$\mathcal{C}^3$: Contrastive Learning for Cross-domain Correspondence in Few-shot Image Generation

Hyuk-Gi Lee **Gi-Cheon Kang** **Chang-Hoon Jeong** **Han-Wool Sul** **Byoung-Tak Zhang**
Seoul National University, Artificial Intelligence institute (AIIS)
{hklee,gckang,chjeong,hwsul,btzhang}@bi.snu.ac.kr

Abstract

Few-shot image generation is a task of generating high-quality and diverse images well fitted to the target domain. The generative model should adapt from the source domain to the target domain given a few images. Despite recent progresses in generative models, cutting edge generative models (e.g., GANs) still suffer from synthesizing high-quality and diverse images in few-shot setting. One of the biggest hurdles is that the number of images from the target domain is too small to approximate the true distribution of the target domain. To this end, the effective approach for the few-shot adaption is required to address the problem. In this paper, we propose a simple yet effective method $\mathcal{C}^3$, **C**ontrastive Learning for **C**ross-domain **C**orrespondence. $\mathcal{C}^3$ method constitutes the positive and negative pairs of images from two different domains and makes the generative model learn the cross-domain correspondence (i.e., semantic mapping from the source domain to the target domain) explicitly via contrastive learning. As a result, our proposed method generates more realistic and diverse images compared to the baseline methods and outperforms the state-of-the-art approaches on photorealistic and non-photorealistic domains.

1 Introduction

Deep generative models (e.g., GANs) have shown remarkable success in various image synthesis domains, including text-to-image generation [1, 2, 3, 4, 5], image-to-image translation [6, 7, 8], and image manipulation [9, 10, 11]. However, cutting edge deep generative models still suffer from synthesizing high-quality and diverse images using a limited number of images. Recent studies [12, 13, 14, 15, 16] have explored such constrained image generation task under the name of few-shot image generation, where the generative model should adapt from the source domain to the target domain given a few images. In few-shot image generation, the generative model is typically trained on a large-scale image dataset (e.g., FFHQ [17]) from the source domain and performs the few-shot adaptation on the target domain.

One of the biggest hurdles in few-shot image generation is that the number of given images from the target domain – typically less than ten images – is too small to approximate the true distribution of the target domain. Accordingly, the efficient domain adaptation mechanism is required to address the problem. Previous works have addressed the problem with transfer learning technique [12], auxiliary network [18], and regularization tricks [13, 14, 15, 16]. Notably, CDC [16] proposed the cross-domain correspondence to preserve the semantic similarity between the source and target domain. Specifically, CDC computes the pairwise distances among images from the source domain and transfers the distribution of pairwise distances to the target domain, encouraging that the distribution from the source domain matches the distribution from the target domain. However, we argue that existing methods for few-shot image generation have shown limited domain adaptation capabilities in

1st CtrlGen: Controllable Generative Modeling in Language and Vision Workshop at NeurIPS 2021.

that they have implicitly attempted to transfer the knowledge from the source to the target domain without any direct mapping from the source to target images.

In this paper, we propose a simple yet effective approach $\mathcal{C}^3$, **C**ontrastive Learning for **C**ross-domain **C**orrespondence, that learns the cross-domain correspondence in an explicit way. Specifically, the generative model (i.e., GANs) firstly generates the target images given the latent vector z sampled from the latent space. Then, $\mathcal{C}^3$ constitutes the positive and negative pairs from the source and target image feature vectors and optimizes the pairwise distances via contrastive learning. From this pipeline, our proposed method makes the generative model to learn the distinctive attributes that the positive pair of images should share. Fig. 1 shows the overview of our approach, $\mathcal{C}^3$.

The main contributions of our paper are as follows. First, we propose $\mathcal{C}^3$ method for few-shot image generation. By leveraging contrastive learning, the generative model learns the semantic similarity between the source and target domain. Second, we validate the effectiveness of our proposed method on photorealistic and non-photorealistic domains by comparing $\mathcal{C}^3$ with the state-of-the-art approaches. Finally, we perform qualitative analysis of our model, demonstrating that $\mathcal{C}^3$ makes the generative model synthesize high-quality and diverse images.

