NOT ALL REGIONS ARE WORTHY TO BE DISTILLED: REGION-AWARE KNOWLEDGE DISTILLATION TOWARDS EFFICIENT IMAGE-TO-IMAGE TRANSLATION

Anonymous authors
Paper under double-blind review

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 learned by a cumbersome teacher model to an efficient student model. However, previous knowledge distillation methods directly train the student to learn teacher knowledge in all the spatial regions of the images but ignore the fact that in image-to-image translation a large number of regions (*e.g.* background regions) should not be translated and teacher features in these regions are not worthy to be distilled. To tackle this challenge, in this paper, we propose *Region-aware Knowledge Distillation* which first localizes the crucial regions in the images with attention mechanism. Then, teacher features in these crucial regions are distilled to students with a region-wise contrastive learning framework. Besides distilling teacher knowledge in features, we further introduce perceptual distillation to distill teacher knowledge in the generated images. Experiments with four comparison methods demonstrate the substantial effectiveness of our method 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 1.36 and 1.16 FID scores on Horse→Zebra and Zebra→Horse, respectively. Codes have been released in the supplementary material and will be released on GitHub soon.

1 INTRODUCTION

Excellent breakthroughs have been attained with state-of-the-art generative adversarial networks (GANs) in generating high-resolution, high-fidelity, and photo-realistic images and videos (Shaham et al., 2019; Brock et al., 2018; Goodfellow et al., 2014; Isola et al., 2017; Zhu et al., 2017). Due to its powerful ability of representation and generation, GANs have evolved to the most dominant model in image-to-image translation. However, the advanced performance of GANs 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 (Romero et al., 2015; Tung & Mori, 2019), some recent works try to directly apply knowledge distillation to image-to-image translation but their improvements are not significant (Li et al., 2020a;c).

In this paper, we argue that the reason leading to failure in previous image-to-image translation knowledge distillation methods is the *spatial redundancy of teacher features.* More specifically, in image-to-image translation, usually only a few regions of the images are required to be translated. For example, in the well-known 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 all the regions in images are required to be translated, there are usually some more crucial regions. However, previous knowledge distillation methods directly employ the student to mimic teacher features in all

the regions while ignoring the fact that not all regions are worthy to be distilled. Since the student has much fewer parameters than their teachers, they are not able to learn all teachers knowledge. As a result, the student should pay more attention to knowledge distillation 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 is no annotations on crucial regions in image-to-image translation, especially unpaired image-to-image translation. Thus, it is still challenging to localize and make good use of these crucial regions.

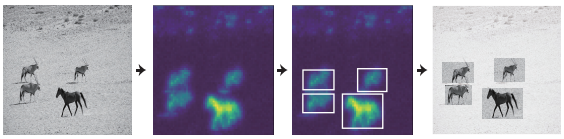


Figure 1: The paradigm of localizing crucial regions in region-aware knowledge distillation on Horse→Zebra with CycleGANs.

To tackle this challenge, in this paper we propose a novel knowledge distillation method referred to as *Region-aware Knowledge Distillation*. Different from previous knowledge distillation methods, the teacher model in our method not only transfers its knowledge in features to students but also tells the student which region should be learned. Concretely, we first propose to localize the crucial regions in images with a parameter-free attention mechanism, where the attention value of a region is decided by its mean absolute value across the channel dimension. As pointed out in previous works (Zhou et al., 2016; Zhang & Kaisheng, 2021; Zagoruyko & Komodakis, 2017), this attention mechanism can reveal the importance of each region with no requirements on additional supervision. Then, we select several the regions with the large attention values as the crucial regions in an image. As shown in Figure 1, in Horse→Zebra, this method can localize the regions of horses while filtering the regions of background. Since no additional labels and parameters are required, it can be easily utilized in all kinds of datasets and models.

After localizing the crucial regions in the image, we then apply a region-wise contrastive learning framework to distill teacher knowledge in crucial regions. Motivated by previous works on contrastive learning and knowledge distillation (Park et al., 2020; Tian et al., 2019), instead of directly minimizing the distance between students and teachers in the feature space, we propose to maximize the mutual information between features of students and teachers in the same crucial region, while pushing away the features of students and teachers in different crucial regions. Concretely, during the training period, the features of students and teachers in the same crucial region are considered as a positive pair and their features in different crucial regions are regarded as negative pairs. Then, by optimizing these pairs with InfoNCE loss (Oord et al., 2018), the distance between positive pairs is minimized while the distance between negative pairs is maximized, which transfers teacher knowledge in the crucial regions to the student.

