Progressive Few-Shot Adaptation of Generative Model with Align-Free Spatial Correlation

Jongbo Moon*, **Hyunjun Kim***, **Jae-Pil Heo**†

Sungkyunkwan University
{mjbox, arith, jaepilheo}@skku.edu

Abstract

In few-shot generative model adaptation, the model for target domain is prone to the mode-collapse. Recent studies attempted to mitigate the problem by matching the relationship among samples generated from the same latent codes in source and target domains. The objective is further extended to image patch-level to transfer the spatial correlation within an instance. However, the patch-level approach assumes the consistency of spatial structure between source and target domains. For example, the positions of eyes in two domains are almost identical. Thus, it can bring visual artifacts if source and target domain images are not nicely aligned. In this paper, we propose a few-shot generative model adaptation method free from such assumption, based on a motivation that generative models are progressively adapting from the source domain to the target domain. Such progressive changes allow us to identify semantically coherent image regions between instances generated by models at a neighboring training iteration to consider the spatial correlation. We also propose an importance-based patch selection strategy to reduce the complexity of patch-level correlation matching. Our method shows the state-of-the-art few-shot domain adaptation performance in the qualitative and quantitative evaluations.

1 Introduction

Generative Adversarial Networks (GANs) have achieved impressive performances in the image synthesis (Brock, Donahue, and Simonyan 2019; Creswell et al. 2018). To train GANs to generate high quality and diverse images, a massive number of training data and a long training time are required (Karras, Laine, and Aila 2019). However, since it is often difficult to collect sufficient amount of training data for a specific domain in practice, learning GANs from scratch for such domain can become hard.

To overcome the challenge of training GANs from scratch due to a lack of data, there has been growing interest in adapting GANs trained on a source domain with abundant data to a target domain that suffers from data scarcity. This approach has been inspired by the success of domain adaptation techniques in various downstream computer vision tasks (Long et al. 2015, 2017). Pioneering methods for the generative model adaptation perform fine-tuning on pre-trained GANs by fixing some layers with target domain data (Mo, Cho, and Shin 2020; Robb et al. 2020) or learn a mapping network between source and target latent distributions (Wang et al. 2020). Although those techniques enable to adapt GANs well-trained on the source domains to the target domains, they still require hundreds of target domain data to generate high-quality images without serious overfitting or mode-collapse problems.

The difficulty of collecting many target domain data (e.g., more than 100 samples) encourages to develop few-shot generative model adaptation techniques. The key idea in this field of research is to match relationship among instances in different domains. The instance-wise relative distances are regularized in IDC (Ojha et al. 2021) to preserve the relationship among samples generated from the same set of latent variables in source and target domains. IDC relieves the over-fitting problem of few-shot adaptation by injecting the sample diversity of source domain into the target domain. However, IDC still suffers from visual artifacts. As the latest work, a relaxed spatial structural alignment (RSSA) (Xiao et al. 2022) is proposed to preserve structural relationship in the source domain by considering spatial-wise relative distances. Specifically, given two images in source and target domains generated from the same latent variable, RSSA aims to match similarities between two patches at the same location in different domains. By matching such local relationship, it significantly enhances the visual quality over IDC in certain scenarios where its strong assumption nicely holds that the source and target domains share a similar spatial layout. However, the assumption is too strict to generalize toward various domain pairs if two domains have misaligned spatial layouts. For example, although the spatial locations of eyes are not aligned in source and target domains, RSSA tries to equalize the similarity between regions of two eyes (in source domain) and the similarity between regions not containing eyes (in target domain). This highly likely to produce visual artifacts, as illustrated in Fig. 2-(a).

In this paper, we mainly address the aforementioned problem of RSSA by modeling a spatial correlation between source and target domains that is free from the assumption on their alignments. Specifically, since the adaptation progressively proceeds with the training iteration, the semantically coherent image regions (i.e., eyes) are also pro-

