Supplementary Material
Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation

1. Implementation Details

For our backbone network we use the ResNet-IR architecture from [2] pretrained on face recognition, which accelerated convergence. We use a *fixed* StyleGAN2 generator trained on the FFHQ [5] dataset. That is, only the pSp encoder network is trained on the given translation task. For all applications, the input image resolution is 256×256, where the generated 1024×1024 output is resized before being fed into the loss functions. Specifically for $\mathcal{L}_{\text{ID}}$, the images are cropped around the face region and resized to 112×122 before being fed into the recognition network. For training, we use the Ranger optimizer, a combination of Rectified Adam [7] with the Lookahead technique [11], with a constant learning rate of 0.001. Only horizontal flips are used as augmentations. All experiments are performed using a single NVIDIA Tesla P40 GPU.

For the StyleGAN inversion task, the λ values are set as $\lambda_1 = 1$, $\lambda_2 = 0.8$, and $\lambda_3 = 0.1$. For face frontalization, we increase the weight of the $\mathcal{L}_{\text{ID}}$, setting $\lambda_3 = 1$ and decrease the $\mathcal{L}_2$ and $\mathcal{L}_{\text{LPIPS}}$ loss functions, setting $\lambda_1 = 0.01$, $\lambda_2 = 0.8$ over the inner part of the face and $\lambda_1 = 0.001$, $\lambda_2 = 0.08$ elsewhere. Additionally, the constants used in the conditional image synthesis tasks are identical to those used in the inversion task except for the omission of the identity loss (i.e. $\lambda_3 = 0$). Finally, λ_4 is set to 0.005 in all applications except for the StyleGAN inversion task, which does not utilize the regularization loss.

2. Dataset Details

We conduct our experiments on the CelebA-HQ dataset [4], which contains 30,000 high-quality images. We use a standard train-test split of the dataset, resulting in approximately 24,000 training images. The FFHQ dataset from [5], which contains 70,000 face images, is used for the StyleGAN inversion and face frontalization tasks.

For the generation of real images from sketches, we construct a dataset representative of hand-drawn sketches using the CelebA-HQ dataset. Given an input image, we first apply a "pencil sketch" filter which retains most facial details of the original image while removing the remaining noise. We then apply the sketch-simplification method by [9], resulting in images resembling hand-drawn sketches. The same approach is also used for generating the sketch images on the AFHQ Cat and AFHQ Dog datasets [1].

3. Application Details

3.1. Super Resolution

In super resolution, the pSp framework is used to construct high-resolution (HR) images from corresponding low-resolution (LR) input images. PULSE [8] approaches this task in an unsupervised manner by traversing the HR image manifold in search of an image that downsamples to the input LR image.

Methodology and details. We train both our model and pix2pixHD [10] in a supervised fashion, where for each input we perform random bi-cubic down-sampling of $\times 1$ (i.e. no down-sampling), $\times 2$, $\times 4$, $\times 8$, $\times 16$, or $\times 32$ and set the original, full resolution image as the target.

Results. Figure 1 demonstrates the visual quality of the resulting images from our method along with those of the previous approaches. Although PULSE is able to achieve very high-quality results due to their usage of StyleGAN to generate images, they are unable to accurately reconstruct the original image even when performing down-sampling of $\times 8$ to a resolution of 32×32. By learning a pixel-wise correspondence between the LR and HR images, pix2pixHD is able to obtain satisfying results even when down-sampled to a resolution of 16×16 (i.e. $\times 16$ down-sampling). However, visually, their results appear less photo-realistic. Contrary to these previous works, we are able to obtain high-quality results even when down-sampling to resolutions of 16×16 and 8×8. Finally, in Figure 1d we generate multiple outputs for a given LR image using our multi-modal technique by performing style-mixing with a randomly sampled $\mathbf{w}$ vector on layers (4-7) with an α value of 0.5. Doing so alters medium-level styles that mainly control facial features.