Besides distilling teacher knowledge in its features, motivated by the well-known perceptual loss utilized in image super-resolution (Johnson et al., 2016), we introduce the perceptual distillation to distill teacher knowledge in the generated images. Instead of directly training the students to mimic the generated images from teachers pixel by pixel, we apply an ImageNet pre-trained model to extract the semantic features of the generated images from students and teachers and then minimize their L_2-norm distance. Compared with the previous pixel-level distillation methods, perceptual distillation is based on differences between high-level image feature representations extracted from the pre-trained models and thus it is more robust and efficient.

In summary, we mainly make the following contributions in this paper.

- We propose *Region-aware Knowledge Distillation* which first localizes the crucial regions in an image depending on the attention values and then distills teacher knowledge in these crucial regions with a region-wise contrastive learning framework.
- We propose perceptual distillation to transfer teacher knowledge in the generated images. Instead of learning the generated images pixel by pixel, the student is trained to generate images which has similar semantic features to images generated from teachers.
- Experiment results with four comparison methods have demonstrated the effectiveness of our method on both paired and unpaired image-to-image translation in terms of both quantitative and qualitative results. For instance, our 7.08× compressed and 6.80× accelerated CycleGAN student outperforms its teacher by 1.36 and 1.16 FID scores on Horse→Zebra and Zebra→Horse, respectively.

2 RELATED WORK

2.1 GANS FOR IMAGE-TO-IMAGE TRANSLATION

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 (Goodfellow et al., 2014). Pix2Pix is proposed to apply conditional GAN (Mirza & Osindero, 2014) to the task of image-to-image translation on paired datasets (Isola et al., 2017). Then, Pix2PixHD improves the resolution of generated images with multi-scale neural networks and boundary maps (Wang et al., 2018b). Based on these efforts, Wang et al. further propose Vid2Vid to perform video-to-video translation (Wang et al., 2018a).

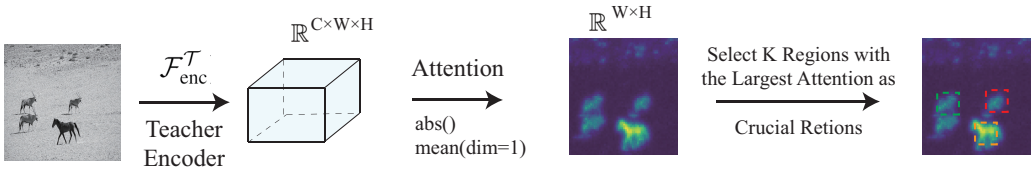
A more challenging task in this domain is how to perform image-to-image translation on unpaired datasets. CycleGAN, DualGAN, and DiscoGAN are proposed to address this issue by regularizing the training of generators with the cycle consistency loss (Zhu et al., 2017; Yi et al., 2017; Kim et al., 2017). Recently, Park et al. propose to replace the cycle consistency loss with a patch-wise contrastive loss, which minimizes the mutual information between the corresponding input and output patches (Park et al., 2020). Besides image style transfer, GANs have also been utilized in the other tasks, such as single image super resolution (Ledig et al., 2017; Wang et al., 2018c), image deblurring (Kupyn et al., 2018) and so on.

The tremendous storage and computation consumption in GAN have promoted the research on its compression. Wang et al. propose a unified GAN compression framework with knowledge distillation, channel pruning, and quantization (Wang et al., 2020). Li et al. propose to compress GANs with once-for-all net architecture search and naive feature knowledge distillation (Li et al., 2020a). Shu et al. propose to investigate and prune the unimportant weights in GANs with a co-evolutionary approach (Shu et al., 2019). Recently, Jin et al. introduce an inception residual block into generators and prune it with a one-step pruning algorithm (Jin et al., 2021).

