EXPLOITING KNOWLEDGE DISTILLATION FOR FEW-SHOT IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot image generation, which trains generative models on limited examples, is of practical importance. The existing pipeline is first pretraining a source model (which contains a generator and a discriminator) on a large-scale dataset and fine-tuning it on a target domain with limited samples. The main challenge is that the few-shot model easily becomes overfitting. It can be attributed to two aspects: the lack of sample diversity for the generator and the failure of fidelity discrimination for the discriminator. In this paper, we treat the diversity and fidelity in the source model as a kind of knowledge and propose to improve the generation results via exploring knowledge distillation. The source model trained on the large-scale dataset is regarded as teacher model and the target model (student) is learned by introducing momentum relation distillation module to produce diverse samples and source discrimination distillation to ensure the fidelity discrimination. With the momentum relation distillation and source discrimination distillation modules, the proposed method outperforms the state-of-the-art by a large margin, $i.e.$, 10% for FFHQ to Sketches, while achieving better diversity.

1 INTRODUCTION

Training an image generation model often requires thousands of images in target domain (Brock et al., 2019; Karras et al., 2019; 2020b; Vahdat & Kautz, 2020). However, in many practical applications, there are only few images in such domains of interest. Few-shot image generation, which is able to use a large-scale dataset from source domain and requires limited samples on target domain, has attracted considerable attention due to its practical importance.

Conventional image generators fail on few-shot scenario, resulting in collapsed generation due to the limited number of images. To alleviate this issue, recent works introduce a two-stage pipeline follows the idea of transfer learning (Mo et al., 2020; Noguchi & Harada, 2019; Wang et al., 2020; 2018). Specifically, a source model is first trained on a large dataset to ensure necessary diversity of the generator. Then the target model inherit the parameters from the source model and is finetuned on a target domain with few images. However, the target model easily suffers from the overfitting problem when facing extremely small number of images such as 10 (Li et al., 2020; Ojha et al., 2021).

We examine the over-fitting problem through two aspects, the diversity and fidelity. On one hand, when only few images are available in the target domain, the discriminator can easily memorize these samples and force the generator to reproduce them. Thus, the diversity of the generated images is undesirable. On the other hand, the optimization of discriminator is difficult due to the limited number of training images in the fine-tuning process. So the discriminator may loss the ability of distinguishing realistic images from malformed ones, resulting in inferior fidelity. Ojha et al. (2021) introduce the regularization of cross-domain correspondence into the finetuning process to alleviate the overfitting problem. However, the problem of how to ensure the diversity and fidelity of few-shot image generator still remains.

In this paper, we treat the source model and target model as teacher and student respectively, and propose to improve both the diversity and fidelity of few-show image generation through knowledge distillation. Specifically, we introduce two novel distillation modules, named Momentum Relation Distillation and Source Discrimination Distillation, into the generator and the discriminator respectively. The Momentum Relation Distillation aims at transferring the image-to-image relation into

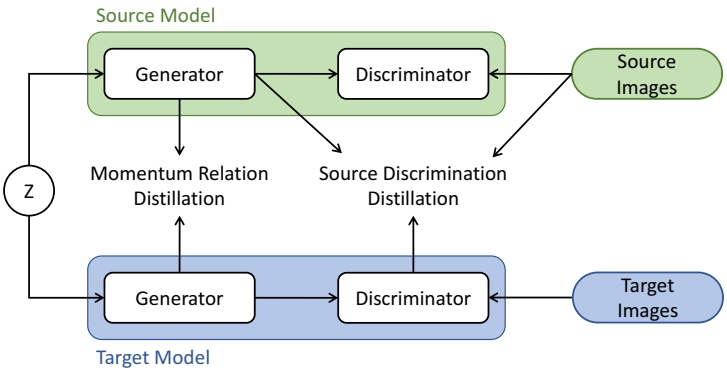

Figure 1: The framework of proposed knowledge distillation for few-show image generation. We treat the source model as teacher and the target model as student. Two novel distillation modules are introduced into the generator and discriminator respectively.

the target model and improving the generation diversity. In order to capture more informative relation, we propose an extend version of SSIM (Wang et al., 2003) as the relation metric and introduce a memory bank in the distillation. As to Source Discrimination Distillation, we force the discriminator to learn to distinguish real-fake images both in the source domain and target domain. In this manner, the discrimination of the target model is enhanced by the rich images in the source domain, leading to better generation fidelity.

Contributions. The main contribution of this paper is the proposed novel distillation framework to improve the few-shot image generation. By treating the source model and target model as teacher and student respectively, we introduce Momentum Relation Distillation and Source Discrimination Distillation to improve both the diversity and fidelity of the generation. Extensive qualitative and quantitative results demonstrate that our method can generate more realistic and diverse images in the target domain.