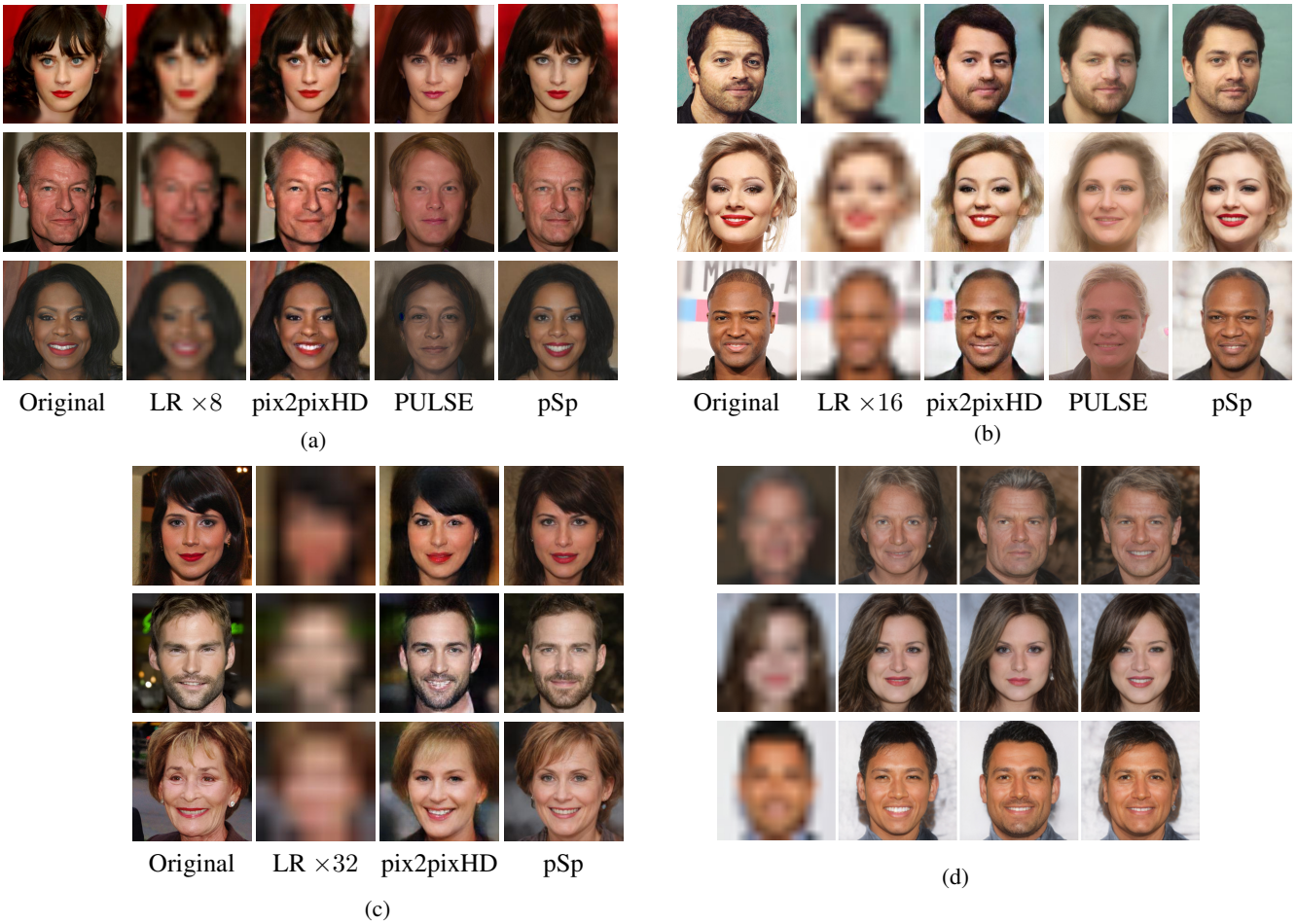


Figure 1: Comparison of super-resolution approaches with (a) $\times 8$ down-sampling, (b) $\times 16$ down-sampling, and (c) $\times 32$ down-sampling on the CelebA-HQ [4] test set. (d) Multi-modal synthesis for super-resolution using pSp with style-mixing.

3.2. Inpainting

In the task of inpainting we wish to reconstruct missing or occluded regions in a given image. Due to their local nature, pix2pix [3] and other local-based translation methods, have shown success in tackling this problem as they can simply propagate non-occluded regions.

Methodology and details We train both pSp and pix2pixHD [10] in a supervised fashion, where each input image is occluded with a symmetric triangular mask.

