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

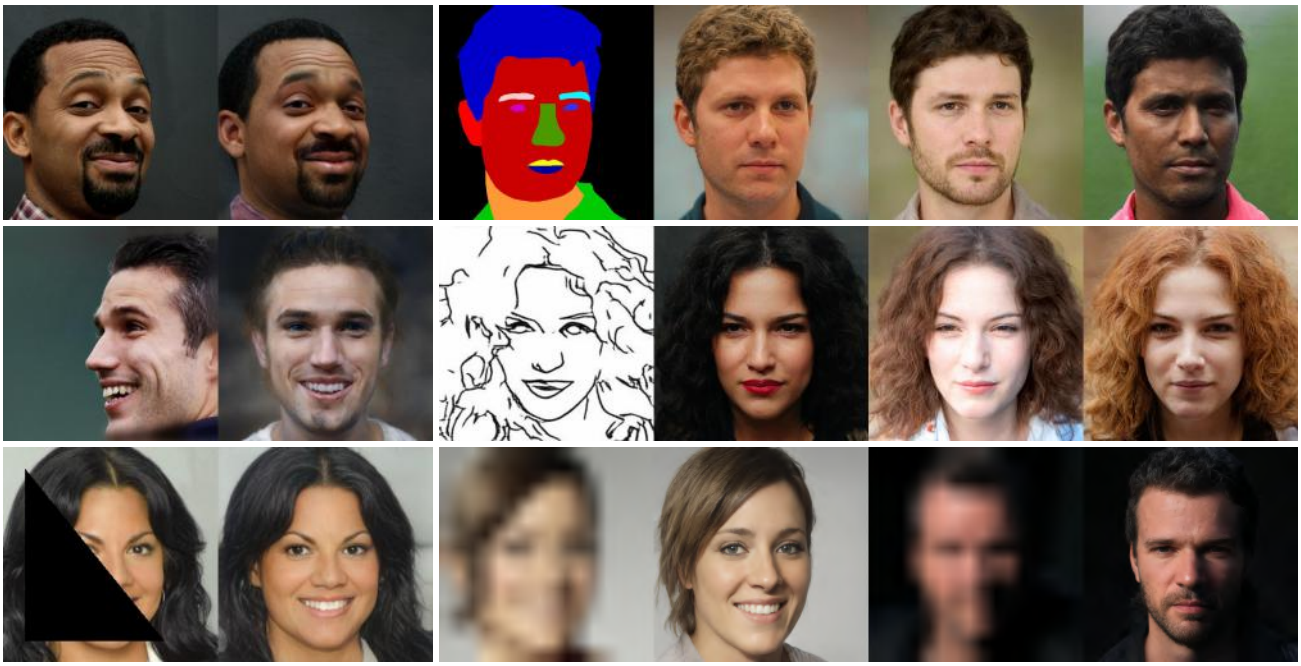

Figure 1. The proposed pixel2style2pixel framework can be used to solve a wide variety of image-to-image translation tasks. Here we show results of pSp on StyleGAN inversion, multi-modal conditional image synthesis, facial frontalization, inpainting and super-resolution.

Abstract

We present a generic image-to-image translation framework, pixel2style2pixel (pSp). Our pSp framework is based on a novel encoder network that directly generates a series of style vectors which are fed into a pretrained StyleGAN generator, forming the extended $\mathcal{W}+$ latent space. We first show that our encoder can directly embed real images into $\mathcal{W}+$, with no additional optimization. Next, we propose utilizing our encoder to directly solve image-to-image translation tasks, defining them as encoding problems from some input domain into the latent domain. By deviating from the standard "invert first, edit later" methodology used with previous StyleGAN encoders, our approach can handle a variety of tasks even when the input image is not represented in the StyleGAN domain. We show that solving translation tasks through StyleGAN significantly simplifies the training process, as no adversary is required, has better support for solving tasks without pixel-to-pixel correspondence, and inherently supports multi-modal synthesis via the resampling of styles. Finally, we demonstrate the potential of our framework on a variety of facial image-to-image translation tasks, even when compared to state-of-the-art solutions designed specifically for a single task, and further show that it can be extended beyond the human facial domain. Code is available at https://github.com/eladrich/pixel2style2pixel.

1. Introduction

In recent years, Generative Adversarial Networks (GANs) have significantly advanced image synthesis, particularly on face images. State-of-the-art image generation methods have achieved high visual quality and fidelity, and can now generate images with phenomenal realism. Most notably, StyleGAN [21, 22] proposes a novel style-

based generator architecture and attains state-of-the-art visual quality on high-resolution images. Moreover, it has been demonstrated that it has a disentangled latent space, $\mathcal{W}$ [44, 7, 39], which offers control and editing capabilities.

Recently, numerous methods have shown competence in controlling StyleGAN's latent space and performing meaningful manipulations in $\mathcal{W}$ [17, 39, 41, 13]. These methods follow an "*invert first, edit later*" approach, where one first inverts an image into StyleGAN's latent space and then edits the latent code in a semantically meaningful manner to obtain a new code that is then used by StyleGAN to generate the output image. However, it has been shown that inverting a real image into a 512-dimensional vector $\mathbf{w} \in \mathcal{W}$ does not lead to an accurate reconstruction. Motivated by this, it has become common practice [1, 2, 4, 48, 3] to encode real images into an extended latent space, $\mathcal{W}+$, defined by the concatenation of 18 different 512-dimensional $\mathbf{w}$ vectors, one for each input layer of StyleGAN. These works usually resort to using per-image optimization over $\mathcal{W}+$, requiring several minutes for a single image. To accelerate this optimization process, some methods [4, 48] have trained an encoder to infer an approximate vector in $\mathcal{W}+$ which serves as a good initial point from which additional optimization is required. However, a fast and accurate inversion of real images into $\mathcal{W}+$ remains a challenge.

In this paper, we first introduce a novel encoder architecture tasked with encoding an arbitrary image directly into $\mathcal{W}+$. The encoder is based on a Feature Pyramid Network [26], where style vectors are extracted from different pyramid scales and inserted directly into a *fixed, pretrained StyleGAN generator* in correspondence to their spatial scales. We show that our encoder can directly reconstruct real input images, allowing one to perform latent space manipulations without requiring time-consuming optimization. While these manipulations allow for extensive editing of real images, they are inherently limited. That is because the input image must be invertible, i.e., there must exist a latent code that reconstructs the image. This requirement is a severe limitation for tasks, such as conditional image generation, where the input image does not reside in the same StyleGAN domain. To overcome this limitation we propose using our encoder together with the pretrained StyleGAN generator as a complete image-to-image translation framework. In this formulation, input images are directly encoded into the desired output latents which are then fed into StyleGAN to generate the desired output images. This allows one to utilize StyleGAN for image-to-image translation even when the input and output images are not from the same domain.

While many previous approaches to solving image-to-image translation tasks involve dedicated architectures specific for solving a single problem, we follow the spirit of pix2pix [16] and define a generic framework able to solve a wide range of tasks, all using the same architecture. Besides the simplification of the training process, as no adversary discriminator needs to be trained, using a pretrained StyleGAN generator offers several intriguing advantages over previous works. For example, many image-to-image architectures explicitly feed the generator with residual feature maps from the encoder [16, 43], creating a strong locality bias [37]. In contrast, our generator is governed only by the styles with no direct spatial input. Another notable advantage of the intermediate style representation is the inherent support for multi-modal synthesis for ambiguous tasks such as image generation from sketches, segmentation maps, or low-resolution images. In such tasks, the generated styles can be resampled to create variations of the output image with no change to the architecture or training process. In a sense, our method performs *pixel2style2pixel* translation, as every image is first encoded into style vectors and then into an image, and is therefore dubbed *pSp*.

The main contributions of this paper are: (i) A novel StyleGAN encoder able to directly encode real images into the $\mathcal{W}+$ latent domain; and (ii) A new methodology for utilizing a pretrained StyleGAN generator to solve image-to-image translation tasks.

2. Related Work

GAN Inversion. With the rapid evolution of GANs, many works have tried to understand and control their latent space. A specific task that has received substantial attention is *GAN Inversion*, which was first introduced by Zhu *et al*. [49]. In this task, the latent vector from which a pretrained GAN most accurately reconstructs a given, known image, is sought. Motivated by its state-of-the-art image quality and latent space semantic richness, many recent works have used StyleGAN [21, 22] for this task. Generally, inversion methods either directly optimize the latent vector to minimize the error for the given image [27, 8, 1, 2], train an encoder to map the given image to the latent space [35, 8, 36, 12, 33], or use a hybrid approach combining both [4, 48]. Typically, methods performing optimization are superior in reconstruction quality to a learned encoder mapping, but require a substantially longer time. Unlike the above methods, our encoder can accurately and efficiently embed a given face image into the extended latent space $\mathcal{W}+$ with no further optimization.

