Few-shot Image Generation via Cross-domain Correspondence Supplementary Material

Utkarsh Ojha[1,2] Yijun Li[1] Jingwan Lu[1] Alexei A. Efros[1,3]
Yong Jae Lee[2] Eli Shechtman[1] Richard Zhang[1]

[1]Adobe Research [2]UC Davis [3]UC Berkeley

1. Overview

This document provides additional information about the proposed method. First, we continue the discussion on training and architecture details of our method and the baselines. We use the official pre-trained models for Church, Cars and Horses [3], and a 256 x 256 resolution pre-trained model for FFHQ.[1]. The generator's (G) and the discriminator's (D_{img}) architecture are the same as StyleGAN2. For D_{pch}, we consider the first l layers of D_{img}, and convert the corresponding feature to a $N \times N$ output, where each member's receptive field corresponds to a patch in the input image. We use Adam optimizer [4], and use the rest of the hyperparameters (e.g. learing rate) from [3]. The ideal training duration for adapting to different target domains is as follows: for domains whose appearance matches very closely with the source, i.e. babies/sunglasses with FFHQ as source, we get good results within 1000 iterations. For other face-based target domains (e.g. Modigliani's paintings), we train our models for 5000 iterations to get decent results. For more complex domains (e.g. landscape drawings), we observe our best results at around 10000 iterations. Note that in eq. 4 (main paper), we have used $z \sim p_z(z)$ instead of $z \sim p_z(z) - Z_{anch}$ for the second term. This is because if an image from Z_{anch} is (globally) realistic, its patches will be realistic as well. Also note that the other way around is not true.

Details about MineGAN We use the publicly available code of MineGAN[2] to produce its results (Fig. 4, Tab 1/2). However, since there is an involvement of an additional *miner* network during the adaptation process, we tried experimenting with smaller networks (than the default) for the extreme few-shot setting (10 training images). The results obtained were similar to the default setting – FID: 96.72, 68.67 for babies and sunglasses domain respectively, suggesting that reducing the complexity of the miner network alone is not sufficient to achieve better results.

FFHQ → face domains Fig. 1 shows the real images used for different target domains in our experiments (apart from those presented in the main paper). Next, we show the results of translating a source model trained on natural faces (FFHQ) to different kinds of target domains. We observe the diversity in the generated images, which come as a result of preserving correspondence between the source and target distribution. Fig. 3 shows more examples for the idea discussed in Fig. 8 of the main paper. These four caricature images are unseen during the adaptation of a source model X (X is FFHQ/Church/Cars/Horses) to the caricature domain. We again observe that adapting FFHQ (natural faces) to the caricature domain best embeds and reconstructs unseen images, indicating that caricature as a domain is most related to FFHQ than any other source domain. Fig. 4 presents an extension of Fig. 9 of main paper, where we study the 1-shot, 5-shot and 10-shot setting for two baselines, and compare it with our method. We notice that both FreezeD and EWC, overfit to the target sample in 1-shot setting, generating virtually identical sketches/scenes. This trend of overfitting for these baselines continues in 5/10-shot settings as well, where generations collapse to small variations around a few modes. Our method, on the other hand, takes the benefit of increasing training data size, by learning to generate more and more diverse samples, different from the images used for training.

Visualizing the clusters Fig. 5 visualizes the clustering-based diversity assessment introduced in Sec. 4.1 of the main paper. We group the generated images from a method into k clusters, with the k training images serving as the cluster center. After this, we study the resulting clusters, where we visualize how similar is (i) the closest member to the center (measured via LPIPS), (ii) the farthest member to the center. The intuition is that a method whose generated images overfit to the training data will result in clus-

[1]We use https://github.com/rosinality/stylegan2-pytorch
[2]https://github.com/yaxingwang/MineGAN/tree/master/styleGAN

ters where the closest member is very similar to the corresponding cluster center. Each column deals with one cluster, where the cluster centers (real images used for training) are shown in the middle. The top half of the figure visualizes the closest members for different methods, whereas the bottom half visualizes the farthest ones. When no images get assigned to a cluster, the concept of closest/farthest members doesn't apply, and we depict this with a red cross. Summarily, we observe that the closest members from TGAN/EWC are much more similar to the corresponding center than our method, whose even closest members are visually distinct. This observation also helps explain the better performance of our method compared to others in Table 2 (main paper).

Hand gestures experiment We find the property of emerging correspondences within seemingly unrelated source/target domains interesting, and hence for creativity purposes, take a further step to explore the idea. We collect images of arbitrary hand gestures being performed over a plain surface, and train a *source* model from scratch using that dataset. Next, we adapt it to various domains such as landscapes, fire, maps. During inference, we observe different aspects of the target domains a pair of hands can control (e.g. structure of river/islands). Please see our teaser video, which shows the correspondence results in this case, as well as better explains the benefits of our method in previously discussed scenarios (e.g. FFHQ $\rightarrow$ caricatures, Church $\rightarrow$ Van Gogh houses).

