, 2019; Peng et al., 2019), task-oriented information (Zhang et al., 2020), decision boundary samples (Heo et al., 2019b), positive feature (Heo et al., 2019a) and frequency-biased information (Zhang et al., 2022) with optimization methods such as L_2-norm distance (Romero et al., 2015; Yim et al., 2017), adversarial learning (Shen et al., 2019; Liu et al., 2019a; Xu et al., 2017), and contrastive learning (Tian et al., 2019; Chen et al., 2020b). Besides image classification, knowledge distillation has already been used in model compression for object detection (Chen et al., 2017; Li et al., 2017; Wang et al., 2019; Bajestani & Yang, 2020; Li et al., 2020b), semantic segmentation (Liu et al., 2019b; Park & Heo, 2020), pre-trained language models (Sanh et al., 2019; Xu et al., 2020)s and so on.

Knowledge Distillation on Image-to-Image Translation A few research has been proposed to perform knowledge distillation on image-to-image translation. Li *et al.* propose the framework of GAN compression, which has applied the classic L_2-norm feature distillation on the intermediate neural layers (Li et al., 2020a). However, their results demonstrate that this application leads to unsatisfying performance improvements. Then, Li *et al.* propose the semantic relation preserving knowledge distillation, which aims to distill the relation between different patches in the generated images instead of the encoded features (Li et al., 2020c). Then, Chen *et al.* propose to distill image-to-image translation models with knowledge distillation not only generators but also the discriminators (Chen et al., 2020a). Similarly, Li *et al.* propose to revisit the discriminator in GAN compression, which transfers the knowledge in the teacher discriminator with L_2-norm and texture loss (Li et al., 2021b). Jin *et al.* introduce the centered kernel alignment as the distance metric in knowledge distillation, which does not require additional layers for feature reshaping. Ren *et al.* propose to train the teacher and student GANs simultaneously, which shows the possibility of online knowledge distillation on image-to-image translation (Ren et al., 2021). Recently, motivated by the fact that tiny GANs work badly in generating high-quality high-frequency information, Zhang *et al.* propose to distill only the high-frequency information decomposed by discrete wavelet transformation in the images generated by teachers (Zhang et al., 2022). Besides image-to-image translation, there are also some knowledge distillation methods designed for GAN compression on the other tasks (Liu et al., 2021b; Wang et al., 2018c; Aguinaldo et al., 2019). Unfortunately, most of these knowledge distillation methods focus on distilling teacher knowledge in their features, and sufficient evidences show that directly training students to mimic the generated images from teachers leads to insufficient and even negative performance (Li et al., 2020c; Zhang et al., 2022). In contrast, this paper firstly shows that naive image-based distillation can also achieve valuable performance boosts.

3 METHODOLOGY

3.1 KNOWLEDGE DISTILLATION

In this section, we firstly revisit the formulation of knowledge distillation on image classification and then simply extend them to image-to-image translation. Given a set of training samples $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ and the corresponding ground truth $\mathcal{Y} = \{y_1, y_2, ..., y_n\}$, by denoting the student function and the pre-trained teacher function as f_s and f_t, then the training loss of classical knowledge distillation method (Hinton et al., 2014) can be formulated as

$$\arg\min_{f_s} \mathbb{E}_{x,y}\Big[(1-\alpha)\cdot \mathrm{CE}(f_s(x), y) + \alpha \cdot \mathrm{KL}(f_s(x)/\tau, f_t(x)/\tau)\Big], \tag{1}$$

where CE and KL indicate cross-entropy loss and the Kullback-Leibler divergence, respectively. τ is the temperature hyper-parameter to soften the probability distribution and α is a hyper-parameter to balance the origin training loss and the knowledge distillation loss. When knowledge distillation is applied to image-to-image translation, since the predictions of students and teachers are the value of pixels instead of probability distributions, KL divergence can be replaced with the L_1-norm loss, which is widely utilized in low-level vision. And the cross-entropy loss for classification should be replaced with the GAN training loss. Taking Pix2Pix (Isola et al., 2017) as an example, the knowledge distillation loss (Hinton et al., 2014) for training the generator can be formulated as

$$\arg\min_{f_s} \mathbb{E}_{x,y}\Big[(1-\alpha)\cdot L_1\big(f_s(x), y\big) + \alpha \cdot L_1\big(f_s(x), f_t(x)\big) + L_{\mathrm{cGAN}}\big(f_s(x)\big)\Big], \tag{2}$$

where L_1 indicates the L_1-norm loss. L_{cGAN} indicates the conditional GAN loss, which measures how the generated images fool the discriminator. Note that we do not introduce $L_{c\text{GAN}}$ and the discriminator of GANs in detail here since they have no direct influence with our method.

3.2 **IYOR**: IMITATE YOUR OWN REFINEMENT

Instead of using two independent neural networks as the student and the teacher we append a refining network f_r after the student network, which is trained to translate the images generated by the student network $f_r(f_s(x))$ to the corresponding ground-truth y. Thus, the "teacher model" in **IYOR** can be written as $f_t = f_s \circ f_r$. In our implementation, f_r has the same architecture as the teacher in traditional KD and hence it has enough learning ability to refine student outputs. Note that the f_r can be discarded after the training period to avoid the additional parameters and computation. Besides, unlike traditional KD where the teacher is first pre-trained and then utilized to teach the student, in **IYOR**, f_r and f_s are trained simultaneously. For simplicity, by denoting $z_s = f_s(x)$ and $z_r = f_r \circ f_s(x)$, the training objective of the refining network f_r can be formulated as

$$\arg\min_{f_r} \mathbb{E}_{x,y}\Big[L_1\big(z_r, y\big) + L_{\text{cGAN}}\big(z_r\big)\Big]. \tag{3}$$

And the training objective of the student f_s can be formulated as

$$\arg\min_{f_s} \mathbb{E}_{x,y}\Big[(1-\alpha)\cdot L_1\big(z_s, y\big) + \alpha \cdot L_1\big(z_s, z_r\big) + L_{\text{cGAN}}\big(z_s(x)\big)\Big]. \tag{4}$$

Note that since **IYOR** only distills the generators of GANs, we omit the description of discriminators here. Besides, **IYOR** can be easily extended to unpaired image-to-image translation models such as CycleGAN by introducing two refining networks to both the two translation directions, respectively.