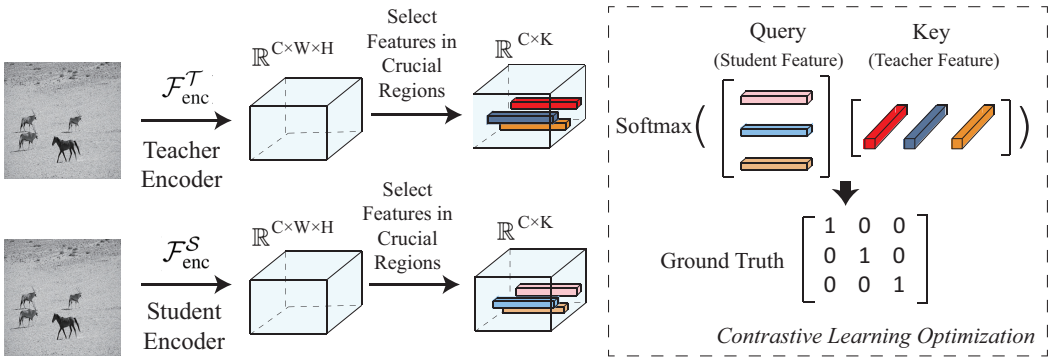
2.2 KNOWLEDGE DISTILLATION

Knowledge distillation has become one of the most effective techniques for model compression (Buciluǎ et al., 2006; Hinton et al., 2014). It first trains a cumbersome teacher model and then transfers its knowledge to a lightweight student model. Previous knowledge distillation usually aims to distill the knowledge in the logits (softmax outputs) (Hinton et al., 2014; Zhang et al., 2018). Then, abundant methods have been proposed to distill the knowledge in the features and its variants, such as attention (Zagoruyko & Komodakis, 2017; Zhang & Kaisheng, 2021) and the gram matrix (Yim et al., 2017). Recently, some research has been proposed to distill the relation between different samples (Park et al., 2019; Tung & Mori, 2019) and pixels (Liu et al., 2020; Li et al., 2020b). Another popular trend in knowledge distillation is to maximize the mutual information between students and teachers with contrastive learning. Tian et al. first propose the contrastive representation distillation framework which regards the representation of the same image from students and teachers as a positive pair in contrastive learning. Then Chen et al. extend this idea with the Wasserstein distance (Chen et al., 2020). In this paper, we extend this framework into the patch-wise contrastive learning (Park et al., 2020) for knowledge distillation on image-to-image translation with GANs.

In the last several years, there has been some research proposed to apply knowledge distillation to the compression of GANs. Li et al. propose to improve the performance of student generators with the naive feature distillation (Li et al., 2020a). Then, Li et al. propose the semantic relation preserving knowledge distillation, which computes and distills the relation between different patches in generators (Li et al., 2020c). Jin et al. propose to distill the knowledge in features with global kernel alignment which enables knowledge distillation without additional adaptation layers (Jin et al., 2021). Recently, Liu et al. propose content-aware GAN compression to compress unconditional GANs, in which the main content such as human faces in an image is first parsed with an additional parser network and then distilled to students. However, the training of the parser network needs additional annotation, which is rare in real-world application.



(a) Find the Crucial Regions in Images with Attention



(b) Region-wise Contrastive Distillation on the Crucial Regions

Figure 2: The overview of Region-aware Knowledge Distillation (best viewed in color). **(a) Step-1**: Find the crucial regions in the to be translated image by applying the attention module to teacher features. Note that the attention module is composed of an absolute value operation and a mean operation on the channel dimension. Then, K regions with the largest attention values are selected as the crucial regions (here K=3). **(b) Step-2**: Based on the crucial regions found in Step-1, select student and teacher features on these crucial regions and discard the features in unimportant regions. Student features and teacher features in the same region are considered as a positive pair (such as ■ and ■) and the others are regarded as negative pairs (such as ■ and ■). All these pairs are optimized in a contrastive learning framework with InfoNCE loss.