Precision and recall metrics A limitation of FID [1] is that it packs two aspects of the generated images, sample quality and diversity, into one score. This makes it difficult to disentangle and study the two properties separately. To overcome this, density and coverage metrics were proposed to evaluate the generative models [7]. In some feature space (e.g. CNN embeddings), density measures how many real-sample neighbourhood regions contain a fake sample. Coverage, in the same space, measures the ratio of real samples whose neighbourhood contains at least one fake sample. In both the definitions, *neighbourhood* is defined as a spherical region around a real sample, with its radius given by the distance from the next nearest real sample. A high score for both the metrics is preferred. Density is unbounded, whereas coverage is bounded by 1. We present evaluation of the baselines fare using these metrics on FFHQ babies dataset in Table 1. We observe that MineGAN achieves a superior *density* score, i.e. quality of the generated image, but suffers in the *coverage* aspect. This is again an indication of mode collapse to a small number of high quality samples. Our method achieves a better balance between the quality as well as diversity of the generated samples. Note that this result is in alignment with the one presented in Table 2 (main paper), which studies diversity among the generated samples in a different way.

	Density	Coverage
TGAN [9]	0.379	0.250
TGAN+ADA [2]	0.434	0.285
FreezeD [6]	0.418	0.217
MineGAN [8]	**0.803**	0.125
EWC [5]	0.301	0.325
Ours	0.690	**0.467**

Table 1: Density (↑) and Coverage (↑) scores for FFHQ babies.

References

[1] 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 *Adv. Neural Inform. Process. Syst.*, 2017.

[2] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Adv. Neural Inform. Process. Syst.*, 2020.

[3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958*, 2019.

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent.*, 2015.

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

[6] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze discriminator: A simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020.

[7] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, 2020.

[8] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[9] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Eur. Conf. Comput. Vis.*, 2018.