3.3 WHY IYOR WORKS

Consider a general knowledge distillation with the L_1-norm, define a function G as

$$G\big(f_s(x), f_t(x)\big) := \mathbb{E}_{x,y}\Big[(1-\alpha)\cdot L_1\big(f_s(x), y\big) + \alpha \cdot L_1\big(f_s(x), f_t(x)\big) + H\big(f_s(x)\big)\Big], \tag{5}$$

where H is a function about the student network. For simplicity, we abbreviate $\mathbb{E}_{x,y}$ as $\mathbb{E}$. The objective function of traditional knowledge distillation (TKD) (2) and **IYOR** (4) are specific cases of equation (5). Let f_s^1 and f_s^2 be the optimal student networks of problem (2) and **IYOR** (4), we will provide an assumption and a theorem to interpret the effectiveness of **IYOR**.

$$\texttt{TKD}: f_s^1 = \arg\min_{f_s} G\big(f_s(x), f_t(x)\big), \qquad \textbf{IYOR}: f_s^2 = \arg\min_{f_s, f_r} G\big(f_s(x), f_r \circ f_s(x)\big). \tag{6}$$

Assumption 3.1 *Since our teacher $f_r \circ f_s(x)$ has more parameters than the traditional teacher $f_t(x)$, we assume that when they achieve the optimal values, the loss of TKD is less than **IYOR**. In other words, denoting f_t^1 and f_t^2 as the optimal teacher networks of TKD and **IYOR**, then we have*

$$G(f_s^2, f_t^2) \leq G(f_s^1, f_t^1). \tag{7}$$

Theorem 3.1 *Under the Assumption (3.1), the L_1 distance between the optimal student network and teacher network in **IYOR** is less than that in TKD, which means*

$$\mathbb{E}\Big[L_1(f_s^2, f_t^2)\Big] \leq \mathbb{E}\Big[L_1(f_s^1, f_t^1)\Big]. \tag{8}$$

Please refer to Appendix A for the proof. Besides, we have also explained why traditional KD methods fail on image-to-image translation with VC theory in Appendix B.

3.4 SIFT DISTILLATION

Scale-Invariant Features Transform (SIFT) is one of the most effective and popular image descriptors in classicial image processing. Usually, SIFT mainly has four steps, including scale-space extrema detection, keypoint localization, orientation assignment and keypoint description. Usually, SIFT features carry rich semantic information of the images while having much lower dimensions than the original image. Thus, distilling the SIFT features is more efficient than directly distilling the pixels of the generated images. By denoting SIFT as $\phi(\cdot)$ and a loss hyper-parameter as β, then the loss function of SIFT distillation in our method can be formulated as

$$\arg\min_{f_s} \mathbb{E}_{x,y}\Big[(1-\alpha)\cdot L_1\big(z_s, y\big) + \alpha \cdot L_1\big(z_s, z_r\big) + L_{\text{cGAN}}\big(z_s(x)\big) + \beta \cdot L_1\big(\phi(z_s), \phi(z_r)\big)\Big]. \tag{9}$$

Table 1: Experiment results on unpaired image-to-image translation on Horse→Zebra and Zebra→Horse with CycleGAN. Δ indicates the performance improvements compared with the origin student. Each result is averaged from 8 trials. **A lower FID is better.**

Horse→Zebra

#Params (M)	FLOPs (G)	Method	Metric		#Params (M)	FLOPs (G)	Method	Metric	
			FID↓	Δ ↑				FID↓	Δ ↑
11.38	49.64	Teacher	$61.34_{\pm 4.35}$	–	11.38	49.64	Teacher	$61.34_{\pm 4.35}$	–
		Origin Student	$85.04_{\pm 6.88}$	–			Origin Student	$70.54_{\pm 9.63}$	–
		Hinton *et al.*	$84.08_{\pm 3.78}$	0.96			Hinton *et al.*	$70.35_{\pm 3.27}$	0.18
		Zagoruyko *et al.*	$81.24_{\pm 2.01}$	3.80			Zagoruyko *et al.*	$67.51_{\pm 4.57}$	3.03
		Li and Lin *et al.*	$83.97_{\pm 5.01}$	1.07			Li and Lin *et al.*	$68.94_{\pm 2.98}$	1.60
		Li and Jiang *et al.*	$81.74_{\pm 4.65}$	3.30			Li and Jiang *et al.*	$68.94_{\pm 2.98}$	1.60
		Jin *et al.*	$82.37_{\pm 8.56}$	2.67			Jin *et al.*	$67.31_{\pm 3.01}$	3.23
0.72	3.35	Ahn *et al.*	$82.91_{\pm 2.41}$	2.13	1.61	7.29	Ahn *et al.*	$69.32_{\pm 5.89}$	1.22
15.81×	14.82×	Ren *et al.*	$77.31_{\pm 6.41}$	7.73	7.08×	6.80×	Ren *et al.*	$64.78_{\pm 5.21}$	5.76
		Li *et al.*	$79.29_{\pm 7.31}$	5.75			Li *et al.*	$66.85_{\pm 6.17}$	3.69
		Zhang *et al.*	$77.04_{\pm 3.52}$	8.00			Zhang *et al.*	$61.65_{\pm 4.73}$	8.89
		Ours	$69.67_{\pm 5.32}$	15.37			Ours	$56.45_{\pm 2.59}$	14.09
		Ours + Ren *et al.*	$67.32_{\pm 4.32}$	17.72			Ours + Ren *et al.*	$55.32_{\pm 2.97}$	15.22
		Ours + Li *et al.*	$68.32_{\pm 5.20}$	16.72			Ours + Li *et al.*	$55.85_{\pm 4.06}$	14.69
		Ours + Zhang *et al.*	$67.21_{\pm 4.91}$	17.83			Ours + Zhang *et al.*	$55.34_{\pm 4.40}$	15.20

