Region-aware Knowledge Distillation for Efficient Image-to-Image Translation

Linfeng Zhang[1] Runpei Dong[2]
Xin Chen[3] Kaisheng Ma[1,4]

[1] Tsinghua Univeristy
[2] Xi'an Jiaotong University
[3] Intel Crop.

Abstract

Recent progress in image-to-image translation has witnessed the success of generative adversarial networks (GANs). However, GANs usually contain a huge number of parameters, which lead to intolerant memory and computation consumption and limit their deployment on edge devices. To address this issue, knowledge distillation is proposed to transfer the knowledge from a cumbersome teacher model to an efficient student model. However, most previous knowledge distillation methods are designed for image classification and lead to limited performance in image-to-image translation. In this paper, we propose Region-aware Knowledge Distillation (ReKo) to compress image-to-image translation models. Firstly, ReKo adaptively finds the crucial regions in the images with an attention module. Then, patch-wise contrastive learning is adopted to maximize the mutual information between students and teachers in these crucial regions. Experiments with nine comparison methods on nine datasets demonstrate the substantial effectiveness of ReKo on both paired and unpaired image-to-image translation. For instance, our 7.08× compressed and 6.80× accelerated CycleGAN student outperforms its teacher by 1.33 and 1.04 FID scores on Horse→Zebra and Zebra→Horse, respectively.

1 Introduction

Tremendous breakthroughs have been attained with the state-of-the-art generative adversarial network (GAN) in generating high-resolution, high-fidelity, and photo-realistic images and videos [4, 9, 12, 27, 38, 40]. Because of its powerful ability of representation and generation, GAN has evolved to the most dominant model in image-to-image translation [2, 5, 6, 7, 14, 31]. However, the advanced performance of GAN is always accompanied by tremendous parameters and computation, which have limited their usage in resource-limited edge devices such as mobile phones.

To address this issue, knowledge distillation is proposed to improve the performance of an efficient student model by mimicking the features and prediction of a cumbersome teacher model. Following previous research on image classification [26, 30], detection [3, 17, 32, 36], and segmentation [19], some recent works have tried to directly apply knowledge distillation to I2IT but earned very limited improvements [15, 18]. In this paper, we first argue that

[4] Corresponding author.

Figure 1: Visualization on the attention results on the Zebra→Horse translation: Attention module in ReKo can localize the to-be-translated objects (zebras) unsupervisedly.

most previous knowledge distillation methods ignore the *spatial redundancy* in I2IT, which results in their failure. More specifically, in I2IT, usually, only a few regions of the images are actually required to be translated. For example, in the Horse→Zebra task, only the regions of horses need to be translated while the regions of background should be preserved. Even in some tasks where the whole image is required to be translated, there are some relatively more crucial regions. However, previous knowledge distillation methods directly employ the student to mimic teacher features in all the regions with the same priority while ignoring the spatial redundancy, which further leads to their unsatisfactory performance.

Since students have much fewer parameters than their teachers, usually they are not able to learn all the knowledge from teachers. Thus, in knowledge distillation, the student should pay more attention to teacher knowledge in the crucial regions instead of learning all the regions with the same priority. Unfortunately, different from the other vision tasks such as object detection, there are no annotations on crucial regions in I2IT, especially unpaired I2IT. Thus, it is still challenging to localize and make good use of these crucial regions. To tackle this challenge, in this paper, we propose a novel knowledge distillation method referred to as Region-aware Knowledge Distillation (ReKo), which contains the following two steps.

Firstly, ReKo localizes the crucial regions in an image with a parameter-free attention module and then only distills teacher features in these crucial regions. As discussed in previous works [35, 36, 39], the attention value in a region shows its importance. The region with a higher attention value usually has more influence on the prediction of the neural network and thus should be considered as a more important region. Hence, we define the importance of a region as its attention value, which is further utilized to decide whether teacher features in this region should be distilled to the student. Visualization results of this attention module have shown in Figure 1. It is observed that this method successfully localizes the regions of horses while filtering the regions of background.

Secondly, ReKo adopts a patch-wise contrastive learning framework to optimize knowledge distillation. Instead of distilling teacher knowledge to students by directly minimizing the L_2-norm distance between their features, we propose to adopt a contrastive learning framework for optimization. Tian *et al.* firstly show that on image classification, knowledge distillation can be performed with contrastive learning to maximize the mutual information between students and teachers [29]. However, their method requires a huge memory bank to contain massive negative samples[1], which is not applicable for I2IT. To address this issue,

[1] 16384 negative samples are required according to their released codes.

