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

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

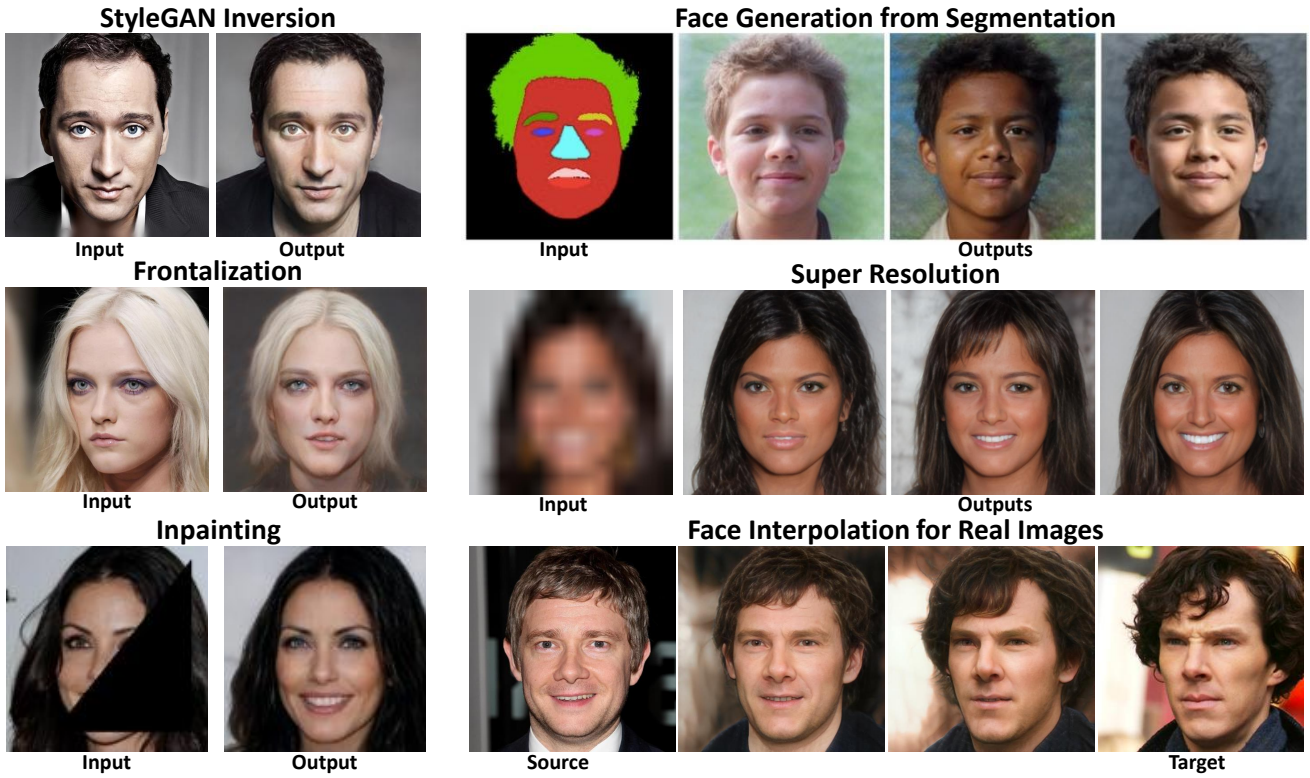

Figure 1: Our pSp framework can be used for a wide variety of image-to-image problems.

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 without any labeled data, generate multi-modal results for ambiguous tasks such as conditional face generation from segmentation maps, and construct high-resolution images from corresponding low-resolution images.

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 [22, 23] proposes a novel style-based generator architecture and attains state-of-the-art vi-

sual quality on high-resolution images. Moreover, it has been demonstrated that it has a disentangled latent space, $\mathcal{W}$ [8, 39, 44], obtained from the initial latent space $\mathcal{Z}$ via a Multi-Layer Perceptron (MLP) mapping network, which may offer control and editing capabilities.

