Smoothness Similarity Regularization for Few-Shot GAN Adaptation

Vadim Sushko[1,2] Ruyu Wang[1] Juergen Gall[2,3]
[1]Bosch Center for Artificial Intelligence [2]University of Bonn
[3]Lamarr Institute for Machine Learning and Artificial Intelligence
vad221@gmail.com ruyu.wang@de.bosch.com gall@iai.uni-bonn.de

Abstract

The task of few-shot GAN adaptation aims to adapt a pre-trained GAN model to a small dataset with very few training images. While existing methods perform well when the dataset for pre-training is structurally similar to the target dataset, the approaches suffer from training instabilities or memorization issues when the objects in the two domains have a very different structure. To mitigate this limitation, we propose a new smoothness similarity regularization that transfers the inherently learned smoothness of the pre-trained GAN to the few-shot target domain even if the two domains are very different. We evaluate our approach by adapting an unconditional and a class-conditional GAN to diverse few-shot target domains. Our proposed method significantly outperforms prior few-shot GAN adaptation methods in the challenging case of structurally dissimilar source-target domains, while performing on par with the state of the art for similar source-target domains.

1. Introduction

Generative adversarial networks (GANs) have been shown to be powerful at various image synthesis tasks [4, 28, 3, 13, 27, 26]. The success of these models is in large part enabled by the availability of large datasets for training, typically consisting of thousands of images. However, there are many applications and computer vision tasks such as one-shot or few-shot learning [1, 33], out-of-distribution detection [24], or long-tailed recognition tasks [8] where the number of available training images is very low.

Since training a GAN from scratch on very few samples does not perform well as shown in Fig. 1, a common strategy is to fine-tune a pre-trained GAN model on the few-shot dataset, typically employing additional regularization losses to penalize the degradation of the diversity [23, 37]. This approach, referred to as few-shot GAN adaptation, performs well when the target domain is structurally very similar to the dataset that has been used for pre-training, e.g., photographs vs. sketches of human faces. However, the performance drastically degrades in case of large dissimilarities between the source and target domain as shown in Fig. 1.

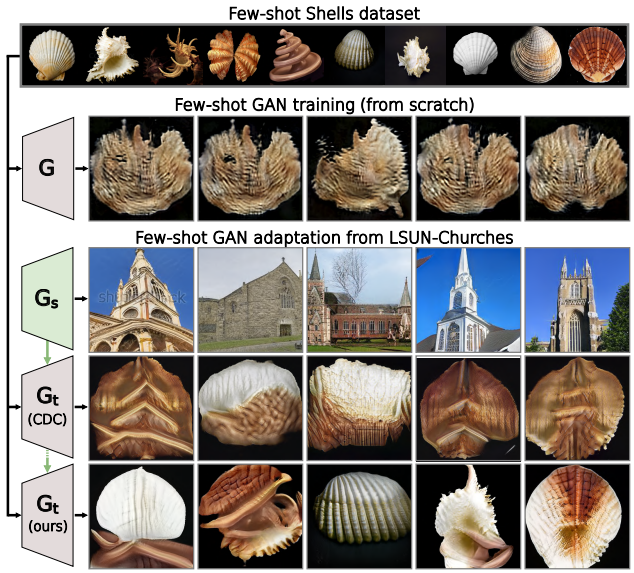

Figure 1. Training a GAN model G on a few-shot dataset (row 1) from scratch fails due to training instabilities (row 2). We thus aim to adapt a GAN G_s that has been pre-trained on a large dataset like LSUN-Church (row 3) to the target few-shot dataset (G_t). While fine-tuning [23] does not perform well either if source and target are dissimilar (row 4), our approach generates diverse and realistic images (row 5) by transferring the smoothness properties of G_s.

Such dissimilarities are a major bottleneck of using GANs in other disciplines like medicine, production, or crop science, where there is a lack of large datasets due to privacy, confidentiality, or simply lack of data. Motivated by this fact, we extend the protocol for few-shot GAN adaptation by investigating also pairs of datasets that are very different like churches and shells as shown in Fig. 1.

To improve few-shot GAN adaptation in the case of structurally dissimilar pairs, we propose a new GAN adaptation strategy. Firstly, we propose a new smoothness similarity regularization for the generator. Our key observation is that pre-trained GAN generators, regardless of the exact structure of objects in the pre-training dataset, learn well-structured and smooth latent spaces. For example, prior works demonstrated that various local shifts in the latent space can lead to interpretable and smooth transitions of

output images, such as translation of objects in the scene or changing their size [34, 9, 30]. As we show in our experiments, the proposed smoothness similarity regularization enables the transfer of this desirable property to other few-shot image domains without compromising the synthesis quality. Secondly, to overcome overfitting issues, we revisit the adversarial loss function of the discriminator and propose a simple yet efficient modification by computing the loss at different layers of the discriminator. This leads to the mitigation of overfitting and a more stabilized adaptation of the model to diverse target domains.