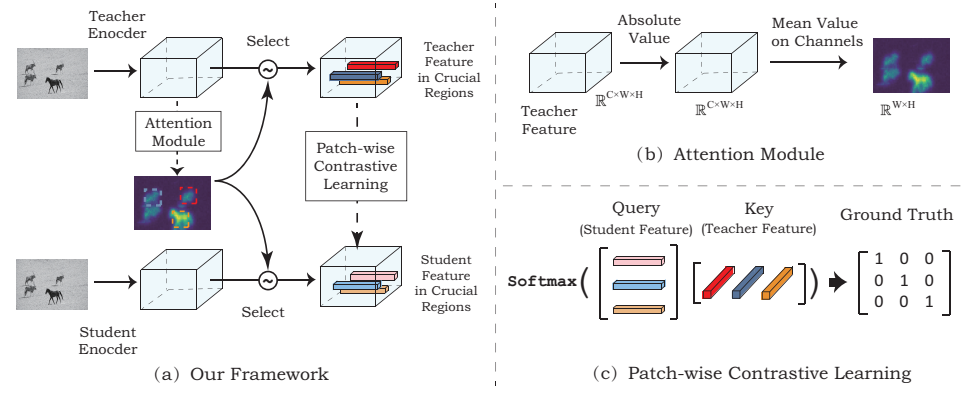

Figure 2: The overview of ReKo (best viewed in color). **Step-1**: Find the crucial regions in the to be translated image by applying the `attention module` to teacher features. Then, K regions with the K largest attention values are selected as the crucial regions (here K=3). **Step-2**: Perform knowledge distillation in these crucial regions with patch-wise contrastive learning. Student features and teacher features in the same region (such as ■ and ■) are considered as a positive pair and the others (such as ■ and ■) are regarded as negative pairs. All these pairs are optimized in a contrastive learning framework with InfoNCE loss, which regards the student features as queries and teacher features as the keys.

we propose to apply patch-wise contrastive learning framework [23] for knowledge distillation, which regards student and teacher features in the same patch as a positive pair and the other features as negative pairs. During the distillation period, by optimizing these pairs with InfoNCE loss [22], the similarity between student and teacher features in the same region is improved, and thus teacher knowledge is distilled to the student.

Extensive experiments on nine datasets with nine comparison methods have been conducted to demonstrate the effectiveness of ReKo both quantitatively and qualitatively. The visualization results between students and teachers show that ReKo has successfully increased the feature similarity between students and teachers in the same region, which indicates knowledge has been effectively distilled. Moreover, we show that ReKo is able to stabilize the training of GANs and thus prevent them from model collapse.

2 Methodology

2.1 Preliminaries

Given two sets of images $\mathcal{X}$ and $\mathcal{Y}$, I2IT aims to find a mapping function $\mathcal{F}$ which maps the images from $\mathcal{X}$ to $\mathcal{Y}$. Usually, $\mathcal{F}$ can be divided into an encoder $\mathcal{E}$ to encode the intermediate features followed by a decoder $\mathcal{D}$ which decodes the intermediate features into the images. Given an image x, then its intermediate feature can be formulated as $\mathcal{E}(x) \in \mathbb{R}^{W\times H\times C}$ where C, W and H denote the number of channels, width, and height, respectively. For convenience, we reshape $\mathcal{E}(x)$ into $\mathbb{R}^{WH\times C}$, where $\mathcal{E}(x)[i]$ indicates the feature of i-th region. Then, the corresponding index set of regions can be formulated as $\mathcal{I} = \{1, 2, 3, ..., WH\}$.

2.2 Patch-wise Contrastive Learning for KD on I2IT

In this paper, we adopt a noise contrastive estimation framework [22] to maximize the mutual information between the features of students and teachers. Given a query v, a positive key v^+ and a set of negative keys $\{v_1^-, v_2^-, ..., v_N^-\}$. The InfoNCE loss can be formulated as

$$L_{\text{InfoNCE}}(v, v^+, v^-) = -\log\left[\frac{\exp(v \cdot v^+/\tau)}{\exp(v \cdot v^+/\tau) + \sum_{n=1}^{N} \exp(v \cdot v_i^-/\tau)}\right], \tag{1}$$

where τ is a temperature hyper-parameter. By regarding the features of students and teachers at the same region (patch) as positive pairs and the other features as the negative pairs, we can extend InforNCE to patch-wise contrastive distillation framework. By distinguishing the student model and the teacher model with the scripts $\mathcal{S}$ and $\mathcal{T}$, respectively, its loss function can be formulated as $L_{\text{RegionDistill}} = \mathbb{E}_{x\sim\mathcal{X}} \sum_{i\in\mathcal{I}} L_{\text{InfoNCE}}(v, v^+, v^-)$, where student feature at i-th patch $v = \mathcal{E}^{\mathcal{S}}(x)[i]$ is the query, the teacher feature at i-th patch $v^+ = \mathcal{E}^{\mathcal{T}}(x)[i]$ is the positive key, and the teacher features at the patches except i-th patch $v^- = \{\mathcal{E}^{\mathcal{T}}[j] \mid j \in \mathcal{I}, j \neq i\}$ is the set of negative keys. During knowledge distillation, the similarity between the query and the positive key (student and teacher features in the same region) is maximized and the similarity between the query and the negative queries (student and teacher feature in different regions) are minimized. Thus, the knowledge in teacher features can be distilled to the student.