Latent Space Manipulation. Recently, numerous papers have presented diverse methods for learning semantic edits of the latent code. One popular approach is to find linear directions that correspond to changes in a given binary labeled attribute, such as young $\leftrightarrow$ old, or no-smile $\leftrightarrow$ smile [39, 11, 10, 3]. Tewari *et al*. [41] utilize a pretrained 3DMM to learn semantic face edits in the latent space. Jahanian *et al*. [17] find latent space paths that correspond to a spe-

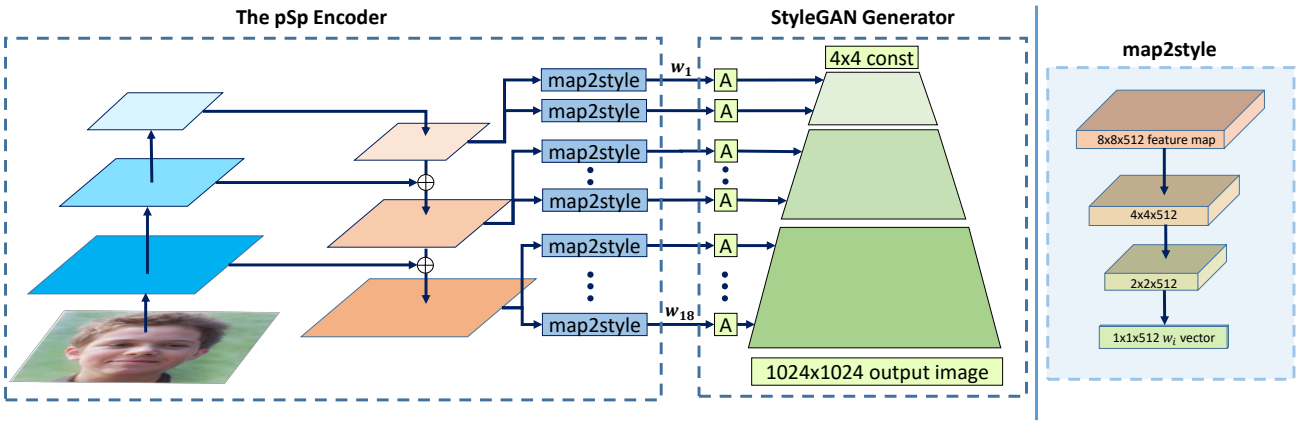

Figure 2. Our pSp architecture. Feature maps are first extracted using a standard feature pyramid over a ResNet backbone. For each of the 18 target styles, a small mapping network is trained to extract the learned styles from the corresponding feature map, where styles (0-2) are generated from the small feature map, (3-6) from the medium feature map, and (7-18) from the largest 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. Each generated 512 vector, is fed into StyleGAN, starting from its matching affine transformation, A.

cific transformation, such as zoom or rotation, in a self-supervised manner. Härkönen *et al*. [13] find useful paths in an unsupervised manner by using the principal component axes of an intermediate activation space. Collins *et al*. [7] perform local semantic editing by manipulating corresponding components of the latent code. These methods generally follow an "*invert first, edit later*" procedure, where an image is first embedded into the latent space, and then its latent is edited in a semantically meaningful manner. This differs from our approach which directly encodes input images into the corresponding output latents, allowing one to also handle inputs that do not reside in the StyleGAN domain.

Image-to-Image. Image-to-Image translation techniques aim at learning a conditional image generation function that maps an input image of a source domain to a corresponding image of a target domain. Isola *et al*. [16] first introduced the use of conditional GANs to solve various image-to-image translation tasks. Since then, their work has been extended for many scenarios: high-resolution synthesis [43], unsupervised learning [30, 50, 23, 28], multi-modal image synthesis [51, 14, 6], and conditional image synthesis [34, 25, 31, 52, 5]. The aforementioned works have constructed dedicated architectures, which require training the generator network and generally do not generalize to other translation tasks. This is in contrast to our method that uses the same architecture for solving a variety of tasks.

3. The pSp Framework

Our pSp framework builds upon the representative power of a pretrained StyleGAN generator and the $\mathcal{W}+$ latent space. To utilize this representation one needs a strong encoder that is able to match each input image to an accurate encoding in the latent domain. A simple technique to embed into this domain is directly encoding a given input image into $\mathcal{W}+$ using a single 512-dimensional vector obtained from the last layer of the encoder network, thereby learning all 18 style vectors together. However, such an architecture presents a strong bottleneck making it difficult to fully represent the finer details of the original image and is therefore limited in reconstruction quality.

In StyleGAN, the authors have shown that the different style inputs correspond to different levels of detail, which are roughly divided into three groups — coarse, medium, and fine. Following this observation, in pSp we extend an encoder backbone with a feature pyramid [26], generating three levels of feature maps from which styles are extracted using a simple intermediate network — map2style — shown in Figure 2. The styles, aligned with the hierarchical representation, are then fed into the generator in correspondence to their scale to generate the output image, thus completing the translation from input *pixels* to output *pixels*, through the intermediate *style* representation. The complete architecture is illustrated in Figure 2.

As in StyleGAN, we further define $\overline{\mathbf{w}}$ to be the average style vector of the pretrained generator. Given an input image, $\mathbf{x}$, the output of our model is then defined as

$$pSp(\mathbf{x}) := G(E(\mathbf{x}) + \overline{\mathbf{w}})$$

where $E(\cdot)$ and $G(\cdot)$ denote the encoder and StyleGAN generator, respectively. In this formulation, our encoder aims to learn the latent code with respect to the average style vector. We find that this results in better initialization.

3.1. Loss Functions

While the style-based translation is the core part of our framework, the choice of losses is also crucial. Our encoder is trained using a weighted combination of several objectives. First, we utilize the pixel-wise $\mathcal{L}_2$ loss,

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

In addition, to learn perceptual similarities, we utilize the LPIPS [46] loss, which has been shown to better preserve image quality [12] compared to the more standard perceptual loss [18]:

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

where $F(\cdot)$ denotes the perceptual feature extractor.

To encourage the encoder to output latent style vectors closer to the average latent vector, we additionally define the following regularization loss:

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

Similar to the truncation trick introduced in StyleGAN, we find that adding this regularization in the training of our encoder improves image quality without harming the fidelity of our outputs, especially in some of the more ambiguous tasks explored below.

Finally, a common challenge when handling the specific task of encoding facial images is the preservation of the input identity. To tackle this, we incorporate a dedicated recognition loss measuring the cosine similarity between the output image and its source,

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

where R is the pretrained ArcFace [9] network.

In summary, the total loss function is defined as

$$\mathcal{L}(\mathbf{x}) = \lambda_1\mathcal{L}_2(\mathbf{x}) + \lambda_2\mathcal{L}_{\text{LPIPS}}(\mathbf{x}) + \lambda_3\mathcal{L}_{\text{ID}}(\mathbf{x}) + \lambda_4\mathcal{L}_{\text{reg}}(\mathbf{x}),$$

where λ_1, λ_2, λ_3, λ_4 are constants defining the loss weights. This curated set of loss functions allows for more accurate encoding into StyleGAN compared to previous works and can be easily tuned for different encoding tasks according to their nature. Constants and other implementation details can be found in Appendix A.

3.2. The Benefits of The StyleGAN Domain

The translation between images through the *style* domain differentiates pSp from many standard image-to-image translation frameworks, as it makes our model operate *globally* instead of *locally*, without requiring pixel-to-pixel correspondence. This is a desired property as it has been shown that the locality bias limits current methods when handling non-local transformations [37]. Additionally, previous works [21, 7] have demonstrated that the disentanglement of semantic objects learned by StyleGAN is due to its layer-wise representation. This ability to independently manipulate semantic attributes leads to another desired property: the support for *multi-modal synthesis*. As some translation tasks are ambiguous, where a single input image may correspond to several outputs, it is desirable to be able to sample these possible outputs. While this requires specialized changes in standard image-to-image architectures [51, 14], our framework inherently supports this by simply sampling style vectors. In practice, this is done by randomly sampling a vector $\mathbf{w} \in \mathbb{R}^{512}$ and generating a corresponding latent code in $\mathcal{W}+$ by replicating $\mathbf{w}$. Style mixing is then performed by replacing select layers of the computed latent with those of the randomly generated latent, possibly with an α parameter for blending between the two styles. This approach is illustrated in Figure 3.

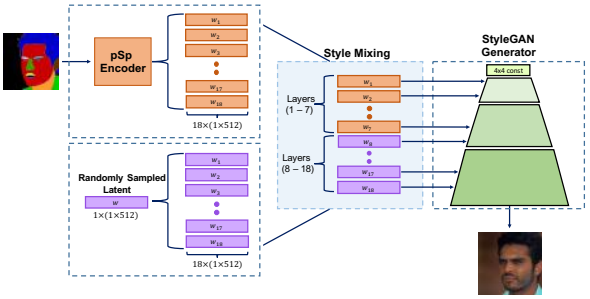

Figure 3. Style-mixing for multi-modal generation.