We evaluate our approach by adapting an unconditional [15] and a class-conditional GAN [2] to diverse few-shot target domains. Our model significantly outperforms previous state-of-the-art methods in image quality and diversity in the challenging case of dissimilar source and target domains, while performing on par with the state of the art on structurally similar dataset pairs. In summary, our contributions are as follows: (i) We extend the evaluation protocol for few-shot GAN adaptation by including new dataset pairs that are structurally much less similar than was considered in prior work. (ii) We propose a new smoothness similarity regularization, which enables diverse synthesis in the target domain by transferring the learned smoothness of a pretrained GAN. (iii) We revisit the adversarial loss function of the discriminator to stabilize few-shot GAN adaptation across diverse target domains. (iv) Our proposed model enables high-quality synthesis in the challenging case of dissimilar source and target domains, significantly outperforming prior methods. In addition, we show that our method can be applied to different classes of GAN architectures, including unconditional and class-conditional GAN models.

2. Related Work

To address the image generation problem in the low data regime, existing works mainly follow three research lines – one-shot, low-shot, and few-shot learning. One-shot generation methods [29, 31] focus on leveraging the internal patch distribution within a single image, however, their extension to capture the distribution of a small collection of images is non-trivial. In low-shot learning [41], several works [41, 12] proposed to mitigate the limited-data-induced overfitting issue by adapting data augmentations to the generative networks. Others [18, 5] stabilized the training process and reduced overfitting by revising the network design. Despite the promising performance in many low data regimes (typically having 100+ images), these low-shot methods fail in the extremely few-shot setting (e.g., 10 images). Our work lies in the scope of few-shot learning.

Few-shot image synthesis. Conventional few-shot learning aims at learning a discriminative classifier under limited data scenarios. In the context of image synthesis with GANs, the goal instead is to produce diverse new images from the learned distribution while preventing overfitting to the few training samples. A straightforward way is to treat it as a domain adaptation problem and incorporate the commonly used transfer learning technique, i.e., fine-tuning, to ease the need for data. However, naive fine-tuning (TGAN) [36] often suffers from overfitting and results in poor performance. Researchers proposed remedies such as mining suitable parts of the latent space before fine-tuning [35] or restricting weight updates, for example, updating only the BatchNorm parameters of the generator [22], penalizing drastic changes in important weights [17], or freezing the earliest layers of the discriminator (FreezeD) [20]. More recent works focused on introducing different regularizations to preserve specific knowledge from the pre-trained model and prevent diversity degradation [42]. For example, CDC [23] proposed to preserve the pair-wise perceptual similarity between samples from the source domain and to transfer it to the target domain, while RSSA [37] designed a novel consistency term to align the structural information between source and target domains. Although the two aforementioned methods constitute the current state of the art in few-shot generative learning, their assumptions impose strong constraints on the structure of the few-shot target domain. As we show in experiments, they fail in the more challenging regime when the source and target domains are not restrictively similar. Most recently, [39] proposed to replace prior knowledge preservation criteria with adaptation-aware kernel modulation (AdAM), which relaxed the source-target proximity requirement of previous methods to some extent. In this work, we take a step further and introduce a new regularization term to preserve the generator's smoothness properties that are not limited to a specific domain, enabling successful adaptation between image domains of unprecedented structural dissimilarity.

Smoothness of image generators. Smooth transitions in the latent space are an important property for generative models, where it is believed to be a sign of a well-conditioned generator. Models trained on large datasets naturally possess this property with or without explicit regularization [2, 15]. For example, StyleGANv2 [15] introduced a regularization based on the perceptual path length measure (PPL) [14], which encourages that a fixed-size step in the latent space results in a fixed-magnitude change in the image space. However, achieving a smooth mapping of the generator is difficult for few-shot image synthesis since there are not enough training samples. Thus, MixDL [16] sought to alleviate the "staircase" latent space interpolations, i.e., jumps between training samples, by introducing a continuous coefficient vector and enforcing smooth interpolations between training images. Although the two above regularizers aim to encourage smoother interpolations between training samples and thus mitigate mode collapse, they are not designed to take advantage of the available pre-training

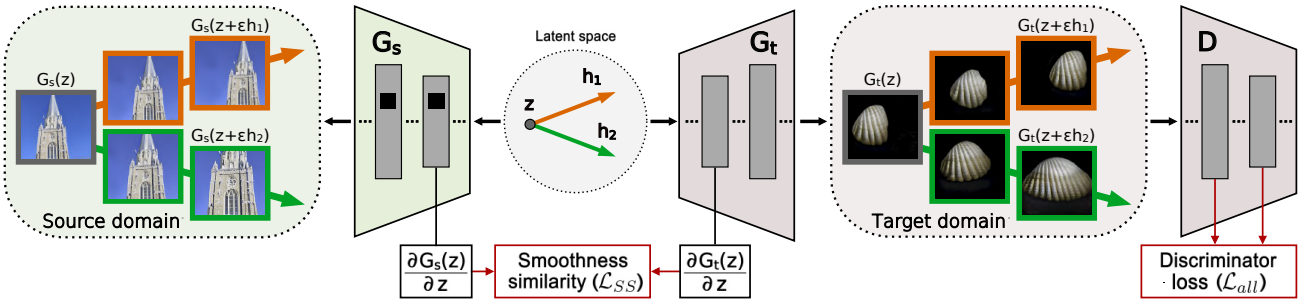

Figure 2. Given a pre-trained generator G_s, the proposed smoothness similarity regularization preserves the learned smoothness of G_s while adapting it to a target domain with very few images. To mitigate overfitting to the target domain, the discriminator loss utilizes features at various layers and automatically adjusts the impact of different semantic scales to the similarity of the source and target domain.