Results Figure 2 presents results for both our method and pix2pixHD. As shown, due to the lack of information in the occluded regions, pix2pixHD is unable to accurately reconstruct the original image and incurs many artifacts. In contrast, since pSp is trained to encode images into realistic face latents, it is able to accurately reconstruct the occluded region, resulting in high-quality outputs with no artifacts.

3.3. Local Editing

Our framework allows for a simple approach to local image editing using a trained pSp encoder where altering specific attributes of an input sketch (e.g. eyes, smile) or segmentation map (e.g. hair) results in local edits of the generated images. We can further extend this and perform local patch editing on real face images. As shown in Figure 3b, pSp is able to seamlessly merge the desired patch into the original image.

3.4. Face Interpolation

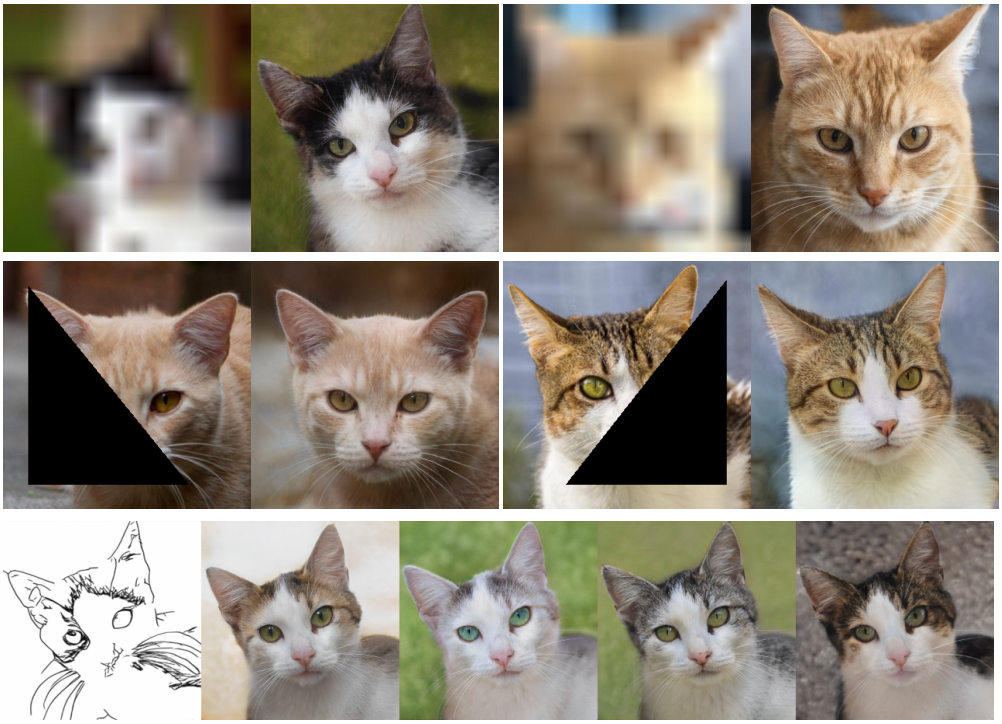
Given two real images one can obtain their respective latent codes $w_1, w_2 \in \mathcal{W}+$ by feeding the images through our encoder. We can then naturally interpolate between the two images by computing their intermediate latent code $w' = \alpha w_1 + (1-\alpha)w_2$ for $0 \leq \alpha \leq 1$ and generate the corresponding image using the new code w'.

Figure 2: Image inpainting results using pSp and pix2pixHD [10] on the CelebA-HQ [4] test set.

