Smoothness Similarity Regularization for Few-Shot GAN Adaptation
Supplementary material

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

A. Additional quantitative analysis

Fig. A shows the progression of the image quality and diversity metrics, FID and LPIPS, for different methods during few-shot adaptation. For these visualizations, we pick a pair of structurally dissimilar source-target domain pairs (*Horses→Pokemons*). The curves in the figure correspond to the results in Table 1 in the main paper.

Our observation from Fig. A is that in the challenging adaptation scenario (*Horses→Pokemons*) prior methods achieve the best performance in FID (left plot) very early during the training, and are then unable to improve the image quality at later stages. For example, the methods without any diversity preserving regularization (TGAN, FreezeD, AdAM) suffer from training instabilities, indicated by an early collapse in the FID curves. On the other hand, the FID of the models that regularize diversity degradation (CDC, RSSA) remains stable without improvements. We hypothesize that these methods can successfully adapt the colors of objects from the source domain to the style of the target domain quickly, but they are not able to learn more high-level properties like shape of objects at later stages. This is confirmed by the visual results in Fig. E, where the images generated by these methods strongly follow the structure of the source domain. In contrast, our method allows to improve FID throughout the whole training and thus achieves higher image quality (yellow curve in the left plot). Next, the diversity evaluation (right plot) demonstrates that the LPIPS of TGAN, FreezeD, and AdAM collapses to low values very quickly, indicating training instabilities. Similarly, CDC and RSSA also suffer from diversity degradation, but it is slowed down with the help of the diversity regularizations used in these methods. Finally, our method allows to maintain high diversity scores throughout the whole training process.

B. The effect of the FID evaluation protocol.

Our FID evaluation protocol differs from prior works in two ways. Firstly, as in the regime of dissimilar source-target domains the best performance can be achieved at later training epochs (e.g., see Fig. A), we extend the duration of the training procedure from 5k to 30k epochs. We additionally evaluate all methods at epochs 500 and 750 since we found this beneficial to achieve superior FID scores for

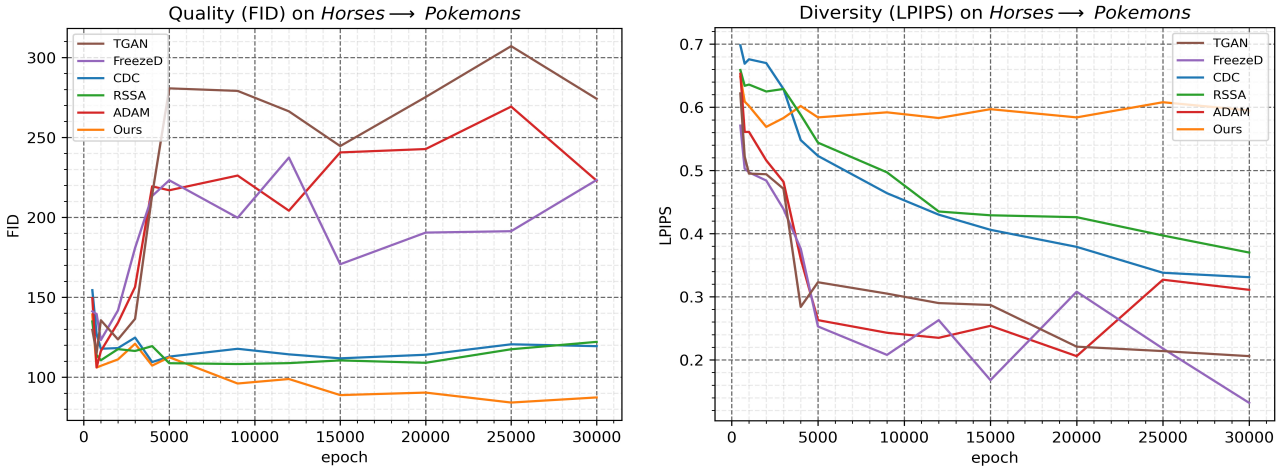

Figure A. The FID and LPIPS curves of different methods for the few-shot adaptation between dissimilar domains *Horses→Pokemons*.

