A Closer Look at Few-Shot Image Generation

Supplementary

Overview

This Supplementary provides additional experiments and results to further support our main finding and proposed method for few-shot image generation. The Supplementary materials are organized as follows:

- Section A: Limitations
- Section B: Potential Social/Ethic Impacts
- Section C: Additional Details for Binary Classifier
- Section D: Pseudo-code for Intra-LPIPS
- Section E: Additional Training Details for DCL
- Section F: Proof of MI Maximization
- Section G: Additional Evaluation Metric
- Section H: Additional Results for Source $\mapsto$ Target Adaptation
- Section I: Effect of Unrelated Source $\mapsto$ Target Adaptation
- Section J: Effect of the Target Data Size
- Section K: Discussion of the Design Choice of DCL

A. Limitations

We follow exactly previous work (*e.g.*, [5]) in the choices of domains and datasets for fair comparison. However, given the extremely wide range of domains to which few-shot image generation can be applied, it is not feasible for us to validate our findings for all possible domains. On the other hand, our comprehensive qualitative and quantitative experiment results supported by our analysis provide supportive evidence that our findings could be generalized for other domains.

Furthermore, similar to existing work [2, 5], our main focus is on related source/target domains, while we also discuss some analysis on unrelated source/target domains, *e.g.*, see details in Figures S3.

B. Potential Social/Ethic Impacts

In this work, we adhere to the general ethical conducts and guidelines, including that we use the publicly available datasets to conduct all of our experiments, without any personally identifiable information or sensitive identifiable information (*e.g.*, name of the human data). However, since real images are used for transfer learning, we hope the community could take the privacy issue carefully and seriously.

C. Additional Details for Binary Classifier

In this section, we provide more details of how to build the binary classifier C (see Sec. 4.2 in the main paper) for quality/realisticness evaluation for different methods, during the few-shot adaptation.

Dataset. As mentioned in the main paper, we aim to build the unbiased binary classifier C by keeping the training data of the source and the target domain balanced. Note that the data used for training the binary classifier is unseen during few-shot adaptation. We summarize the dataset setups and the data link in Table S1.

Optimization. In the training phase of C, we randomly initialize the AlexNet with official Pytorch implementation, and we employ the Adam optimizer with binary cross-entropy loss to optimize the weights. We train it until convergence on each dataset.

Table S1. We provide the training data of different source $\mapsto$ target adaptation setups for training the binary classifier C.

Source $\mapsto$ Target	Source data	Target data	Size
FFHQ $\mapsto$ Sketches	Link	Link	$\sim$ 300
FFHQ $\mapsto$ Babies	Link	Link	$\sim$ 2700
FFHQ $\mapsto$ Sunglasses	Link	Link	$\sim$ 2500

D. Pseudo-code for Intra-LPIPS

Intra-LPIPS [5] evaluates to what extent the generated images collapse to the few-shot target data. The detailed text description of Intra-LPIPS can be found in Sec. 4.3 in the main paper. In this section, we provide the pseudo-code to compute the intra-LPIPS for evaluating the diversity-degradation during few-shot adaptation, as Algorithm 1.

E. Additional Training Details for DCL

We follow the previous work [2, 3, 5] to use the architecture of StyleGAN-V2 with pytorch implementation [1]. We

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

Algorithm 1 Pseudo-code of Intra-LPIPS

```
# Input: 1. Generated images X=[x1, ..., xn];
#        # suppose we have 2-shot target samples;
#        2. Cluster center: c0, c1;
#        3. Cluster_0, Cluster_1 = [], []
# Output: Avg Intra-LPIPS over 2 clusters
# ------------------------------------------ #
# Step 0. Define the LPIPS function
lpips_fn = lpips.LPIPS(net='vgg')

# Step 1. Assign images to the closet center
for X[i] in X:
    dist0 = lpips_fn(X[i], c0)
    dist1 = lpips_fn(X[i], c1)
    if dist0 < dist1:
        Cluster_0.append(X[i])
    else:
        Cluster_1.append(X[i])

# Step 2. Compute Intra-LPIPS
lpips_dist = []
While not done:
    for img_i, img_j in Cluster_0:
        lpips_dist.append(lpips_fn(img_i, img_j))
    for img_i, img_j in Cluster_1:
        lpips_dist.append(lpips_fn(img_i, img_j))
return lpips_dist.mean()
# ------------------------------------------ #
```

use Adam optimizer to optimize the generator and the discriminator, and use the same hyperparameters and settings in [5], including the non-saturating loss $\mathcal{L}_{adv}$. For the image resolution applied in this work, except for the adaptation setup "Cars $\rightarrow$ Wrecked cars", in which we adopts the 512 $\times$ 512 (this is because the GAN pretrained on LSUN Cars adopts the 512 $\times$ 512 image resolution), we use 256 $\times$ 256 for other adaptation setups in both the pretraining and the adaptation stage. We run our experiments (including those in Sec. 4) on a single Tesla V100 GPU.