*These authors contributed equally.
†Corresponding author

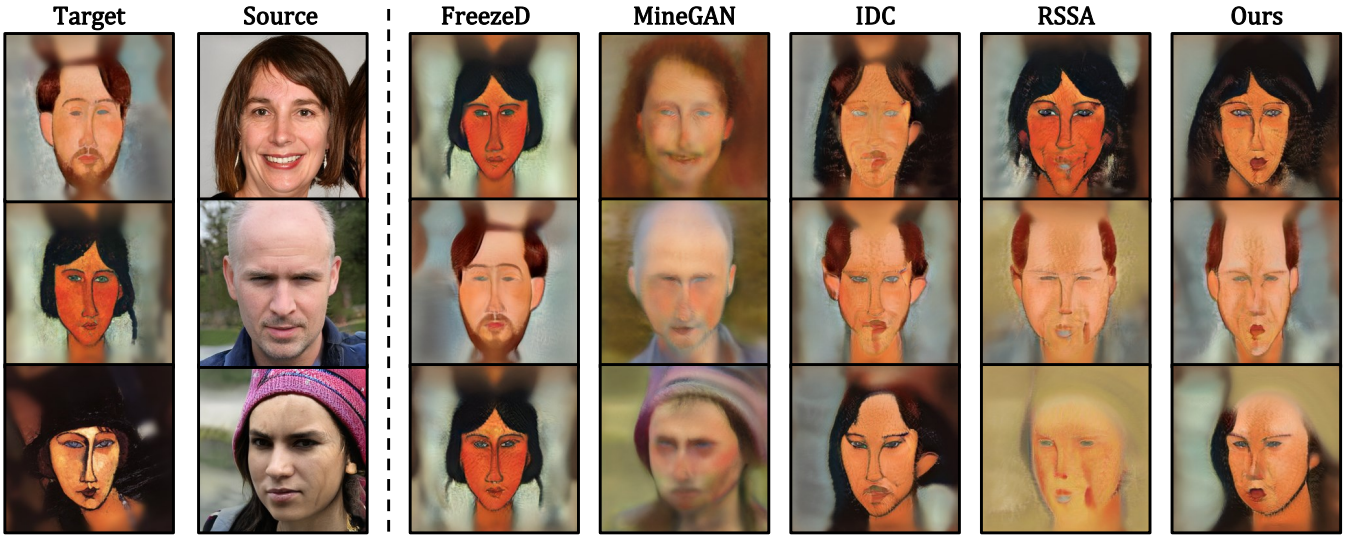


Figure 1: Qualitative results for a generative model adaptation task on FFHQ → Amedeo (FFHQ and Amedeo as source and target domains, respectively), only with 5 images of Amedeo as the available training data (5-shot). It is extremely challenging task to generate high-quality target domain samples due to the significant domain gap and scarcity of training data. As shown, our method preserves the contents of source domain, and spatial structure and style of target domain. On the other hand, IDC generates more visual artifacts while RSSA hardly reflects the spatial structures of target domain. Note that, images at the second column are generated by GANs trained with sufficient number of samples for the comparison, while other results are produced by models adapted only with 5-shot examples. The images at each row are generated from the same latent variable.

gressively shifting during the adaptation. This enables us to match such image regions within nearby training iterations and formulate a training objective with the consistency of similarity between those regions. Furthermore, we also introduce an importance patch sampling criteria to reduce the complexity of the adaptation process.

Our main contributions are summarized as follows:

- We propose an align-free spatial correlation for few-shot generative model adaptation, to deal with source and target domains whose spatial layouts are different, via progressive matching of regions of interests.
- To reduce the complexity of adaptation process, we introduce an importance sampling criterion to be considered in the spatial correlation.
- Our proposed method achieves the state-of-the-art results in both quantitative and qualitative evaluations including an extensive user study for the few-shot generative model adaptation task.

2 Background

Problem Definition In the few-shot generative model domain adaptation task, we have a source generator G_s pretrained on a large source dataset $\mathcal{D}_s$, and K-shot target domain data $\mathcal{D}_t$ (K is very small, for instance 5 or 10). The goal is to train a generator $G_{s\to t}$ for the target domain by adapting G_s. Formally, for a latent vector z sampled from the latent space $p(z)$ for the target domain, the generating distribution of $G_{s\to t}(z)$ should be similar to the data distribution of the target domain.

Previous Few-Shot Generative Model Adaptation Methods We first briefly describe very recent methods IDC (Ojha et al. 2021) and RSSA (Xiao et al. 2022) closely related to our method.

IDC proposes to preserve the pairwise feature-level distances among instances within a batch for source and target domains. Specifically, for a set of latent variables $\{z_b\}_1^B$ within a batch size B, instance-wise probability distributions ρ for i^{th} latent vector z_i is defined as follows:

$$\rho_s^{l,i} = \sigma\left(\left\{\text{sim}\left(F_s^l\left(i\right), F_s^l\left(j\right)\right)\right\}_{\forall i\neq j}\right), \tag{1}$$

where $F_s^l(i) \in \mathbb{R}^{C_l\times H_l\times W_l}$ represent feature maps of $G_s(z_i)$ at l^{th} layer, while C_l, H_l, and W_l denote the number of channels, heights, and width of the feature map, respectively. $\text{sim}(\cdot)$ and $\sigma(\cdot)$ are the cosine similarity and softmax functions, respectively.

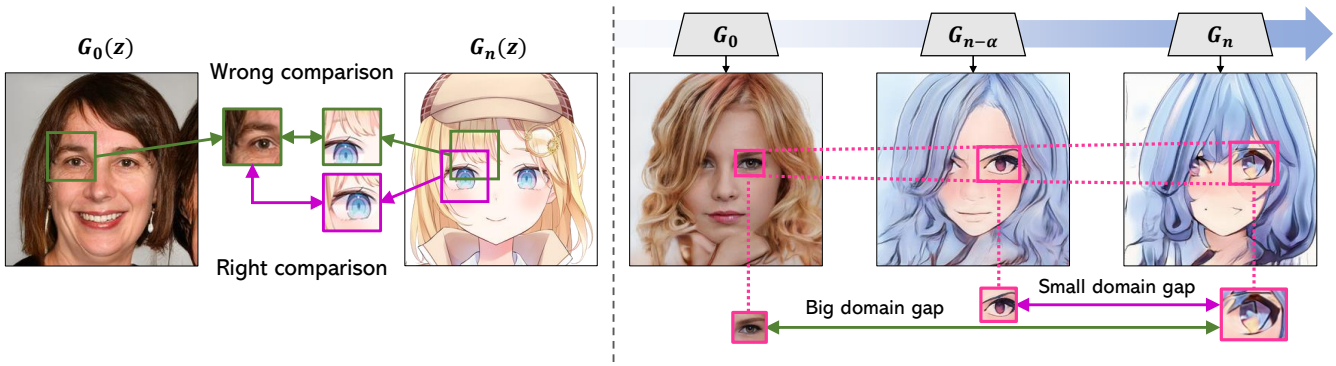
To preserve the consistency of cross-domain distances within a batch, the corresponding objective function is formulated with the KL-divergence as follows:

$$\mathcal{L}_{\text{idc}}\left(G_s, G_{s\to t}\right) = \mathbb{E}_{z\sim p(z)}\sum_{l,i} D_{KL}\left(\rho_s^{l,i} \parallel \rho_{s\to t}^{l,i}\right) \tag{2}$$