Method	Face→Sketch		Face→Sunglasses	
	$FID_{val}\downarrow$	$FID_{[6]}\downarrow$	$FID_{val}\downarrow$	$FID_{[6]}\downarrow$
TGAN [7]	54.2	47.3	36.8	36.2
FreezeD [5]	48.8	40.8	32.0	31.9
CDC [6]	54.2	46.8	30.5	31.2
RSSA [8]	61.4	51.8	36.3	35.9
AdAM [10]	56.3	47.8	31.1	29.7
Ours	**45.2**	**39.9**	**27.5**	**27.0**

Table A. Effect of a different FID evaluation protocol. The differences between the protocols are described in Sec. B.

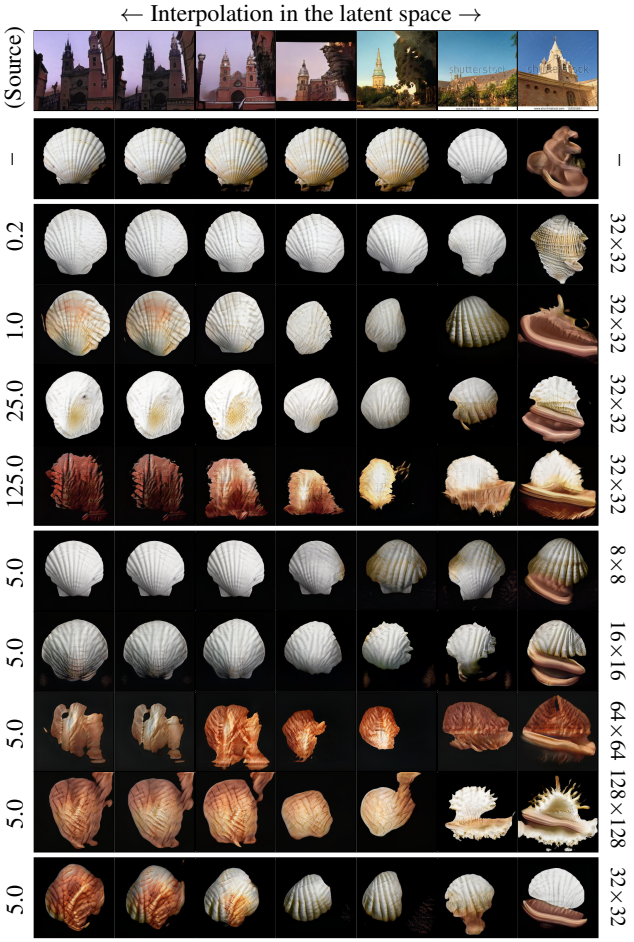

Figure B. Latent space interpolations of the source generator and the ablation models from Table B. Leftmost and rightmost columns show the used λ_{SS} and the resolution of G^l.

some close source-target domain pairs.

Secondly, we resort to evaluating FID between the sets of generated and real images of the same size, which is a standard practice in the community. In contrast, prior work [6] proposed to compute FID between 5000 generated images and the whole validation set, which often contains a significantly smaller amount of images. We note that computing FID between sets of different sizes is generally not advisable due to a mismatch in estimation of variances of the first two moments between real and generated distributions [1]. The difference in evaluation results between our (FID_{val}) and prior protocols ($FID_{[6]}$) is demonstrated in Table A, where the Sketch and Sunglasses domains have 290 and 2683 images, respectively. Due to a larger generated set, $FID_{[6]}$ tends to output consistently lower scores than our reported numbers, but it does not change the ranking of the models.

C. Additional qualitative results

We provide additional visual results with StyleGANv2 for the dissimilar source-target domains *Face→Cats* [10] and *Horses→Pokemons* in Fig. E. In both cases, our method generates diverse images that inherit the variation of the source images and flexibly combine features of different target images. In contrast, in most cases for the prior methods we observe inferior performance due to either memorization issues, training instabilities, or inability to learn the shape of objects in the target domain. We note that while the generation results of our method in the *Pokemons* domain exhibit the most realistic shapes and the largest variation in colors, it is still challenging to generate fully realistic new pokemons in the 10-shot regime. Further improvement of few-shot synthesis for such challenging datasets is an interesting direction for future work.

Fig. F shows results for the more similar domain pairs *Face→Babies* and *Churches→Haunted Houses*. We find that the results are consistent with Fig. 4: our method successfully adapts images of churches to a new style or converts adult faces into babies, performing on par with previous state-of-the-art approaches.

