(Supplementary Material)
Improving Diversity in Zero-Shot GAN Adaptation with Semantic Variations

A. Limitations

Our proposed directional moment loss $\mathcal{L}_{dm}$ is designed to align the image directions with the text directions. However, it could be very hard or even infeasible to achieve complete alignment between two modalities in some scenarios with large domain gaps, e.g., “Dog-to-Joker” that share few common visual concepts and require significant content changes. Meanwhile, the adaptation process requires the expert intervention to determine the language description of the source domain as well as appropriate training iterations for the desired result, due to the lack of quality measures. In addition, since our directional moment loss $\mathcal{L}_{dm}$ solely depends on the guidance provided by CLIP [12] text encoder, there might exist risks of inheriting underlying biases which can cause fairness and privacy issues. Capturing and alleviating the bias for GAN adaptation is an important research topic that is beyond the scope of this work.

B. Quantitative evaluation.

For quantitative quality evaluation, we measure Fréchet Inception Distance (FID) [3] in the “dog-to-cat” scenario between the generated samples and the AFHQ-Cat [1] training split. With 5 independent runs using different random seeds, we generated the same number of images with the real dataset, and then computed the mean and standard deviation of FID. As shown in Table 1, we achieved better FID in zero-shot setting with the help of learned semantic variations as well as source knowledge for diverse characteristics. Moreover, we accomplished the state-of-the-art in the 10-shot setting, demonstrating scalability and superiority. To further evaluate quality and diversity, we respectively plot the precision and recall [14, 15, 9] for each truncation rate in Figure 1. Our framework reports comparable precision with the baseline while showing improvements in recall rate, which is closely related to sample diversity.

Table 1. FIDs [3] under the “Dog-to-Cat” scenario on AFHQ [1].

Methods	FID (↓)
StyleGAN-NADA [2]	$70.225_{\pm 0.242}$
Ours	$61.347_{\pm 0.317}$
Ojha et al. (10-shot) [11]	$57.196_{\pm 0.249}$
StyleGAN-NADA (10-shot) [2]	$60.500_{\pm 0.191}$
Ours (10-shot)	$\mathbf{45.084}_{\pm 0.108}$

Table 2. FIDs [3] under the “Dog-to-Cat” scenario on AFHQ [1].

FID (↓)					
λ_{EWC}		λ_{rel}		K (number of z^i)	
10^5	94.317	1	82.900	2	81.840
10^6	77.898	10	76.590	4	71.793
10^7 (Ours)	**61.347**	100 (Ours)	**61.347**	6 (Ours)	**61.347**
10^8	81.758	1000	81.758	8	66.698

C. Hyperparameter ablations.

The quantitative analyses on hyperparameters are provided in Table 2. Note that λ_{EWC} affects the level of source characteristic preservation, while λ_{rel} is related to the degree of maintained semantic relationship between images before and after adaptation. We meticulously set these weighting factors to guarantee that all losses ($\mathcal{L}_{dm}$, $\mathcal{L}_{EWC}$, $\mathcal{L}_{rel}$) are balanced in terms of their magnitude. In addition, enlarging the number of semantic variations K showed increased fidelity with the excavated semantics of the target domain. However, setting a large K makes the semantic variation learning more challenging, resulting in a slightly quality decrease. Meanwhile, the perturbation strength ϵ is an important parameter for optimizing semantic variation. A large ϵ makes $\mathcal{L}_{cons}$ optimization difficult, while a small ϵ causes fast convergence and minimal semantic differences after perturbation. To discover meaningful semantic variations, we empirically set the value ϵ to $||E_T(t_{trg})||_2$, which has been shown to enable sufficient convergence within 2,000 iterations.

D. Few-shot GAN adaptation.

To further verify the effectiveness, we extended our framework to few-shot GAN adaptation, *i.e.*, 10-shot, 5-shot and 1-shot settings. For each target image, we extract feature from the image encoder and conduct semantic variation learning process. As reported in Table 1, our framework achieved FID score of 45.084 in 10-shot setting, considerably surpassing both StyleGAN-NADA [2] and the

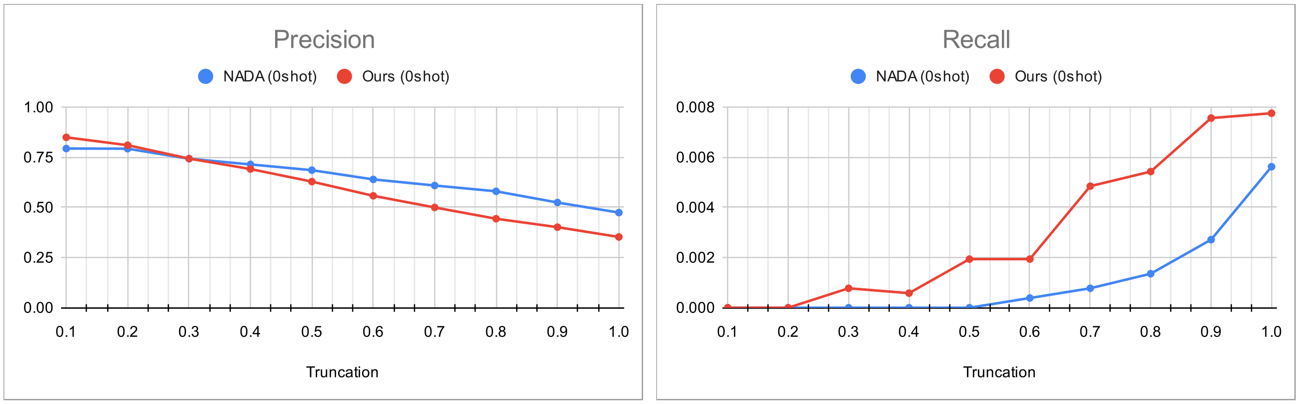


Figure 1. Precision and Recall [9] with varying truncation rates in the “Dog-to-Cat” scenario on AFHQ [1].