2 Related Work

Few-shot learning. Few-shot learning (Lake et al., 2015) is originally explored in discriminative works to learn to distinguish between novel visual categories from only few labelled samples. Recently, more and more attention has been paid to few-shot image generation. The most common and effective method is fine-tuning, where the initial model is pretrained well in the source domain with sufficient training samples and then adapted to the target domain with limited data. Some works employ model regularization to prevent model from over fitting to the training set (Mo et al., 2020; Li et al., 2020) or introduce new parameters to transform the distribution from source domain to target domain (Noguchi & Harada, 2019; Wang et al., 2020). Zhao et al. (2020) and Karras et al. (2020a) apply data augmentation to alleviate the lack of data, but it works ineffectively with extremely few samples ($e.g.$, 10 images). Ojha et al. (2021) propose cross-domain correspondence (CDC) to preserve distance between generated samples. Based on the prior works, we focus on relationships of generated images between source domain and target domain, with only few samples.

Distance preservation. DistanceGAN (Benaim & Wolf, 2017) introduces distance constraints between inputs and outputs to improve the distance preserving capability of the model and lessen mode collapse. The same idea has been applied to unconditional (Tran et al., 2018; Liu et al., 2019) and conditional (Mao et al., 2019; Yang et al., 2019) generation tasks to improve diversity of generated samples. In this work, our goal is to preserve the diversity of the model adapted from the source domain to the target domain.

Memory bank. Memory bank (Wu et al., 2018) has been widely used in contrastive learning to achieve more negative examples in one epoch. Some works (Munkhdalai & Yu, 2017; Santoro et al., 2016; Oreshkin et al., 2018; Mishra et al., 2017) have employed memory network methods

in the field of transfer learning to retain "experience" in the source domain. He et al. (2020) propose momentum update to keep the consistency of dictionary keys despite its evolution, where rapidly changing encoder will reduce the key representations' continuity. We introduce this idea to domain translation tasks, which will improve the consistency of the model adaptation from the source domain to the target domain.

Knowledge distillation. Knowledge distillation is the main non-invasive method used to transfer knowledge from a teacher network to a student network (Hinton et al., 2015), and has been widely applied to model compression in computer vision tasks (Chen et al., 2017; Luo et al., 2016; Yim et al., 2017). A recent survey (Gou et al., 2021) divides knowledge distillation into three categories: response-based, feature-based and relation-based. We leverage relation-based distillation in few-shot image generation to preserve diversity from source to target domain by Momentum Relation Distillation and Source Discrimination Distillation.

3 METHOD

Given a pretrained model G_S, which is capable of generating high-quality images in source domain D_S, our goal is fine-tuning it to a small target dataset D_T. In our work, we follow the architecture of StyleGAN2, a state-of-the-art architecture on image generation. Accordingly, the original loss function of StyleGAN2 is employed, which consists of an adversarial loss $\mathcal{L}_{adv}$ as follows:

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{x}\sim p_{data}(x)}[D(x)] - \mathbb{E}_{\mathbf{z}\sim p_{data}(z)}[D(G(z))], \tag{1}$$

where x, z denote real images and input noises respectively. However, the direct adaptation results in over-fitting when the number of training data in the target domain is extremely limited. Several works have been proposed to alleviate this issue by applying extra constraint on the adapted model. In the rest of this section, we first give an overview on the existing state-of-the-art algorithm, then we introduce our two novel distillation modules elaborately.

3.1 OVERVIEW

To overcome the issue of overfitting, Ojha et al. (2021) proposed correspondence between source and target domains for the generator, and relax the constraint of realism by introducing patch-discriminator. We choose this work as our baseline due to its outstanding generation results.

Formally, given a pretrained source model (G_S, D_S), we finetune it into the target model (G_T, D_T) with k training images. In one finetuning step, we first sample a mini-batch of latent code $Z = \{z_i\}$. Then we extract the intermediate feature maps of layer l from the generators, denoted as $G_S^l(z_i)$ and $G_T^l(z_i)$. Distances between feature maps are estimated through cosine similarity, and transferred into probability distribution. For latent code z_i, the distribution on others is formulated as

$$\begin{aligned} P_S^{i,l} &= \text{Softmax}(\{\cos(G_S^l(z_i)', G_S^l(z_j)')|\forall j \neq i\}) \\ P_T^{i,l} &= \text{Softmax}(\{\cos(G_T^l(z_i)', G_T^l(z_j)')|\forall j \neq i\}), \end{aligned} \tag{2}$$

where $'$ denotes reshaping into 1-dimensional vector. Then a correspondence loss is estimated through the KL-divergence between $P_S^{i,l}$ and $P_T^{i,l}$ as

$$\mathcal{L}_{corr} = \mathbb{E}_{z_i\sim p_z(z)} \sum_{i,l} D_{KL}(P_S^{i,l}||P_T^{i,l}). \tag{3}$$

We further sample k latent codes at the start of the finetuning and denote them as anchors Z^*. For z_i sampled from Z^*, we use the inherited image discriminator. Otherwise, we append extra classifier after the middle layer, and force adversarial loss on each patch. So the adversarial loss is computed as

$$\mathcal{L}^*_{adv}(G_T, D_T, D_P) = \mathbb{E}_{x\sim\mathcal{D}_T}[\mathbb{E}_{z\sim Z^*}\mathcal{L}_{adv}(G_T, D_T) + \mathbb{E}_{z\sim p_z(z)}\mathcal{L}_{adv}(G_T, D_P). \tag{4}$$

In this paper, we exploit knowledge distillation to improve generation results, by treating the source model as teacher and the target model as student. Specifically, we handle this problem from two aspects, the diversity and fidelity, and introduce Momentum Relation Distillation and Source Discrimination Distillation, into the generator and the discriminator respectively.

