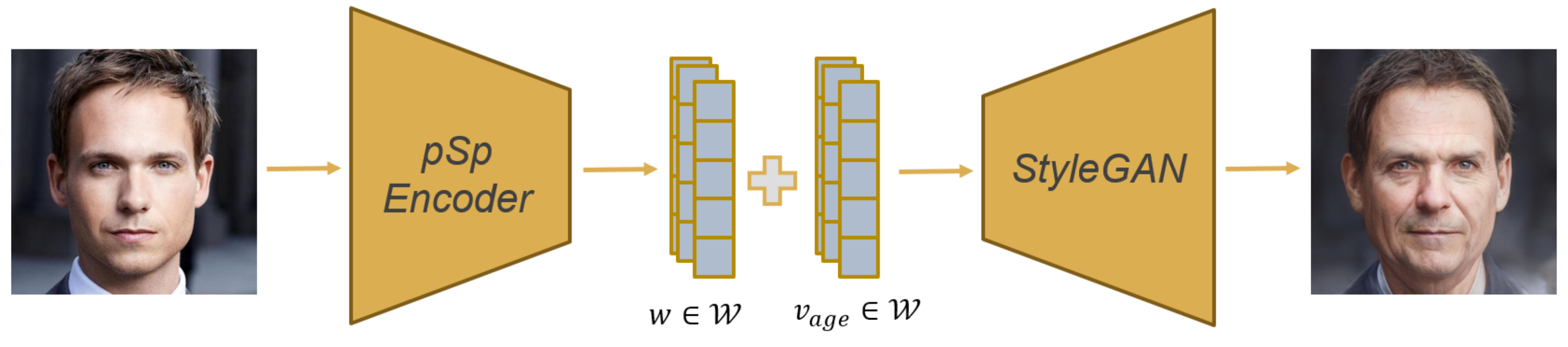
Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation

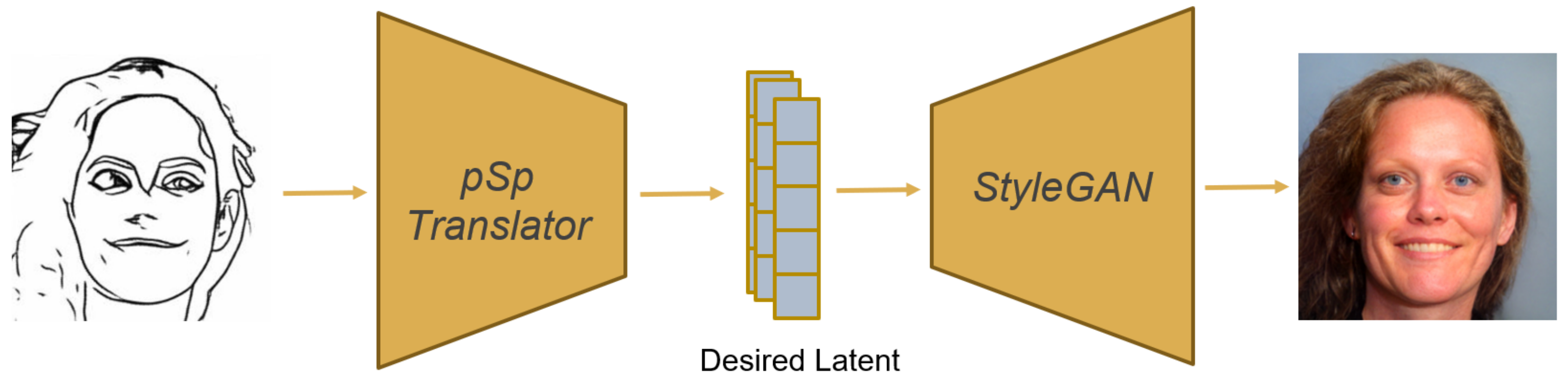
Elad Richardson[1] Yuval Alaluf[1,2] Or Patashnik[1,2] Yotam Nitzan[2] Yaniv Azar[1] Stav Shapiro[1] Daniel Cohen-Or[2]
[1]Penta-AI [2]Tel-Aviv University

Introducing pSp

The pixel2style2pixel (pSp) framework provides a fast and accurate solution for encoding images into the latent space of a pretrained StyleGAN. This encoding can then be used to easily manipulate and edit real images directly in the latent space.