Ojha et al. [11]. It is also notable that we scored 61.347 FID in zero-shot setting comparable to few-shot methods.

Moreover, in 5-shot and 1-shot settings, we present qualitative results respectively in Figure 2 and 3. Remarkably, our method shows more visually favorable results with diverse facial characteristics, e.g., emotional expressions, compared to previous methods [11, 2] as shown in Figure 2. Also compared to MTG [18] and DynaGAN [8] in 1-shot setting (Figure 3), it is notable that our method has strength in both utilizing diverse semantic information of source contents and maintaining original identity, while still being able to capture the style of target samples (*e.g.*, face of the joker, hairstyle of the doctor brown).

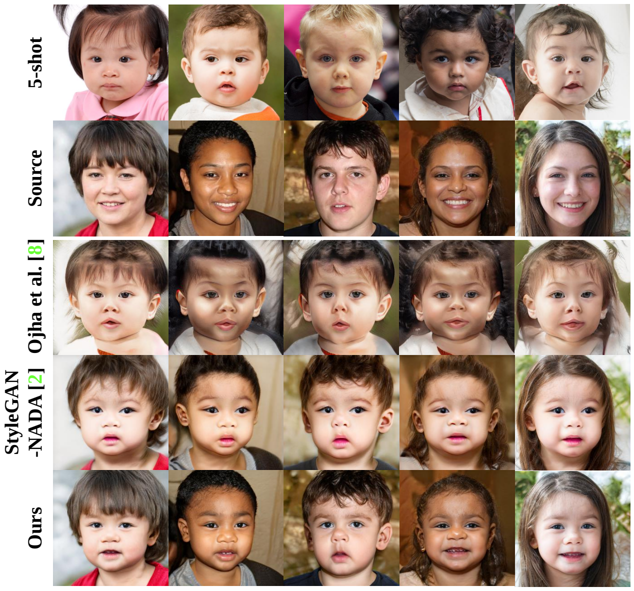


Figure 2. Results on 5-shot GAN adaptation scenario with the StyleGAN2 trained on FFHQ.

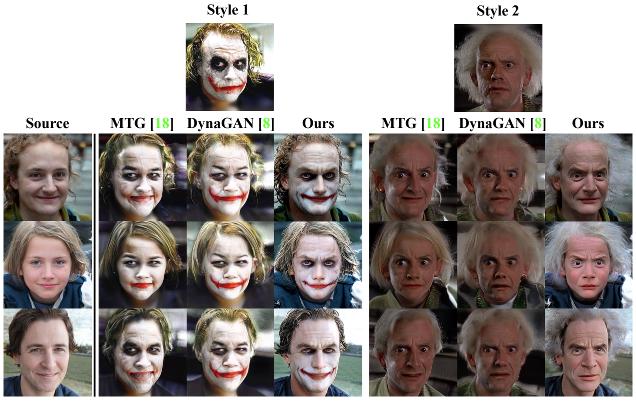


Figure 3. Results on 1-shot GAN adaptation scenarios with the StyleGAN2 trained on FFHQ.

E. Additional Qualitative Results

In this section, we provide further qualitative results in various GAN adaptation scenarios which are not included in the main manuscript in Figures 4 through 9. Most of the hyperparameters are kept unchanged except for λ_{ewc}, which is adjusted from 10^7 to 10^6 for artistic texture manipulation, e.g., “Photo-to-Sketch” and “Photo-to-Caricature”. We can observe that structural layouts are reflected in the results of StyleGAN-NADA and ours. Noticeably, the proposed framework synthesizes buildings with diverse characteristics with the help of $\mathcal{L}_{dm}$ which guides the model with explored semantic variations. Also, the contexts are well preserved by inheriting useful knowledge from the source generator via $\mathcal{L}_{EWC}$ and $\mathcal{L}_{rel}$.

F. Questionnaire for User Study

In Fig. 10, we present the screenshots of questionnaires for the user study. The questionnaire is composed of 25 questions for quality evaluation and 6 questions for diversity assessment. Each question contains the generated samples from Ojha et al [11], StyleGAN-NADA [2], and ours in the

"Cat-to-Dog" scenario from the same latent codes. To evaluate the quality, we requested users to select the best result that looks most like a "Dog" while preserving the content information of the source image (Fig. 10 (a)). For diversity assessment, we provided the participants with a set of 4 images from each method and asked to pick the one showing the most diverse characteristics of dogs (Fig. 10 (b)). Note that we manually shuffled the order of methods for the reliability of the survey.

G. License

In Table 1, we specify the source and licenses of the models and datasets used in our work. Note that the FFHQ [5] dataset consists of facial images collected from Flickr, which are under permissive licenses for non-commercial purposes.

