Supplemental Material: Fix the Noise: Disentangling Source Feature for Controllable Domain Translation

Dongyeun Lee[1,2] Jae Young Lee[1] Doyeon Kim[1] Jaehyun Choi[1]
Jaejun Yoo[3] Junmo Kim[1]
[1]KAIST [2]Klleon AI Research [3]UNIST

A. Evaluation protocol

Comparison with unconditional GANs based method. We adopt Fréchet Inception Distance (FID) [3] and Kernel Inception Distance (KID) [1] to measure the generation quality and diversity of generated images. FID and KID are computed between 50K generated images and the entire training samples. We use a modified version of Perceptual Smoothness (PS) [12] to measure the smoothness of interpolation between different domain features. Instead of the style code which is used in the original paper [12], we use $\mathbf{z} \in \mathcal{Z}$ for target interpolation latent. Note that this modification is due to the architectural difference. Same as FID and KID, 50K samples are used to compute PS. In order to ensure that the implementation difference does not affect performance, we compare all methods above the official Pytorch [13] implementation of StyleGAN2-ADA[1] [6].

Comparison on domain translation method. We use FID and KID to evaluate generated images. 20K images are randomly sampled from the source domain. For our approach, we project source domain images to $\mathbf{z} \in \mathcal{Z}$ and provide them to the target model. For the other domain translation methods, source domain images and corresponding randomly sampled style latent codes are used to generate images. Note that 20K generated images and the entire target domain images are used for evaluation.

B. Additional results

Evaluation on anchor point n_{anch}. We evaluate the proposed method using different anchor points in FFHQ → Metfaces setting. We train our model from scratch 10 times and report Perceptual Smoothness (PS) [12], FID [3], and KID [1] in Table 1. The anchor point n_{anch} is randomly sampled from the Gaussian distribution for each experiment.

Noise interplation. We provide additional noise interpolation results of the proposed method on FFHQ → MetFaces (Figure 2), FFHQ → AAHQ (Figure 3) and LSUN Church → WikiArt Cityscape (Figure 4).

Setting	FFHQ → MetFaces		
α	PS	FID	KID $_{(\times 10^3)}$
1	0.884 ± 0.04	38.69 ± 3.23	14.36 ± 1.93
0		20.08 ± 0.30	3.54 ± 0.37

Table 1. Experiment on different anchor point n_{anch}. We report the mean and standard deviation of metrics over 10 runs.

Comparison with unconditional GANs based method. An additional qualitative comparison of controlling preserved source features is shown in Figure 5, 6, and 7. Freeze G [11] that requires new training for each source degree shows an inconsistent transition of the preserved source features. Layer-swap [14] and UI2I StyleGAN2 [10] that convert weights of a source model also show the inconsistent transition. Specifically, unnatural color transitions from the source domain are observed in Figure 5. Additionally, several artifacts and changes in the human identity are observed in Figure 6. We believe that this phenomenon occurs due to the long training time of the target model (*e.g.* FFHQ → AAHQ are trained for 12000K images). The long training time causes more changes in the target model weights, and this may disturb the combined models to generate realistic images. For example, the identity changes seen in the result of layer swap (Figure 6) seem to be caused by a large change in the mapping function that transforms $\mathbf{z} \in \mathcal{Z}$ to $\mathbf{w} \in \mathcal{W}$. The color transition problems and inconsistent transition are less observable in LSUN Church → WikiArt Cityscape, due to the artistic target dataset and the spatial difference between the source and target domain, respectively. Nevertheless, these methods require models for each degree of preserved source features, while the proposed method can control in a single model.

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

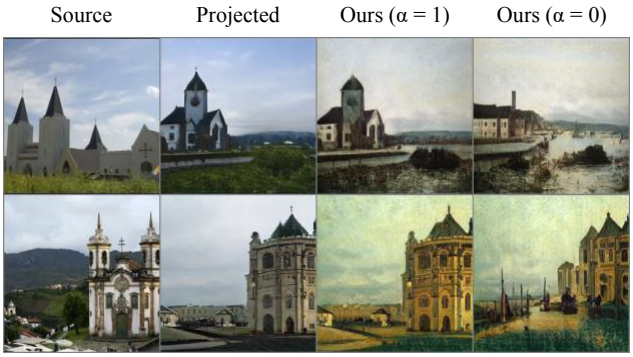

Figure 1. Domain translation results for Church → Cityscape. Incorrectly projected images cause our method to generate uncorrelated target images with source.

Comparison on domain translation method. An additional qualitative comparison of controlling preserved source features are shown in Figure 11, 12. The latent inversion method is only used for our approach. Modified version of the inversion method in StyleGAN2 [8] is used. We embed real images into the $\mathcal{Z}$ space of the source model with truncation psi of 0.7 following StyleAlign [17]. In the comparison, we use the exact same latent code obtained by the inversion method for the target model. However, please note that our method can also be multimodal like MUNIT [4] and StarGAN-v2 [2] by combining early latent code from the projected latent code with the late latent code from the others [17].

Latent inversion failure cases. Projecting images into the $\mathcal{Z}$ space of the StyleGAN often fails to accurately reconstruct original images when the dataset becomes larger and more diverse. The inversion and translated results on Church → Cityscape are shown in Figure 13. The result shows a strong correlation between projected and translated images. However, the incorrectly acquired latent codes lead to uncorrelated target domain images. It would be interesting to integrate our method well with the inversion method for other spaces (*e.g.* $\mathcal{Z}+$, $\mathcal{W}$, and $\mathcal{W}+$), or to improve the performance of the inversion method for $\mathcal{Z}$ space.

C. Latent modulation

Recently, several works [5, 15, 16] observe that StyleGAN can effectively adjust semantic attributes of images by modulating latent codes in interpretable directions. Additionally, StyleSpace [16] revealed that the $\mathcal{S}$ space is the most disentangled among the three latent spaces $\mathcal{Z}$, $\mathcal{W}$, and $\mathcal{S}$ of StyleGAN [8, 9], and it is possible to change various semantic attributes of generated images just by adjusting a value of the single dimension of $\mathcal{S}$. Based on this observation, we examine latent modulation effects on our proposed method. The latent modulation effects on different interpolation weights are shown in Figure 8, 9, and 10. The latent modulation effects of the source model are highly aligned in anchored subspace ($\alpha = 1$). As α decreases, some latent modulation effects remain, while the rest gradually weakens or disappears. This phenomenon may occur as the preserved source features gradually vanish.