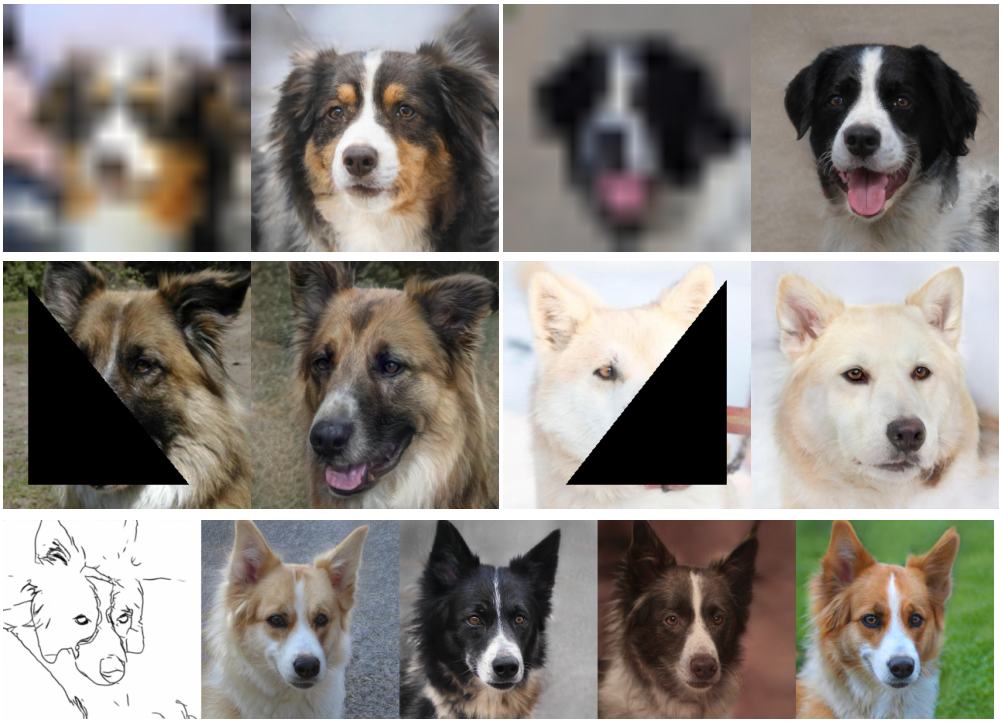
(a) (b)

Figure 3: Local patch editing results using pSp on sketches (a) and real face images (b).

Figure 4: Image interpolation results using pSp on the CelebA-HQ [4] test set.



(a)



(b)

Figure 5: Results of pSp on the AFHQ Cat and AFHQ Dog datasets [1] on super resolution, inpainting, and image generation from sketches.

