Supplementary Material:3D-Aware Multi-Class Image-to-Image Translation with NeRFs

Senmao Li[1] Joost van de Weijer[2] Yaxing Wang[1*]
Fahad Shahbaz Khan[3,4] Meiqin Liu[5] Jian Yang[1]
[1]VCIP,CS, Nankai University, [2]Universitat Autònoma de Barcelona
[3]Mohamed bin Zayed University of AI, [4]Linkoping University, [5]Beijing Jiaotong University
senmaonk@gmail.com {yaxing,csjyang}@nankai.edu.cn joost@cvc.uab.es
fahad.khan@liu.se mqliu@bjtu.edu.cn

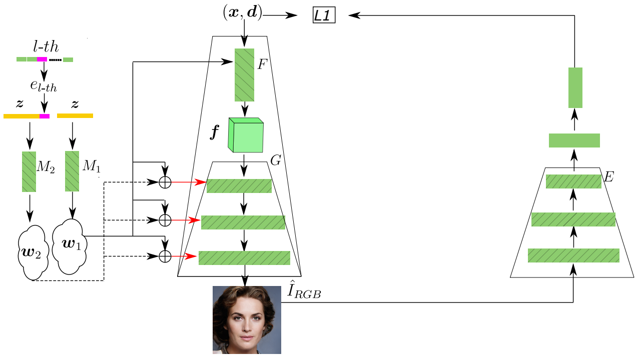

Figure 1. We predict 3D location and 2D viewing direction, and align with the input ones.

1. Network detail

1.1. U-net-like adaptor

Inspired by Pix2pix [2], we design a U-net-like adaptor between the encoder and the decoder. Encoder: C64-C128-C256-C512-C512 Decoder: C512-C256-C128-C64-C64 To specific, we apply Batch-norm and leaky-ReLUs (0.2) after the convolution in the encoder, and Batch-norm and ReLU after the convolution in the decoder except for the last layer. The last layer of the decoder is a convolution of which the number of output channels is 64. The U-Net architecture is identical except for the skip connection between each layer k in the encoder and the layer $n - k$ in the decoder, where n is the number of the decoder.

1.2. Test detail

At inference time, like StarGANv2, we use 50000 images to compute FID. To evaluate the view-consistency, we use 100 videos to compute *TC*, and report the mean value.

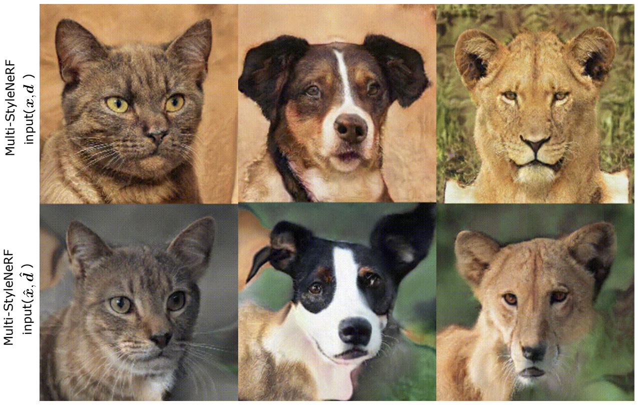
Multi-StyleNeRF input$(\boldsymbol{x}, \boldsymbol{d}$)
Multi-StyleNeRF input$(\hat{\boldsymbol{x}}, \hat{\boldsymbol{d}}$)

Figure 2. The synthesised images with the input $(\boldsymbol{x}, \boldsymbol{d})$ and the predicted $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{d}})$.

2. Ablation

2.1. Alignment with both the 3D location and 2D viewing direction

Instead of outputting the feature map $\hat{\boldsymbol{f}}$, we use the adaptor to output the 3D location $\hat{\boldsymbol{x}}$ and 2D viewing direction $\hat{\boldsymbol{d}}$, and align with the input the 3D location $\boldsymbol{x}$ and 2D viewing direction $\boldsymbol{d}$ (Figure 1). As shown in Figure 2, when utilizing the predicted 3D location $\hat{\boldsymbol{x}}$ and 2D viewing direction $\hat{\boldsymbol{d}}$, we are not able to generate high-quality images like the ones synthesized by the location $\boldsymbol{x}$ and 2D viewing direction $\boldsymbol{d}$.

[*] The corresponding author.

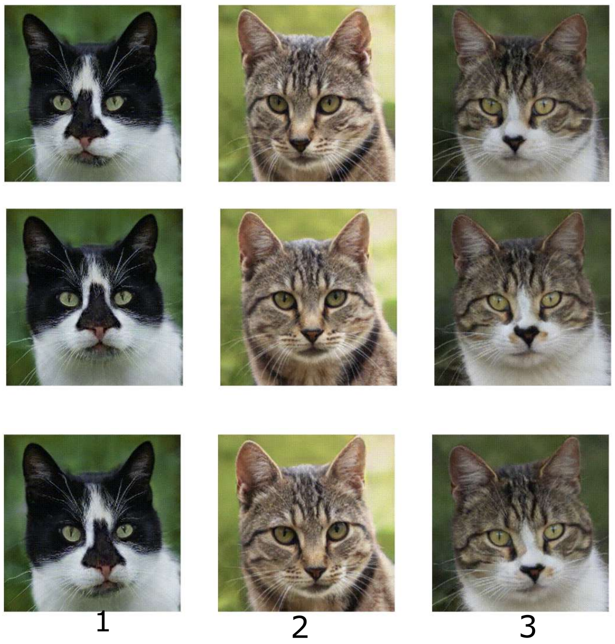

Figure 3. User study example. We conduct a user study and ask subjects to select the results that is *Which video has the best view-consistency? please select one*.

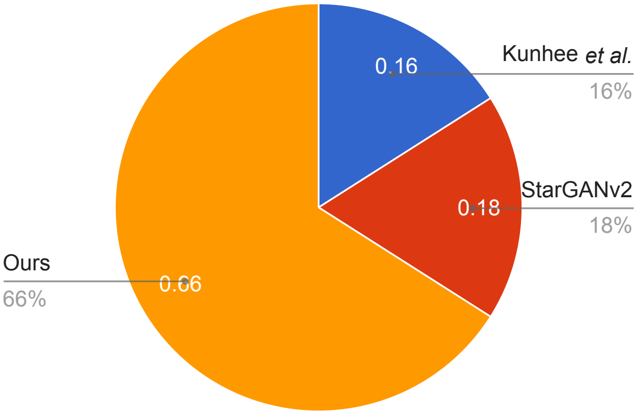

Figure 4. User study.

3. Unconditional 3D-aware I2I translation

We also explore the unconditional 3D-aware I2I translation. As shown in Figure 5 (left), we design the unconditional 3D-aware I2I translation architecture. To be specific, we share the NeRFF, and device the domain-specific generators: G_1 and G_2. Figure 5 (right) shows we test our model. Note, we train this system like multi-class 3D-aware I2I translation. Figure 6 shows the generated images, which indicates the effectiveness of the proposed method.

Method \ Dataset	CelebA-HQ			AFHQ		
	TC↓	vLPIPS↑	TL↓	TC↓	vLPIPS↑	TL↓
StarGANv2	10.250	**0.032**	**0.328**	3.025	0.121	0.366
GP-UNIT	3.065	0.123	0.377	2.073	**0.193**	0.4
Ours (3D)	**3.743**	0.101	0.378	**2.067**	0.165	**0.341**

Table 1. Comparison with baselines on vLPIPS, the temporal loss (TL) and the temporal consistency (TC), where TC=TL/vLPIPS.

4. Additional results

4.1. Quantitative results

Table. 1 reports the result of the additional metrics: vLPIPS and TL. Although StarGANv2 has better performance, it fails to preserve 3D Consistent. We have best score on *TC*, which is corresponding to consistency.

4.2. User study

We conduct a user study and ask subjects to select the results that is *Which video has the best view-consistency? please select one* (Figure 3). We apply triplet comparisons (forced choice) with 14 users (10 triplets/user) for 3D-aware I2I translation. Experiments are performed on images from the AFHQ dataset. Fig. 4 shows that our method considerably outperforms the other methods.