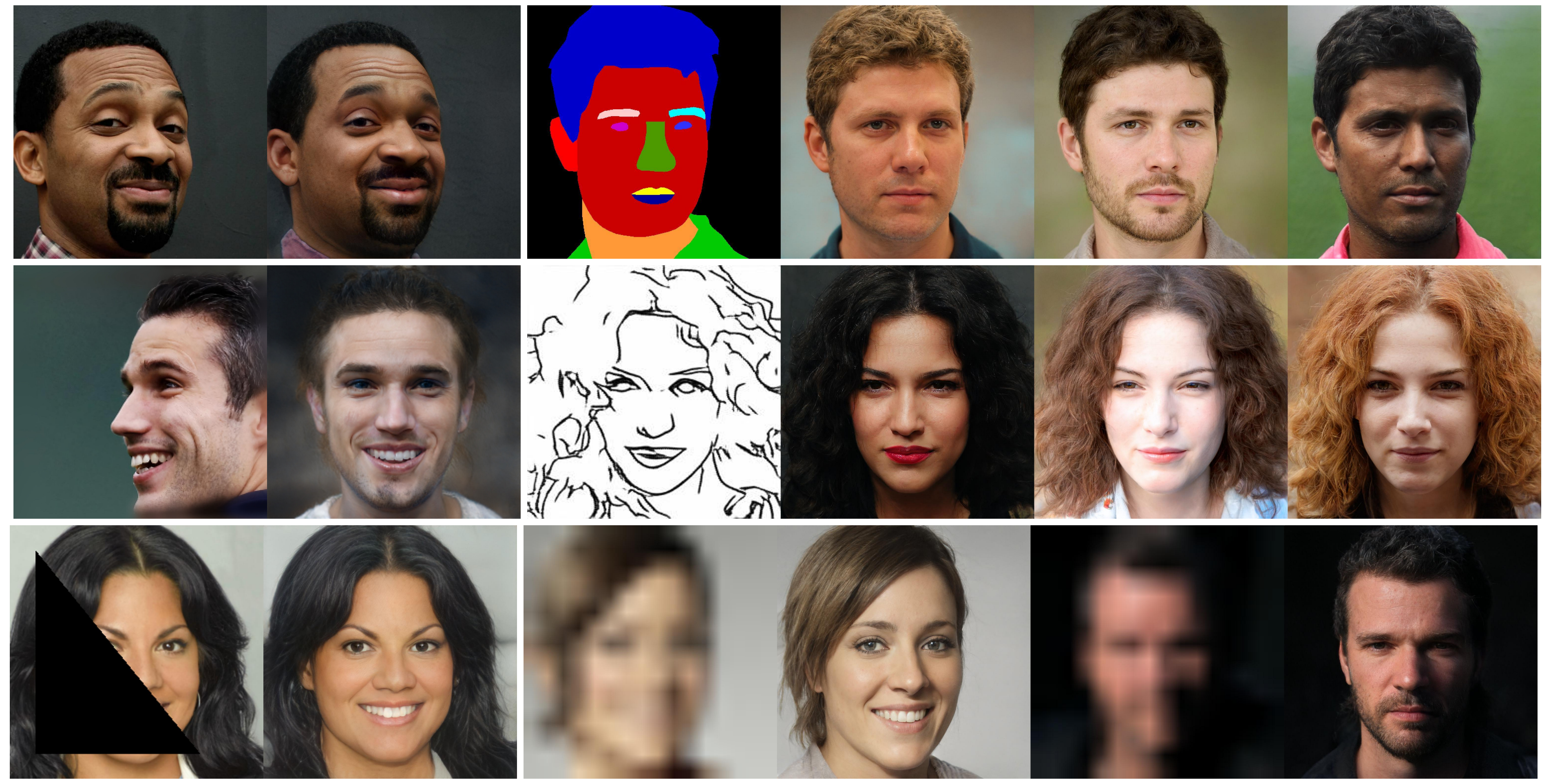


What makes pSp even more interesting is that it can be applied to more general image-to-image translation tasks by directly encoding the input image into the latent code corresponding to the desired output image. Using this technique one can perform image-to-image translation even when the input image cannot be encoded into the latent space of our pretrained StyleGAN.



What can it do?

The proposed pixel2style2pixel framework can be used to solve a wide variety of image-to-image translation tasks, including StyleGAN inversion, multi-modal conditional image synthesis, face frontalization, inpainting and super-resolution.