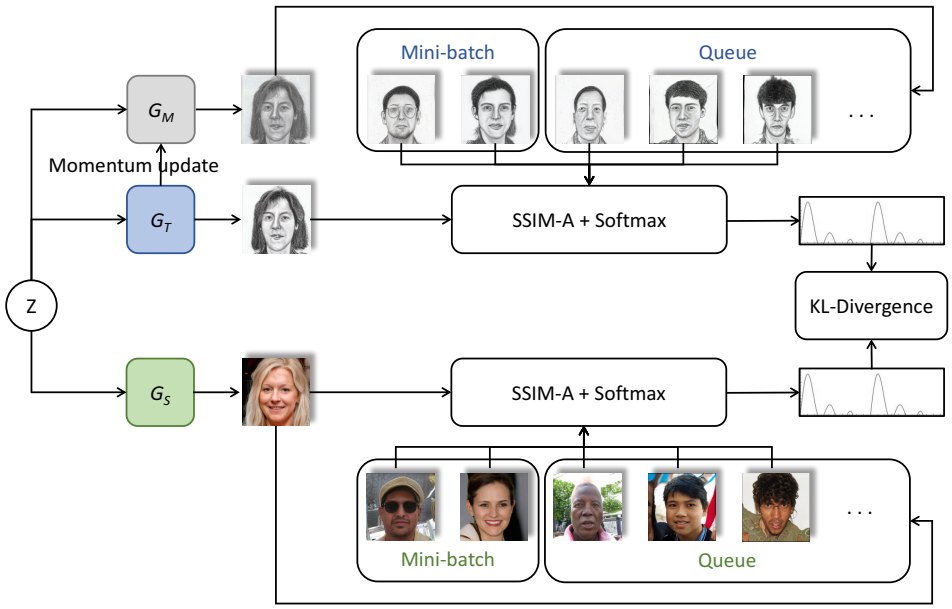

Figure 2: An illustration of the proposed Momentum Relation Distillation. Combined with the SSIM-A metric, the queue of instances generated by momentum model helps to capture more useful information.

3.2 Momentum Relation Distillation

The source model pretrained on large scale dataset has the ability to generate various images, and the relation between generated images directly captures this diversity. Following the baseline method, we treat the relation between generated images as a kind of knowledge, and introduce an effective way to distill this knowledge to the target model.

Definition of relation. To quantitatively define the relation, we utilize the similarity between intermediate activations of the generators to model the relation. However, it is hard to effectively measure the similarity of two high-dimensional feature maps. Inspired by the success of SSIM on estimating similarity of two images, we extend SSIM to activations, named SSIM-A, to capture semantic and spatial information simultaneously.

Specifically, given two feature maps $A_1 \in \mathbb{R}^{c\times h\times w}$ and $A_2 \in \mathbb{R}^{c\times h\times w}$, we estimate the similarity as

$$\text{SSIM-A}(A_1, A_2) = \frac{(2\mu_{A_1}\mu_{A_2} + c_1)(2\sigma_{A_1\cdot A_2} + c_2)}{(\mu_{A_1}^2 + \mu_{A_2}^2 + c_1)(\sigma_{A_1}^2 + \sigma_{A_2}^2 + c_2)}, \tag{5}$$

where μ_{A_1}, μ_{A_2} are the mean of a fixed window centered at each pixel, $\sigma_{A_1}, \sigma_{A_2}$ are standard deviations of A_1 and A_2 respectively. $\sigma_{A_1\cdot A_2}$ indicates the covariance. c_1, c_2 are constants to maintain stability.

Momentum relation. In order to transfer the knowledge of pairwise relation into the target model, we utilize the query-key formulation and converts similarities across instances into probability distributions as in (Ojha et al., 2021; Chen et al., 2020; Ye et al., 2021). Formally, given a list of activations as keys K, the probability can be estimated as

$$\begin{aligned} P_S^{i,l} &= \text{Softmax}(\{\text{SSIM-A}(G_S^l(z_i), K_S^{i,l})\}) \\ P_T^{i,l} &= \text{Sofrmax}(\{\text{SSIM-A}(G_T^l(z_i), K_T^{i,l})\}). \end{aligned} \tag{6}$$

Naturally, we use other samples in one mini-batch as keys. However, the batch size for image generation model is often small due to GPU memory limitation. Inspired by recent methods in contrastive learning, good information can be extracted with a large key list, while the generator for

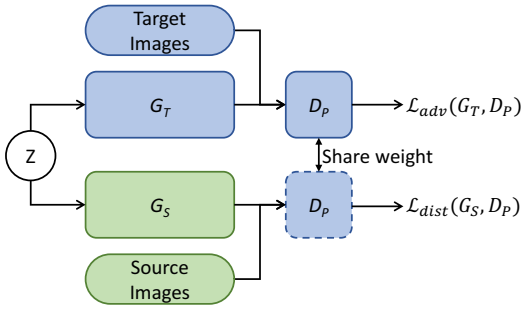

Figure 3: An illustration of the proposed Source Discrimination Distillation. The patch-discriminator of the target model is trained on both the source and target domain to enhance discrimination.