4.3. T-SNE

We investigate the latent space of the generated images. We randomly sample images, and translate them into the target class. Specifically, given the generated images we firstly perform Principal Component Analysis (PCA) [1] to the extracted feature. Then, we conduct the T-SNE [4] to visualize the generated images in a two-dimensional space. As illustrated in Figure 7, the T-SNE plot shows that our method has a similar distribution as the training set.

4.4. Additional Qualitative Results

We provide additional 3D-aware mult-class I2I translation interpolation (Figure 8) and results for models trained on AFHQ and CelebA-HQ datasets in Figures 9, 10, 11, 12, 13, 14, 15, 16 , 17 , 18, 19.

We also provide the demo for StarGANv2, Kunhee *et al.* [3] and ours. While Kunhee *et al.* [3] obtains the best FID score on CelelbA-HQ dataset, it suffers from the view-consistency problem. Please see the accompanying video for more results.

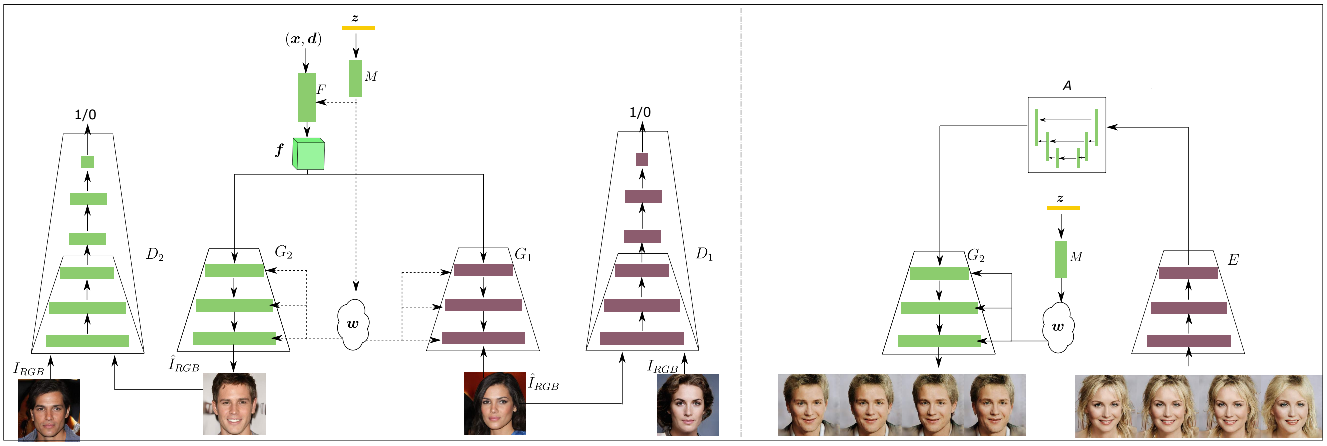

Figure 5. The unconditional 3D-aware I2I translation. (Left) we design unconditonal 3D-aware generative model. Here we share the NeRF mode F. (Right) Usage of proposed model at inference time. Note we train this system like multi-class 3D-aware I2I translation.

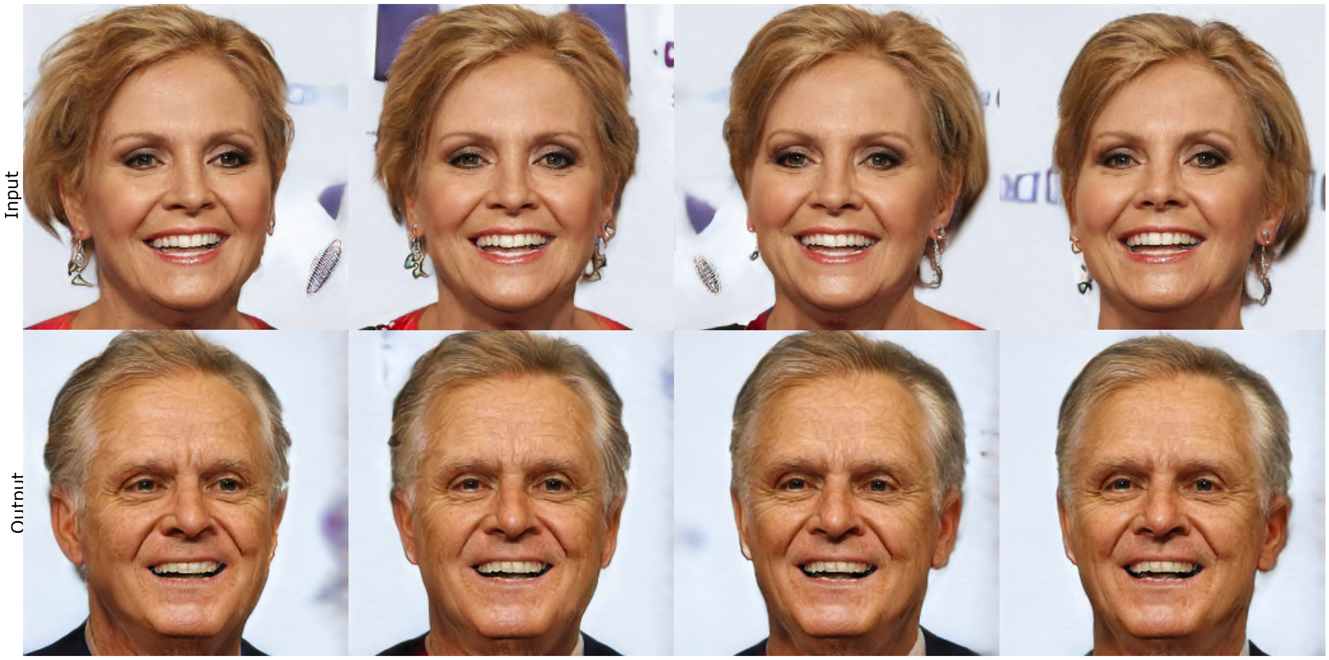

Figure 6. The synthesized images by the proposed unconditional 3D-aware I2I translation.