Zebra→Horse

#Params (M)	FLOPs (G)	Method	Metric		#Params (M)	FLOPs (G)	Method	Metric	
			FID↓	Δ ↑				FID↓	Δ ↑
11.38	49.64	Teacher	$138.07_{\pm 4.01}$	–	11.38	49.64	Teacher	$138.07_{\pm 4.01}$	–
		Origin Student	$152.57_{\pm 9.63}$	–			Origin Student	$141.86_{\pm 1.57}$	–
		Hinton *et al.*	$148.64_{\pm 1.62}$	4.03			Hinton *et al.*	$142.03_{\pm 3.27}$	-0.17
		Zagoruyko *et al.*	$148.92_{\pm 1.20}$	3.75			Zagoruyko *et al.*	$141.23_{\pm 1.88}$	0.63
		Li and Lin *et al.*	$151.32_{\pm 2.31}$	1.35			Li and Lin *et al.*	$141.32_{\pm 1.27}$	0.54
		Li and Jiang *et al.*	$151.09_{\pm 3.67}$	1.58			Li and Jiang *et al.*	$141.16_{\pm 1.31}$	0.70
		Jin *et al.*	$149.73_{\pm 3.94}$	2.94			Jin *et al.*	$140.98_{\pm 1.41}$	0.88
0.72	3.35	Ahn *et al.*	$150.31_{\pm 3.55}$	2.36	1.61	7.29	Ahn *et al.*	$141.50_{\pm 2.51}$	0.36
15.81×	14.82×	Ren *et al.*	$147.34_{\pm 2.98}$	5.23	7.08×	6.80×	Ren *et al.*-	$140.87_{\pm 2.03}$	0.99
		Li *et al.*	$148.30_{\pm 1.53}$	4.27			Li *et al.*	$140.92_{\pm 2.31}$	0.94
		Zhang *et al.*	$146.01_{\pm 1.80}$	6.66			Zhang *et al.*	$138.84_{\pm 1.47}$	3.02
		Ours	$144.20_{\pm 2.78}$	8.37			Ours	$137.98_{\pm 2.90}$	3.88
		Ours + Ren *et al.*	$143.01_{\pm 3.11}$	9.56			Ours + Ren *et al.*	$137.57_{\pm 1.40}$	4.29
		Ours + Li *et al.*	$143.16_{\pm 2.87}$	9.40			Ours + Li *et al.*	$137.60_{\pm 1.57}$	4.26
		Ours + Zhang *et al.*	$143.08_{\pm 2.10}$	9.48			Ours + Zhang *et al.*	$137.18_{\pm 1.89}$	4.68

4 EXPERIMENT

4.1 EXPERIMENT SETTINGS

Models and Datasets In this paper, we mainly evaluate the performance of our method with CycleGAN (Zhu et al., 2017a) for unpaired image-to-image translation, and Pix2Pix (Isola et al., 2017) and Pix2PixHD (Wang et al., 2018b) for paired image-to-image translation. The refining network in our method has an identical architecture to the original model before compression. The students in our experiments have the same network depth as the original model before compression except for fewer channels. Five datasets are utilized for quantitative evaluation, including Horse→Zebra, Maps, Edge→Shoe, Summer→Winter, and Apple→Orange.

Comparison Methods We have compared our methods with nine knowledge distillation methods, including three of them which are firstly proposed for image classification and then adopted by us to image-to-image translation (Hinton et al., 2014; Ahn et al., 2019; Zagoruyko & Komodakis, 2017), and six of them which are designed for image-to-image translation (Li et al., 2020a; Jin et al., 2021; Zhang et al., 2022; Li et al., 2021b; Ren et al., 2021; Li et al., 2020c). Note that some comparison methods have both knowledge distillation and neural network pruning. Following the setting of the previous work (Zhang et al., 2022), we only compare our method with their knowledge distillation algorithms for a fair comparison.

Training and Evaluation Settings We adopt the same training setting from the origin implementation of CycleGAN and Pix2Pix. Models for Edge→Shoe and the other datasets are trained by 50 and 200 epochs, respectively. Following previous works, we adopt *Frechet Inception Distance (FID)* as the performance metric for all datasets. A lower FID indicates that the distribution of the generated

Table 2: Experimental results on paired image-to-image translation on the Edge→Shoe dataset with Pix2Pix and Pix2PixHD. Δ indicates the performance improvements compared with the origin student. Each result is averaged from 8 trials. "Origin Student" indicates the student trained without knowledge distillation. **A lower FID is better**.

Pix2Pix on Edge→Shoe

#Params (M)	FLOPs (G)	Method	Metric	
			FID↓	Δ ↑
54.41	6.06	Teacher	$59.70_{\pm 0.91}$	–
		Origin Student	$85.06_{\pm 0.98}$	–
		Hinton *et al.*	$86.97_{\pm 3.49}$	-1.91
		Zagoruyko *et al.*	$84.25_{\pm 2.08}$	0.81
		Li and Lin *et al.*	$83.63_{\pm 3.12}$	1.43
		Li and Jiang *et al.*	$84.01_{\pm 2.31}$	1.05
		Jin *et al.*	$84.39_{\pm 3.62}$	0.67
13.61	1.56	Ahn *et al.*	$84.92_{\pm 0.78}$	0.14
4.00×	3.88×	Ren *et al.*	$80.31_{\pm 2.59}$	4.75
		Li *et al.*	$81.24_{\pm 3.74}$	3.82
		Zhang *et al.*	$80.13_{\pm 2.18}$	4.93
		Ours	$77.90_{\pm 2.20}$	7.16
		Ours + Ren *et al.*	$76.34_{\pm 1.84}$	8.72
		Ours + Li *et al.*	$77.09_{\pm 2.34}$	7.97
		Ours + Zhang *et al.*	$76.27_{\pm 3.21}$	8.79

Pix2PixHD on Edge→Shoe

#Params (M)	FLOPs (G)	Method	Metric	
			FID↓	Δ ↑
11.38	49.64	Teacher	$41.59_{\pm 0.42}$	–
		Origin Student	$44.64_{\pm 0.54}$	–
		Hinton *et al.*	$45.31_{\pm 0.63}$	-0.67
		Zagoruyko *et al.*	$44.21_{\pm 0.72}$	0.43
		Li and Lin *et al.*	$44.03_{\pm 0.41}$	0.61
		Li and Jiang *et al.*	$43.90_{\pm 0.36}$	1.28
		Jin *et al.*	$43.97_{\pm 0.17}$	1.21
1.61	1.89	Ahn *et al.*	$44.53_{\pm 0.48}$	0.11
28.23×	25.59×	Ren *et al.*	$42.98_{\pm 0.34}$	1.66
		Li *et al.*	$43.21_{\pm 0.35}$	0.29
		Zhang *et al.*	$42.53_{\pm 0.29}$	2.11
		Ours	$41.37_{\pm 0.67}$	3.27
		Ours + Ren *et al.*	$41.02_{\pm 0.48}$	3.62
		Ours + Li *et al.*	$41.28_{\pm 0.81}$	3.36
		Ours + Zhang *et al.*	$40.90_{\pm 0.42}$	3.74