On the other hand, RSSA proposes a self-correlation consistency loss that inherits local relationship of source instances to the target domain. Specifically, h-th row and w-th column of the self-correlation matrix of the position X at l^{th} layer $\pi^{l,X} \in \mathbb{R}^{H_l\times W_l}$ is defined as follows:

$$\pi_s^{l,X} = \text{sim}\left(f_s^l\left(X\right), f_s^l\left(h_l, w_l\right)\right), \tag{3}$$

where $f_s^l(\cdot) \in \mathbb{R}^{C_l}$ is a local feature at given position.



(a) **Matching Problem** (b) **Progressive Adaptation**

Figure 2: (a) When the source and target domains have a significant domain gap, the image regions at the same position hardly contain the comparable semantics. However, the prior technique, RSSA, assumes the positional identity in source and target domains so thus it is hard to learn the spatial structure of the target domain. This motivates us to match regions having the comparable semantics to generate images structurally consistent with the target domain. (b) Our motivation is hard to achieve if we consider the source and target domains directly (i.e., the distributions of G_0 and G_n), since it is very challenging to match semantically similar regions between two domains in unsupervised setting. However, we are inspired from the fact that the model is progressively adapted from source to target domains. During the progressive adaptation, the two models at nearby training iterations (i.e., $G_{n-\alpha}$ and G_n with a constant α small enough) have a much smaller domain gap so thus we have a high change to match such regions since the distribution shifts highly likely to be minor.

Based on the self-correlation matrix, RSSA formulates a self-correlation consistency objective with the smooth-ℓ_1 loss as follows:

$$\mathcal{L}_{\text{scc}}\left(G_s, G_{s\to t}\right) = \mathbb{E}_{z\sim p(z)} \sum_{l,X} \text{s-}\ell_1 \left(\pi_s^{l,X}, \pi_{s\to t}^{l,X}\right) \quad (4)$$

Based on this objective, RSSA can generate more diverse images than IDC by maintaining the spatial correlation of images generated by the source model. However, this approach could produce undesired visual artifacts if there are spatial mismatches between the target and source domains.

3 Method

3.1 Overview

In a generative model adaptation task with extremely few-shot samples, as shown in Fig.1, early approaches such as FreezeD (Mo, Cho, and Shin 2020) and MineGAN (Wang et al. 2020) suffer from severe mode-collapse or produce low-quality images. To solve this problem, IDC (Ojha et al. 2021) and RSSA (Xiao et al. 2022) have proposed advanced methods that preserve cross-domain distance consistency either instance-wise or patch-wise, but these approaches still produce visual artifacts as shown in Fig. 1.

Although the patch-wise cross-domain distance consistency enables more precise adaptation for some domain pairs, it is relied on a strong assumption that the source and target domain instances share identical semantics at the same position on their image space. However, the assumption hardly holds when the source and target domains have a large domain gap. As shown in Fig.2-(a), the image regions of two domains are not semantically aligned so thus the aforementioned consistency brings inappropriate regularization in the adaptation. To develop an few-shot generative model adaptation method handling such difference in spatial layouts across domains, while preserving the fidelity, identity and diversity of the source model, we explicitly match the semantically coherent image regions between domains. However, matching the similar semantic regions between different domains is very challenging due to the domain gaps (Lee, Moon, and Heo 2022). To alleviate such difficulty, we pay attention to the property of progressive adaptation process that two models in training at the nearby iterations have a small domain gap. As shown in Fig.2-(b), domain gap between models at the current and neighboring iterations is much smaller than ones from the initial source model. Based on this, we propose a progressive adaptation with align-free spatial correlation in Sec. 3.2. Furthermore, we introduce an importance feature sampling strategy to be considered in the consistency regularization to select semantically meaningful regions and reduce the complexity of learning in Sec. 3.3.

3.2 Progressive Adaptation with Align-Free Spatial Correlation

Instead of directly considering G_s and $G_{s\to t}$, we view the adaptation process as a progressive learning with a total $N+1$ iterations. If we denote G_n as the generator at the n^{th} iteration where $n \in [0, N]$, the G_s and $G_{s\to t}$ correspond to G_0 and G_N, respectively. Our key insight is that the domain gap between two distributions of G_n and $G_{n-\alpha}$ expect to be manageable if a constant α is small enough. Note that, we define $G_{n-\alpha}$ parameters based on the exponential moving average (EMA) (Yazıcı et al. 2019) so that it is not necessary to keep the parameters of every iteration but they are

