Supplementary Material
for ZIGNeRF: Zero-shot 3D Scene Representation with Invertible Generative Neural Radiance Fields

Kanghyeok Ko Minhyeok Lee*
Chung-Ang University
Seoul, South Korea
{dogworld12, mlee}@cau.ac.kr

1. Introduction of Supplementary Material

In this supplemental document, we experimentally substantiate infeasibility of optimization-based GAN inversion approaches and suitability of our learning-based GAN inversion design in Sec. 2. We offer a detailed overview of the various architectural elements within the network – including the feature fields, the neural renderer, and the discriminator, all discussed in Sec. 3. Furthermore, we bring forth additional qualitative findings on datasets such as CelebA-HQ [2], CompCar [3], and AFHQ [1].

2. Comparison with Optimization-based GAN inversion Approach

We observe the limitations of optimization-based GAN inversion approaches for this specific application. Generally, optimization-based GAN inversion methods tend to restore target images more accurately, albeit at a slower speed. However, in the setting of 3D-aware generative pipeline, optimization-based methods are contingent upon the availability of camera parameter for training images to align the viewing direction of the restored image. For evaluation, we performed optimization-based GAN inversion using a gradient descent approach with the Adam optimizer. Our experimental evidence, as depicted in Fig. 1, shows that optimization-based method does not properly restore the input images. In contrast, our model operates effectively without the need for camera parameters.

Furthermore, we conduct additional experiments with LPIPS [4] metric for evaluating the input preservation. The LPIPS measures perceptual similarity using AlexNet pretrained with ImageNet. We compared 1000 pairs of the generated images and the corresponding target images. As shown in Tab. 1, our learning-based GAN inversion method is superior to the optimization-based GAN inversion approaches for all datasets.

3. Network Architectures

In this section, we provide the details of network architecture: feature fields, neural renderer, and the discriminator as exhibited in Fig. 2 and Fig. 3.

Figure 2 presents a detailed overview of the architecture underpinning the feature fields and the neural renderer. The construct of the feature fields is parameterized via multi-layer perceptrons, colloquially referred to as MLPs, a feature vividly displayed in subfigure (a). This setup maps a three-dimensional point, the viewing direction, along with latent codes into a volume density and a feature. Subfigure (b) unravels the process behind the neural renderer blocks, demonstrating how these blocks transform a volume-rendered feature image, into the ultimate synthesized image.

Figure 3 explicates the architecture of the discriminator network, emphasizing the steps involved in processing the input image. Initially, the image is subjected to a series of residual convolution blocks, which are fortified with spectral normalization. This is followed by the execution of an average pooling operation. The process culminates with the derivation of the output probability, which is obtained post the final linear layer, again, involving spectral normalization.

4. Additional Qualitative Results

Figure 4, 5, and 6 deliver additional examples on CelebA-HQ [2], AFHQ [1], and CompCar [3] datasets.

We embark on rigorous evaluation of our model using a diverse range of input images sourced from varied datasets. With the CelebA dataset, we assess the model's performance using faces of different genders, ages, and ethnic backgrounds, all of which yield impressive quality in output. In the context of the AFHQ dataset, we utilize images from a variety of categories as input for our testing phase. It is worth noting that these results, encompassing distinct

*Corresponding author.

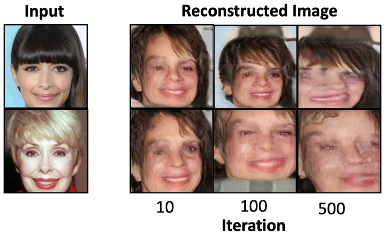


Figure 1. Qualitative experiments of optimization based method.

Method	Cats		CelebA(HQ)		CompCar		AFHQ	
	128^2	256^2	128^2	256^2	128^2	256^2	128^2	256^2
Optimization	0.46	0.43	0.34	0.40	0.56	0.53	0.48	0.48
Learning(ours)	**0.22**	**0.24**	**0.23**	**0.25**	**0.40**	**0.43**	**0.23**	**0.26**

Table 1. Input preservation comparison to optimization-based with LPIPS(↓) [4]. Optimization and Learning denote optimization-based GAN inversion and learning-based GAN inversion, respectively

categories, are obtained using a single model with different conditional vector inputs, thereby highlighting the large capacity of our model.