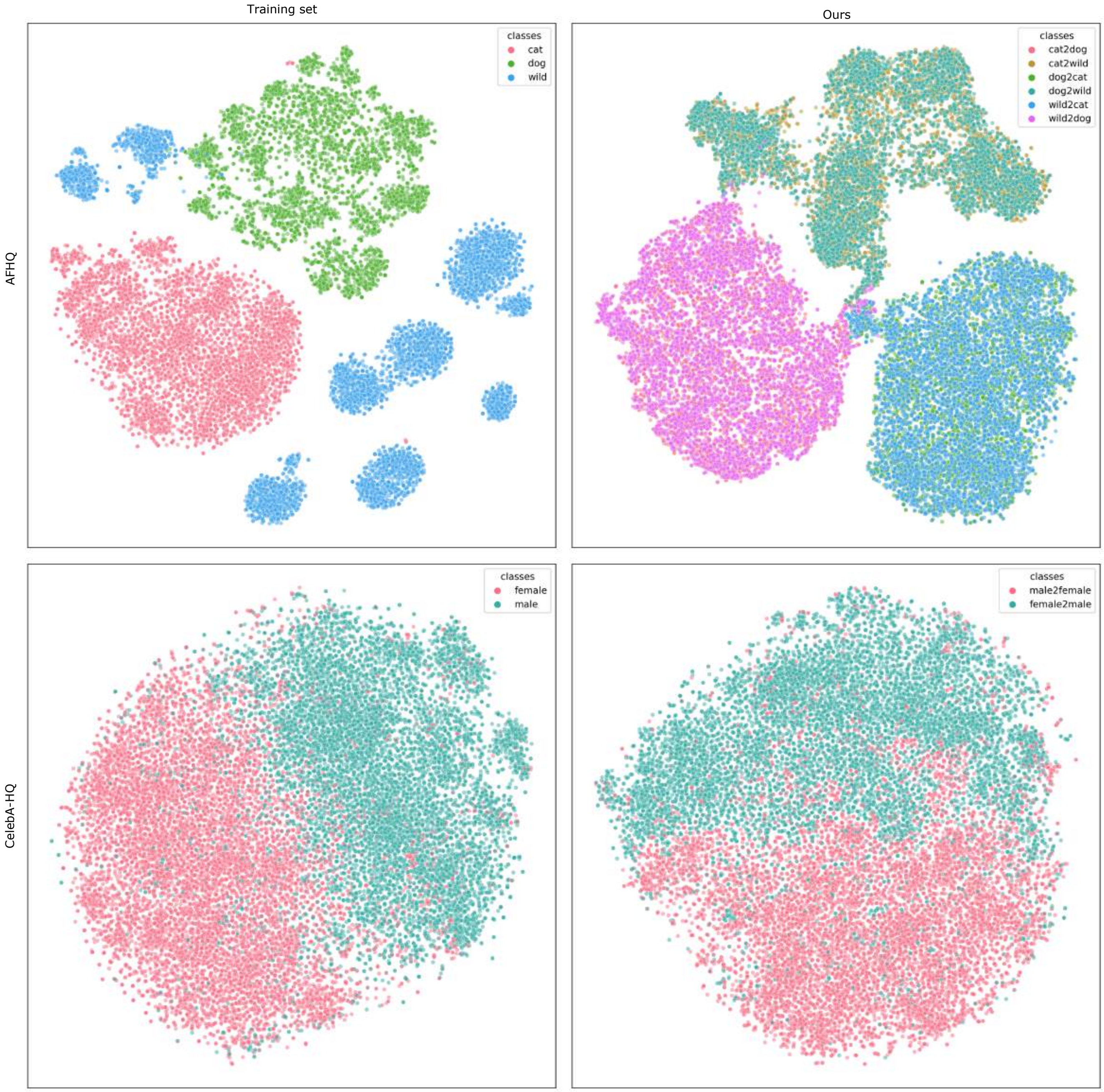

Figure 7. We show T-SNE of both the training sett and the proposed method on two datasets. We exhibits the translation for each category (top right). We observe that there is the similar distribution when the target domain is same.

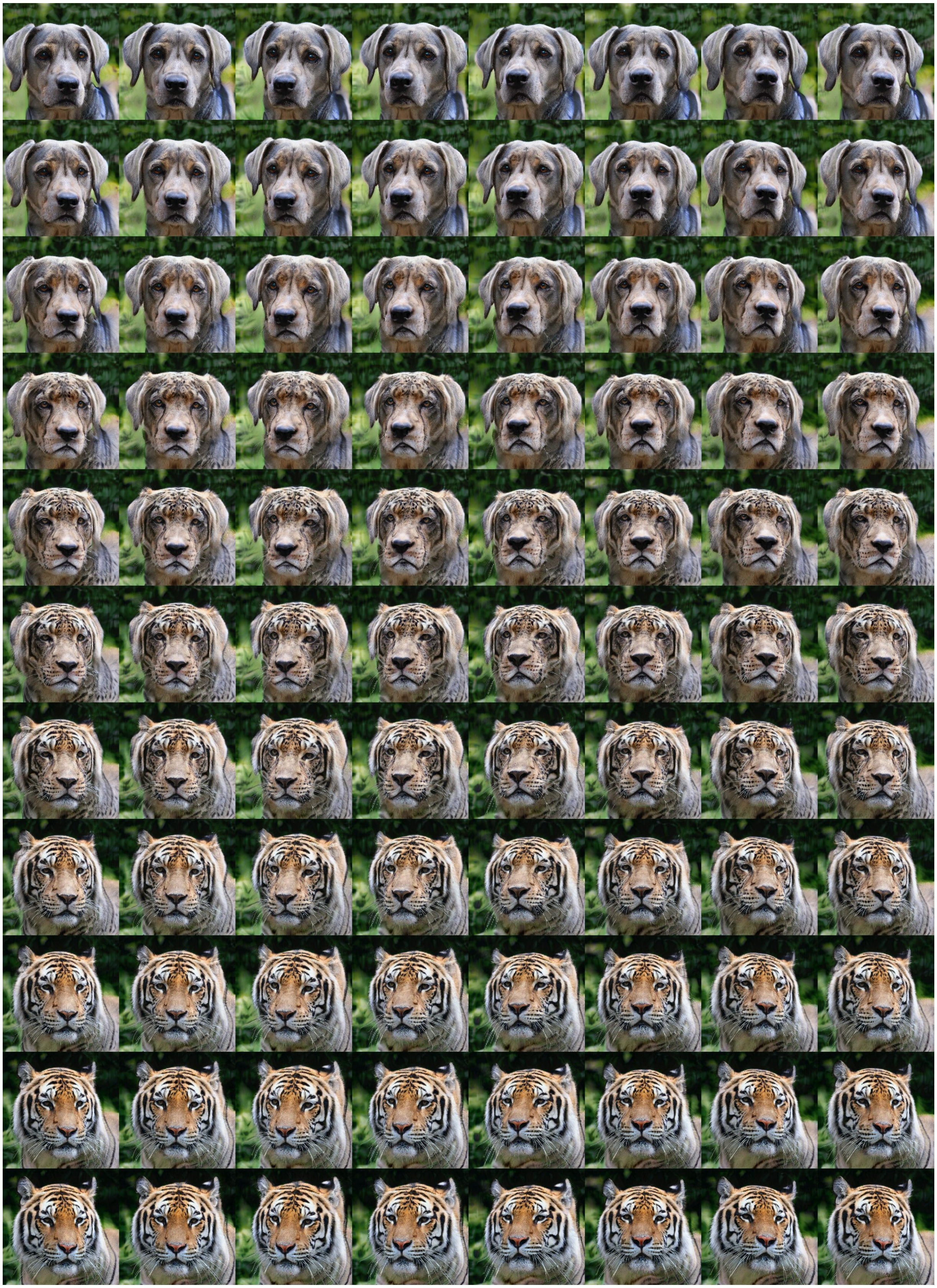

Figure 8. Interpolation between dog and wild on AFHQ 256^2.

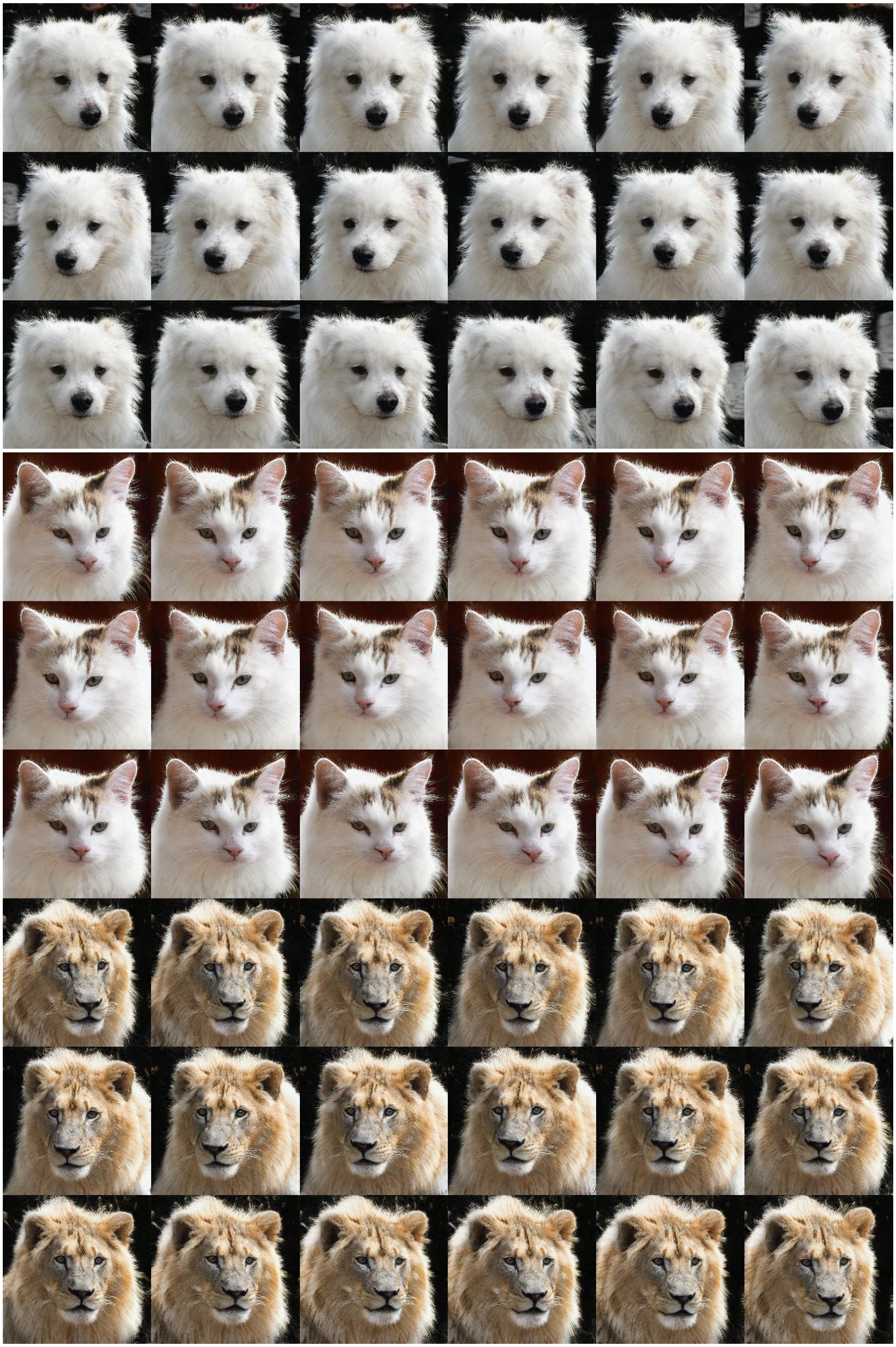

Figure 9. Example of 3D-aware I2I translation of dog into cat and wild on AFHQ 256^2.

