Few Shot Generative Model Adaption via Relaxed Spatial Structural Alignment

Jiayu Xiao[1,2], Liang Li[1]*, Chaofei Wang[3]†, Zheng-Jun Zha[4], Qingming Huang[1,2,5]
[1]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China, [3]Department of Automation, Tsinghua University
[4]University of Science and Technology of China, China, [5]Peng Cheng Laboratory, Shenzhen, China
jiayu.xiao@vipl.ict.ac.cn, liang.li@ict.ac.cn,
wangcf18@mails.tsinghua.edu.cn, zhazj@ustc.edu.cn, qmhuang@ucas.ac.cn

Abstract

Training a generative adversarial network (GAN) with limited data has been a challenging task. A feasible solution is to start with a GAN well-trained on a large scale source domain and adapt it to the target domain with a few samples, termed as few shot generative model adaption. However, existing methods are prone to model overfitting and collapse in extremely few shot setting (less than 10). To solve this problem, we propose a relaxed spatial structural alignment (RSSA) method to calibrate the target generative models during the adaption. We design a cross-domain spatial structural consistency loss comprising the self-correlation and disturbance correlation consistency loss. It helps align the spatial structural information between the synthesis image pairs of the source and target domains. To relax the cross-domain alignment, we compress the original latent space of generative models to a subspace. Image pairs generated from the subspace are pulled closer. Qualitative and quantitative experiments show that our method consistently surpasses the state-of-the-art methods in few shot setting. Our source code: https://github.com/StevenShaw1999/RSSA.

1. Introduction

Generative adversarial networks (GANs) have achieved promising results in various computer vision scenarios such as natural image synthesis [3, 10], image to image translation [42] and image inpainting [35, 38]. Meanwhile, GANs are notoriously hard to train, and training an image generative model generally requires thousands of images and tens of hours of training time. Actually, for many real-world applications, data acquisition is difficult or expensive. For example, in the artistic domain, it is impossible to hire artists to make thousands of creations. Without enough training data, GANs are prone to overfit and collapse.

To address this issue, researchers begin to focus on effective GAN training with limited samples. Most of them follow the route of few shot generative model adaption that starting with a model pre-trained on a large dataset of a source domain, and adapting to the target domain with limited data, as shown in Fig. 1. Wang *et al*. [32] leverage the fine-tuning strategy to directly model the distribution of target domain. Some works either impose strong regularization [15, 39] or modify the network parameters with a slight perturbation [22, 25, 30] to avoid overfitting to the limited target samples. In addition, some data augmentation methods [28, 40, 41] are proposed to enlarge the amount of the training data so as to improve the robustness of generative models. However, these methods are only suitable for scenarios with more than 100 training images. When the number of training images is reduced to just a few (less than 10), the generative model usually generates images with poor quality and suffers from early collapse.

Recently, as the pioneer work, intuited by the contrastive learning, Ojha *et al*. [23] proposed to preserve the relative

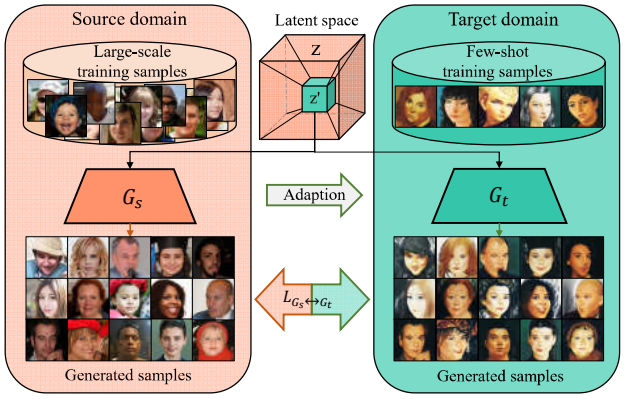

Figure 1. Few shot generative model adaption and our motivation. Start with a pre-trained model of the source domain G_s and adapt to the target domain to get G_t by using extremely few (such as 5) training images of the target domain. We compress the latent space to a subspace close to target domain, and align the spatial structural information of synthesis pairs generated by G_s and G_t.

*Corresponding author.
†Equal contribution.

similarities and differences between instances in the source domain via an instance distance consistency loss (IDC for short). Given only 10 images, this method can generate more diverse and realistic images for the target domain. Although IDC has made great strides, the generated images still undergo identity degradation and unnatural distortions or textures. The main reason is that IDC can not guarantee the inherent structure of each image, leading to the drift of the samples in the space of target domain (see Sec. 3.2 for a detailed discussion).

In this work, we propose a relaxed spatial structural alignment (RSSA) method to cope with the few shot generative model adaption task. It leverages richer spatial structure priors of images from source domain to address the identity degradation problem of the generative model. Specifically, we design a cross-domain spatial structural consistency loss, which consists of *self-correlation consistency loss* and *disturbance correlation consistency loss*. The former helps align the self-correlation information of feature maps of the synthesis pairs generated by the source and target generators, so as to constrain the cross-domain consistency of inherent structural information. The latter helps align the spatial mutual correlation between samples adjacent to each other in the latent space, in order to constrain the cross-domain consistency of variation tendency of a specific instance.

With the help of the cross-domain spatial structural consistency loss, the samples from the target generator maintain original self-correlation and disturbance correlation properties inherited from the source domain during adaption. However, straightforward alignment may result in the dominant of the attributes from the source domain in the optimization phase, and slow down the model convergence. Thus, we propose to compress the latent space into a subspace which is close to the target domain. This can relax the above alignment because synthesis pairs generated from the subspace get closer to each other.

To better evaluate the few shot generative model adaption methods, besides the traditional quantitative metric and qualitative visualization, we design a structural consistency score (SCS) which measures the structural similarities of synthesis pairs from the source and target domains. Moreover, compared with Inception Score (IS [27]) or Fréchet Inception Distance (FID [8]), SCS can better reflect image identity preservation in few shot adaption.

The main contributions are summarized as follows:

- We propose RSSA, a relaxed spatial structural alignment method, to transfer rich spatial structural information of the large-scale source domain to the few shot target domain with better identity preservation.
- We introduce the latent space compression to relax the cross-domain alignment via pulling synthesis pairs generated from the compressed subspace closer to each other, and accelerate the training procedure.
- We design a metric to evaluate the quality of synthesis images from the structural perspective, which can serve as an alternative supplement to the current metrics. Qualitatively and quantitatively, our method outperforms existing competitors in a variety of settings.