The CompCars dataset allows us to experiment with 360-degree image generation using real image inputs representing various car models, colours, and camera poses. It is important to note that a significant advantage of our model is the freedom it provides in the longitudinal movement of objects, along with the capacity to alter the background. This flexibility underpins the model's capacity for highly controllable image synthesis, an attribute that holds immense potential for a wide array of applications.

Figure 7 and 8 exhibit additional style-mixed 3D image synthesis examples on CelebA-HQ and AFHQ datasets. The latent vectors of the style-mixed images are obtained from two distinct source image. The results signify that our inverter successfully disentangle features of source images.

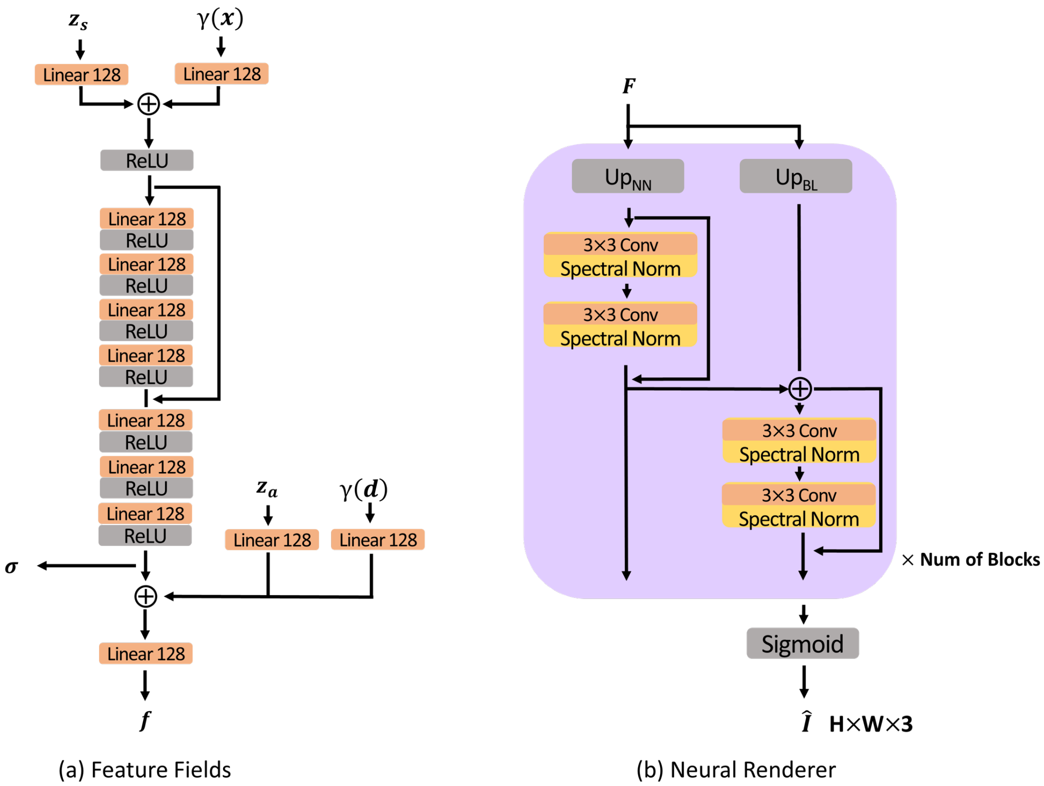
References

[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.

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[3] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3973–3981, 2015.

[4] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.



(a) Feature Fields

(b) Neural Renderer

Figure 2. Architecture of the feature fields and neural renderer. The feature fields are parameterized with multi-layer perceptrons (MLPs) as shown in the (a). The 3D point **x**, viewing direction **d**, and latent codes $\mathbf{z_s}$, $\mathbf{z_a}$ are mapped into a volume density σ and feature **f**. In (b), the neural renderer blocks depict the transformation of the volume-rendered feature image F into final synthesized image $\hat{I}$. UP_{NN} and UP_{BL} symbolize the nearest neighbour upsampling and bilinear upsampling, respectively.