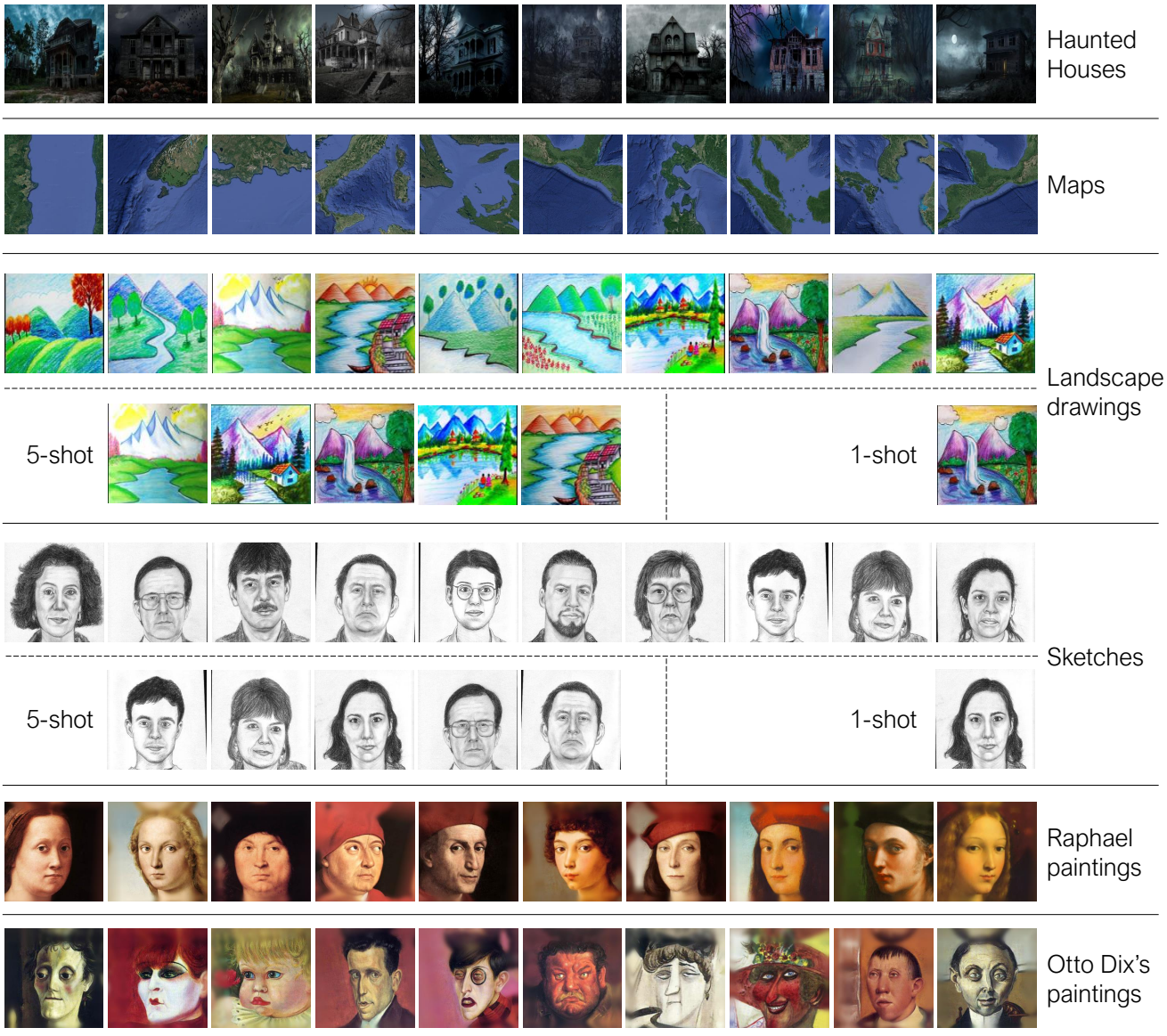

Figure 1: Real images used for translating a source model to different target domains. For landscape drawings and sketches, we have shown the images used in 1-shot and 5-shot scenario, for the experiment presented in Fig. 9 of the main paper, and Fig. 4 of this document.

Figure 2: Translating the FFHQ source model to different target domains. The noise vector is kept same across the columns, so that we can study the relation between the corresponding source and target image.

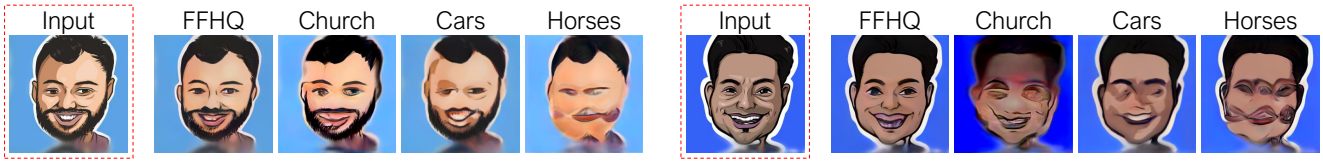

Figure 3: Embedding unseen caricature images into models adapted from different source to the same target domain (caricature). We observe that FFHQ $\rightarrow$ caricatures best captures the caricature properties, resulting in best reconstructions.

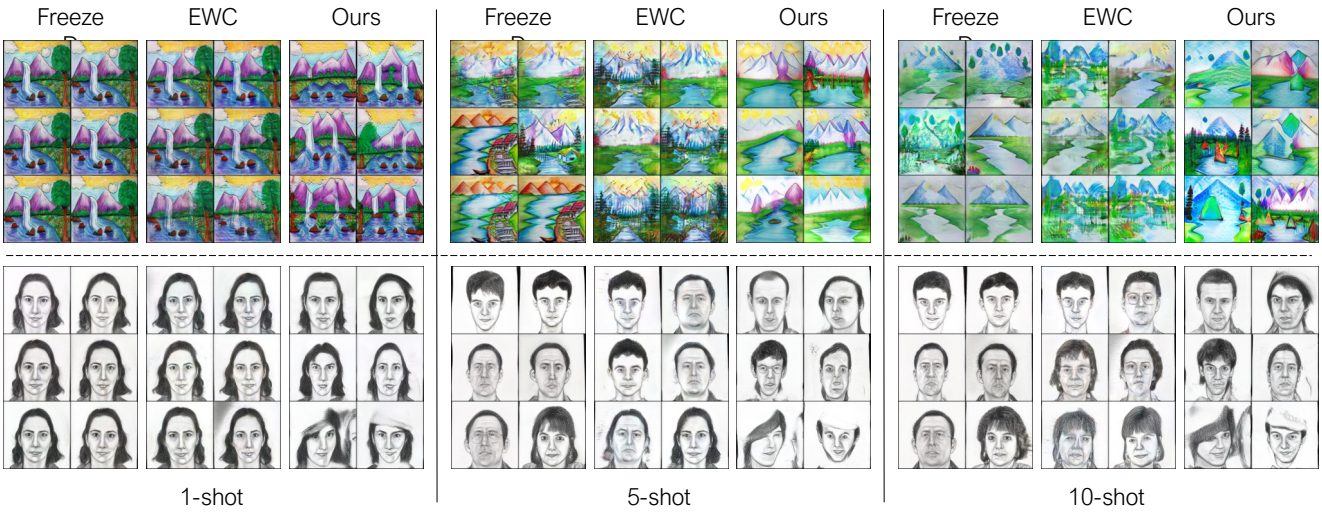

Figure 4: Comparison of our method compared to EWC [5] and FreezeD [6] in 1-shot, 5-shot and 10-shot setting.

Figure 5: Visualizing the clusters formed using the technique described in Sec. 4.1 of the main paper. The closest members produced by TGAN/EWC are much more similar to the corresponding cluster center than our method, indicating that the generations using the proposed method possess more diversity.