ENCODING IN STYLE: A STYLEGAN ENCODER FOR IMAGE-TO-IMAGE TRANSLATION

Anonymous authors
Paper under double-blind review

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. We further introduce a dedicated identity loss which is shown to achieve improved performance in the reconstruction of an input image. We demonstrate pSp to be a simple architecture that, by leveraging a well-trained, *fixed* generator network, can be easily applied on a wide-range of image-to-image translation tasks. Solving these tasks through the style representation results in a global approach that does not rely on a local pixel-to-pixel correspondence and further supports multi-modal synthesis via the resampling of styles. Notably, we demonstrate that pSp can be trained to align a face image to a frontal pose with no labeled data and generate multi-modal results for ambiguous tasks such as conditional face generation from sketches and segmentation maps.

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 (Karras et al., 2019; 2020) 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, $\mathcal{W}$ (Yang et al., 2019; Collins et al., 2020; Shen et al., 2020), which may offer control and editing capabilities.

Recently, numerous methods have shown competence in controlling StyleGAN's latent space and performing meaningful manipulations in $\mathcal{W}$ (Jahanian et al., 2019; Shen et al., 2020; Tewari et al., 2020; Härkönen et al., 2020). To perform such edits on real images, one needs to invert the image into StyleGAN's latent space, i.e., retrieve the latent code that reconstructs the 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 (Abdal et al., 2019; 2020; Baylies, 2019; Zhu et al., 2020a; Adbal et al., 2020) 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. Nevertheless, many methods resort to using per-image optimization over $\mathcal{W}+$, requiring several minutes for a single image. To accelerate this optimization process, some methods (Baylies, 2019; Zhu et al., 2020a) 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, direct, and accurate learned inversion of real images into $\mathcal{W}+$ remains a challenge.

In this paper, we focus on the broader task of *latent space embedding*, which aims to retrieve the latent vector that generates a desired, not necessarily known, image. We do so by introducing a novel encoder architecture tasked with encoding an arbitrary image directly into $\mathcal{W}+$. The encoder is based on a Feature Pyramid Network (Lin et al., 2017), where style feature vectors are extracted from different pyramid scales and inserted directly into a *fixed, pretrained StyleGAN generator* in correspondence to their spatial scales. Our encoder into $\mathcal{W}+$, together with the StyleGAN decoder, form a generic encoder-decoder network that benefits many image-to-image translation tasks. Focusing on face images, we first demonstrate our method's ability to successfully reconstruct a given image while preserving identity and other attributes. We then present numerous image-to-image translation applications. 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*.

While many previous approaches to solving image-to-image translations tasks involve dedicated architectures specific for solving a single problem, we follow the spirit of pix2pix (Isola et al., 2017) 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. Many image-to-image architectures explicitly feed the generator with residual feature maps from the encoder (Isola et al., 2017; Wang et al., 2018), creating a strong locality bias (Richardson & Weiss, 2020). In contrast, our generator is governed only by the styles with no direct spatial input. The advantage of such a global approach is most evident in the task of *Face Frontalization*, where our encoder can be trained to align a given face image to a frontal pose with no labeled data. Another notable advantage of the intermediate style representation is the inherent support for multi-modal synthesis for ambiguous tasks such as face 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.

The main contributions of this paper are: (i) a novel StyleGAN encoder able to directly encode real face images into the $\mathcal{W}+$ target latent domain; and (ii) a generic end-to-end framework for solving image-to-image translation tasks.

2 Related Work

Latent Space Embedding 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* — where 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 (Karras et al., 2019; 2020) for this task. Generally, inversion methods either directly optimize the latent vector to minimize the error for the given image (Lipton & Tripathi, 2017; Creswell & Bharath, 2018; Abdal et al., 2019; 2020), train an encoder to map the given image to the latent space (Perarnau et al., 2016; Creswell & Bharath, 2018; Pidhorskyi et al., 2020; Guan et al., 2020; Nitzan et al., 2020), or use a hybrid approach combining both (Baylies, 2019; Zhu et al., 2020a). 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 $\mathcal{W}+$ of a *fixed*, pretrained StyleGAN generator, with no further optimization.

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. (2017) 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 (Wang et al., 2018), unsupervised learning (Liu et al., 2017; Zhu et al., 2017a; Katzir et al., 2019; Lira et al., 2020), multi-modal image synthesis (Zhu et al., 2017b; Huang et al., 2018; Choi et al., 2020), and conditional image synthesis (Park et al., 2019; Li et al., 2019; Liu et al., 2019b; Zhu et al., 2020b; Chen et al., 2020). The aforementioned works have constructed dedicated architectures, which require training the generator network.

Latent-Space Manipulation Recently, numerous papers have presented diverse methods to learn semantic edits of the latent code. A popular approach is finding linear directions that correspond to changes in a given binary labeled attribute, such as young $\leftrightarrow$ old, or no-smile $\leftrightarrow$ smile (Shen et al., 2020; Goetschalckx et al., 2019; Denton et al., 2019; Adbal et al., 2020). Tewari et al. (2020) utilize a pretrained 3DMM to learn semantic face edits in the latent space. Jahanian et al. (2019) find latent space paths that correspond to a specific transformation, such as zoom or rotation, in a self-supervised manner. Härkönen et al. (2020) find useful paths in an unsupervised manner by using the principal component axes (PCA) of an intermediate activation space. Finally, Collins et al. (2020) perform local semantic editing by manipulating corresponding components of the latent code.

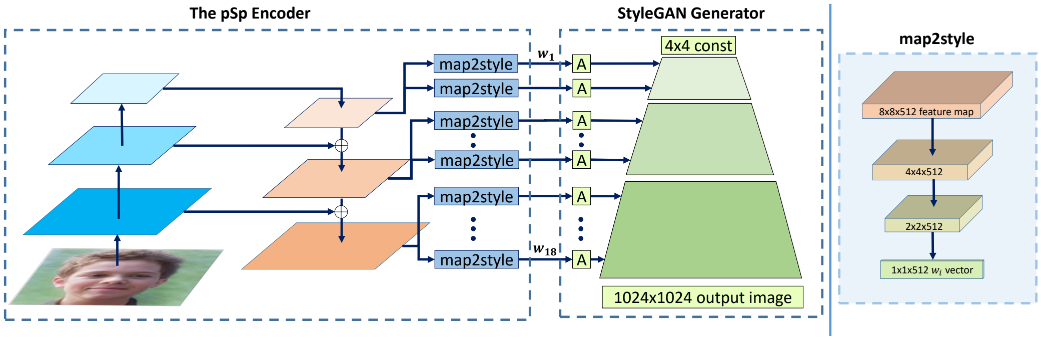


Figure 1: Our pSp architecture. Feature maps are first extracted using a standard feature pyramid over a ResNet backbone. For each of the 18 target styles, a small mapping network is trained to extract the learned styles from the corresponding feature map, where styles (0-2) are generated from the small 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.

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 therefore limiting the 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 (Lin et al., 2017), generating three levels of feature maps from which styles are extracted using a simple intermediate network — map2style — shown in Figure 1. 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. Therefore, our architecture, pSp, is an end-to-end image-to-image translation framework. The complete architecture is illustrated in Figure 1.