Recently, numerous methods have shown competence in controlling StyleGAN's latent space and performing meaningful manipulations in $\mathcal{W}$ [39, 18, 41, 14, 3]. 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 $w \in \mathcal{W}$ does not lead to an accurate reconstruction. Motivated by this, it has become common practice [1, 2, 49, 4] to encode real images into an extended latent space, $\mathcal{W}+$, defined by the concatenation of 18 different 512-dimensional w vectors, one for each input layer of StyleGAN. Nevertheless, encoding into $\mathcal{W}+$ is difficult and many methods resort to using a per-image optimization over $\mathcal{W}+$, requiring several minutes for a single image. To accelerate this optimization process, some methods [49, 4] 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 at the retrieval of 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 [27], 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, see Figure 1. Focusing on face images, we first demonstrate our method's ability to successfully reconstruct a given image while preserving the 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 [17] and define a generic framework able to solve a wide range of image-to-image 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 the residual feature maps from the encoder [17, 42], creating a strong locality bias [38]. 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 in a fully unsupervised manner to align a given face image to a frontal pose with a neutral expression. Another notable advantage of the intermediate style representation is the inherent support of multi-modal synthesis for ambiguous tasks such as face generation from segmentation maps or low-resolution images. In such tasks, the generated styles can be resampled to create variations of the output image without any change to the architecture or training process.

The main contributions of this paper are:

- A novel StyleGAN encoder able to directly encode real face images into the $\mathcal{W}+$ latent domain.
- 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 for this task [22, 23]. Generally, inversion methods either directly optimize the latent vector to minimize the error for the given image [1, 2, 9, 28], train an encoder to map the given image to the latent space [36, 9, 13], or use a hybrid approach combining both [4, 49]. Typically, methods performing optimization are superior in reconstruction quality to a learned encoder mapping, but are costly and require a substantially longer time.

Focusing on the more general task of *latent space embedding*, Nitzan *et al*. [34] trained an encoder to infer a latent vector from which StyleGAN can directly generate an image with the identity of one image and the pose, expression, and illumination of another. While this shows the potential of latent embedding, their method solves only a specific application and cannot be used to solve other image-to-image translation tasks.

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*. [17] 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 [42], unsupervised learning [31, 50, 24, 29], multi-modal image synthesis [7, 15, 51], multi-domain image synthesis [6, 7], and conditional image synthesis [35, 53, 32, 5, 26].

The aforementioned works have constructed dedicated architectures for their tasks which require training the generator network. This is in contrast to our method that utilizes a fixed pretrained StyleGAN generator, enjoying its state-of-the-art image quality.

Latent-Space Manipulation Recently, the use of the latent spaces of pretrained GANs has surged as an alternative approach for solving image-to-image tasks. This approach is motivated by the image quality generated by unconditional GANs and the high semantics of their latent space, all without the heavy requirement of training the generator.

All approaches generally follow the same procedure which we call *Invert and Edit*. Specifically, to control the generated image, these methods take a two-step test-time approach. First, inverting a given image into the latent space, then editing the inverted latent code in a semantically meaningful manner to obtain a new code used by the unconditional GAN to generate the output image.

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 [39, 12, 11, 3]. Tewari *et al*. [41] utilize a pretrained 3DMM to learn semantic face edits in the latent space. Jahanian *et al*. [18] find latent space paths that correspond to a specific image transformation, such as zoom or rotation in a self-supervised manner. Hrknen *et al*. [14] find useful paths in a completely unsupervised manner by using the principal component axes (PCA) of an intermediate activation space. Abdal *et al*. [3] learn a transformation between vectors in $\mathcal{W}+$, modifying a set of predetermined labeled attributes. Finally, Collins *et al*. [8] perform local semantic editing by manipulating corresponding components of the latent code.