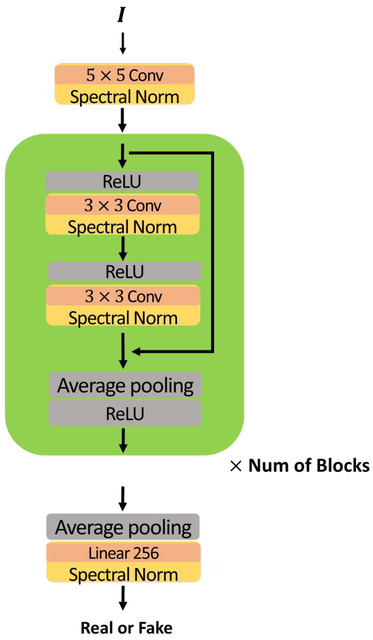


Figure 3. Architecture of the discriminator. The input image is processed through residual convolution blocks fortified with spectral normalization, and an average pooling operation. The output probability is derived after the final linear layer with spectral normalization.

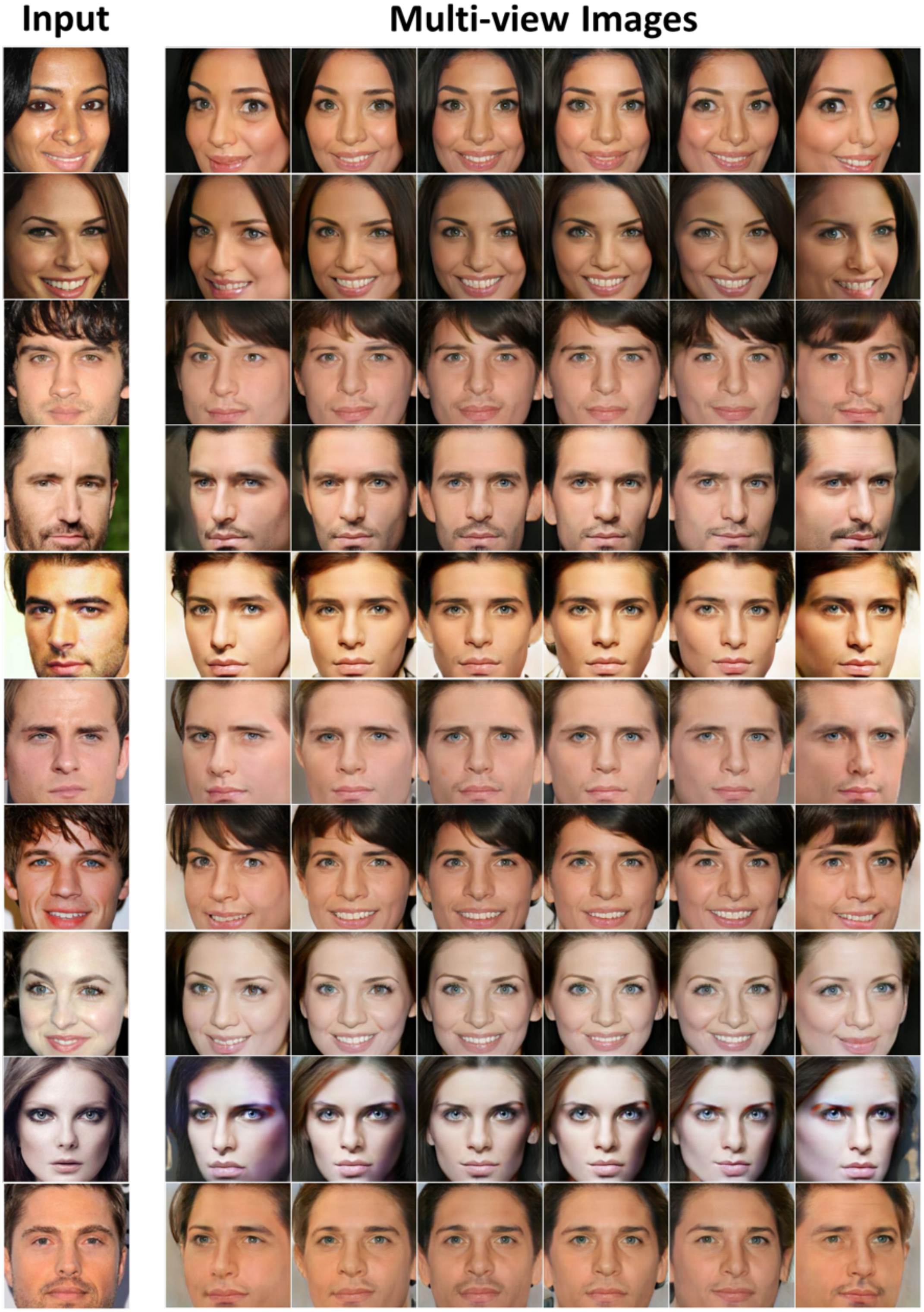


Figure 4. Supplementary results with $\mathbf{256^2}$ CelebA-HQ image inputs.