Table 3. Sources and licenses of the utilized models and datasets

Models	License
StyleGAN2 [6]	Nvidia Source Code License
scikit-learn-extra [16]	BSD 3-Clause
CLIP [12]	MIT License
StyleGAN2-pytorch [7]	MIT License
StyleGAN-ADA [4]	Nvidia Source Code License
psp [13]	MIT License
Ojha *et al*. [11]	Adobe Research License
StyleGAN-NADA [2]	MIT License
Datasets	
FFHQ [5]	CC BY-NC-SA 4.0
AFHQ [1]	CC BY NC 4.0
LSUN [17]	No License
CelebA [10]	CC BY-NC-SA 4.0

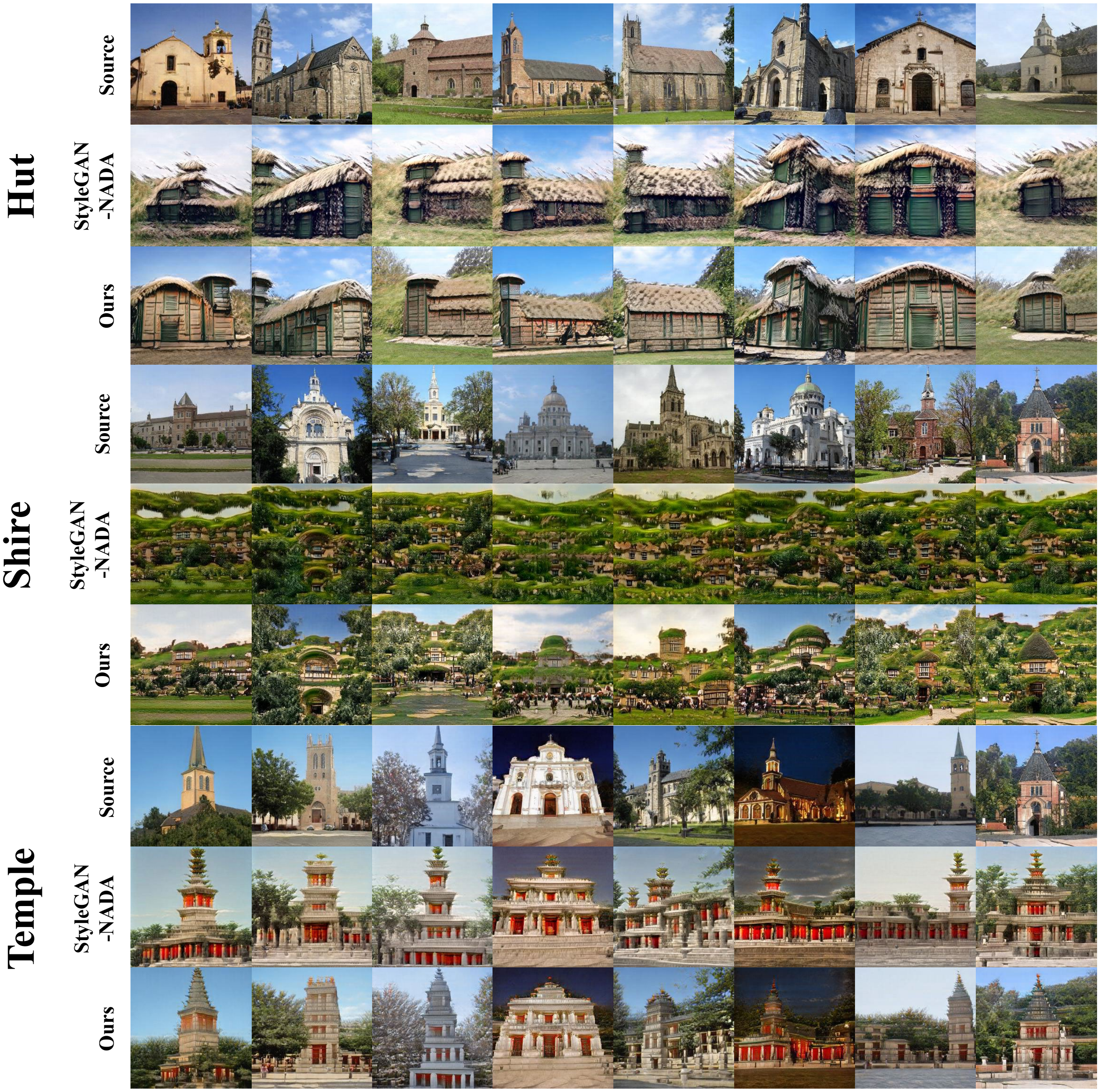


Figure 4. Qualitative results in the adaptation scenarios of from the source generator trained on LSUN-Church dataset [17] to different buildings, i.e, "Hut" and "Temple", and to fictitious village of "Shire". The results of StyleGAN-NADA [2] are sharing a specific design or global characteristic of the target domain, e.g., green doors and walls of huts. On the other hand, the proposed framework synthesizes buildings with more diverse textural details. Moreover, various contexts of the source domain are well inherited and fully utilized to generate satisfactory results.