F. Proof of MI Maximization

Under mild assumptions, our proposed DCL (see Sec. 5) maximizes the lower bound of mutual information (MI) between generated samples with the same noise input, of the source and the target generator, respectively [6].

In this section, we show the proof of this statement in the main paper. We use $\mathcal{L}_{CL_1}$ (with expectation) for example and show that, $\text{MI}(G_t(z_i); G_s(z_i)) \geq \log[N] - \mathcal{L}_{CL_1}$, where

$$\mathcal{L}_{CL_1} = \mathbb{E}_X - \log \frac{f(G_t(z_i), G_s(z_i))}{\sum_{j=1}^{N} f(G_t(z_i), G_s(z_j))} \tag{1}$$

To make it concise, in this section we omit the layer index l used in the main paper. We let $X = \{G_t(z_1), G_t(z_2), \ldots, G_t(z_N)\}$. Follow [6], we write the optimal probability of this objective function (Eqn. 1) as $p(d = i|X, G_s(z_i))$ where $[d = i]$ indicates that the sample X_i is the 'positive' sample $G_t(z_i)$ that corresponds to $G_s(z_i)$. The probability of the generated image which is sampled from $p(G_t(z_i)|G_s(z_i))$, rather than the random generated image distribution, can be shown as follows:

$$p(d = i|X, G_s(z_i)) = \frac{p(d = i, X|G_s(z_i))}{\sum_{j=1}^{N} p(d = j, X|G_s(z_i))} \tag{2}$$

$$= \frac{p(X_i|G_s(z_i)) \prod_{k \neq i} p(X_k)}{\sum_{j=1}^{N} p(X_i|G_s(z_i)) \prod_{k \neq j} p(X_k)} \tag{3}$$

$$= \frac{\frac{p(X_i|G_s(z_i))}{p(X_i)}}{\sum_{j=1}^{N} \frac{p(X_j|G_s(z_i))}{p(X_j)}}. \tag{4}$$

Therefore, $f(X_i, G_s(z_i))$ is proportional to $\frac{p(X_i|G_s(z_i))}{p(X_i)}$. Then, we argue that DCL is a lower bound of the MI between $G_s(z_i)$ from the source generator, and $G_t(z_i)$ from the target generator, which adopt the same noise vector z_i. This can be shown as follows.

$$\mathcal{L}_{CL_1} = \mathbb{E}_X - \log \left\{ \frac{\frac{p(X_i|G_s(z_i))}{p(X_i)}}{\sum_{j=1}^{N} \frac{p(X_j|G_s(z_i))}{p(X_j)}} \right\} \tag{5}$$

$$= \mathbb{E}_X - \log \left\{ \frac{\frac{p(X_i|G_s(z_i))}{p(X_i)}}{\frac{p(X_i|G_s(z_i))}{p(X_i)} + \sum_{x \in Neg} \frac{p(X_j|G_s(z_i))}{p(X_j)}} \right\} \tag{6}$$

$$= \mathbb{E}_X \log \left\{ 1 + \frac{p(X_i)}{p(X_i|G_s(z_i))} \sum_{x \in Neg} \frac{p(X_j|G_s(z_i))}{p(X_j)} \right\} \tag{7}$$

$$= \mathbb{E}_X \log \left\{ 1 + \frac{p(X_i)}{p(X_i|G_s(z_i))} \mathbb{E} \frac{p(X_j|G_s(z_i))}{p(X_j)} (N-1) \right\} \tag{8}$$

$$= \mathbb{E}_X \log \left\{ 1 + \frac{p(X_i)}{p(X_i|G_s(z_i))} (N-1) \right\} \tag{9}$$

$$= \mathbb{E}_X \log \left\{ \frac{p(X_i|G_s(z_i)) - p(X_i)}{p(X_i|y)} + N \frac{p(X_i)}{p(X_i|G_s(z_i))} \right\} \tag{10}$$

$$\geq \mathbb{E}_X \log \left\{ N \frac{p(X_i)}{p(X_i|G_s(z_i))} \right\} \tag{11}$$

$$= \log[N] - MI(G_t(z_i); G_s(z_i)) \tag{12}$$

Therefore, we have $MI(G_s(z_i); G_s(z_i)) \geq \log[N] - \mathcal{L}_{CL_1}$, which means the Eqn. 1 is a lower bound of the mutual information between $G_s(z_i)$ and $G_t(z_i)$.