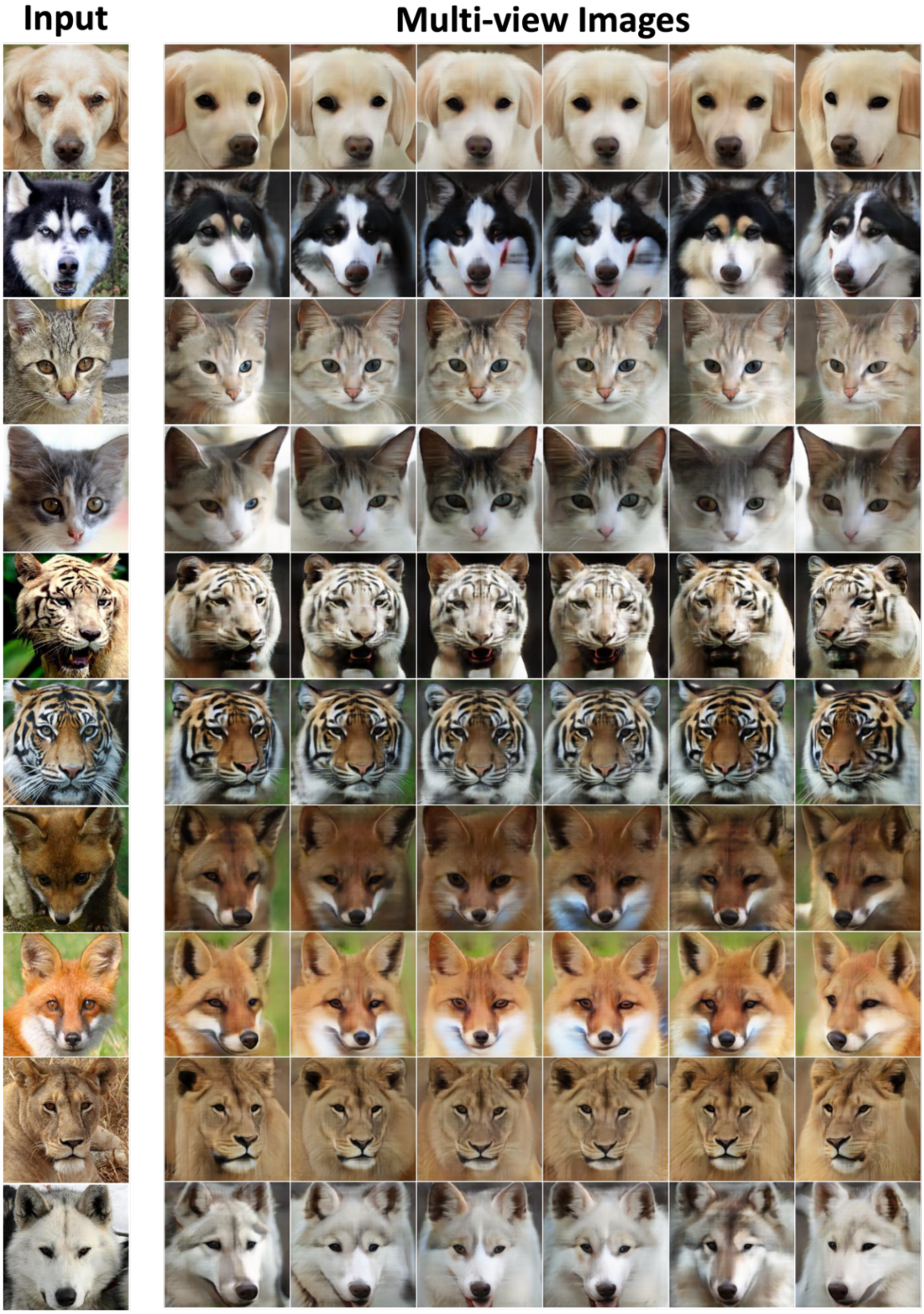


Figure 5. Supplementary results with $\mathbf{256^2}$ AFHQ image inputs.