4. Applications and Experiments

To explore the effectiveness of our approach we evaluate pSp on numerous image-to-image translation tasks.

4.1. StyleGAN Inversion

We start by evaluating the usage of the pSp framework for StyleGAN Inversion, that is, finding the latent code of real images in the latent domain. We compare our method to the optimization technique from Karras *et al*. [22], the ALAE encoder [36] and to the encoder from IDInvert [48]. The ALAE method proposes a StyleGAN-based autoencoder, where the encoder is trained alongside the generator to generate latent codes. In IDInvert, images are embedded into the latent domain of a pretrained StyleGAN by first encoding the image into $\mathcal{W}+$ and then directly optimizing over the generated image to tune the latent. For a fair comparison, we compare with IDInvert where no further optimization is performed after encoding.

Results. Figure 4 shows a qualitative comparison between the methods. One can see that the ALAE method, operating in the $\mathcal{W}$ domain, cannot accurately reconstruct the input images. While IDInvert [48] better preserves the image attributes, it still fails to accurately preserve identity and the finer details of the input image. In contrast, our

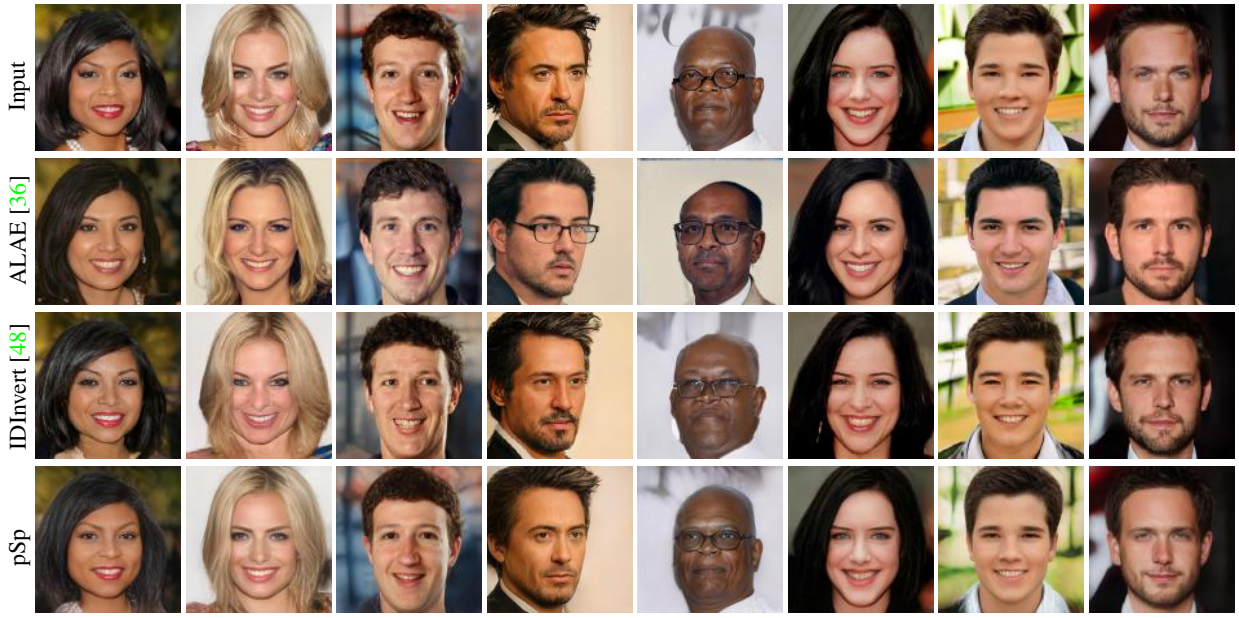

Figure 4. Results of pSp for StyleGAN inversion compared to other encoders on CelebA-HQ.

Method	↑ Similarity	↓ LPIPS	↓ MSE	↓ Runtime
Karras *et al.* [22]	0.77	0.11	0.02	182.1
ALAE [36]	0.06	0.32	0.15	0.207
IDInvert [48]	0.18	0.22	0.06	0.032
$\mathcal{W}$ Encoder	0.35	0.23	0.06	0.064
Naive $\mathcal{W}+$	0.49	0.19	0.04	0.064
pSp w/o ID	0.19	0.17	0.03	0.105
pSp	0.56	0.17	0.03	0.105

Table 1. Quantitative results for image reconstruction.

method is able to preserve identity while also reconstructing fine details such as lighting, hairstyle, and glasses.

Next, we conduct an ablation study to analyze the effectiveness of the pSp architecture. We compare our architecture to two simpler variations. First, we define an encoder generating a 512-dimensional style vector in the $\mathcal{W}$ latent domain, extracted from the last layer of the encoder network. We then expand this and define an encoder with an additional layer to transform the 512-dimensional feature vector to a full 18×512 $\mathcal{W}+$ vector. Figure 5 shows that while this simple extension into $\mathcal{W}+$ significantly improves the results, it still cannot preserve the finer details generated by our architecture. In Figure 6 we show the importance of the identity loss in the reconstruction task.

Finally, Table 1 presents a quantitative evaluation measuring the different inversion methods. Compared to other encoders, pSp is able to better preserve the original images in terms of both perceptual similarity and identity. To make sure the similarity score is independent of our loss function, we utilize the CurricularFace [15] method for evaluation.

Figure 5. Ablation of the pSp encoder over CelebA-HQ.

Figure 6. The importance of identity loss.

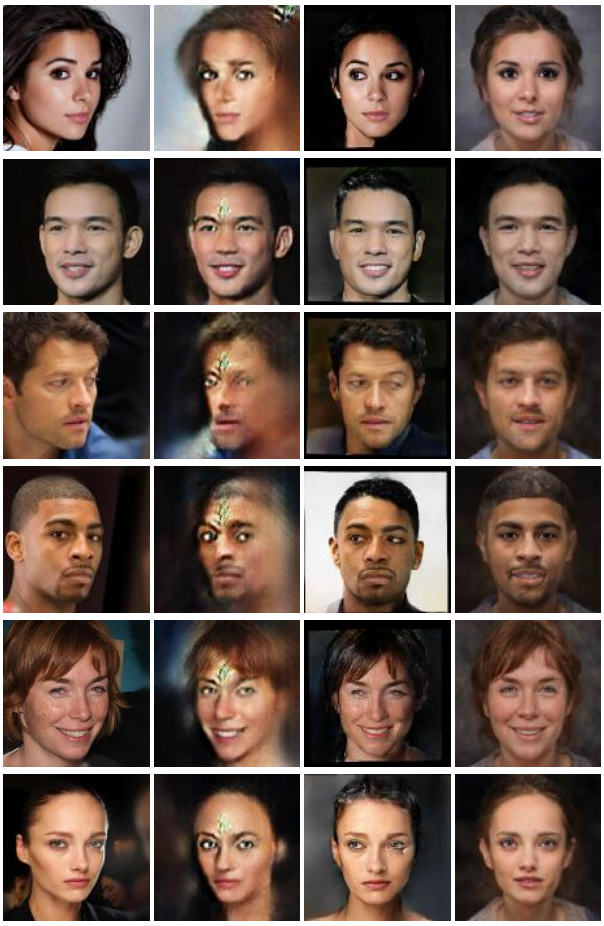

Figure 7. Comparison of face frontalization methods.

4.2. Face Frontalization

Face frontalization is a challenging task for image-to-image translation frameworks due to the required non-local transformations and the lack of paired training data. Rotate-AndRender (R&R) [47] overcome this challenge by incorporating a geometric 3D alignment process before the translation process. Alternatively, we show that our style-based translation mechanism is able to overcome these challenges, even when trained with no labeled data.

Methodology. For this task, training is the same as the encoder formulation with two important changes. First, we randomly flip the target image during training, effectively forcing the model to output an image that is close to both the original image and the mirrored one. The underlying idea behind this augmentation is that it guides the model to converge to a fixed frontal pose. Next, we increase $\mathcal{L}_{\text{ID}}$ and decrease the $\mathcal{L}_2$ and $\mathcal{L}_{\text{LPIPS}}$ losses for the outer part of the image. This change is based on the fact that for frontalization we are less interested in preserving the background region compared to the face region and the facial identity.

Method	↑ Similarity				↓ Runtime
	90°	70°	50°	30°	
R&R	**0.34**	**0.56**	**0.66**	**0.7**	1.5s
pSp	0.32	0.52	0.60	0.63	**0.1s**

Table 2. Results for Face Frontalization on the FEI Face Database split by rotation angle of the face in the input.