D. Comparison with SmoothingLatentSpace

Our approach allows smooth interpolation between the source and target features in the transfer-learned model. We additionally compare our approach with SmoothingLatentSpace [12] which tries to smooth the interpolation between the source and target domain. For SmoothingLatentSpace, We interpolate latent codes from source images $\mathbf{s}_s$ and randomly sampled noise $\mathbf{s}_{rand}$, $\alpha \cdot \mathbf{s}_s + (1 - \alpha) \cdot \mathbf{s}_{rand}$, and generate target images with content from source images and interpolated latent codes. Figure 14 and 15 show interpolation results between the source and target features. The results show that SmoothingLatentSpace frequently generates severe artifacts during the interpolation between latent codes from source images and randomly sampled noise. In addition, compared to our method, SmoothingLatentSpace generates less smooth interpolation results.

E. Limitations

Despite our method achieved compelling results, it is not without limitations. Although our method is easily applicable to StyleGAN 1 [8] and 2 [9], it is hard to directly incorporate our method into architectures that do not contain the noise input such as StyleGAN3 [7] which removed the noise input to achieve equivariances. Second, inversion methods for $\mathcal{Z}$ space cannot accurately reconstruct finer details of real images, which interferes with the consistency between the source and target images in domain translation. For example, the results in Figure 6 in the main paper show slight changes in the face identity due to inaccurately obtained latent codes. Additionally, this phenomenon is exacerbated when the dataset becomes larger and more diverse. As shown in Figure 1, the latent inversion method causes significant changes in overall reconstructed images in LSUN church, which leads our method to generate target images uncorrelated with source images. In the future, it might be interesting to design inversion methods that overcome the above issues.

F. Broader impact

Translating one image to other domains has received tremendous attention from the community and has been used in a variety of applications. In addition to generating various images from one image (multimodal), it is also

very important to determine how much of the source features are preserved. For example, users may obtain results in which the desired degree of characteristics is preserved in the applications. As such, we see great potential for our technology to be utilized in various applications.

However, since our method is based on data-driven generative modeling, it faces various ethical issues arising from bias in the training data. For example, a target model fine-tuned from a source model pre-trained on FFHQ tends to generate more light-skinned images than dark-skinned ones. In addition, the phenomenon of changing dark-skinned images to light-colored skin was also observed. At a time when data-driven modeling is getting a lot of attention, the community needs a lot of effort and discussion about data bias.

References

[1] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. 1

[2] 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. 2

[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] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 2

[5] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Proc. NeurIPS*, 2020. 2

[6] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020. 1

[7] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021. 2

[8] 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. 2

[9] 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. 2

[10] Sam Kwong, Jialu Huang, and Jing Liao. Unsupervised image-to-image translation via pre-trained stylegan2 network. *IEEE Transactions on Multimedia*, 2021. 1, 7, 8, 9

[11] Bryan Lee. Freeze g. http://github.com/bryandlee/FreezeG, 2020. Accessed Jan. 2022. 1, 7, 8, 9

[12] Yahui Liu, Enver Sangineto, Yajing Chen, Linchao Bao, Haoxian Zhang, Nicu Sebe, Bruno Lepri, Wei Wang, and Marco De Nadai. Smoothing the disentangled latent style space for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10785–10794, June 2021. 1, 2, 16, 17

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[14] Justin NM Pinkney and Doron Adler. Resolution dependent gan interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334*, 2020. 1, 7, 8, 9

[15] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1532–1540, 2021. 2

[16] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 2

[17] Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. Stylealign: Analysis and applications of aligned styleGAN models. In *International Conference on Learning Representations*, 2022. 2