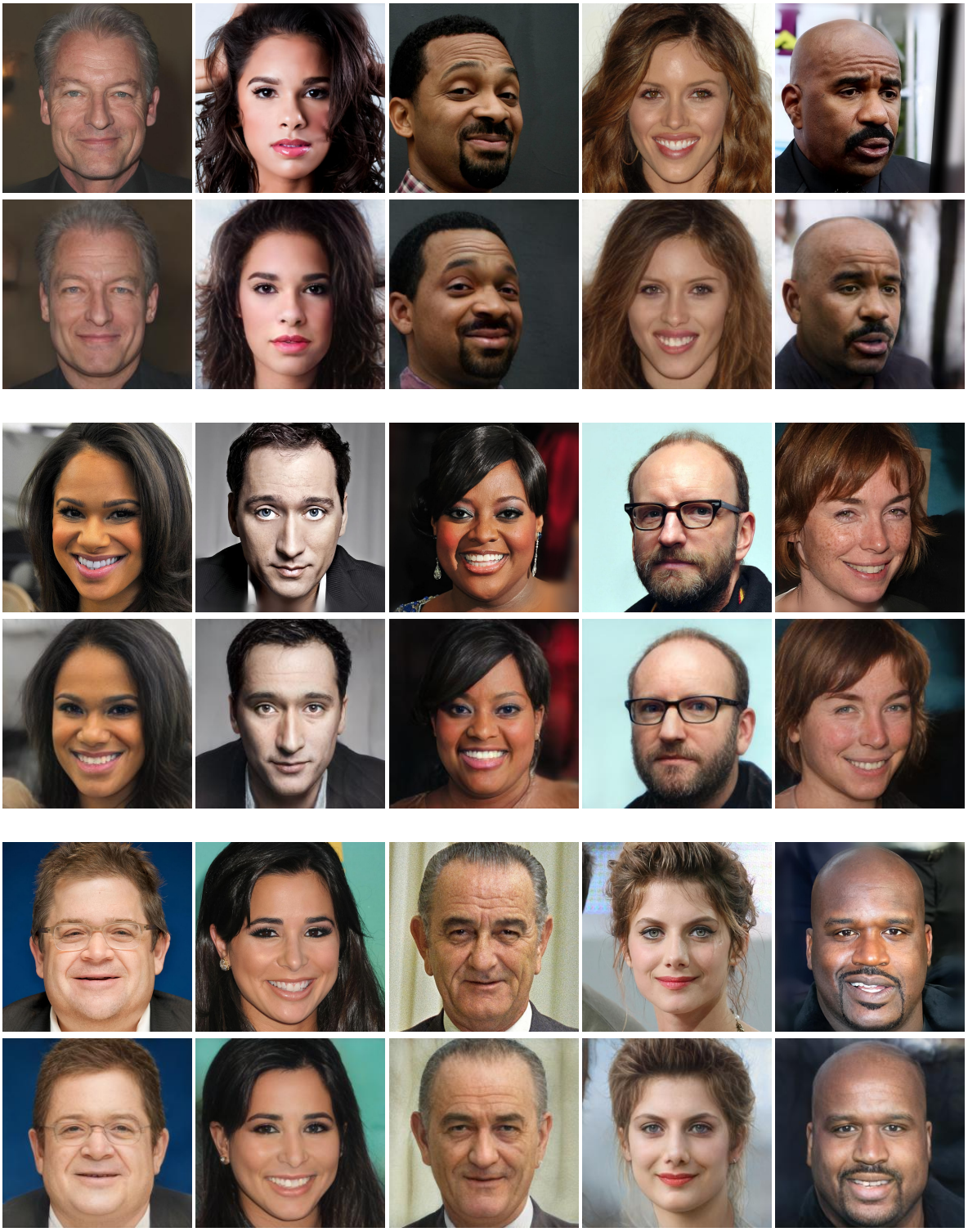


Figure 6: Additional StyleGAN inversion results using pSp on the CelebA-HQ [4] test set.

Figure 7: Additional face frontalization results using pSp on the CelebA-HQ [4] test set.

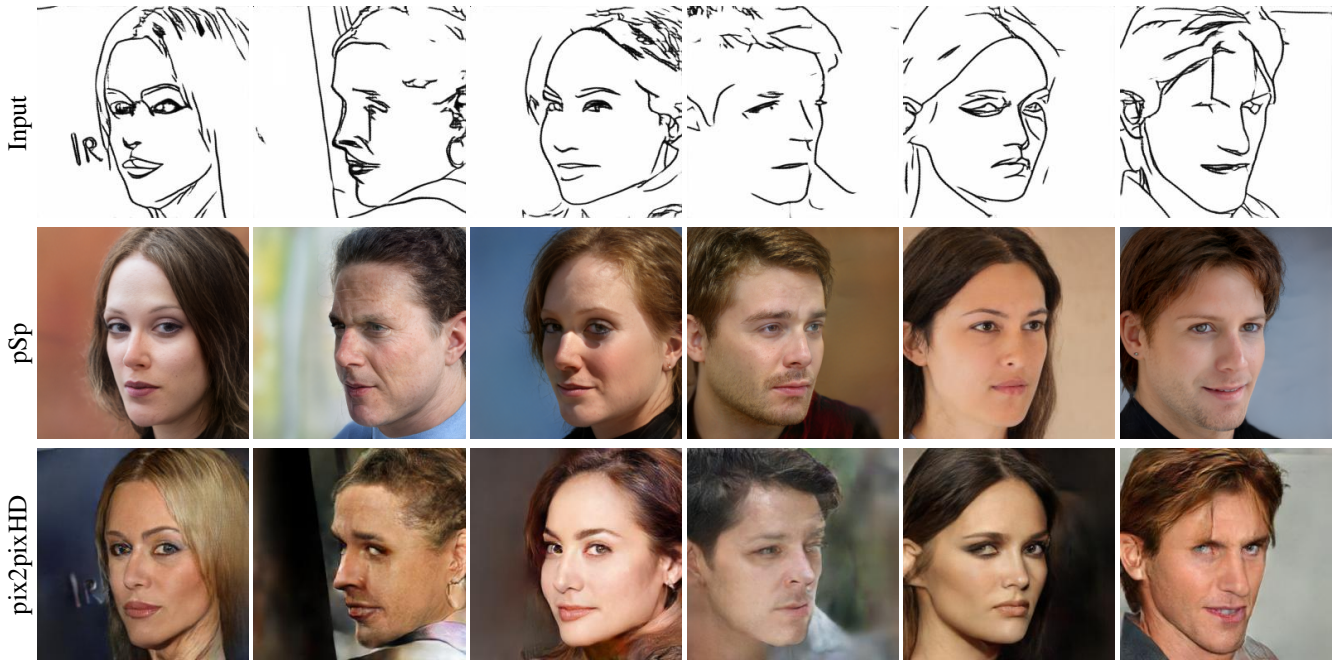


Figure 8: Even for challenging, non-frontal face sketches, pSp is able to obtain high-quality, diverse outputs.