knowledge. In contrast, in this work we develop a new smoothness similarity regularization that leverages the well-structured latent space of a pre-trained GAN generator. In effect, our approach enables high-quality few-shot image synthesis by transferring smooth and realistic image transitions of pre-trained GANs to diverse few-shot domains.

3. Method

In the task of few-shot GAN adaptation, we are given a small target dataset T and a pre-trained GAN model, consisting of a discriminator D and a generator G_s, which produces an image $x = G_s(z)$ from a continuous input variable z, such as a random noise vector or a continuous class embedding. The goal is to adapt the generator to the target dataset such that it generates diverse and realistic images in the domain of the target dataset as shown in Fig. 1. We denote the adapted target generator by G_t.

To achieve few-shot synthesis with a high image quality and diversity, our model should adhere to the following two properties. Firstly, the generator G_t should not only memorize and generate the target images, which will be addressed by the smoothness similarity regularization (Sec. 3.1). Secondly, the discriminator D must avoid overfitting to the few target images in order to provide useful supervision for G_t (Sec. 3.2). The overview of our method is shown in Fig. 2.

3.1. Smoothness similarity regularization for G_t

In a low data regime like ours, G_t can easily overfit to the target dataset T and collapse to reproducing only the few modes represented in the training data. When walking in the latent space of such a generator, one would observe "staircase" patterns, where minor shifts in the latent space cause discontinuous transitions in the output image space (as shown in row 4 of Fig. 5). Naturally, to achieve a synthesis of high diversity, it is desirable for G_t to avoid such discontinuities, as having smoother image transitions allows to generate intermediate samples that can exhibit novel features. Therefore, in our model we aim to encourage G_t to produce smooth latent space interpolations, in which all the intermediate images are realistic.

Our approach is based on the observation that GANs trained on large datasets tend to have a well-structured latent space [34, 9, 30], in which different latent space directions can lead to smooth and interpretable image transitions. For example, in a generator pre-trained on a large dataset of churches, latent directions can emerge causing smooth zooming or translation of churches (see Fig. 2). Our observation is that the nature of such image transitions (e.g., zooming or translation) is remarkably general. Thus, we propose a regularizer that utilizes this smoothness property of the source generator G_s as a cue while adapting it to another image domain, which can be very different from the domain that was used for pre-training. For example, as shown in Fig. 2, the same latent directions of churches can cause similar zooming or translation effects on shells.

Mathematically, the smoothness of the generator can be represented via a Jacobian matrix $J_{G^l}(z) = ||\partial G^l(z)/\partial z||$, quantifying how the generator's intermediate features after the l-th block change under local shifts in the latent space. As we want the same latent shift to cause perceptually similar image transitions in the source and target domains, we design a regularization term that brings the Jacobian matrices of G^l_s and G^l_t closer together. As the computation of full Jacobian matrices is expensive, we use an unbiased estimator of their products with a Gaussian vector [6, 15], which can be computed with standard back-propagation:

$$J^T_{G^l}(z) \cdot y = \mathbb{E}_{(y)\sim N(0,1)} \nabla_z \langle G^l(z), y \rangle, \tag{1}$$

where y is a Gaussian tensor of the same shape as G^l. Our smoothness similarity regularization is then expressed as:

$$\mathcal{L}_{SS} = \lambda_{SS} \cdot \mathbb{E}_{(z,y)\sim N(0,1)} ||\nabla_z \langle G^l_s(z), y \rangle - \nabla_z \langle G^l_t(z), y \rangle||_2, \tag{2}$$

where λ_{SS} steers the impact of the regularizer. As shown in Fig. 2, the smoothness similarity regularization depends on both generators, but only G_t is updated. It is interesting to note that the Jacobian matrix is also used for the path length regularization [15], which forces $J_G(z)$ to be orthogonal up to a global scale at any z. While this alternative regularizer also induces some form of smoothness, it does not transfer the inherently learned smoothness

of a pre-trained GAN. We show in Sec. 4.1 that it struggles to enforce the realism of intermediate images. Furthermore, our approach shares the motivation with some prior regularization approaches that use noise perturbations to enforce diversity [23, 37]. In contrast to Eq. 2, these approaches incorporate non-gradient components, e.g., assuming similarity of images $G_s(z) \leftrightarrow G_t(z)$ or distributions $d(G_t(z_1), G_t(z_2)) \leftrightarrow d(G_s(z_1), G_s(z_2))$. As such assumptions are violated when source and target domains are dissimilar, they perform poorly compared to our smoothness similarity regularization $\mathcal{L}_{SS}$ as shown in the experiments.