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,

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

In addition, to learn perceptual similarities, we utilize the LPIPS (Zhang et al., 2018) loss, which has been shown to better preserve image quality (Guan et al., 2020) compared to the more standard perceptual loss (Johnson et al., 2016):

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

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:

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

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.

The Identity Loss One of the main challenges of face generation tasks is the ability to preserve identity between the input and output images. Since identity preservation is a crucial part of face reconstruction tasks, it is important to integrate this objective into the overall loss function. Therefore, we incorporate a dedicated recognition loss measuring the cosine similarity between the output image and its source,

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

where R is a pretrained ArcFace (Deng et al., 2019) network for face recognition. The input, $\mathbf{x}$, and output, $pSp(\mathbf{x})$, are cropped around the face and resized to 112×112 before being fed into R.

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. Constants and other implementation details can be found in Appendix A.1.

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 (Richardson & Weiss, 2020). Moreover, previous works (Karras et al., 2019; Collins et al., 2020) 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 (Zhu et al., 2017b; Huang et al., 2018), 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 is illustrated in Figure 7a in Appendix A. There, layers 1-7 are selected from the input latent while layers 8-18 are taken from the sampled vector allowing one to obtain outputs with similar coarse and medium features, but varying fine features.

4 Applications and Experiments

To explore the effectiveness of our approach we evaluate our pSp framework 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 ALAE encoder (Pidhorskyi et al., 2020) and to the encoder from IDInvert (In-Domain Invert) (Zhu et al., 2020a). The ALAE method proposes a StyleGAN-based autoencoder, where the encoder is trained alongside the generator to generate latent codes. In IDInvert, real 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 with our method, we compare with IDInvert where no further optimization is performed after computing the encoding of a given image.

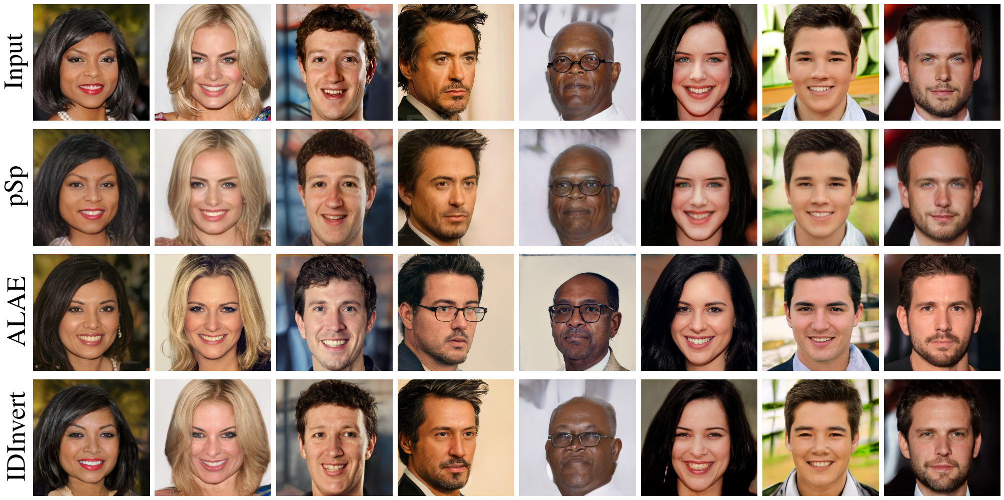
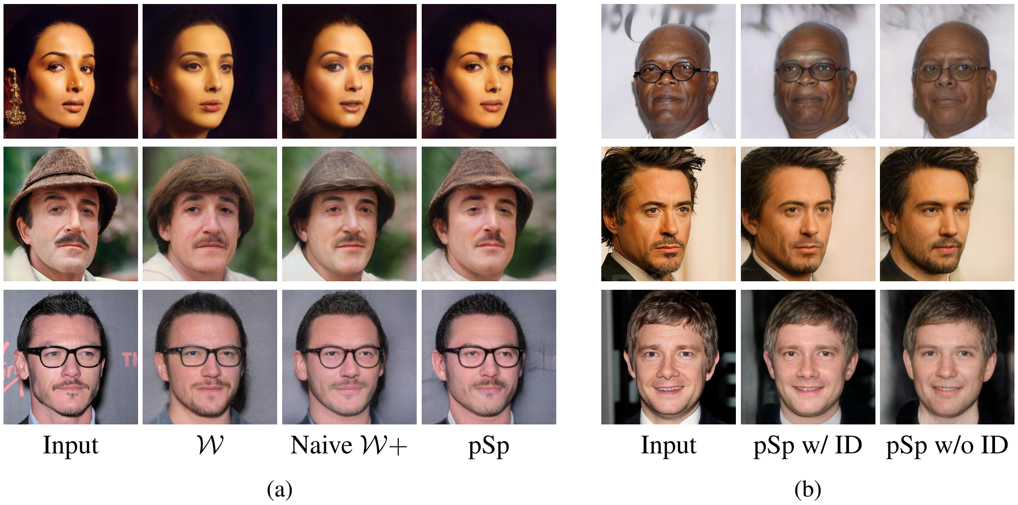


Figure 2: Results of pSp for StyleGAN inversion compared to other approaches on CelebA-HQ.



(a)

(b)

Figure 3: (a) Ablation of the pSp encoder over CelebA-HQ. (b) The importance of the identity loss.

Results Figure 2 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 (Zhu et al., 2020a) better preserves the image attributes, it still fails to accurately preserve identity and the finer details of the input image. In contrast, our 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 3a 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 3b we show the importance of the identity loss in the reconstruction task.

Finally, Table 4a presents a quantitative evaluation measuring the different encoders examined above. Our pSp model 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 (Huang et al., 2020) method for evaluation.

Method	↑ Similarity	↓ LPIPS	↓ MSE	↓ Runtime
ALAE	0.06	0.32	0.15	0.207
IDInvert	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	**0.56**	**0.17**	**0.03**	0.105

(a)

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

(b)

Figure 4: (a) Quantitative results for image reconstruction on CelebA-HQ. (b) Results for Face Frontalization on the FEI Face Database split by rotation angle of the face in the input image.