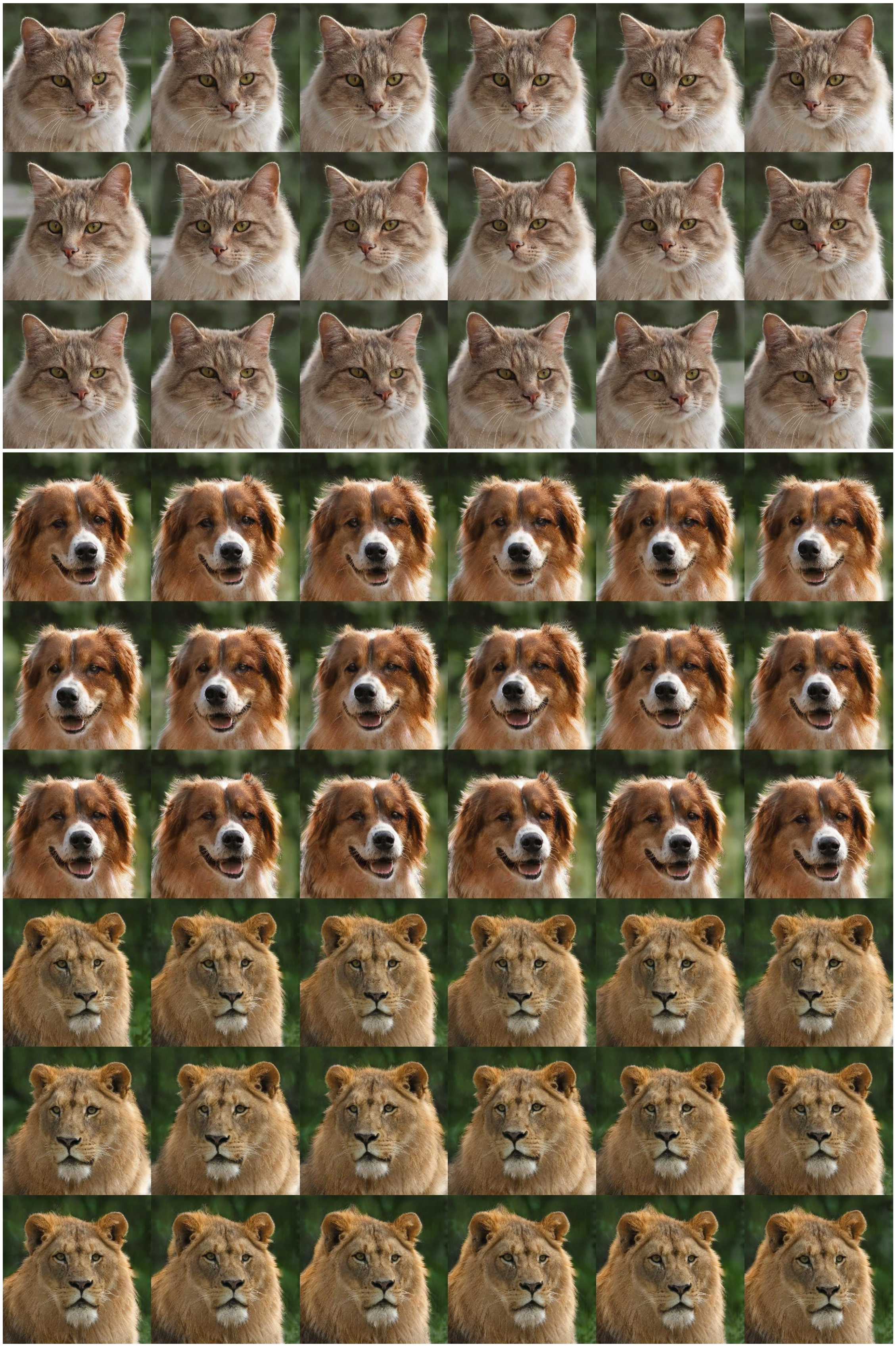

Figure 10. Example of 3D-aware I2I translation of cat into dog and wild on AFHQ 256^2.

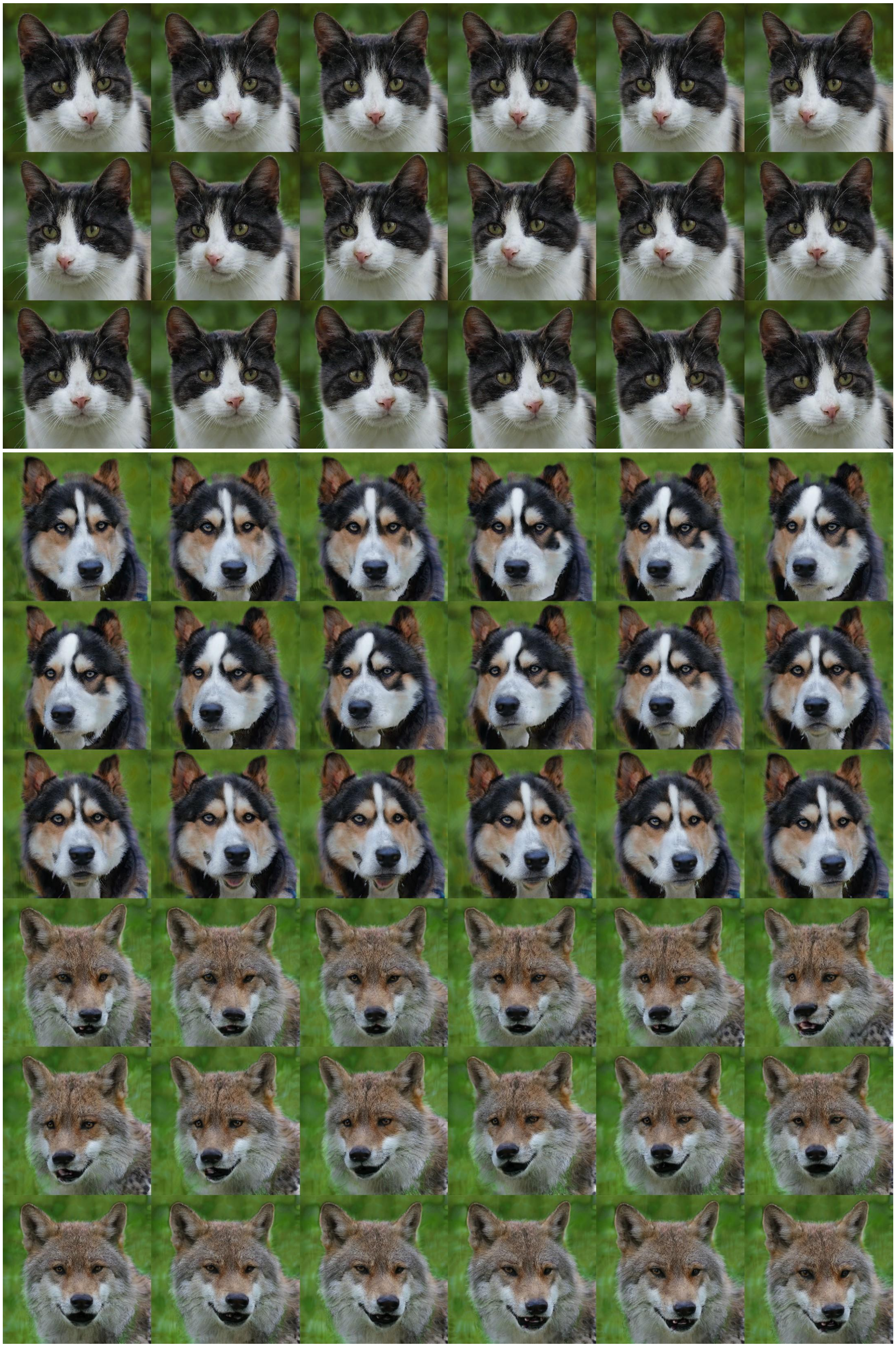

Figure 11. Example of 3D-aware I2I translation of cat into dog and wild on AFHQ 256^2.