3.2. Revisiting the D adversarial loss

To identify what kind of image transitions look realistic for the target domain, G_t requires strong supervision from the discriminator on image realism at different semantic scales. This includes the colors and textures of objects, as well as object shapes, especially if their distribution is different from the shapes of objects in the source domain. Learning the concept of image realism in low data regimes is, however, challenging due to the problem of overfitting.

Typically, a GAN discriminator consists of several consecutive blocks $\{D^i\}_{i=1}^N$ and computes for each given image x a real/fake logit after the last block $l = s^N \circ D^N(x)$, where s^N is a final processing layer such as a convolution. When adapting such a discriminator to a very small dataset, it is prone to memorizing the training set [32], leading to mode collapse and poor diversity of synthesized images [23]. A possible solution [23, 37] to overcome memorization is to use variants of the PatchGAN discriminator [11], discarding the latest discriminator layers: $l = s^k \circ D^k(x), k < N$. This solution allows to adapt colors and textures of generated images to the target domain while avoiding the memorization problem. However, it naturally has a limited capacity to learn more high-level semantic scene properties such as the shapes of objects, which we show in the experiments.

In order to avoid memorization, and yet to balance the adaptation of colors, textures, and shapes of generated objects to a new domain, we hypothesize that a more flexible attention to different levels of image realism is required by the discriminator. To this end, we perform a simple yet efficient modification to the loss function of the discriminator. Given a discriminator $\{D^i\}_{i=1}^N$ and its adversarial loss function $\mathcal{L}_{\mathcal{D}}(l)$ used for pre-training (e.g., cross-entropy or hinge loss), we design the discriminator to produce real/fake logits after *each* discriminator's block, and correspondingly compute the loss as the average across all blocks:

$$\mathcal{L}_{all}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\mathcal{D}}[l^i(x)], \quad l^i(x) = s^i \circ D^i(x). \quad (3)$$

With the new objective, D is given more freedom to utilize the features extracted at different scales to compute the loss. Our finding is that D dynamically adapts the magnitude of the loss at each scale to the target domain, without explicit supervision (see Fig. 6). Consequently, we observe a strong overall stabilization effect on the adaptation performance across diverse source-target dataset pairs.

4. Experiments

To demonstrate that our approach for few-shot GAN adaptation can be applied to unconditional and class-conditional GANs, we selected for each category a popular GAN architecture: unconditional StyleGANv2 [15] and class-conditional BigGAN [2]. For both models, we test our approach on a variety of source-target domain pairs. We focus on 10-shot target adaptation in the main paper, but we provide results for 1-shot and 5-shot adaptation in the supplementary material. For fair comparisons with prior works, most of our ablations and comparisons are conducted with StyleGANv2.

4.1. Adaptation of unconditional GAN

Datasets. In contrast to previous works that mostly considered pairs of similar datasets like *Face→Sketch* and *Face→Sunglasses*, we extend the protocol by including structurally dissimilar pairs of source and target domains, which is a more challenging task and is our primary interest. As source generators, we use StyleGANv2 checkpoints pretrained on FFHQ [14], LSUN-Church, and LSUN-Horse [38]. For the target datasets, we selected 10-shot subsets of various commonly used few-shot datasets, such as Anime-Face, Shells, or Pokemons [41, 18]. Results on more datasets are shown in the supplementary material.

Training details. We fine-tune StyleGANv2 using the $\mathcal{L}_{SS}$ and $\mathcal{L}_{all}$ loss terms as presented in Sec. 3. For the smoothness similarity regularization, we use the intermediate features G^l at resolution (32×32) and set $\lambda_{SS} = 5.0$. We follow [23] in choosing all the other hyperparameters, such as image resolution (256×256), learning rates, and batch size. Our experiments across all datasets use the same model configuration and set of hyperparameters.

Baselines. We compare our method to most recent few-shot GAN adaptation approaches: TGAN [36], FreezeD [20], CDC [23], RSSA [37], and AdAM [39]. In addition, we compare our proposed smoothness similarity regularizer $\mathcal{L}_{SS}$ to other regularization techniques: path length regularization (PPL) [15] and MixDL [16].

Evaluation. In low data regimes, it is necessary to judge results both in quality and diversity aspects, as there is a trade-off between them [25, 32]. We measure the quality with FID [10] between a held-out validation set and a generated set of the same size. Following [23], we evaluate diversity with the intra-LPIPS, clustering generated images according to their nearest training samples and computing

Figure 3. Visual comparison to prior methods on *Face→Anime* and *Church→Shells*, the source-target dataset pairs with a dissimilar structure (e.g., shapes of objects). In this challenging regime, we observe that prior methods suffer from training instabilities, memorization issues, or inability to adapt the shapes of objects to the new domain. In contrast, our method generates images that look realistic, flexibly combine features of different target images, and transfer the variation of images from the source domain to the target domain.

the average LPIPS [40] of all the clusters. We train all models for 30k epochs in case of dissimilar domain pairs and for 5k on closer domains, evaluating metrics every 1k epochs. Final checkpoints in all experiments correspond to best FID.