2 Related Work

Few-shot generative models Recently, several studies on few-shot learning have been conducted for generative tasks[19, 20, 18, 14, 13, 16, 15]. In generative task, a few-shot learner aims to generate high-quality and diverse images given in small amount of examples while preventing the generative model over-fitting. Basically, most works follow a training phase, where transfer prior knowledge of pre-trained model on source domain to target model for fitting to a smaller target domain[18, 14, 13, 16, 15]. In previous works, [14], [13] reduce the number of network parameters that changed during adapting to target domain avoiding over-fitting. [18] introduces a small network for transforming source latent space to other latent space, which is more relevant to target domain. [15] uses a regularization term that is applied differently to learnable parameters depending on the importance during adaptation. [16] also uses regularization term, which enforcing relative similarity between source and corresponding target images. In contrast to prior work, our approach directly applies contrastive learning for semantic correspondence between source and target domains.

Contrastive learning for image synthesis Contrastive learning is to learn a metric space where similar sample pairs stay close to each other while dissimilar ones are distant. Contrastive learning has shown effectiveness for self-supervised representation learning[21, 22, 23, 24, 25]. Due to its powerful representation learning performance, contrastive learning has been used extensively in image synthesis such as conditional image generation[26, 27], text-to-image generation [4, 28, 5], image-to-image translation [29, 30]. In terms of maintaining the structure of the source domain during adaptation, image-to-image translation is more relevant to few-shot image generation than other tasks. In that works, [29] uses patchwise contrastive loss to maximize mutual information between the corresponding patches of source image and target image. [30] uses contrastive loss in both generator and discriminator for learning representations between real, reference, and augmented images. To our best knowledge, this work first uses contrastive learning for unconditional image generation given on few shot target data.

3 Method

3.1 Problem formulation

In this subsection, we formally describe the problem of few-shot image generation. Given a few images from a target domain $\mathcal{D}_t$, the target generator $G_{s\to t}$ adapts from the source domain to the target domain. The target generator is typically trained on a large-scale source dataset (e.g., FFHQ [17]) before the few-shot adaptation. Formally,

$$G_{s\to t} = \mathbb{E}_{z\sim p_z(z), x\sim\mathcal{D}_t} \arg\min_G \max_D \mathcal{L}_{adv}(G, D) \tag{1}$$

$$\mathcal{L}_{adv} = \mathbb{E}_{x\sim p_{data}(x)} \left[\log D(x)\right] + \mathbb{E}_{z\sim p_z(z)} \left[\log\left(1 - D\left(G(z)\right)\right)\right] \tag{2}$$

where $z \sim p_z(z)$ and $x \sim p_{data}(x)$ denote the latent vectors and the samples from real data, respectively. Note that the same noise vectors z are used in generating the source and target images.