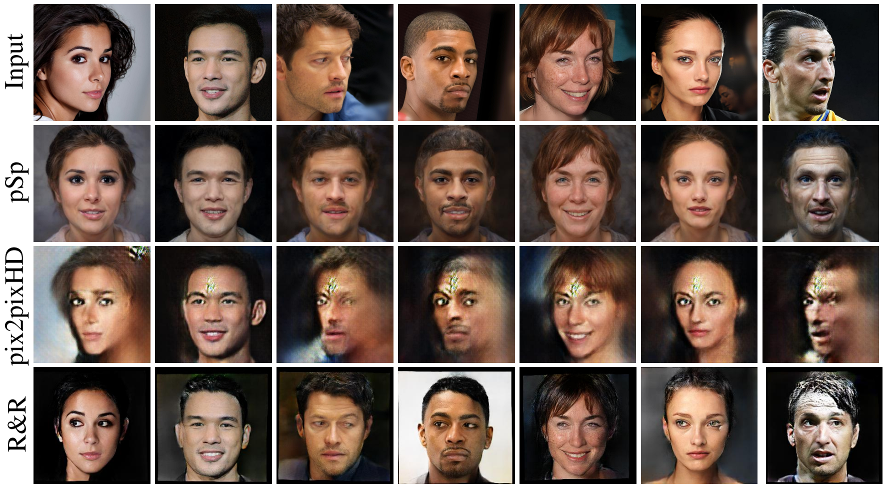


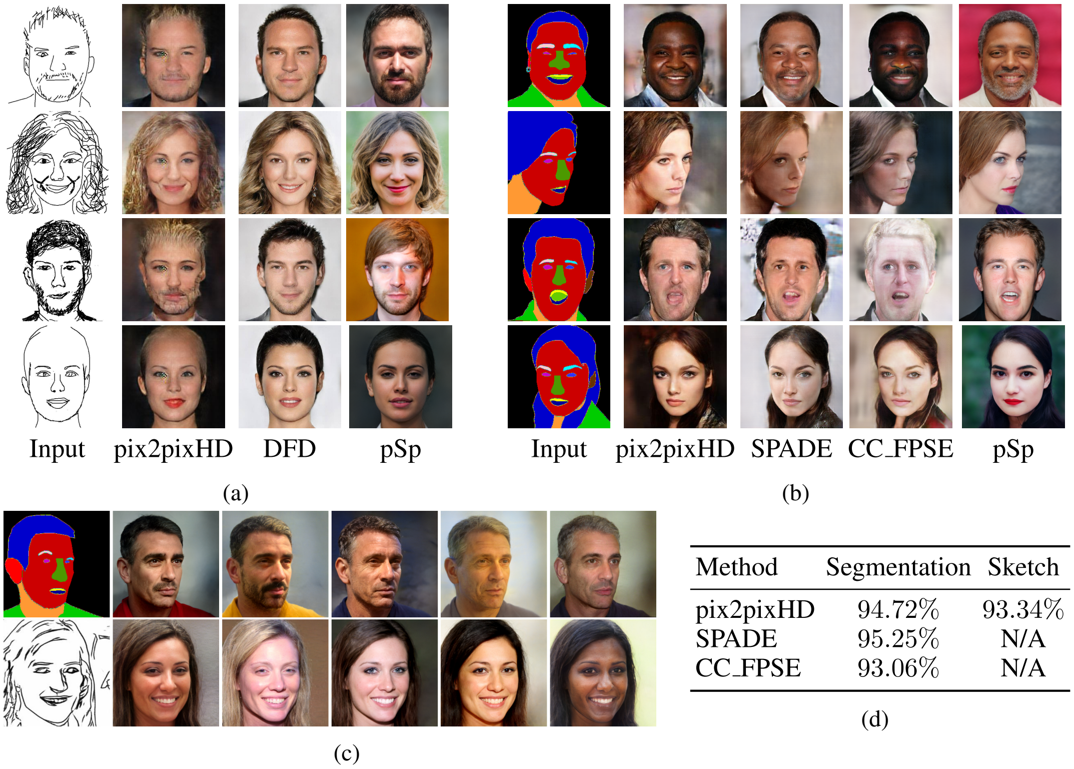
Figure 5: Comparison of face frontalization methods.

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. RotateAndRender (R&R) (Zhou et al., 2020) 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 overcome these challenges, even when trained with no labeled data.

Methodology and details For this task, training is the same as the encoder formulation with two important changes. First, we randomly flip the target image, thus creating inconsistencies in terms of pose compared to the input image. This guides the model towards generating a frontalized face, as the true target pose is unknown. While this may seem minor, without this augmentation the model would simply learn to encode the input image, matching its pose as well as identity. Next, in frontalization, as we are less interested in the background region compared to the face region and its identity, we also change the weights of the loss objective. In particular, we decrease the weights of the LPIPS and L_2 loss functions, and give more weight to the losses computed on the inner part of the face, focusing the model on the inner region while reducing the importance of background preservation. As shown below, these changes to the training objective are enough for the model to generate realistic frontal faces, while also preserving identity.

Results Results are illustrated in Figure 5. When trained with the same data, 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 RotateAndRender 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 4b provides a quantitative evaluation on the FEI Faces Database (Thomaz & Giraldi, 2010). While R&R outperforms pSp, our simple approach provides an elegant alternative, without requiring specialized alignment steps.



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

Figure 6: (a) Comparison of sketches presented in DeepFaceDrawing. (b) Comparisons to other label-to-image methods on CelebAMask-HQ. (c) Multi-modal outputs using pSp with style-mixing. (d) 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.

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 label maps. We demonstrate that, with only minimal changes, our encoder successfully utilizes the expressiveness of StyleGAN to generate high-quality and diverse outputs. Additionally, an ideal mapping framework should be able to generate multiple diverse outputs for a given input. To achieve this, we utilize the multi-modal synthesis approach described in Section 3.2.

Methodology and details The training of the two conditional generation tasks is identical to that of the encoder for StyleGAN inversion except for the omission of the identity loss and the addition of the regularization loss. To generate multiple images at inference time, we perform style-mixing, 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 sketches. DeepFaceDrawing (Chen et al., 2020) address this using a set of dedicated mapping networks. We show that pSp provides a simple alternative to past approaches.

Dataset Construction 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 A.2.

Results Figure 6a 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. Due to the hard constraints of pix2pixHD, they are unable to handle the abstract sketches and obtain poor visual results. 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). Another limitation of DeepFaceDrawing is its focus on frontal images. We therefore illustrate our model's ability to generate high-fidelity outputs from non-frontal sketches in Figure 13. As we are unable to directly evaluate DeepFaceDrawing on our constructed dataset, we compare our results only to those of pix2pixHD, trained and evaluated with the same data.

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 (Park et al., 2019), and CC_FPSE (Liu et al., 2019b), both of which are based on pix2pixHD.

Results In Figure 6b 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 are visually similar. 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 6c with additional examples in Appendix C.

4.3.3 Human Perceptual Study

We additionally perform a human evaluation to compare the visual quality of each method presented above. Here, 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 6d shows that pSp significantly outperforms the other respective methods in both synthesis tasks.

5 Discussion and Conclusions

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 7b in Appendix A presents some examples of such reconstruction failures.

In this work, we proposed a novel encoder architecture that can be used to directly map a face image into the $\mathcal{W}+$ latent space with no optimization required. The encoder architecture, motivated by StyleGAN, consists of a hierarchy of three levels that correspond to the coarse, medium, and fine groupings of the 18 style vectors defining the input in the $\mathcal{W}+$ latent space. Styles are then extracted from the encoder in a hierarchical fashion and fed into the corresponding inputs of a fixed StyleGAN generator. Notably, our network is trained with an ID similarity loss, which encourages better preservation of identity compared to previous direct approaches. Combining our encoder with a StyleGAN decoder, we present a general framework for solving various image-to-image translation tasks. In contrast to previous methods, which tackle such tasks using a local "pixel-to-pixel" approach, our framework takes a global approach, which we show can be used to solve a wide variety of image-to-image translation problems.