2. Related Work

2.1. Few shot image generation

Few-shot image generation aims to generate diversified and high-quality images in a new domain with a small amount of training data. The most straightforward approach is to fine-tune a pre-trained GAN [2, 4, 16, 32]. However, fine-tuning the entire network weights often leads to poor results. Researchers proposed to modify part of the network weights [21, 25] or batch statistics [22], and besides leverage different forms of regularization [15, 39] to avoid overfitting. Wang *et al*. [30] introduced a miner network to steer the sampling of the latent distribution to the target distribution. Data augmentation strategies [28, 40, 41] were introduced to enlarge the amount of the training data to improve the robustness of the generative model. However, most of them fail in the extremely few shot setting (less than 10 images). Recently, Ojha *et al*. [23] proposed to preserve the relative similarities and differences between instances in the source domain via an instance distance consistency loss. Different from work [23], we explore align the distributions of the source and target domains from the perspective of spatial structural consistency, and solve the problems of identity degradation and image distortion during model adaption.

2.2. Image to image translation

The goal of the image to image translation is to convert an input image from a source domain to a target domain with the intrinsic source content preserved and the extrinsic target style transferred [24]. Variational autoencoders (VAEs) and GANs are most commonly used and efficient deep generative models in the image-to-image translation tasks [6, 9, 12, 17, 19, 20]. However, most methods require a large amount of training data for both source and target domains. Furthermore, generally they are not suitable for the few-shot scenarios. Recent works [18, 26, 31] have begun to address this issue via learning to separate the content and style factors, but require a large amount of labeled data (class or style labels). Different from image to image translation, we explore model-level adaption to the few shot target domain rather than image-level translation. In addition, we do not rely on additional labeled data.

3. Method

In this section, we first overview the definition of few shot generative model adaption, and propose the framework of our method. Then, the two key components of RSSA, cross-domain spatial structural consistency loss and latent space

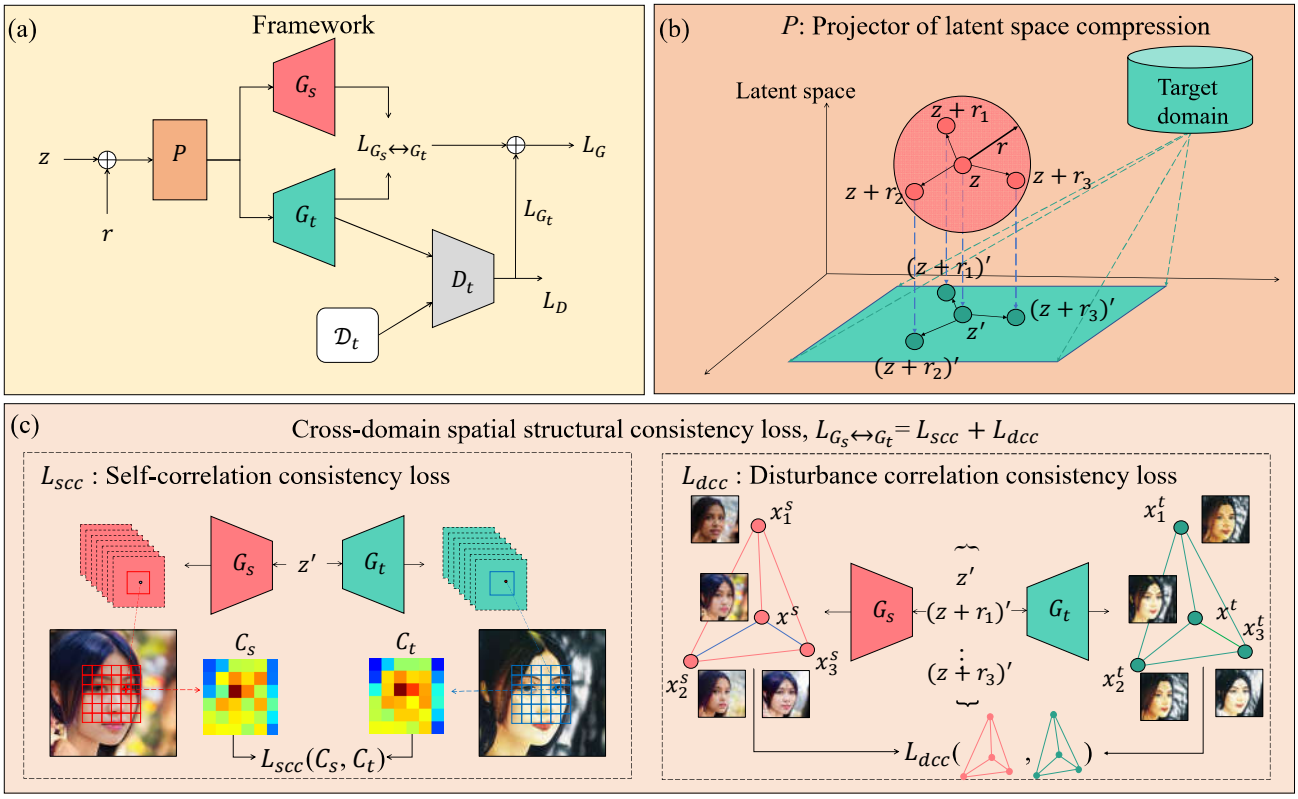

Figure 2. The framework of our method. (a) an overview. r is a small disturbance corresponding to z. P is a projector of compressing the latent space as shown in (b). $L_{G_s \leftrightarrow G_t}$ is a cross-domain spatial structural consistency loss, consisting of self-correlation consistency and disturbance correlation consistency loss, as shown in (c).

compression are interpreted in detail in Sec. 3.2 and 3.3, the optimization strategy is described in Sec. 3.4. Finally, a novel metric of structural consistency score is proposed to better evaluate few shot generation methods in Sec. 3.5.