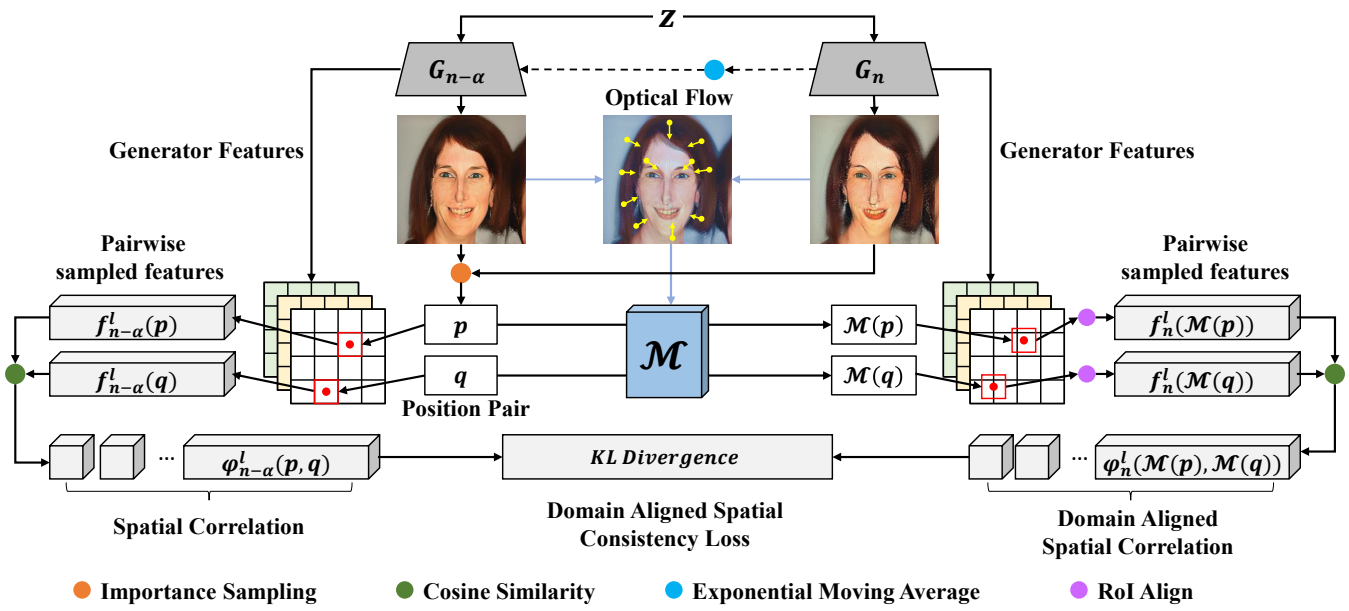


Figure 3: Overall framework of our approach. We are mainly consider a pair of generators $G_{n-\alpha}$ and G_n during the progressive adaptation from source to target domains, where G_n is the generator at the n^{th} training iteration and α is a small constant so that $G_{n-\alpha}$ and G_n have a small domain gap. The parameters of those generators can be efficiently approximated without explicit storing of them thanks to utilization of exponential moving average for the parameters. We first match the semantically comparable image regions generated by two generators at the nearby iterations based on the optical flow since the images are close to each other. Once such image regions are identified, we define the correlation between image patches within each generated image and formulate a loss function to encourage the correlations in different images to have a similar distribution via KL-Divergence. Moreover, we also perform an importance sampling for the selection of reliable image regions to reduce undesired visual artifacts and computational complexity.

efficiently estimated on the fly with a negligible cost.

Through our progressive adaptation process, the small domain gap between the models at nearby iterations G_n and $G_{n-\alpha}$ allows us to identify semantically similar regions in $G_n(z_i)$ and $G_{n-\alpha}(z_i)$ generated from the same latent variable z_i. In other word, given an image patch in $G_n(z_i)$, we can identify its corresponding patch in $G_{n-\alpha}(z_i)$. Specifically, we utilize a dense optical flow for this, since the pixel-level displacements between $G_n(z_i)$ and $G_{n-\alpha}(z_i)$ are expected to be small. As a result, we construct a pixel-wise positional mapping table $\mathcal{M}^{z_i}_{n,\alpha} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, as follows:

$$\mathcal{M}^{z_i}_{n,\alpha}(p) = \mathcal{H}\left(G_{n-\alpha}\left(z_i\right), G_n\left(z_i\right), p\right), \tag{5}$$

where $\mathcal{H}\left(I_1, I_2, p\right)$ is a predicted position in the image I_2 corresponding to the given position p in the image I_1, computed by the Farneback optical flow estimation (Farnebäck 2003).

Based on the $\mathcal{M}^{z_i}_{n,\alpha}$, we identify the matching positions between two samples generated by $G_{n-\alpha}\left(z_i\right)$ and $G_n\left(z_i\right)$. For each latent variable z_i, let us sample M_l positions at l^{th} layer, $P^{l,i}_{n-\alpha} = \{p^{i,j}_{n-\alpha}\}^{M_l}_{j=1}$ and determine its pairing positions $Q^{l,i}_{n-\alpha} = \{q^{i,j}_{n-\alpha}\}^{M_l}_{j=1}$ from $G_{n-\alpha}\left(z_i\right)$. Then we can approximate their aligned positions $P^{l,i}_n = \{p^{i,j}_n\}^{M_l}_{j=1}$ and $Q^{l,i}_n = \{q^{i,j}_n\}^{M_l}_{j=1}$ in $G_n\left(z_i\right)$ with $\mathcal{M}^{z_i}_{n,\alpha}$ as follows:

$$p^{i,j}_n \approx \mathcal{M}^{z_i}_{n,\alpha}(p^{i,j}_{n-\alpha}) \quad \text{and} \quad q^{i,j}_n \approx \mathcal{M}^{z_i}_{n,\alpha}(q^{i,j}_{n-\alpha}). \tag{6}$$