Results. Results are illustrated in Figure 7. When trained with the same data and methodology, pix2pixHD is unable to converge to satisfying results as it is much more dependent on the correspondence between the input and output pairs. Conversely, our method is able to handle the task successfully, generating realistic frontal faces, which are comparable to the more involved R&R approach. This shows the benefit of using a pretrained StyleGAN for image translation, as it allows us to achieve visually-pleasing results even with weak supervision. Table 2 provides a quantitative evaluation on the FEI Database [42]. While R&R outperforms pSp, our simple approach provides a fast and elegant alternative, without requiring specialized modules, such as R&R's 3DMM fitting and inpainting steps.

4.3. Conditional Image Synthesis

Conditional image synthesis aims at generating photo-realistic images conditioned on certain input types. In this section, our pSp architecture is tested on two conditional image generation tasks: generating high-quality face images from sketches and semantic segmentation maps. We demonstrate that, with only minimal changes, our encoder successfully utilizes the expressiveness of StyleGAN to generate high-quality and diverse outputs.

Methodology and details. The training of the two conditional generation tasks is similar to that of the encoder, where the input is the conditioned image and the target is the corresponding real image. To generate multiple images at inference time we perform style-mixing on the fine-level features, taking layers (1-7) from the latent code of the input image and layers (8-18) from a randomly drawn $\mathbf{w}$ vector.

4.3.1 Face From Sketch

Common approaches for sketch-to-image synthesis incorporate hard constraints that require pixel-wise correspondence between the input sketch and generated image, making them ill-suited when given incomplete, sparse sketches. DeepFaceDrawing [5] address this using a set of dedicated mapping networks. We show that pSp provides a simple alternative to past approaches. As there are currently no publicly available datasets representative of hand-drawn face sketches, we elect to construct our own dataset, which we describe in Appendix B.

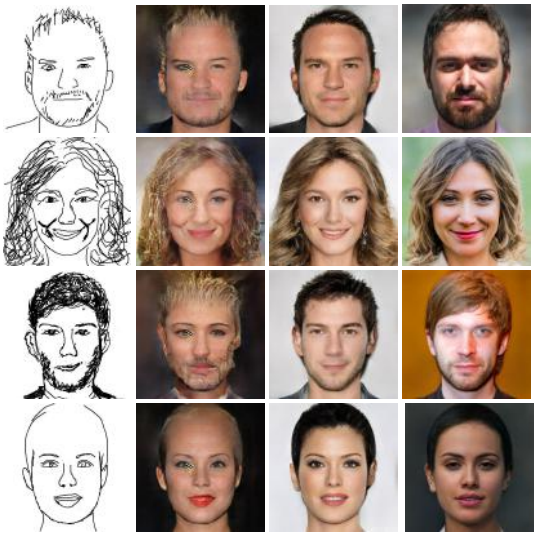

Figure 8. Comparison of sketches presented in DeepFace-Drawing [5].

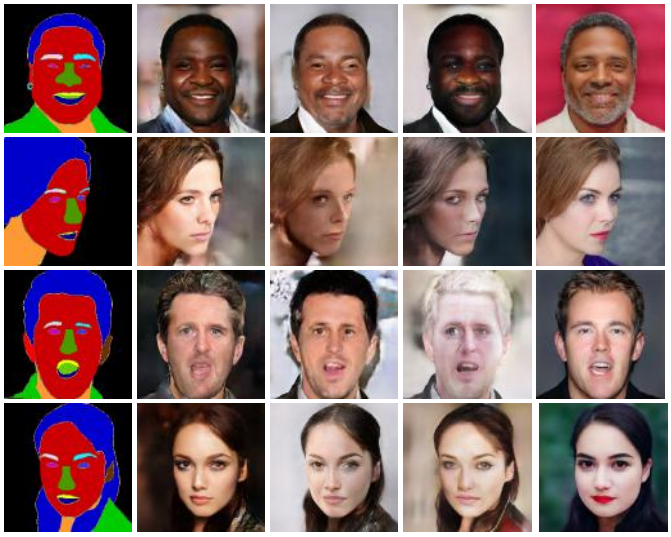

Figure 9. Comparisons to other label-to-image methods.

Results. Figure 8 compares the results of our method to those of pix2pixHD and DeepFaceDrawing. As no code release is available for DeepFaceDrawing, we compare directly with sketches and results published in their paper. While DeepFaceDrawing obtain more visually pleasing results compared to pix2pixHD, they are still limited in their diversity. Conversely, although our model is trained on a different dataset, we are still able to generalize well to their sketches. Notably, we observe our ability to obtain more diverse outputs that better retain finer details (e.g. facial hair). Additional results, including those on non-frontal sketches are provided in the Appendix.

4.3.2 Face from Segmentation Map

Here, we evaluate using pSp for synthesizing face images from segmentation maps. In addition to pix2pixHD, we compare our approach to two additional state-of-the-art label-to-image methods: SPADE [34], and CC_FPSE [31], both of which are based on pix2pixHD.

Results. In Figure 9 we provide a visual comparison of the competing approaches on the CelebAMask-HQ dataset containing 19 semantic categories. As the competing methods are based on pix2pixHD, the results of all three suffer from similar artifacts. Conversely, our approach is able to generate high-quality outputs across a wide range of inputs of various poses and expressions. Additionally, using our multi-modal technique, pSp can easily generate various possible outputs with the same pose and attributes but varying fine styles for a single input semantic map or sketch image. We provide examples in Figure 1 with additional results in the Appendix.

Task	pix2pixHD	SPADE	CC_FPSE
Segmentation	94.72%	95.25%	93.06%
Sketch	93.34%	N/A	N/A

Table 3. Human evaluation results on CelebA-HQ for conditional image synthesis tasks. Each cell denotes the percentage of users who favored pSp over the listed method.

Human Perceptual Study. We additionally perform a human evaluation to compare the visual quality of each method presented above. Each worker is given two images synthesized by different methods on the same input and is given an unlimited time to select which output looks more realistic. Each of our three workers reviews approximately $2,800$ pairs for each task, resulting in over $8,400$ human judgements for each method. Table 3 shows that pSp significantly outperforms the other respective methods in both synthesis tasks.

4.4. Extending to Other Applications

Besides the applications presented above, we have found pSp to be applicable to a wide variety of additional tasks with minimal changes to the training process. Specifically, we present samples of super-resolution and inpainting results using pSp in Figure 1 with further details and results presented in Appendix C. For both tasks, paired data is generated and training is performed in a supervised fashion. Additionally, we show multi-modal support for super-resolution via style-mixing on medium-level features and evaluate pSp on several image editing tasks, including image interpolation and local patch editing.

StyleGAN Inversion and Reconstruction

Image Generation from Sketches

Figure 10. Results of pSp on the AFHQ Dataset for StyleGAN Inversion and the sketch-to-image tasks. For reconstruction, the input (left) is shown alongside the reconstructed output (right). For sketch-to-image, multiple outputs are generated via style-mixing.

4.5. Going Beyond the Facial Domain

In this section we show that our pSp framework can be trained to solve the various tasks explored above without relying on the advantages provided by the identity loss in the facial domain. While our method does require a pretrained StyleGAN generator, recent works [20, 38] have shown that such a generator can be easily trained with significantly fewer examples than required in the past.

Figure 20 shows the results on the AFHQ Cat and AFHQ Dog datasets [6] for the StyleGAN inversion and sketch-to-image tasks. For these tasks, we use a pretrained StyleGAN-ADA [20] model for each of the two domains and train our pSp encoder using only the $\mathcal{L}_2$, $\mathcal{L}_{\text{LPIPS}}$, and $\mathcal{L}_{\text{reg}}$ losses with the same λ values as those used for the facial domain. As shown, we are able to generalize well to the examined domains, obtaining high-quality, accurate reconstruction results while also supporting multi-modal synthesis via our style-mixing approach. The accompanying Appendix provides additional results for super-resolution and inpainting on these domains.

Figure 11. Challenging cases for StyleGAN Inversion.

5. Discussion

Although our suggested framework for image-to-image translation achieves compelling results in various applications, it has some inherent assumptions that should be considered. First, the high-quality images that are generated by utilizing the pretrained StyleGAN come with a cost — the method is limited to images that can be generated by StyleGAN. Thus, generating faces which are not close to frontal, or have certain expressions may be challenging if such examples were not available when training the StyleGAN model. Also, the global approach of pSp, although advantageous for many tasks, does introduce a challenge in preserving finer details of the input image, such as earrings or background details. This is especially significant in tasks such as inpainting or super-resolution where standard image-to-image architectures can simply propagate local information. Figure 11 presents some examples of such reconstruction failures.

6. Conclusion

In this work, we propose a novel encoder architecture that can be used to directly map a real image into the $\mathcal{W}+$ latent space with no optimization required. There, styles are extracted in a hierarchical fashion and fed into the corresponding inputs of a fixed StyleGAN generator. Combining our encoder with a StyleGAN decoder, we present a generic framework for solving various image-to-image translation tasks, all using the same architecture. Notably, in contrast to the "*invert first, edit later*" approach of previous StyleGAN encoders, we show pSp can be used to directly encode these translation tasks into StyleGAN, thereby supporting input images that do not reside in the StyleGAN domain. Additionally, differing from previous works that typically rely on dedicated architectures for solving a single translation task, we show pSp to be capable of solving a wide variety of problems, requiring only minimal changes to the training losses and methodology. We hope that the ease-of-use of our approach will encourage further research into utilizing StyleGAN for real image-to-image translation tasks.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE international conference on computer vision*, pages 4432–4441, 2019.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020.