the keys is kept as consistent as possible. Thus, we introduce an extra generator G_M with momentum update, and extend the key list with a queue of activations generated by G_M. Thus, the key list $K_S^{i,l}$ and $K_T^{i,l}$ becomes

$$\begin{aligned} K_S^{i,l} &= \{G_S^l(z_j)|\forall j \neq i\} \bigcup \{G_S^l(Z), G_S^l(Z^{-1}), \ldots\} \\ K_T^{i,l} &= \{G_T^l(z_j)|\forall j \neq i\} \bigcup \{G_M^l(Z), G_M^{l,-1}(Z^{-1}), \ldots\}, \end{aligned} \tag{7}$$

where G_M^{-n} and Z^{-n} represent the momentum generator and latent code that are n steps before. Note that the source model G_S keeps the same along the finetuning process. In practice, we use the EMA model in StyleGAN2 as G_M for simplicity.

Relation distillation. Finally, KL-divergence is used to transfer the knowledge of relation from the source model into the target one, and the loss function is formulated as

$$\mathcal{L}_{dist}(G_T, G_S) = \mathbb{E}_{z_i \sim p_z(z)} \sum_{i,l} D_{KL}(P_S^{i,l} || P_T^{i,l}). \tag{8}$$

3.3 Source Discrimination Distillation

As few samples are available in the target domain, the discriminator easily becomes overfitting by simply memorizing all target images. Introducing patch-discriminator D_P helps to alleviate this issue. However, D_P is still difficult to optimize due to the limited number of real patches. The under-optimized D_P may fail to generalise well, leading the generator to create malformed images. We further treat the discrimination as a kind of knowledge, and propose to transfer the discrimination from source model to the finetuned one. Specifically, we force the target discriminator to distinguish source images from the output of source generator along the finetuning process. The discrimination distillation loss is defined as

$$\mathcal{L}_{dist}(G_S, D_P) = \mathbb{E}_{\mathbf{z} \sim p_z(z)} \mathcal{L}_{adv}(G_S, D_P). \tag{9}$$

3.4 Full Objective

The full objective contains three components:

$$\mathcal{L} = \mathcal{L}_{adv}^*(G_T, D_T, D_P) + \alpha \mathcal{L}_{dist}(G_T, G_S) + \beta \mathcal{L}_{dist}(G_S, D_P), \tag{10}$$

where α and β control the relative weight of different losses.

4 Experiments

Datasets: For real faces to other artistic faces, we use Flickr-Faces-HQ(FFHQ) (Karras et al., 2019) as the source domain, and employ the pre-trained model on FFHQ as G_S. Several other

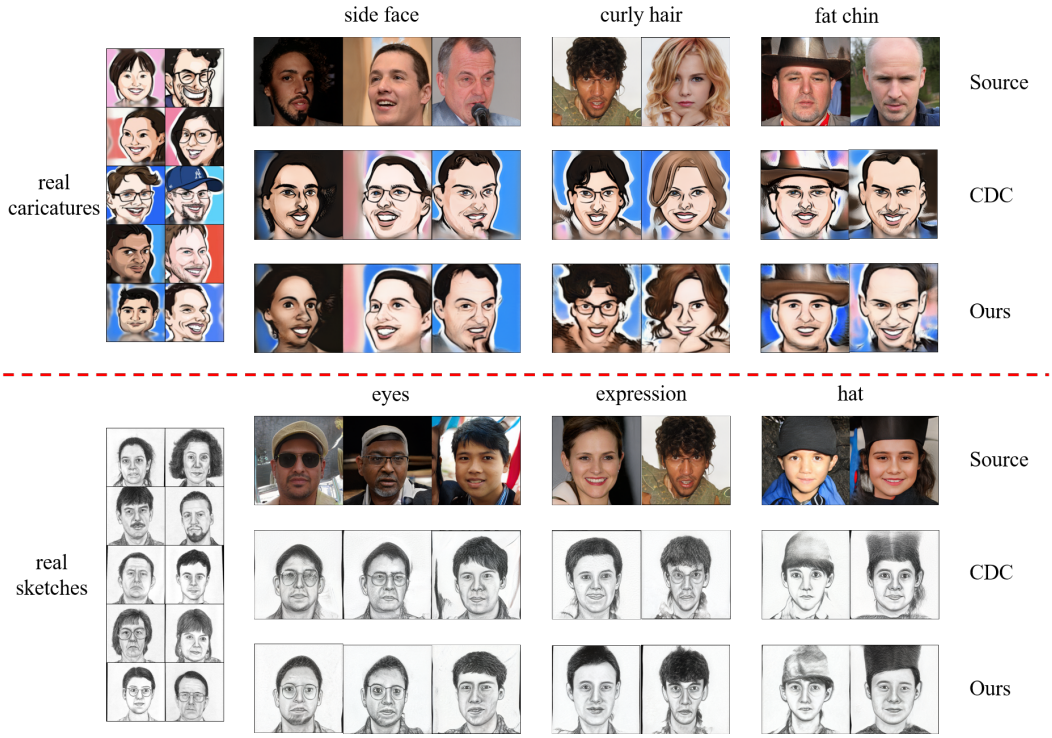

Figure 4: Visual results of 10-shot image adaptation, compared between CDC and our method. With the same latent code, our method can preserve more details from source domain with higher quality.

face databases are chosen as the target domain: (1) face sketches (Wang & Tang, 2008), (2) face paintings by Amedeo Modigliani (Yaniv et al., 2019), (3) FFHQ-babies, (4) FFHQ-sunglasses, (5) face caricatures (Ojha et al., 2021) and (6) face paintings by Rapheal. We also evaluate our methods on Church to Hauntedchurch and LSUN Cat (Yu et al., 2015) to AFHQ Dog (Choi et al., 2020). All the experiments are based on 256×256 resolution images if there is no special statement. It is worth mentioning that network can only access 10 images of the target domain during training.