3 Methodology

3.1 Region-wise Contrastive Learning for Knowledge Distillation

Given two set of images $\mathcal{X}$ and $\mathcal{Y}$, image-to-image translation aims to find a mapping function $\mathcal{F}: \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^{C\times H\times W}$ which maps images in $\mathcal{X}$ to $\mathcal{Y}$. Note that C, H, W indicates the number of channels, height and width of the image, respectively. Usually, $\mathcal{F}$ can be divided into an encoder $\mathcal{F}_{\text{enc}}$ followed by a decoder $\mathcal{F}_{\text{dec}}$. Given an image x, then its intermediate feature can be formulated as $\mathcal{F}_{\text{enc}}(x) \in \mathbb{R}^{c\times w\times h}$ where c, w and h denotes its number of channels, width and height respectively. For convenience, we reshape it into $\mathbb{R}^{c\times wh}$, where $\mathcal{F}_{\text{enc}}(x)[:, i]$ indicates the feature of i-th region. The corresponding index set of regions can be formulated as $S = \{1, 2, 3, ..., wh\}$.

In this paper, we adopt a noise contrastive estimation framework (Oord et al., 2018) to maximize the mutual information between the features between 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

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

where τ is a temperature hyper-parameter. Regarding the features of students and teachers at the same region as positive pairs and the other features as the negative pair, we can extend InforNCE to

a region-wise contrastive distillation framework, which can be formulated as

$$\mathcal{L}_{\text{RegionDis}}(\mathcal{X}, \mathcal{F}^{\mathcal{S}}_{\text{enc}}, \mathcal{F}^{\mathcal{T}}_{\text{enc}}) = \mathbb{E}_{x\sim\mathcal{X}} \sum_{i=1}^{wh} \mathcal{L}_{\text{InfoNCE}}(\underbrace{\mathcal{F}^{\mathcal{S}}_{\text{enc}}(x)[:,i]}_{\text{Query}}, \underbrace{\mathcal{F}^{\mathcal{T}}_{\text{enc}}(x)[:,i]}_{\text{Positive Key}}, \underbrace{\{\mathcal{F}^{\mathcal{T}}_{\text{enc}}[:,j] \mid j \in S, j \neq i\}}_{\text{Negative Keys}}), \tag{2}$$

where the scripts $\mathcal{S}$ and $\mathcal{T}$ are utilized to distinguish students and teachers.

3.2 Region-aware Knowledge Distillation

It is generally acknowledged that the attention value of each pixel shows its importance (Zhang & Kaisheng, 2021). 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

$$\mathcal{A} : \mathbb{R}^{c\times wh} \xrightarrow{\text{abs()}} \mathbb{R}^{c\times wh} \xrightarrow{\text{mean(dim=1)}} \mathbb{R}^{wh}. \tag{3}$$

Then, given a teacher feature, $\mathcal{F}^{\mathcal{T}}_{\text{enc}}(x)$, its attention map can be denoted as $\mathcal{A}(\mathcal{F}^{\mathcal{T}}_{\text{enc}})(x)$. Then, we select K regions with the largest attention values as the crucial regions in this image. Denote the index set of regions as P_K, then the feature of the crucial regions can be formulated as $\mathcal{G}(x) = \text{stack}(\{\mathcal{F}_{\text{enc}}[:,i]\}, i \in P_k) \in \mathbb{R}^{c\times K}$. Denote its index set as $S' = \{1, 2, ..., K\}$, then our region-aware knowledge distillation can be formulated as

$$\mathcal{L}_{\text{RegionAware}}(\mathcal{X}, \mathcal{G}^{\mathcal{S}}, \mathcal{G}^{\mathcal{T}}) = \mathbb{E}_{x\sim\mathcal{X}} \sum_{i=1}^{K} \mathcal{L}_{\text{InfoNCE}}(\mathcal{G}^{\mathcal{S}}(x)[:,i],\ \mathcal{G}^{\mathcal{T}}(x)[:,i],\ \{\mathcal{G}^{\mathcal{T}}[:,j] \mid j \in S', j \neq i\}), \tag{4}$$

It is observed that the main difference between Equation 2 and Equation 4 is that Equation 4 applies knowledge distillation to only the K crucial regions found by $\mathcal{A}$ instead of all the regions.