3.1. Overview and framework

We have a generator G_s pre-trained on a large-scale dataset in the source domain $\mathcal{D}_s$. It can be considered as a function that maps a noise vector z sampled from the d-dimentional latent space $z \sim p(z) \subset \mathbb{R}^d$ to a generated image $G_s(z)$ in the pixel space. The goal of few shot adaption is to adapt G_s from the source domain to the target domain and obtain G_t, using a few samples in the target domain. A standard fine-tuning approach is done by initializing G_t with G_s and fine-tuning G_t on the dataset in the target domain $\mathcal{D}_t$ with an adversarial training procedure. The optimization objectives are as below:

$$L_G = -\mathbb{E}_{z \sim p(z)}[\log(D(G_t(z)))] \quad (1)$$

$$L_D = \mathbb{E}_{x \sim \mathcal{D}_t}[\log(1 - D(x)] + \mathbb{E}_{z \sim p(z)}[\log(D(G_t(z)))], \quad (2)$$

where D represents a learnable discriminator.

Most of fine-tuning methods are easy to overfit in the extremely few shot setting, because the discriminator can memorize the few examples and force the generator to reproduce them. To solve this problem, we consider two strategies. One is preserving the useful structure priors of images from the source domain to restrict the generated images, so as to avoid identity degradation during adaption. The other is compressing the latent space to a subspace where synthesis images of the source and target domains are pulled closer to each other to relax the structural constraint. Fig. 2(a) shows our pipeline. During the model adaption, we conduct relaxed spatial structural alignment. A cross-domain spatial structural consistency loss $L_{G_s \leftrightarrow G_t}$ (Fig. 2(c)) and a projector P of latent space compression (Fig. 2(b)) are introduced.

3.2. Cross-domain spatial structural consistency loss

IDC [23] preserves the relative distances between instances in the source domain and achieves the state-of-the-art (SOTA) performance in few shot setting. However, the structure of generated images distorts, leading to the problem of identity degradation. Abundant evidences can be found in Fig. 5, 6 and 7. The main reason is that IDC can not guarantee the inherent structure of each image, leading to the drift of the generated samples in the target domain space. As shown in Fig. 3(a), the generated images of the target domain (the green points) maintain the correct instance distances, but deviate from their correct positions (the red points). To avoid this drift, we propose a cross-domain spatial structural consistency loss, which preserves the inherent spatial struc-

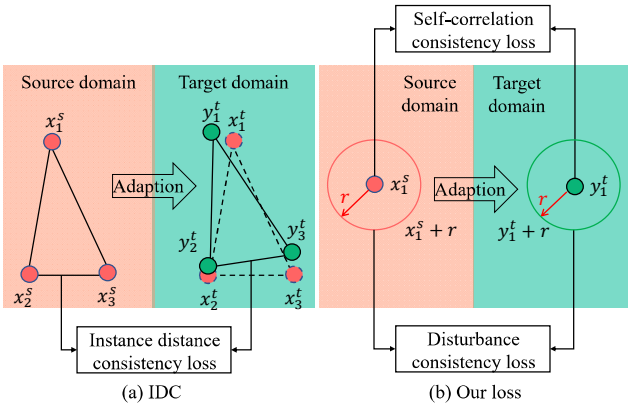

Figure 3. Illustration of the spatial structural alignment. (a) instance distance consistency loss from IDC [23]. (b) our cross-domain spatial structural consistency loss.

ture and the variation tendency of images from the source domain as shown in Fig. 3(b). Specifically, we design a self-correlation consistency loss to constrain the inherent structure of the images, and a disturbance correlation consistency loss to constrain the variation tendency under a certain disturbance of the images.

Self-correlation consistency loss. We are inspired by some relevancy mining methods [13, 14, 19] and adopt the self-correlation matrices of feature maps at each each convolutional layer to formulate the inherent structure information of the image. Each pair of self-correlation matrices from the same layers of G_s and G_t are constrained with the smooth-ℓ_1 loss [7], so as to ensure that the images in source and target domains have similar inherent structure. Define $f^l \in \mathbb{R}^{c\times w\times h}$ as the feature maps at the l^{th} layer. $f^l(x,y)$ is a c dimensional vector. Each entry of the self-correlation matrix $C^l_{x,y} \in \mathbb{R}^{w\times h}$ of the position (x,y) at the l^{th} layer can be calculated as below:

$$C^l_{x,y}(i,j) = \cos(f^l(x,y), f^l(i,j)), \tag{3}$$

where $\cos(\cdot)$ denotes the cosine similarity function, (i,j) is the corresponding position in f^l. The spatial self-correlation consistency loss between G_s and G_t can be calculated as

$$L_{\text{scc}}(G_t, G_s) = \mathbb{E}_{z_i\sim p(z)} \sum_l \sum_{x,y} \text{smooth-}\ell_1(C^{t,l}_{x,y}, C^{s,l}_{x,y}), \tag{4}$$

where $C^{t,l}_{x,y}$ and $C^{s,l}_{x,y}$ indicate the self-correlation matrices of the position (x,y) at the l^{th} layer for G_t and G_s.

Note that computation of the self-correlation matrices is an $O((w\cdot h)^2)$ operation. For feature maps with high resolution, we first aggregate the adjacent feature vectors by adopting average pooling and break the whole feature map into patches to compute local self-correlation matrices.

Disturbance correlation consistency loss. The latent space of a generative model is continuous rather than discrete, hence we propose to model the variation tendency under certain disturbances of the images(essentially the gradient information around each instance). Specifically, we take an input noise vector as an anchor point, and then sample a batch of vectors from a small neighborhood of this anchor point. The spatial similarities between these samples are calculated and transferred from the source domain to the target domain.

For an input noise z_i, define a neighborhood with radius r, $U(z_i, r) = \{z | |z - z_i| < r\}$. We sample N noise vectors from $U(z_i, r)$ and form a batch of $N+1$ vectors $\{z_n\}_1^{N+1}$ to represent the neighborhood. Define D^l_{jk} as the pixel-wise spatial mutual correlation for the l^{th} layer feature map f^l between any two samples z_j and z_k from $\{z_n\}_1^{N+1}$. D^l_{jk} at position (x,y) is denoted by the softmax of similarities between feature vector at (x,y) in f^l_j and a small corresponding region $Q = \{(m,n) | x - \frac{\delta}{2} < m < x + \frac{\delta}{2}, y - \frac{\delta}{2} < n < y + \frac{\delta}{2}\}$, where δ is the width of a slide window, in f^l_k as below:

$$D^l_{jk}(x,y) = \text{Softmax}(\{\cos(f^l_j(x,y), f^l_k(m,n))\}_{(m,n)\in Q}), \tag{5}$$

where $\cos(\cdot)$ denotes the cosine similarity function.