Once we identify the positional correspondences between $G_{n-\alpha}(z_i)$ and $G_n(z_i)$, we encourage the relational consistency. We formulate the spatial correlation for a latent representation z_i as follows:

$$\varphi^{l,i}_{n-\alpha} = \sigma\left(\left\{\text{sim}\left(f^{l,i}_{n-\alpha}\left(p^{i,j}_{n-\alpha}\right), f^{l,i}_{n-\alpha}\left(q^{i,j}_{n-\alpha}\right)\right)\right\}_{\forall j}\right),$$
$$\varphi^{l,i}_{n} = \sigma\left(\left\{\text{sim}\left(\mathcal{R}\left(f^{l,i}_{n}\left(p^{i,j}_{n}\right)\right), \mathcal{R}\left(f^{l,i}_{n}\left(q^{i,j}_{n}\right)\right)\right)\right\}_{\forall j}\right), \tag{7}$$

where $\mathcal{R}\left(\cdot\right)$ is a function for the RoI align (He et al. 2017) for extracting the interpolated local feature, since the displaced location $p^{i,j}_n$ and $q^{i,j}_n$ hardly coincide with the feature map grid.

Therefore we compare aligned pairs of cross-domain local features via alignment-free spatial correlation loss $\mathcal{L}_{\text{asc}}$ between two nearby training iterations with the KL-divergence as described follows:

$$\mathcal{L}_{\text{asc}}\left(G_{n-\alpha}, G_n\right) = \mathbb{E}_{z\sim p(z)} \sum_{l,i} D_{KL}\left(\varphi^{l,i}_{n-\alpha} \parallel \varphi^{l,i}_{n}\right) \tag{8}$$

3.3 Importance Sampling

As presented in the last section, our align-free spatial correlation loss is designed to compare aligned feature pairs across domains. Considering all the local features is compu-

tationally expensive and could disturb the adaptation. For instance, a correlation between background or visual artifacts would not be helpful in the adaptation. For those reasons, we introduce a feature selection algorithm.

As a criterion for the high importance, we utilize the edge locations in the image since the positions where significant intensity changes are relatively more important. We also exclude locations whose displacement vector magnitudes are greater than a threshold, since those can be caused by errors in optical flow estimation. Thus, for a sample $G_{n-\alpha}(z_i)$ and a position set $Y = \{(1,1), ..., (H_l, W_l)\}$, the probability of each feature at a location $y \in Y$ to be selected is proportional to ν_y^i defined as follows:

$$\nu_y^i = \begin{cases} 0, & \text{if } |\mathcal{M}_{n,\alpha}^{z_i}(y) - y| > \beta^i \\ \Phi(G_{n-\alpha}(z_i), y), & \text{Otherwise,} \end{cases} \quad (9)$$

where $\Phi(I, y)$ denotes the value at the position y of Sobel filtered I, and β^i is a flow threshold for the image $G_{n-\alpha}(z_i)$ that η-quantile of flow distribution.

As a result, each position has a probability proportional to its value of the edge response function. Specifically, each feature located at y has its selection probability of $\nu_y^i / \sum_{y \in Y} \nu_y^i$. On the other hand, the pairing feature locations (i.e., $q_{n-\alpha}^{i,j}$ for $p_{n-\alpha}^{i,j}$) are uniformly selected.

3.4 Objective Function

Our objective function at n^{th} training iteration is defined as:

$$\mathcal{L}_{\text{prog}} = \lambda_{\text{idc}} \mathcal{L}_{\text{idc}}\left(G_{n-\alpha}, G_n\right) + \lambda_{\text{asc}} \mathcal{L}_{\text{asc}}\left(G_{n-\alpha}, G_n\right), \quad (10)$$

where λ_{idc} and λ_{asc} are the balancing factors for two loss terms.

Our adversarial loss is defined similarly with the IDC. Anchor variables z_{anch} are defined by adding a small Gaussian noise to $|\mathcal{D}_t|$ latent vectors randomly sampled from $p(z)$. A patch discriminator D^{patch} (Isola et al. 2017) is used to evaluate the generated samples except for the anchors, whereas the full image discriminator D^{img} is applied to the anchor samples. Thus adversarial loss at the n^{th} training iteration is defined as follows:

$$\mathcal{L}'_{\text{adv}}(G_n, D_n^{\text{img}}, D_n^{\text{patch}}) = \mathbb{E}_{x \sim \mathcal{D}_t}\left[\mathbb{E}_{z \sim p(z_{\text{anch}})} \mathcal{L}_{\text{adv}}(G_n, D_n^{\text{img}}) + \mathbb{E}_{z \sim p(z), z \nsim p(z_{\text{anch}})} \mathcal{L}_{\text{adv}}(G_n, D_n^{\text{patch}})\right], \quad (11)$$

where $\mathcal{L}_{\text{adv}}(G, D) = D(G(z)) - D(x)$.

So, our final adaptation objective including adversarial loss at the n^{th} training iteration is:

$$G_n^* = \arg\min_{G_n} \max_{D_n^{\text{img}}, D_n^{\text{patch}}} \mathcal{L}'_{\text{adv}}\left(G_n, D_n^{\text{img}}, D_n^{\text{patch}}\right) + \lambda_{\text{prog}} \mathcal{L}_{\text{prog}}\left(G_{n-\alpha}, G_n\right), \quad (12)$$

where λ_{prog} weights Eq. 10.

4 Experiment

In this section, we evaluate our method by quantitative and qualitative comparisons with previous techniques FreezeD (Mo, Cho, and Shin 2020), IDC (Ojha et al. 2021), and RSSA (Xiao et al. 2022). Note that, more experimental results on various source and target domains not presented in the paper are available in our supplementary report.