3.3 Perceptual Distillation

Perceptual distillation is proposed to distill teacher knowledge in the generated images. Usually, the native knowledge distillation methods directly minimize the L_1 norm distance between each pixel, which can be formulated as

$$\mathcal{L}_{\text{Naive Distill}} = \mathbb{E}_{x\sim\mathcal{X}} \left|\mathcal{F}^{\mathcal{S}}(x) - \mathcal{F}^{\mathcal{T}}(x)\right|_1 \tag{5}$$

In contrast, motivated by previous research in image super-resolution, in this paper we propose perceptual distillation, which minimizes the difference between students and teachers on the semantic features extracted by a ImageNet pre-trained model. Denote the pre-trained model as $\mathcal{J}(\cdot)$, then its loss function can be formulated as

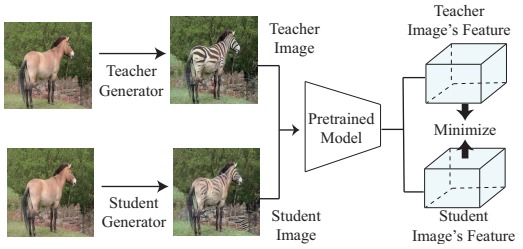


Figure 3: The overview of perceptual distillation. A ImageNet pre-trained model is utilized to extract the features of images generated by students and teachers. Then, the distance between these extracted features are minimized for distillation.

$$\mathcal{L}_{\text{Percep. Distill}} = \mathbb{E}_{x\sim\mathcal{X}} \left|\mathcal{J}\circ\mathcal{F}^{\mathcal{S}}(x) - \mathcal{J}\circ\mathcal{F}^{\mathcal{T}}(x)\right|_2. \tag{6}$$

Based on Equation 4 and 6, now we can formuate the overall loss function as

$$\mathcal{L}_{\text{overall}} = \alpha\cdot\mathcal{L}_{\text{RegionAware}} + \beta\cdot\mathcal{L}_{\text{Percep. Distill}} + \mathcal{L}_{\text{Origin}}, \tag{7}$$

where $\mathcal{L}_{\text{Origin}}$ indicates the origin training loss of GANs. α and β are two hyper-parameters introduced to balance different loss functions. Sensitivity studies on them have been conducted and shown in Appendix D. For the image-to-image translation models which have two generators such as CycleGAN, the distillation loss are applied to the two directions, respectively.

4 Experiments

4.1 Experiment Setting

Models, Datasets and Comparison Methods We evaluate our method on three image-to-image translation models including CycleGAN, Pix2Pix and Pix2PixHD, and two datasets including

Horse⟺Zebra and Edges2Shoes. Horse⟺Zebra is an unpaired image-to-image translation dataset which translates images of horses to zebras and vice versa. Edges2Shoes is a paired image-to-image dataset which maps the edges of shoes to their natural images. Besides, experiments on Cityscapes have also been conducted and shown in Appendix A. The students in our experiments have the same neural network depth but fewer channels compared with their teachers. Four GAN knowledge distillation methods are utilized for comparison. Note that some of these methods includes both neural network pruning and knowledge distillation and we only compare our method with the knowledge distillation parts in these comparison methods for fairness. To obtain more reliable results, we run 8 trials for each experiment and report their average and standard deviation.

Evaluation Settings *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 indicates the synthetic images have better quality. Please refer to the codes in the supplementary material for more details.

Figure 4: Qualitative results on Horse→Zebra and Zebra →Horse. A 15.81× compressed CycleGAN is utilized as the student. Results on Edges2Shoes are shown in Appendix B.

4.2 Experiment Results

Quantitative Result Quantitative results of our methods compared with four knowledge distillation methods have been shown in Table 1. It is observed that: **(a)** Our method leads to consistent and

significant performance improvements (FID reduction) on various datasets and models. On average, it leads to 12.65 and 5.08 FID reduction on unpaired and paired image-to-image translation tasks, respectively. **(b)** Our method outperforms the other four kinds of image-to-image translation knowledge distillation methods by a large margin. On average, it outperforms the second-best method by 7.77 FID. **(c)** Not all the knowledge distillation methods work well on GAN for image-to-image translation. Directly applying the naive Hinton knowledge distillation (Hinton et al., 2014) leads to very limited and even sometimes negative effects. For example, it leads to 1.91 FID increment on the Pix2Pix student for the Edges2Shoes task. **(d)** Compared with paired image-to-image translation, there are more performance improvements on unpaired image-to-image translation with our method.

Table 1: Quantitative comparison between different knowledge distillation methods. Numbers in the brackets indicate the ratio of compression and acceleration. **A lower FID indicates better performance.** Δ indicates the relative increment compared with the student trained without knowledge distillation (higher is better). Each experiment is averaged from 8 trials.