On basis of the computed pixel-wise correlation distribution, we impose the disturbance correlation consistency constraint by minimizing the L1 distance:

$$L_{\text{dcc}}(G_t, G_s) = \mathbb{E}_{z_i\sim p(z)} \sum_{l,j,k,x,y} \left\| D^{t,l}_{jk}(x,y) - D^{s,l}_{jk}(x,y) \right\|_1, \tag{6}$$

The spatial structural consistency loss $L_{G_s\leftrightarrow G_t}$ is composed of the self-correlation consistency loss L_{scc} and disturbance correlation consistency loss L_{dcc}. It is calculated as below:

$$L_{G_s\leftrightarrow G_t} = \alpha L_{\text{scc}} + \beta L_{\text{dcc}}. \tag{7}$$

where α, β are ratio parameters.

3.3. Latent space compression

The spatial structural consistency loss helps align the generated images of the source and target domains. However, straightforward alignment may cause the dominant of the attributes from the source domain, and thus slow down model adaption. Therefore, we propose to compress the latent space to a subspace close to the target domain to relax the cross-domain alignment. Specifically, we first invert the few samples from target domain $\{x^t_i\}_{i=1}^n$ to the W^+ space of G_s by utilizing Image2StyleGAN [1]. Given n target samples, an inverted latent code set at l^{th} layer is denoted as $\{w^l_i\}_{i=1}^n$. Define a n column matrix A^l which is composed of $\{w^l_i\}_{i=1}^n$ by $A^l_{*i} = w^l_i$, and hence we obtain a subspace $\mathcal{X}^l$ of the l^{th} latent space, where $\mathcal{X}^l$ is equivalent to the column space of A^l. Given an input noise z_j, the corresponding latent code in the l^{th} layer is w^l_j, the corresponding modulation coefficient is α^l, we modulate w^l_j and project it onto the l^{th} sub-plane $\mathcal{X}^l$ via least square method:

$$\begin{aligned} \overline{w^l_j} &= A^l({A^l}^\top A^l)^{-1}{A^l}^\top w^l_j \\ \hat{w^l_j} &= \alpha^l \overline{w^l_j} \frac{\left\| w^l_j \right\|}{\left\| \overline{w^l_j} \right\|} + (1-\alpha^l) w^l_j, \end{aligned} \tag{8}$$

Figure 4. Generated images of G_s with input latent codes sampled from different spaces. Top: original latent space. Bottom: compressed subspace. Setting: Flickr-Faces $\rightarrow$ Sketches.

where $\hat{w_j^l}$ is the projected code at l^{th} for z_j.

In this way, we compress the original latent space into a narrow subspace close to the target domain. Images generated by G_s with the latent codes sampled from the compressed subspace imply some characteristics of the target domain. As shown in Fig. 4, generated images (bottom row) shows some characteristics of sketch on the texture and color. By sampling latent codes from the compressed subspace before the alignment, we are capable to stabilize and accelerate the whole training procedure. Note that latent codes of different layers modulate the output images at distinct semantic levels in StyleGAN [10] architecture. We set large α_i at the top layers and small ones at the bottom layers to alter the generated images' attributes at high semantic levels while maintaining the original spatial structure to a great extent.

3.4. Optimization

We follow an adversarial optimization procedure. The objective of G_t is a combination of L_G (Eq.1) and $L_{G_s \leftrightarrow G_t}$ (Eq.7). We simply set the hyper-parameters α and β as 1 for all experiments. The objective of D is the same with [23] by utilizing a combination of image-wise and patch-wise discriminant loss. Standard path regularization loss to G_t and gradient penalization loss to D are also adopted at every several iterations.

3.5. Evaluation metric

In addition to the intuitively visual evaluation, IS [27] and FID [8] are the most widely used quantitative evaluation metrics for image generative models. However, both of them map the generated images to the feature space by an Inception network, which can not quantify the quality of the spatial structure of the generated images. Meanwhile, in few shot setting, many fine-tuning based methods are inclined to simply synthesis images similar with the training samples given arbitrary input noises. Yet, this may obtain high IS in some cases which is counter-intuitive, see Table 1. Furthermore, computing FID requires a large number of realistic images of the target domain, which is impractical in the few shot setting. Therefore, we adopt IS as a general evaluation metric and propose a novel spatial structure evaluation metric, termed structural consistency score, to cover the shortage of IS.

Structural Consistency Score (SCS). For a image pair $\langle x^s, x^t \rangle$ generated by $\langle G_s, G_t \rangle$ with the same input z_i, we claim that x^t preserves structural consistency when it can be easily recognized as a derivative sample from x^s. Inspired by [36], we extract a meaningful edge map of one image to represent its structural information by HED [33]. The SCS of a generated image of the target domain x^t is computed with the dice similarity coefficient [5] between the edge maps of x^t and x^s. The formalization is as below:

$$\text{SCS}(x^t) = \frac{2|H(x^t) \cap H(x^s)|}{|H(x^t)| + |H(x^s)|}, \tag{9}$$

where $H(\cdot)$ denotes HED function. $|H(x^t) \cap H(x^s)|$ is calculated by the pixel-wise inner product of $H(x^t)$ and $H(x^t)$. $|H(x^t)|$ and $|H(x^s)|$ is calculated by the sum of squares of matrix elements. Then, the SCS of the target GAN G_t can be quantified into the expectation as below:

$$\text{SCS}(G_t) = \mathbb{E}_{z_i \sim p(z)}[\frac{2|H(G_t(z_i)) \cap H(G_s(z_i))|}{|H(G_t(z_i))| + |H(G_s(z_i))|}]. \tag{10}$$

Higher SCS means better spatial structural consistency between G_t and G_s. It is remarkable that the SCS of an overfitted model will be very low, because it can not generate images with structures similar to the source domain images.