G. Additional Evaluation Metric

Standard LPIPS. In the main paper, we use intra-LPIPS [5] to evaluate the diversity (degradation) of the target generator for different methods. Here, we provide the standard

LPIPS (↑) results in order to have a comprehensive comparison. In Table S2, we show that, we still outperform other models. Different from intra-LPIPS, the standard LPIPS only evaluate if the generated images are different from each other, and does not evaluate if they collapse to the few-shot training samples, hence we do not include the result of standard LPIPS in the main paper.

Table S2. Standard Pair-wise LPIPS distance (↑) of generated fake images. We firstly generate abundant data using the adapted generator on the target domain, then we compute the average perceptual distance between randomly paired images [9].

	Church ↦ Haunted house	FFHQ ↦ Amedeo's paintings	FFHQ ↦ Sketches
TGAN [8]	0.57 ± 0.06	0.58 ± 0.12	0.44 ± 0.07
TGAN+ADA [1]	0.60 ± 0.05	0.61 ± 0.11	0.45 ± 0.08
BSA [4]	0.47 ± 0.05	0.45 ± 0.07	0.32 ± 0.05
FreezeD [3]	0.55 ± 0.08	0.55 ± 0.13	0.42 ± 0.09
MineGAN [7]	0.56 ± 0.10	0.59 ± 0.12	0.46 ± 0.09
EWC [2]	0.59 ± 0.06	0.60 ± 0.09	0.45 ± 0.06
CDC [5]	0.61 ± 0.03	0.62 ± 0.06	0.47 ± 0.05
DCL (Ours)	$\mathbf{0.63 \pm 0.03}$	$\mathbf{0.64 \pm 0.06}$	$\mathbf{0.51 \pm 0.05}$

Table S3. Standard LPIPS (↑) evaluation. We use the same setup as evaluating the intra-LPIPS (see Table 2 in the main paper.)

H. Additional Results of Source ↦ Target Adaptation

In this section, we perform additional source ↦ target adaptation experiments to visualize the effectiveness of our method. In Figure S2, compared to the source domain images, the generated samples on the target domain preserve rich semantic features (*e.g.*, hair style, hat, building structure) on the source, but capture the style (and accessories) of the few-shot target set, which further confirm our ideas proposed in this work.

I. Effect of Unrelated Source ↦ Target Adaptation

Background. In this work, we mainly focus on the few-shot image generation (with GAN adaptation) where the source domain and the target domain are related, similar to all existing methods [2, 3, 5, 8]. However, the case where the source and the target domain are unrelated should be included in the discussion, *e.g.*, transferring from FFHQ (human face) to Haunted Houses.

Experiments. In this section, we compare with other methods (see related works in the main paper) with the setup that the source domain and the target domain are unrelated. Note that all other settings are identical to Sec. 6 in the main paper. In Figure S3, we adapt two source domains (FFHQ, LSUN Church) to three different target domains (Haunted house, Amedeo's paintings and Van Gogh's house). The differences between these methods are more obvious, as discussed in Figure S3.

Nevertheless, these methods cannot accurately capture the target domain distribution **with much diversity knowledge**, as what we expect when the source domains and the target domains are related. To the best of our knowledge, there is no existing work that focus on this issue and we leave this open problem as our future work.

J. Effect of the Target Data Size

In the main paper, we mainly focus on the 10-shot adaptation setups, in both Sec. 4 and Sec. 6. Here, we extend our analysis to 5-shot and 1-shot setups. As Figure S4 and Figure S5, we show that, our main analysis is still hold for 1-shot and 5-shot adaptation case: while some methods have disproportionate focus on diversity preserving which impede quality improvement, they will achieve almost the identical realisticness on the target domain, **even for 1-shot and 5-shot adaptation**. Therefore, we argue that the main focus of few-shot image generation method should be on reducing the diversity degradation during few-shot adaptation.