The aforementioned approaches suffer from a few critical issues. First and foremost, the input image must be invertible, i.e., there must exist a latent code that reconstructs the image. Therefore, the input image domain must typically be the same domain the GAN was trained on. This requirement is a severe limitation for many image-to-image tasks that translate between two different data domains such as segmentation maps, sketches, edges, etc. This limitation may be bypassed by optimizing to an enlarged space, such as $\mathcal{W}+$ [1, 2, 49]. However, in this case, the latent space does not contain rich semantics for an unknown data domain. For example, when translating between sketches and natural face images, even if the input sketches could be inverted into $\mathcal{W}+$ of a StyleGAN trained on faces, the latent code would not be semantically meaningful. Second, GAN inversion remains difficult. In particular, while direct encoding methods have achieved limited success in reconstruction quality, optimization and hybrid methods are costly and require significant time to converge.

Inspired by this realization, our end-to-end pSp framework directly encodes a given real image into the desired latent vector, thus enabling one to directly solve the image-to-image task while enjoying the benefits of a pretrained StyleGAN without incurring unnecessary limitations.

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 [27], 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. Therefore, our architecture, pSp, is an end-to-end image-to-image translation framework. The complete architecture is illustrated in Figure 2. We note that while we found the feature pyramid to best match the StyleGAN architecture, other possible variations could also work. For example, generating all the style vectors from the largest feature map would mostly affect the model size without hindering model accuracy. Conversely, generating the style vectors from the smallest feature map is also feasible without limiting performance as long as its dimensionality is large enough.

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

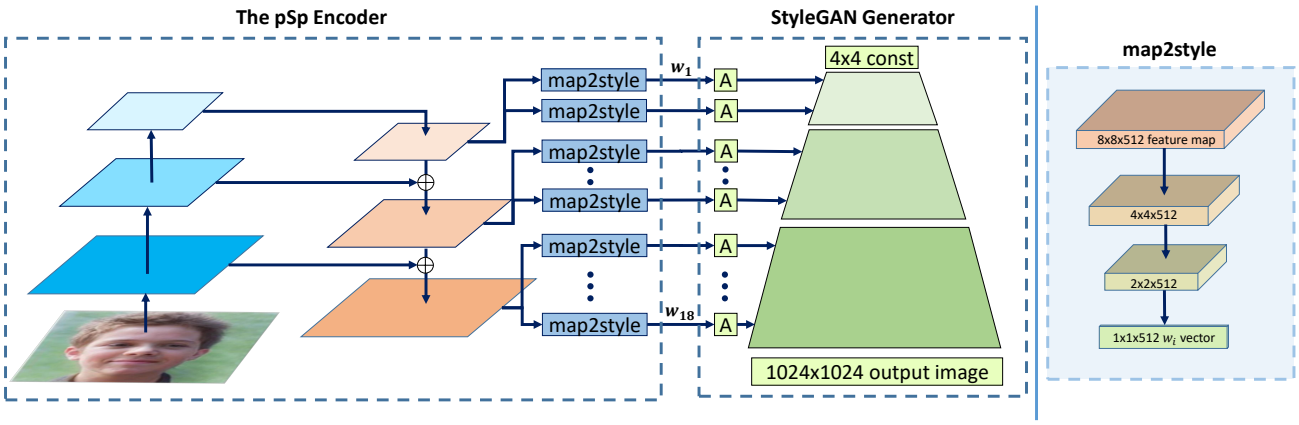

Figure 2: Our pSp architecture. Feature maps are first extracted using a standard feature pyramid over a ResNet backbone. Then, for each of 18 target styles, a small mapping network is trained to extract the learned styles from the corresponding feature map, where styles $(0-2)$ are generated from the 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 then fed into StyleGAN, starting from its matching affine transformation, denoted A.

is trained using a weighted combination of several objectives. First, we utilize the pixel-wise $\mathcal{L}_2$ loss,