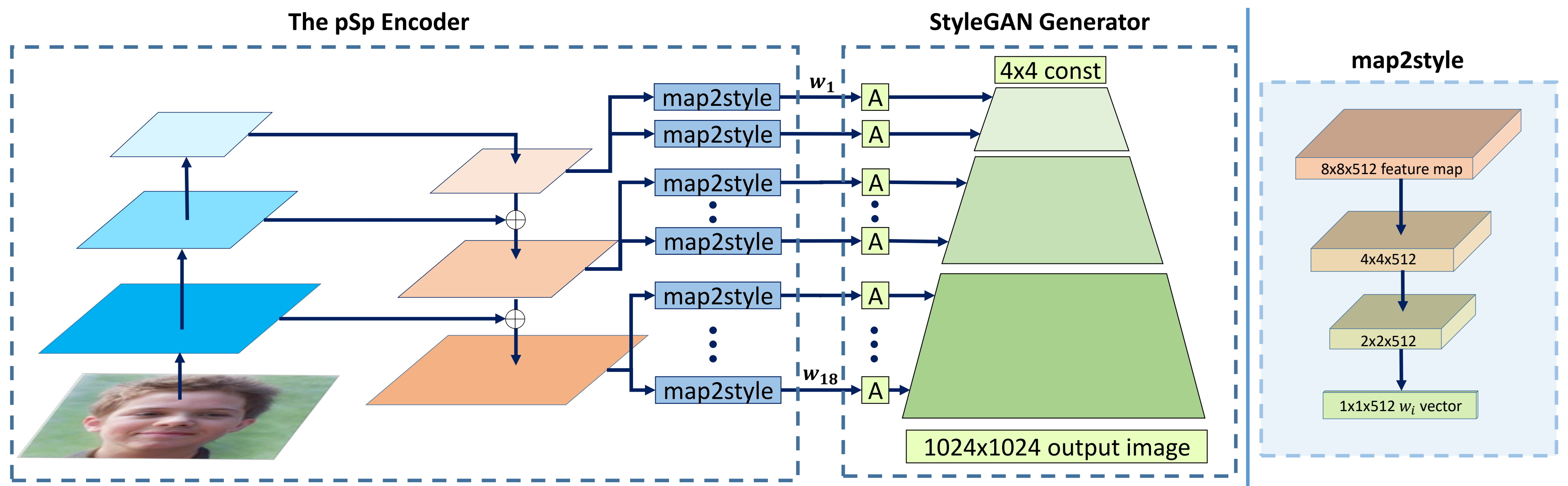


The benefits of StyleGAN

Using a pretrained StyleGAN and performing the translation between images through the *style* domain differentiates pSp from many standard image-to-image translation frameworks with several benefits:

- Simplification of the training process, as no adversary discriminator needs to be trained.
- Better results for non-local translations, as the generator is governed only by the styles with no direct spatial input.
- Inherent support for multi-modal synthesis for ambiguous tasks such as image generation from sketches or super-resolution, thanks to the ability to resample styles.

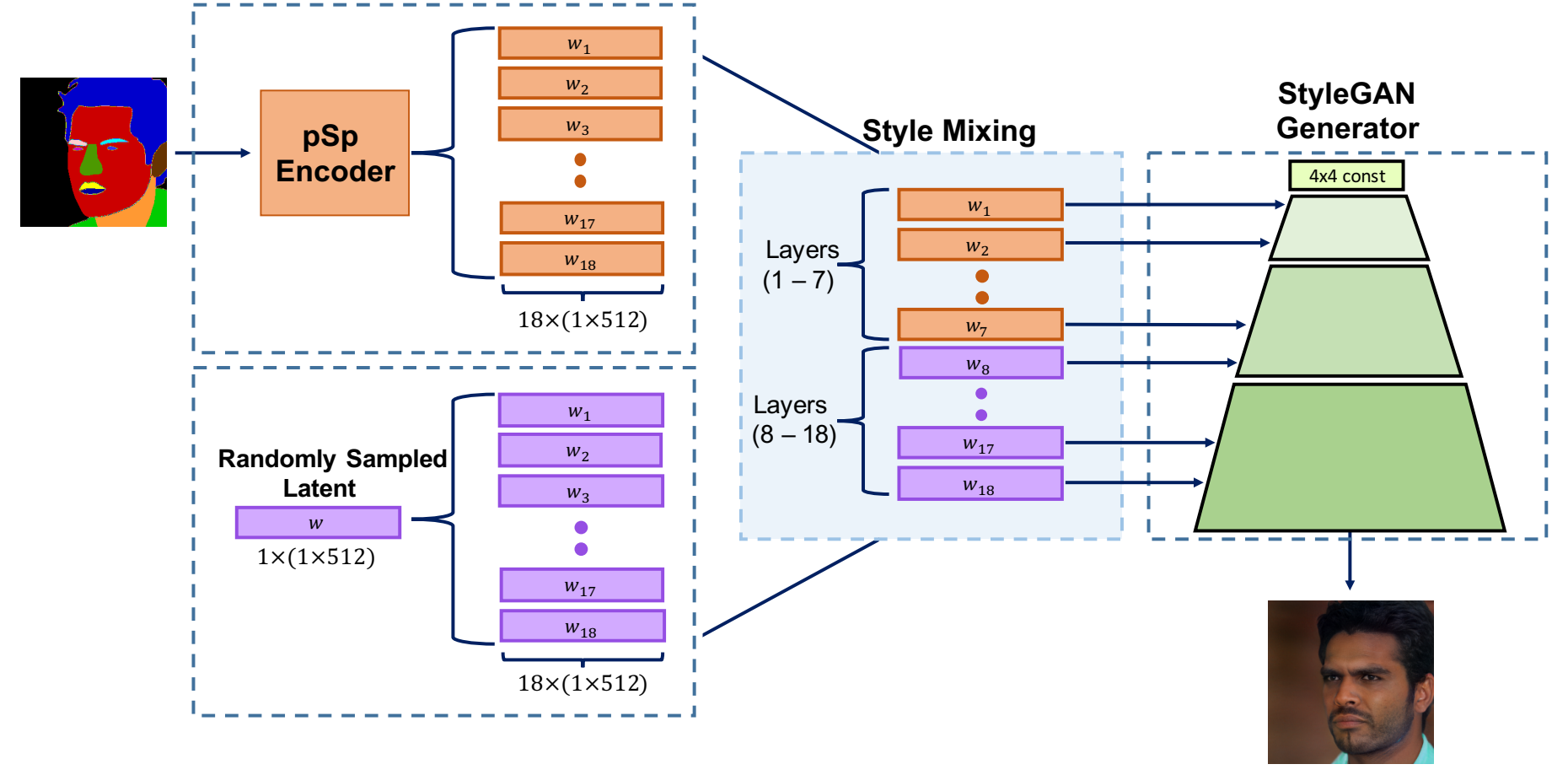
The Architecture



Starting from an input image, the pSp architecture outputs the corresponding output image:

- Feature maps are first extracted using a standard feature pyramid over a ResNet backbone.
- For each of the 18 target styles, a small mapping network is trained to extract the learned styles from the corresponding feature map. The mapping network, *map2style*, is a small fully convolutional network, which gradually reduces spatial size using a set of 2-strided convolutions followed by LeakyReLU activations.
- The generated 512 vectors are fed into a pretrained StyleGAN which then generates the output image.