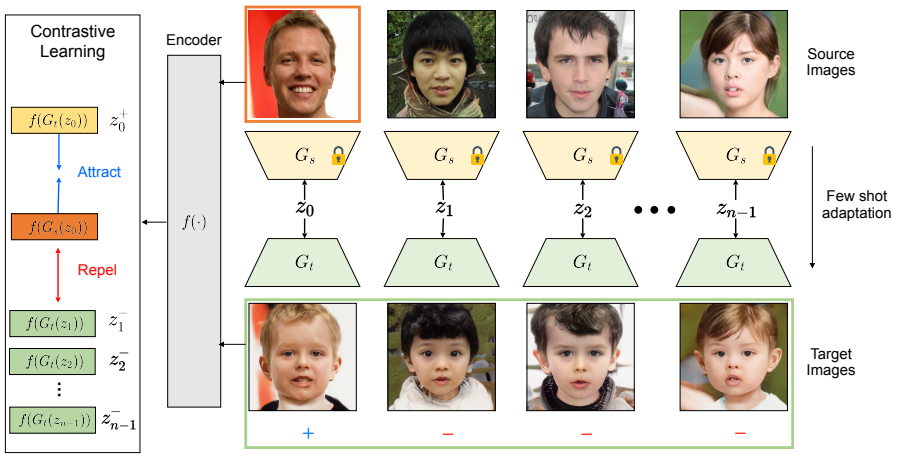

Figure 1: An overview of our approach, $\mathcal{C}^3$. The source image is an anchor (orange box). Positive pair consists of an anchor and target image (+ on green box) from the same latent vector mapped to anchor. Negative pairs consist of anchor and others (- on green box) from different latent vectors within minibatch. $\mathcal{C}^3$ makes semantic similarity of corresponding pair (positive pair) high during adaptation, otherwise vice versa.

$\mathcal{L}_{adv}(G, D)$ is the adversarial loss for Generative Adversarial Networks (GANs). However, as previous works [15, 16] demonstrated, a naive domain transfer described above tends to be over-fitting to target domain in few shot settings. Over-fitted target model reproduces training data, not generating diverse images. In other words, an effective regularization technique is required to mitigate the over-fitting problem. In the following sections, we carefully describe the contrastive learning method for cross-domain correspondence (Sec. 3.2.) and final objective functions for $\mathcal{C}^3$ (Sec. 3.3.).

3.2 Contrastive loss for cross-domain correspondence

In few shot image generation, over-fitting occurs when many latent vectors are mapped to just few images. This is because the goal of the generator in GANs is just fooling the discriminator with the high-quality generated images. The *diversity* is not explicitly considered in the goal. We then hypothesize that given on few target data, if source and target images from same latent vectors are semantically well-aligned while adaptation, target generator can generate diverse images avoiding over-fitting. To achieve this, we maximize the mutual information between the corresponding pair which should maintain high semantic similarity. Since it is difficult to directly maximize mutual information, we alternatively use contrastive loss for maximizing the lower bound of the mutual information. Given source image $G_s(z)$ and target image $G_t(z)$ from same latent variable $z \sim p_z(z)$, we define a score function following previous work [21, 22, 23] on contrastive learning.

$$S_{sim}(G_s(z), G_t(z)) = cos(f(G_s(z)), f(G_t(z)))/\tau \tag{3}$$

where $cos(u, v) = u^T v/\|u\|\|v\|$ denotes cosine similarity and τ denotes a temperature hyper-parameter. f is an encoder network to extract feature maps from source and target images. we use pre-trained VGG16[31] network as f. The contrastive loss between $G_s(z_i)$ and $G_t(z_i)$ is computed as:

$$\mathcal{L}_{con}\left(G_s(z_i), G_t(z_i)\right) = -\log \frac{\exp\left(\cos\left(f\left(G_s(z_i)\right), f\left(G_t(z_i)\right)\right)/\tau\right)}{\sum_{j=1}^{M} \exp\left(\cos\left(f\left(G_s(z_i\right), f\left(G_t(z_j)\right)\right)/\tau\right)} \tag{4}$$

Contrastive loss takes output of the encoder for source and target image from same latent vector as corresponding pair(positive pair) otherwise as negative pairs within minibatch size M. This loss makes target generator keeping semantic similarity with source generator, which helps prevent over-fitting while adapting to target domain.

3.3 Final objective

Our final objective consists of $\mathcal{L}_{adv}$ and $\mathcal{L}_{con}$:

$$G_{s\to t} = \arg\min_{G}\max_{D_{patch}} \mathcal{L}_{adv}(G, D_{patch}) + \lambda_{con}\mathcal{L}_{con}(G, G_s) \tag{5}$$

L_{adv} is an adversarial loss for training Generative Adversarial Networks(GANs), and $\mathcal{L}_{con}$ is a contrastive loss for keeping source-target semantic similarity while adapting to target domain. λ_{con} is regularization weight to balance between adversarial loss and contrastive loss. D_{patch} means patch discriminator [32], which tries to discriminate if each N × N patch in an image is real or fake. If N is smaller than the full size of the image, the patch discriminator classifies the whole image as real or fake based on the part of the image. In few-shot generative setting, using patch discriminator can give degree of freedom to generator generating diverse images compared with when using image discriminator [16]. We use patch discriminator following previous work [16]. However, the difference between the previous work and our work is that we do not use patch discriminator and image discriminator together, but only patch discriminator simply. Only using patch discriminator shows the competitive results.