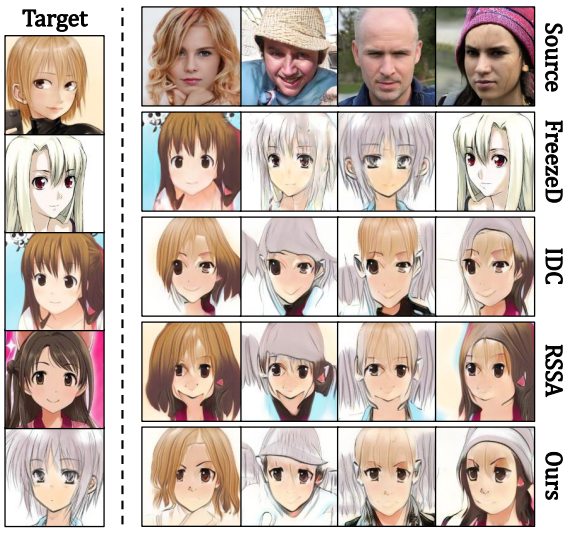


Figure 4: Qualitative results of tested methods on FFHQ → Animation with 5-shot. While tested baselines produce structural artifacts or even suffers from mode-collapse, our method generates more diverse images in a higher-quality with structural coherency. Note that, images at each column of right section are generated from the same latent vector, while images corresponding to Source is produced by the generator trained with sufficient data for the comparison.

Implementation Detail We use StyleGAN2 (Karras et al. 2020) architecture, and our training strategy starts with a StyleGAN2 model pretrained on a 256×256 resolution source domains, and adapts it into target domains with 5 or 10 target domain images. The utilized hyperparameters are identical to previous works, IDC and RSSA. We set the total number of training iteration N = 5000, batch size B = 4, and use the Adam optimizer with learning rate of 0.002. The actual implementation of $G_{n-\alpha}$ is based on EMA as discussed earlier, its decay factor is 0.9978. The number of sampled positions $M_l = 2\sqrt{H_l \times W_l}$. The hyperparameter η of quantile is set to 0.6, and the factors λ_{idc}, λ_{asc}, and λ_{prog} are 1, 200, and 1000, respectively.

Datasets We use a pre-trained model on FFHQ dataset (Karras, Laine, and Aila 2019) as a source domain, and adapt the source domain model to the following target domains; Amedeo, Sketches (R.Sharma and Prakash.R.Devale 2012), FFHQ-babies, FFHQ-sunglasses, and Animation Faces (Li et al. 2021). For 10-shots of Amedeo, Sketches, FFHQ-babies, FFHQ-sunglasses, we use the same dataset used in IDC and RSSA; in case of 5-shot, we randomly select 5-shots from 10-shot of each domain training set. Otherwise, we randomly select 5 or 10-shot from each target domain images.

Evaluation Metric To evaluate the adaptation performance, we measure diversity using IS (Inception Score) (Salimans et al. 2016) and Intra-cluster pairwise LPIPS (IC-LPIPS) (Ojha et al. 2021), and FID (Heusel et al. 2017). And we conduct a user study to subjectively evaluate

Metric	Method	Face→Amedeo		Face→Sketches		Face→Babies		Face→Sunglasses		Face→Anime	
		10-shot	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot	5-shot
IC-LPIPS	FreezeD	0.40∗	0.39	0.39∗	0.38	0.43	0.30	0.46	0.44	0.47	0.43
	IDC	0.60∗	0.56	0.45∗	0.41	0.56	0.51	0.55	0.53	0.55	0.49
	RSSA	0.58	0.56	0.43	0.41	**0.58**	0.52	0.56	0.55	0.56	0.47
	Ours	**0.61**	**0.61**	**0.46**	**0.45**	0.56	**0.54**	**0.57**	**0.58**	**0.58**	**0.55**
IS-mean	FreezeD	2.68	2.45	1.43	1.50	2.68	1.95	1.28	1.36	2.51	2.30
	IDC	2.41	3.14	1.52	1.58	2.61	2.46	1.24	1.13	2.25	1.59
	RSSA	2.67	3.73	1.62	1.57	**3.00**	2.88	1.52	1.33	2.32	1.62
	Ours	**3.13**	**3.97**	**1.86**	**1.80**	2.77	**2.95**	**1.62**	**1.36**	**2.73**	**2.30**
FID	FreezeD	-	-	46.54∗	69.04	117.18	119.79	67.93	79.14	106.48	126.97
	IDC	-	-	**45.67**∗	**56.71**	57.73	100.83	38.96	65.87	89.09	95.13
	RSSA	-	-	54.17	57.07	64.81	89.97	35.21	51.03	**83.82**	86.83
	Ours	-	-	54.92	67.72	**56.99**	**78.70**	**32.44**	**44.46**	88.33	**84.34**

Table 1: Quantitative comparison by Intra-cluster pairwise LPIPS (↑), Inception Score (↑), and FID (↓). Note that, the FID of Amedeo dataset is not measured because the dataset consists of only 10 images. Denoted with ∗ is taken from the paper of IDC.

identity, diversity, and fidelity in a human perspective.

Additionally, we experiment reconstruction task and process an instance-wise comparison as original domain. To make sure the identity preservation on reconstruction task, we are following CurricularFace (Huang et al. 2020) evaluation metrics.