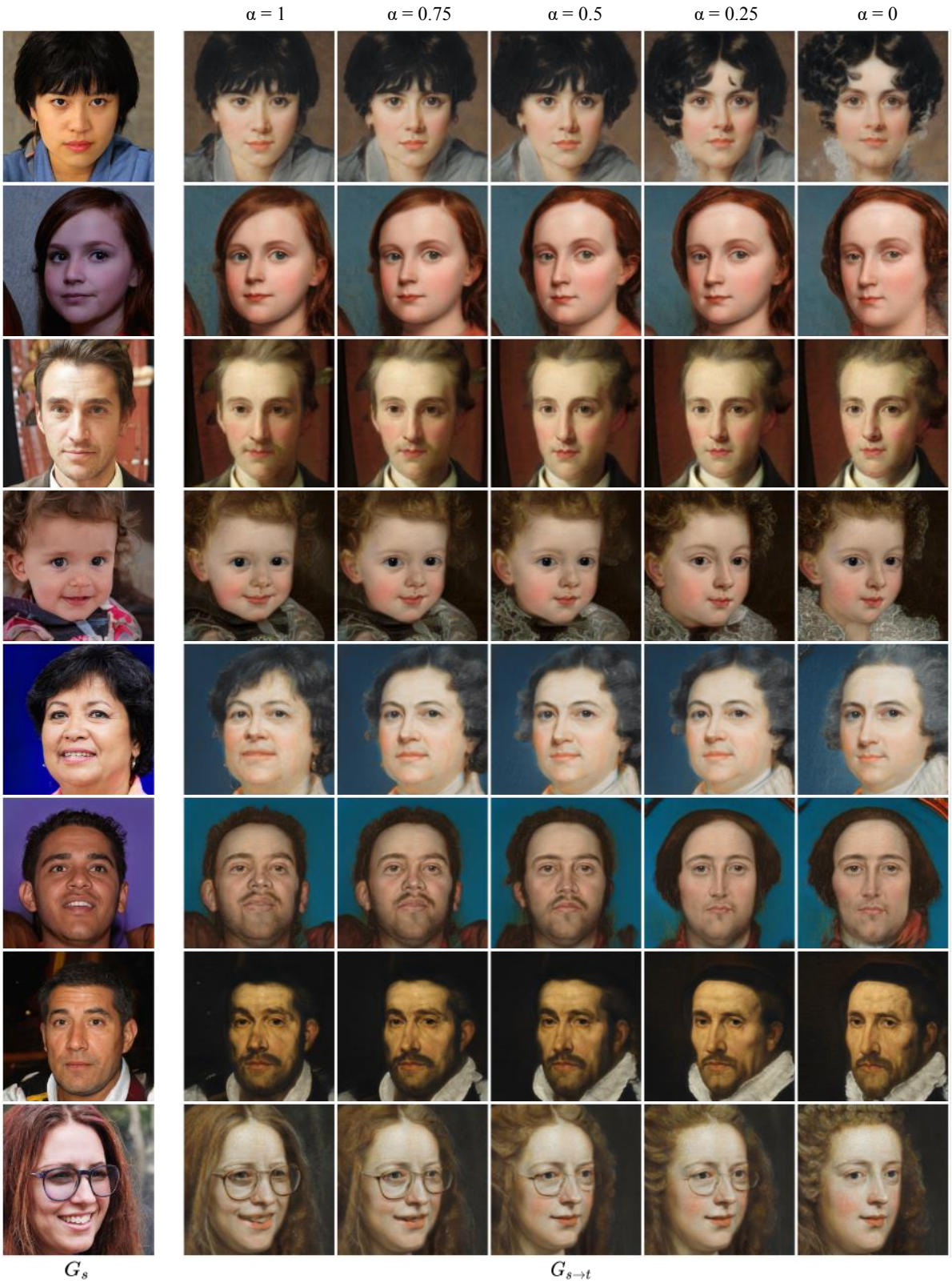

Figure 2. **[FFHQ → MetFaces]** Visualizing the effects of the noise interpolation. The interpolation weight α is presented above each column.

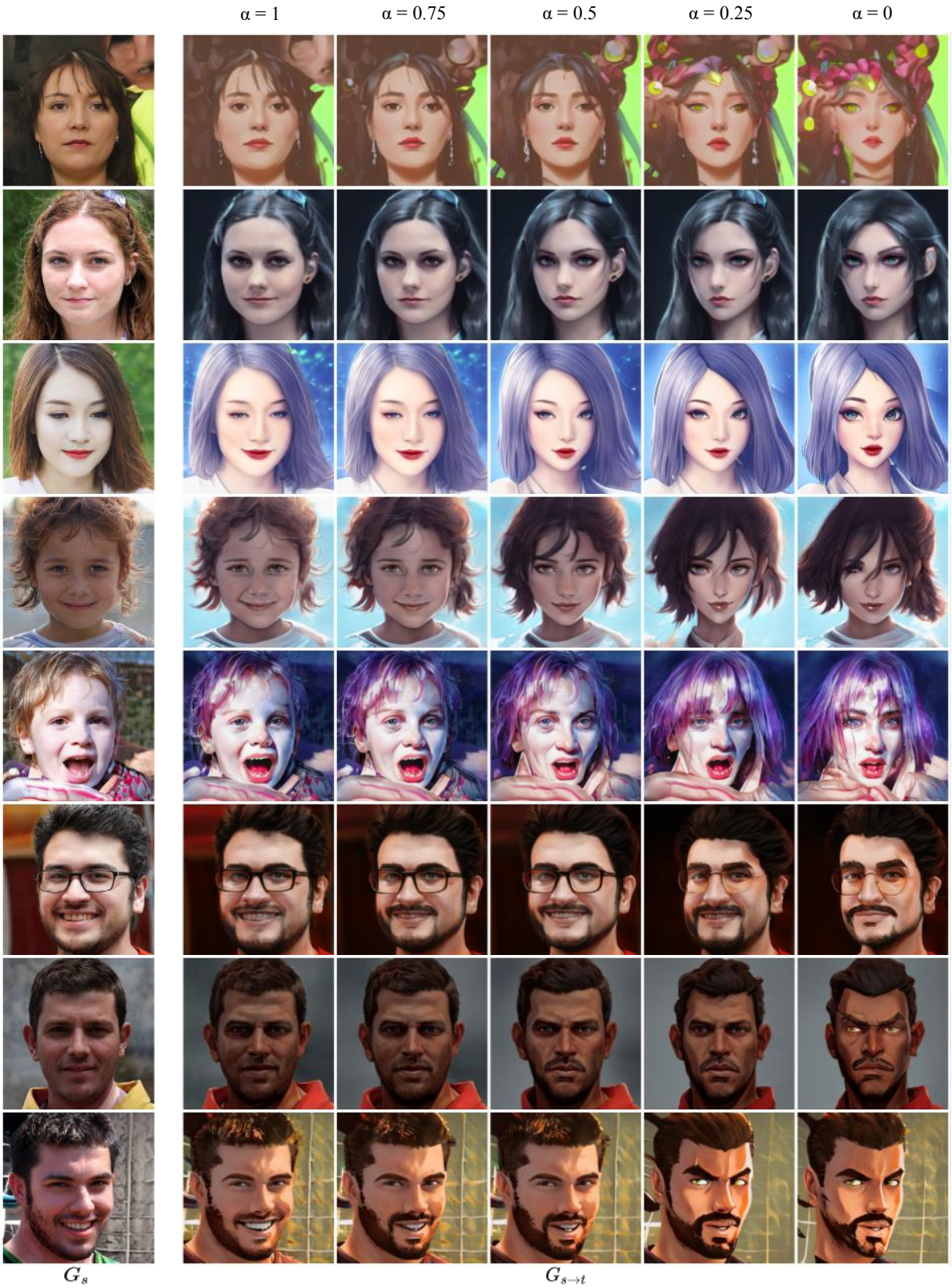

Figure 3. **[FFHQ → AAHQ]** Visualizing the effects of the noise interpolation. The interpolation weight α is presented above each column.