REFERENCES

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*, pp. 4432–4441, 2019.

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*, pp. 8296–8305, 2020.

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.

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

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.

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*, pp. 8188–8197, 2020.

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*, pp. 5771–5780, 2020.

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.

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*, pp. 4690–4699, 2019.

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

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

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.

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

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)*, pp. 172–189, 2018.

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*, pp. 5901–5910, 2020.

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*, pp. 1125–1134, 2017.

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

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

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. URL https://openreview.net/forum?id=Hk99zCeAb.

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*, pp. 4401–4410, 2019.

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*, pp. 8110–8119, 2020.

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

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 (eds.), *Computer Vision – ECCV 2012*, pp. 679–692, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33712-3.

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*, pp. 2323–2331, 2019.

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*, pp. 2117–2125, 2017.

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

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.

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*, 2019a.

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

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*, pp. 570–580, 2019b.

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.

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.

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*, pp. 2337–2346, 2019.

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

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

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

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*, pp. 9243–9252, 2020.

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. doi: 10.1145/2897824.2925972.

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.

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. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2009.11.005. URL `http://www.sciencedirect.com/science/article/pii/S0262885609002613`.

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*, pp. 8798–8807, 2018.

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

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*, pp. 9597–9608, 2019.

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*, pp. 586–595, 2018.

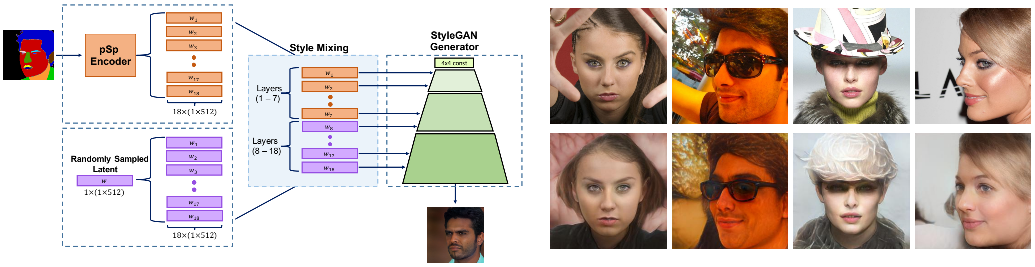
Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5911–5920, 2020.

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

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*, pp. 2223–2232, 2017a.

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*, pp. 465–476, 2017b.

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*, pp. 5104–5113, 2020b.



(a)

(b)

Figure 7: (a) To generate multiple outputs for a single input image, style-mixing is performed over pSp. (b) Challenging cases for StyleGAN Inversion.

A Additional Details

A.1 Implementation Details

Training Details For our backbone network we use the ResNet-IR architecture from (Deng et al., 2019) pretrained on face recognition, which accelerated convergence. We use a *fixed* StyleGAN2 generator trained on the FFHQ (Karras et al., 2019) dataset. That is, only the pSp encoder network is trained on the given image-to-image 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. For training, we use the Ranger optimizer, a combination of Rectified Adam (Liu et al., 2019a) with the Lookahead technique (Zhang et al., 2019), with a constant learning rate of 0.001. Only horizontal flips are used as augmentations during training. All experiments are performed using a single NVIDIA Tesla P40 GPU.

For the StyleGAN inversion task, the λ values are set as $\lambda_1 = 1$, $\lambda_2 = 0.8$, $\lambda_3 = 0.1$. For face frontalization, we increase the weight of the identity loss, setting $\lambda_3 = 1$, and decrease the LPIPS and L_2 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. we set $\lambda_3 = 0$). Finally, λ_4 is set to 0.005 in all applications except for the StyleGAN inversion task, which does not utilize the regularization loss.

A.2 Datasets

We conduct our experiments on the CelebA-HQ dataset (Karras et al., 2018), 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 (Karras et al., 2019), which contains 70,000 face images, is used for the StyleGAN inversion and face frontalization tasks.

For the generation of face images from sketches, we construct a dataset representative of hand-drawn sketches using the CelebA-HQ dataset (Karras et al., 2018). 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 Simo-Serra et al. (2016), resulting in images resembling hand-drawn sketches.

B Additional Applications

B.1 Super Resolution

Here we show that our framework can be used to construct high-resolution (HR) facial images from corresponding low-resolution (LR) input images. PULSE (Menon et al., 2020) 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. In this work we focus on applying pSp in a supervised manner as obtaining paired data is immediate. We show that our method achieves comparable results to PULSE and other previous works.

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

Results Figure 9 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 retain identity 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, we generate multiple outputs for a given LR image using our multi-modal technique by perform style-mixing on layers (4-7) with an α value of 0.5 with a randomly sampled $\mathbf{w}$ vector, which alters medium-level styles that mainly control facial features. Figure 10 illustrates the results.

B.2 Even more applications

To better show the flexibility of our pSp framework, We present three additional applications, which are summarized in Figure 8.

Local Editing Our framework allows for a simple approach to local image editing 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.

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 $w' = \lambda w_1 + (1 - \lambda) w_2$ for $0 \leq \lambda \leq 1$ and generate the corresponding image using the new code w'.

Inpainting Finally, we show the ability of our framework to reconstruct missing parts of an image using a *simple, symmetric* triangular mask. Our approach is able to accurately reconstruct the occluded areas while preserving the identity with respect to the original image.