Parameter-free Attention Module It is generally acknowledged that the attention value of each pixel shows its importance [36]. In this paper, we define the attention value of a region as its absolute mean value across the channel dimension, which can be formulated as $\mathcal{A} : \mathbb{R}^{c\times wh} \xrightarrow{\text{absolute}} \mathbb{R}^{c\times wh} \xrightarrow{\text{mean on channel}} \mathbb{R}^{wh}$.Note that since the absolute operation and the mean computation operation do not have any trainable parameters, indicating that it can be directly applied to any neural network to find the desired importance score.

Knowledge Distillation on Crucial Regions Given a teacher feature, $\mathcal{E}^{\mathcal{T}}(x)$, its attention map can be denoted as $\mathcal{A}(\mathcal{F}_{\text{enc}}^{\mathcal{T}})(x)$. Then, we select K regions with the K largest attention values as the crucial regions in this image. Denote the index set of regions as $\mathcal{I}_K$, the proposed region-aware knowledge distillation can be formulated as

$$L_{\text{ReKo}} = \mathbb{E}_{x\sim\mathcal{X}} \sum_{i\in\mathcal{I}_K} L_{\text{InfoNCE}}(v, v^+, v^-), \tag{2}$$

where $v = \mathcal{E}^{\mathcal{S}}(x)[i]$ is the query, i-th patch $v^+ = \mathcal{E}^{\mathcal{T}}(x)[i]$ is the positive key, and $v^- = \{\mathcal{E}^{\mathcal{T}}[j] \mid j \in \mathcal{I}, j \neq i\}$ is the set of negative keys. It is observed that the main difference between $L_{\text{RegionDis}}$ and L_{ReKo} is that L_{ReKo} applies knowledge distillation only to the K crucial regions found by $\mathcal{A}$ instead of all the regions. Based on the above formulation, we can introduce the overall training loss of students as

$$L_{\text{Student}} = \alpha \cdot L_{\text{ReKo}} + L_{\text{Origin}}, \tag{3}$$

where L_{Origin} is the original training loss of I2IT models. For instance, in Pix2Pix, L_{Origin} indicates the adversarial learning loss and the mean square loss between the ground truth and the generated images. We do not introduce it here in detail since it has no direct influence with our method. α is a hyper-parameter to balance the two loss functions.