4 Experimental Results

In this section, we describe the details of our experiments. We then compare our method and baselines in qualitative and quantitative. We use the StyleGANv2 architecture [33][1], pre-trained on a each large dataset (e.g., FFHQ[17], LSUN-Church) as our source model following [16]. Resolution images are also 256 $\times$ 256 both on source and target images. Adaptation is also done on 10 images from the target domain following [16]. We use effective patch size of patch discriminator as range from 61 $\times$ 61 to 189 $\times$ 189 and use batch size of 4 and temperature parameter τ as 0.1. Empirically, we observe that λ_{con}, from 0.05 to 0.2, to work well.

Dataset We use several source/target domains of datasets for experiments. We choose source domains as 3 categories: 1) Real faces (FFHQ [17]), 2) Real Object (LSUN-Church), 3) Real Animal (LSUN-Cat). We then adapt to 4 categories of datasets as target domain based on distance with source domain: 1) Real faces (FFHQ-Babies, FFHQ-sunglasses), 2) Artistic faces (face sketches [34], face paintings by Modigliani [35], face paintings by Raphael), 3) Artistic Object (haunted houses, Van Gogh's house paintings), 4) Real animal faces (AFHQ-dog [36])

Baselines We set baselines among methods using pre-trained model as source model to adapt to target domains with limited data. 1) TGAN[12] : fine-tunes without regularization to target model from pre-trained source model. 2) TGAN+ADA[37] : fine-tunes to target model with only data augmentation 3) BSA[13] : Except parameters of other layers in source model, update only scale and shift parameters in the normalization layers. 4) FreezeD[14] : freezes the lower layers of discriminator for reducing parameter changed during adaptation 5) MineGAN[18][2] : introduces auxiliary network called miner for transforming source latent space to another latent space. This is for finding latent space relevant with target for adaptation. 6) EWC[15] : applies regularization term as Elastic Weight Consolidation[38] by penalizing important parameters of source model being should not to change large while adapting to target domain. 7) CDC[16][3] : avoiding over-fitting, computes the distribution of the pairwise distances among images from the source domain to enforce cross-domain correspondence between source model and target model.

Evaluation metrics Following previous work [16], we report two standard metrics for quantitative results between proposed method and baselines: Frechet Inception Distance(FID)[39] and Learned Perceptual Image Patch Similarity(LPIPS) [40]. For calculating Frechet Inception Distance(FID), we generate 10,000 images randomly by adapted target model and use the entire target dataset, not 10 images for training. This is for assessing that target model learns true target distribution well or not. But FID score as few-shot metric does not capture over-fitting problem [41], so we report LPIPS score for assessing diversity. As in [16], we generate 1,000 images randomly and assign each image

[1] https://github.com/rosinality/stylegan2-pytorch

[2] https://github.com/yaxingwang/MineGAN/tree/master/styleGANv2

[3] https://github.com/utkarshojha/few-shot-gan-adaptation

into i-th cluster based on the lowest LPIPS among k clusters where k equals the number of training samples. Then we compute the average LPIPS score within members of the same cluster and average over k clusters. If target generator just reproduces training data, then LPIPS score will be almost zero.

Figure 2: Adaptation results for different main methods to target domains given on 10-shot target data. we observe that some baselines generate images similar to training data and others generate diverse images but are less realistic. Our method generates better quality images while keeping correspondence to the source domain.

4.1 Qualitative and quantitative results

Qualitative results Fig. 2 shows the results of several primary methods on three target domains depending on distance with source domain. All methods starting from same pre-trained source model, which is initialized on FFHQ adapt to FFHQ-babies (top), FFHQ-sunglasses (middle), and Face Sketches (bottom). FFHQ-babies and FFHQ-sunglasses are relatively near to source domain compared to Face sketches because Face sketches are also face domain but have quite different textures, styles with source domain.