Training details: The basic model of our method is StyleGANv2 (Karras et al., 2020b), pre-trained on a large dataset (e.g. FFHQ) as our source model. The batch size of our experiments is 4. We find the results are robust to different α and β, and we use $\alpha = 10^4$ and $\beta = 0.1$ by default.

Competing methods: There are several existing works on few-shot image generation, and we compare the proposed method with: (1) Transferring GANs (TGAN) (Wang et al., 2018), (2) Data augmentations (Karras et al., 2020a), (3) Batch Statistics Adaptation (BSA) (Noguchi & Harada, 2019), (4) Freeze Discriminator (Freeze-D) (Mo et al., 2020), (5) MineGAN (wang2020minegan) (Wang et al., 2020), (6) Elastic Weight Consolidation (EWC) (Li et al., 2020), and (7) Cross-domain Correspondence (CDC) (Ojha et al., 2021).

4.1 Qualitative comparisons

Figure 4 shows result comparison between our method and CDC, which is the state-of-the-art in terms of few-shot image generation with only 10 target images. We observe that our method can generate more realistic and diverse images. Taking FFHQ→caricatures as an example, we find that:

- CDC prefers to frontalizing faces, while our method keep the face angle compared with results of G_S.
- Images generated by CDC have similar eyes (e.g. eye color, eye size) and hair (e.g. curly hair) which are close to the training set, while those of our results have more variance and keep diversity from source domain.

Table 1: FID scores (↓) for domains with abundant data. Standard deviations are computed across 5 runs.

	Babies	Sunglasses	Sketches
TGAN (Wang et al., 2018)	104.79 ± 0.03	55.61 ± 0.04	53.14 ± 0.02
TGAN+ADA (Karras et al., 2020a)	102.58 ± 0.12	53.64 ± 0.08	66.99 ± 0.01
BSA (Noguchi & Harada, 2019)	140.34 ± 0.01	76.12 ± 0.01	69.32 ± 0.02
FreezeD (Mo et al., 2020)	110.92 ± 0.02	51.29 ± 0.05	46.54 ± 0.01
MineGAN (Wang et al., 2020)	98.23 ± 0.03	68.91 ± 0.03	64.34 ± 0.02
EWC (Li et al., 2020)	87.41 ± 0.02	59.73 ± 0.04	71.25 ± 0.01
CDC (Ojha et al., 2021)	74.39 ± 0.03	42.13 ± 0.04	45.67 ± 0.02
Ours	$\mathbf{68.67 \pm 0.02}$	$\mathbf{34.61 \pm 0.03}$	$\mathbf{35.87 \pm 0.02}$

Table 2: Intra-cluster pairwise LPIPS distance (↑). Standard deviation is computed across the k clusters, where k is the number of training samples).

	Caricatures	Amedeo's paintings	Sketches
TGAN (Wang et al., 2018)	0.39 ± 0.06	0.41 ± 0.03	0.39 ± 0.03
TGAN+ADA (Karras et al., 2020a)	0.50 ± 0.05	0.51 ± 0.04	0.41 ± 0.05
BSA (Noguchi & Harada, 2019)	0.35 ± 0.01	0.39 ± 0.04	0.35 ± 0.01
FreezeD (Mo et al., 2020)	0.37 ± 0.01	0.40 ± 0.03	0.39 ± 0.03
MineGAN (Wang et al., 2020)	0.39 ± 0.07	0.42 ± 0.03	0.40 ± 0.05
EWC (Li et al., 2020)	0.47 ± 0.03	0.52 ± 0.03	0.42 ± 0.03
CDC (Ojha et al., 2021)	0.53 ± 0.01	0.60 ± 0.01	0.45 ± 0.02
Ours	$\mathbf{0.57 \pm 0.02}$	$\mathbf{0.62 \pm 0.01}$	$\mathbf{0.47 \pm 0.02}$

- It can be found that the lower part of faces for CDC are generally fatter, which is not a common feature of the training set.

Similar improvements of our method can also be observed in FFHQ→Sketches. In general, our method performs better on different target domains both in quality and diversity. More qualitative results can be found in appendix.

4.2 Quantitative comparisons

In terms of image generation, there are two widely used evaluation to measure generative model, Fréchet Inception Distance (FID) (Heusel et al., 2017) for the quality of generated images and LPIPS metric (Zhang et al., 2018) for the variance of results. In our work, we also evaluate our method through the fidelity and diversity of G_T's generating space on the target domain.

Quality evaluation. For the measurement of image quality, adequate real samples in the target domain are needed. In our experiments, the FFHQ-babies, FFHQ-sunglasses and Sketches datasets contain enough real data. We evaluate FID score in these experiments the target images except the 10 images randomly selected for training.