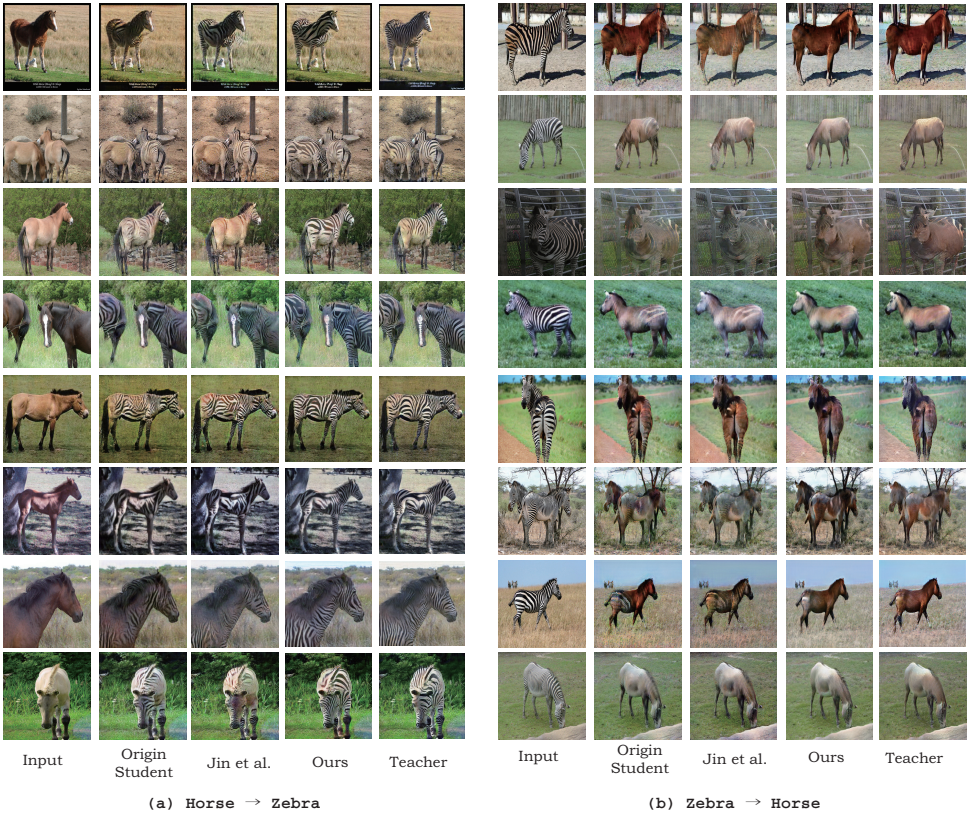

Figure 3: Qualitative results on Horse→Zebra and Zebra→Horse. The students are 15.81× compressed CycleGAN.

3 Experiments

3.1 Experiment Setting

We evaluate ReKo on three models including CycleGAN [40] for unpaired I2IT, Pix2Pix [12] and Pix2PixHD [34] for paired I2IT. Three datasets including Horse↔Zebra, Edge→Shoe and Cityscapes are utilized for quantitative evaluation. Horse↔Zebra is an unpaired I2IT dataset that translates images of horses to zebras and vice versa. Edge→Shoe is a paired I2IM dataset which maps the edges of shoes to their natural images. Cityscapes is a dataset that translates the segmentation result to its origin image [8]. Besides, we also conduct qualitative experiments on Facades, Maps, Summer↔Winter, Apple↔Orange, Photo↔Monet and Photo↔Vangogh. We build the student models in our experiments by reducing the number of channels from 64 to 48, 32, and 16. On Cityscapes, we evaluate the quality of generated images with the mIoU score of a pre-trained segmentation model. On the other datasets, *Fréchet Inception Distance (FID)*, which measures the distance between the distribution of features extracted from the real and the synthetic images, is utilized as the metric for all the experiments. A lower FID and a higher mIoU indicate that the synthetic images have better quality. Moreover, to obtain more reliable results, we run 8 trials for each experiment and report their average and standard deviation.

Table 1: Experimental results on unpaired I2IT 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 from 8 trials.

Horse→Zebra					Zebra→Horse				
#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	$138.07_{\pm 4.01}$	–
0.72 15.81×	3.35 14.82×	Origin Student	$85.04_{\pm 6.88}$	–	0.72 15.81×	3.35 14.82×	Origin Student	$152.67_{\pm 9.63}$	–
		Hinton *et al.* [10]	$84.08_{\pm 3.78}$	0.96			Hinton *et al.* [10]	$148.64_{\pm 1.62}$	4.03
		Zagoruyko *et al.* [36]	$81.24_{\pm 2.01}$	3.80			Zagoruyko *et al.* [36]	$148.92_{\pm 1.20}$	3.75
		Li and Lin *et al.* [15]	$83.97_{\pm 5.01}$	1.07			Li and Lin *et al.* [15]	$151.32_{\pm 2.31}$	1.35
		Li and Jiang *et al.* [13]	$81.74_{\pm 4.65}$	3.30			Li and Jiang *et al.* [13]	$151.09_{\pm 3.67}$	1.58
		Jin *et al.* [12]	$82.37_{\pm 8.56}$	2.67			Jin *et al.* [12]	$149.73_{\pm 3.94}$	2.94
		Ahn *et al.* [1]	$82.91_{\pm 2.41}$	2.13			Ahn *et al.* [1]	$150.31_{\pm 3.55}$	2.36
		Ren *et al.* [24]	$77.31_{\pm 6.41}$	7.73			Ren *et al.* [24]	$147.34_{\pm 2.98}$	5.23
		Li *et al.* [14]	$79.29_{\pm 7.31}$	5.75			Li *et al.* [14]	$148.30_{\pm 1.53}$	4.27
		Zhang *et al.* [34]	$77.04_{\pm 3.52}$	8.00			Zhang *et al.* [34]	$146.01_{\pm 1.80}$	6.66
		ReKo (Ours)	$\mathbf{71.21}_{\pm 6.17}$	**13.83**			**ReKo (Ours)**	$\mathbf{142.58}_{\pm 4.27}$	**10.09**
1.61 7.08×	7.29 6.80×	Origin Student	$70.54_{\pm 9.63}$	–	1.61 7.08×	7.29 6.80×	Origin Student	$141.86_{\pm 1.57}$	–
		Hinton *et al.* [10]	$70.35_{\pm 3.27}$	0.19			Hinton *et al.* [10]	$142.03_{\pm 1.61}$	-0.17
		Zagoruyko *et al.* [36]	$67.51_{\pm 4.57}$	3.03			Zagoruyko *et al.* [36]	$141.23_{\pm 1.27}$	0.63
		Li and Lin *et al.* [15]	$68.58_{\pm 4.31}$	1.96			Li and Lin *et al.* [15]	$141.32_{\pm 1.27}$	0.54
		Li and Jiang *et al.* [13]	$68.94_{\pm 2.98}$	1.60			Li and Jiang *et al.* [13]	$151.09_{\pm 3.67}$	1.58
		Jin *et al.* [12]	$67.31_{\pm 3.01}$	3.23			Jin *et al.* [12]	$140.98_{\pm 1.41}$	0.88
		Ahn *et al.* [1]	$69.32_{\pm 5.89}$	1.22			Ahn *et al.* [1]	$141.50_{\pm 2.51}$	0.36
		Ren *et al.* [24]	$64.78_{\pm 5.21}$	5.76			Ren *et al.* [24]	$140.87_{\pm 2.03}$	0.99
		Li *et al.* [14]	$66.85_{\pm 6.17}$	3.69			Li *et al.* [14]	$140.92_{\pm 2.31}$	0.94
		Zhang *et al.* [34]	$61.65_{\pm 4.73}$	8.89			Zhang *et al.* [34]	$138.84_{\pm 1.47}$	3.02
		ReKo (Ours)	$\mathbf{60.01}_{\pm 5.22}$	**10.53**			**ReKo (Ours)**	$\mathbf{137.03}_{\pm 3.03}$	**4.83**