In near distance case (top, middle), we observe that images generated by MineGAN have no semantic similarity such as pose, expressions with source images, and have poor quality. In contrast, images generated by other methods maintained structural consistency with source images. But, results for EWC show less realistic as of blurry. CDC performs better than EWC by generating more realistic images. However, some attributes (e.g., chin or beard) are not properly changed (row 1, 4, 5 in top), or some images are somewhat artificial (row 2, 4 in middle). So they look unnatural than images generated by our method.

In far distance case (bottom), MineGAN just reproduces training data as in near distance case. EWC, unlike in near distance case, cannot generate images well (row 2) or generate similar images even if source images are quite different each other (row 1, 5). In all cases, our method outperforms other methods in generating higher quality and diverse images.

To show that the effectiveness of our method is not limited to real human face domain, other domain adaptation results by our method are shown in Fig. 3. a) LSUN-Church -> Van Gogh's house paintings, haunted house. b) FFHQ[17] -> Modigliani [35], Raphael. c) LSUN-Cat -> AFHQ-dog [36]. We observe that other domain adaptation results are also well-showing correspondence with source images. Especially in the case of LSUN-Cat → AFHQ-Dog, as shown in Fig. 3, LSUN-Cat is not a dataset containing only well-exposed animal faces, but the entire body or poorly exposed faces. But, even in that case, our method can capture semantic correspondence with source images while adaptation(row 1).

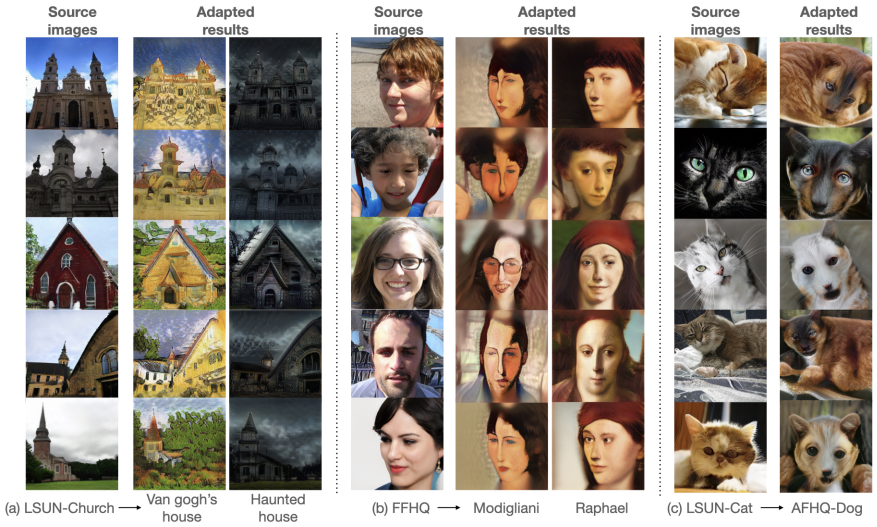

Figure 3: Other domain adaptation results by our method. target data is also 10-shot. In all above domains, generated images resemble structure of source images such as face pose, building structure depending on source domain.

Quantitative results Table 1 shows the quantitative comparison results between different methods for assessing image quality. Compared to other methods, it can be seen that our method is the most adaptable to target domains as it records the lowest FID score with the most realistic result in qualitative results. However, as pointed out by previous studies [41, 16], in a few shot settings, the over-fitting problem cannot be captured, so the LPIPS score is shown in Table 2 to assess the diversity of generated images.

	Babies	Sunglasses	Sketches
TGAN [12]	104.79	55.61	53.41
TGAN+ADA [37]	102.58	53.64	66.99
BSA [13]	140.34	76.12	69.32
FreezeD [14]	110.92	51.29	46.54
MineGAN [18]	98.23	68.91	64.34
EWC [15]	87.41	59.73	71.25
CDC [16]	74.39	42.13	45.67
Ours	$\mathbf{67.55 \pm 2.23}$	$\mathbf{36.69 \pm 2.63}$	$\mathbf{41.50 \pm 1.64}$

Table 1: FID scores (↓) for target domains with entire target data. Reported scores of baselines are referenced in [16]. Standard deviations are computed across 5 random runs.

