FEW-SHOT CROSS-DOMAIN IMAGE GENERATION VIA INFERENCE-TIME LATENT-CODE LEARNING APPENDIX

Arnab Kumar Mondal[*], **Piyush Tiwary**[†], **Parag Singla**[*] **& Prathosh AP**[†]
[*] IIT Delhi, [†] IISc Benglauru

A IMPACT OF k ON GENERATION FOR k-SHOT ADAPTATION

In the main paper, we have provided quantitative comparison of generation quality as k in k-shot generation is varied from 1 to 10. Here, we present some visual samples of 1-shot and 5-shot adaptation for FFHQ $\rightarrow$ Babies. Figure 1d and 2d presents newly generated baby samples using the proposed method under 1-shot and 5-shot settings respectively. As can be seen variation in the generated images increases in the 5-shot adaptation. As compared to the current SOTA CDC Ojha et al. (2021) and RSSA Xiao et al. (2022), our method generates images that are more crisp and exhibits more variety.

Further we quantitatively analyze the impact of higher number of shot on generation quality using babies dataset. From Table 1, we note that though the relative gain of our method compared to baselines decreases as the number of shots is increased (which is expected, since other approaches can now better fine-tune the source generator), we still outperform the baselines. This points to the robustness of our approach across varying number of shots. We note that the best relative advantage of our approach is seen at a very small number of shots.

Table 1: Impact of higher number of shots on the generation quality.

Method	20-shot	100-shot
CDC Ojha et al. (2021)	57.69	39.32
RSSA Xiao et al. (2022)	59.14	41.56
Proposed Method	**54.98**	**39.27**

B IMPACT OF UNRELATED SOURCE DOMAIN

To analyze the effect of an unrelated source domain on adaptation to a target domain, we consider the following two settings: (i) FFHQ $\rightarrow$ Haunted House and (ii) Church $\rightarrow$ Sketches. Figure 3 presents the qualitative results. As can be seen, our method fails to capture the target domain distribution correctly. When the inference time optimization is solved for 250 steps the model underfits the 10-shot target domain examples and generates meaningless images. However, upon solving the inference time optimization for more iterations the model starts to overfit and starts mimicking the few show examples.

C QUALITATIVE RESULTS FOR FFHQ $\rightarrow$ ARTISTIC FACE ADAPTATION

In this supplementary material we present more visual samples of different target domains (different from what has been presented in the main paper) for qualitative evaluation. Specifically, we adapt a FFHQ source generator to the artistic style of Raphael (Fig 4), and Otto Dix (Fig 5) using 10 training examples. In Figure 4 and 5, sub-figure (a) presents 10 shot training examples, sub-figure (b) presents samples generated using CDC Ojha et al. (2021), sub-figure (c) presents samples generated using RSSA Xiao et al. (2022), and sub-figure (d) presents samples generated using our proposed method. As can be seen, the images generated using our proposed method are crisp and diverse and mimics the target domain closely.

(a) 1-shot training example of baby for adaptation.

(b) Images generated using CDC Ojha et al. (2021).

(c) Images generated using RSSA Xiao et al. (2022).

(d) Images generated using proposed method.

Figure 1: Examples of images generated after 1-shot adaptation.

D EDITING USING THE LEARNT LATENT CODES

We have performed experiments to edit generated images using a closed-form unsupervised algorithm, SeFa Shen & Zhou (2021). The objective of this experiment is to see check whether the discovered latent manifold is semantically meaningful, smooth and support editing. The results of our experiment on two datasets (babies and sketches) are presented respectively in Figure 6 and Figure 7.

(a) 5-shot training example of baby for adaptation.

(b) Images generated using CDC Ojha et al. (2021).

(c) Images generated using RSSA Xiao et al. (2022).

(d) Images generated using proposed method.

Figure 2: Examples of images generated after 5-shot adaptation.

E GENERALIZATION ACROSS MULTIPLE IMAGE RESOLUTION

To validate whether the proposed method generalizes well across different resolutions and complex datasets, we have performed several additional experiments.