[3] Rameen Adbal, Pie Zhu, Niloy J. Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *arXiv preprint arXiv:*, 2020.

[4] Baylies. stylegan-encoder. https://github.com/pbaylies/stylegan-encoder, 2019. Accessed: April 2020.

[5] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. DeepFaceDrawing: Deep generation of face images from sketches. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2020)*, 39(4):72:1–72:16, 2020.

[6] 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.

[7] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5771–5780, 2020.

[8] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 30(7):1967–1974, 2018.

[9] 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.

[10] Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439*, 2019.

[11] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties, 2019.

[12] Shanyan Guan, Ying Tai, Bingbing Ni, Feida Zhu, Feiyue Huang, and Xiaokang Yang. Collaborative learning for faster stylegan embedding. *arXiv preprint arXiv:2007.01758*, 2020.

[13] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020.

[14] 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.

[15] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5901–5910, 2020.

[16] 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.

[17] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019.

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[19] 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.

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

[21] 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.

[22] 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.

[23] Oren Katzir, Dani Lischinski, and Daniel Cohen-Or. Cross-domain cascaded deep feature translation. *arXiv*, pages arXiv–1906, 2019.

[24] 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.

[25] Yuhang Li, Xuejin Chen, Feng Wu, and Zheng-Jun Zha. Linestofacephoto: Face photo generation from lines with conditional self-attention generative adversarial networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2323–2331, 2019.

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[27] Zachary C Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks. *arXiv preprint arXiv:1702.04782*, 2017.

[28] Wallace Lira, Johannes Merz, Daniel Ritchie, Daniel Cohen-Or, and Hao Zhang. Ganhopper: Multi-hop gan for

unsupervised image-to-image translation. *arXiv preprint arXiv:2002.10102*, 2020.

[29] 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.

[30] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[31] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang 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.

[32] 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.

[33] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Disentangling in latent space by harnessing a pretrained generator. *arXiv preprint arXiv:2005.07728*, 2020.

[34] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[35] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[36] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14104–14113, 2020.

[37] Eitan Richardson and Yair Weiss. The surprising effectiveness of linear unsupervised image-to-image translation. *ArXiv*, abs/2007.12568, 2020.

[38] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-shot adaptation of generative adversarial networks. *arXiv*, 2020.

[39] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020.

[40] 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.

[41] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. *arXiv preprint arXiv:2004.00121*, 2020.

[42] Carlos Eduardo Thomaz and Gilson Antonio Giraldi. A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing*, 28(6):902 – 913, 2010.

[43] 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.

[44] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis, 2019.

[45] 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.

[46] 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.

[47] Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang 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.

[48] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. *arXiv preprint arXiv:2004.00049*, 2020.

[49] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016.

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[51] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017.

[52] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020.

A. Implementation Details

For our backbone network we use the ResNet-IR architecture from [9] pretrained on face recognition, which accelerated convergence. We use a *fixed* StyleGAN2 generator trained on the FFHQ [21] 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 [29] with the Lookahead technique [45], 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.

B. Dataset Details

We conduct our experiments on the CelebA-HQ dataset [19], 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 [21], 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 [40], 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 [6].

C. Application Details

C.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 [32] 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.

Figure 12. Comparison of super-resolution approaches with $\times 8$ down-sampling on the CelebA-HQ [19] test set.

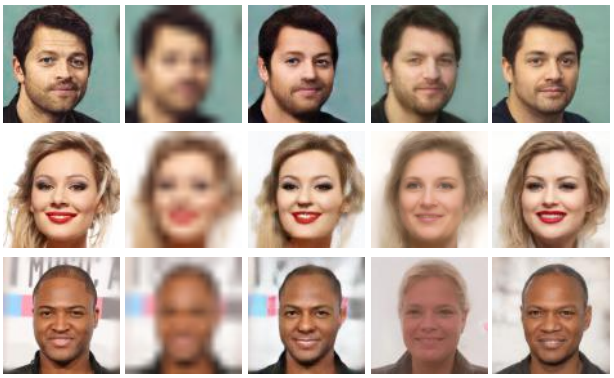

Figure 13. Comparison of super-resolution approaches with $\times 16$ down-sampling on the CelebA-HQ [19] test set.

Methodology and details. We train both our model and pix2pixHD [43] 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. Figures 12-14 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

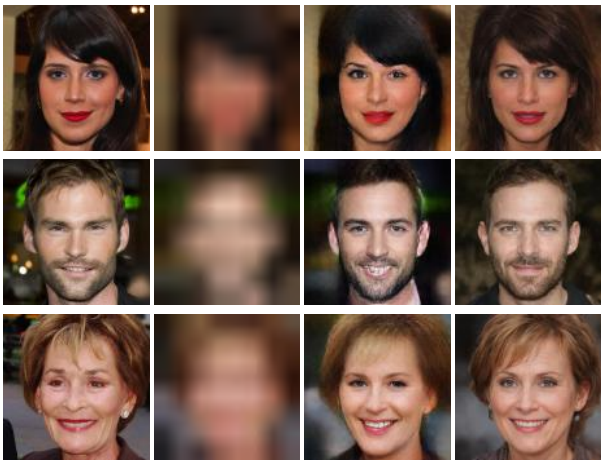

Figure 14. Comparison of super-resolution approaches with $\times 32$ down-sampling on the CelebA-HQ [19] test set.

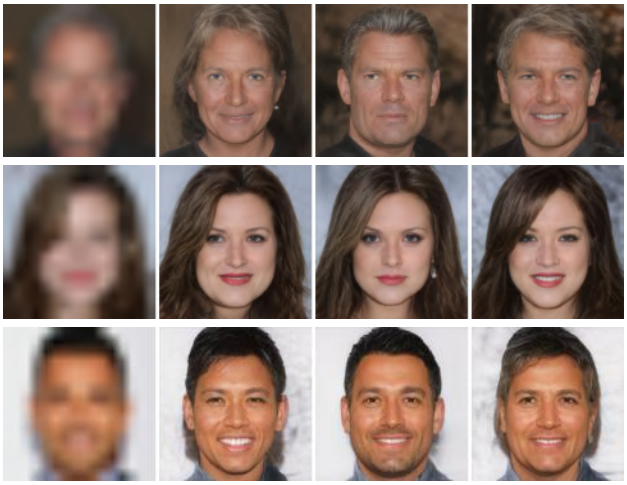

Figure 15. Multi-modal synthesis for super-resolution using pSp with style-mixing.

are able to obtain high-quality results even when down-sampling to resolutions of 16×16 and 8×8. Finally, in Figure 15 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.

C.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 [16] and other local-based translation methods, have shown success in tackling this problem as they can simply propagate non-occluded regions.

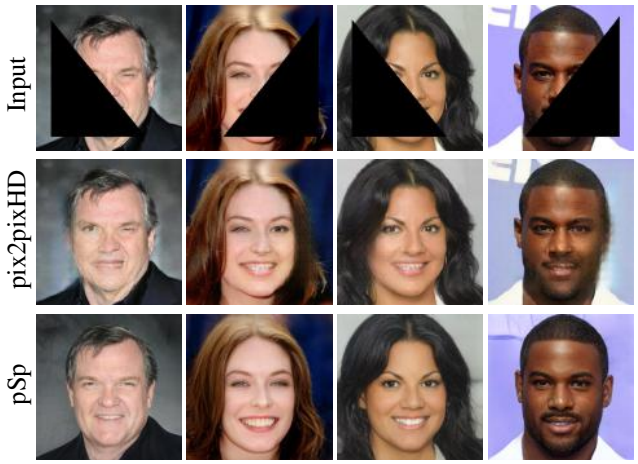

Figure 16. Image inpainting results using pSp and pix2pixHD [43] on the CelebA-HQ [19] test set.

Methodology and details We train both pSp and pix2pixHD [43] in a supervised fashion, where each input image is occluded with a symmetric triangular mask.

Results Figure 16 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.

C.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 18, pSp is able to seamlessly merge the desired patch into the original image.

C.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'.