	Babies	Sunglasses	Sketches
MineGAN [18]	0.52 ± 0.03	0.43 ± 0.04	0.40 ± 0.05
EWC [15]	$\mathbf{0.58 \pm 0.01}$	$\mathbf{0.58 \pm 0.01}$	0.42 ± 0.03
CDC [16]	0.57 ± 0.02	0.57 ± 0.02	$\mathbf{0.45 \pm 0.02}$
Ours	$\mathbf{0.58 \pm 0.02}$	0.56 ± 0.01	$\mathbf{0.45 \pm 0.03}$

Table 2: LPIPS scores(↑) for adapted results. Standard deviations is computed across the target samples(In this case 10) following in [16]

In Table 2, MineGAN scored the lowest LPIPS score for all datasets. This is because MineGAN generates target images similar to training data without any correspondence with source images. On the other hand, EWC, CDC, and our method show similar scores for all datasets. In short, according to Table 1 and Table 2, These methods can generate diverse images but, our method can generate more realistic target images than other methods.

4.2 N-shot settings

In this subsection, we explore the cases when target data is less than 10-shot. Because the number of training data is a key factor of few shot image generation, it's need to investigate how the dataset size affects the quality and diversity of the adaptation results. We set up this exploration using Face sketches[34] as target domain. Fig. 4 shows the results when target data is given on 1-shot, 5-shot, and 10-shot. In the case of 1-shot, the appearance of the images is almost the same, but only the poses and facial expressions have limited variations to a target data. In the case of 5-shot setting, generated images have distinct characteristics by appearance not limited to small changes such as pose, expression. In 10-shot setting, results show more diverse and detailed images than when there are fewer target data.

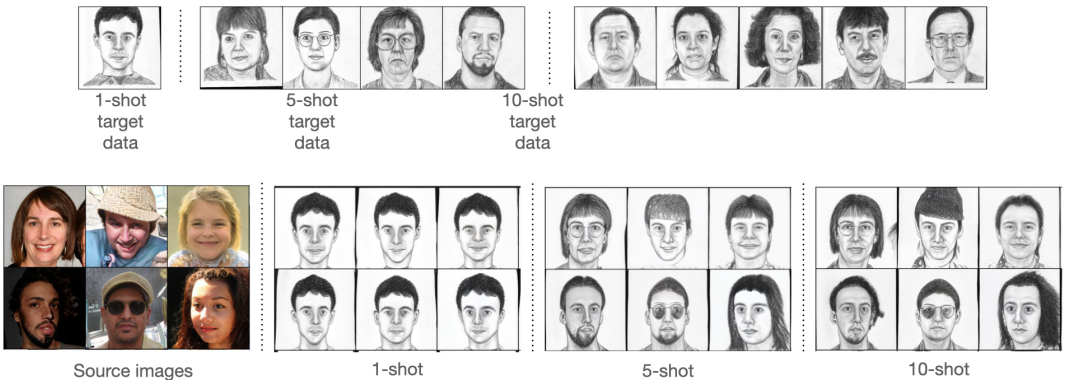

Figure 4: Adaptation results on different target data size. The larger target data size, our method can generate more diverse and detailed images. Even if given on 1-shot target data, generated images reflect weak correspondence like pose and expressions with source images.

4.3 Effect of λ_{con} value

The λ_{con} parameter controls the balance between adaptation and keeping semantic similarity with source domain. In this subsection, we explore the effect of λ_{con} using two target domains: Face paintings by Modigliani[35] and Raphael. As shown in Fig. 5, The larger the λ_{con}, the more parts like the pose, expressions, and visual features of the source images remain strong on adaptation results. On the contrary, the smaller the λ_{con}, the weaker the tendency to maintain correspondence with the source, which may cause over-fitting to target domain. It is shown in row 4, 5 in Fig. 5. Source images (row 4, 5) are quite different with each other, but adapted results are almost the same as each other. However, when λ_{con} is larger than 0, the generated images show diversity with correspondence to source images.