4. Experiments

In this section, we demonstrate the effectiveness of RSSA in few shot setting. Qualitative and quantitative comparisons between our method and several baselines, TGAN [32], FreezeD [21], MineGAN [30], IDC [23]. As the SOTA method, IDC [23] is our primary comparison method in most experiments.

We adopt the StyleGANv2 [11] pre-trained on three different datasets: (i) Flickr-Faces-HQ (FFHQ) [10], (ii) LSUN Churches [37], (iii) LSUN Cars [37]. We adapt the source GANs to various target domains including: (i) face sketches [29], (ii) face paintings by Van Gogh [34], (iii) face paintings by Moise Kisling [34], (iv) haunted houses [23], (v) village painting by Van Gogh [23], (vi) wrecked/abandoned cars [23]. Model adaptions are done in 10-shot, 5-shot and 1-shot settings.

4.1. Performance evaluation

Qualitative comparison. Fig. 5 shows results on FFHQ $\rightarrow$ sketches using different adaption methods. We can observe that TGAN overfits strongly to the samples of the target domain. Compared with TGAN, FreezeD and MineGAN do not improve the results. This indicates that although they could play a positive role in a small dataset with more than

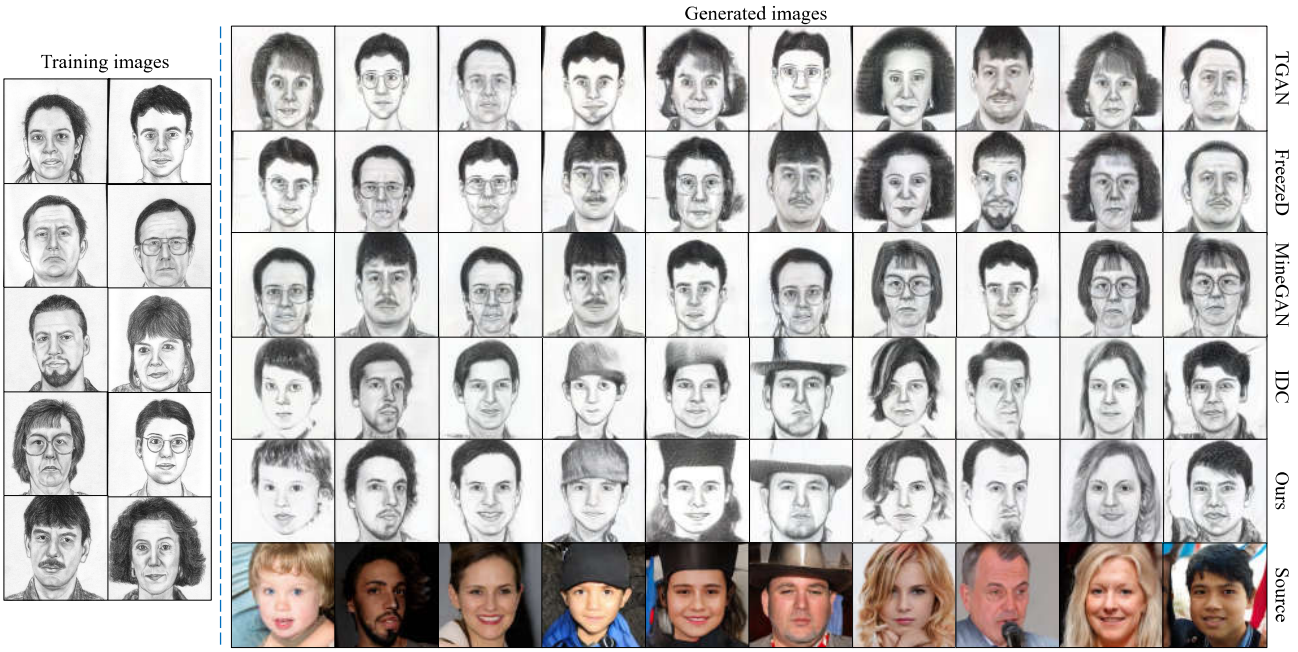

Figure 5. Comparison results with different methods on Flickr-Faces → Sketches (10-shot).

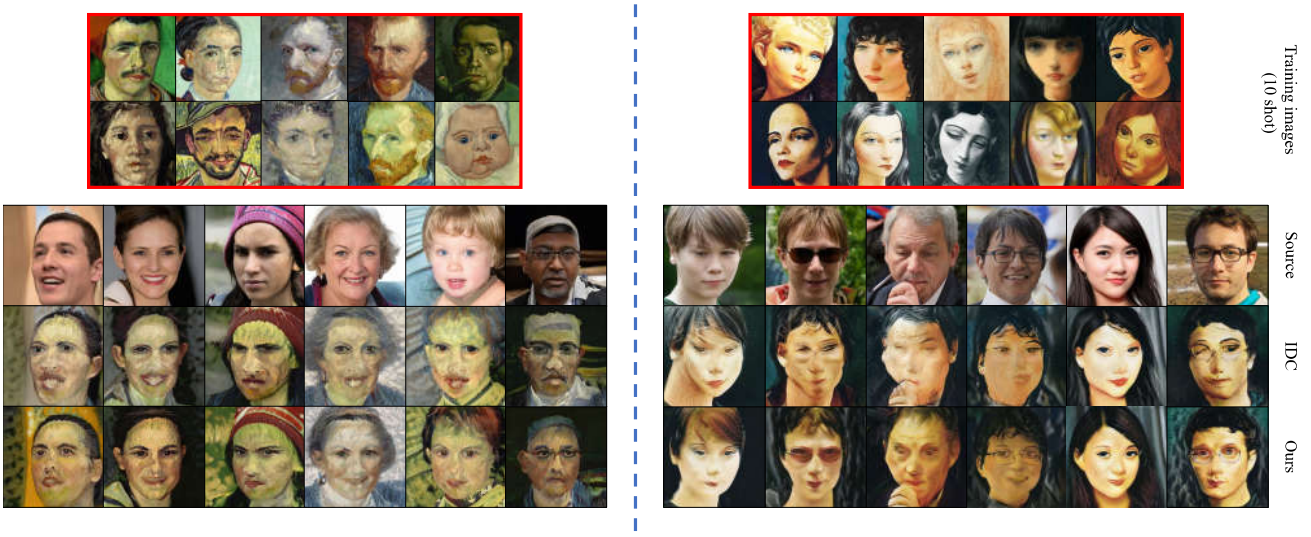

Figure 6. Comparison results between IDC and our method on Flickr-Faces → Vincent van Gogh (left), Moise Kisling (right) (10-shot).