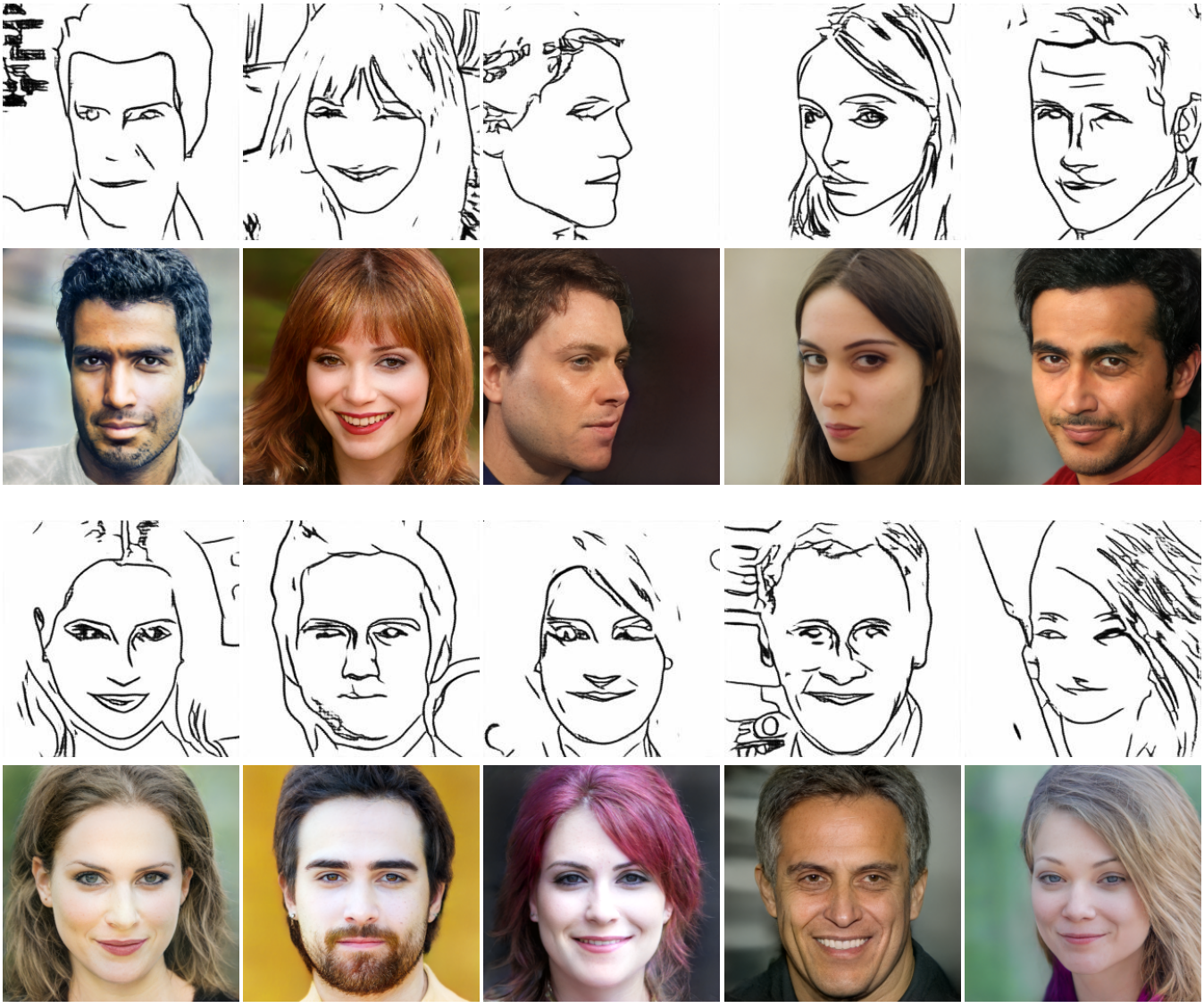


Figure 9: Additional results using pSp for the generation of face images from sketches on the CelebA-HQ [4] test dataset.