K. Discussion of the Design Choice of DCL

Choice of coefficients. In Eqn. 8 in the main paper, there are two coefficients: λ_1 and λ_2 in the loss term of DCL. We perform a grid search to tune these hyperparameters, depending on the performance on diversity and FID score. In experiments, we empirically find that the setting $\lambda_1 = 2$, $\lambda_2 = 0.5$ achieves the best result.

Choice of batch size. For our proposed DCL, in Generator CL, the batch size depends on how many noise vectors we sample in each iteration. In Discriminator CL, the batch size depends on how many real samples we have in the few-shot adaptation. Therefore, for fair comparison, we sample 4 noise vectors as input in each iteration, which is identical to other methods, while we sample all few-shot real target images (*e.g.*, 10-shot) to perform Discriminator CL.

Choice of negative samples. The negative samples can be selected from various sources for both Generator CL and Discriminator CL. In Generator CL, the negative samples are $G_s(z_{j\neq i})$ where z_i is used to produce G_{z_i} (**Setup A**). However, the negative samples can also be $G_t(z_{j\neq i})$ to prevent all generated images of the adapted generator collapsing to the same mode (**Setup B**). Empirically, we find that the both setups has similar performance on reducing the loss of diversity during adaptation, as we show the change of intra-LPIPS in Figure S1.

For Discriminator CL, we aim to prevent the generated images collapsing to real target data, as observed in other methods (*e.g.*, TGAN). Therefore, we sample discriminating features from real target data (*i.e.*, $D_t(x)$) as negative samples to regularize the few-shot adaptation process. Potentially, the negative samples can also come from the generated images. However, in experiments we do not observe better performance with this setup.

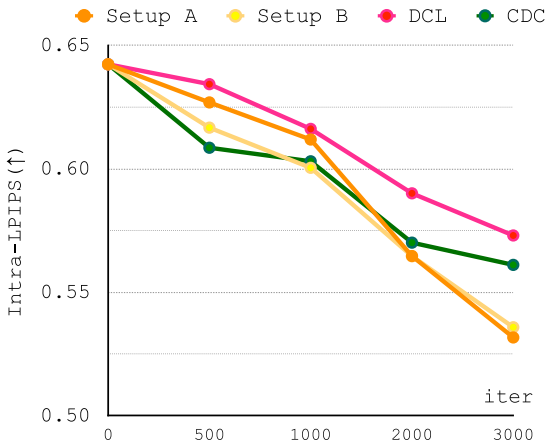


Figure S1. Transferring from FFHQ $\mapsto$ 10-shot Amedeo's paintings (the same setup as Figure 1 in the main paper). We show that both Setup A and Setup B have similar performance on mitigating the loss of diversity during few-shot adaptation. Note that we do not use Discriminator CL in this ablation study.

References

[1] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020. 3

[2] Yijun Li, Richard Zhang, Jingwan (Cynthia) Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15885–15896. Curran Associates, Inc., 2020. 1, 3

[3] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. In *CVPR AI for Content Creation Workshop*, 2020. 1, 3

[4] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2750–2758, 2019. 3

[5] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021. 1, 2, 3

[6] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[7] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341, 2020. 3