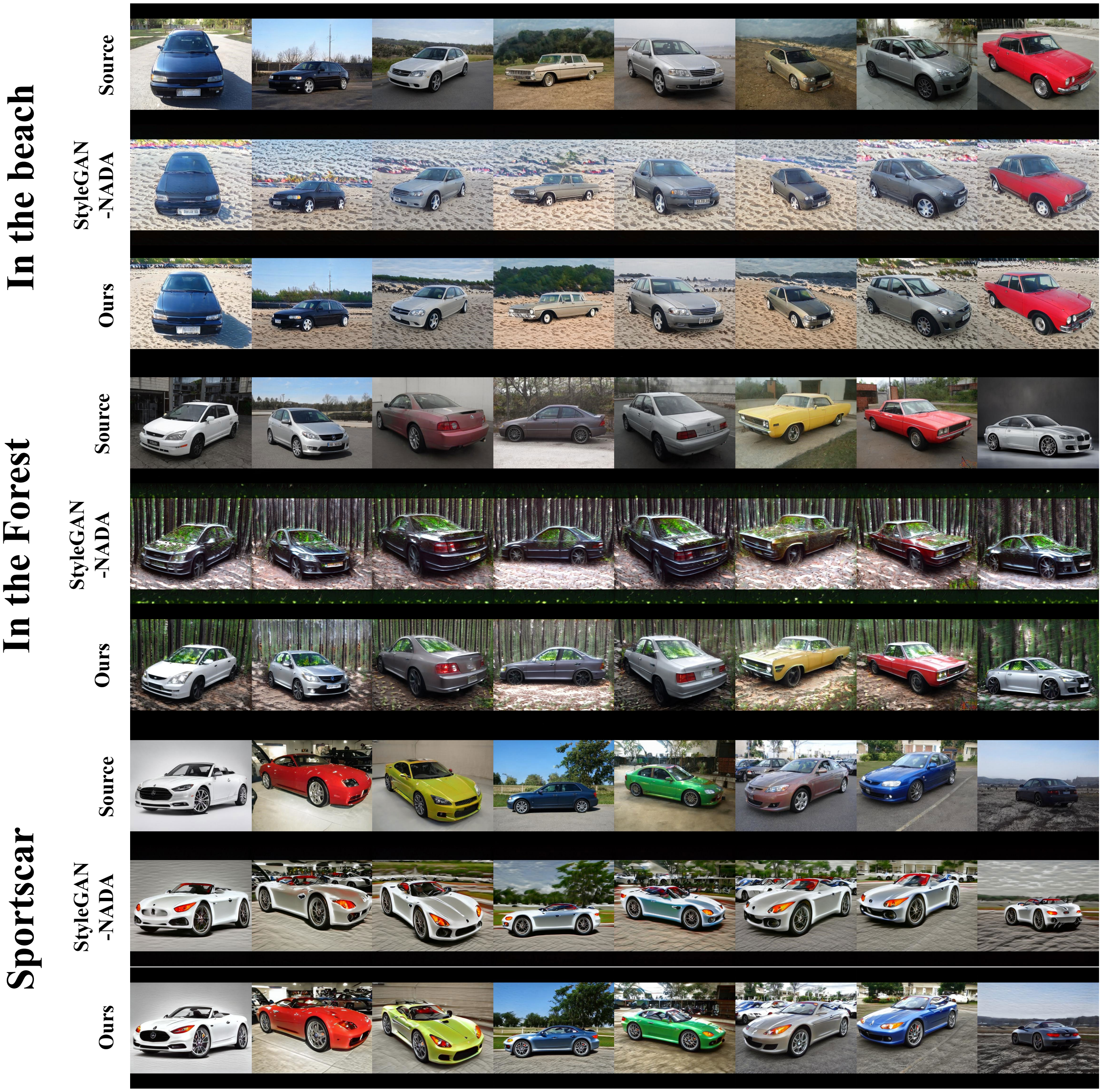


Figure 5. Qualitative results in the adaptation scenarios of from the source generator trained on LSUN-Car dataset [17] to different backgrounds, i.e, “Car in the beach” and “Car in the forest”, and appearance translation to “Sportscar”. StyleGAN-NADA [2] samples fails in succeeding context of the source domain, filling the overall image even in the sky region. Also in “Car-to-Car in the forest” and “Car-to-Sportscar”, the underlying bias of the target text prompt are reflected to StyleGAN-NADA, e.g., dark car in the forest and same the sportscar design. In contrary, our framework generates results with more diverse designs and characteristics.