1. **Example 1 (64 x 64 images)** - To validate our method in lower resolution images, we have performed 10-shot adaptation to CeleB-A (Liu et al., 2015) Female dataset using FFHQ source model. We have used 64 x 64 sized images for this experiment. The results can be found in Figure 8.

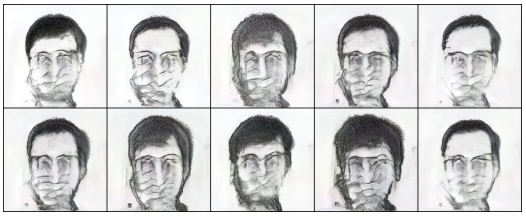
(a) Source Generator trained using FFHQ images.

(b) Source Generator trained using Church images.

(c) 10-shot training examples of haunted houses.

(d) 10-shot training examples of haunted houses.

(e) FFHQ → Haunted House. After 250 iterations.



(f) Church → Sketches. After 250 iterations.

(g) FFHQ → Haunted House. After 1000 iterations.

(h) Church → Sketches. After 1000 iterations.

(i) FFHQ → Haunted House. After 2000 iterations.

(j) Church → Sketches. After 2000 iterations.

Figure 3: Visualizing the impact of unrelated source domains.

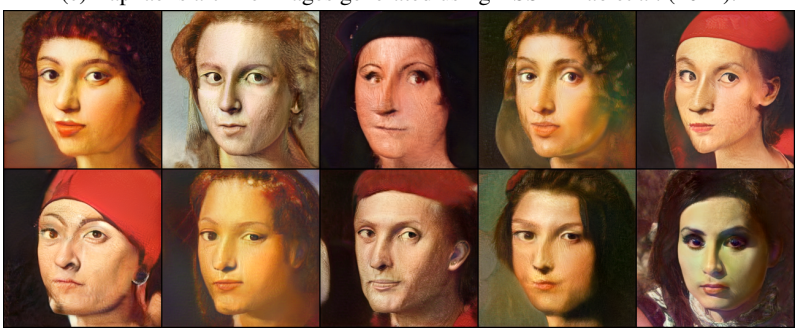
2. **Example 2 (512 x 512 images)** - Next, we have adapted a source GAN trained to generate CARs[1] to target domain wrecked cars (Ojha et al., 2021). Images in this experimental setting are of 512 x 512 resolution. The visual examples of this experiment are presented in Figure 9.

[1] `https://nvlabs-fi-cdn.nvidia.com/stylegan2/networks/stylegan2-car-config-f.pkl`

(a) 10-shot training examples of Raphael's art.

(b) Raphael's art like images generated using CDC Ojha et al. (2021).

(c) Raphael's art like images generated using RSSA Xiao et al. (2022).



(d) Raphael's art like images generated using our proposed method.

Figure 4: FFHQ $\rightarrow$ Raphael's Art 10-shot adaptation.

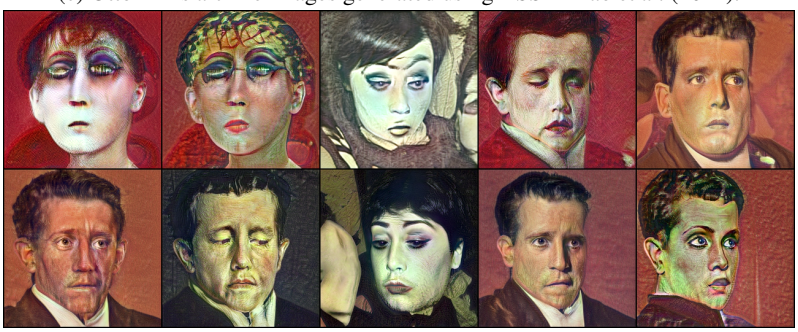
3. **Example 3 (1024 x 1024 images)** - We have used the StyleGAN2 trained on the map-dataset[2] as the source and floor-plan as the target. In this experiment the images are of size 1024 x 1024. Refer to Figure 10 for visual samples.

[2] https://github.com/justinpinkney/awesome-pretrained-stylegan2#maps

(a) 10-shot training examples of Otto Dix's art.