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

where $\mathbf{x}$ denotes the input image and $pSp(\mathbf{x}) = G(E(\mathbf{x}))$ is the output returned by pSp defined by the encoder network, $E(\cdot)$, and generator network, $G(\cdot)$. In addition, to learn perceptual similarities, we utilize the LPIPS [46] loss, which has been shown to better preserve image quality [13] compared to the more standard perceptual loss [19]. Formally,

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

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

3.1.1 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, as the aforementioned loss functions are less sensitive to the preservation of facial identity, we incorporate a dedicated recognition loss measuring the cosine similarity between the output image and its source,

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

where R is a pretrained ArcFace [10] network for face recognition. The input image, $\mathbf{x}$, and corresponding generated image, $pSp(\mathbf{x})$, are cropped around the face and resized to 112×112 before being fed into R.

In summary, the total loss function is defined as

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

where λ_1, λ_2, and λ_3 are constants defining the loss weights.

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 [38].

Moreover, previous works [22, 8] 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 image-to-image translations 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, 15, 52], our framework inherently supports this by simply sampling style vectors. In practice, this is done by first randomly sampling a vector $w \in \mathbb{R}^{512}$ and generating a corresponding latent code in $\mathcal{W}+$. We then perform style mixing between the randomly generated latent code, w_R, and the computed latent code of the input image, w_I, by inserting select layers of w_R into

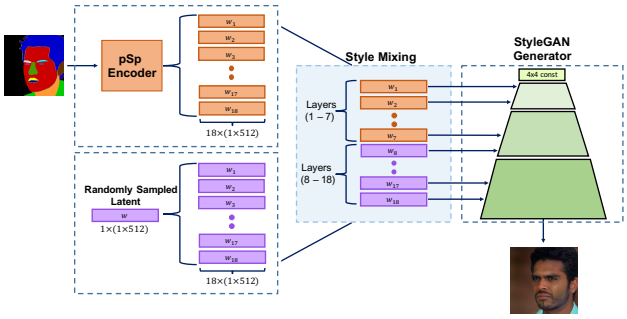

Figure 3: To generate multiple outputs for a single input image, style-mixing is performed over pSp.

the corresponding layers of w_I, possibly with some α parameter for blending between the two styles. This approach is illustrated in Figure 3. There, layers $(1-7)$ are selected from the latent code of the input image while layers $(8-18)$ are taken from the sampled vector, allowing one to obtain multiple outputs with similar coarse and medium features, but varying fine features.

3.3. Implementation Details

For our backbone network we use the ResNet-IR architecture from [10] pretrained on face recognition, which accelerated convergence. We use a *fixed* StyleGAN2 generator trained on the FFHQ [22] dataset. That is, only the pSp encoder network is trained on the given image-to-image translation task. 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 [30] with the Lookahead technique [45], with a constant learning rate of 0.001. Only horizontal flips are used as augmentations during training. Unless stated otherwise, the λ values are set as $\lambda_1 = 1$, $\lambda_2 = 0.8$, and $\lambda_3 = 0.1$. All experiments are performed using a single NVIDIA Tesla P40 GPU.

4. Applications and Experiments

To explore the effectiveness of our approach we evaluate our pSp framework on numerous image-to-image translation tasks including face frontalization, conditional face synthesis, and super-resolution. As the same framework is used for all these tasks there are many common elements between them.

Datasets We conduct our experiments on the CelebA-HQ dataset [20], 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 [22], which contains 70,000 face images, is also used for the StyleGAN inversion task.

Baselines To validate our method we compare it to the general image-to-image translation framework of pix2pixHD [42], which we train for each task with the same data used by our method. Additionally, we compare to state-of-the-art methods for each specific task, showing that our method can achieve comparable, or better, results.

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 [37] and to the encoder from IDInvert (In-Domain Invert) [49]. 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.