D. Ablation on the parameters of the smoothness similarity regularization

Our smoothness similarity regularization has two parameters: the regularization strength λ_{SS} and the resolution of features G^l. All the experiments in the main paper were conducted with $\lambda = 5.0$ and G^l at resolution (32×32). In Fig. B and Table B we provide an ablation on both these parameters. Firstly, we observe the effect of λ_{SS} (rows 3-6 in Fig. B and rows 2-5 in Table B). As seen from the ablation study, compared to the model without any regularization, our smoothness similarity regularization helps to overcome memorization and achieve diverse synthesis. The effect of $\mathcal{L}_{SS}$ is, as expected, higher when λ_{SS} is increased, which is indicated by increasing LPIPS scores. Yet, we find that setting a high λ_{SS} starts to compromise the image quality, as the loss starts to overtake the adversarial loss supervision. We found that $\lambda = 5.0$ consistently achieves a good trade-off between image quality and diversity across many source-target domains.

Furthermore, we observe the effect of using features at

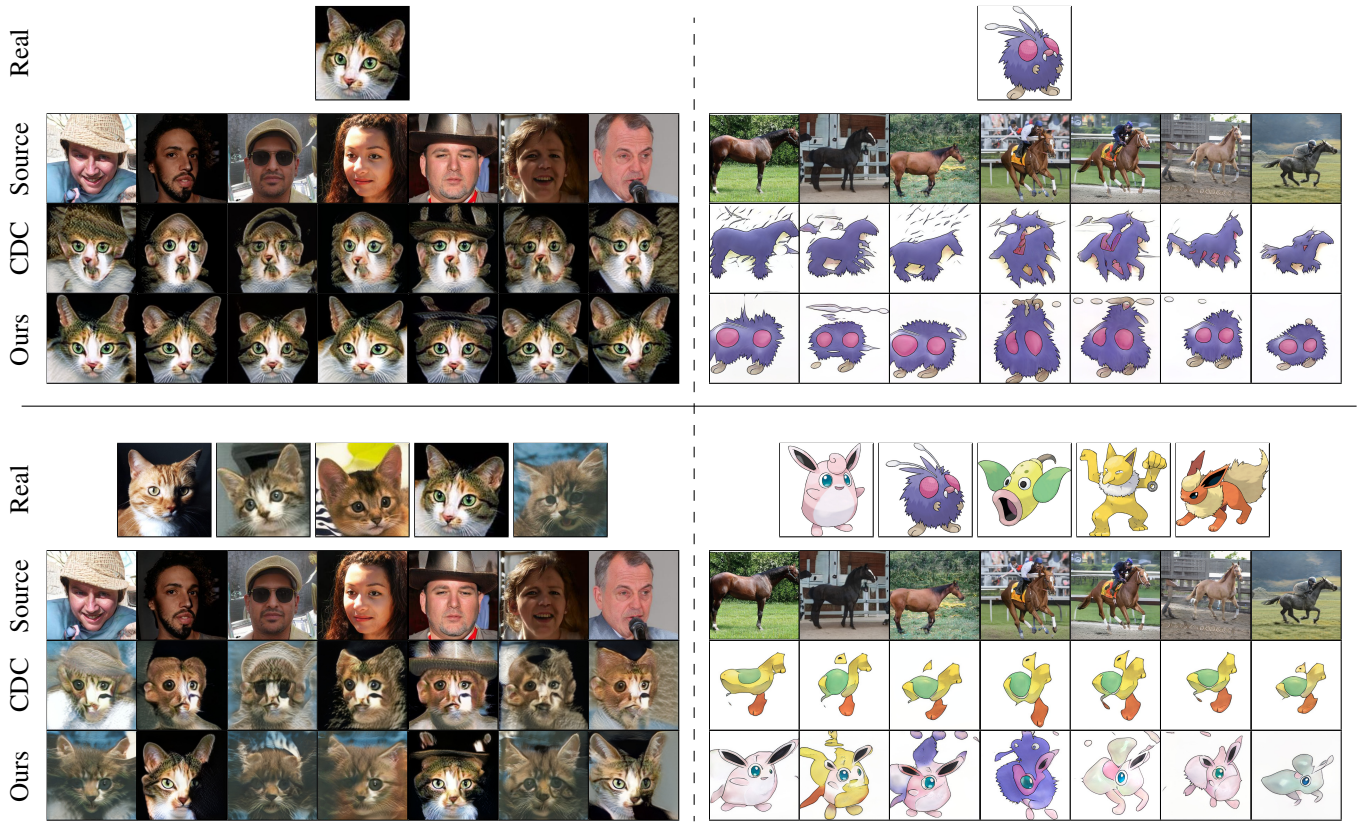

Figure C. 1-shot and 5-shot GAN adaptation results on the Cats and Pokemons datasets.