100 training samples, they would be ineffective in extremely few shot setting (less than 10). By contrast, IDC improves the correspondence between the source domain and the target domain, and shows similar visual patterns between the synthesis pairs. Further, as for images synthesised by RSSA, one can easily recognize the corresponding source domain images with only few glances. This is because our method acquires visual attributes from the target domain and meanwhile greatly preserves the spatial structural information of images from the source domain.

In order to comprehensively compare IDC with our method, we extend comparison experiments to multiple target domains with different few shot setting as shown in Fig. 6 (10-shot) and Fig. 7 (5-shot). We can observe the distorted attributes of human faces (Fig. 6) and texture degradation of churches (Fig. 7) in IDC's results, but there are almost no similar phenomenons in our results. In addition, we take a bold stab at 1-shot scenarios as shown in Fig. 8, and obtain some decent results. Good visual results are mainly attributed to the cooperation of latent space compression and spatial structure alignment, the former helps acquire attributes from the target domain faster, the latter helps preserve the structural knowledge from the source domain.

Quantitative comparison. To quantify the quality and di-

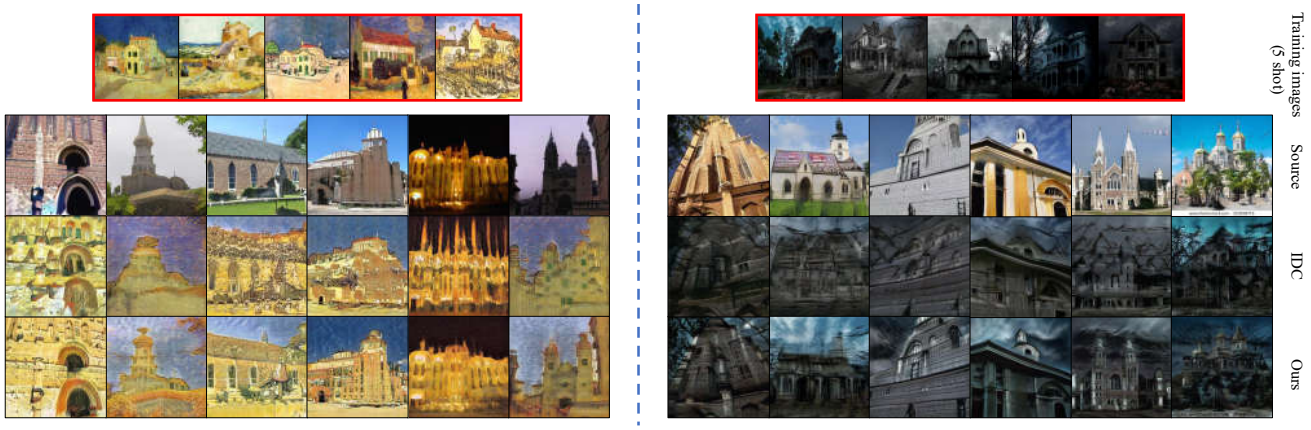

Figure 7. Comparison results between IDC and our method on Churches → Van Gogh Village (left), Haunted Houses (right) (5-shot).

Metric	Method	f→S		f→V		f→M		c→V		c→H	
		10-shot	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot	5-shot
IS	TGAN	$2.00^{\pm 0.04}$	$2.09^{\pm 0.05}$	$\mathbf{2.69}^{\pm 0.18}$	$1.65^{\pm 0.07}$	$1.88^{\pm 0.08}$	$1.98^{\pm 0.04}$	$2.82^{\pm 0.09}$	$2.28^{\pm 0.05}$	$5.16^{\pm 0.19}$	$4.47^{\pm 0.14}$
	FreezeD	$1.98^{\pm 0.05}$	$2.05^{\pm 0.06}$	$2.46^{\pm 0.06}$	$\mathbf{1.77}^{\pm 0.03}$	$1.91^{\pm 0.06}$	$1.85^{\pm 0.04}$	$2.78^{\pm 0.07}$	$2.32^{\pm 0.09}$	$5.29^{\pm 0.15}$	$4.60^{\pm 0.17}$
	MineGAN	$1.92^{\pm 0.03}$	$2.12^{\pm 0.07}$	$2.52^{\pm 0.09}$	$1.71^{\pm 0.04}$	$1.84^{\pm 0.03}$	$1.87^{\pm 0.08}$	$2.51^{\pm 0.05}$	$2.18^{\pm 0.08}$	$5.22^{\pm 0.11}$	$4.55^{\pm 0.18}$
	IDC	$1.95^{\pm 0.02}$	$2.15^{\pm 0.04}$	$1.55^{\pm 0.04}$	$1.52^{\pm 0.02}$	$1.97^{\pm 0.03}$	$1.76^{\pm 0.06}$	$2.86^{\pm 0.11}$	$2.78^{\pm 0.06}$	$5.44^{\pm 0.21}$	$5.15^{\pm 0.13}$
	Ours	$\mathbf{2.10}^{\pm 0.03}$	$\mathbf{2.41}^{\pm 0.03}$	$1.77^{\pm 0.06}$	$1.61^{\pm 0.04}$	$\mathbf{2.17}^{\pm 0.07}$	$\mathbf{2.09}^{\pm 0.05}$	$\mathbf{3.54}^{\pm 0.10}$	$\mathbf{3.62}^{\pm 0.13}$	$\mathbf{6.26}^{\pm 0.18}$	$\mathbf{5.70}^{\pm 0.20}$
SCS	TGAN	0.287	0.285	0.380	0.375	0.344	0.350	0.355	0.343	0.211	0.218
	FreezeD	0.289	0.288	0.384	0.372	0.347	0.346	0.353	0.349	0.216	0.212
	MineGAN	0.294	0.290	0.340	0.332	0.339	0.341	0.375	0.350	0.211	0.214
	IDC	0.437	0.422	0.594	0.568	0.524	0.494	0.551	0.535	0.490	0.424
	Ours	**0.511**	**0.507**	**0.702**	**0.685**	**0.644**	**0.618**	**0.702**	**0.681**	**0.573**	**0.582**

Table 1. Quantitative evaluation of methods by IS and SCS. f and c represent faces and churches source domains. S,V,M,H represent four target domains: sketches, Van Gogh's paintings, Moise Kisling's paintings and haunted houses. Best results are **bold**.