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 [49] 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 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 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 that when training the reconstruction task using pSp without the additional identity loss, the identity is not well-preserved. However, by incorporating the identity loss, identity is preserved across the reconstruction task.

Finally, Table 1 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 that the similarity score is independent from our loss function, we measure it using the state-of-the-art Curricularface [16] method.

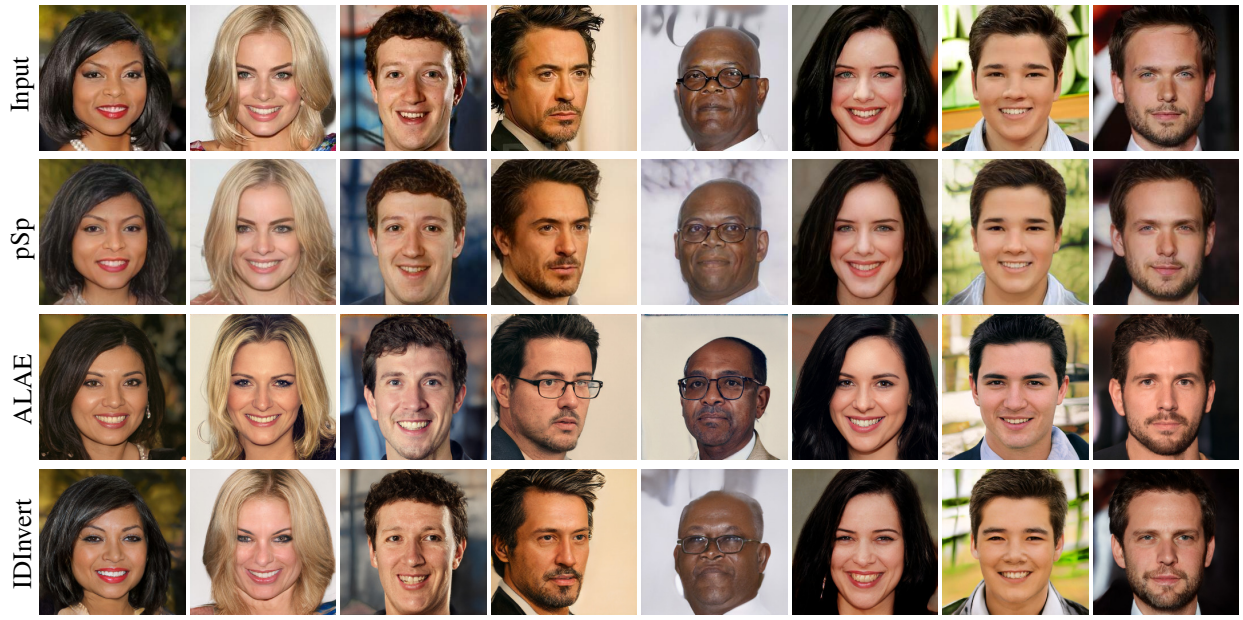

Figure 4: Results of our pSp framework for StyleGAN inversion compared to other approaches on CelebA-HQ.

Method	↑ Similarity	↓ LPIPS	↓ MSE	↓ Runtime
ALAE [37]	0.06	0.32	0.15	0.207
IDInvert [49]	0.18	0.22	0.06	**0.032**
$\mathcal{W}$ Encoder	0.09	0.31	0.11	0.063
Naive $\mathcal{W}+$	0.48	0.23	0.06	0.063
pSp	**0.58**	**0.19**	**0.04**	0.106

Table 1: Quantitative results on CelebA-HQ.