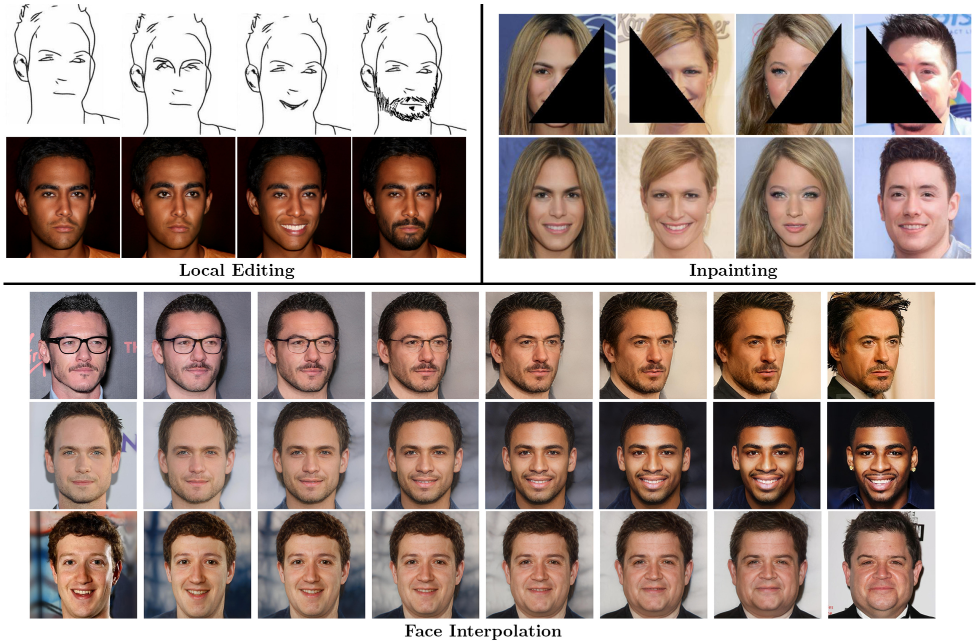
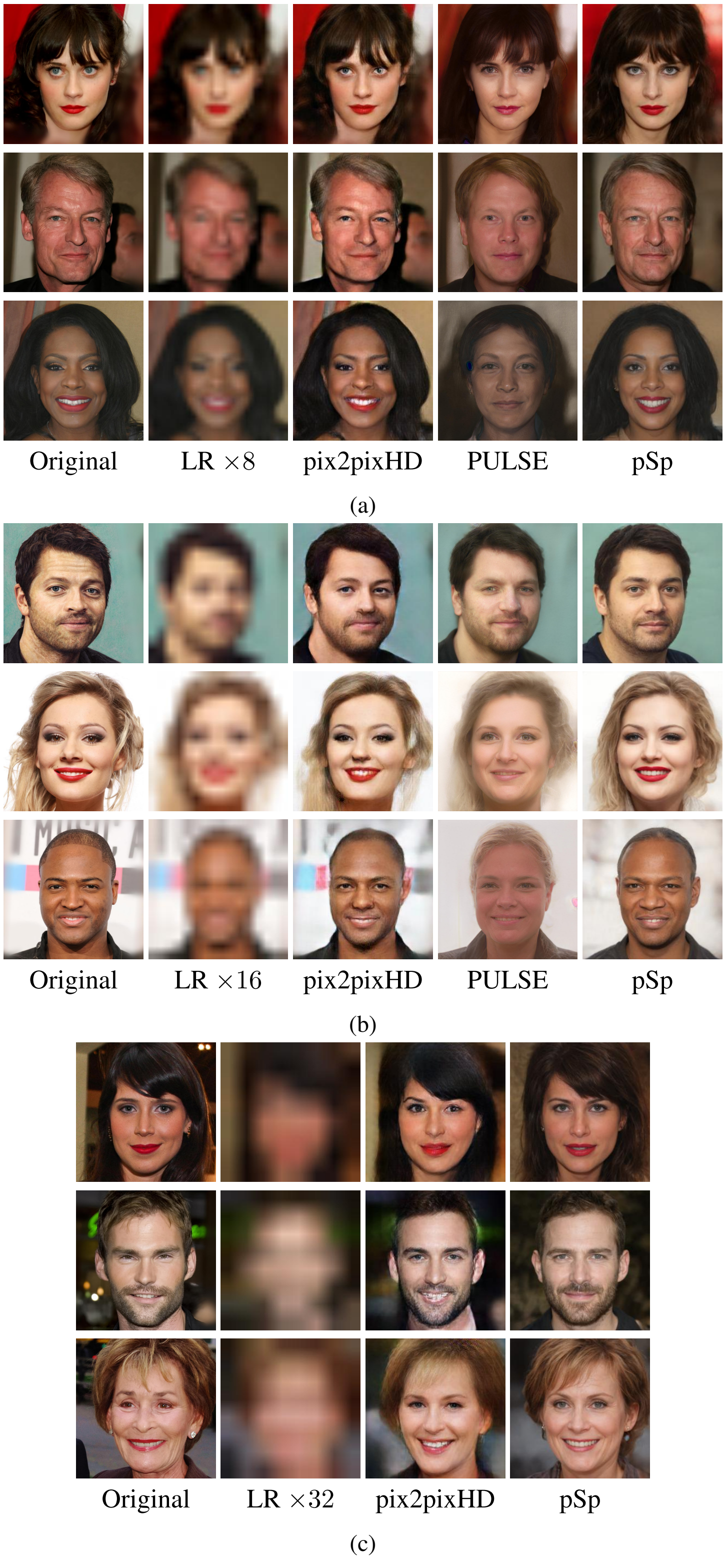


Figure 8: Additional applications for the pSp framework.



(a)

(b)

(c)

Figure 9: Comparison of super-resolution approaches with (a) ×8 down-sampling, (b) ×16 down-sampling, and (c) ×32 down-sampling.

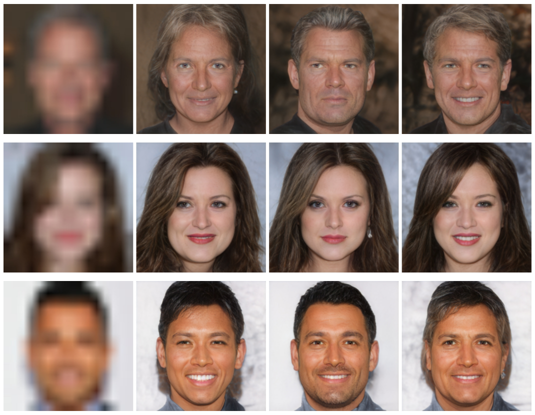


Figure 10: Multi-modal synthesis for super-resolution using pSp with style-mixing.