Style Mixing



Our model supports multi-modal image generation by resampling styles and mixing them with the original encoding.

The Losses

$$\mathcal{L}_2\left(\mathbf{x}\right) = ||\mathbf{x} - pSp(\mathbf{x})||_2. \tag{1}$$

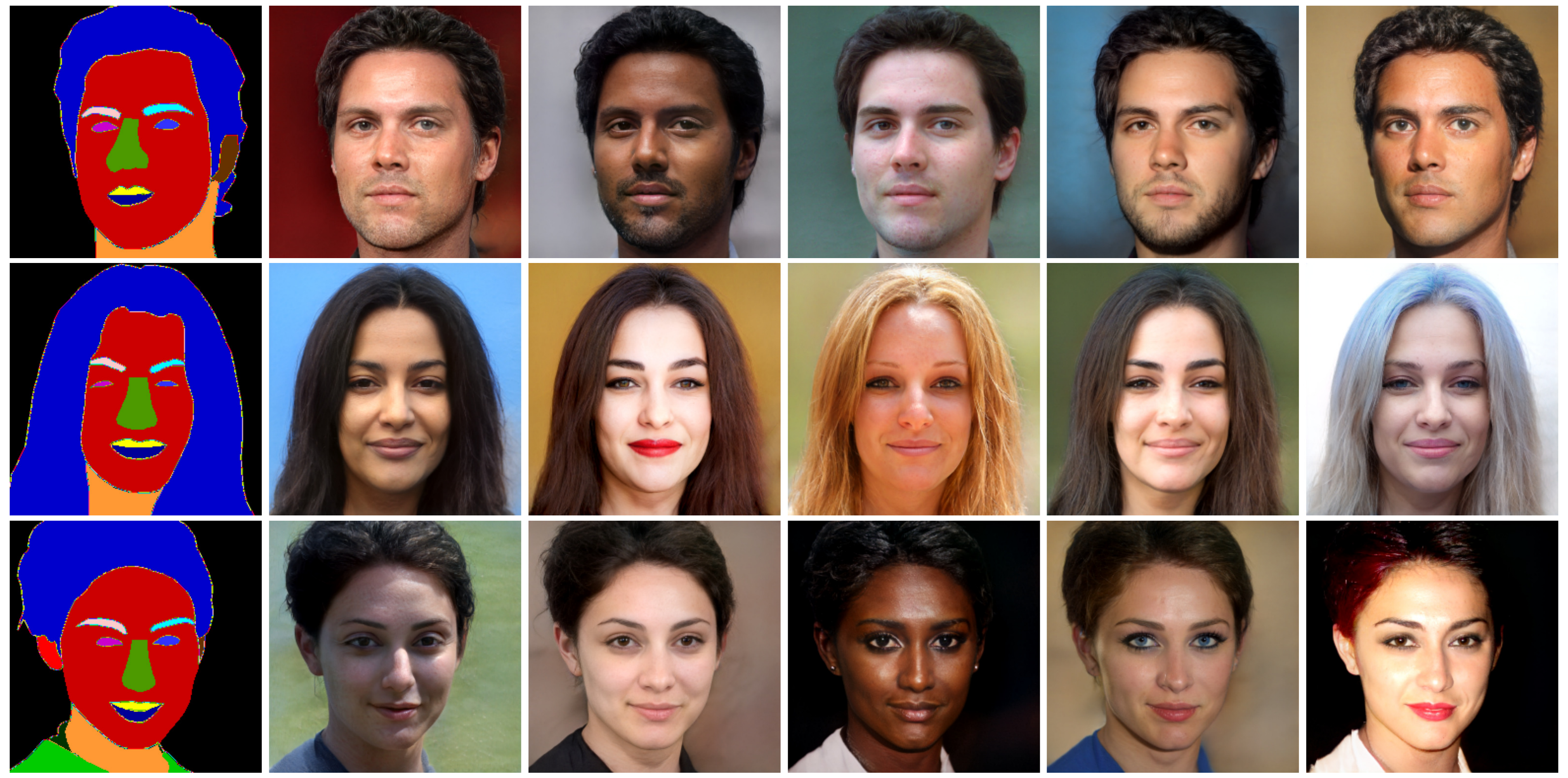
$$\mathcal{L}_{\text{LPIPS}}\left(\mathbf{x}\right) = ||F(\mathbf{x}) - F(pSp(\mathbf{x}))||_2, \tag{2}$$

$$\mathcal{L}_{\text{reg}}\left(\mathbf{x}\right) = ||E(\mathbf{x}) - \overline{\mathbf{w}}||_2. \tag{3}$$