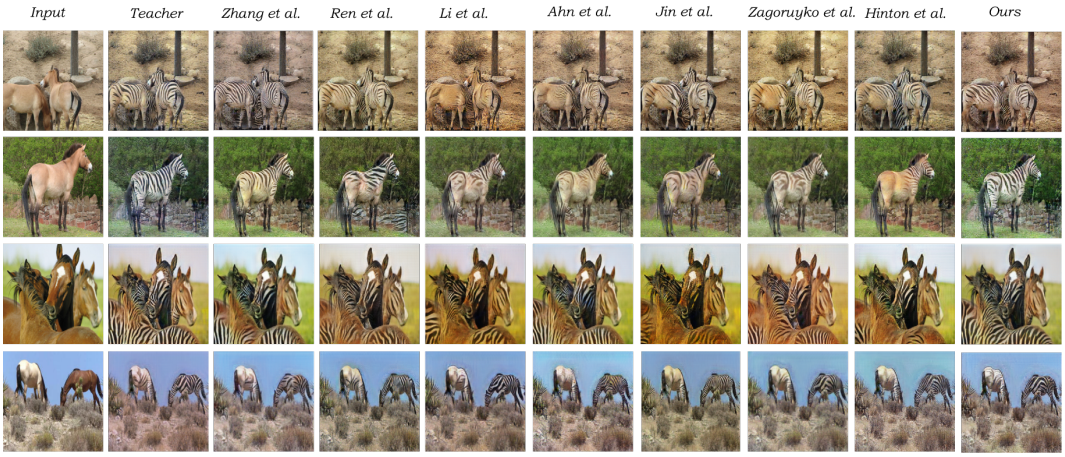

Figure 2: Qualitative comparison between our methods and previous knowledge distillation methods with 14.82× accelerated and 15.81 × compressed CycleGAN students on Horse→Zebra.

images and the real images have a lower distance, and thus the generated images have better quality. On paired image-to-image translation, we report model performance at the last epoch. On unpaired image-to-image translation, since the performance for different epochs is unstable, we compute the FID for every five epochs and report the lowest one. For both paired and unpaired image-to-image translation, FID are computed over only the images in the test set.

4.2 Experiment Results

Quantitative Results Quantitative comparison with previous knowledge distillation methods on unpaired image-to-image translation and paired image-to-image translation datasets are shown in Table 1 and Table 2, respectively. It is observed that: (i) Directly applying the naive image-based knowledge distillation (Hinton et al., 2014) leads to very limited and even negative performance. For instance, it leads to 1.91 and 0.67 FID increments (performance drop) on Edge→Shoe with Pix2Pix and Pix2PixHD, respectively. (ii) In contrast, by replacing the teacher in naive image-based with the refining network in our method, knowledge distillation leads to consistent performance improvements. On average, 5.12 and 10.43 FID decrements (performance improvements) can be gained in paired and unpaired image-to-image translation, respectively. (iii) Combining our method with previous feature-based knowledge distillation leads to further performance improvements. For instance, on the 14.82× compressed and 6.80× compressed Horse→Zebra students, combining our method with the method of Ren *et al.* leads to 2.35 and 1.13 further FID decrements. (iv) Table 3 further demonstrates the effectiveness of our method in more compression ratios and more datasets. These

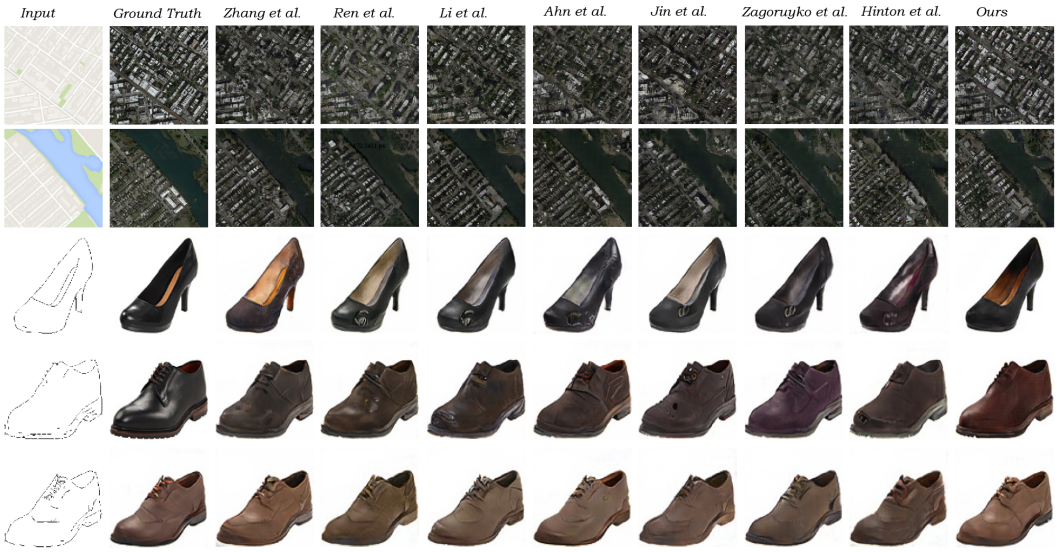

Figure 3: Qualitative results between our methods and previous knowledge distillation methods with 3.88 $\times$ accelerated and 4.00 $\times$ compressed Pix2Pix students on Maps and Edge→Shoe.

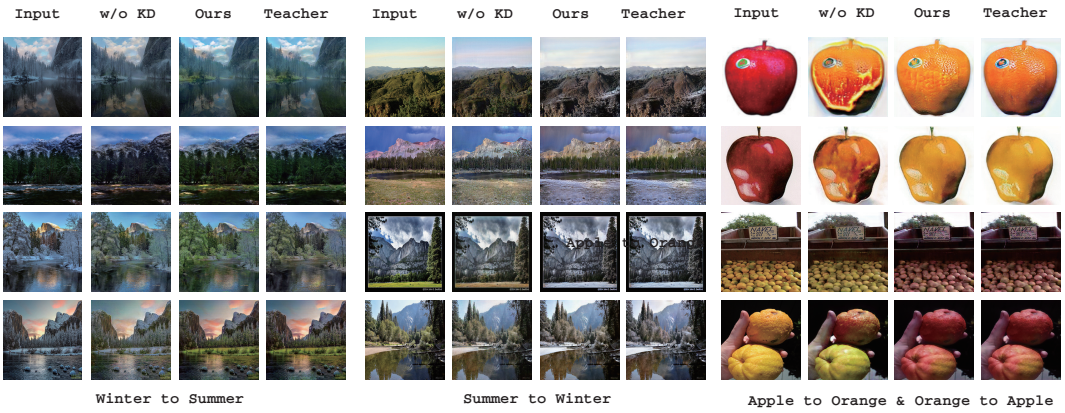

Figure 4: Qualitative results on Winter→Summer, Summer→Winter, Orange→Apple and Apple→Orange with with 14.82$\times$ accelerated and 15.81 $\times$ compressed CycleGAN students.

observations demonstrate that our method can significantly improve the performance of lightweight image-to-image translation models in a wide range of settings.