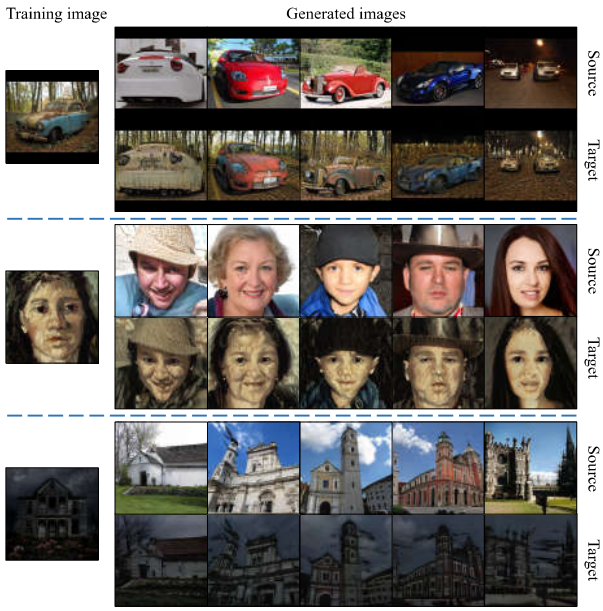

Figure 8. Results of our method on Cars → Wrecked Cars, Faces → Vincent van Gogh, and Churches → Haunted houses (1-shot).

versity of the synthesis images, we evaluate all methods with IS and SCS. All quantitative experiments are conducted in 10-shot and 5-shot settings. For IS, we calculate means and variances over 10 runs on 10000 randomly sampled images. As shown in Table 1, our method achieves best scores in most cases due to the high diversity and quality of synthesis images. However, TGAN, FreezeD and MineGAN outperforms IDC and RSSA on face→Van Gogh's paintings. The reason is that they simply overfit to the few training samples, while images generated by IDC and RSSA tend to remix the textures and colors of the paintings. This indicates that IS sometimes fails to handle the overfitting problem in few shot setting. For SCS, we randomly sample 500 noise vectors as inputs of G_s and G_t, then form the synthesis pairs and calculate their mean score. As shown in Table 1, TGAN, FreezeD and MineGAN overfit to the training samples and obtain lower results for all settings. IDC performs much better by preserving the distances of instances. RSSA consistently surpasses all the comparison methods by a large margin due to the better spatial structure preservation. To visually understand SCS, Fig. 9 shows two groups of edge extraction examples of on FFHQ → sketches and Churches → Van Gogh Village. Obviously, our generated images retain more accurate edge information than those from IDC, thus obtain higher SCS.

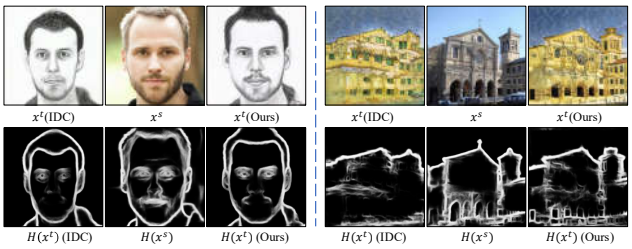

Figure 9. Comparison of edge maps between IDC and our method. Top row shows generated images, bottom raw shows corresponding edge maps. Left: IDC. Middle: Source. Right: Ours

4.2. Ablation study

Effect of the spatial structural consistency loss. We conduct ablation experiments to verify the effectiveness of the two components of our proposed spatial structural consistency loss. As shown in Fig. 10, L_{scc} and L_{dcc} greatly improve the visual quality of synthesis images separately and get the best visual results in cooperation. Consistently, quantitative conclusion can be drawn in Table 2. The reason is that they introduce inherent and neighbouring structural constraint separately and share good compatibility.

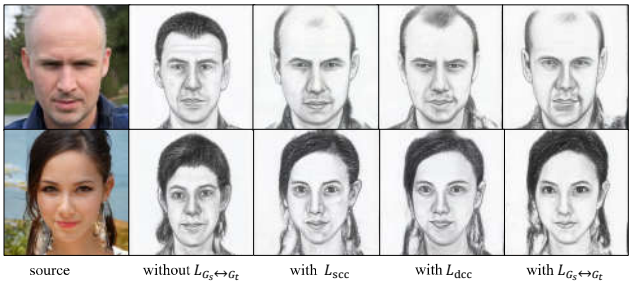

Figure 10. Qualitative ablation of the loss function on FFHQ → sketches in 10-shot setting.

Method	IS		SCS	
	f→V	c→H	f→V	c→H
without $L_{G_s \leftrightarrow G_t}$	$1.40^{\pm 0.03}$	$5.38^{\pm 0.09}$	0.449	0.369
with L_{scc}	$1.69^{\pm 0.05}$	$6.18^{\pm 0.17}$	0.684	0.544
with L_{dcc}	$1.64^{\pm 0.05}$	$6.13^{\pm 0.14}$	0.671	0.548
with $L_{G_s \leftrightarrow G_t}$	$\mathbf{1.77}^{\pm 0.06}$	$\mathbf{6.26}^{\pm 0.18}$	**0.702**	**0.573**

Table 2. Quantitative ablation of the loss function in 10-shot setting. f and c represent faces and churches source domains, V and H represent Van Gogh's paintings and haunted houses target domains.

Effect of the latent space compression. We conduct the model adaption on FFHQ → sketches in 10-shot setting. As shown in Fig. 11, we observe that some target-domain irrelevant attributes (e.g. background and color) degrade faster when adopting latent space compression. This demonstrates that (1) the latent space compression copes with the dominant of attributes from source domain well so as to relax the spatial structural alignment; (2) it helps speed up the adaption procedure of G_t.