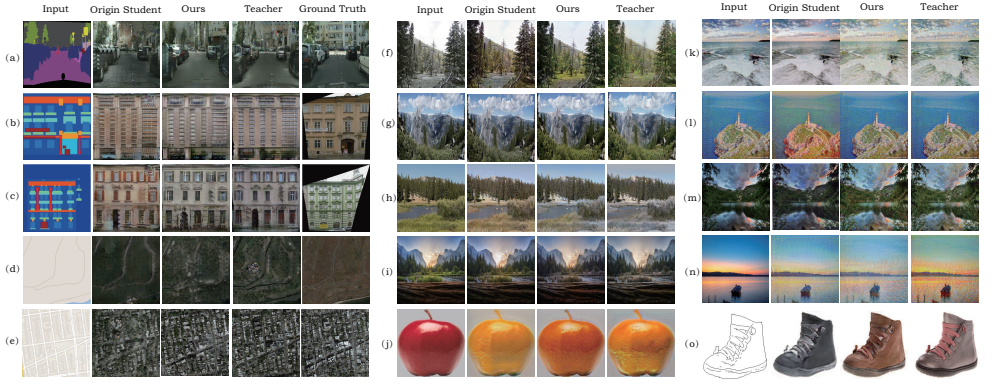

Figure 4: Qualitative experiments on the other datasets: Cityscapes (a), Facades (b-c), Maps→Aerial Photos (d-e), Edge→Shoe (f-i) with Pix2Pix for paired I2IT and Winter →Summer (j-k), Summer→Winter (i-m), Apple→Orange (n), Photo→Monet (o-p), Photo→Vangogh (q-r) for unpaired I2IT. Pix2Pix students on Cityscapes, Edge→Shoe, and the other datasets are 4.00×, 4.00× and 28.32× compressed, respectively. CycleGAN students are 15.81× compressed.

3.2 Experimental Result

Quantitative Result Quantitative experimental results of ReKo and the other nine knowledge distillation methods on Horse→Zebra, Edge→Shoe and Cityscapes are shown in Table 1, Table 2 and Table 3, respectively. Besides, quantitative results of our method with the students pruned with the methods from Li *et al.* [15] and Jin *et al.* [13] are shown in Table 4.

Table 2: Experimental results on paired I2IT on Edge→Shoe with Pix2Pix and Pix2PixHD. A lower FID is better performance. Δ indicates the performance improvements compared with the origin student trained without KD. Each result is averaged from 8 trials.

Pix2PixHD

#Params (M)	FLOPs (G)	Method	Metric	
			FID↓	Δ↑
45.59	48.36	Teacher	$41.59_{\pm 0.42}$	–
1.61 28.32×	1.89 25.59×	Origin Student	$44.64_{\pm 0.54}$	–
		Hinton *et al.* [10]	$45.31_{\pm 0.63}$	-0.67
		Zagoruyko *et al.* [5]	$44.21_{\pm 0.72}$	0.43
		Li and Lin *et al.* [3]	$44.03_{\pm 0.41}$	0.61
		Li and Jiang *et al.* [4]	$43.90_{\pm 0.36}$	0.74
		Jin *et al.* [7]	$43.97_{\pm 0.17}$	0.67
		Ahn *et al.* [1]	$44.53_{\pm 0.48}$	0.11
		Ren *et al.* [24]	$42.98_{\pm 0.34}$	1.66
		Li *et al.* [18]	$43.21_{\pm 0.35}$	0.29
		Zhang *et al.* [31]	$42.53_{\pm 0.29}$	2.11
		ReKo (Ours)	**$42.31_{\pm 0.17}$**	**2.33**
		ReKo + Ren*et al.* [24]	$41.25_{\pm 0.54}$	3.39
		ReKo + Li*et al.* [18]	$41.88_{\pm 0.53}$	2.76