(b) Otto Dix's art like images generated using CDC Ojha et al. (2021).

(c) Otto Dix's art like images generated using RSSA Xiao et al. (2022).



(d) Otto Dix's art like images generated using our proposed method.

Figure 5: FFHQ $\rightarrow$ Otto Dix's Art 10-shot adaptation.

The results (Figure 8, 9, and 10) demonstrates that generative domain adaptation using the proposed method is possible across multiple resolutions and complex datasets.

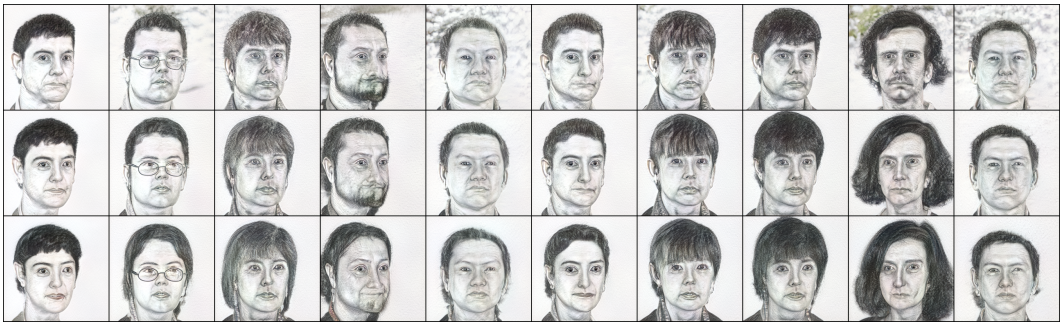
(a) The middle row presents the generated babies image using our method. In the top row smile appeared in the face and in the bottom row fear appeared in the expression.

(b) The middle row presents the generated babies image using our method. The pose is changing from left orientation (top row) of face to right orientation of face (bottom row).

(c) The middle row presents the generated babies image using our method. Age is changing from older (top row) to younger (bottom row) in each row.

Figure 6: Editing generated baby images. The middle row are the images generated using our method. The top row and bottom row are edited images.

(a) The middle row presents the generated sketch images using our method. Age is changing from older (top row) to younger (bottom row) in each row.



(b) The middle row presents the generated sketch images using our method. The gender is changing from male (top row) of face to female (bottom row).

(c) The middle row presents the generated sketch images using our method. Face cut is changing in each row from wide (top row) to narrow (bottom row).

Figure 7: Editing generated sketches. The middle row are the images generated using our method. The top row and bottom row are edited images.

(a) 10-shot examples of CelebA female.

(b) CelebA females generated using the proposed method.

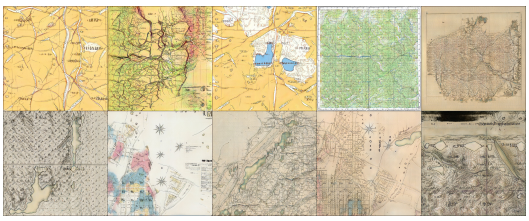
Figure 8: 10 shot adaptation to CelebA (Liu et al., 2015) Female. Image resolution is 64×64.

(a) Generated images using a source GAN, trained on cars dataset.

(b) Wrecked cars 10-shot training examples.

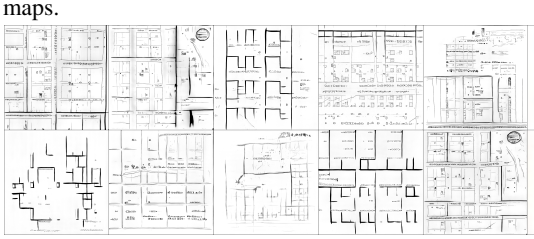
(c) Wrecked cars generated using proposed method.

Figure 9: 10 shot adaptation to wrecked cars (Ojha et al., 2021) from cars. Resolution of images is 512×512.