Results with dissimilar source-target domains. We first present our results on the source-target domain pairs with dissimilar structure: *Face→Anime*, *Church→Shells*, and *Horse→Pokemon* (see Fig. 3 and supplementary material). Our general observation from Fig. 3 is that in this challenging regime prior methods suffer either from training instabilities, memorization issues, or inability to adapt the shape of objects to the new domain. For example, for *Face→Anime*, despite an apparent correspondence between the two domains, none of the prior methods success-

Figure 4. Visual comparison to most recent prior methods on *Face→Sketch* and *Church→Sunglasses*, the dataset pairs depicting similar image domains. In this regime, our method performs on par with previous state of the art. See Table 2 for a quantitative comparison.

Method	**Face→Anime**		**Church→Shells**		**Horse→Pokemons**	
	FID↓	LPIPS↑	FID↓	LPIPS↑	FID↓	LPIPS↑
TGAN [36]	153.2	0.29	205.3	0.22	115.0	0.52
FreezeD [20]	112.4	0.22	180.8	0.27	123.3	0.49
CDC [23]	140.2	0.50	187.9	0.48	109.5	0.55
RSSA [37]	133.2	0.37	182.4	0.44	117.3	0.54
AdAM [39]	116.4	0.42	152.4	0.28	106.5	0.55
Ours	**97.3**	**0.57**	**140.5**	**0.53**	**84.1**	**0.61**

Table 1. Comparison of the adaptation performance in case of dissimilar source-target domains. Bold denotes best performance.

Method	**Face→Sketch**		**Face→Sunglasses**	
	FID↓	LPIPS↑	FID↓	LPIPS↑
TGAN [36]	54.2	0.38	36.8	0.56
FreezeD [20]	48.8	0.32	32.0	0.59
CDC [23]	54.2	0.40	30.5	0.59
RSSA [37]	61.4	**0.45**	36.3	0.58
AdAM [39]	56.3	0.37	31.1	**0.60**
Ours	**45.2**	0.44	**27.5**	**0.60**

Table 2. Comparison in case of structurally close source-target domains. Bold denotes best performance.

fully transfers the distribution of head poses to the anime style, e.g., overfitting too strongly to the 10 provided samples (FreezeD), failing to adapt the shape of faces to the style of anime (CDC), or not generating high-quality anime-faces due to instabilities (TGAN, RSSA, AdAM). Similarly, for *Church→Shells*, we observe that prior methods produce only copies of the example shells (FreezeD, AdAM), generate shells of unrealistic church-like shapes (CDC, RSSA), or suffer from instabilities (TGAN). In contrast, our method achieves high-quality synthesis, in which the generated images (i) look like realistic anime-faces and shells; (ii) flexibly combine features observed in different target images (e.g., anime hair color can be combined with various eye colors or background styles); and (iii) meaningfully transfer the variation of images from the source domain (e.g., generated shells adjust to the positions and shapes of churches).

The quantitative comparison in Table 1 confirms our analysis, where our method achieves the best quality and diversity scores across all datasets. We note a high average relative improvement of more than 18% and 11% in FID and LPIPS compared to the highest scores achieved by prior methods. Overall, we conclude that our method significantly improves over prior works on few-shot GAN adaptation with dissimilar source and target domains.

Results with close source-target domains. Next, we follow the evaluation of prior works and compare the models on similar source and target domains, such as adaptation of human faces to a different style. The visual results for *Face→Sketch* and *Face→Sunglasses* are shown in Fig. 4. Our method successfully performs the few-shot adaptation in this setting, adapting the colors and textures of faces to the gray-scale sketch domain, or adding a novel attribute (sunglasses). We note that our method is not explicitly designed to transfer all the details of a face from the source domain, thus changes in the generated images like facial hair are expected. Yet, we observe that our method generally does not lose distinctive features of faces in source images, performing on par with previous state-of-the-art methods. The quantitative comparison is provided in Table 2[1]: on both datasets our method achieves the best FID scores and performs on par with the best performer in LPIPS.

Ablations. We demonstrate the importance of our proposed loss terms in Fig. 5, which shows latent space interpolations of trained models and their similarity to the pre-

[1]FID evaluation differs from prior works (discussed in suppl. material).

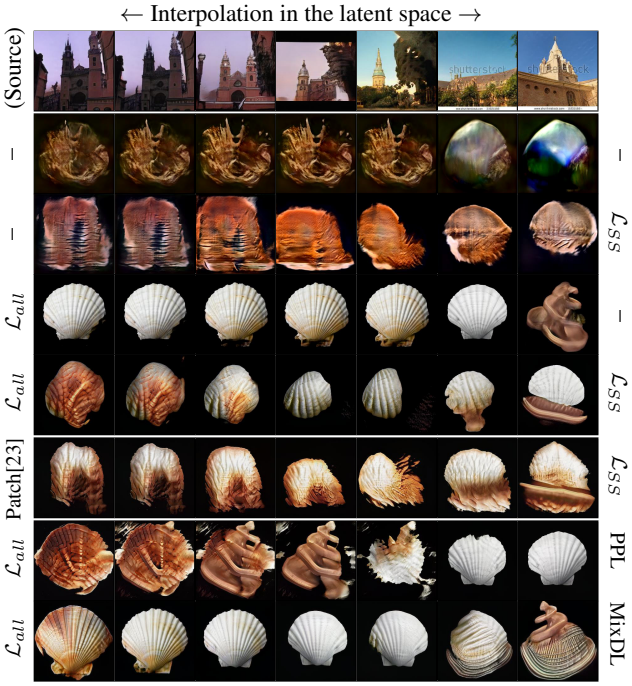

Figure 5. Latent space interpolations of the source generator and the ablation models from Tables 3-4. Leftmost and rightmost columns show the used D loss and G smoothness regularization.