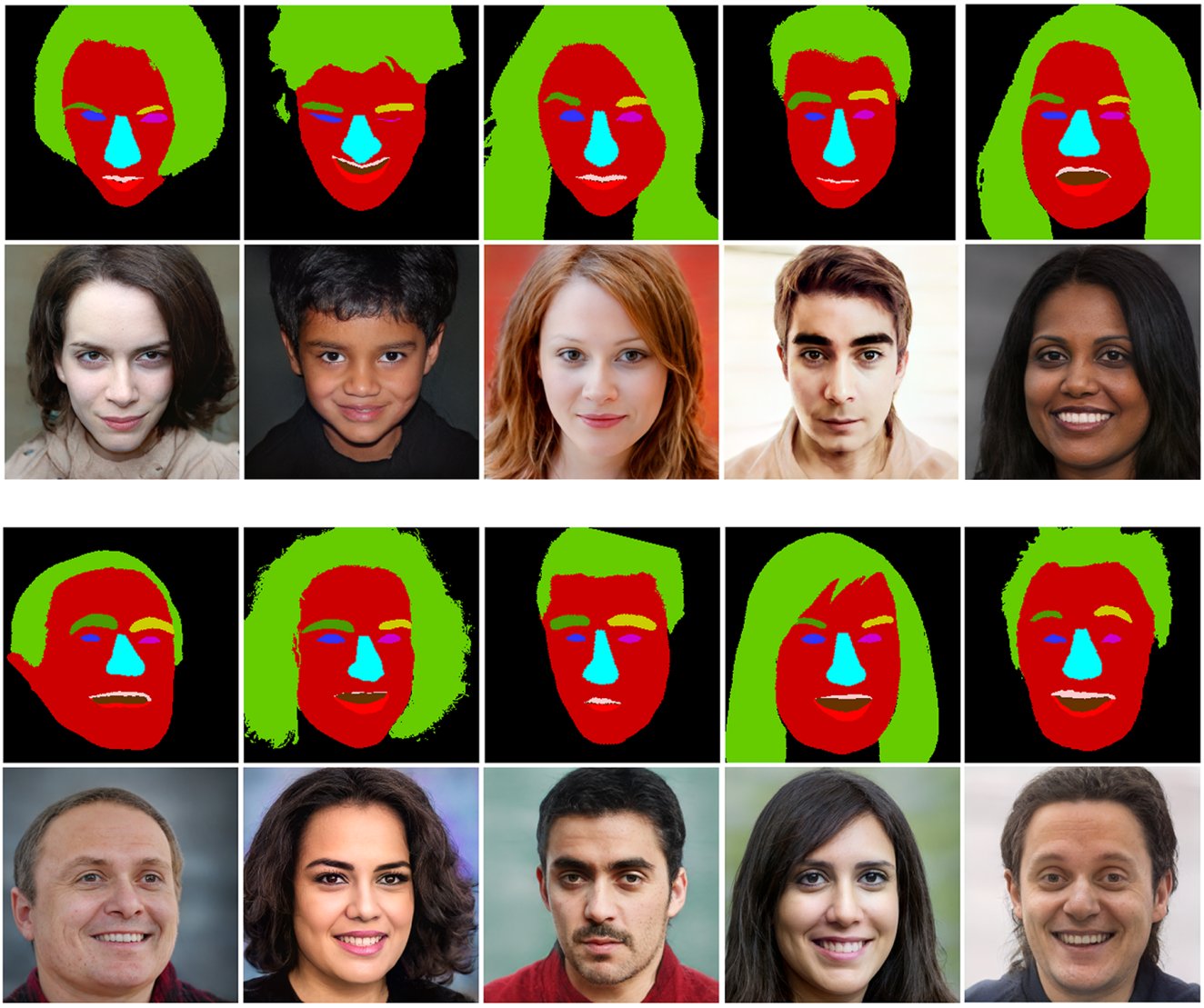


Figure 10: Additional results on the Helen Faces [6] dataset using our proposed segmentation-to-image method.

Figure 11: Additional results on the CelebAMask-HQ [4] test set using our proposed segmentation-to-image method.