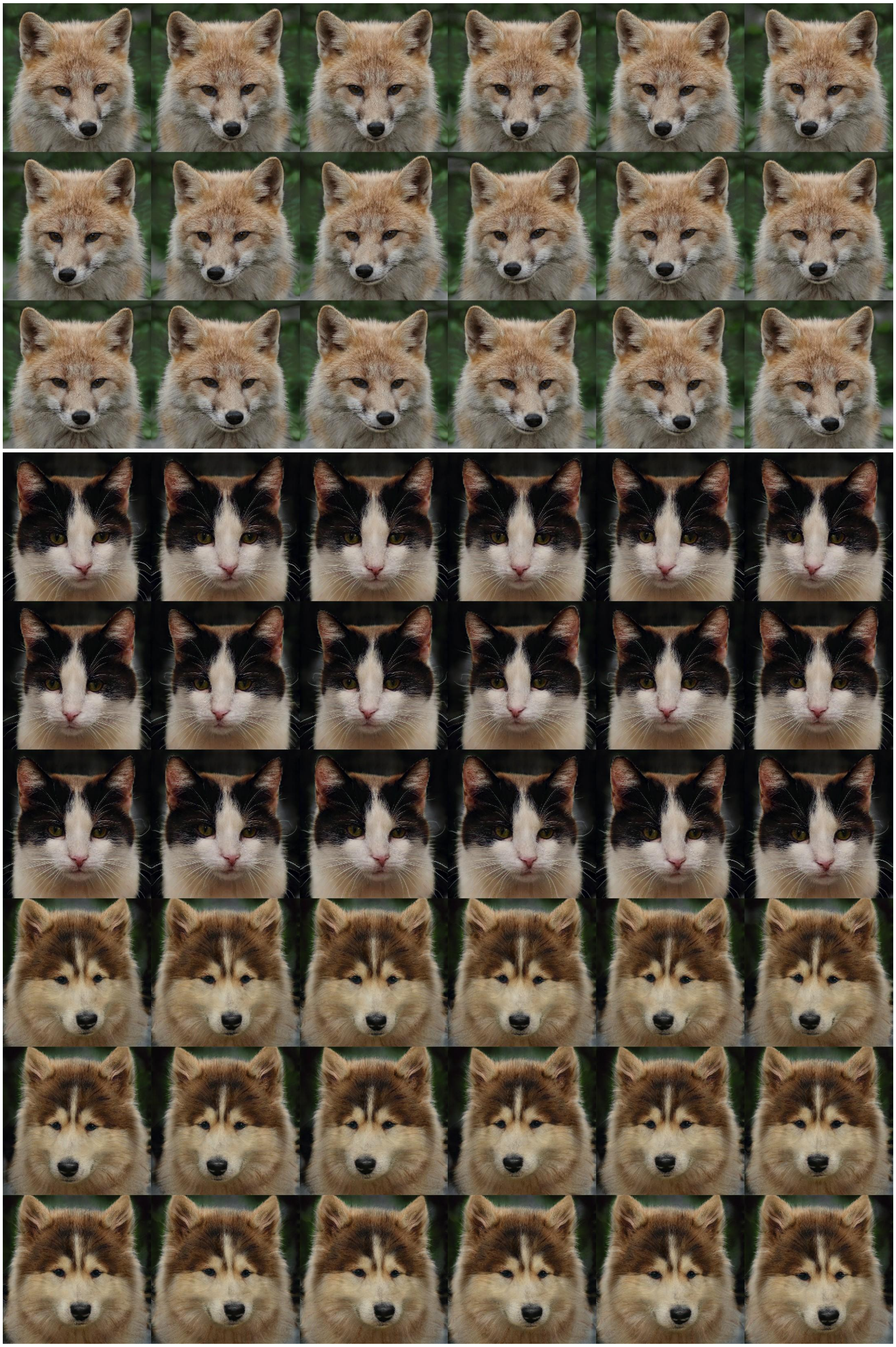

Figure 12. Example of 3D-aware I2I translation of wild into cat and dog on AFHQ 256^2.

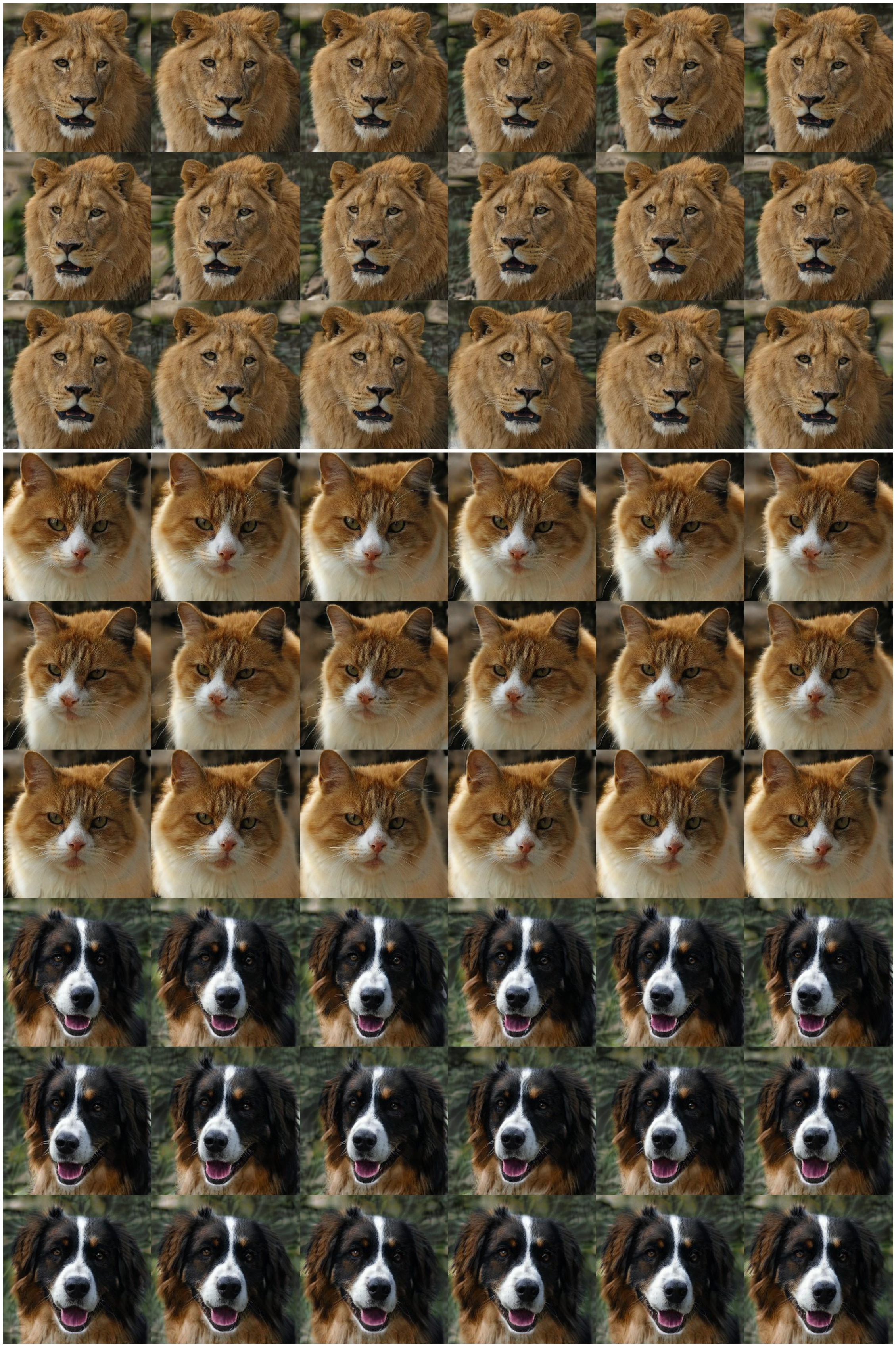

Figure 13. Example of 3D-aware I2I translation of wild into cat and dog on AFHQ 256^2.