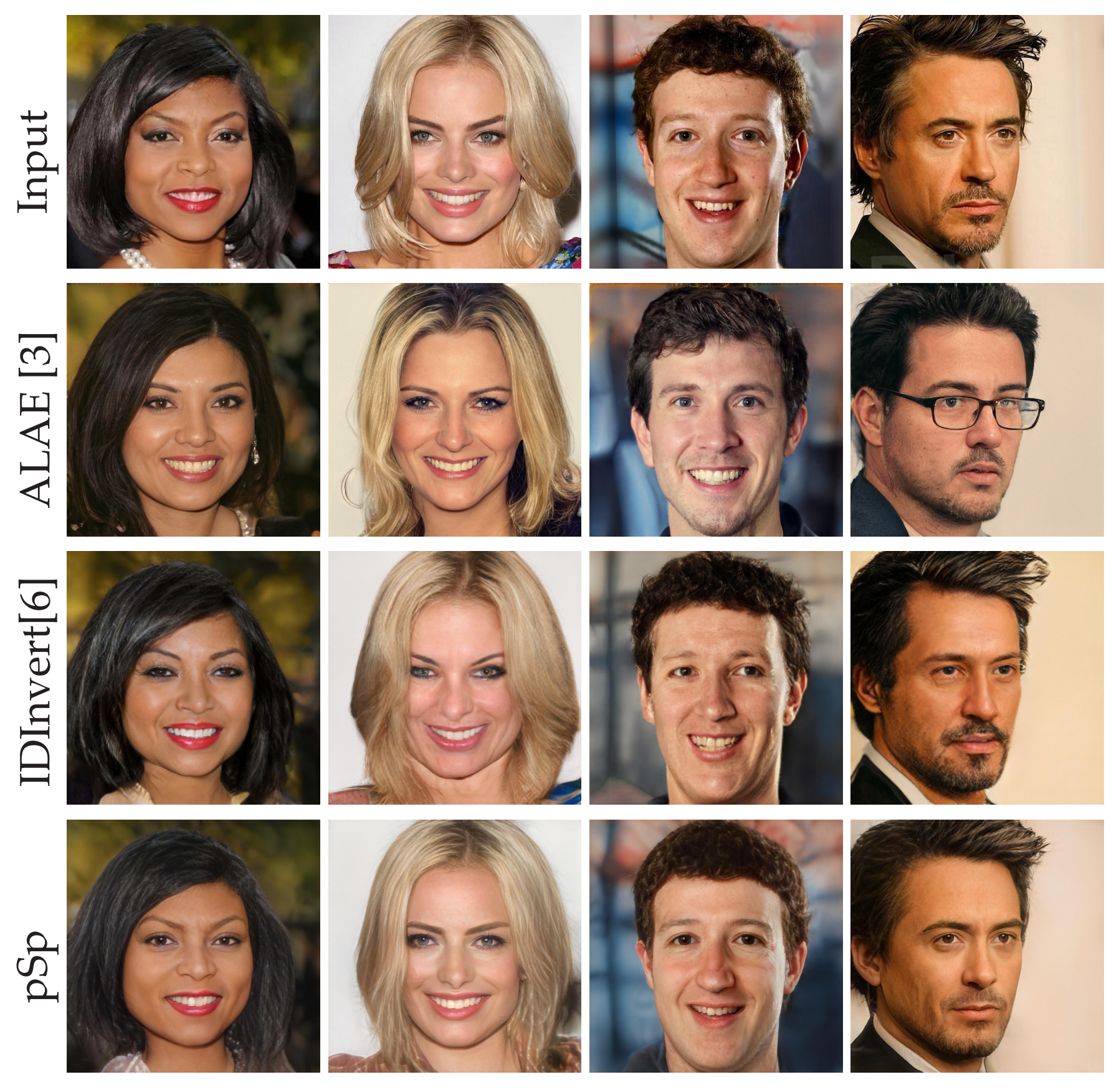
$$\mathcal{L}_{\text{ID}}\left(\mathbf{x}\right) = 1 - \langle R(\mathbf{x}), R(pSp(\mathbf{x}))\rangle\,, \tag{4}$$

A curated set of losses allows pSp to learn accurate encodings. $\mathcal{L}_{\text{LPIPS}}$ adds perceptual similarity on top of the pixel-wise $\mathcal{L}_2$ loss. We found $\mathcal{L}_{\text{ID}}$ to be important for preserving identity in facial images. $\mathcal{L}_{\text{reg}}$ serves as a regularization for training, similarly to the truncation trick in StyleGAN.