Models	Dataset	#Params (M)	FLOPs (G)	Method	Metric	
					FID↓	$\Delta\uparrow$
CycleGAN	Horse→Zebrea	11.38	49.64	Teacher	61.34±4.35	–
		0.72 (15.81×)	3.35 (14.82×)	Origin Student	85.04±6.88	–
				Hinton *et al.* 2014	84.08±3.78	0.96
				Li and Lin *et al.* 2020a	83.97±5.01	1.07
				Li and Jiang *et al.* 2020c	81.74±4.65	3.30
				Jin *et al.* 2021	82.37±8.56	2.67
				Ours	**71.04±6.21**	**14.00**
		1.61 (7.08×)	7.29 (6.80×)	Origin Student	70.54±9.63	–
				Hinton *et al.* 2014	70.35±3.27	0.18
				Li and Lin *et al.* 2020a	68.58±4.31	1.96
				Li and Jiang *et al.* 2020c	68.94±2.98	1.60
				Jin *et al.* 2021	67.31±3.01	3.23
				Ours	**59.98±5.48**	**10.56**
CycleGAN	Zebra→Horse	11.38	49.64	Teacher	138.07±4.01	–
		0.72 (15.81×)	3.35 (14.82×)	Origin Student	152.67±5.07	–
				Hinton *et al.* 2014	148.64±1.62	4.03
				Li and Lin *et al.* 2020a	151.32±2.31	1.35
				Li and Jiang *et al.* 2020c	151.09±3.67	1.58
				Jin *et al.* 2021	149.73±3.94	2.94
				Ours	**142.39±4.40**	**10.28**
		1.61 (7.08×)	7.29 (6.80×)	Origin Student	141.86±1.57	–
				Hinton *et al.* 2014	142.03±1.61	-0.17
				Li and Lin *et al.* 2020a	141.32±1.27	0.54
				Li and Jiang *et al.* 2020c	141.16±1.31	0.70
				Jin *et al.* 2021	140.98±1.41	0.88
				Ours	**136.91±2.90**	**15.76**
Pix2Pix	Edges2Shoes	54.41	6.06	Teacher	59.70±0.91	–
		13.61 (4.00×)	1.56 (3.88×)	Origin Student	85.06±0.98	–
				Hinton *et al.* 2014	86.97±3.49	-1.91
				Li and Lin *et al.* 2020a	83.63±3.12	1.43
				Li and Jiang *et al.* 2020c	84.01±2.31	1.05
				Jin *et al.* 2021	84.39±3.62	0.67
				Ours	**77.51±3.28**	**7.55**
Pix2PixHD	Edges2Shoes	45.59	48.36	Teacher	41.59±0.42	–
		1.61 (28.23×)	1.89 (25.59×)	Origin Student	44.64±0.54	–
				Hinton *et al.* 2014	45.31±0.63	-0.67
				Li and Lin *et al.* 2020a	44.03±0.41	0.61
				Li and Jiang *et al.* 2020c	43.90±0.36	1.28
				Jin *et al.* 2021	43.97±0.17	1.21
				Ours	**42.03±0.20**	**2.61**

Figure 5: The visualization of the learned similarity between student features in one region and teacher features in all regions. For each line, we plot the similarity when six different regions of students are selected. Three of these regions come from horses and the others come from the background. A Lighter region indicates a higher similarity.

This may be caused by the fact that there is less labeled supervision in unpaired image-to-image translation and thus the knowledge from teachers is more helpful. **(e)** A high ratio of acceleration and compression can be achieved by replacing the teacher model with the distilled student model. For example, our method leads to 7.08× compression and 6.80× acceleration on CycleGAN students in terms of the number of parameters and FLOPs. The compressed students outperform their teachers by 1.36 and 1.16 FID on Horse→Zebra and Zebra→Horse, respectively.

Qualitative results Qualitative results of our method on Horse→Zebra and Zebra→Horse have been shown in Figure 4. It is observed that the student model trained without knowledge distillation always can not translate the whole body of horses and zebras. In contrast, the student model trained with our methods does not have this issue. Moreover, on Horse→Zebra, the student model trained by our method sometimes outperforms its teacher on the effect of removing the stripes in zebras.