C Additional Results

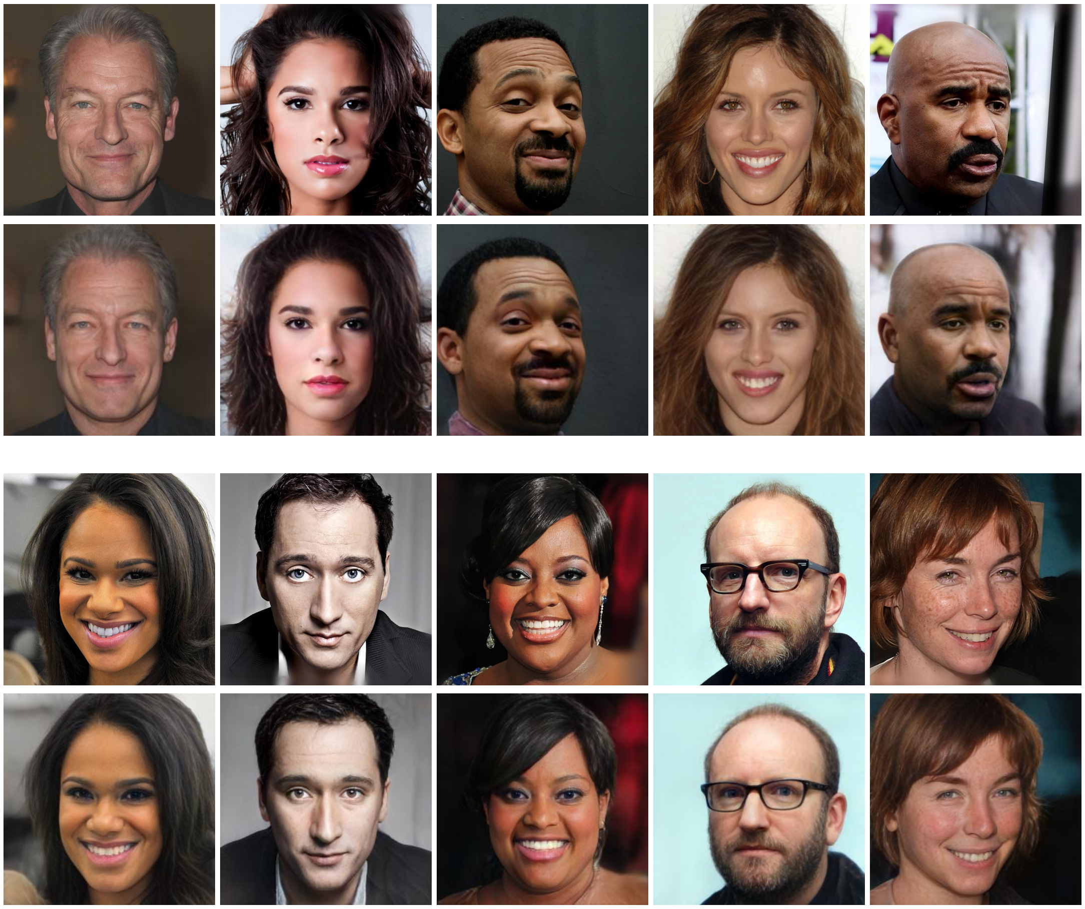


Figure 11: Additional StyleGAN inversion results using pSp on the CelebA-HQ (Karras et al., 2018) test set.

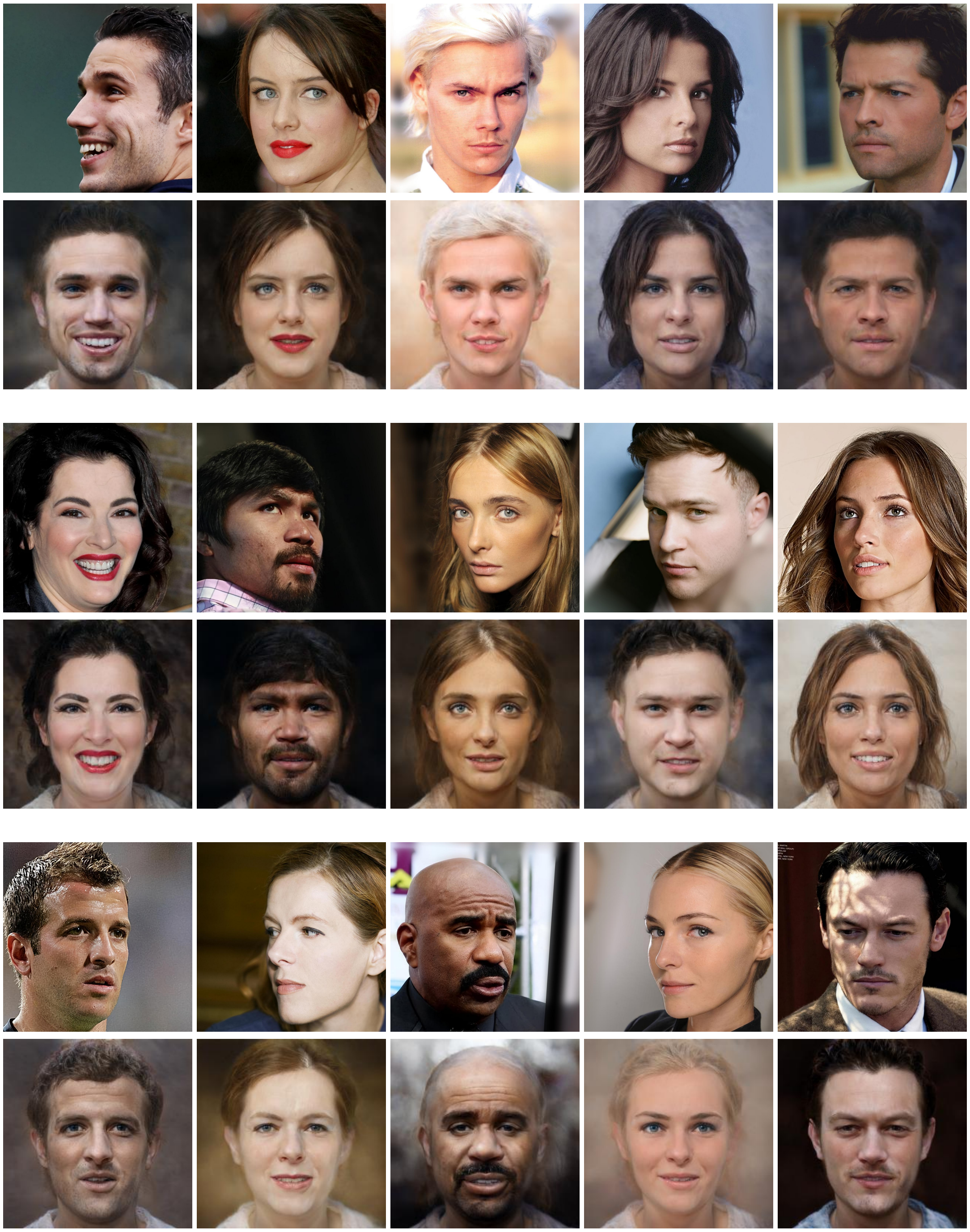


Figure 12: Additional face frontalization results using pSp on the CelebA-HQ (Karras et al., 2018) test set.

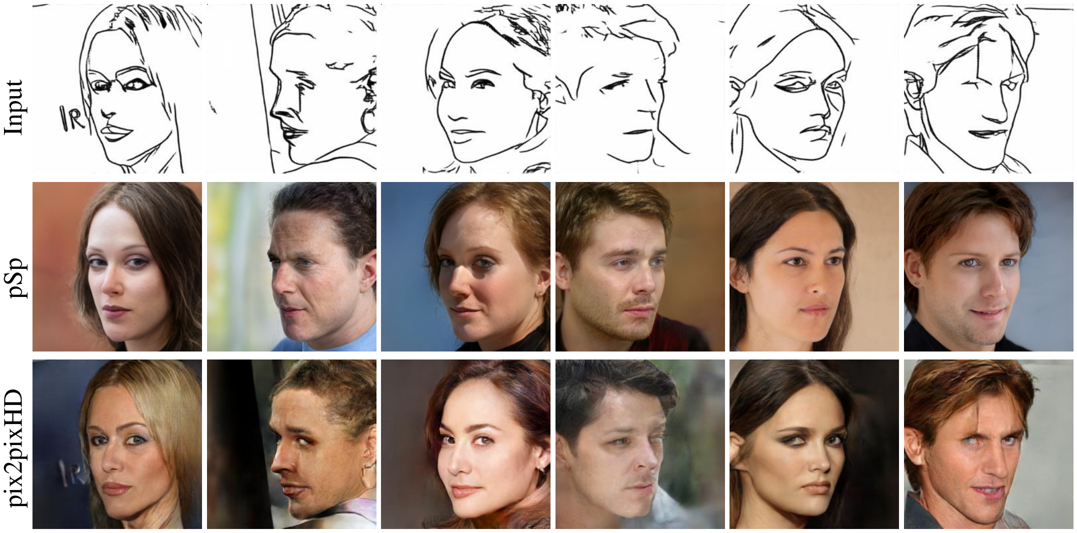


Figure 13: Even for challenging, non-frontal face sketches, pSp is able to obtain high-quality, diverse outputs.