D loss	Smooth reg. for G	**Face→Anime** FID↓	LPIPS↑	**Church→Shells** FID↓	LPIPS↑
StyleGANv2	-	178.0	0.21	243.8	0.17
StyleGANv2	SS (ours)	180.7	**0.61**	252.8	**0.62**
PatchGAN [23]	-	145.2	0.37	183.1	0.31
PatchGAN [23]	SS (ours)	132.2	0.55	184.2	0.56
$\mathcal{L}_{all}$ (ours)	-	116.4	0.36	175.4	0.43
$\mathcal{L}_{all}$ (ours)	SS (ours)	**97.3**	0.57	**140.5**	0.53

Table 3. Impact of $\mathcal{L}_{all}$ and $\mathcal{L}_{SS}$. Bold denotes best performance.

trained source model G_s (row 1). Firstly, we note that the plain StyleGANv2 model (row 2) suffers from instabilities in our low data regime, achieving poor image quality and diversity and having "staircase"-like latent space interpolations. Applying $\mathcal{L}_{SS}$ without $\mathcal{L}_{all}$ (row 3) helps to achieve diverse synthesis with smooth interpolations, but it is not enough to achieve good image quality. On the other hand, using $\mathcal{L}_{all}$ (row 4) helps to overcome instabilities and improve image quality, but it cannot maintain smooth interpolations and high diversity. Finally, our full model (row 5) allows a higher-quality, diverse synthesis with smooth and realistic latent space interpolations. Note how the image transitions mimic the behaviour of the source model (churches and shells change shapes and positions similarly), allowing to achieve diverse and realistic synthesis.

The effect of $\mathcal{L}_{all}$ is further demonstrated in Fig. 6, where we show the contribution of different D blocks to the adversarial loss at different epochs. We note the ability of the discriminator to identify correct loss contributions

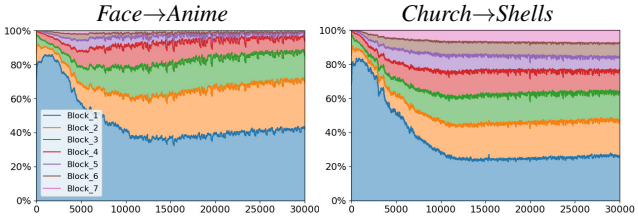

Figure 6. The contribution of features at different D blocks to the adversarial loss function $\mathcal{L}_{all}$. For two closer image domains (the left plot), the network concentrates mostly on earlier layers to compute the loss, while for less similar domains the network learns to use the later layers representing more high-level D features.

D loss	Smooth reg. for G	**Face→Anime** FID↓	LPIPS↑	**Church→Shells** FID↓	LPIPS↑
$\mathcal{L}_{all}$ (ours)	-	116.4	0.36	175.4	0.43
$\mathcal{L}_{all}$ (ours)	PPL [14]	107.8	0.46	179.4	0.44
$\mathcal{L}_{all}$ (ours)	MixDL [16]	105.9	0.50	150.4	0.51
$\mathcal{L}_{all}$ (ours)	SS (ours)	**97.3**	**0.57**	**140.5**	**0.53**

Table 4. Comparison of smoothness similarity regularization $\mathcal{L}_{SS}$ with other regularizers. Bold denotes best performance.

adaptively for different source-target domain pairs. For example for *Face→Anime*, the network concentrates mostly on the earliest D blocks to adapt the colors and textures of faces to a new style. In contrast, for the more distant domains *Church→Shells*, the network learns to attribute a higher weight to the later blocks to also adapt higher-level features, such as shapes of objects. In effect, we observe a stabilized adaptation of colors, textures, and shapes of objects across diverse source-target pairs. Using PatchGAN [23] as discriminator loss does not achieve such a balance as it focuses mostly on lower-scale features (row 6 in Fig. 5).

Our observations are confirmed by the quantitative study in Table 3: without $\mathcal{L}_{SS}$ the model does not achieve high diversity (high LPIPS), while $\mathcal{L}_{all}$ is necessary for high image quality (low FID). We conclude that both our proposed loss terms are important to achieve high-quality synthesis. More ablations on $\mathcal{L}_{SS}$ and $\mathcal{L}_{all}$ can be found in the supplementary material.