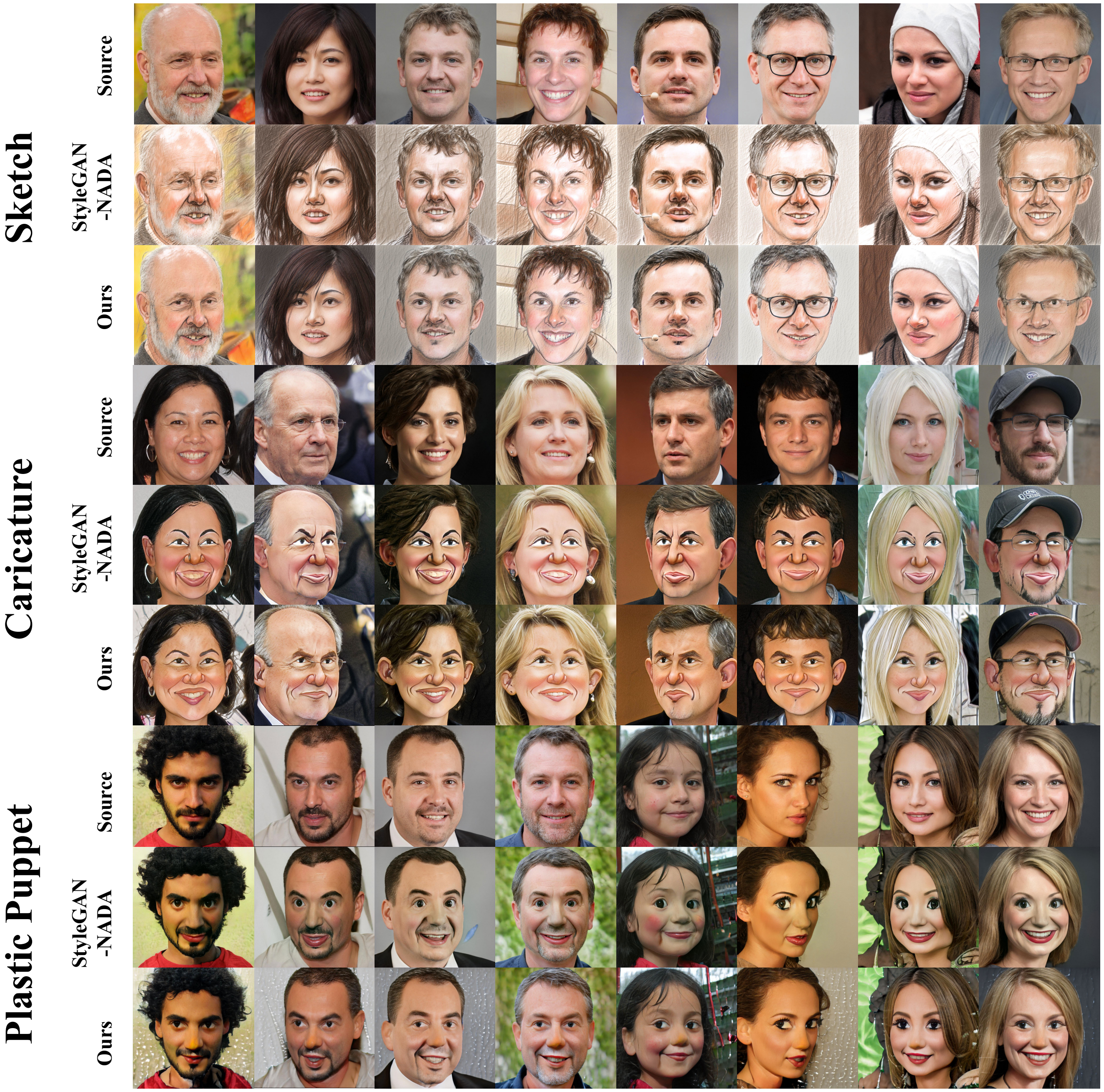


Figure 6. Generated samples from StyleGAN2 [6] trained with FFHQ [5] in various adaptation scenarios.