Table 1 shows the quality evaluation between different methods with 10-shot adaptation. Our method achieve the lowest FID score for all target domains compared with other methods, and supasses the existing state-of-the-art by a large margin.

Diversity evaluation. Besides the quality of generated images, diversity is also of great importance. For example, if G_T only output images that are identical to the k samples of training data, the FID score will be very low, while the result is not desirable. In order to measure the image diversity of different models, we generate 1K images and divide them into k clusters according to the LPIPS distance with k training samples, following CDC (Ojha et al., 2021). Then the LPIPS distances are calculated inside each cluster and the final score is averaged over k clusters. The higher final score, the more diverse the output distribution of G_T is.

Table 3: Ablation Study on FFHQ→Sketches, evaluated both from image quality and diversity. ✓ represents adaptation with the method.

	SSIM-A	Momentum	Discrimination	LPIPS	FID
Baseline				0.45 ± 0.02	45.67 ± 0.02
ours	✓			0.46 ± 0.02	41.19 ± 0.02
	✓	✓		0.46 ± 0.01	37.64 ± 0.03
	✓	✓	✓	$\mathbf{0.47 \pm 0.02}$	$\mathbf{35.87 \pm 0.02}$

Table 4: Ablation Study on FFHQ→Sketches, evaluated both from image quality and diversity.

queue size	Baseline	4	8	12	16
LPIPS	0.45 ± 0.02	0.46 ± 0.03	0.47 ± 0.02	0.47 ± 0.02	$\mathbf{0.47 \pm 0.02}$
FID	45.67 ± 0.02	38.98 ± 0.01	36.99 ± 0.01	36.43 ± 0.03	$\mathbf{35.87 \pm 0.02}$

The distances of different methods are shown in the table 2. Our method always scores the highest LPIPS distance over the three domains. The FFHQ→Sketches is both evaluated with FID score and LPIPS distance, and test model is the same model. So the proposed algorithm can improve the fidelity and diversity simultaneously.

4.3 Ablation Study

Effectiveness of modules. We evaluate the effectiveness of the proposed modules through both FID and LPIPS trained from FFHQ to Sketches. As shown in Table 3, each part of our method can improve the distillation performance. By combining the proposed MRD and SDD, we achieve the best results both in FID and LPIPS.

Influence of queue size. The queue in the proposed Momentum Relation Distillation plays an important role in capturing informative relation. We conduct ablation study on the choice of queue size and show the result in Table. 4. It can be seen that both the diversity and fidelity becomes better as the queue size increases. The improvement is marginal when we increase the queue size from 12 to 16, so we choose $k = 16$ for other experiments as default.

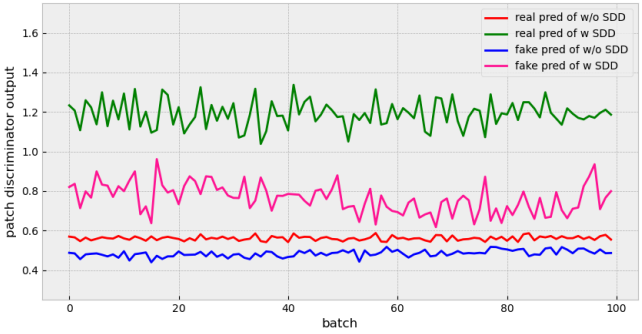

Figure 5: Patch Discriminator outputs of with and without SDD. The patch discriminator of with SDD has better discrimination than that of without SDD.

Impact of SDD. We conduct comparative experiments to analyze the role of Source Discrimination Distillation. We show the averaged value of the patch-discriminator outputs after softplus operation. Note that the discriminator loss is calculated by $D_{fake} - D_{real}$, which measures the discrimination. It can be observed from Figure 5 that the gap between D_{fake} and D_{real} of the model with SDD is greater than that without SDD. In other words, SDD can help the patch-discriminator to distinguish real/fake images, and lead the generator to create high quality images.

Figure 6: Results with different sizes of training set.

Training data size. We explore our method on different number of training data. As shown in Figure 6, the model overfits when only one image is available, but performs well enough with 5 target images. Quantitatively, we achieve 40.87 FID and 0.46 LPIPS under 5-shot setting, which is even better than CDC under 10-shot. This further demonstrates the superior of our method.

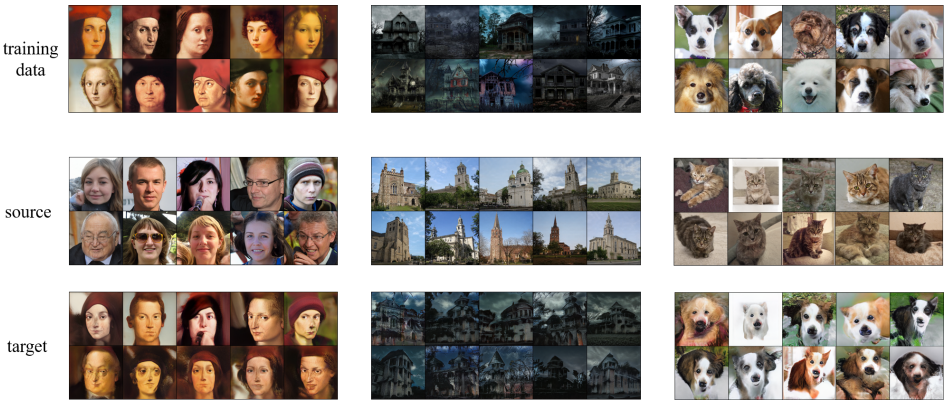

Figure 7: Generalization on different domains.