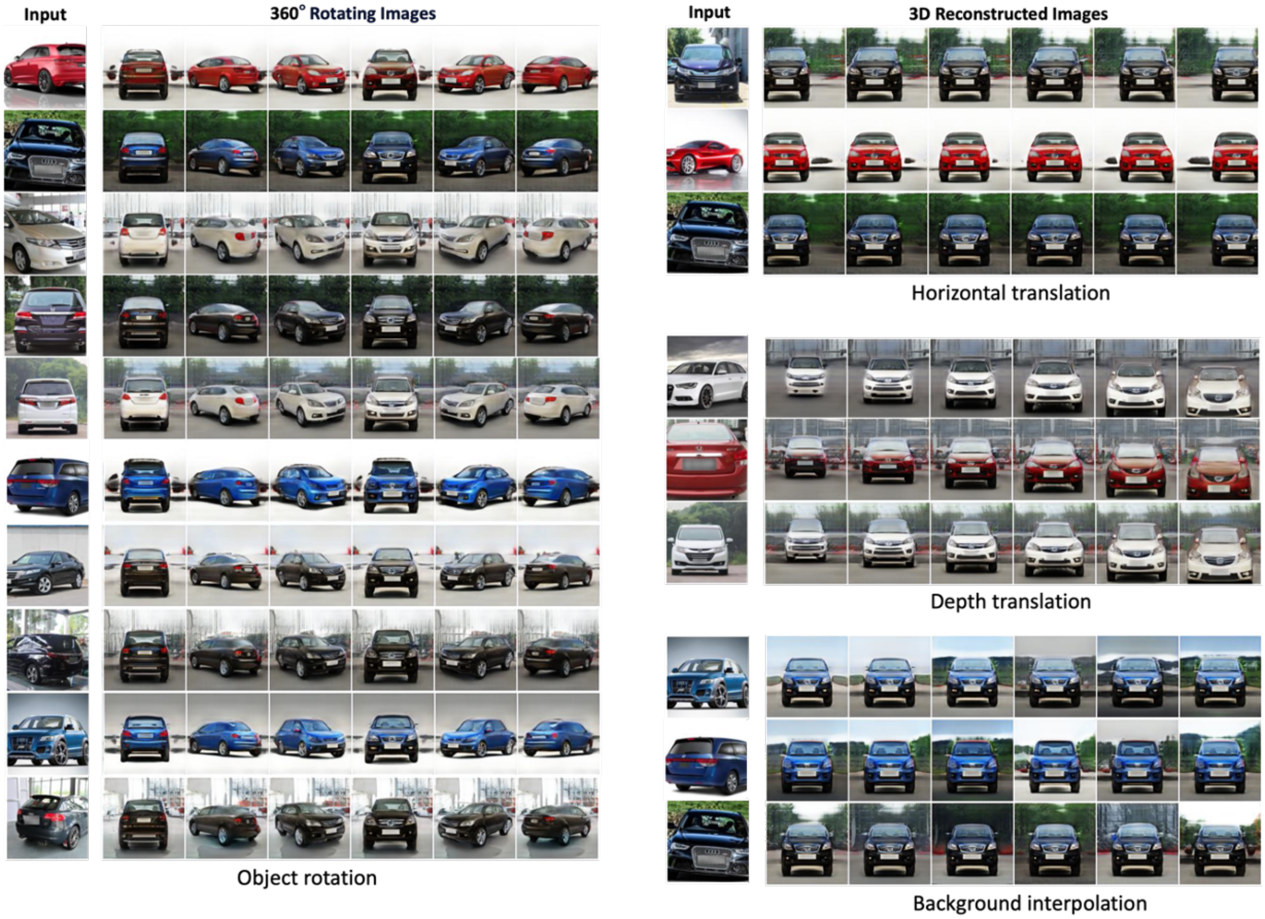


Figure 6. Controllable image synthesis with 128^2 CompCars image inputs.