Pix2Pix

#Params (M)	FLOPs (G)	Method	Metric	
			FID↓	Δ↑
54.41	6.06	Teacher	$59.70_{\pm 0.91}$	–
13.61 4.00×	1.56 3.88×	Origin Student	$85.06_{\pm 0.98}$	–
		Hinton *et al.* [10]	$86.97_{\pm 3.49}$	-1.91
		Zagoruyko *et al.* [5]	$84.25_{\pm 2.08}$	0.81
		Li and Lin *et al.* [3]	$83.63_{\pm 3.12}$	1.43
		Li and Jiang *et al.* [4]	$84.01_{\pm 2.31}$	1.05
		Jin *et al.* [7]	$84.39_{\pm 3.62}$	0.67
		Ahn *et al.* [1]	$84.92_{\pm 0.78}$	0.14
		Ren *et al.* [24]	$80.31_{\pm 2.59}$	4.75
		Li *et al.* [18]	$81.24_{\pm 3.74}$	3.82
		Zhang *et al.* [31]	$80.13_{\pm 2.18}$	4.93
		ReKo (Ours)	**$77.69_{\pm 3.14}$**	**7.37**
		ReKo + Ren *et al.* [24]	$74.24_{\pm 2.48}$	10.85
		ReKo + Li *et al.* [18]	$75.21_{\pm 3.15}$	9.85

We mainly have the following observations: (i) ReKo leads to consistent and significant performance improvements (FID reduction) on all kinds of datasets and models. On average, it leads to 9.2 and 4.85 FID reduction on unpaired and paired I2IT tasks, respectively. (ii) ReKo outperforms the other eight kinds of I2IT knowledge distillation methods by a large margin. For instance, on Horse→Zebra, it outperforms the second-best method by 3.7 FID, on average. (iii) Not all the knowledge distillation methods work well on GAN for I2IT. Directly applying the naïve Hinton knowledge distillation [10] leads to limited and sometimes even negative effects. For instance, it leads to 1.91 FID increment (performance drop) on the Pix2Pix student for Edge→Shoe. (iv) Compared with paired I2IT, there are more performance improvements on unpaired I2IT with ReKo. This observation may be caused by the fact that there is less labeled supervision in unpaired I2IT. Thus the knowledge from teachers is more helpful. (v) A high ratio of acceleration and compression can be achieved by replacing the teacher model with the distilled student model. For example, ReKo leads to 7.08× compression and 6.80× acceleration on CycleGAN students. Besides, the compressed students outperform their teachers by 1.33 and 1.04 FID on the tasks of Horse→Zebra and Zebra→Horse, respectively. (vi) As shown in Table 2, our method can also be utilized with previous methods together to achieve better performance. For instance, 6.07 and 6.03 FID reduction can be observed on Edge→Shoe with Pix2Pix by combining the methods of Ren *et al.* and Li *et al.*, respectively. (vii) As shown in Table 4, significant performance gains can also be obtained on pruned models, indicating that our method can be utilized with the other model compression methods for better performance.

Table 3: Experimental results on Cityscapes with Pix2Pix. Δ indicates the performance boosts compared with the origin student. Each result is averaged from 8 trials. **A higher mIoU is better**.

#Params (M)	FLOPs (G)	Method	Metric	
			mIoU↑	Δ↑
54.41	96.97	Teacher	$46.51_{\pm 0.32}$	–
13.61 4.00×	24.90 3.88×	Origin Student	$41.35_{\pm 0.22}$	–
		Hinton *et al.* [10]	$40.49_{\pm 0.41}$	-0.86
		Zagoruyko *et al.* [5]	$40.17_{\pm 0.36}$	-1.18
		Li and Lin *et al.* [3]	$41.52_{\pm 0.34}$	0.17
		Li and Jiang *et al.* [4]	$41.77_{\pm 0.30}$	0.42
		Jin *et al.* [7]	$41.29_{\pm 0.51}$	-0.06
		Ahn *et al.* [1]	$41.88_{\pm 0.45}$	0.53
		Ren *et al.* [24]	$42.31_{\pm 0.31}$	0.96
		Li *et al.* [18]	$41.75_{\pm 0.42}$	0.40
		Zhang *et al.* [31]	$42.93_{\pm 0.25}$	1.58
		ReKo(Ours)	**$43.57_{\pm 0.25}$**	**2.22**