4.4 Generalization of Our Method

In order to verify the generalization and robustness of our method, we also conduct experiments on different source and target domain pairs: (1) FFHQ→Rapheal painting, (2) Church→Haunted church and (3) LSUN Cat→AFHQ Dog. The results of our method are shown in Figure 7, where the source and target results are generated with the same latent codes. It can be found that target images preserve the structures (e.g. pose, shape) of source images while render the style of target domain. Given that data are difficult to obtain in some target domains, we can achieve better generation by knowledge distillation between the relevant source and target domain.

5 Conclusion

In this work, we propose to exploit knowledge distillation image generation with few data. It is based on Momentum Relation Distillation and Source Discrimination Distillation, for generator and discriminator respectively. Momentum Relation Distillation creates a dictionary of feature tensors with momentum update for relation distillation, and Source Discrimination Distillation makes use of source data for discrimination distillation. We demonstrate that our method can work well on different relevant domains and performs better against other methods.

REFERENCES

Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. *arXiv preprint arXiv:1706.00826*, 2017.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *ArXiv*, abs/1809.11096, 2019.

Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

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.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020b.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*, 2020.

Shaohui Liu, Xiao Zhang, Jianqiao Wangni, and Jianbo Shi. Normalized diversification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10306–10315, 2019.

Ping Luo, Zhenyao Zhu, Ziwei Liu, Xiaogang Wang, and Xiaoou Tang. Face model compression by distilling knowledge from neurons. In *Thirtieth AAAI conference on artificial intelligence*, 2016.

Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1429–1437, 2019.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017.

Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020.

Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *International Conference on Machine Learning*, pp. 2554–2563. PMLR, 2017.

Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2750–2758, 2019.

Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10743–10752, 2021.

Boris N Oreshkin, Pau Rodriguez, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *arXiv preprint arXiv:1805.10123*, 2018.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pp. 1842–1850. PMLR, 2016.

Ngoc-Trung Tran, Tuan-Anh Bui, and Ngai-Man Cheung. Dist-gan: An improved gan using distance constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 370–385, 2018.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *ArXiv*, abs/2007.03898, 2020.

Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(11):1955–1967, 2008.

Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *ECCV*, 2018.

Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9332–9341, 2020.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. *arXiv preprint arXiv:1901.09024*, 2019.

Jordan Yaniv, Yael Newman, and Ariel Shamir. The face of art: landmark detection and geometric style in portraits. *ACM Transactions on graphics (TOG)*, 38(4):1–15, 2019.

Han-Jia Ye, Lu Ming, De-Chuan Zhan, and Wei-Lun Chao. Few-shot learning with a strong teacher. *arXiv preprint arXiv:2107.00197*, 2021.

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.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *arXiv preprint arXiv:2006.10738*, 2020.

A APPENDIX

We conduct abundant experiments on different pairwise source and target domains. Figure 8 shows our results on FFHQ to other face domains. The results show excellent effects both in structure preservation and style transformation.

We also leverage our method on animal face domains. Figure 7 (LSUN Cat $\rightarrow$ AFHQ Dog) and 9 (LSUN Cat $\rightarrow$ AFHQ wild (Choi et al., 2020)) show that animals' faces can simply transfer to each other, but it is hard for human faces to animal faces. As shown in Figure 9, although target images of FFHQ $\rightarrow$ LSUN Dog present style of dogs, but the structure of human faces still exist. That is because there is too much difference between human face and animal face, while animals (e.g. cat, dog, tiger, fox) have similar faces.