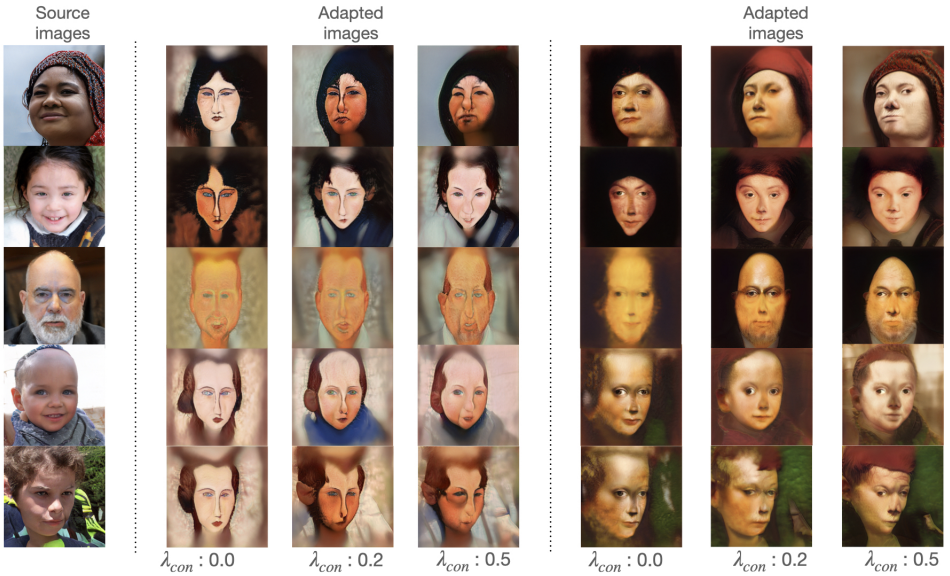

Figure 5: Effect of λ_{con} on adaptation results. The larger λ_{con}, visual features of source images remain strong on adapted images. Conversely, the smaller λ_{con}, the weaker the adaptation results reflect the correspondence with the source images.

5 Conclusion

In few-shot image generation, over-fitting to the few target data can easily happen, which hinders to generate diverse and high-quality images on target domain. To alleviate this problem, we propose $\mathcal{C}^3$ method to enforce the cross-domain correspondence directly between source and target domain in few-shot image generation. By transferring the prior knowledge of a pre-trained model while keeping cross-domain correspondence, it is possible for the adapted model to generate new images of the target domain, avoiding over-fitting. Experimental results on multiple datasets demonstrate the effectiveness of our approach. We believe that $\mathcal{C}^3$ can be seamlessly applicable to other few-shot image generation models.

Acknowledgments and Disclosure of Funding

This work was partly supported by the Institute of Information Communications Technology Planning Evaluation (2015-0-00310-SW.StarLab/20%, 2017-0-01772-VTT/20%, 2018-0-00622-RMI/20%, 2019-0-01371-BabyMind/20%, 2021-0-02068-AIHub/10%) grant funded by the Korean government and CARAI (UD190031RD/10%) grant funded by the DAPA and ADD.

References

[1] Zhang, H., T. Xu, H. Li, et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, 2017.

[2] Xu, T., P. Zhang, Q. Huang, et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks, 2017.

[3] Zhu, M., P. Pan, W. Chen, et al. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis, 2019.

[4] Yin, G., B. Liu, L. Sheng, et al. Semantics disentangling for text-to-image generation, 2019.

[5] Zhang, H., J. Y. Koh, J. Baldridge, et al. Cross-modal contrastive learning for text-to-image generation, 2021.

[6] Zhu, J.-Y., T. Park, P. Isola, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.

[7] Huang, X., M.-Y. Liu, S. Belongie, et al. Multimodal unsupervised image-to-image translation. In *ECCV*. 2018.