[8] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 218–234, 2018. 3

[9] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3

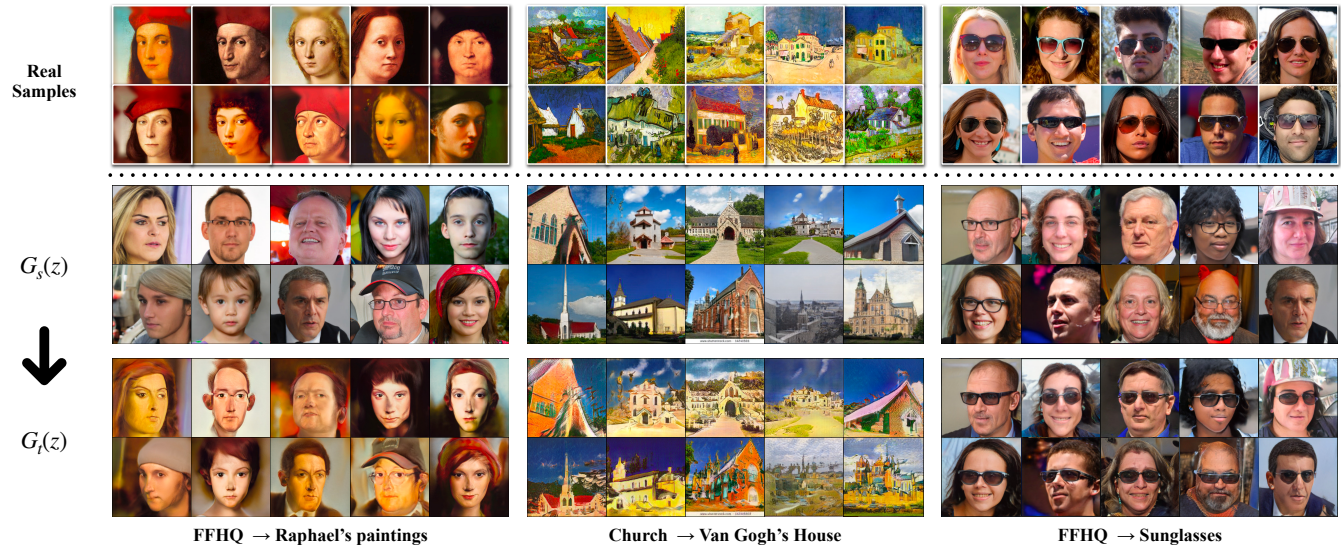
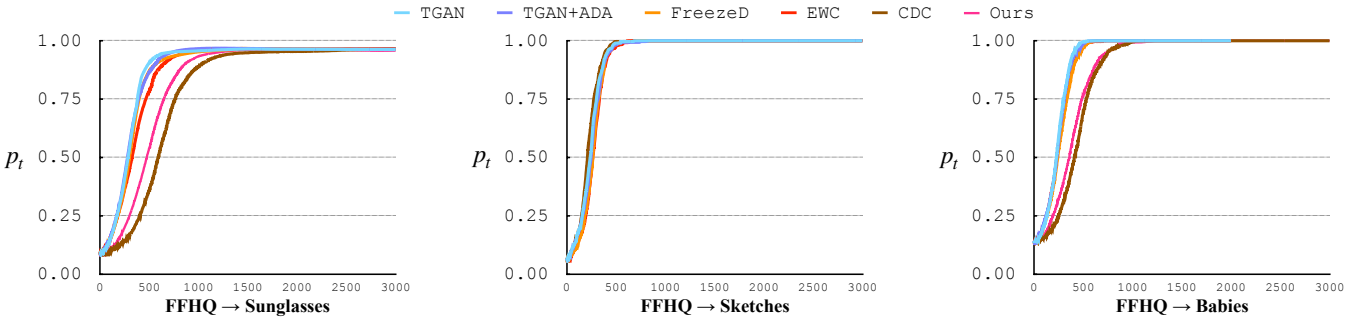
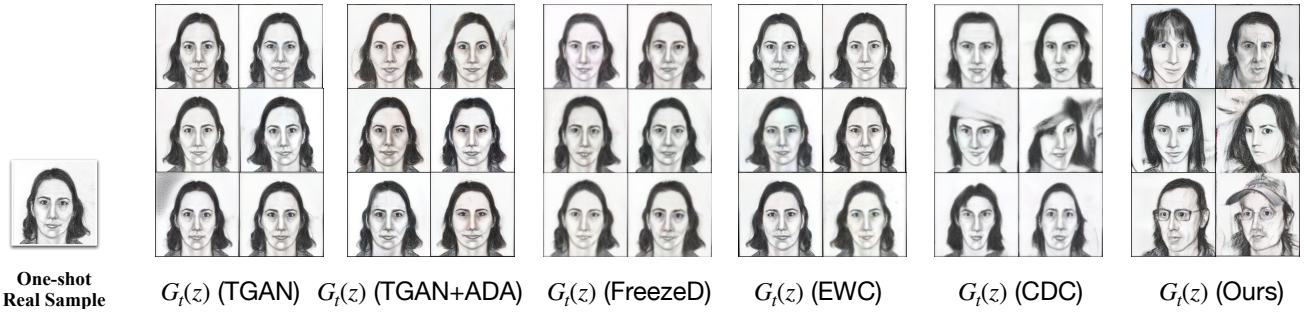


Figure S2. Generated images with additional source → target adaptation setups, which is an extension to the Figure 5 in the main paper. We show that, while preserving the meaningful semantic features on the source domain (high level: rough structures, face appearance, and huamn postures, middle level: hair style, hat and color), the generated images after adaptation are able to capture good style or texture information on the target domain.