4.1 Qualitative Comparison

To qualitatively compare our approaches with the baselines, we randomly generate 5-shots of Animation faces. Fig. 4 shows results on FFHQ to the Animation Faces domain using different adaptation methods, FreezeD, IDC, and RSSA.

FreezeD suffers from a serious mode-collapse problem, as it generates similar images with given few-shot training images. The instance-wise cross-domain correlation (IDC) shows more diverse images than FreezeD but produces visual artifacts and unbalanced shapes between the eyes. Plus, IDC hardly preserves the identify of the source domain. The patch-level correlation (RSSA) shows that it better preserves identity compared to other baselines, however, it still generates some artifacts like miss-aligned chin and lips. On the other hands, our methods generates images highly preserving identity and characteristics of source domain with much less visual artifacts.

4.2 Quantitative Comparison

We adapt the pre-trained source domain generative model to target domains with 5 or 10-shot images, and evaluate the results in terms of diversity and distribution similarity.

Diversity and Fidelity IS-mean and IC-LPIPS scores in Table. 1 show that our method is more effective than baselines for almost all tested domains in terms of preserving diversity. To evaluate the fidelity, we use FID to measure the distribution similarity. The FID is evaluated for only domains that have sufficient population (Sketches, Babies, Sunglasses, and Animation faces). FID scores in Table. 1 show the effectiveness of our method. Note that, since the sketch domain has too small number of images, transferring rich semantics of source model may cause even worse FID scores.

Metric	Method	Domain		
		Amedeo	Sketches	Anime
Similarity	IDC	0.793	0.793	0.799
	RSSA	0.841	0.851	0.830
	Ours	**0.848**	**0.864**	**0.856**
LPIPS	IDC	0.661	0.666	0.667
	RSSA	0.609	0.581	0.617
	Ours	**0.577**	**0.563**	**0.578**
MSE	IDC	0.443	0.446	0.432
	RSSA	0.443	0.315	0.364
	Ours	**0.310**	**0.268**	**0.250**

Table 2: Quantitative evaluation of IDC, RSSA, and our method by the reconstruction performance in terms of feature similarity(↑), LPIPS(↓), and MSE(↓) on different domains with 5-shot.

Reconstruction Evaluation Our reconstruction experiments is performed as following steps. We first adapt a pre-trained source model to a target domain with 5-shot training images. We then adapt it back to the source domain with 5-shot images generated by the pre-trained source model. This step results in a reconstructed generative model for the source domain. The quantitative performances are measured between the original source model and the reconstructed model. As reported in Table. 2, our method shows significantly higher scores compared to the baselines. These results clearly validate the merits of our proposed progressive adaptation technique based on the explicit spatial matching.

4.3 User Study

Pre-trained feature extractors for quantitative evaluation, there is a difference in cognition, such as focusing more on style rather than content, unlike the human perspective (Lee, Kim, and Nam 2019). Since diversity and fidelity of adaptation results are not sufficiently evaluated only with IS and

Metric	Method	Domain		
		Amedeo	Sunglasses	Anime
Identity	IDC	5.56%	11.11%	5.56%
	RSSA	22.22%	13.89%	18.52%
	Ours	**72.22%**	**75.00%**	**75.93%**
Diversity	IDC	3.70%	3.70%	1.85%
	RSSA	24.07%	18.52%	7.41%
	Ours	**72.22%**	**77.78%**	**90.74%**
Style	IDC	31.48%	24.07%	11.11%
	RSSA	24.07%	22.22%	14.81%
	Ours	**44.44%**	**53.70%**	**74.07%**
Fidelity	IDC	14.81%	16.67%	5.56%
	RSSA	12.96%	25.93%	7.41%
	Ours	**72.22%**	**57.41%**	**87.04%**

Table 3: Results of user studies on four questions: identity, diversity, style, and fidelity. We count the voting if each method is rated by the most identity, diversity, style, and fidelity. Note that, the examples generated on 5-shot settings are shown to users.

FID, an experiment that reflects the human perspective was needed. Therefore, we conduct a user study on 3 types of domains: Amedeo, Animation, and Sunglasses with 4 questions online to validate subjectively our approaches with previous works on preserving identity, diversity, and fidelity. 54 users are required to select one of 3 images generated from methods, IDC, RSSA and Ours.

Questions To evaluate whether the identity of the source model is maintained well after adaptation, we provide images generated from the same latent of each adapted model and ask this question, "Please choose the one that resembles the one shown source image". In the diversity validation case, we provided a group of images from each model with the same latent and asked the user to "choose the group that produced the most diverse images". In case of fidelity validation, we shows 5-shot target domain training images and each group of generated images of methods. And then append 2 questions, "Please choose the one group that generated the image in the style that most closely resembles this target images." and "Please choose the image group with the highest fidelity and quality."

Results As shown in Table.3, we observe that more than 77% of participants judged that our method is most capable of preserving identity and diversity in these few-shot adaptation tasks. And over 64% of participants choose our proposed method as producing the most target-like styled and high fidelity images. In particular, from the human perspective, most of the adaptation results of our method are well-evaluated.

4.4 Ablation Study

We further investigate each component of our proposed methods, $i.e.$, aligned spacial distance consistency loss ($\mathcal{L}_{\text{ASC}}$), progressive adaptation (PA), and edge-based importance sampling (ImpS) on Animation dataset.