Qualitative Results Qualitative comparison between our methods and previous methods on unpaired and paired image-to-image translation datasets are shown in Figure 2 and Figure 3 , respectively. Besides, Figure 4 further shows the performance of our method on the other two datasets. It is observed that: (i) Compared with the model before compression (the teacher model), a significant performance drop can be observed on the student model trained without knowledge distillation. For instance, on Horse→Zebra, most student models can not transform the whole body of horses into stripes. Some previous knowledge distillation methods (*e.g.* Zhang *et al.*, Ren *et al.*) can alleviate this problem while our method leads to much better performance. (ii) On the Maps translation task, the buildings and the roads generated by students trained with previous knowledge distillation methods are fuzzy. In contrast, our method can generate clearer shapes and edges for buildings, roads, and rivers. (iii) On Edge→Shoe, the images generated by the students trained without knowledge distillation usually have severe corruption such as the holes in high-heeled shoes. In contrast, the images generated by our methods have better quality in terms of highlights, shapes, and colors. (iv) On Winter→Summer, the model trained by our method can successfully remove the snow on the plants. On Apple→Orange, the images generated by our method have much less corruption than the baseline model. These results demonstrate that students trained by **IYOR** achieve better performance in terms of not only statistical scores but also human vision.

Table 3: Experiment results of CycleGAN with and without **IYOR** on different compression ratios and datasets. Each result is averaged from 8 trials. The reported number is FID (lower is better).

Dataset	Param(M)	FLOPs(G)	without KD	with KD
Horse→Zebra	11.37	49.64	$61.34_{\pm 4.35}$	–
	1.61	7.29	$70.54_{\pm 9.63}$	$56.45_{\pm 2.59}$
	1.11	4.84	$76.09_{\pm 3.89}$	$59.25_{\pm 3.30}$
	0.72	3.35	$85.04_{\pm 6.88}$	$69.67_{\pm 5.32}$
	0.40	1.74	$104.30_{\pm 8.40}$	$91.53_{\pm 0.98}$
	0.28	1.21	$121.45_{\pm 16.78}$	$98.06_{\pm 6.88}$
	0.17	0.77	$134.51_{\pm 14.6}$	$111.75_{\pm 10.77}$
Apple→Orange	11.37	49.64	$117.59_{\pm 1.65}$	–
	2.84	12.41	$124.34_{\pm 1.87}$	$118.09_{\pm 2.82}$
	1.61	7.29	$135.70_{\pm 3.12}$	$120.53_{\pm 3.43}$
Summer→Winter	11.37	49.64	$81.61_{\pm 3.80}$	–
	2.84	12.41	$93.18_{\pm 2.73}$	$82.61_{\pm 1.16}$
	1.61	7.29	$104.51_{\pm 4.53}$	$85.30_{\pm 2.65}$

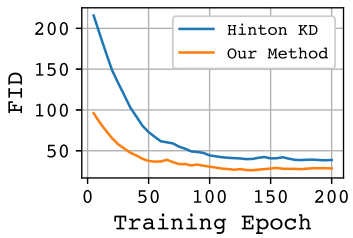

Figure 5: Comparison between our method and Hinton KD on the FID between students and teachers on Horse→Zebra with CycleGAN.

Table 4: Ablation study on SIFT distillation and the usage of the refining network on Horse→Zebra with CycleGAN students.

#Params	FLOPs	Refining	SIFT	FID↓	Δ ↑
1.61	7.29	×	×	$70.54_{\pm 9.63}$	–
		✓	×	$59.31_{\pm 2.89}$	11.23
		×	✓	$63.17_{\pm 3.66}$	7.37
		✓	✓	$56.45_{\pm 2.59}$	14.09
0.72	3.35	×	×	$85.04_{\pm 6.88}$	–
		✓	×	$72.53_{\pm 3.15}$	12.51
		×	✓	$78.11_{\pm 1.71}$	6.93
		✓	✓	$69.67_{\pm 5.32}$	15.37

5 Discussion

5.1 Ablation Study

In this paper, we mainly propose two knowledge distillation techniques, including (a) *learning from a refining network instead of a teacher network* and (b) *SIFT KD*. Table 4 shows the ablation study of the two techniques on Horse→Zebra with CycleGAN. It is observed that on the 7.08× and 15.81× compressed students: (i) 11.23 and 12.51 FID decrements can be observed by replacing the teacher network in traditional knowledge distillation with a refining network, respectively. (ii) 7.37 and 6.93 FID decrements can be observed by applying SIFT KD, respectively. (iii) 14.09 and 15.37 FID decrements can be obtained by combining the two techniques together, respectively. These observations indicate that both the two techniques have their own merits and their benefits are orthogonal.

5.2 Student-Teacher Similarity

In this subsection, we show that the refining network in **IYOR** has more consistent outputs with the student than the teacher in traditional KD. The FID between images generated by students and refining network in our method and the traditional KD method is shown in Figure 5. Note that A lower FID here indicates a larger student-teacher similarity. It is observed that our method leads to lower FID during the whole training period, indicating that compared with the teachers in traditional KD, the images generated by the refining network in our method are more likely to be consistent with images generated by the students. Besides, since FID measures the distance between the distribution of images generated by the student and the teacher, this observation also implies that the student in our method can learn teacher knowledge more effectively.

6 Conclusion

Due to the ill-posed property of image-to-image translation, directly applying traditional knowledge distillation usually leads to unsatisfactory and even negative impacts. To address this problem, we propose a new knowledge distillation method, named **IYOR** (imitate your own refinement), in which a refining network replaces the teacher network in traditional KD. During the training phase, the refining network strives to improve the quality of images generated by the students instead of generating images from the inputs. Hence, the refined results can be better learning targets than the teacher outputs that are used in traditional KD. Extensive quantitative and qualitative results have demonstrated that **IYOR** outperforms existing nine approaches in both paired and unpaired translation. Besides, SIFT knowledge distillation is also introduced to improve the effectiveness of knowledge distillation by extracting the distinctive and scale-invariant features of images and then distilling them from teachers to students. Furthermore, we have analyzed why traditional KD fails and **IYOR** works well on image-to-image translation theoretically.