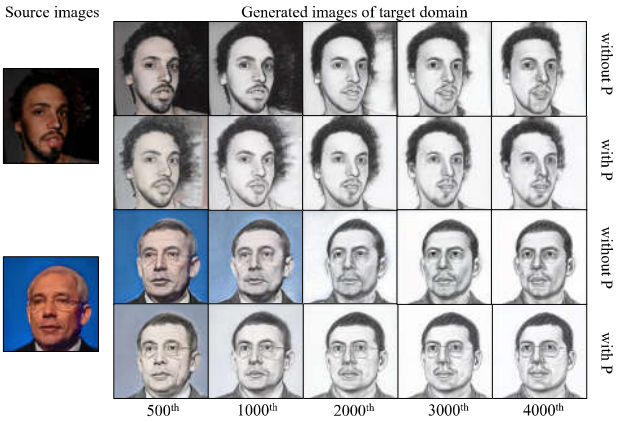

Figure 11. Ablation of the latent space compression, which shows the output images from the same input noises at different training iterations. P denotes the projector of compression.

5. Conclusion and Limitations

In this paper, we propose a novel few shot generative model adaption method, relaxed spatial structural alignment (RSSA). By aligning the generative distributions of the source and target domains via a cross-domain spatial structural consistency loss, the inherent structure information and spatial variation tendency of images from the source domain can be well preserved and transferred to the target domain. The original latent space is compressed to a narrow subspace close to the target domain, which relaxes the cross-domain alignment and accelerates the convergence rate of the target domain generator. In addition, we design a novel metric, SCS, to assess the structural quality of generated images. It may serve as an alternative supplement to the current metrics in the few shot generation scenarios.

Although our method can deal with the settings of extremely few shot training samples well and generate compelling visual results, it also has some limitations. The spatial structural consistency loss is not friendly to some abstract domains, such as paintings from Amedeo Modigliani, who is known for portraits in a modern style characterized by a surreal elongation of faces. Nevertheless, we believe that more data-efficient generative models will be proposed in the near future. The application of these models will in turn facilitate a wide range of downstream tasks such as few shot image classification.

Acknowledgement. This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102000, and in part by the National Natural Science Foundation of China: U21B2038, 61931008, 61732007, and CAAI-Huawei MindSpore Open Fund, Youth Innovation Promotion Association of CAS under Grant 2020108, CCF-Baidu Open Fund.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, pages 4432–4441, 2019. 4

[2] Sergey Bartunov and Dmitry Vetrov. Few-shot generative modelling with generative matching networks. In *AISTATS*, pages 670–678. PMLR, 2018. 2

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2018. 1

[4] Louis Clouâtre and Marc Demers. Figr: Few-shot image generation with reptile. *arXiv preprint arXiv:1901.02199*, 2019. 2

[5] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 5

[6] Yuke Fang, Weihong Deng, Junping Du, and Jiani Hu. Identity-aware cyclegan for face photo-sketch synthesis and recognition. *Pattern Recognition*, 102:107249, 2020. 2

[7] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 4

[8] 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. *NeurIPS*, 30, 2017. 2, 5

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 2

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1, 5

[11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 5

[12] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, pages 35–51, 2018. 2

[13] Liang Li, Xingyu Gao, Jincan Deng, Yunbin Tu, Zheng-Jun Zha, and Qingming Huang. Long short-term relation transformer with global gating for video captioning. *TIP*, 2022. 4

[14] Liang Li, Chenggang Clarence Yan, Xing Chen, Chunjie Zhang, Jian Yin, Baochen Jiang, and Qingming Huang. Distributed image understanding with semantic dictionary and semantic expansion. *Neurocomputing*, 174:384–392, 2016. 4

[15] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*, 2020. 1, 2

[16] Weixin Liang, Zixuan Liu, and Can Liu. Dawson: A domain adaptive few shot generation framework. *arXiv preprint arXiv:2001.00576*, 2020. 2

[17] Jianxin Lin, Yingxue Pang, Yingce Xia, Zhibo Chen, and Jiebo Luo. Tuigan: Learning versatile image-to-image translation with two unpaired images. In *ECCV*, pages 18–35. Springer, 2020. 2

[18] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *ICCV*, pages 10551–10560, 2019. 2

[19] Zhenahuan Liu, Liang Li, Huajie Jiang, Xin Jin, Dandan Tu, Shuhui Wang, and Zheng-Jun Zha. Unsupervised coherent video cartoonization with perceptual motion consistency. In *AAAI*, 2022. 2, 4

[20] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. In *ICLR*, 2018. 2

[21] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020. 2, 5

[22] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *ICCV*, pages 2750–2758, 2019. 1, 2

[23] 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 *CVPR*, pages 10743–10752, 2021. 1, 2, 3, 4, 5

[24] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *arXiv preprint arXiv:2101.08629*, 2021. 2

[25] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-shot adaptation of generative adversarial networks. *arXiv preprint arXiv:2010.11943*, 2020. 1, 2

[26] Kuniaki Saito, Kate Saenko, and Ming-Yu Liu. Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder. In *ECCV*, pages 382–398. Springer, 2020. 2

[27] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 29:2234–2242, 2016. 2, 5

[28] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. On data augmentation for gan training. *TIP*, 30:1882–1897, 2021. 1, 2

[29] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *PAMI*, 31(11):1955–1967, 2008. 5

[30] 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 *CVPR*, pages 9332–9341, 2020. 1, 2, 5

[31] Yaxing Wang, Salman Khan, Abel Gonzalez-Garcia, Joost van de Weijer, and Fahad Shahbaz Khan. Semi-supervised learning for few-shot image-to-image translation. In *CVPR*, pages 4453–4462, 2020. 2

[32] 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*, pages 218–234, 2018. 1, 2, 5

[33] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, pages 1395–1403, 2015. 5

[34] Jordan Yaniv, Yael Newman, and Ariel Shamir. The face of art: landmark detection and geometric style in portraits. *TOG*, 38(4):1–15, 2019. 5

[35] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *CVPR*, pages 5485–5493, 2017. 1

[36] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Unpaired portrait drawing generation via asymmetric cycle mapping. In *CVPR*, pages 8217–8225, 2020. 5

[37] 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. 5

[38] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, pages 5505–5514, 2018. 1

[39] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *ICLR*, 2019. 1, 2

[40] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *NeurIPS*, 2020. 1, 2

[41] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for gan training. *arXiv preprint arXiv:2006.02595*, 2020. 1, 2

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 1