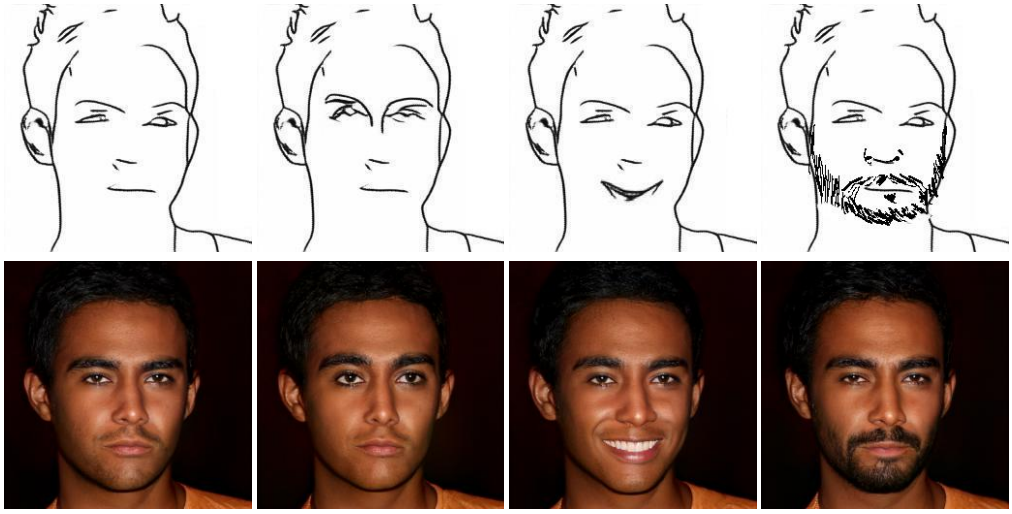

Figure 17. Sketch-Based Local Editing

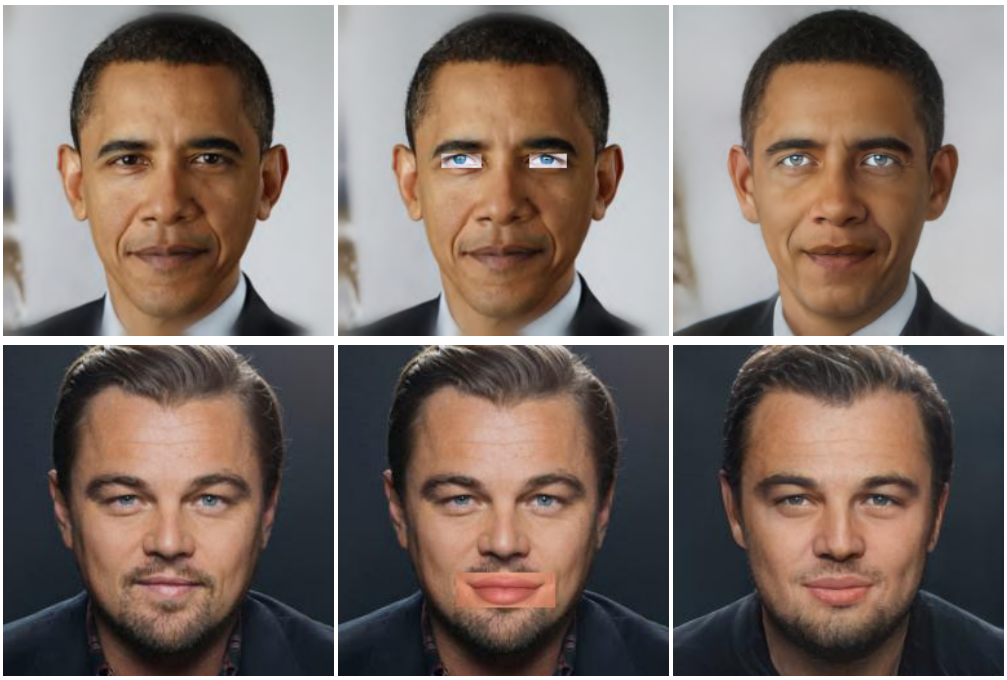

Figure 18. Local patch editing results using pSp on real images.

Figure 19. Image interpolation results using pSp on the CelebA-HQ [19] test set.

Figure 20. Results of pSp on the AFHQ Cat and Dog datasets [6] on super resolution, inpainting, and image generation from sketches.

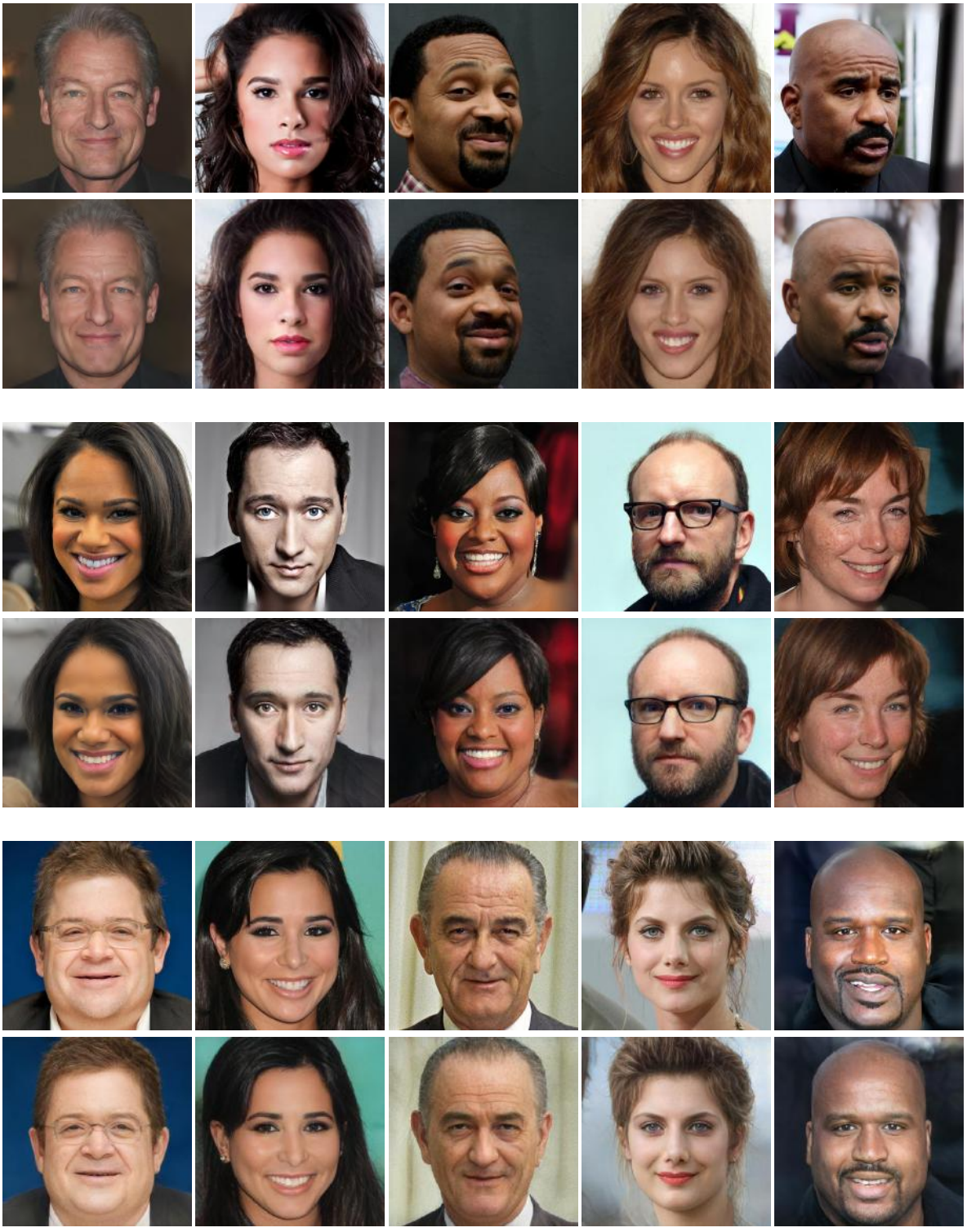

Figure 21. Additional StyleGAN inversion results using pSp on the CelebA-HQ [19] test set.