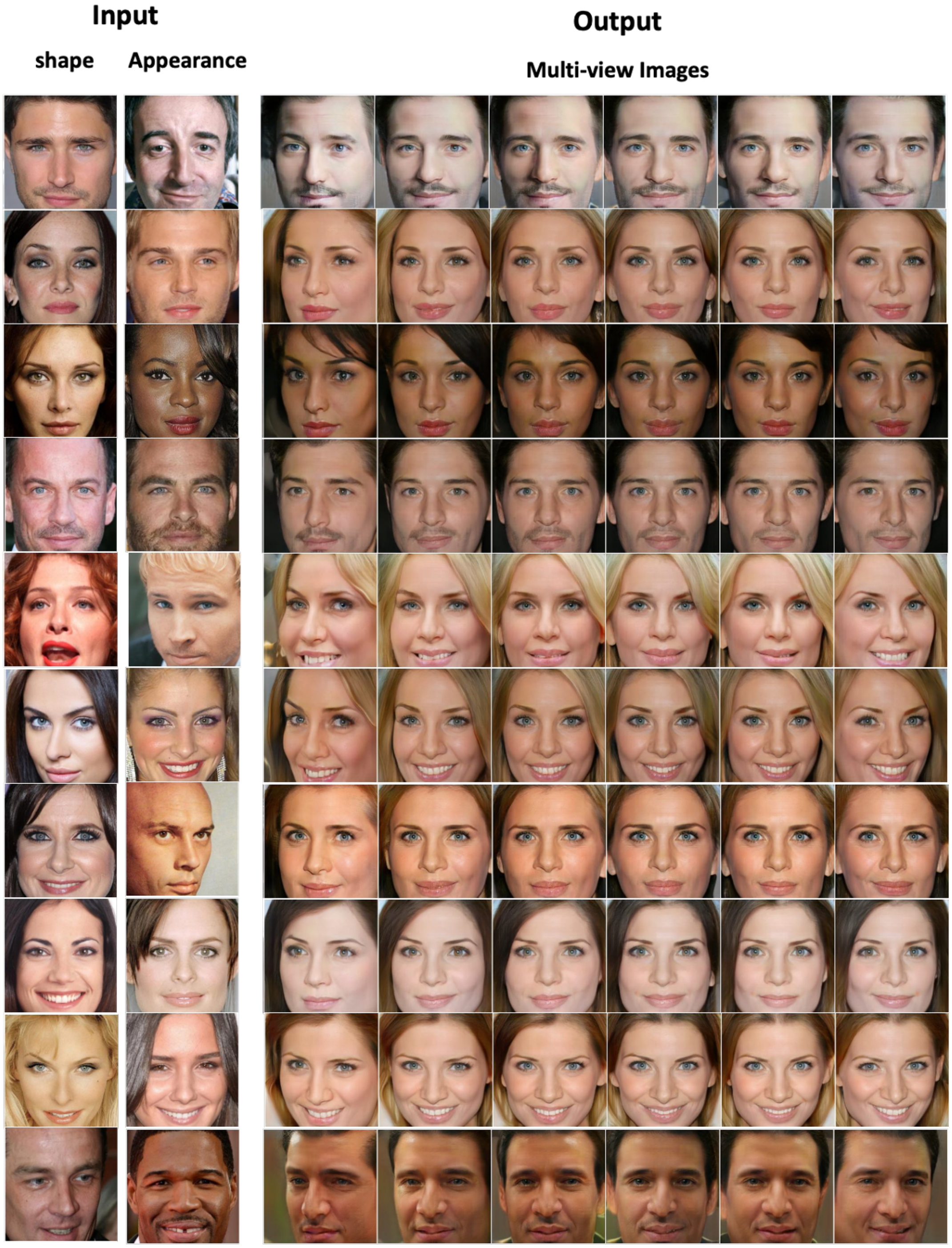


Figure 7. Multi-view images with style mixing of two CelebA-HQ input images.