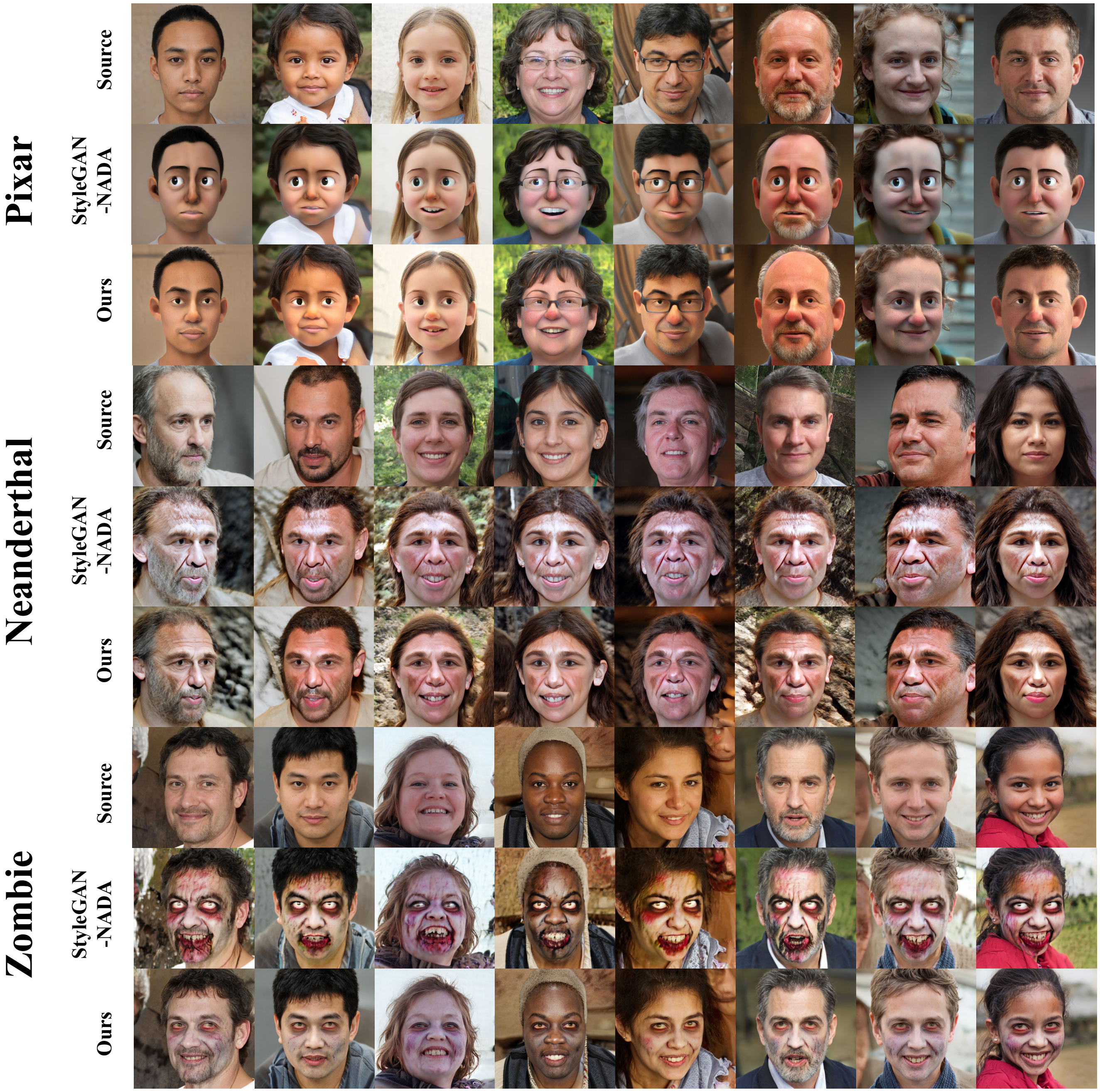


Figure 7. Generated samples from StyleGAN2 [6] trained with FFHQ [5] in various adaptation scenarios.

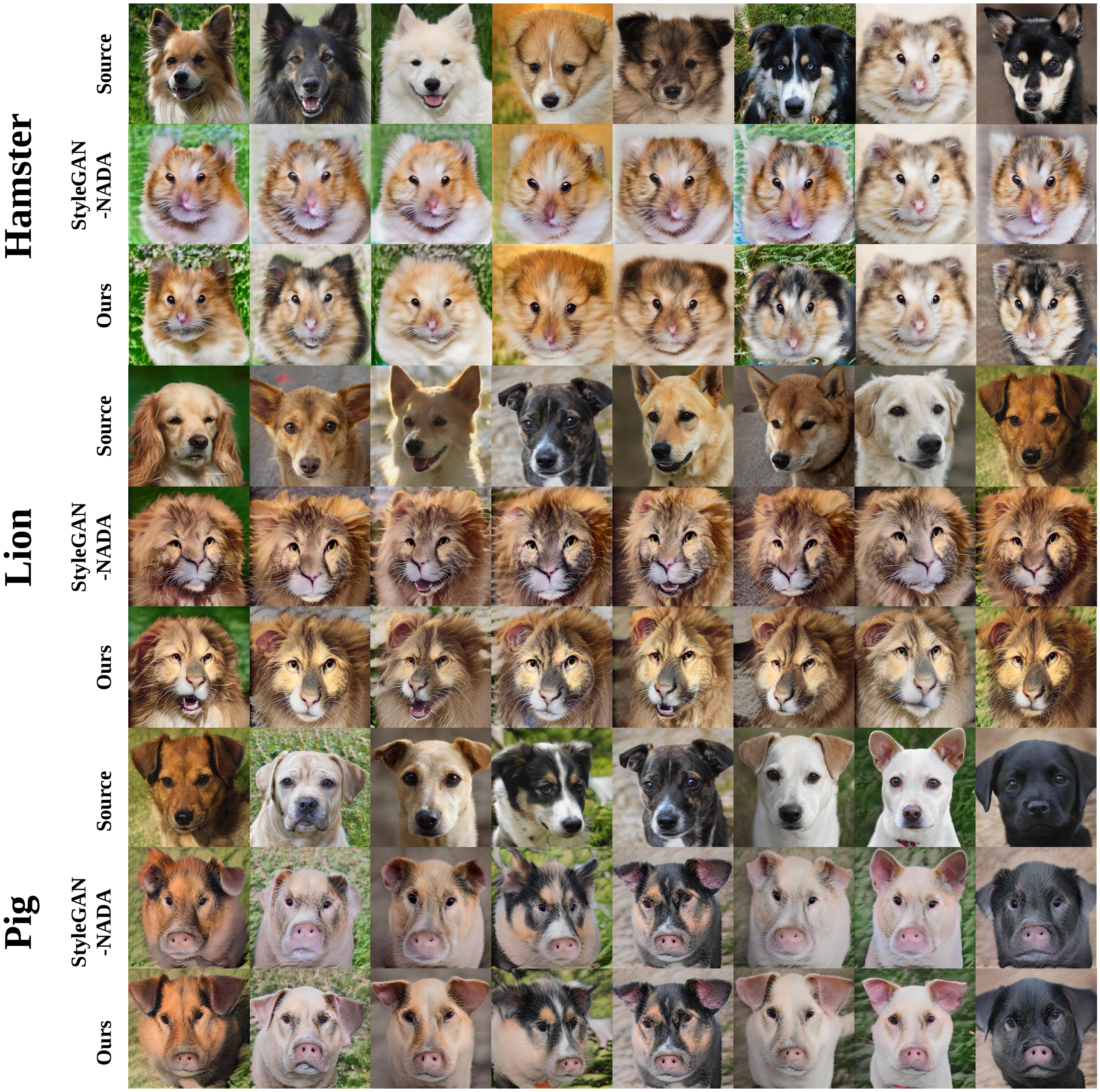


Figure 8. Generated samples from StyleGAN2 [6] trained with AFHQ-Dog [1] adapted to various target domains.