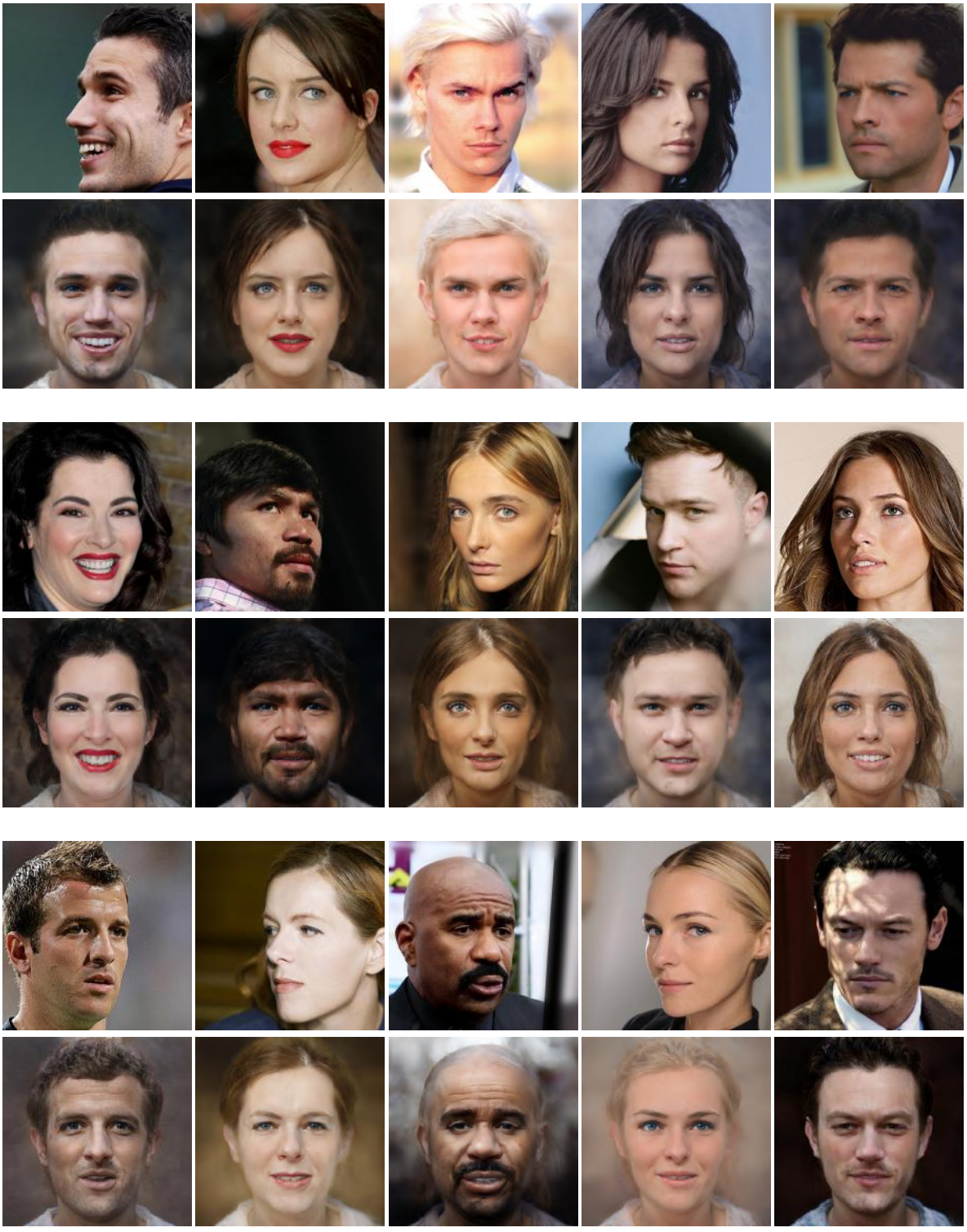

Figure 22. Additional face frontalization results using pSp on the CelebA-HQ [19] test set.

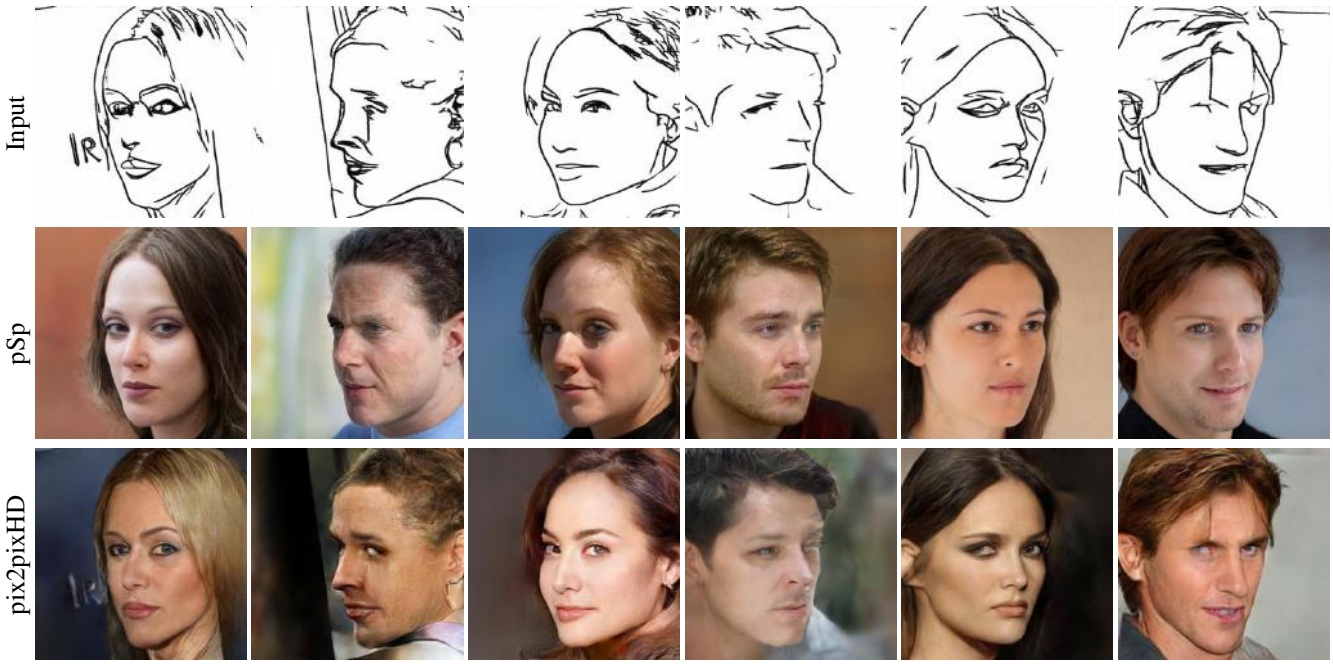

Figure 23. Even for challenging, non-frontal face sketches, pSp is able to obtain high-quality, diverse outputs.

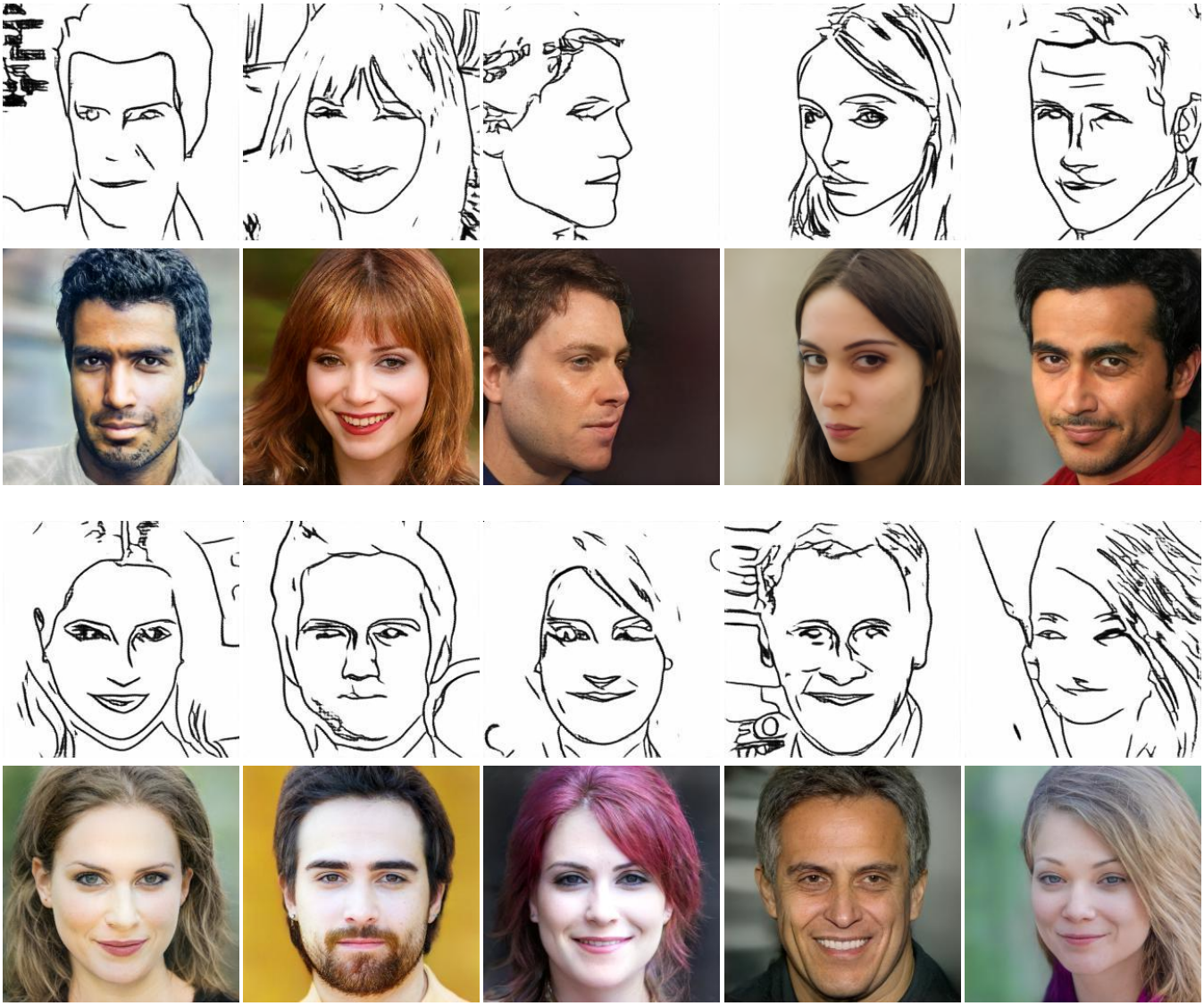

Figure 24. Additional results using pSp for the generation of face images from sketches on the CelebA-HQ [19] test dataset.