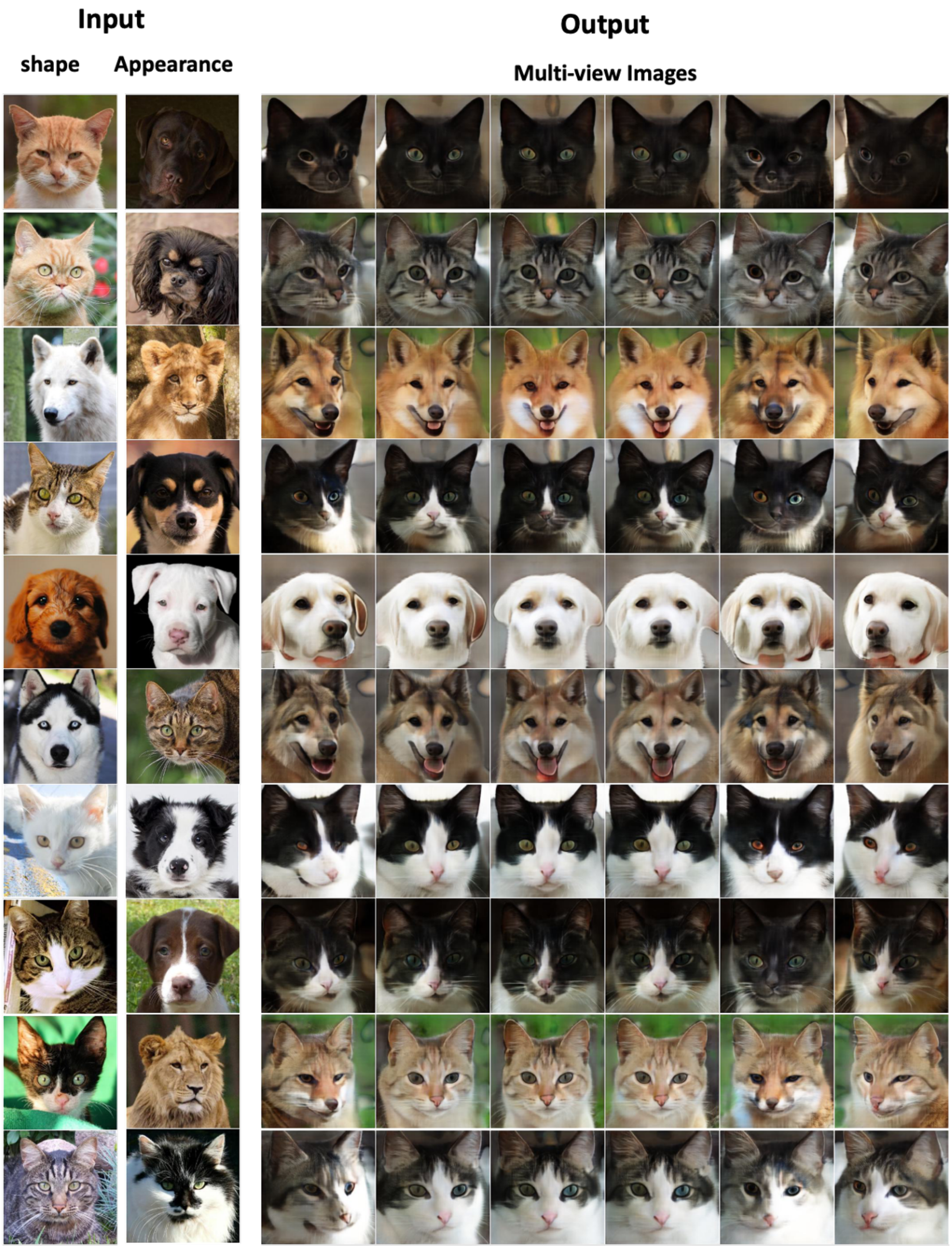


Figure 8. Multi-view images with style mixing of two AFHQ input images.