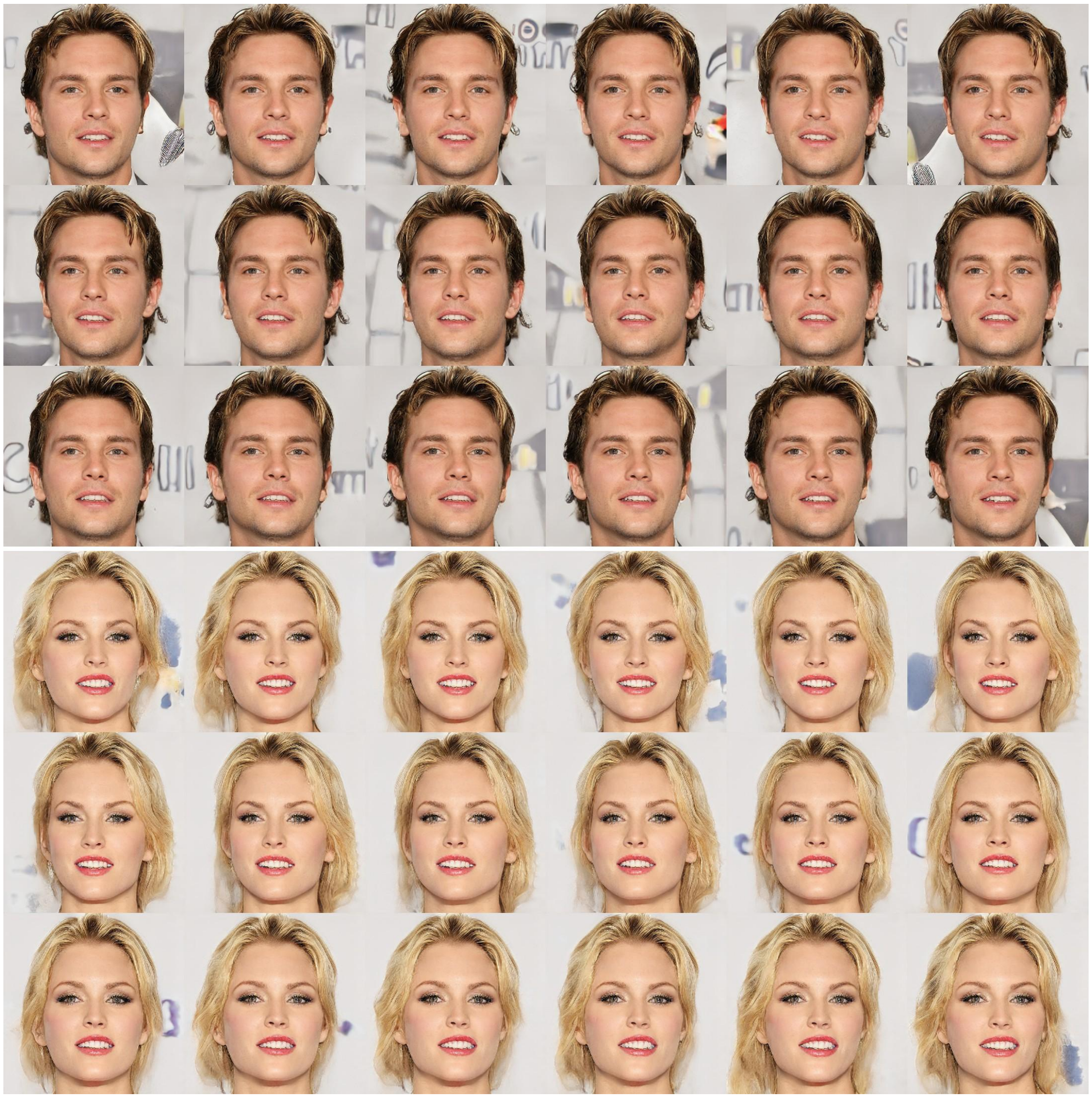

Figure 14. Example of 3D-aware I2I translation of male into female on Celeba-HQ 256^2.

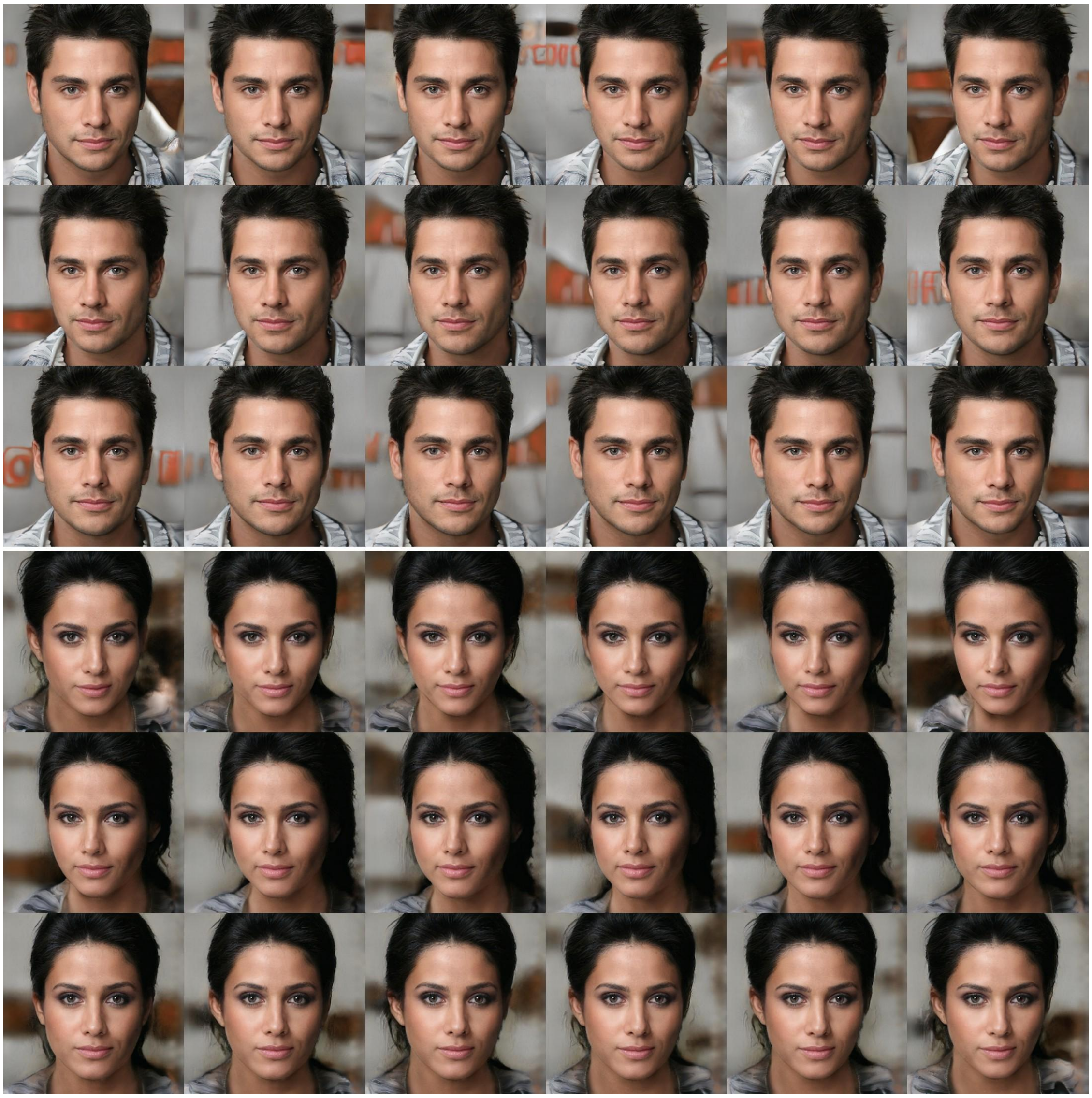

Figure 15. Example of 3D-aware I2I translation of male into female on Celeba-HQ 256^2.