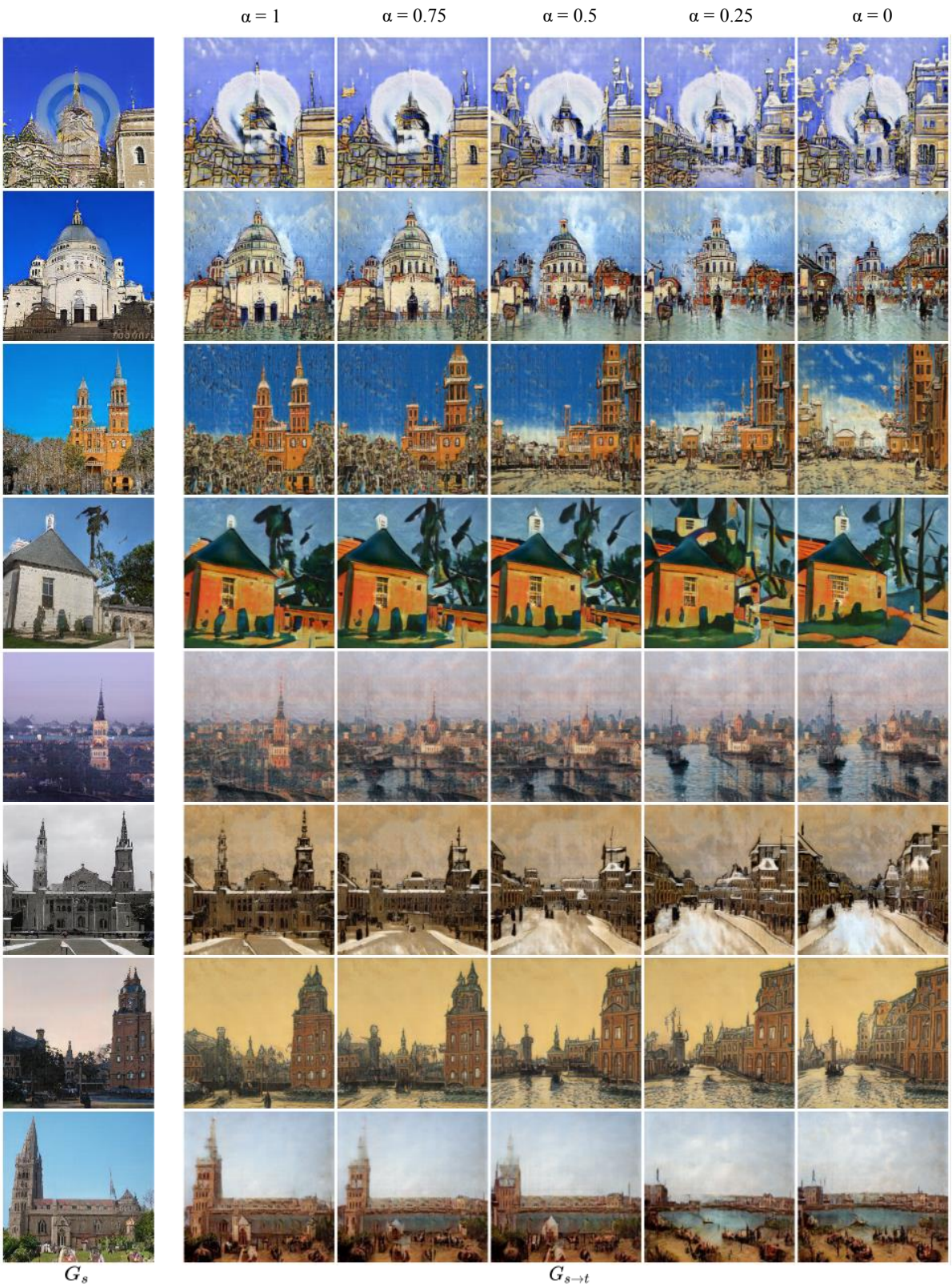

Figure 4. **[Church → Cityscape]** Visualizing the effects of the noise interpolation. The interpolation weight α is presented above each column.

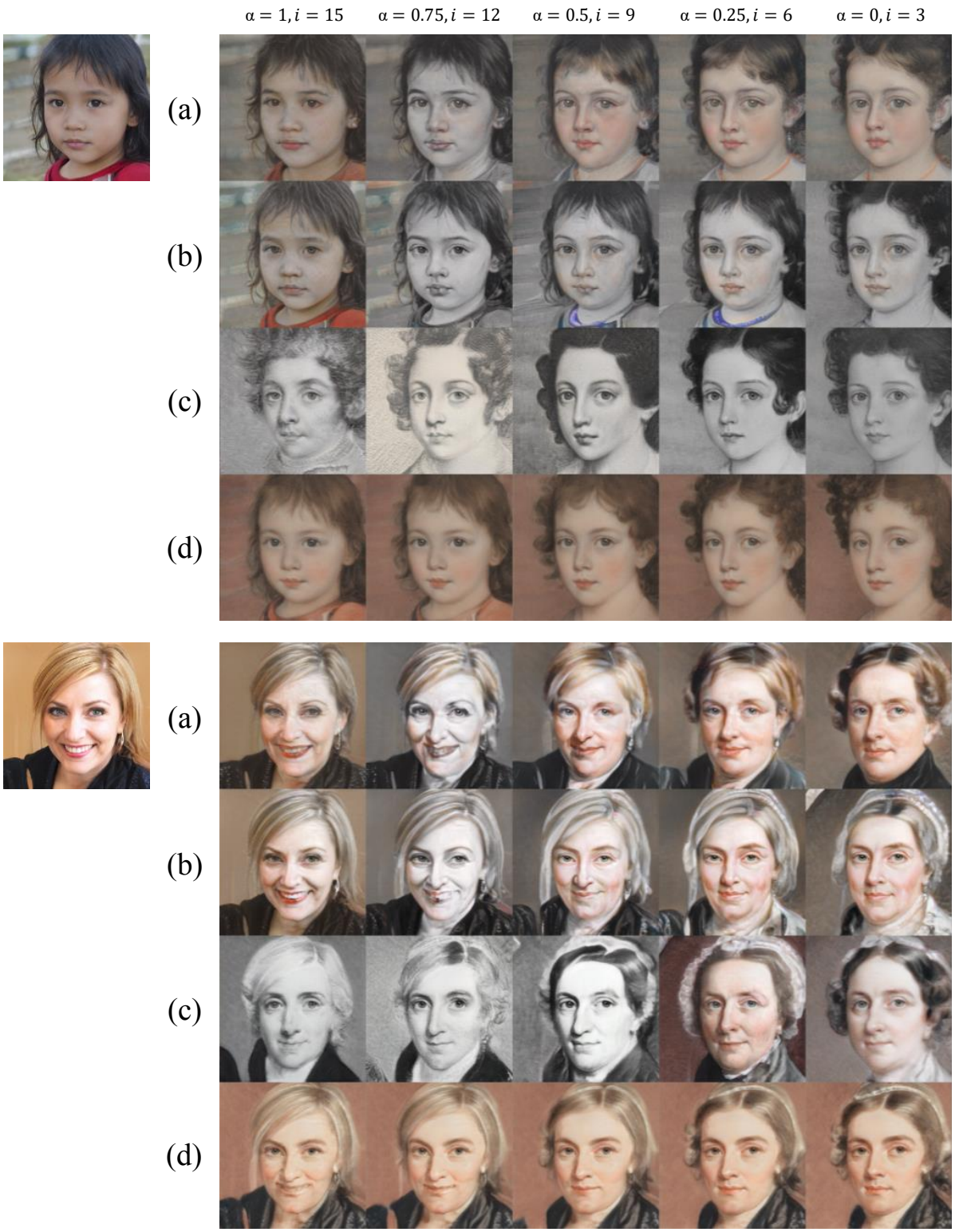

Figure 5. **[FFHQ → MetFaces]** Qualitative comparison on controlling preserved source features: (a) Layer-swap [14], (b) UI2I StyleGAN2 [10], (c) Freeze G [11], (d) ours. The interpolation weight α and swap / freeze layer i are presented above each column.