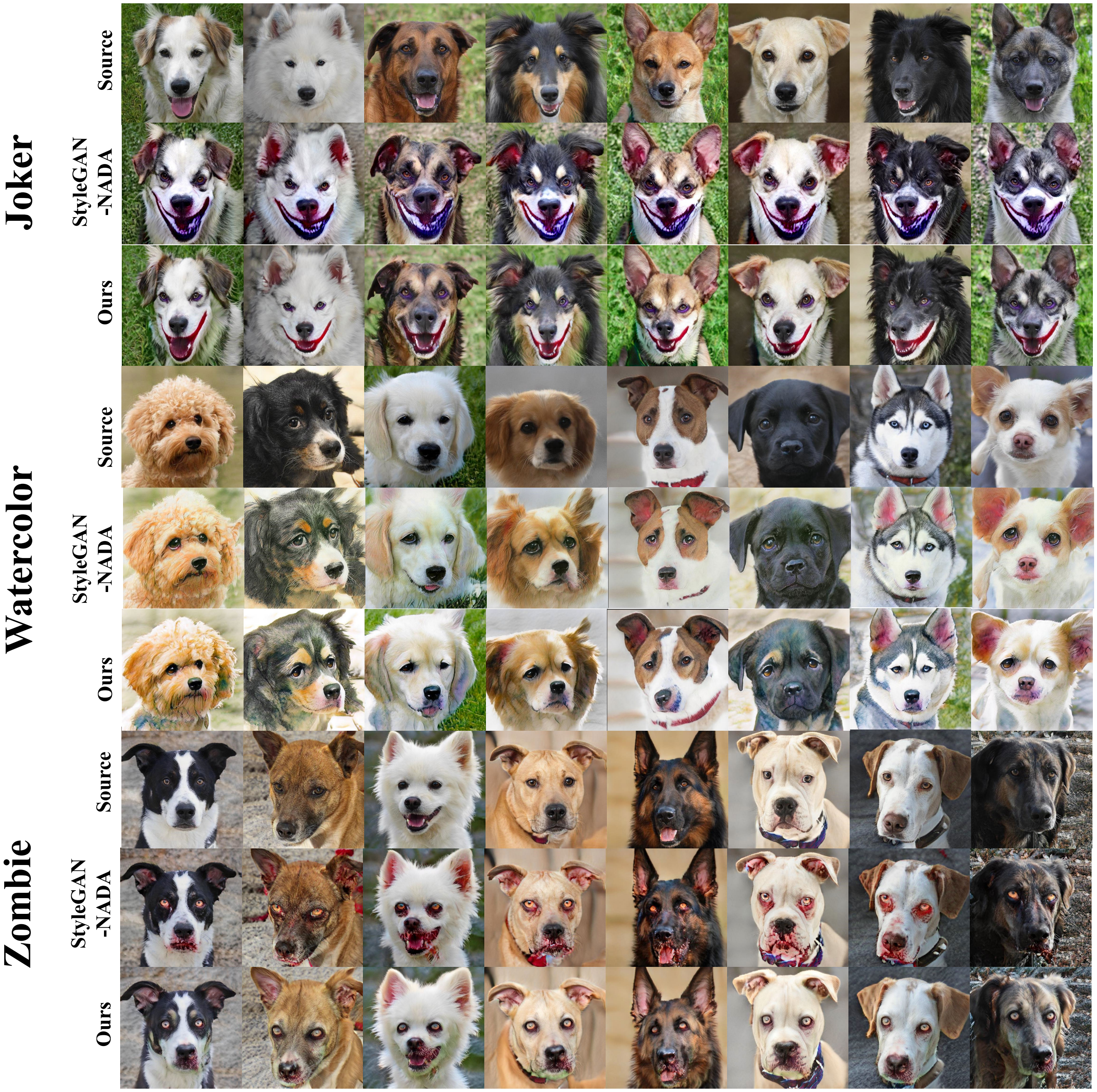
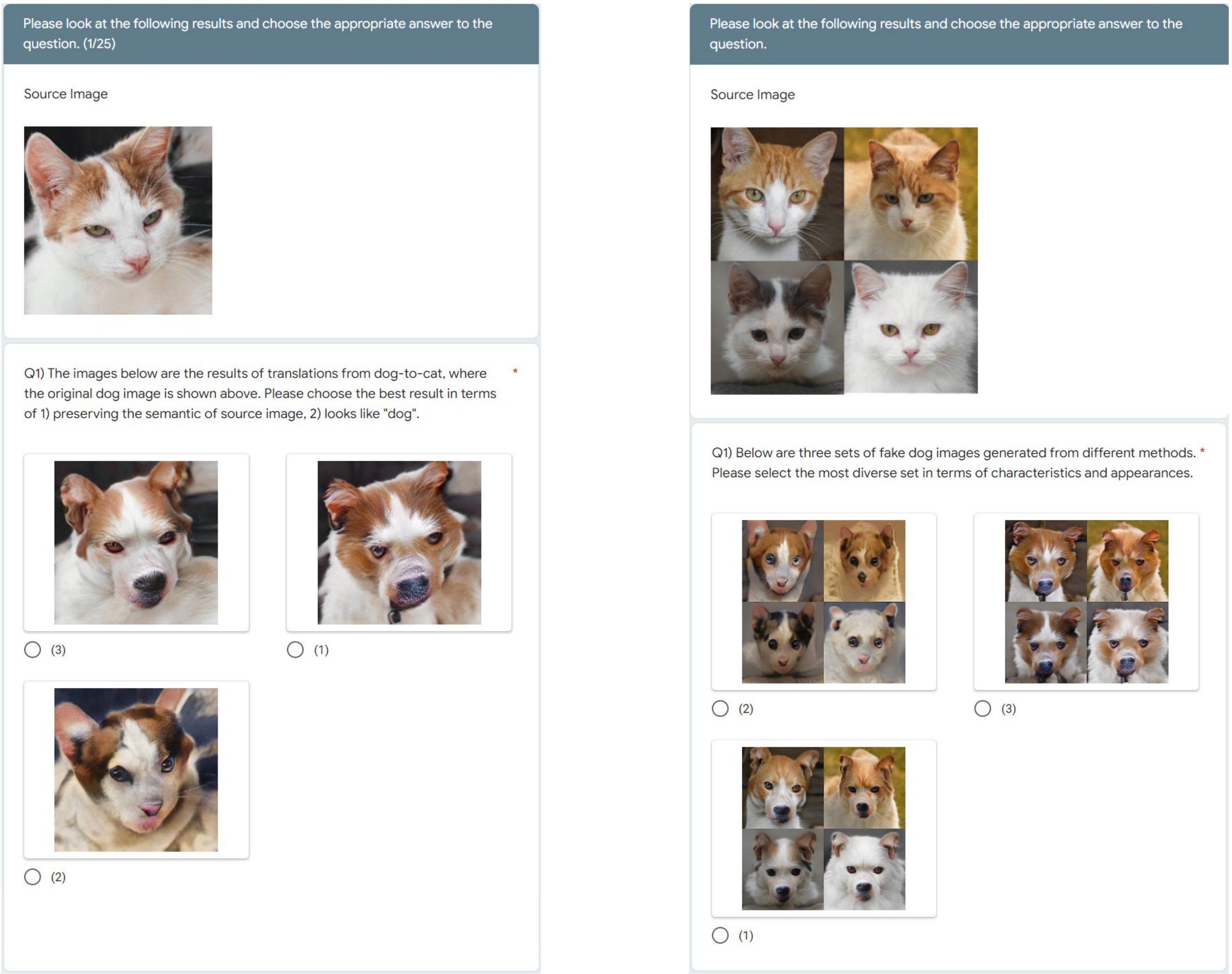