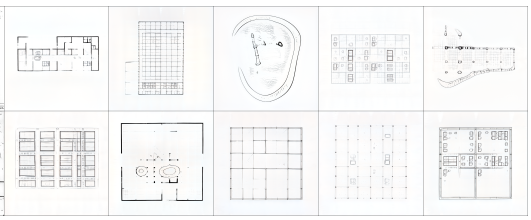


(a) Generated images using a source GAN, trained on maps.

(b) Floor plans 10-shot training examples.



(c) Floor plans generated using proposed method.



(d) Floor plans generated using a StyleGAN trained on large-scale floor plan dataset.

Figure 10: 10 shot adaptation to floor plans from maps. Resolution of images is 1024×1024.

F Details of Image Translation

For the task of Image Translation we have used three additional loss components apart from the original style loss and adversarial loss for the training of latent learner. These three components are as follows:

$$\mathcal{L}_{struct}^{source} = \mathop{\mathbb{E}}_{\chi}\left[\sum_{l} \zeta_l \left[\|S^l(L_{\theta_L}(\chi)) - S^l(\boldsymbol{w}_s)\|_2\right]\right] \tag{1}$$

$$\mathcal{L}_{struct}^{target} = \mathop{\mathbb{E}}_{\chi}\left[\sum_{l} \zeta_l \left[\|S^l(L_{\theta_L}(\chi)) - S^l(\boldsymbol{w}_t)\|_2\right]\right] \tag{2}$$

$$\mathcal{L}_{SSIM} = 1 - \mathop{\mathbb{E}}_{\chi}\left[[l_R(x,y)]^{\rho_R} \cdot \prod_{j=1}^{R} [c_j(x,y)]^{\varphi_j} [s_j(x,y)]^{\tau_j}\right] \tag{3}$$

where,

$$x = G_{\theta_G}(L_{\theta_L}(\chi)) \tag{4}$$

$$y = G_{\theta_G}(\boldsymbol{w}_s) \tag{5}$$

$$l_R(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \qquad \text{at scale } R \tag{6}$$

$$c_j(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \qquad \text{at scale } j \tag{7}$$

$$s_j(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \qquad \text{at scale } j \tag{8}$$

Here, $\mathcal{L}_{struct}^{source}$ is the Structural loss computed by extracting the features from layers of the generator (denoted by $S^l(\cdot)$) for the generated image and the source image and taking the difference between them. $\boldsymbol{w}_s$ is the source embedding. $\mathcal{L}_{struct}^{target}$ is the Structural loss computed by extracting the features from layers of the generator (denoted by $S^l(\cdot)$) for the generated image and the target image and taking the difference between them. $\boldsymbol{w}_t$ is the target embedding. ζ_l is a hyper-parameter which is set to 1. $\mathcal{L}_{SSIM}$ is the muli-scale strucural similarity measure. μ_x, σ_x^2, σ_{xy} are mean of x, variance of x and covariance of x and y respectively. $C_1 = (K_1 L)^2$, $C_2 = (K_2 L)^2$ and $C_3 = (K_3 L)^2$ are constants with fixed values of K_1 and K_2, and the value of L is 255. The exponents ρ_R, φ_j and τ_j are used to adjust the relative effect of each term. Hence, the overall optimization objective is given as follows:

$$\theta_M^*, \theta_D^* = \arg\min_{\theta_M}\max_{\theta_D}\left(\mathcal{L}_{style} + \mathcal{L}_{adv} + \mathcal{L}_{struct}^{source} + \mathcal{L}_{struct}^{target} + \mathcal{L}_{SSIM}\right) \tag{9}$$

where, $\mathcal{L}_{style}$ and $\mathcal{L}_{adv}$ are same as defined earlier.

G Computation Resources

We have used a machine with Intel® Xeon® Gold 6142 CPU, 376GiB RAM, and Zotac GeForce® GTX 1080 Ti 11GB Graphic Card for all of our experiments.

References

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. of ICCV*, 2015.

Utkarsh Ojha, Yijun Li, Cynthia Lu, Alexei A. Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proc. of CVPR*, 2021.

Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proc. of CVPR*, pp. 1532–1540, 2021.

Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. Few shot generative model adaption via relaxed spatial structural alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11204–11213, 2022.