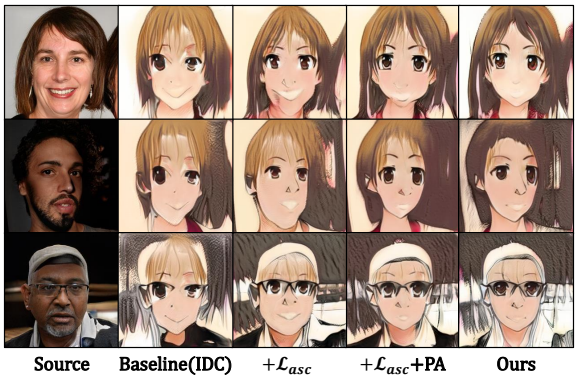


Figure 5: Qualitative ablation study of our method to verify each component of our method.Those images are from FFHQ → Animation task. Note that, our full model, Ours, is Baseline (IDC) + $\mathcal{L}_{\text{asc}}$ + PA + ImpS.

Metric	Method	Domain
		Face→Anime
FID	Baseline (IDC)	95.13
	+ $\mathcal{L}_{\text{asc}}$	93.19
	+ $\mathcal{L}_{\text{asc}}$ + PA	86.26
	+ $\mathcal{L}_{\text{asc}}$ + PA + ImpS	**84.34**

Table 4: Quantitative ablation study of our method on Animation dataset. PA and ImpS represent the progressive adaptation and importance sampling, respectively.

As shown in Fig. 5, our cross-domain aligned spatial distance consistency method is effective in preserving identity, but solely utilizing it is hard to generate an image similar to the target domain because of the large positional difference due to the domain gap. Applying progressive adaptation improves the accuracy of the feature mapping method of $\mathcal{L}_{\text{asc}}$, so generated image follows the spatial structure of the target domain. Finally, the fidelity and quality are further improved when adding importance sampling.

5 Conclusion

We proposed a novel generative model adaptation approach, progressive adaptation with align-free spatial correlation, that addresses the big domain gap between the source and target domains through a progressive learning. Specifically, we identified semantically coherent image regions between images generated by models in training at nearby iteration by exploiting their small domain gaps, and formulated to preserve the spatial correlation to enable high-quality adaptation. Moreover, we also proposed an importance sampling strategy to exclude undesired image regions for the adaptation such backgrounds and reduce the training complexity. In our extensive experiments, our method outperforms the state-of-the-art methods, in terms of diversity, fidelity, and identity preservation of the source model with much less visual artifacts.

Acknowledgements

This work was supported in part by MSIT/IITP (No. 2022-0-00680, 2020-0-00973, 2020-0-01821, and 2019-0-00421), and MSIT&KNPA/KIPoT (Police Lab 2.0, No. 210121M06).

References

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*, 35: 53–65.

Farnebäck, G. 2003. Two-Frame Motion Estimation Based on Polynomial Expansion. In *Scandinavian conference on Image analysis*, 363–370.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*.

Huang, Y.; Wang, Y.; Tai, Y.; Liu, X.; Shen, P.; Li, S.; Li, J.; and Huang, F. 2020. CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5900–5909.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4396–4405.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8107–8116.

Lee, H.; Kim, H.-E.; and Nam, H. 2019. SRM: A Style-Based Recalibration Module for Convolutional Neural Networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1854–1862.

Lee, S.; Moon, W.; and Heo, J.-P. 2022. Task Discrepancy Maximization for Fine-Grained Few-Shot Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5331–5340.

Li, B.; Zhu, Y.; Wang, Y.; Lin, C.-W.; Ghanem, B.; and Shen, L. 2021. AniGAN: Style-Guided Generative Adversarial Networks for Unsupervised Anime Face Generation. *ArXiv*, abs/2102.12593.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning Transferable Features with Deep Adaptation Networks. In *International Conference on Machine Learning*, 97–105. PMLR.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep Transfer Learning with Joint Adaptation Networks. In *International Conference on Machine Learning*, 2208–2217. PMLR.

Mo, S.; Cho, M.; and Shin, J. 2020. Freeze the Discriminator: a Simple Baseline for Fine-Tuning GANs. In *CVPR AI for Content Creation Workshop*.

Ojha, U.; Li, Y.; Lu, J.; Efros, A. A.; Lee, Y. J.; Shechtman, E.; and Zhang, R. 2021. Few-shot Image Generation via Cross-domain Correspondence. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10738–10747.

Robb, E.; Chu, W.-S.; Kumar, A.; and Huang, J.-B. 2020. Few-Shot Adaptation of Generative Adversarial Networks. *ArXiv*, abs/2010.11943.

R.Sharma, A.; and Prakash.R.Devale, P. 2012. Face Photo-Sketch Synthesis and Recognition. *International Journal of Applied Information Systems*, 1: 46–52.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. *Advances in Neural Information Processing Systems*, 29.

Wang, Y.; Gonzalez-Garcia, A.; Berga, D.; Herranz, L.; Khan, F. S.; and van de Weijer, J. 2020. MineGAN: Effective Knowledge Transfer From GANs to Target Domains With Few Images. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9329–9338.

Xiao, J.; Li, L.; Wang, C.; Zha, Z.-J.; and Huang, Q. 2022. Few Shot Generative Model Adaption via Relaxed Spatial Structural Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11204–11213.

Yazıcı, Y.; Foo, C.-S.; Winkler, S.; Yap, K.-H.; Piliouras, G.; and Chandrasekhar, V. 2019. The Unusual Effectiveness of Averaging in GAN Training. In *International Conference on Learning Representations*.