REFERENCES

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. 4

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*, pp. 9163–9171, 2019. 6, 17

Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. *Advances in neural information processing systems*, 31, 2018. 3

Mohammad Farhadi Bajestani and Yezhou Yang. Tkd: Temporal knowledge distillation for active perception. In *The IEEE Winter Conference on Applications of Computer Vision*, pp. 953–962, 2020. 4

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

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*, pp. 535–541. ACM, 2006. 1, 3

George Cazenavette and Manuel Ladron De Guevara. Mixergan: An mlp-based architecture for unpaired image-to-image translation. *arXiv preprint arXiv:2105.14110*, 2021. 3

Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems*, pp. 742–751, 2017. 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*, volume 34, pp. 3585–3592, 2020a. 4

Liqun Chen, Zhe Gan, Dong Wang, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. *arXiv preprint arXiv:2012.08674*, 2020b. 4

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)*, pp. 164–180, 2018. 3

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 conference on computer vision and pattern recognition*, pp. 8789–8797, 2018. 3

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8188–8197, 2020. 3

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. 17

Hajar Emami, Majid Moradi Aliabadi, Ming Dong, and Ratna Babu Chinnam. Spa-gan: Spatial attention gan for image-to-image translation. *IEEE Transactions on Multimedia*, 23:391–401, 2020. 3

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. 3

Yang He, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang. Soft filter pruning for accelerating deep convolutional neural networks. *arXiv preprint arXiv:1808.06866*, 2018a. 1

Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1389–1397, 2017. 1

Yihui He, Ji Lin, Zhijian Liu, Hanrui Wang, Li-Jia Li, and Song Han. Amc: Automl for model compression and acceleration on mobile devices. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 784–800, 2018b. 1

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 International Conference on Computer Vision*, pp. 1921–1930, 2019a. 4

Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge distillation with adversarial samples supporting decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3771–3778, 2019b. 4

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS*, 2014. 1, 3, 4, 6, 7, 15, 17

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. *arXiv preprint arXiv:1905.02244*, 2019. 1

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In *CVPR*, 2017. 1

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 172–189, 2018. 3

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*, pp. 1125–1134, 2017. 1, 3, 4, 6

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*, pp. 13600–13611, 2021. 6, 16, 17

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pp. 1857–1865. PMLR, 2017. 3

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020. 1

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*, pp. 4681–4690, 2017. 3

Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1

Bo Li, Bin Chen, Yunxiao Wang, Tao Dai, Maowei Hu, Yong Jiang, and Shutao Xia. Knowledge distillation via channel correlation structure. In *International Conference on Knowledge Science, Engineering and Management*, pp. 357–368. Springer, 2021a. 4

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*, pp. 5284–5294, 2020a. 4, 6, 17

Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 6356–6364, 2017. 4

Ruihui Li, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-gan: a point cloud upsampling adversarial network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7203–7212, 2019. 1

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, 2021b. URL https://arxiv.org/abs/2110.14439. 4, 6, 17

Xiaojie Li, Jianlong Wu, Hongyu Fang, Yue Liao, Fei Wang, and Chen Qian. Local correlation consistency for knowledge distillation. In *European Conference on Computer Vision*, pp. 18–33. Springer, 2020b. 4

Zeqi Li, Ruowei Jiang, and Parham Aarabi. Semantic relation preserving knowledge distillation for image-to-image translation. In *European Conference on Computer Vision*, pp. 648–663. Springer, 2020c. 2, 4, 6, 15, 17

Li Liu, Qingle Huang, Sihao Lin, Hongwei Xie, Bing Wang, Xiaojun Chang, and Xiaodan Liang. Exploring inter-channel correlation for diversity-preserved knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8271–8280, 2021a. 4

Peiye Liu, Wu Liu, Huadong Ma, Tao Mei, and Mingoo Seok. Ktan: Knowledge transfer adversarial network. In *AAAI*, 2019a. 4

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*, pp. 2604–2613, 2019b. 4

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*, pp. 12156–12166, 2021b. 4

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *ICLR*, 2016. 15, 16

David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pp. 1150–1157. Ieee, 1999. 3

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 3

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 116–131, 2018. 1

Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1

Sangyong Park and Yong Seok Heo. Knowledge distillation for semantic segmentation using channel and spatial correlations and adaptive cross entropy. *Sensors*, 20(16):4616, 2020. 4

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!*, pp. 1–1. 2019a. 17

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019b. 4

Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5007–5016, 2019. 4

Yuxi Ren, Jie Wu, Xuefeng Xiao, and Jianchao Yang. Online multi-granularity distillation for gan compression. *CoRR*, abs/2108.06908, 2021. URL https://arxiv.org/abs/2108.06908. 2, 4, 6, 17

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015. 4

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018. 1

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. 4

Zhiqiang Shen, Zhankui He, and Xiangyang Xue. Meal: Multi-model ensemble via adversarial learning. In *AAAI*, 2019. 4

Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5311–5320, 2021. 4

Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 3

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 4

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1365–1374, 2019. 4

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29:613–621, 2016. 1

Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4692–4701, 2021. 3

Hongwei Wang, Jia Wang, Jialin Wang, Miao Zhao, Weinan Zhang, Fuzheng Zhang, Xing Xie, and Minyi Guo. Graphgan: Graph representation learning with generative adversarial nets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018a. 1

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4933–4942, 2019. 4

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*, pp. 8798–8807, 2018b. 3, 6

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Kdgan: Knowledge distillation with generative adversarial networks. *Advances in Neural Information Processing Systems*, 31, 2018c. 4

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018d. 3

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. 4

Zheng Xu, Yen-Chang Hsu, and Jiawei Huang. Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks. *arXiv preprint arXiv:1709.00513*, 2017. 4

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

Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pp. 2849–2857, 2017. 3

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4133–4141, 2017. 4

Donggeun Yoon, Jinsun Park, and Donghyeon Cho. Lightweight deep cnn for natural image matting via similarity-preserving knowledge distillation. *IEEE Signal Processing Letters*, 27:2139–2143, 2020. 4

Zhi Yuan, Peimin Yan, and Sheng Li. Super resolution based on scale invariant feature transform. In *2008 International Conference on Audio, Language and Image Processing*, pp. 1550–1554. IEEE, 2008. 3

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 4, 6, 17

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017. 3

Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *ICLR*, 2021. 1, 4

Linfeng Zhang, Yukang Shi, Zuoqiang Shi, Kaisheng Ma, and Chenglong Bao. Task-oriented feature distillation. In *NeurIPS*, 2020. 1, 4

Linfeng Zhang, Xin Chen, Xiaobing Tu, Pengfei Wan, Ning Xu, and Kaisheng Ma. Wavelet knowledge distillation: Towards efficient image-to-image translation. In *CVPR*, 2022. 2, 4, 6, 15, 17