λ_{SS}	Res. of G^l	**Face→Anime** FID↓	LPIPS↑	**Church→Shells** FID↓	LPIPS↑
-	-	116.4	0.36	175.4	0.43
0.2	32×32	110.0	0.41	160.2	0.44
1.0	32×32	**96.4**	0.51	144.5	0.50
25.0	32×32	105.2	0.58	171.0	0.55
125.0	32×32	131.3	**0.64**	188.5	**0.57**
5.0	8×8	104.1	0.44	156.6	0.45
5.0	16×16	101.4	0.55	150.2	0.48
5.0	64×64	114.7	0.59	165.5	0.54
5.0	128×128	128.2	0.60	182.2	**0.57**
5.0	32×32	97.3	0.57	**140.5**	0.53

Table B. Ablation on λ_{SS} and the resolution of G^l used for the smoothness similarity regularization.

different resolutions, corresponding to different generator blocks (rows 7-10 in Fig. B and rows 6-9 in Table B). We find that using later generator blocks at higher resolution increases the impact of the regularization. However, we also observe that using a very high resolution leads to the transfer of image transitions from the source domain at more fine-grained level, which can compromise image quality, for example transferring minor details that do not look realistic in the target domain. Based on the results in Table B), we concluded that the resolution (32×32) provides a good quality-diversity trade-off as it transfers high-level, more interpretable image variations without compromising the high-level coherency of objects in the target domain.

E. Additional analysis on $\mathcal{L}_{all}$

The second component of our model is a new way to compute the D's loss. As discussed in Sec. 4.1, allowing the discriminator to compute the loss at different layers is strongly beneficial for improving the quality of synthesized images. Interestingly, even though the formulation of $\mathcal{L}_{all}$ in Eq. 3 has equal weights for all layers, the final contributions can still be different because activations $s^i \circ D^i(x)$ can have different magnitudes for different layers. In effect, this leads to an automatic discovery of the correct loss contribution of each layer depending on the source-target domains, as shown in Fig. 6.

The fact that optimal contributions of different layers in Fig. 6 are different suggests using alternative weighting schemes rather than using equal weights for all layers. For comparison, we consider two alternative strategies: assigning higher weights on earlier or later D layers. For this, instead of the uniform weights [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]/7 in Eq. 3, we use either the weighting [1.6, 1.4, 1.2, 1.0, 0.8, 0.6, 0.4]/7 or [0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6]/7,