Lastly, Table 4 provides a comparison of our proposed $\mathcal{L}_{SS}$ loss term to other regularizers: path length regularization (PPL) [14] and MixDL [16]. While all regularizers help to achieve smoother latent space interpolations and thus improve the quality and diversity metrics, our smoothness similarity regularization enables the highest performance in both FID and LPIPS. While our approach transfers the learned smoothness of the source generator to the target domain, PPL and MixDL resort to gradually interpolating between the provided training samples, which leads to latent space interpolations that either look unrealistic or lack diversity (rows 7-8 in Fig. 5). This demonstrates that transferring smoothness from a pre-trained generator is beneficial to enforce image transitions that are realistic and diverse.

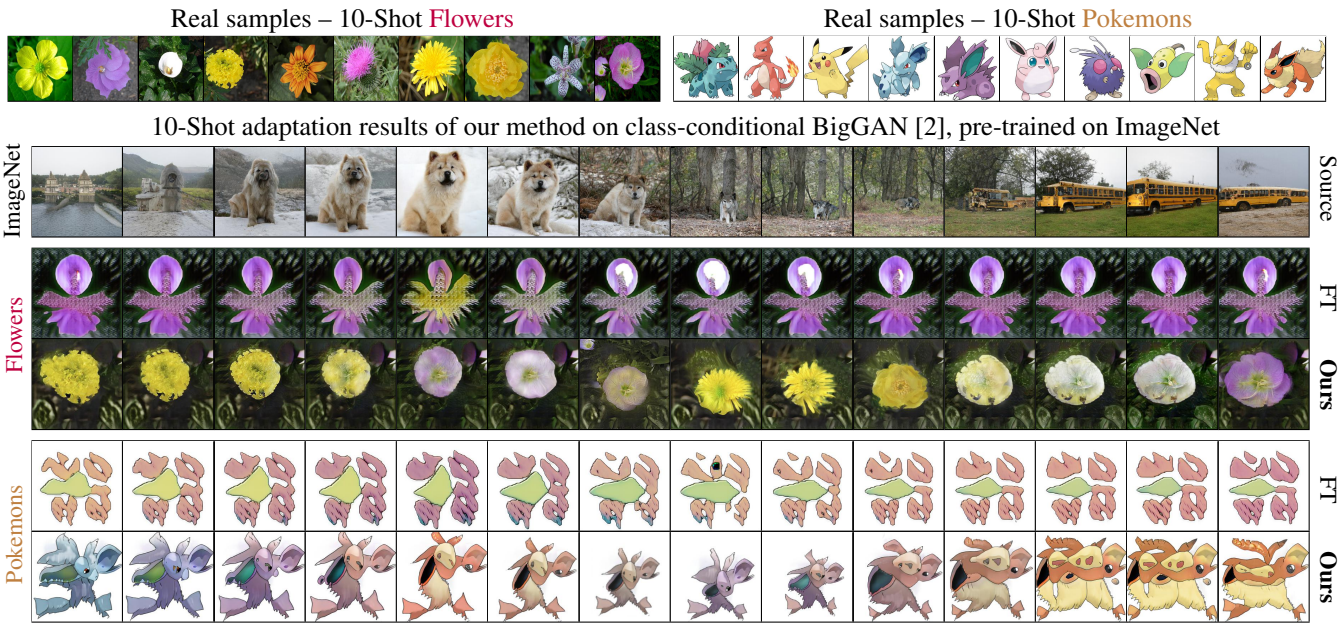

Figure 7. 10-shot adaptation results for the class-conditional BigGAN [2] pre-trained on ImageNet. While simple fine-tuning (FT) suffers from training instabilities and mode collapse, our method helps to achieve much higher image quality and diversity, transferring smooth and realistic image transitions from the source domain, e.g., objects smoothly changing their locations, size, and shape.

D loss	Smooth reg. for G	**ImageNet→Flowers** FID↓	LPIPS↑	**ImageNet→Pokemons** FID↓	LPIPS↑
BigGAN	-	213.3	0.29	226.8	0.15
BigGAN	SS (ours)	225.6	0.47	208.3	0.47
$\mathcal{L}_{all}$ (ours)	-	123.9	0.28	129.4	0.27
$\mathcal{L}_{all}$ (ours)	SS (ours)	**106.4**	**0.55**	**89.6**	**0.56**

Table 5. Ablation on the performance when adapting the class-conditional BigGAN model [2] pre-trained on ImageNet.

4.2. Adaptation of class-conditional GAN

Our approach is not limited to unconditional GANs, but it can also be applied to a class-conditional GAN model. We selected BigGAN [2] for our experiments as it is a popular backbone architecture for class-conditional image synthesis on ImageNet [7]. We make two modifications to enable the adaptation of the model to unconditional few-shot datasets. Firstly, we remove the conditioning of the discriminator via the projection layer [19]. Secondly, we treat the generator's learned continuous class embedding as part of the latent space, thus sampling a Gaussian vector in the joint noise-class space at each fine-tuning epoch. This way, the generator produces an image based on a single input vector in an unconditional fashion. We then fine-tune the pre-trained model using our loss terms $\mathcal{L}_{SS}$ and $\mathcal{L}_{all}$ as presented in Sec. 3. We use image resolution 256×256 and batch size of 32. The hyperparameters for $\mathcal{L}_{SS}$ are the same as for StyleGANv2: intermediate features G^l at resolution (32×32) and $\lambda_{SS} = 5.0$. We train for 30k epochs and select checkpoints by best FID.