Figure 9. Generated samples from StyleGAN2 [6] trained with AFHQ-Dog [1] adapted to various target domains.



Please look at the following results and choose the appropriate answer to the question. (1/25)

Source Image

Q1) The images below are the results of translations from dog-to-cat, where the original dog image is shown above. Please choose the best result in terms of 1) preserving the semantic of source image, 2) looks like "dog". *

○ (3)

○ (1)

○ (2)

(a) Quality Assessment

Please look at the following results and choose the appropriate answer to the question.

Source Image

Q1) Below are three sets of fake dog images generated from different methods. * Please select the most diverse set in terms of characteristics and appearances.

○ (2)

○ (3)

○ (1)

(b) Diversity Assessment

Figure 10. Screenshots of user study on (a) quality assessment and (b) diversity assessment.

References

[1] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 1, 2, 3, 8, 9

[2] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 1, 2, 3, 4, 5

[3] 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. *Advances in neural information processing systems*, 30, 2017. 1

[4] 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

[5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3, 6, 7

[6] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3, 6, 7, 8, 9

[7] Seonghyeon Kim. Stylegan2-pytorch. *https://github.com/rosinality/stylegan2-pytorch*, 2020. 3

[8] Seongtae Kim, Kyoungkook Kang, Geonung Kim, Seung-Hwan Baek, and Sunghyun Cho. Dynagan: Dynamic few-shot adaptation of gans to multiple domains. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 2

[9] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[10] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 3

[11] 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

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3

[13] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021. 3

[14] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in n eural information processing systems*, 31, 2018. 1

[15] Loic Simon, Ryan Webster, and Julien Rabin. Revisiting precision recall definition for generative modeling. In *International Conference on Machine Learning*, pages 5799–5808. PMLR, 2019. 1

[16] Gaël Varoquaux, Lars Buitinck, Gilles Louppe, Olivier Grisel, Fabian Pedregosa, and Andreas Mueller. Scikit-learn: Machine learning without learning the machinery. *GetMobile: Mobile Computing and Communications*, 19(1):29–33, 2015. 3

[17] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 3, 4, 5

[18] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. *arXiv preprint arXiv:2110.08398*, 2021. 2