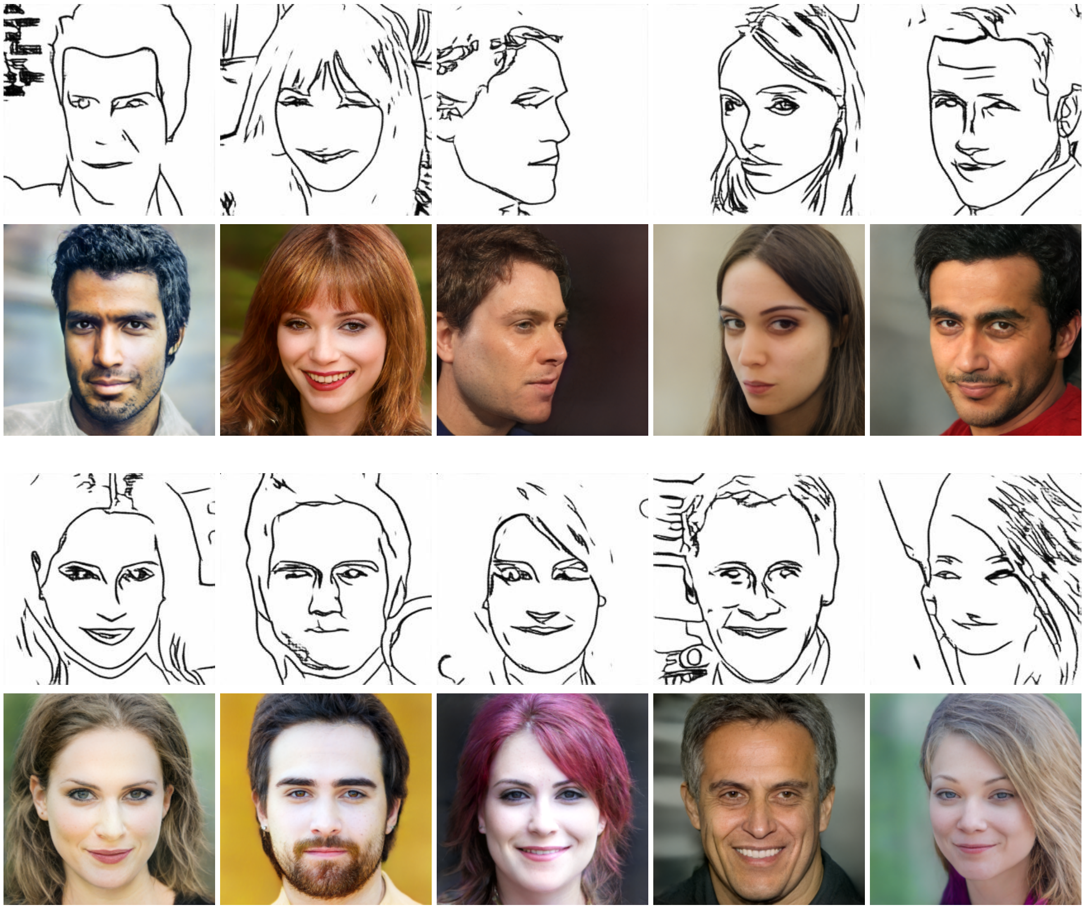


Figure 14: Additional results using pSp for the generation of face images from sketches constructed from the CelebA-HQ (Karras et al., 2018) test dataset.

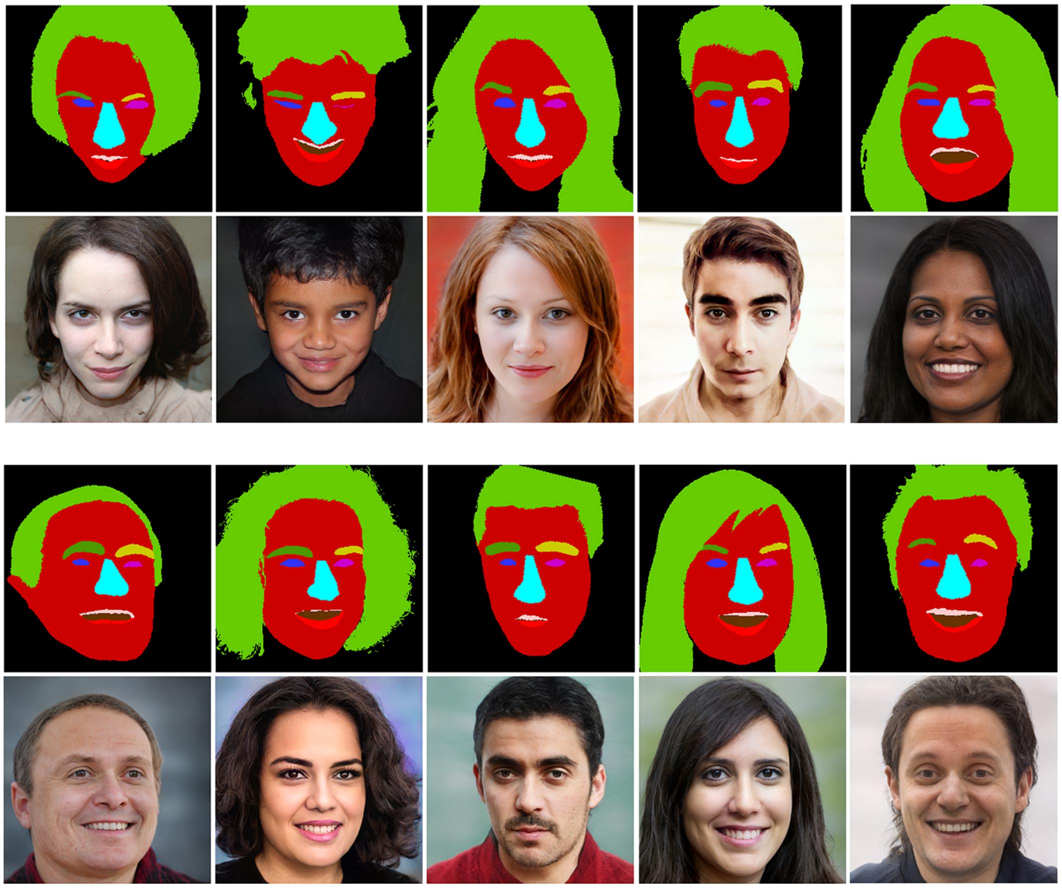


Figure 15: Additional results on the Helen Faces (Le et al., 2012) dataset using our proposed label-to-image method.