Datasets. As the source generator, we use the BigGAN checkpoint pre-trained on class-conditional ImageNet at resolution 256×256. We demonstrate 10-shot adaptation results with two commonly used few-shot generation datasets: Oxford-Flowers [21] and Pokemons [18]. We use the same model configuration for both datasets.

Results. Fig. 7 demonstrates latent space interpolations of the source and target generators. We note that a simple fine-tuning of BigGAN suffers from training instabilities and mode collapse. In contrast, our method successfully adapts BigGAN to generate diverse images in the target domains. We highlight that our method transfers smooth and realistic image transitions from the well-learned BigGAN's noise-class space, despite significant dissimilarities between ImageNet and the few-shot datasets, in particular Pokemons. For example, it can be noticed how the latent space interpolations in the target domains mimic the source domain, e.g., the generated flowers and pokemons change their position and size similarly to dogs and wolves (5^{th}-10^{th} columns in Fig. 7) or stretch their shape to mimic the proportions of busses (11^{th}-14^{th} columns).

Table 5 shows the importance of our proposed loss terms. Our observations are consistent with the ablations with StyleGANv2: $\mathcal{L}_{all}$ is necessary to avoid instabilities and achieve a good image quality (low FID), while $\mathcal{L}_{SS}$ is required to achieve smooth latent space interpolations and good diversity (high LPIPS). We conclude that our method successfully extends to the adaptation of class-conditional models, where target domains benefit from the rich noise-class space learned on a multi-class dataset such as ImageNet. More details and results are provided in the supplementary material.

5. Conclusion

In this work, we presented a new method for few-shot adaptation of GAN models. It transfers the smooth latent space of a pre-trained GAN, which was trained on a large dataset, to a new domain with very few images. We addressed the case of few-shot GAN adaptation when the source and target domains are structurally dissimilar, which is a common issue in applications. Our extensive results demonstrate that in this setting our approach outperforms previous works in terms of image quality and diversity.

Acknowledgement

The work has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2070 -390732324 and the ERC Consolidator Grant FORHUE (101044724). We thank Jinhui Yi for providing and analyzing the sugar beet data used in supplementary experiments.

References

[1] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

[3] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[5] Kaiwen Cui, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, Fangneng Zhan, and Shijian Lu. Genco: Generative co-training for generative adversarial networks with limited data. In *Conference on Artificial Intelligence (AAAI)*, 2021.

[6] Yann Dauphin, Harm De Vries, and Yoshua Bengio. Equilibrated adaptive learning rates for non-convex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[8] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[10] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Tero Karras, Miika Aittala, Janne Hellsten, S. Laine, J. Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[13] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[16] Chaerin Kong, Jeesoo Kim, Donghoon Han, and Nojun Kwak. Few-shot image generation with mixup-based distance learning. In *European Conference on Computer Vision (ECCV)*, 2022.

[17] Yijun Li, Richard Zhang, Jingwan Cynthia Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[18] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

[19] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations (ICLR)*, 2018.

[20] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze discriminator: A simple baseline for fine-tuning gans. In *CVPR AI for Content Creation Workshop*, 2020.

[21] Maria-Elena Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[22] Atsuhiro Noguchi and T. Harada. Image generation from small datasets via batch statistics adaptation. In *International Conference on Computer Vision (ICCV)*, 2019.

[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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[24] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[25] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-shot adaptation of generative adversarial networks. *arXiv:2010.11943*, 2021.

[26] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv preprint arXiv:2301.09515*, 2023.

[27] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH*, 2022.

[28] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

[29] Tamar Rott Shaham, Tali Dekel, and T. Michaeli. SinGAN: Learning a generative model from a single natural image. In *International Conference on Computer Vision (ICCV)*, 2019.

[30] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[31] Vadim Sushko, Juergen Gall, and Anna Khoreva. One-Shot GAN: Learning to generate samples from single images and videos. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

[32] Vadim Sushko, Dan Zhang, Juergen Gall, and Anna Khoreva. Learning to generate novel scene compositions from single images and videos. *arXiv:2103.13389*, 2021.

[33] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[34] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning (ICML)*, 2020.

[35] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, L. Herranz, F. Khan, and Joost van de Weijer. Minegan: Effective knowledge transfer from gans to target domains with few images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[36] Yaxing Wang, Chenshen Wu, L. Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and B. Raducanu. Transferring gans: generating images from limited data. In *European Conference on Computer Vision (ECCV)*, 2018.

[37] Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. Few shot generative model adaption via relaxed spatial structural alignment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

[39] Zhao Yunqing, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-man Cheung. Few-shot image generation via adaptation-aware kernel modulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[41] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[42] Yunqing Zhao, Henghui Ding, Houjing Huang, and Ngai-Man Cheung. A closer look at few-shot image generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.