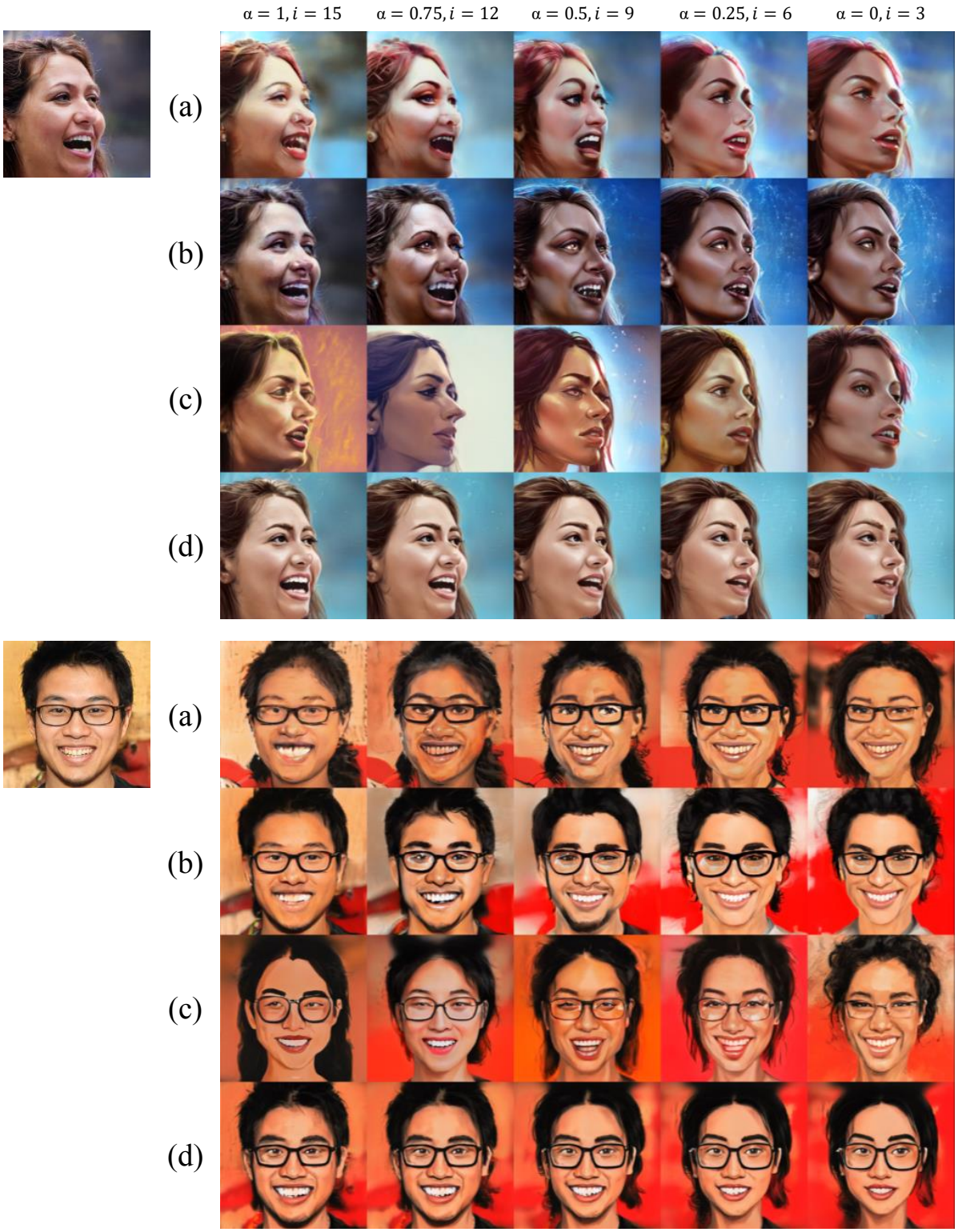

Figure 6. **[FFHQ → AAHQ]** Qualitative comparison on controlling preserved source features: (a) Layer-swap [14], (b) UI2I StyleGAN2 [10], (c) Freeze G [11], (d) ours. The interpolation weight α and swap / freeze layer i are presented above each column.