5 Discussion

5.1 Ablation Study

There are mainly three modules in the proposed region-aware knowledge distillation, including (a) localizing the crucial regions in images with attention mechanism (b) performing knowledge distillation with region-wise contrastive learning, and (c) distilling knowledge in the generated images with perceptual distillation. A series of ablation studies have been conducted to demonstrate their effectiveness. As shown in Table 2: (i) The basic framework of applying contrastive learning to knowledge distillation is beneficial even without the other two modules. (ii) By only distilling the features in the crucial regions, 5.01 FID reduction can be achieved. (iii) Individual usage of perceptual distillation leads to 3.02 FID reduction and applying it to the other two modules reduces FID from 61.10 to 59.98. These observations demonstrate that each module in our method is indispensable.

Table 2: Ablation studies of the three main modules in our method are Horse→Zebra with CycleGAN students. Each experiment is averaged from 8 trials. Reported results are FID (lower is better).

(a) Crucial Region	×	×	✓	×	✓
(b) Contrastive Distillation	×	✓	✓	×	✓
(c) Perceptual Distillation	×	×	×	✓	✓
Horse→Zebra	70.54	65.53	61.10	67.52	59.98

Ablations on Distilling the Crucial Regions To further show the effectiveness of only distilling the crucial regions, we have compared the following three schemes: (a) distilling regions with the largest attention (our scheme) (b) distilling the regions with the least attention and (opposite to our scheme) (c) randomly choose regions for knowledge distillation. Our experiments show that the three schemes achieve 59.98, 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. These results show that there is a positive relation between the attention value of a region and the benefits from distilling this region.

5.2 Visualizing the Similarity between Students and Teacher

The similarity of features from students and teacher have been visualized in Figure 5. For each line, we select student features of six different regions in an image. Note that three of the student regions are selected from the body of the horses and the other three regions comes from the background. Then, we compute the similarity between teacher features in all the regions and the student feature in the selected region. It is observed that when computing the similarity with respect to student features in regions of horses, teacher regions in the horse body have a much higher value than the background regions. When computing the similarity with respect to student features of background regions, teacher regions of the horses become are and teacher regions in the background are light. This result shows that there is a high similarity between student features and teacher features on the same location, which demonstrates the effectiveness of knowledge distillation.

5.3 Knowledge Distillation Can Stabilize GAN Training

The training of GAN is usually not stable due to their 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 early 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 more stable during the whole training period. Its undulations are much smaller than the student trained without knowledge distillation.

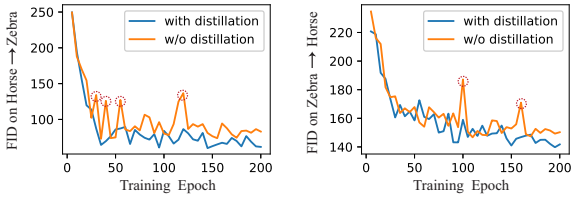


Figure 6: The FID curve of CycleGAN students trained with and without knowledge distillation on Horse→Zebra and Zebra→Horse.

6 Conclusion

Motivated by the observation that a large number of regions in image-to-image translation are not worthy to be distilled, this paper proposes region-aware knowledge distillation. First, attention mechanism is utilized to localize the crucial regions in the to be translated images. Then, a region-wise contrastive learning framework is employed for knowledge distillation, which maximizes the mutual information between the features of students and teachers in the same region. Besides, perceptual distillation is also introduced to transfer teacher knowledge in the generated images. Abundant experiments with four comparison methods have been conducted to demonstrate the effectiveness of our method. On average, 12.65 FID and 5.08 FID reduction can be observed on unpaired and paired image-to-image translation tasks, respectively. Our 7.08× compressed and 6.80× accelerated CycleGAN student outperforms its teacher by 1.36 and 1.16 FID on Horse→Zebra and Zebra→Horse respectively. In the discussion period, detailed ablation studies results have further shown the effectiveness of each module in our method. Besides, visualization results and FID curves during the training period show that our knowledge distillation method enables students to learn the similarity between different regions and stabilize the training of GANs.

REFERENCES

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

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.

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

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.

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

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.

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.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

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.

Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8183–8192, 2018.

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.

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.

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.

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.

Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

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, 2021.