[8] Lee, H.-Y., H.-Y. Tseng, Q. Mao, et al. Drit++: Diverse image-to-image translation viadisentangled representations. *International Journal of Computer Vision*, pages 1–16, 2020.

[9] Härkönen, E., A. Hertzmann, J. Lehtinen, et al. Ganspace: Discovering interpretable gan controls, 2020.

[10] Shen, Y., C. Yang, X. Tang, et al. Interfacegan: Interpreting the disentangled face representation learned by gans, 2020.

[11] Patashnik, O., Z. Wu, E. Shechtman, et al. Styleclip: Text-driven manipulation of stylegan imagery, 2021.

[12] Wang, Y., C. Wu, L. Herranz, et al. Transferring gans: generating images from limited data, 2018.

[13] Noguchi, A., T. Harada. Image generation from small datasets via batch statistics adaptation, 2019.

[14] Mo, S., M. Cho, J. Shin. Freeze the discriminator: a simple baseline for fine-tuning gans, 2020.

[15] Li, Y., R. Zhang, J. Lu, et al. Few-shot image generation with elastic weight consolidation, 2020.

[16] Ojha, U., Y. Li, J. Lu, et al. Few-shot image generation via cross-domain correspondence, 2021.

[17] Karras, T., S. Laine, T. Aila. A style-based generator architecture for generative adversarial networks, 2019.

[18] Wang, Y., A. Gonzalez-Garcia, D. Berga, et al. Minegan: effective knowledge transfer from gans to target domains with few images, 2020.

[19] Clouâtre, L., M. Demers. Figr: Few-shot image generation with reptile. *CoRR*, abs/1901.02199, 2019.

[20] Liang, W., Z. Liu, C. Liu. Dawson: A domain adaptive few shot generation framework, 2020.

[21] He, K., H. Fan, Y. Wu, et al. Momentum contrast for unsupervised visual representation learning, 2020.

[22] Chen, T., S. Kornblith, M. Norouzi, et al. A simple framework for contrastive learning of visual representations, 2020.

[23] van den Oord, A., Y. Li, O. Vinyals. Representation learning with contrastive predictive coding, 2019.

[24] Chen, X., H. Fan, R. Girshick, et al. Improved baselines with momentum contrastive learning, 2020.

[25] Li, J., P. Zhou, C. Xiong, et al. Prototypical contrastive learning of unsupervised representations, 2021.

[26] Zhao, Z., Z. Zhang, T. Chen, et al. Image augmentations for gan training, 2020.

[27] Kang, M., J. Park. Contragan: Contrastive learning for conditional image generation, 2021.

[28] Ye, H., X. Yang, M. Takac, et al. Improving text-to-image synthesis using contrastive learning, 2021.

[29] Park, T., A. A. Efros, R. Zhang, et al. Contrastive learning for unpaired image-to-image translation, 2020.

[30] Lee, H., J. Seol, S. goo Lee. Contrastive learning for unsupervised image-to-image translation, 2021.

[31] Simonyan, K., A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[32] Isola, P., J.-Y. Zhu, T. Zhou, et al. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[33] Karras, T., S. Laine, M. Aittala, et al. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*. 2020.

[34] Wang, X., X. Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2009.

[35] Yaniv, J., Y. Newman, A. Shamir. The face of art: landmark detection and geometric style in portraits. *ACM Transactions on Graphics*, 38:1–15, 2019.

[36] Choi, Y., Y. Uh, J. Yoo, et al. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2020.

[37] Karras, T., M. Aittala, J. Hellsten, et al. Training generative adversarial networks with limited data, 2020.

[38] Kirkpatrick, J., R. Pascanu, N. Rabinowitz, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[39] Heusel, M., H. Ramsauer, T. Unterthiner, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

[40] Zhang, R., P. Isola, A. A. Efros, et al. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*. 2018.

[41] Robb, E., W.-S. Chu, A. Kumar, et al. Few-shot adaptation of generative adversarial networks, 2020.