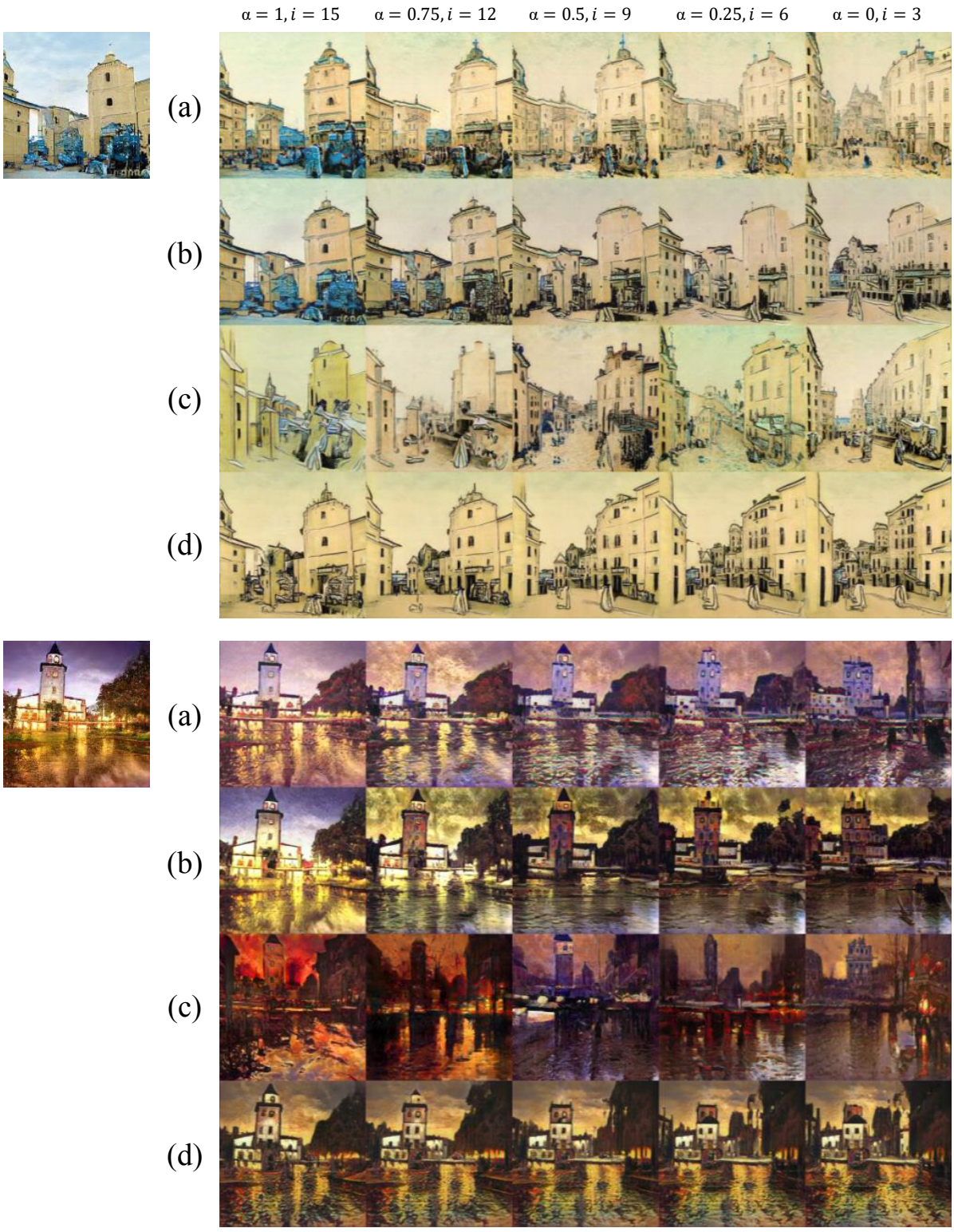

Figure 7. **[Church → Cityscape]** Qualitative comparison on controlling preserved source features: (a) Layer-swap [14], (b) UI2I StyleGAN2 [10], (c) Freeze G [11], (d) ours. The interpolation weight α and swap / freeze layer i are presented above each column.

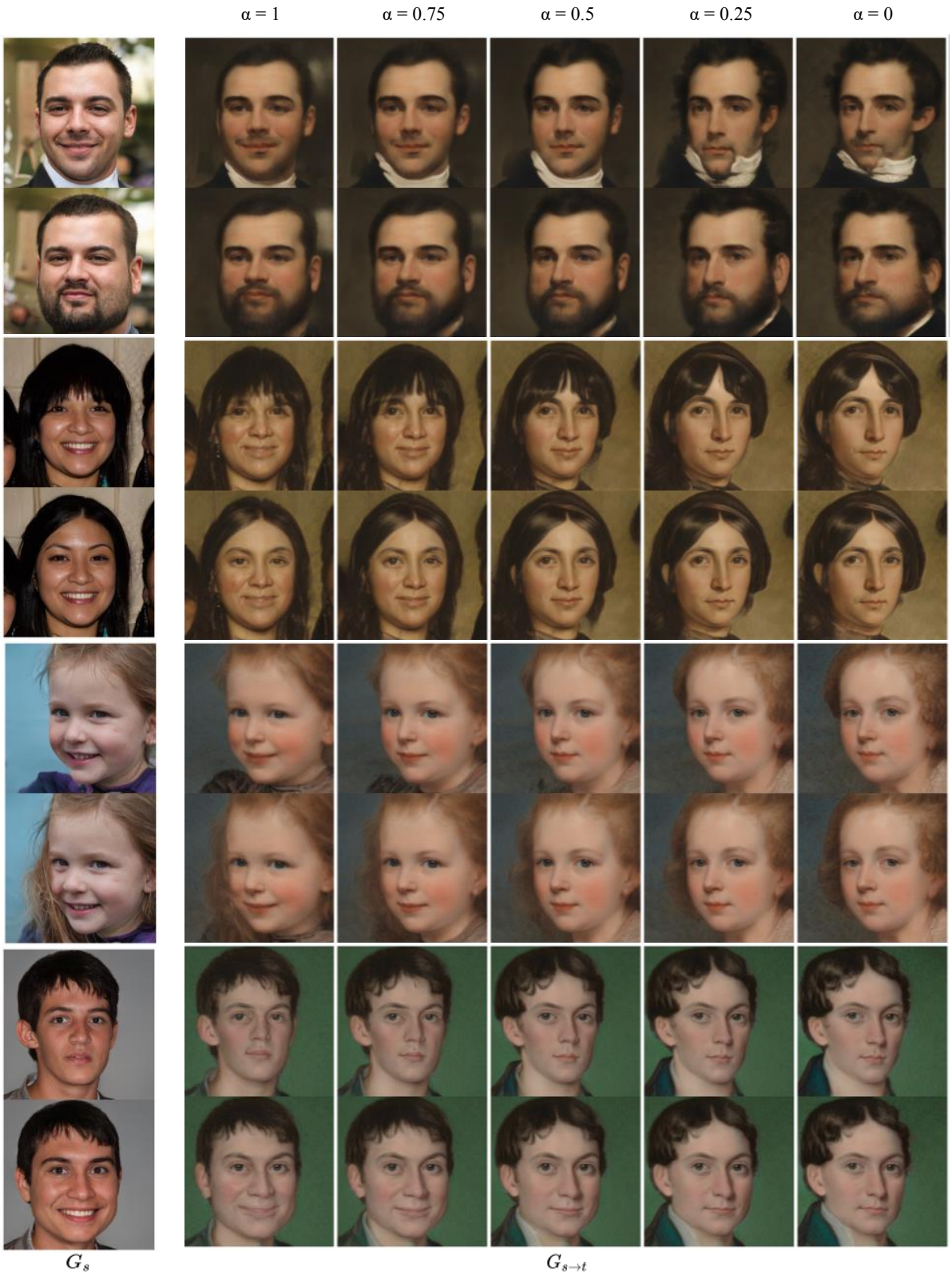

Figure 8. **[FFHQ → MetFaces]** Visualizing the effects of the latent modulation on different interpolation weight. Each of the two adjacent columns is the result of modulating the latent in a different direction $(+/-)$. The interpolation weight α is presented above each column.