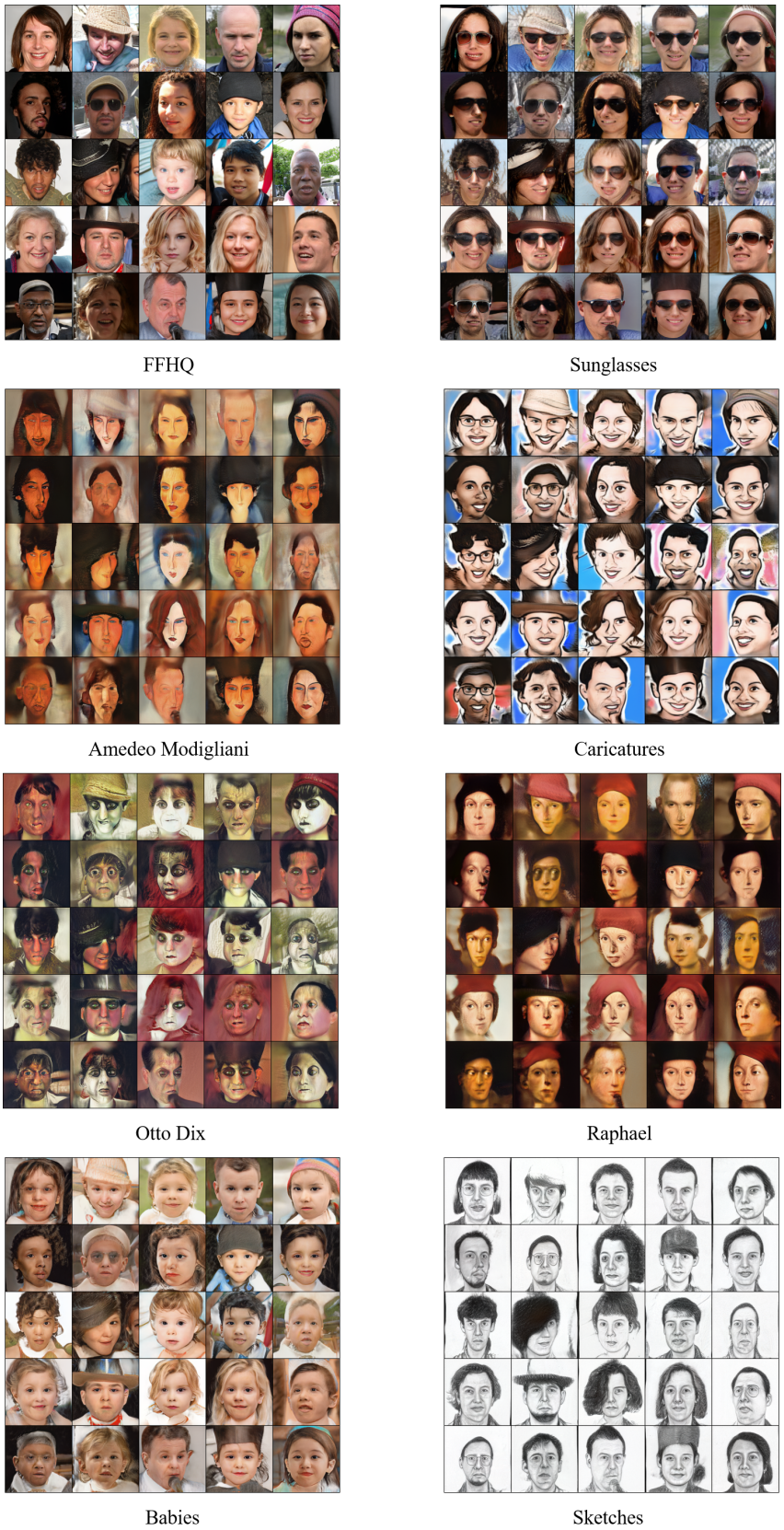

Figure 8: Experiments on FFHQ to other human face domains.

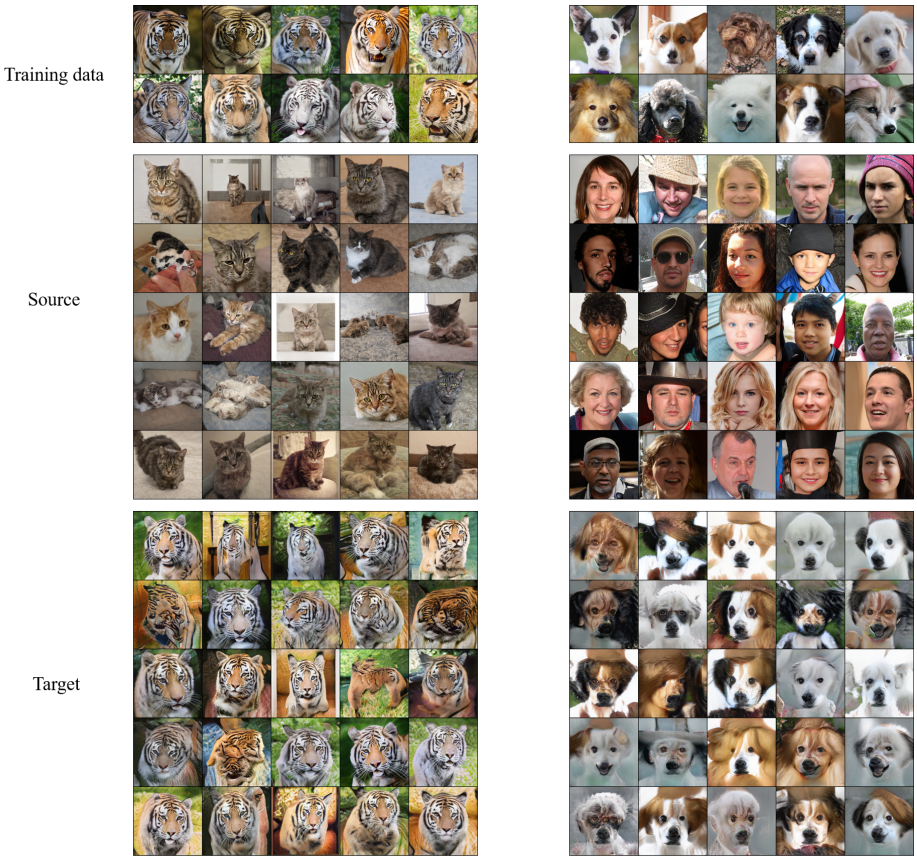

Figure 9: Experiments on face domains.