Qualitative Result Qualitative results on Horse →Zebra and the other datasets are shown in Figure 3 and Figure 4, respectively. It is observed that: (i) The student model trained without KD can not always translate the whole body of horses and zebras while the student model

Figure 5: Visualization on the similarity between teacher features and student features in twelve selected patches. On the left six gray-scale sub-figures, the student patches are selected from the body of the horse. On the right six gray-scale sub-figures, the student patches are selected from the background. Each pixel in a gray-scale figure shows the similarity between the teacher feature at this pixel and the student feature in the selected patch. A whiter pixel indicates a higher similarity.

trained with ReKo does not have this issue. Moreover, on Horse→Zebra, the student model trained by ReKo sometimes outperforms its teacher on the effect of removing the stripes in zebras. (ii) As shown in Figure 4, ReKo also leads to consistent and significant image quality improvements on all of them. Specifically, on the tasks which all the image should be translated such as Summer→Winter, Cityscapes and Map→Aerial, ReKo still leads to significant improvements on the generated images, indicating that ReKo is also effective in the task where all the pixels of images should be translated. We suggest that the effect of ReKo in these tasks is caused by the fact that there are still some relatively more important pixels, such as the color of trees and the snow of mountains in the season translation tasks.

4 Discussion

4.1 Visualization on Similarity

As shown in Figure 2(c), ReKo distills teacher knowledge to the student by improving their feature similarity on the same region. To further verify whether KD has been successfully optimized during training, we visualize the feature similarity between a trained student and its teacher in Figure 5. In the gray-scale sub-figures of (a)-(d) and (e)-(h), the student patches are selected from the body of the horse and the background, respectively. Each pixel in a gray-scale figure shows the similarity between the teacher feature at this pixel and the student feature in the selected patch. It is observed that when the student patch is selected from the body of the horse, it has a higher similarity with the teacher features of the pixels of the horse body. Similarly, the student feature of the patches in the background has a higher similarity with the teacher features in the background. This observation indicates that the feature similarity in the same

Table 4: Results on Horse→Zebra with pruned CycleGANs. **A lowe FID is better**.

#Params (M)	FLOPs (G)	Pruned Method	KD Method	Metric	
				FID↓	Δ↓
11.3	56.8	w/o Pruning	w/o KD	61.53	–
0.34	2.67	Li *et al.* [CVPR]	Li *et al.* [CVPR]	71.81	-10.28
			ReKo (Ours)	**62.21**	**-0.68**
0.32	2.55	Jin *et al.* [CVPR]	Jin *et al.* [CVPR]	60.18	1.35
			ReKo (Ours)	**58.23**	**3.30**

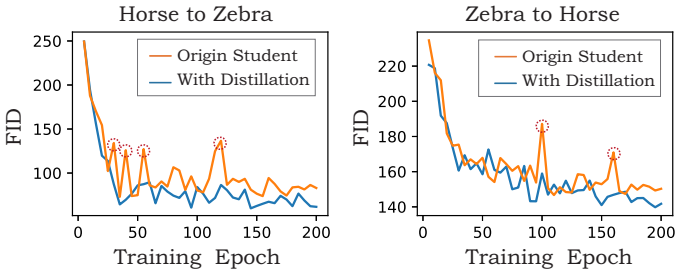

Figure 6: FID curves of two 14.82× compressed CycleGAN students trained with and without KD during the training period.

region is successfully increased by optimizing the KD loss, and thus teacher knowledge has been effectively distilled to the students.

4.2 Ablation Study

Ablation studies on the individual effectiveness of *distilling only the **C**rucial **R**egions* (CR) and *patch-wise **C**ontrastive learning for knowledge **D**istillation* (CD) are shown in Table 5. Note that when CD is disabled but CR is used, we directly minimize the L_2-norm distance between teacher and student features in crucla regions for optimization. It is observed that 5.01 and 3.23 FID reduction can be obtained by using only CD and CR, respectively. Besides, combining the two methods further leads to a 5.52 FID reduction. These observations indicate that both modules are effective, and their merits are orthogonal.

Table 5: Ablation studies on CycleGAN for Horse→Zebra. **CR**: **C**rucial **R**egions. **CD**: Patch-wise **C**ontrastive learning for knowledge **D**istillation. A lower FID indicates better performance.

CR	×	×	✓	✓
CD	×	✓	×	✓
FID (↓)	70.54	65.53	67.31	60.01

Ablation study on Distilling the Crucial Regions To further verify that the attention module in ReKo can localize the important regions for knowledge distillation, we have compared the following three schemes: (a) Distilling regions with the K-largest attention (the scheme in ReKo) (b) Distilling the regions with the K-least attention and (opposite to ReKo) (c) Randomly choose K regions for knowledge distillation. Our experiments show that the three schemes achieve 60.01, 72.54, and 65.53 FID on Horse→Zebra with 7.08× compressed CycleGAN students, respectively. It is observed that our scheme (a) and its opposite scheme (c) achieves the best and the worst performance, respectively, indicating that there is a strong positive correlation between the attention value of a region and the benefits from distilling this region.