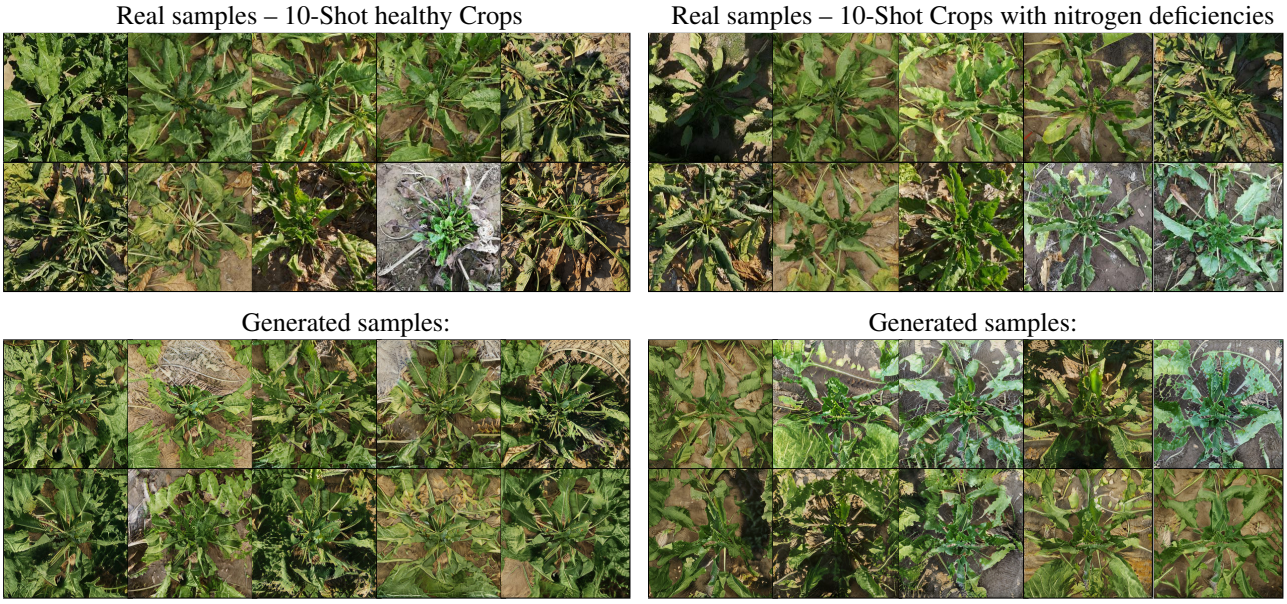

Figure D. Above: 10-shot subsets of the DND-SB dataset [9], depicting healthy sugar beat crops or crops with nitrogen nutrition deficiency. Below: few-shot adaptation results with StyleGAN pre-trained on FFHQ.

$\mathcal{L}_{all}$ weighting	**Face→Anime**		**Church→Shells**	
	FID↓	LPIPS↑	FID↓	LPIPS↑
"Earlier"	96.8	**0.59**	157.4	0.52
Uniform (ours)	97.3	0.57	140.5	**0.53**
"Later"	**93.2**	0.53	**138.4**	0.48

Table C. Effect of using different weights for different layers in L_{all}. Bold denotes the best performance.

referred to as "Earlier" or "Later" in Table C.

We note that there exists a trade-off. On one hand, while using higher weights for earlier layers is beneficial for the closer domains *Face→Anime* (improved FID and LPIPS), it also leads to degraded image quality for the more distant domains *Church→Shells* (higher FID). On the other hand, the "later" strategy universally improves the image quality, but leads to memorization of training images and thus lower LPIPS scores. For this reason, we select the uniform weighting as it is the simplest solution which already allows D to adjust the contributions of different layers, while providing a reasonable balance between image quality and diversity for diverse source-target domain pairs.

F. 1-shot and 5-shot adaptation performance

In the main paper, we mainly focus on the 10-shot target datasets. Following prior work, we extend our analysis to 5-shot and 1-shot setups. Consistent with Sec. 4, our main focus is on the challenging case of structurally dissimilar source and target domains. We thus construct 1-shot and 5-shot scenarios of the adaptation between *Face→Cats* and *Horses→Pokemons* (10-shot results for these datasets are shown in Fig. E). We compare our method to CDC [6], which is a popular baseline from the literature. Our observations from Fig. C are consistent with the main paper: while the prior method CDC cannot learn the shapes of objects in the new domain, our method achieves more realistic synthesis, successfully transferring meaningful high-level image variations even from structurally dissimilar datasets.

G. Application: detection of nutrient deficiencies of crops

We investigate the application of our model to the task of visual detection of nutrient deficiencies in crop science [9]. In agriculture, this task is important to enable timely actions to prevent major losses of crops caused by lack of nutrients, such as nitrogen. From the data collection perspective, this task refers to restricted image domains, since it typically requires manual photographing of growing crops and expert knowledge for obtaining correct annotations. Therefore, we explore whether our model can be trained on a limited set of images depicting sugar beats (see Fig. D).

For our experiments, we pick two random 10-shot subsets of the DND-SB dataset [9], consisting of images with healthy sugar beats and crops suffering from nitrogen nutrient deficiencies (see Fig. D). We use the StyleGANv2 checkpoint pre-trained on FFHQ [3]. Despite using such a dissimilar source domain, we observe that our model still achieves photorealistic synthesis of new crops, for example changing the shape or locations of leaves of the training examples.

To verify that our generated images preserve the characteristics of interest of the training images, we take a classification network, which was pre-trained to perform

Smooth. reg. w.r.t.:		ImageNet→Flowers		ImageNet→Pokemons	
Noise	Class	FID↓	LPIPS↑	FID↓	LPIPS↑
✗	✗	123.9	0.28	129.4	0.27
✓	✗	114.0	0.39	104.6	0.41
✓	✓	**106.4**	**0.55**	**89.6**	**0.56**

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

the *healthy-deficient* binary classification on images of the same resolution (256×256). We observe that generated images from the healthy subset were identified correctly in 98.9% cases, while nitrogen deficiencies were detected correctly for 95.6% of the images generated from the second subset. We consider this experiment as a promising example which suggests future utilization of our model for data augmentation in restricted image domains.