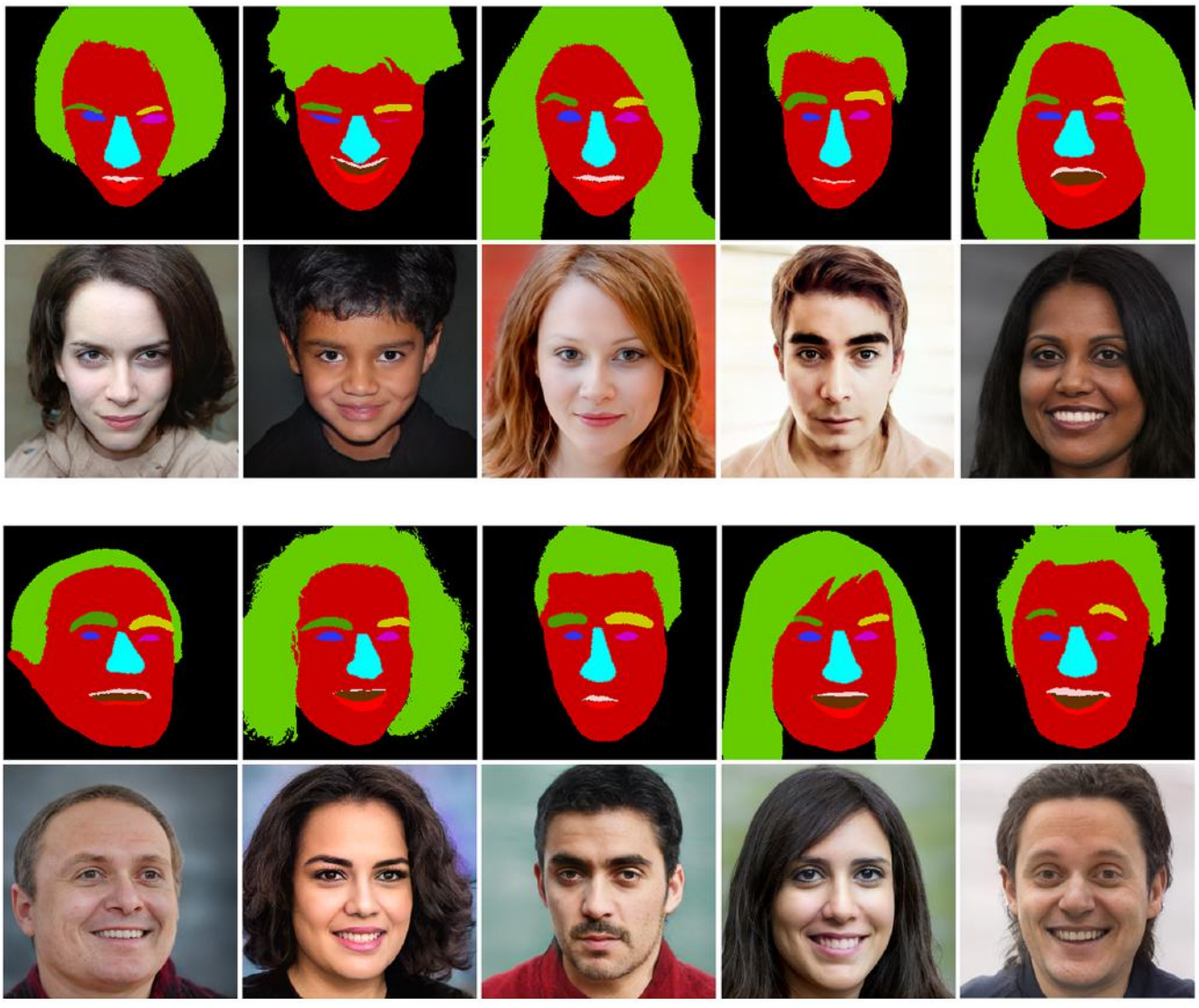

Figure 25. Additional results on the Helen Faces [24] dataset using our proposed segmentation-to-image method.

Figure 26. Additional results on the CelebAMask-HQ [19] test set using our proposed segmentation-to-image method.

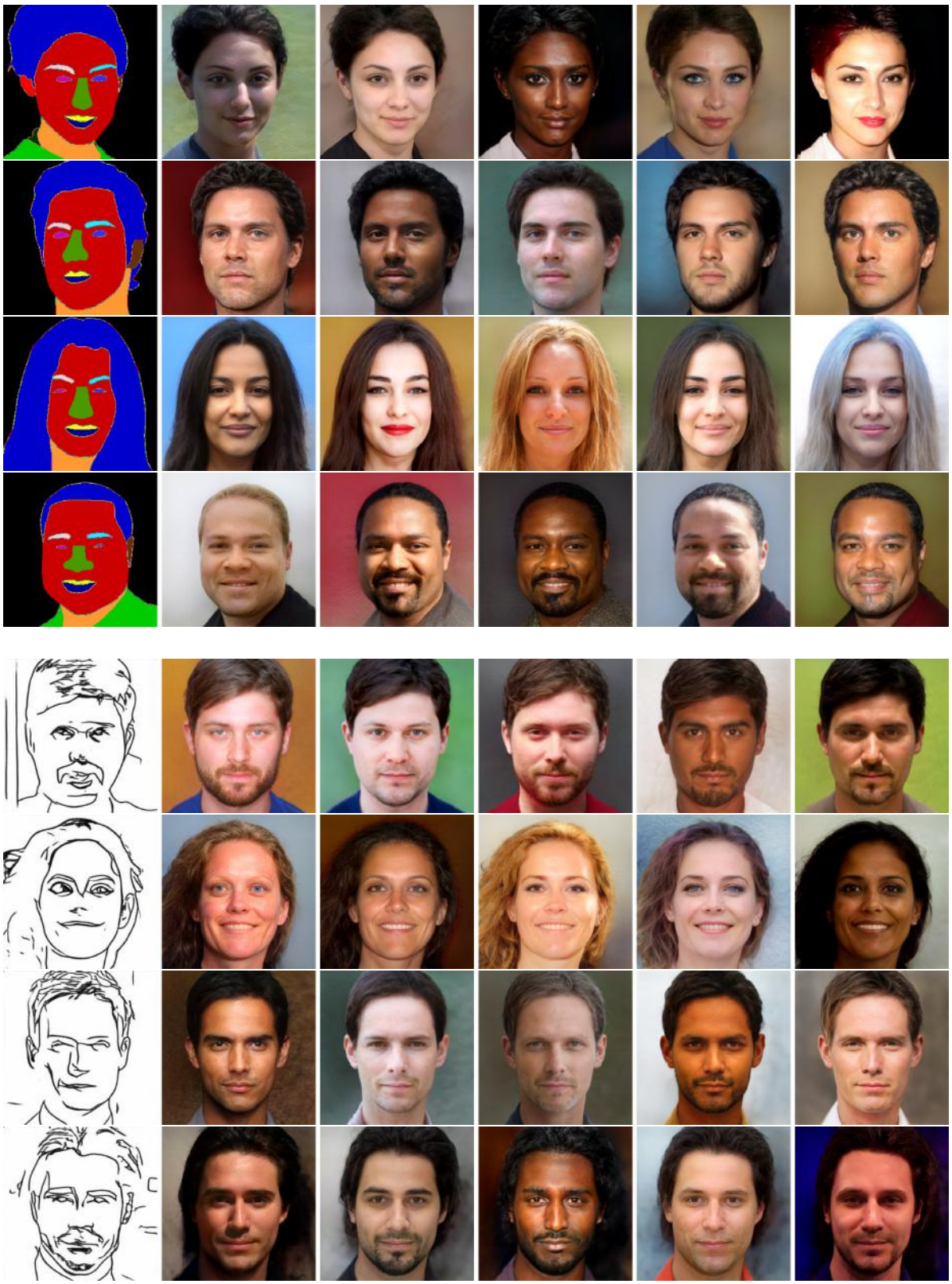

Figure 27. Conditional image synthesis results from sketches and segmentation maps displaying the multi-modal property of our approach.