4.3 KD Stabilizes GAN Training

The training of GAN is usually not stable due to its complex network architectures and loss functions. In this paper, we find that the proposed knowledge distillation can alleviate this problem. Figure 6 shows the FID curves of CycleGAN students in different training epochs on Horse→Zebra and Zebra→Horse. It is observed that: **(a)** Both the training of students with and without knowledge distillation are stable in the first several epochs. **(b)** After the early epochs, the training of the student without knowledge distillation becomes

unstable and sometimes collapses (marked with circles). In contrast, the distilled student is consistently stable during the whole training period. Its FID undulations are much smaller than the student trained without knowledge distillation.

4.4 Find Crucial Regions with other Methods

In ReKo, the attention of the teacher network is utilized to find the crucial regions in the to-be-translated image. In this subsection, we further investigate the performance of ReKo with the following four different schemes: (1)*student attention* - localizing crucial regions with attention of the student; (2)*VGG attention* - localizing crucial regions with the attention from a ImageNet pre-trained VGG model; (3)*VGG Grad-CAM* localizing crucial regions with the Grad-CAM result from a ImageNet pretrained VGG respect to the to-be-transformed object (*e.g.* horses and zebras); (4)*salient detection* - localizing crucial regions with unsupervised salient detection. Experiments with 7.08× compressed CyCleGAN on Horse→Zebra show that our scheme (with teacher attention) and the above four schemes achieve 60.01, 61.46, 62.91, 63.46 and 65.09 FID, respectively, indicating that teacher attention is the most effective metric in ReKo to localize the crucial regions for knowledge distillation and student attention and Grad-CAM are also two effective solutions.

5 Related Work

Generative Adversarial Network (GAN), which is composed of a generator for image generation and a discriminator for discriminating the real and generated images, have become the most popular model in image-to-image translation [9]. Pix2Pix is proposed to apply conditional GAN [20] to image-to-image translation on paired datasets [12]. Then, Pix2PixHD improves the resolution of generated images with multi-scale neural networks and boundary maps [34]. Based on these efforts, Wang *et al.* further propose Vid2Vid to perform video-to-video translation [33]. Recent, diffusion models have achieved significant breakthroughs in image synthesis [11, 21]. Latent diffusion is proposed to accelerate diffusion process by replacing the image space with latent space [25]. DDIM is introduced to sample pre-trained diffusion models with any timesteps [28].

6 Conclusion

This paper proposes region-aware knowledge distillation (ReKo) for the compression of I2IT models. Firstly, instead of distilling the features of all the spatial positions, ReKo is proposed to distill only the features of patches with higher importance, which is defined by a parameter-free attention module. Then, patch-wise contrastive learning is employed for knowledge distillation, which maximizes the mutual information between the features of students and teachers in the same region. The effectiveness of ReKo demonstrates that it is necessary to design specific KD algorithm for I2IM based its property. We hope this work can promote more research on specific knowledge distillation methods for GANs for I2IM.

7 Acknowledgement

This research was partially supported by National Key R&D Program of China (2022YFB 2804103), Key Research and Development Program of Shaanxi (2021ZDLGY01-05), Tsinghua University Dushi Program, National Natural Science Foundation of China (2021 1710187), and Tsinghua University Talent Program.

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.

[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.

[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.

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

[5] Prashanth Chandran, Gaspard Zoss, Paulo F. U. Gotardo, Markus Gross, and Derek Bradley. Adaptive convolutions for structure-aware style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7972–7981. Computer Vision Foundation / IEEE, 2021.

[6] 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.

[7] 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.

[8] 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.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. volume 27, 2014.

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS*, 2014.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[12] 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.

[13] 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.

[14] Dmytro Kotovenko, Matthias Wright, Arthur Heimbrecht, and Björn Ommer. Rethinking style transfer: From pixels to parameterized brushstrokes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12196–12205. Computer Vision Foundation / IEEE, 2021.

[15] 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.

[16] 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. URL https://arxiv.org/abs/2110.14439.

[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*, pages 18–33. Springer, 2020.

[18] 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.

[19] 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.

[20] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[21] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[23] 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.

[24] 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.

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

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

[27] 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.

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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

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

[31] 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.

[32] 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*, pages 4933–4942, 2019.

[33] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

[34] 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.

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

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

[37] 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.

[38] Yang Zhao and Changyou Chen. Unpaired image-to-image translation via latent energy transport. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 16418–16427. Computer Vision Foundation / IEEE, 2021.

[39] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[40] 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.