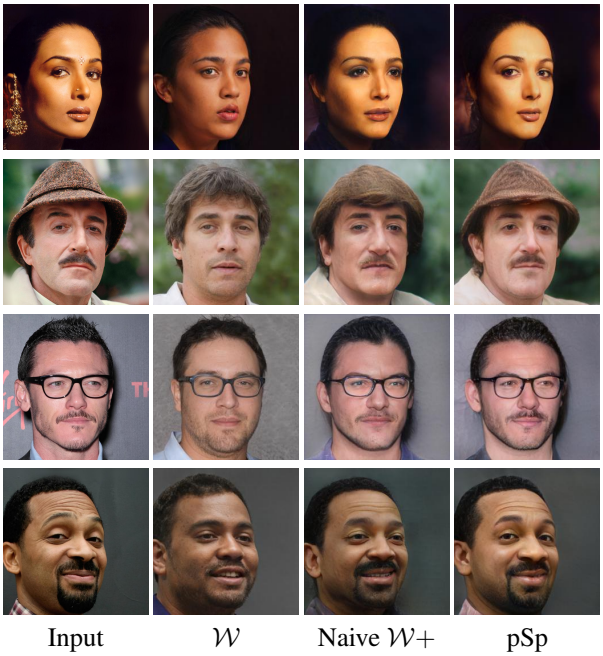

Figure 5: Ablation of the pSp encoder over CelebA-HQ.

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) [48] 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 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, focusing the model on the inner region while reducing the

Figure 6: The importance of identity loss.

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 7. One can see that when trained with the same data, the pix2pixHD method 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 Rotate-AndRender 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.

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. Specifically, given an input face sketch or label map, we wish to generate a corresponding *high-fidelity*, natural face image *semantically aligned* with the input. We demonstrate that, with only minimal changes, our encoder successfully utilizes the expressive power of StyleGAN to generate high-quality and diverse outputs from a given face sketch or semantic label.

Additionally, as conditional image synthesis is a one-to-many mapping, 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.

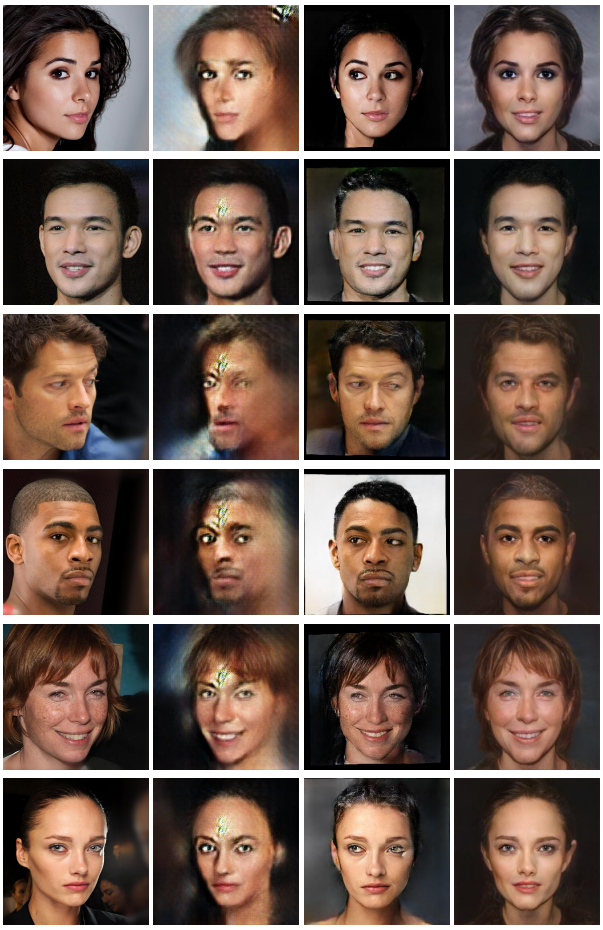

Figure 7: Comparison of face frontalization methods.

Methodology and details Due to the one-to-many mapping described above, 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. In particular, only the LPIPS and $\mathcal{L}_2$ loss functions are utilized for training. 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 w vector.