H. Additional details in the class-conditional GAN setting

For our experiments in Sec. 4.2, we pre-train the class-conditional BigGAN model [2] (without BigGAN-deep extensions) on ImageNet at the image resolution of (256×256). The model achieves FID of 9.23 on the ImageNet validation set. We then fine-tune both the pre-trained generator and discriminator on the provided few-shot dataset using our proposed loss terms as presented in Sec. 3. We use batch size of 32, decay of 0.999 for the generator's exponential moving averages, and learning rates of 2e-4 and 8e-4 for the generator and discriminator, respectively, while preserving all the other hyperparameters that were used for pre-training.

The generator of BigGAN takes two inputs, a noise vector and a class label. The input label is then projected into a continuous embedding space via a learnable linear mapping. To enable the adaptation of the generator to unconditional few-shot datasets, we do not inject class labels in our approach but directly operate with the pre-learnt continuous class embedding. At each fine-tuning epoch, we therefore sample a Gaussian vector in a joint noise-class space.

The discriminator of BigGAN takes a class label only at the final layer, where it is processed via a linear projection layer [4] and added to the output features of the last discriminator's block. In our experiments, we remove this conditioning mechanism and simply pass unmodified features after the last block to the final layer to compute the adversarial loss. This way, our whole model can be trained on the provided dataset in an unconditional fashion.

We apply our smoothness similarity regularization using the generator's features at resolution (32×32). We explored two different ways for the implementation of the regularization, considering smoothness with respect to only the noise space or the joint noise-class space. We found that using the class embeddings for the regularization is important, as achieving a high synthesis diversity without it is difficult (see Table D). We hypothesize that this happens because a large part of transferable image variations in the source domains is contained not only in the interpolations between different noise vectors, but also in the interpolations between different classes.

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations (ICLR)*, 2018.

[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] 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.

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

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

[6] 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.

[7] 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.

[8] 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.

[9] Jinhui Yi, Lukas Krusenbaum, Paula Unger, Hubert Hüging, Sabine J Seidel, Gabriel Schaaf, and Juergen Gall. Deep learning for non-invasive diagnosis of nutrient deficiencies in sugar beet using rgb images. *Sensors*, 2020.

[10] 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.