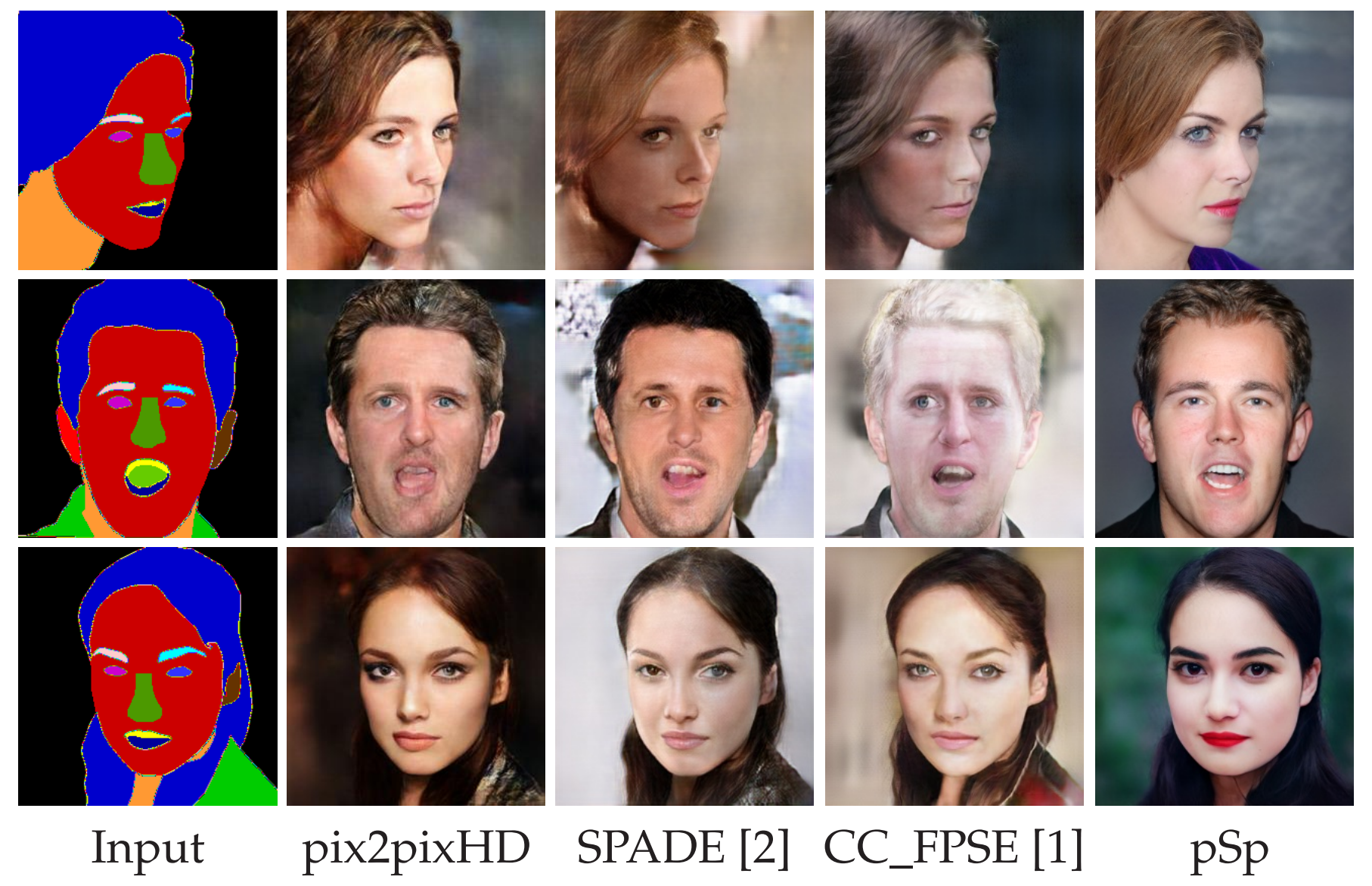
Multi-Modality Results



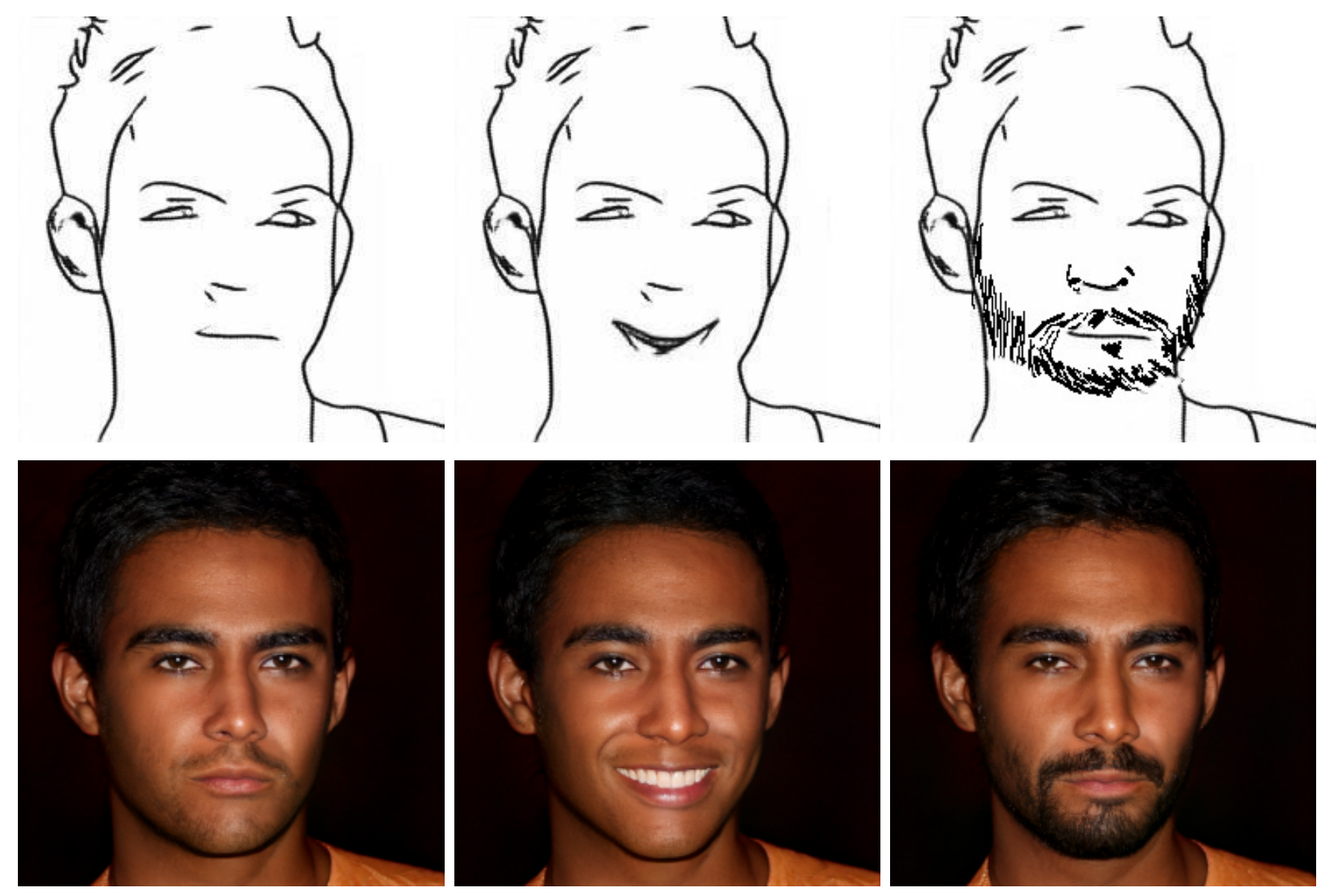
Encoding Results



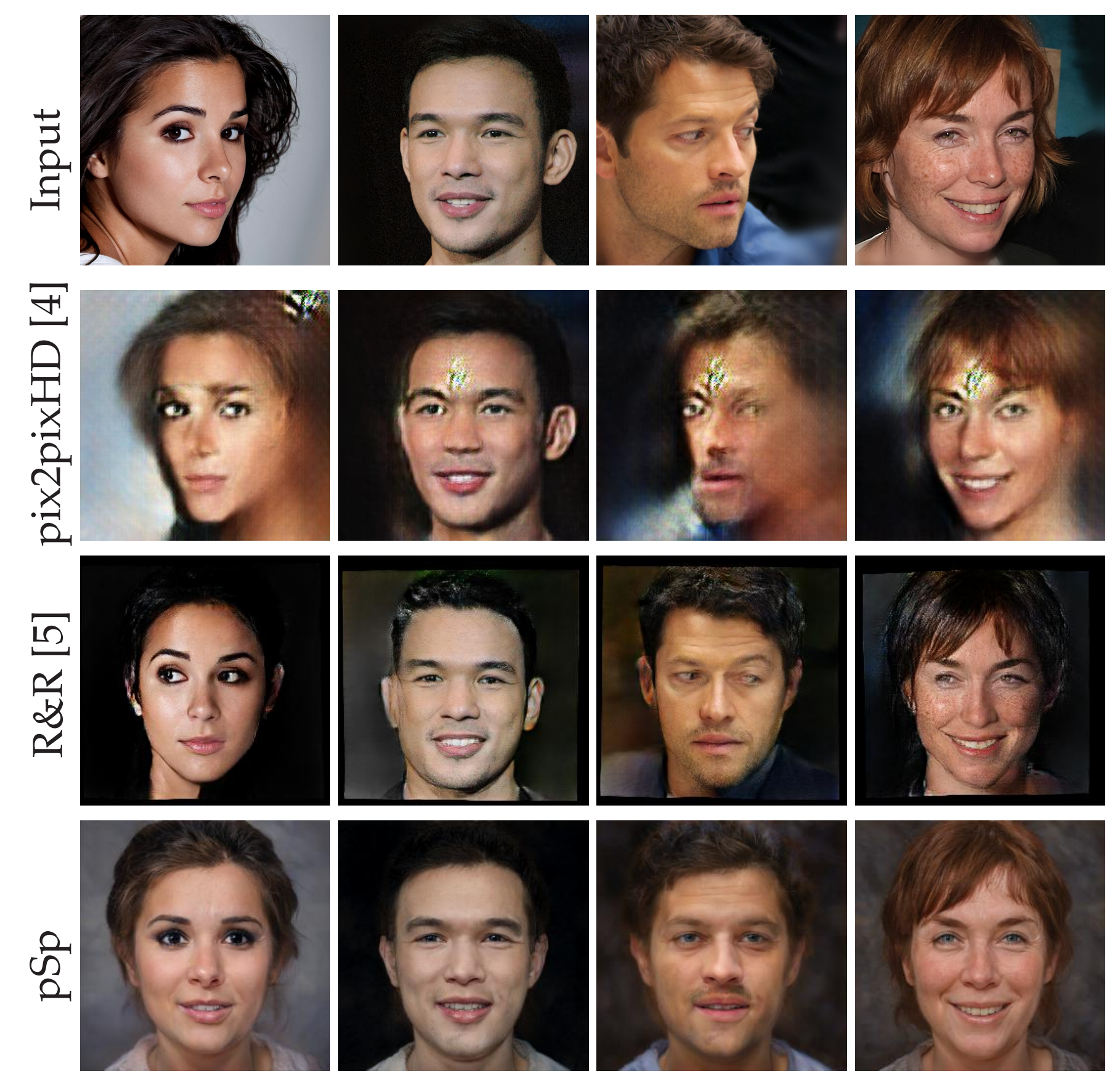
Conditional Image Generation



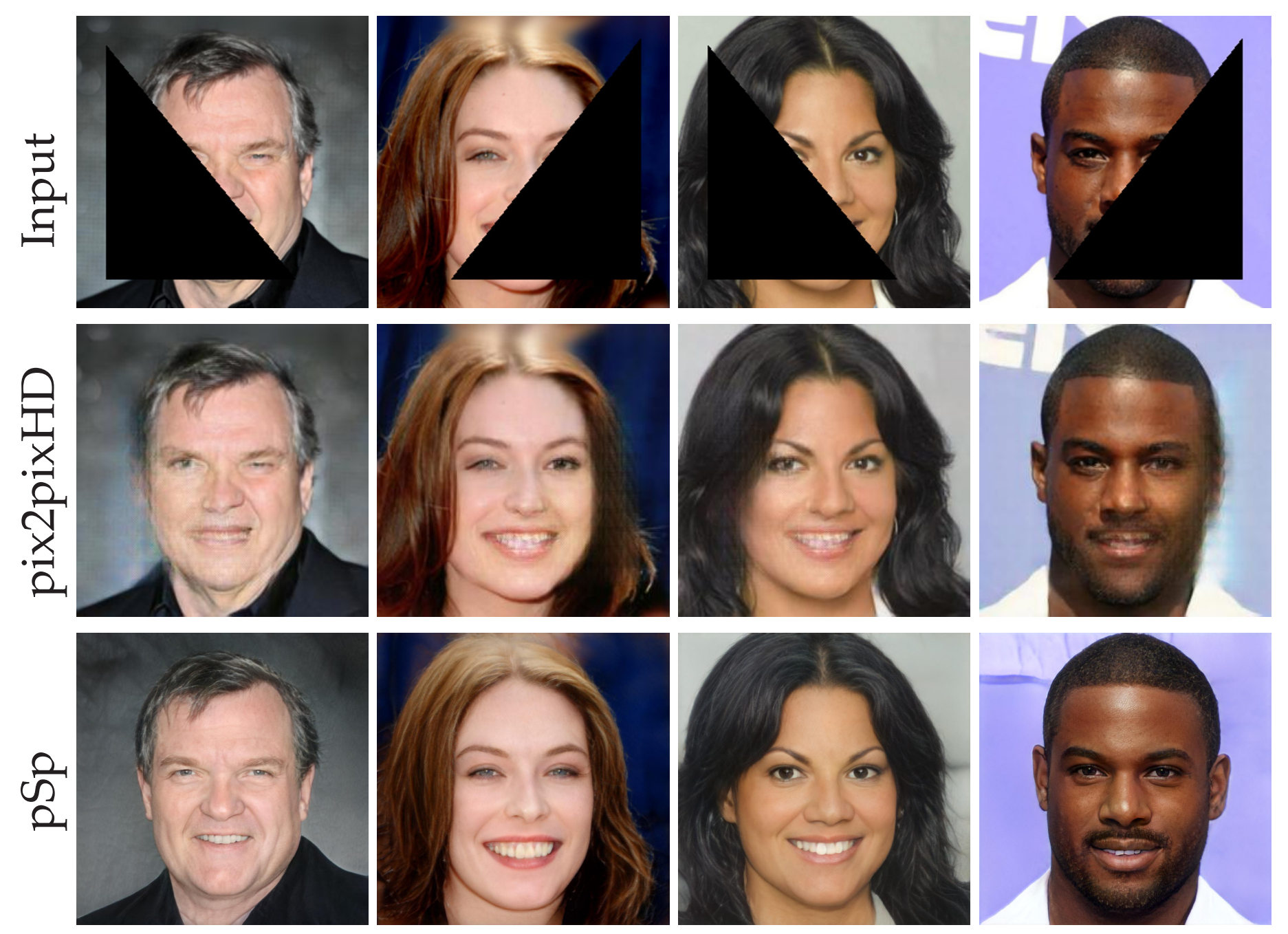
Local Editing



Frontalization Results



Inpainting Results



Additional Domains

References

[1] X. Liu, G. Yin, J. Shao, X. Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Advances in Neural Information Processing Systems*, pages 570–580, 2019.

[2] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[3] S. Pidhorskyi, D. A. Adjeroh, and G. Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14104–14113, 2020.

[4] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. 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.

[5] H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5911–5920, 2020.

[6] J. Zhu, Y. Shen, D. Zhao, and B. Zhou. In-domain gan inversion for real image editing. *arXiv preprint arXiv:2004.00049*, 2020.