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

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

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*, pp. 319–345. Springer, 2020.

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, 2019.

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

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*, pp. 4570–4580, 2019.

Han Shu, Yunhe Wang, Xu Jia, Kai Han, Hanting Chen, Chunjing Xu, Qi Tian, and Chang Xu. Co-evolutionary compression for unpaired image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3235–3244, 2019.

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

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

Haotao Wang, Shupeng Gui, Haichuan Yang, Ji Liu, and Zhangyang Wang. Gan slimming: All-in-one gan compression by a unified optimization framework. In *European Conference on Computer Vision*, pp. 54–73. Springer, 2020.

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*, 2018a.

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.

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, 2018c.

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.

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.

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

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

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pp. 4320–4328, 2018.

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.

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, 2017.

Table 3: Quantitative comparison between different knowledge distillation methods on Cityscapes with Pix2Pix. Numbers in the brackets indicate the ratio of compression and acceleration. **A higher mIoU indicates better performance.** Δ indicates the relative increment compared with the student trained without knowledge distillation (higher is better). Each experiment is averaged from 8 trials.

Models	Dataset	#Params (M)	FLOPs (G)	Method	Metric	
					mIoU↑	$\Delta\uparrow$
Pix2Pix	Cityscapes	54.41	96.97	Teacher	46.51±0.32	–
		13.61 (4.00×)	24.90 (3.88×)	Origin Student	41.35±0.22	–
				Hinton *et al.* 2014	40.49±0.41	-0.86
				Li and Lin *et al.* 2020a	41.52±0.34	0.17
				Li and Jiang *et al.* 2020c	41.77±0.30	0.42
				Jin *et al.* 2021	41.29±0.51	-0.06
				Ours	**42.41±0.25**	**1.06**

Figure 7: Qualitative results on Edges2Shoes. "w/o" indicates "without".

A Experiments on Cityscapes

Experiments on Cityscapes with Pix2Pix are shown in Table 3. Following previous works (Jin et al., 2021), we adopt the mIoU of a pre-trained segmentation model on the generated images as the performance metric on Cityscapes. A high mIoU indicates better performance. It is observed that the Pix2Pix student trained with our method leads to 1.06 mIoU improvements compared with the baseline, which outperforms the second-best knowledge distillation method by 0.64 mIoU.

B Qualitative Results on Edges2Shoes

Qualitative results on Edges2Shoes are shown in Figure 7. It is observed that the distilled student outperforms the student trained without knowledge distillation by a large margin. The distilled student has much better details such as the shoe string and the highlight on the shoes.

C Tricks: Random Projection Heads

As pointed out by many previous works on contrastive learning, the architecture and training methods of the projection heads have a significant influence on the performance of contrastive learning. In this paper, we fix the parameters of projection head and do not train them during the whole training period. Surprisingly, we find this trick can stabilize student training and lead to better performance.

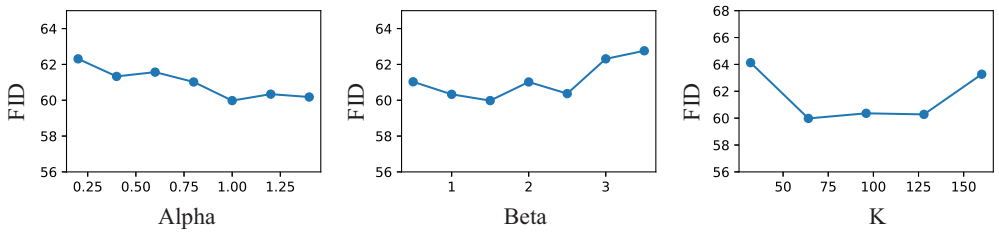


Figure 8: Sensivity studies on the three hyper-parameters with CycleGAN on Horse→Zebra.

D SENSITIVITY STUDY

There are mainly three hyper-parameter α, β and K introduced in our method. α and β are utilized to balance the magnitude of different loss functions and K is the number of crucial regions selected in an image. The sensitivity studies results on Horse→Zebra with CycleGAN students have been shown in Figure 8. It is observed that all our method is not sensitive to the choice of hyper-parameters. Even in the worst siatuation, our methods still outperforms the baseline (70.54 FID) and the second-best method (67.31 FID) by a clear margin.