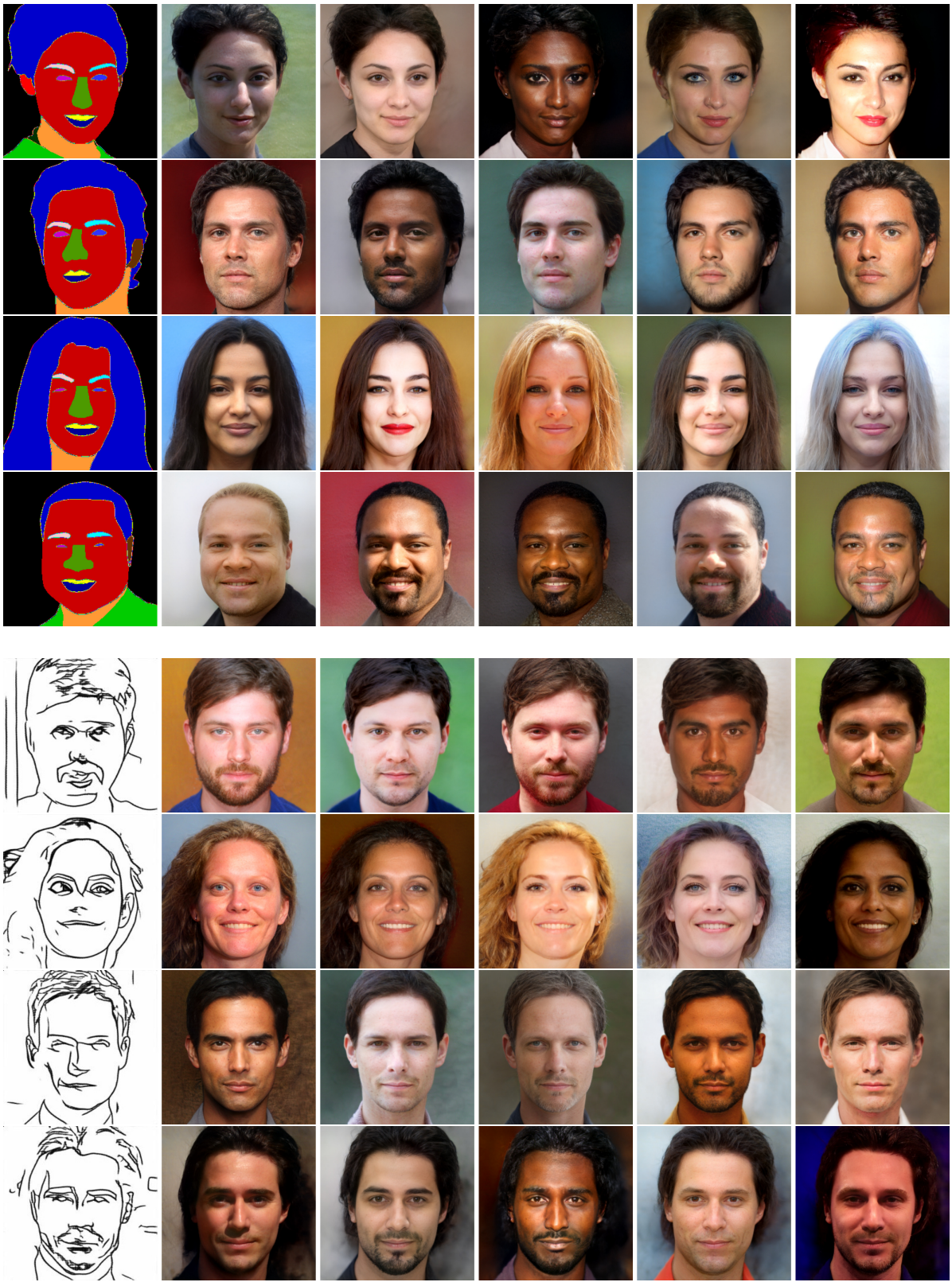


Figure 12: Conditional image synthesis results from sketches and segmentation maps displaying the multi-modal property of our approach.

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] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

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

[6] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S. Huang. Interactive facial feature localization. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 679–692, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[7] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

[8] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[9] Edgar Simo-Serra, Satoshi Iizuka, Kazuma Sasaki, and Hiroshi Ishikawa. Learning to simplify: fully convolutional networks for rough sketch cleanup. *ACM Transactions on Graphics*, 35:1–11, 07 2016.

[10] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[11] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, pages 9597–9608, 2019.