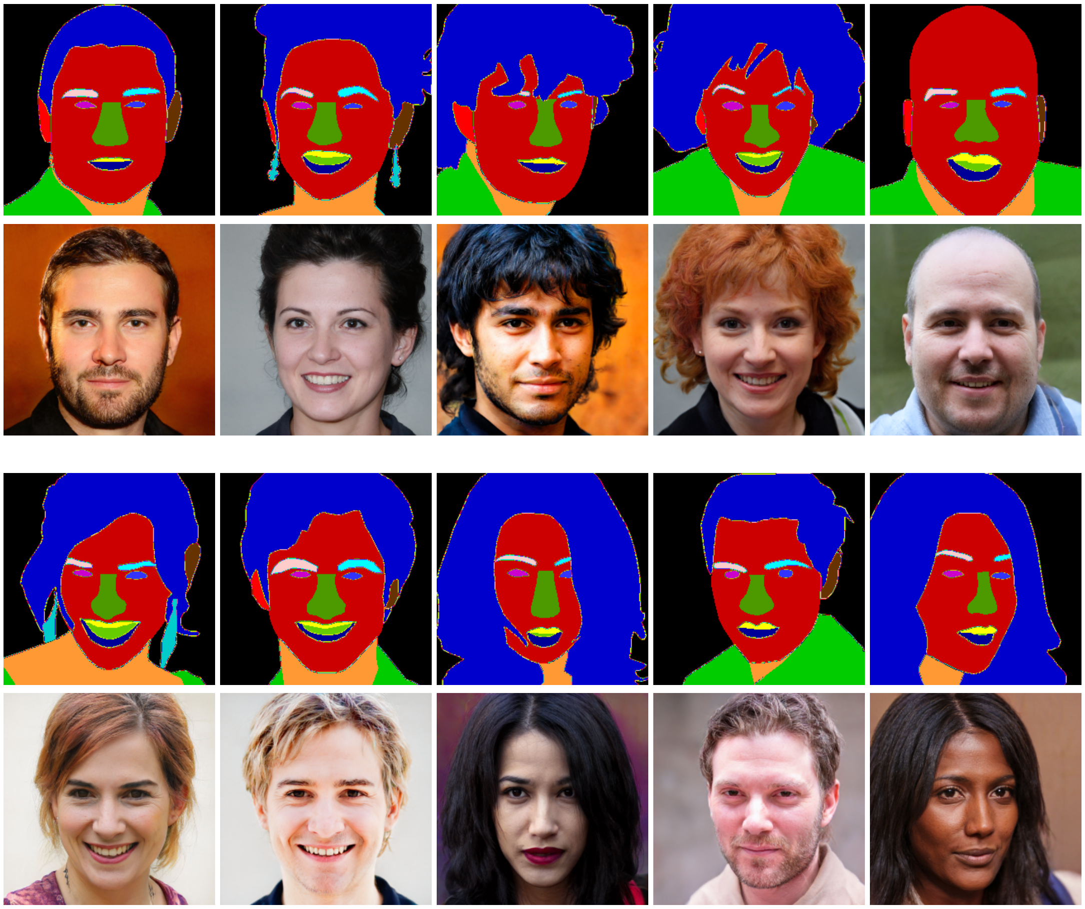


Figure 16: Additional results on the CelebAMask-HQ (Karras et al., 2018) test set using our proposed label-to-image method.

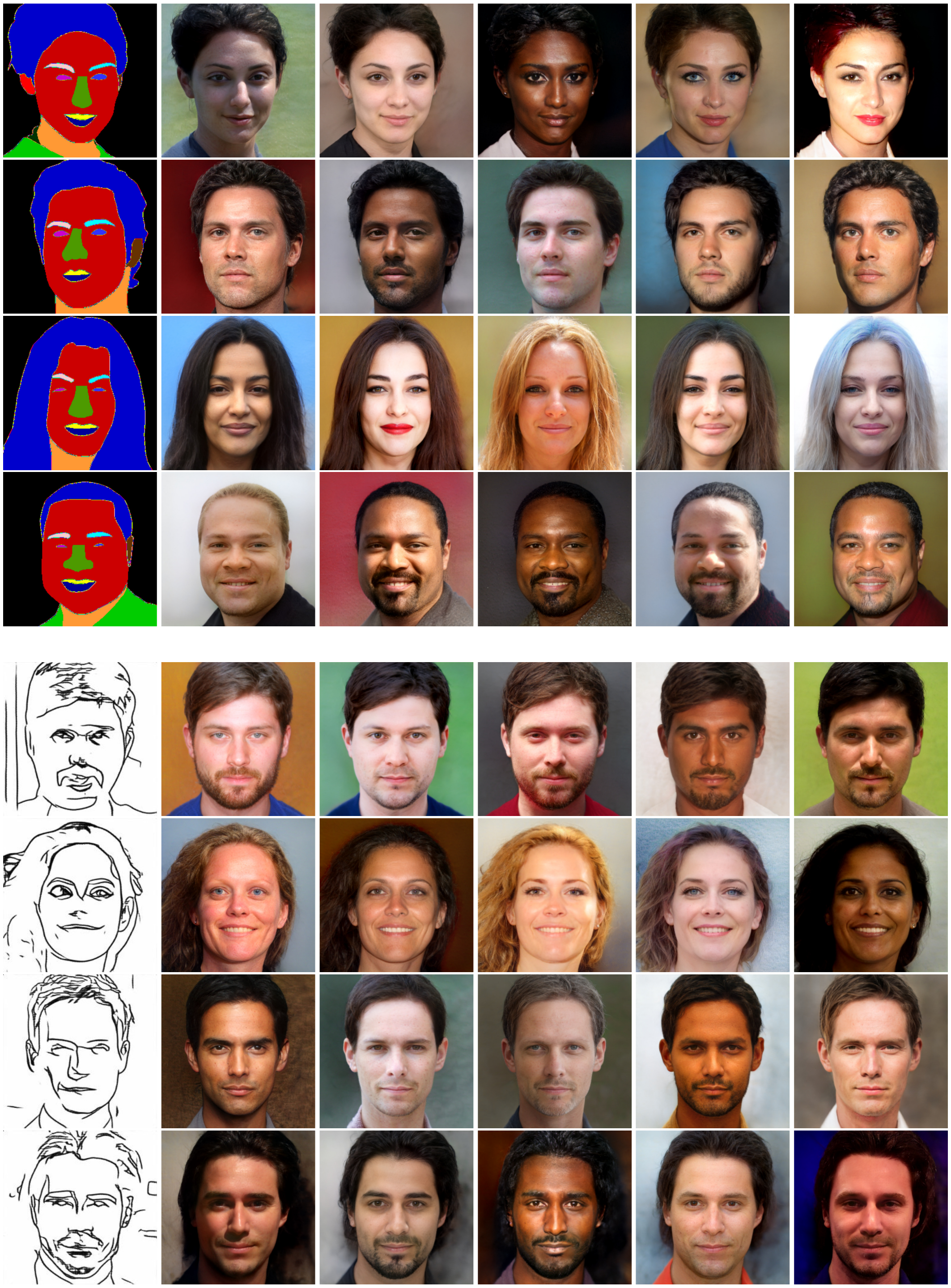


Figure 17: Conditional image synthesis results from sketches and segmentation maps displaying the multi-modal property of our approach.