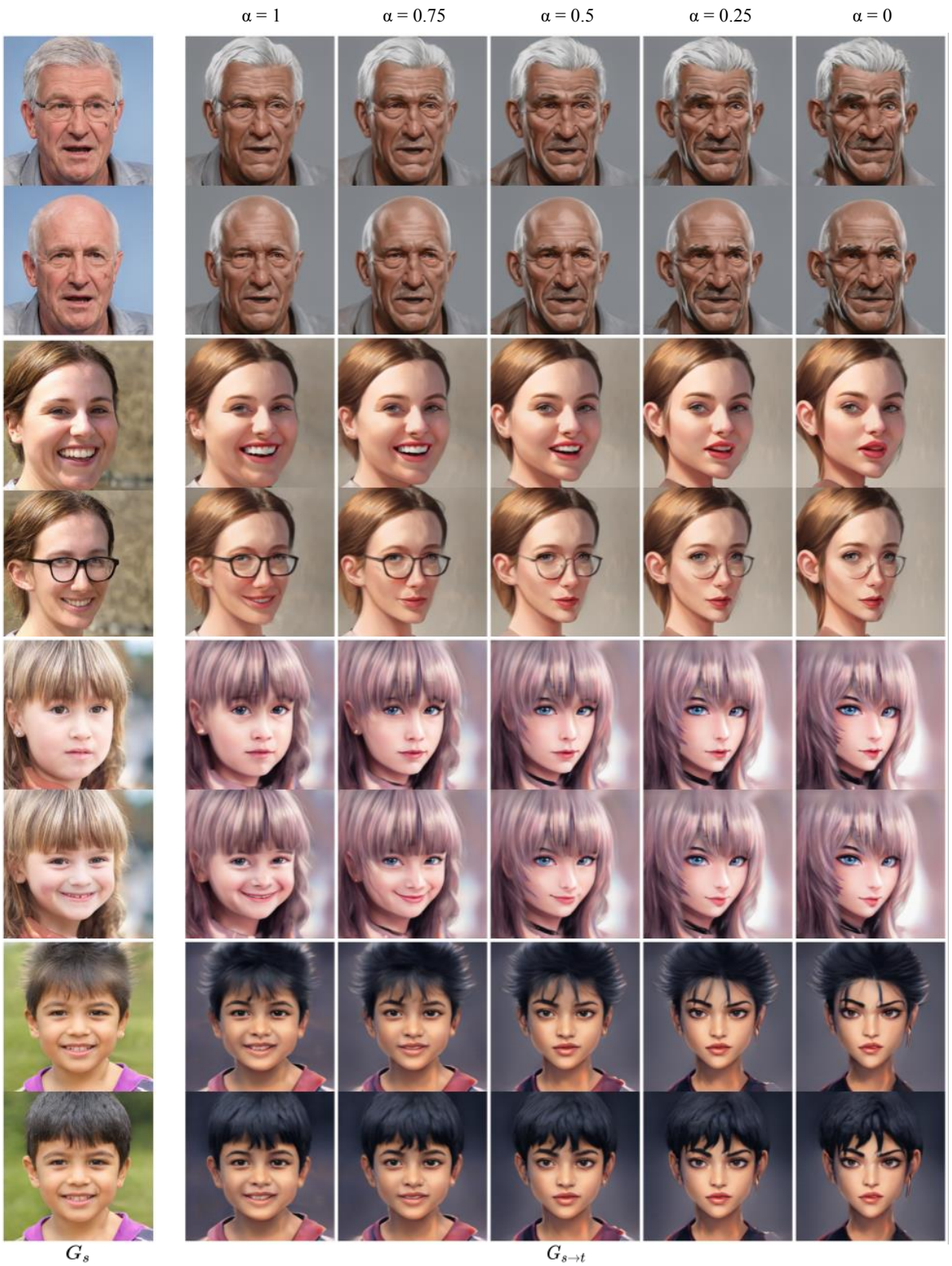

Figure 9. **[FFHQ → AAHQ]** Visualizing the effects of the latent modulation on different interpolation weight. Each of the two adjacent columns is the result of modulating the latent in a different direction $(+/-)$. The interpolation weight α is presented above each column.

Figure 10. **[Church → Cityscape]** Visualizing the effects of the latent modulation on different interpolation weight. Each of the two adjacent columns is the result of modulating the latent in a different direction $(+/-)$. The interpolation weight α is presented above each column.

Figure 11. **[FFHQ → MetFaces]** Qualitative comparison on domain translation. Our method is not only qualitatively best, but also can control source features in a single model.

Figure 12. **[FFHQ → AAHQ]** Qualitative comparison on domain translation. Our method is not only qualitatively best, but also can control source features in a single model.

Figure 13. Domain translation results for Church → Cityscape. Incorrectly projected images cause our method to generate uncorrelated target images with source.