Figure S3. Generated images with **unrelated** source → target adaptation setups. We show that, TGAN still overfits the few-shot target set regardless of the source domain knowledge; CDC preserves the distance between instances in the source, therefore it captures the part-level correspondence between the source and the target, *e.g.*, the eyes and the teeth are roughly mapped to the doors and windows in the target domain. In contrast, since DCL (Ours) emphasizes on the connection to the generated image on the source domain with the same latent code, our results preserve more refined details (*e.g.*, glasses, hair style) and the structure appearance is not thoroughly destroyed when transferring to the target domain.

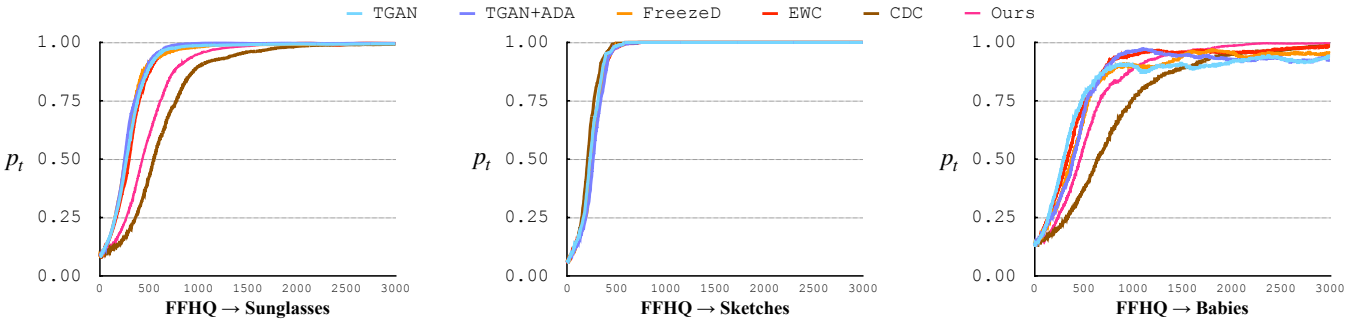


(a) 1-shot quality/realisticness evaluation.

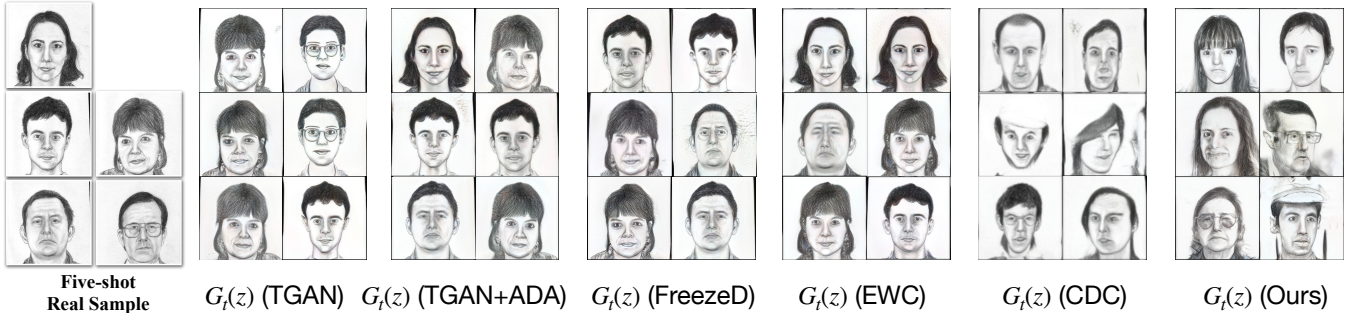


(b) 1-shot visualization from FFHQ ↦ Sketches.

Figure S4. Generated images with 1-shot adaptation (the same setup as Section 4 in the main paper), which is an extension to the Figure 2 in the main paper.



(a) 5-shot quality/realisticness evaluation.



(b) 5-shot visualization from FFHQ ↦ Sketches.

Figure S5. Generated images with 5-shot adaptation (the same setup as Section 4 in the main paper), which is an extension to the Figure 2 in the main paper.