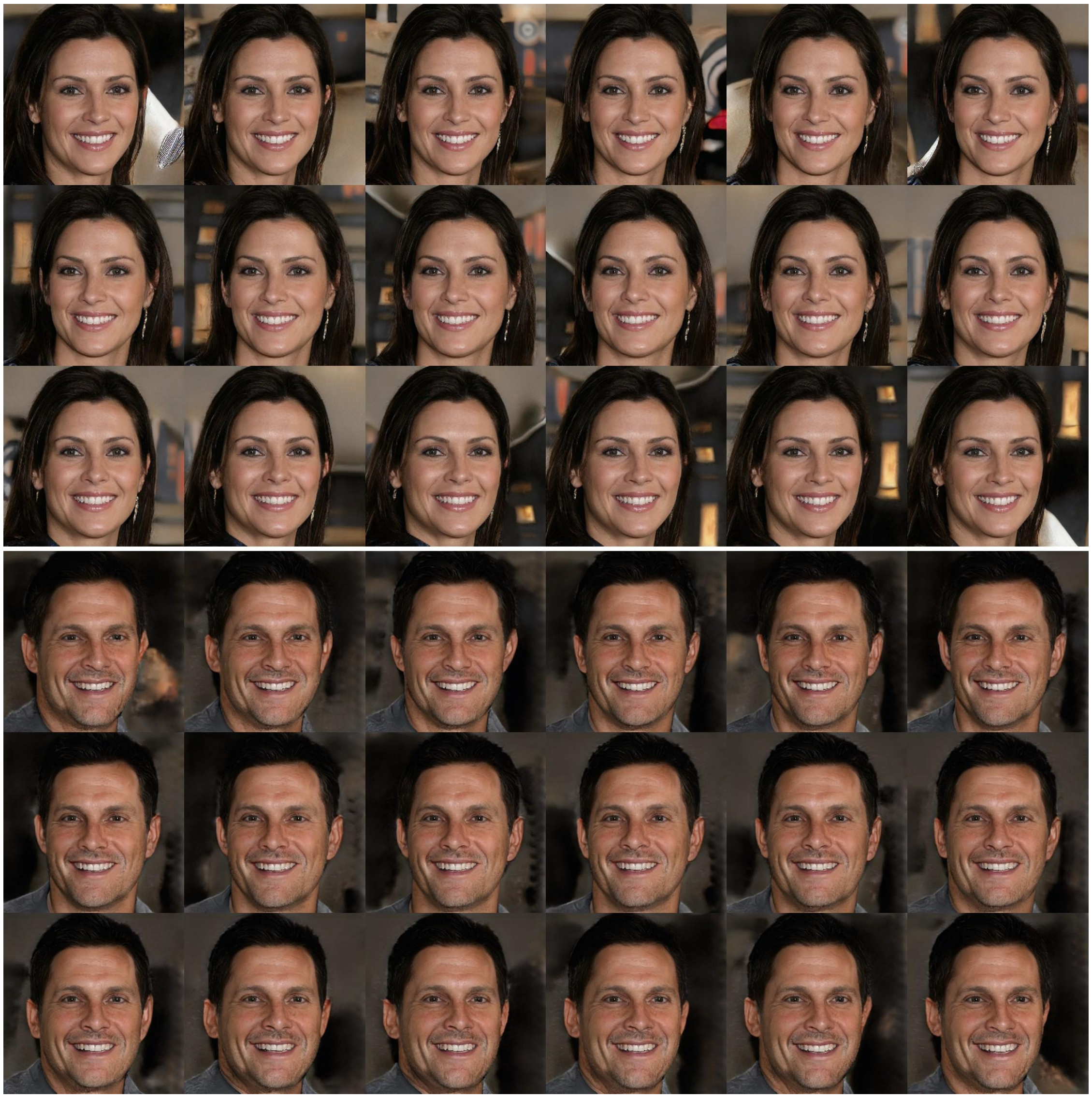

Figure 16. Example of 3D-aware I2I translation of female into male on Celeba-HQ 256^2.

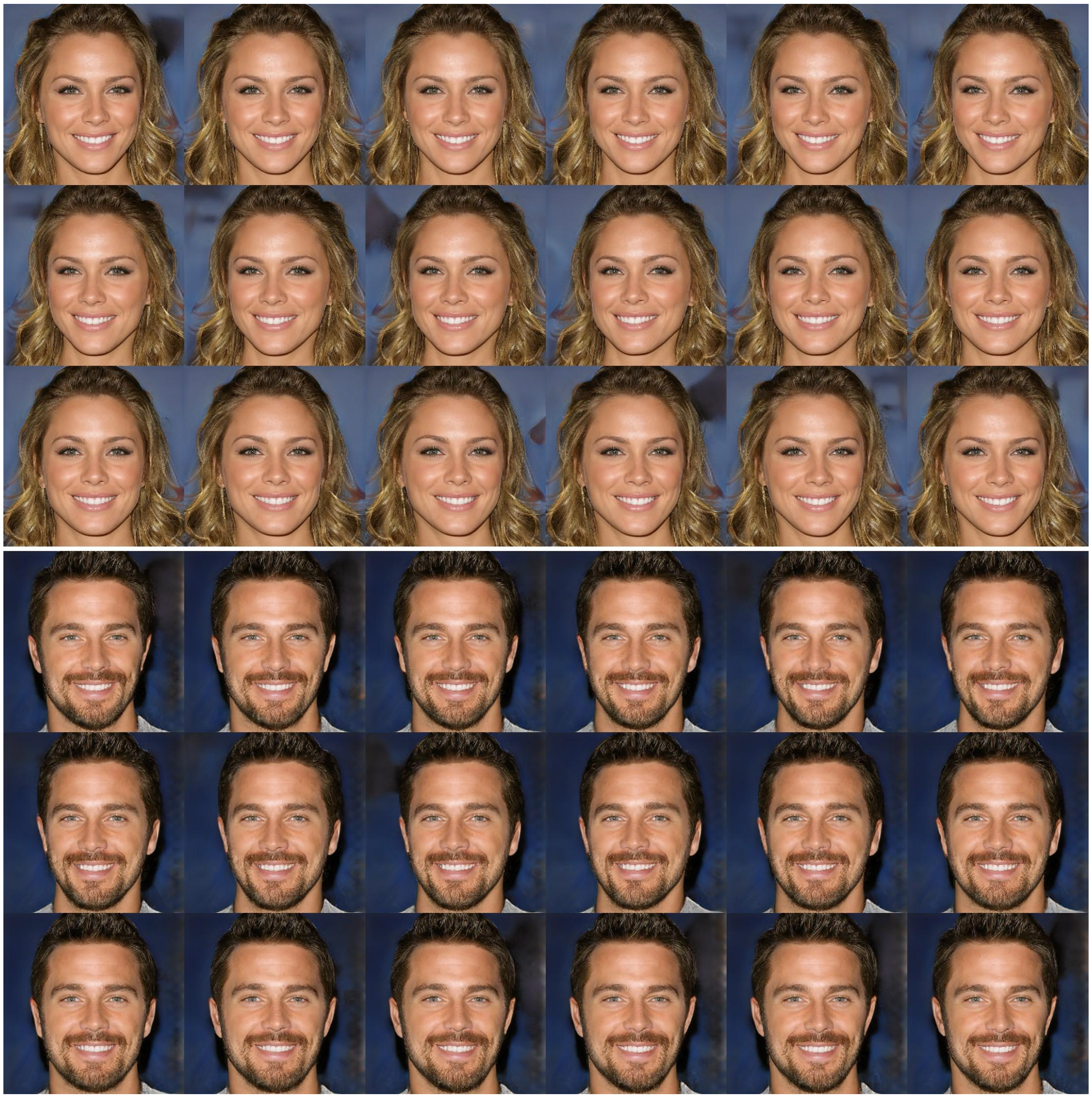

Figure 17. Example of 3D-aware I2I translation of female into male on Celeba-HQ 256^2.