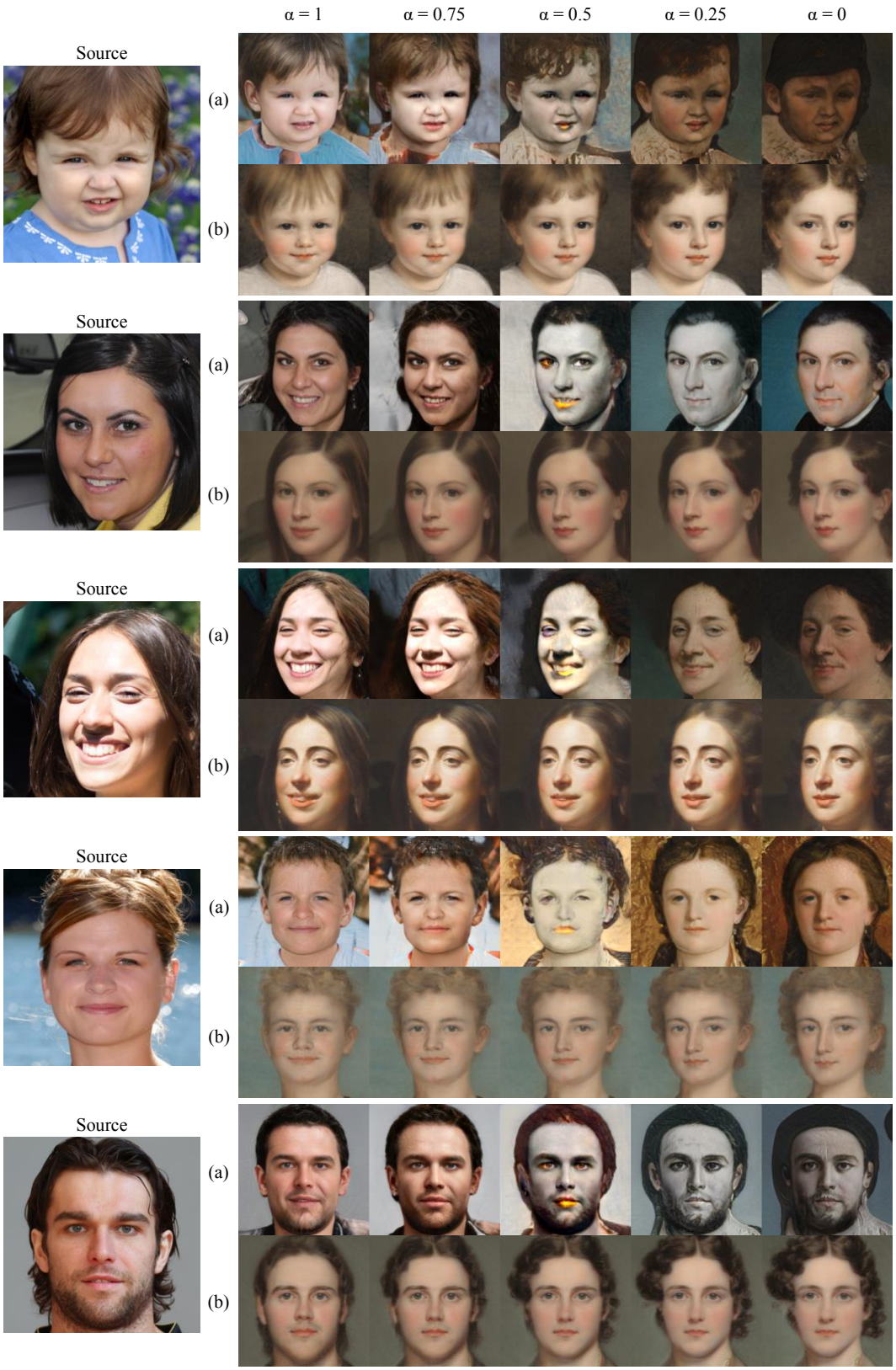

Figure 14. **[FFHQ → MetFaces]** Qualitative comparison on interpolation between source and target features: (a) SmoothingLatentSpace [12], (b) ours. The interpolation weight α is presented above each column.

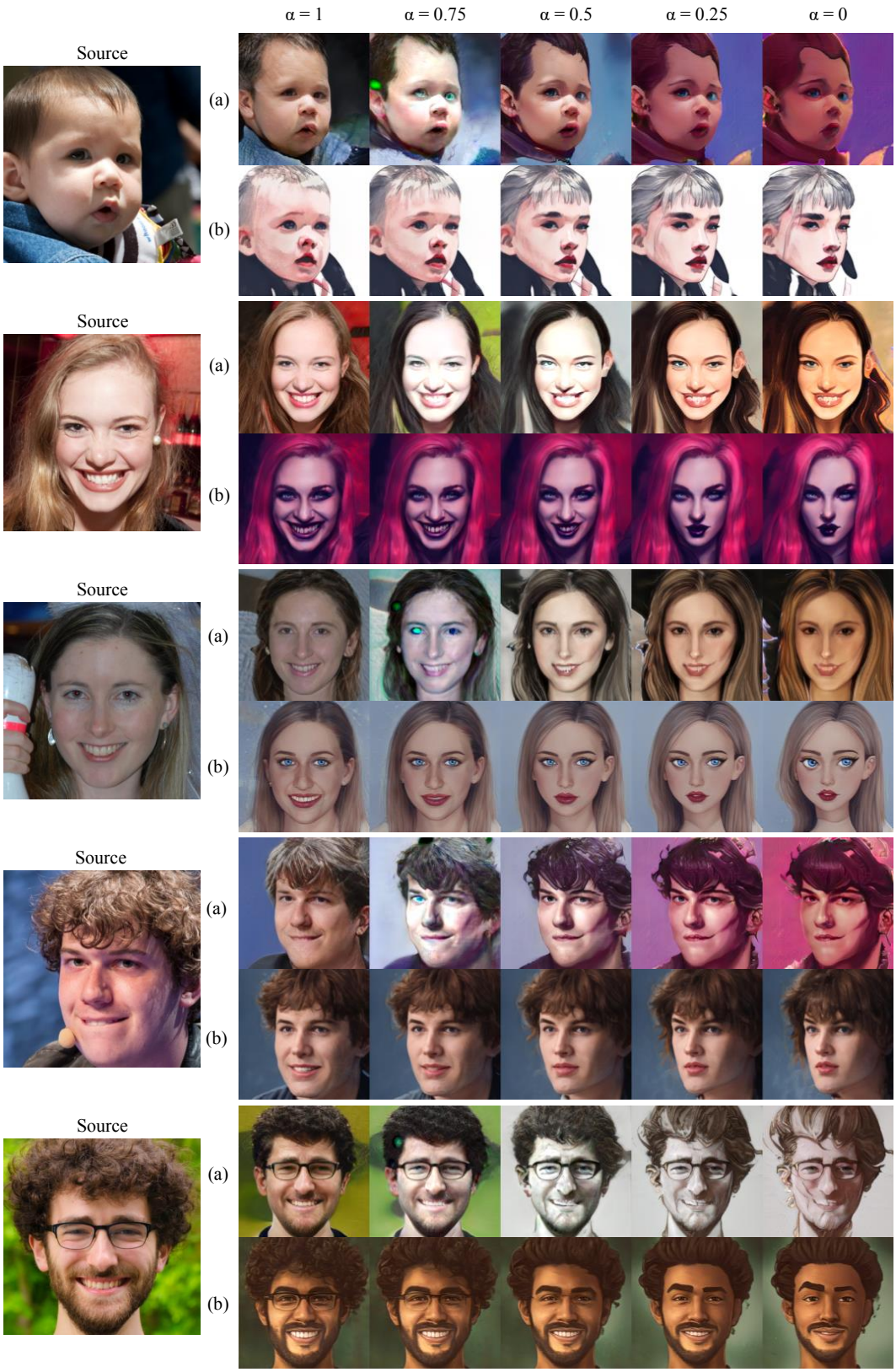

Figure 15. **[FFHQ → AAHQ]** Qualitative comparison on interpolation between source and target features: (a) SmoothingLatentSpace [12], (b) ours. The interpolation weight α is presented above each column.