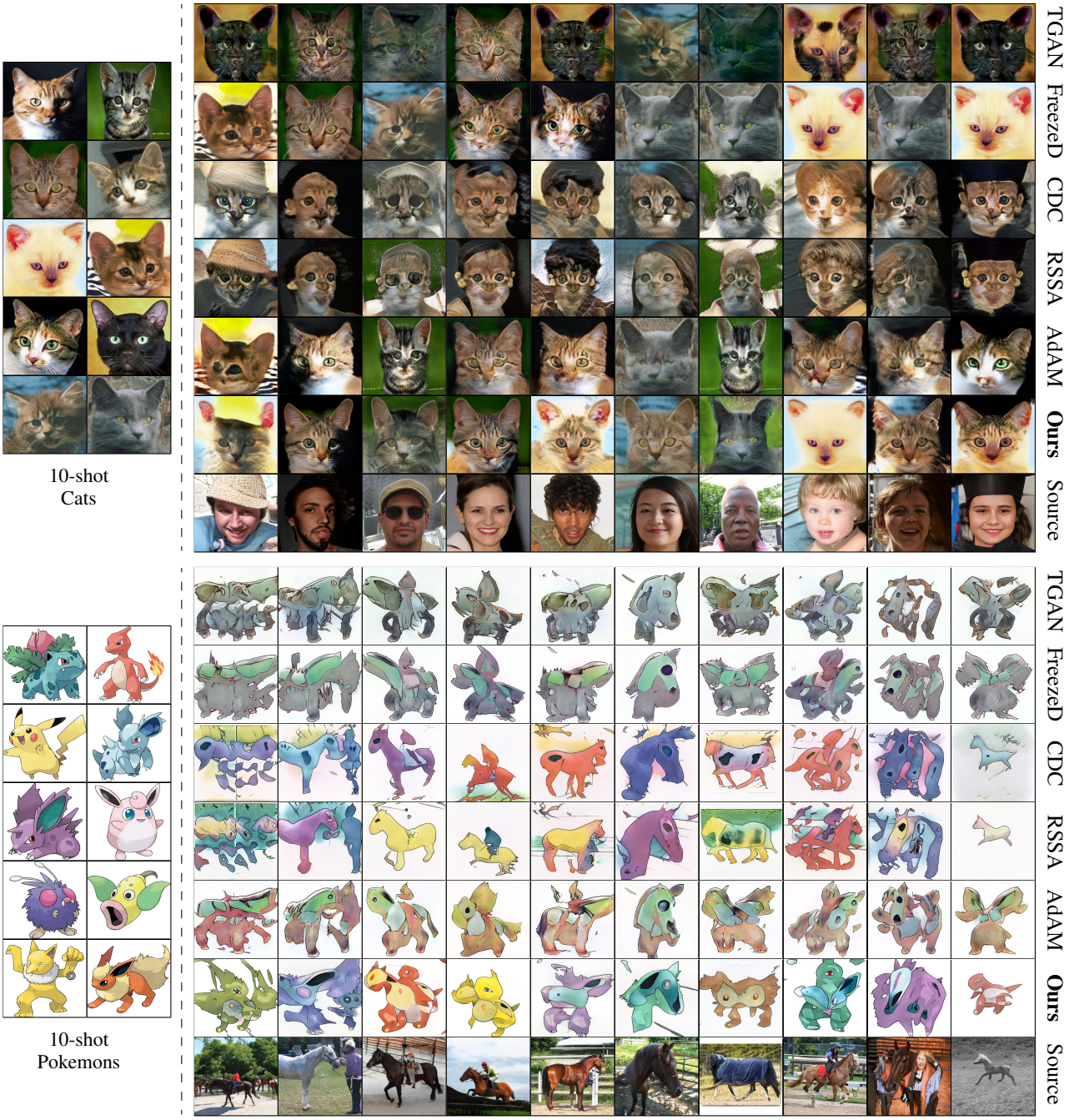

Figure E. Additional visual comparison to prior methods on *Face→Cats* and *Horses→Pokemons*, the source-target dataset pairs with a dissimilar structure (e.g., shapes of objects).

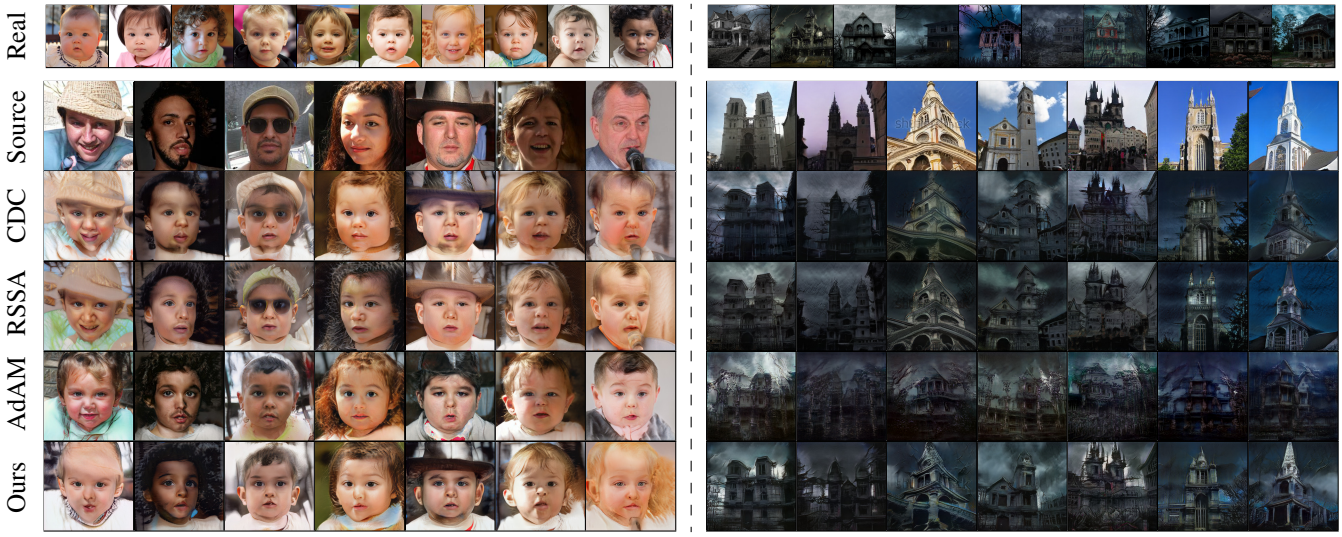

Figure F. Additional visual comparison to most recent prior methods on related domains.