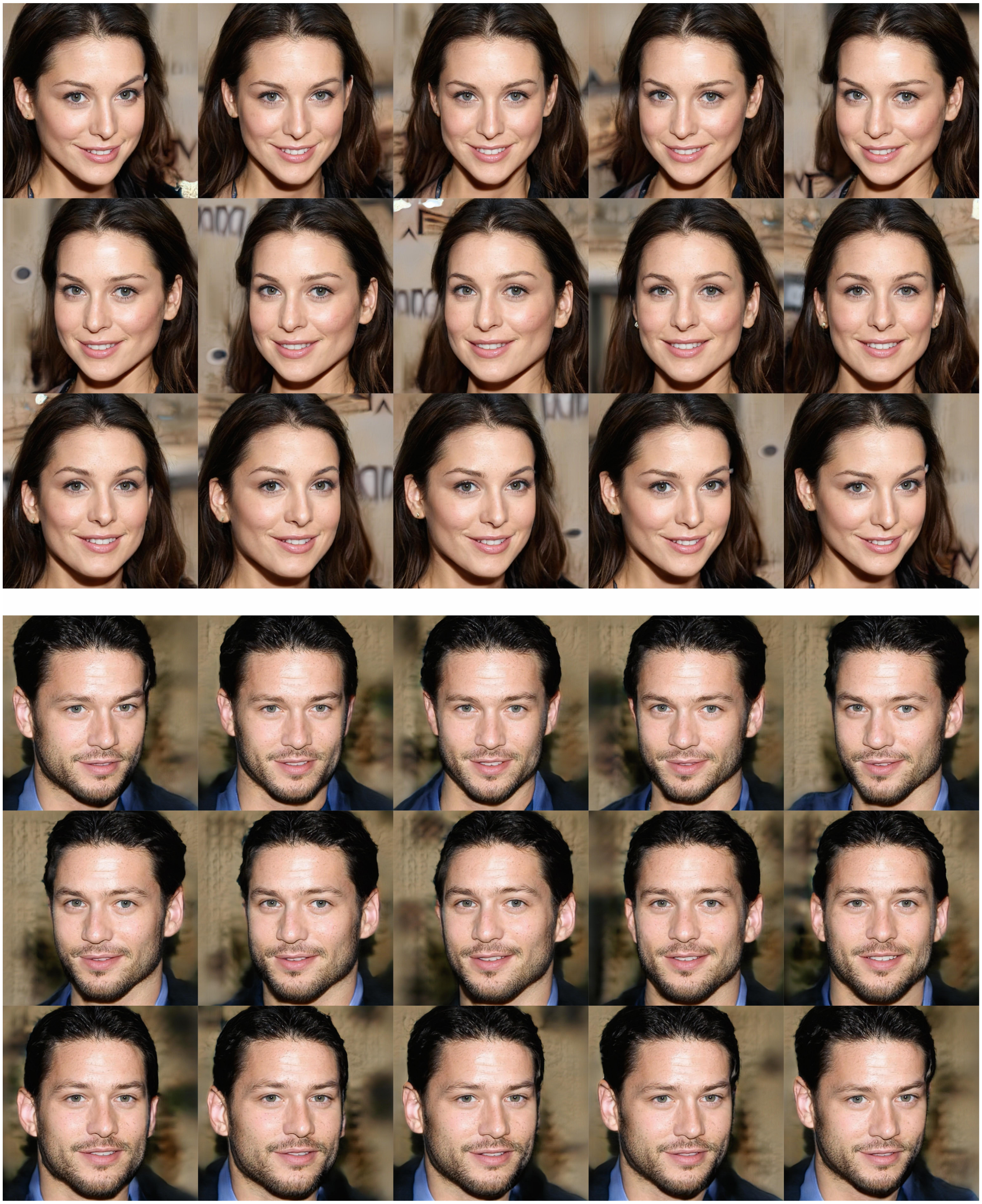

Figure 18. Example of 3D-aware I2I translation of female into male on Celeba-HQ 1024^2.

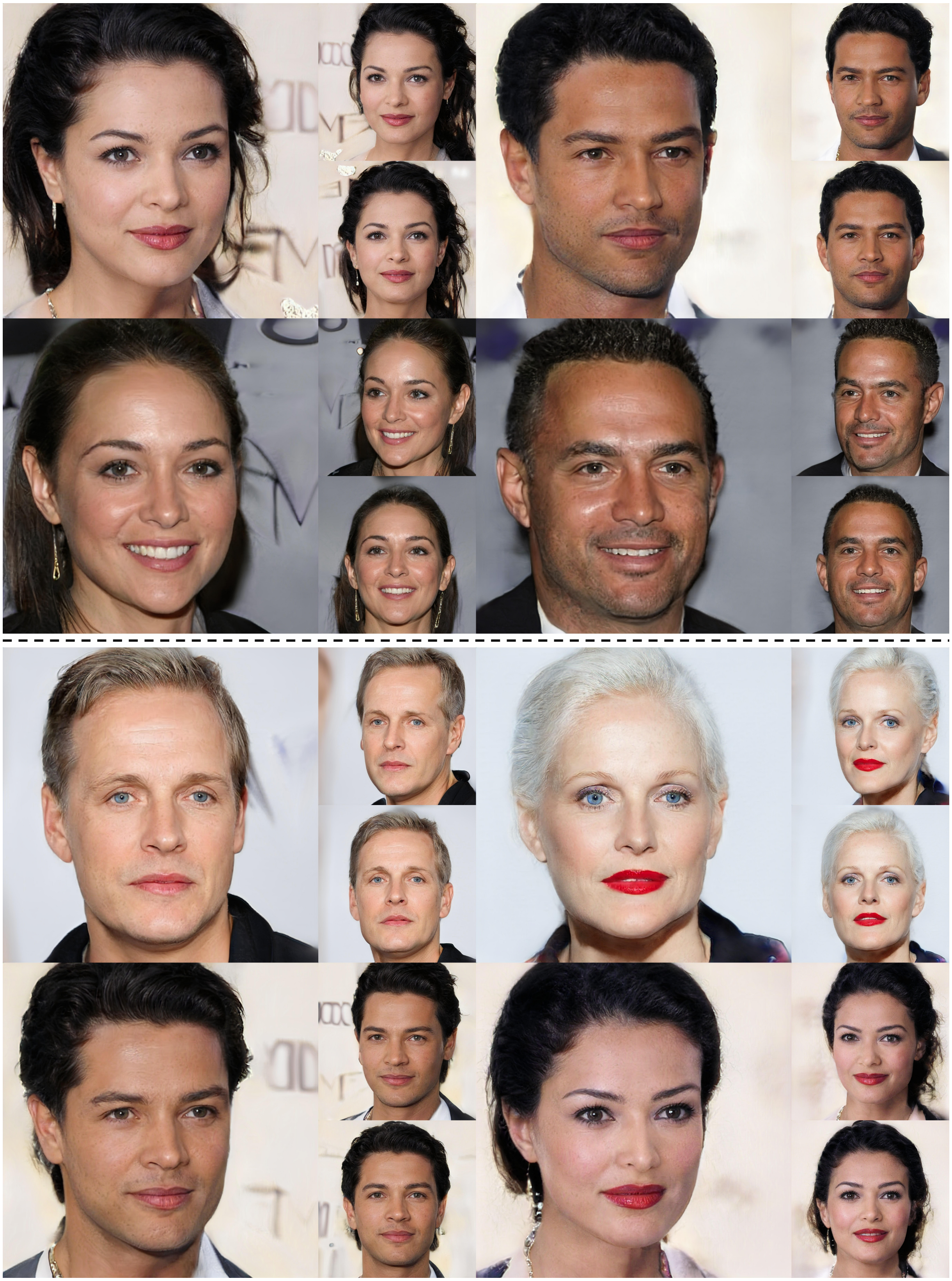

Figure 19. Example of 3D-aware I2I translation of female into male (top) and male into female (bottom) on Celeba-HQ 1024^2.

References

[1] Rasmus Bro and Age K Smilde. Principal component analysis. *Analytical Methods*, 6(9):2812–2831, 2014. 2

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 1

[3] Kunhee Kim, Sanghun Park, Eunyeong Jeon, Taehun Kim, and Daijin Kim. A style-aware discriminator for controllable image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18239–18248, 2022. 2

[4] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 2