Yizhe Zhang, Zhe Gan, and Lawrence Carin. Generating text via adversarial training. In *NIPS workshop on Adversarial Training*, volume 21, pp. 21–32. academia. edu, 2016. 1

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*, pp. 2223–2232, 2017a. 1, 3, 6, 17

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017b. 1

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017c. 3

A THE PROOF FOR THEOREM 3.1

Proof According to the Assumption 3.1, we have

$$\begin{aligned}&\mathbb{E}\Big[(1-\alpha)\cdot L_1\big(f_s^2,y\big)+\alpha\cdot L_1\big(f_s^2,f_t^2\big)+H(f_s^2)\Big]\\&\leq\mathbb{E}\Big[(1-\alpha)\cdot L_1\big(f_s^1,y\big)+\alpha\cdot L_1\big(f_s^1,f_t^1\big)+H(f_s^1)\Big]\end{aligned}\tag{10}$$

Since f_s^1 is the optimal solution for TKD (6), and $G(f_s^1,f_t^1)\leq G(f_s^2,f_t^1)$ implies

$$\begin{aligned}&\mathbb{E}\Big[(1-\alpha)\cdot L_1\big(f_s^1,y\big)+\alpha\cdot L_1\big(f_s^1,f_t^1\big)+H(f_s^1)\Big]\\&\leq\mathbb{E}\Big[(1-\alpha)\cdot L_1\big(f_s^2,y\big)+\alpha\cdot L_1\big(f_s^2,f_t^2\big)+H(f_s^2)\Big]\end{aligned}\tag{11}$$

Combining equation (10) and equation (11), we have

$$\mathbb{E}\Big[L_1(f_s^2,f_t^2)\Big]\leq\mathbb{E}\Big[L_1(f_s^1,f_t^1)\Big].\tag{12}$$

□

B ANALYSING KNOWLEDGE DISTILLATION WITH VC THEORY

Recent evidences show that directly applying the naive Hinton *et al.* knowledge distillation (Hinton et al., 2014; Zhang et al., 2022; Li et al., 2020c) to image-to-image translation usually leads to limited and even negative performance. In this subsection, we try to explain this observation from the perspective of VC theory based on generalized knowledge distillation (Lopez-Paz et al., 2016). Denoting a function class as $\mathcal{F}$, then the student function, the teacher function and the oracle real target function can be written as $f_s\in\mathcal{F}_s$, $f_t\in\mathcal{F}_t$, and $f\in\mathcal{F}$, respectively. Given n training samples, we can assume that the student function f_s and the teacher function f_s may learn the true function f at a rate of α_s and α_t, which can be formulated as

$$R(f_s)-R(f)\leq O(\frac{|\mathcal{F}_s|_C}{n^{\alpha_s}})+\varepsilon_s,\quad\text{and}\quad R(f_t)-R(f)\leq O(\frac{|\mathcal{F}_t|_C}{n^{\alpha_t}})+\varepsilon_t\quad\text{respectively,}\tag{13}$$

where $O(\cdot)$ term is the estimation error, ε_s and ε_t are the approximation error of the student function class $\mathcal{F}_s$ and the teacher function class $\mathcal{F}_t$ with respect to $f\in\mathcal{F}$. A higher α indicates the learning problem is easier to be solved. Then, we can assume that the student learns from the teacher at the rate α_{kd} with the approximation error ε_{kd}, which can be formulated as

$$R(f_s)-R(f_t)\leq O(\frac{|\mathcal{F}_s|_C}{n^{\alpha_{kd}}})+\varepsilon_{kd}.\tag{14}$$

As pointed out by Lopez-Paz *et al.* (Lopez-Paz et al., 2016), since the teacher model has more parameters than the student, we can assume the teacher function can learn the true function with a higher rate, indicating $\alpha_t>\alpha_s$ and $\alpha_t>\alpha_{kd}$. By combining (13) and (14), we have the following inequality.

$$\begin{aligned}R(f_s)-R(f)&=R(f_s)-R(f_t)+R(f_t)-R(f_s)\\&\leq O(\frac{|\mathcal{F}_s|_C}{n^{\alpha_{kd}}})+\varepsilon_{kd}+O(\frac{|\mathcal{F}_t|_C}{n^{\alpha_t}})+\varepsilon_t\\&\leq O(\frac{|\mathcal{F}_s|_C+|\mathcal{F}_t|_C}{n^{\alpha_{kd}}})+\varepsilon_{kd}+\varepsilon_t\end{aligned}\tag{15}$$

Thus, given a learning task, now we can study whether knowledge distillation works well in this task by analyzing whether the following inequality

$$O(\frac{|\mathcal{F}_s|_C+|\mathcal{F}_t|_C}{n^{\alpha_{kd}}})+\varepsilon_{kd}+\varepsilon_t\leq O(\frac{|\mathcal{F}_s|_C}{n^{\alpha_s}})+\varepsilon_s\tag{16}$$

holds. Since the teacher model usually have more parameters than the student model, $|\mathcal{F}_s|_C+|\mathcal{F}_t|_C\leq|\mathcal{F}_s|_C|$ usually does not hold in knowledge distillation. Thus, the inequality highlights that the benefits of knowledge distillation arise because of $\varepsilon_{kd}+\varepsilon_t\leq\varepsilon_s$ and $\alpha_{kd}>\alpha_s$.

In image classification, as pointed out by Lopez-Paz *et al.* (Lopez-Paz et al., 2016), since soft labels $f_t(x)$ (the probability distribution) of teachers contain more information than the one-hot label y, it allows students to learn teachers at a higher rate than learning the true function, indicating that $\alpha_{kd} > \alpha_s$ (Lopez-Paz et al., 2016). Besides, since the label for an input image is unique, learning the true function does not conflict with learning the teacher function, and thus it is safe to assume that $\varepsilon_s \geq \varepsilon_t + \varepsilon_{kd}$. In contrast, on image-to-image translation, since the prediction of students and teachers are values of pixels instead of the probability distribution, there is no additional information in $f_t(x)$ compared with the ground truth. Thus $\alpha_{kd} > \alpha_s$ does not hold. Moreover, since image-to-image translation is an ill-posed problem, the prediction of students and teachers may be different but correct answers for the same input image, indicating that $\varepsilon_s \geq \varepsilon_t + \varepsilon_{kd}$ also does not hold. These observations demonstrate that the inequality (16) does not hold in image-to-image translation, which can explain the limited performance of directly applying Hinton *et al.* knowledge distillation to image-to-image translation.

Instead of distilling the generated images, some recent knowledge distillation methods have been proposed to distill teacher knowledge in their features. Since there is more information contained in teacher features than ground-truth images, these methods can be considered as a guarantee for $\alpha_{kd} > \alpha_s$. In contrast, **`IYOR`** aims to improve knowledge distillation by addressing the ill-posed property, implying $\varepsilon_s \geq \varepsilon_t + \varepsilon_{kd}$. Since **`IYOR`** and previous feature-based methods have different perspectives to support inequality (16), their benefits are orthogonal and can be combined.

C Detailed Experiment Settings

We follow the official codes of CycleGAN and Pix2Pix[1] to conduct our experiments. Models on Edge→Shoe are trained by 50 epochs. Models on the other datasets are trained by 200 epochs. The momentum of Adam optimizer is 0.5. During the all the experiments, we set $\alpha = 1$ and $\beta = 1$. The initial learning rate is 0.0002. LSGAN is used as the generator of the model. The discriminator is a 70x70 PatchGAN. In the experiments of CycleGAN, the backbone in the generators of both students and teachers (refining networks) are ResNet with six blocks. Their main difference is that the student backbone has much less channels than the teacher. Batch size is set to 1 for both training and inference. We compute the FID scores based on Pytorch-FID [2], a well known python package. We find that some of previous works compute the FID for unpaired image-to-image translation by using the images in both training set and test set to achieve more stable performance. However, we believe this behavior that access the test images during training is not reasonable. Hence, we choose to compute the FID on only the test set. As claimed by previous research Jin et al. (2021)[3], this makes the FID scores in our experiments around 5-6 lower than the previous works which report FID on the both training and test set.

D Influence from Hyper-parameters

In this paper, we mainly have two hyper-parameters α and β to balance the magnitudes of knowledge distillation loss and the original GAN training loss. Hyper-parameters sensitivity study on Horse→Zebra with 15.81× compressed students is introduced in Figure 6. Note that the reported value is FID (lower is better). It is observed that: (i) With the worst α, our method achieve 70.21 FID, which is still 14.83 lower than the student trained without KD, and 6.83 lower than the second-best KD method. (ii) With the worst β, our method achieve 70.54 FID, which is still 14.50 lower than the student trained without KD, and 6.50 lower than the second-best KD method. These observations indicate that our method is not sensitive to the value of hyper-parameters.

E Influence from the Size of the Refining Network

In our experiments, the refining network has the same architecture as the teacher network in traditional KD, which is also same as the image-to-image translation model before compression. In

[1] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix/

[2] https://github.com/mseitzer/pytorch-fid

[3] https://github.com/snap-research/CAT

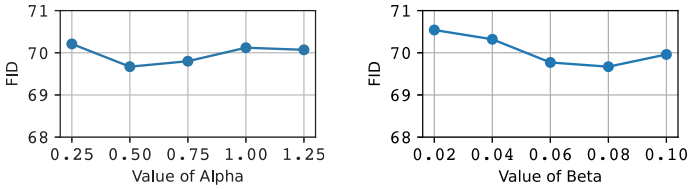

Figure 6: Hyper-parameter sensitivity study on 15.81× compressed CycleGAN.

Table 5: Influence from the number of parameters and FLOPs in the refining network. Note that the student is a 15.81× compressed CycleGAN.

#Params of Refining Network	FLOPs	Refining FID	Student FID
11.37	49.64	56.31	69.67
2.84	12.41	62.59	70.56
1.61	7.29	63.41	73.28

Table 6: Experimental results on Cityscapes with Pix2Pix. Δ indicates the performance improvements compared with the origin student. Each experiment is averaged from 8 trials. "Origin Student" indicates the student trained without knowledge distillation. **A higher mIoU is better**.

Model	#Params (M)	FLOPs (G)	Method	Metric	
				mIoU↑	Δ ↑
Pix2Pix	54.41	96.97	Teacher without KD	$46.51_{\pm 0.32}$	–
	13.61 $_{4.00\times}$	24.90 $_{3.88\times}$	Origin Student without KD	$41.35_{\pm 0.22}$	–
			Hinton et al. (2014)	$40.49_{\pm 0.41}$	-0.86
			Zagoruyko & Komodakis (2017)	$40.17_{\pm 0.36}$	-1.18
			Li et al. (2020a)	$41.52_{\pm 0.34}$	0.17
			Li et al. (2020c)	$41.77_{\pm 0.30}$	0.42
			Jin et al. (2021)	$41.29_{\pm 0.51}$	-0.06
			Ahn et al. (2019)	$41.88_{\pm 0.45}$	0.53
			Ren et al. (2021)	42.31 ± 0.31	0.96
			Li et al. (2021b)	41.75 ± 0.42	0.40
			Zhang et al. (2022)	$42.81_{\pm 0.25}$	1.46
			Ours	$43.52_{\pm 0.41}$	0.71

this section, we study the influence from the size of the refining network. As shown in Table 5: (i) With more parameters, the refining network can achieve a very low FID, which indicates that the refinement has good quality. And at the same time, the student can also be trained better, which achieve relative lower FID. (ii) When the refining network does not have enough parameters, the refinement has a relative higher FID and the effectiveness of knowledge distillation is not very significant. These observations indicate that a refining network with enough parameters can make a positive influence to the performance of knowledge distillation. In contrast, when the refining network does not have enough parameters, it can not successfully refine the image generated by the student, which leads to limited knowledge distillation performance.

F Experiments on Cityscapes

Following previous research Zhu et al. (2017a); Park et al. (2019a), we have also evaluated our method on Cityscapes (Cordts et al., 2016). Cityscapes is original proposed as a dataset for autonomous driving, including tasks such as detection and segmentation. In our experiments, we take the semantic segmentation mask as the input and take the natural images of the street as the label to train the image-to-image translation models. Then, we adopt the mIoU of a pre-trained FCN model on the generated images as the performance metric. A higher mIoU indicates that the image-to-image translation model has better performance. Our experimental results are shown in Table 6. It

is observed that there are 2.17 mIoU improvements on the student trained with our method, which is 0.71 higher than the second-best method.

G PYTHON-STYLE PSEUDO CODE

The following code block presents a brief implementation of **IYOR**.

```
# The pseudo code of IYOR
def IYOR(x, student, refiner, sift):
    # x: the input image, student: the student network
    # refiner: the refining network
    # sift: a function to extract sift features
    student_output = student(x)
    refinement = refine(student_output.detach())
    # pixel-wise imitating
    kd_loss = l1_loss(student_output, refinement)
    # sift distillation
    kd_loss += l1_loss(sift(student_output), sift(refinement))
    return kd_loss
```