4.3.1 Face From Sketch

In this section, we explore the task of generating high-quality face images from input sketches. Common approaches to this image translation task incorporate hard constraints that require pixel-wise correspondence between the input sketch and generated image, resulting in outputs that strictly align with the input. These approaches are therefore ill-suited for the sketch-to-image task when inputs are incomplete, hand-drawn sketches. DeepFaceDrawing [5] ad-

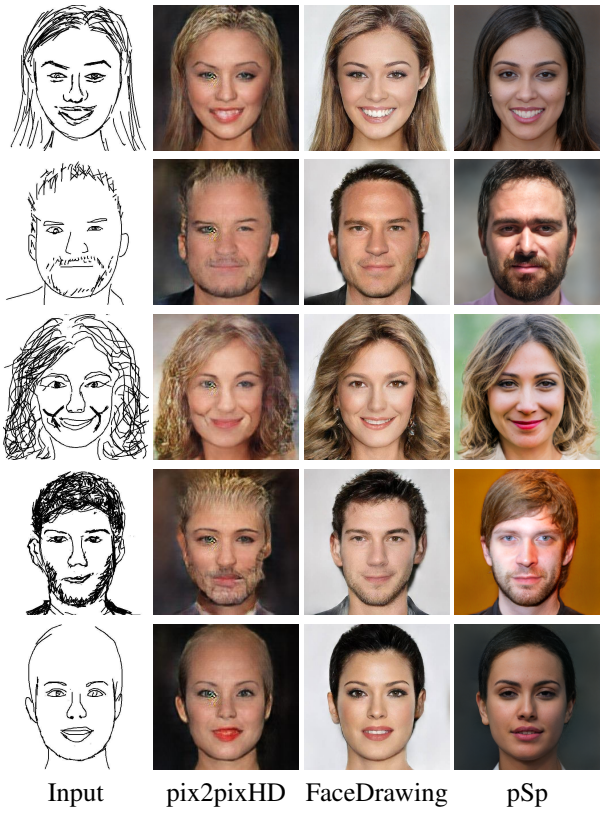

Figure 8: Comparison of sketches presented in DeepFaceDrawing [5].

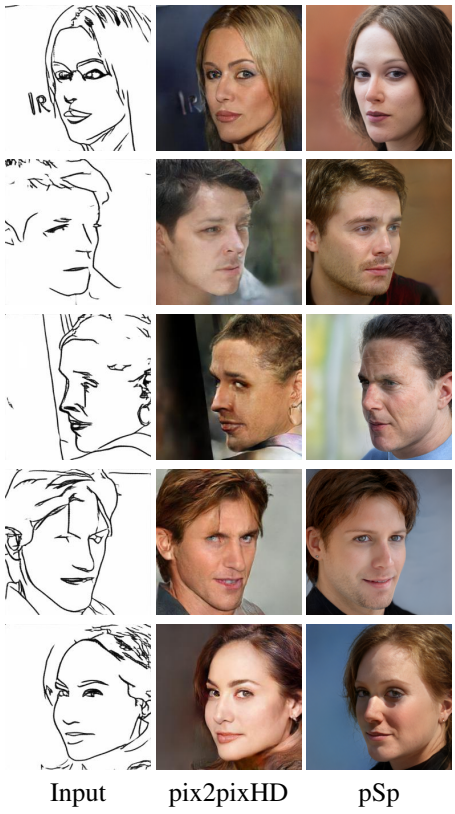

Figure 9: Even for challenging, non-frontal face sketches, pSp is able to obtain high-quality, diverse outputs.

dress this challenge using a set of dedicated mapping networks. We show that pSp provides a simple method for approaching this image translation task that compares favorably to these previous approaches.

Dataset Construction There are several datasets such as the CUHK face sketch database [43, 47] that are commonly used for the task of sketch-to-image synthesis. However, such datasets involve highly-detailed sketches, instead of the more challenging sparse sketches. We therefore elect to use a similar approach to that used by DeepFaceDrawing [5] in order to construct a dataset that is more representative of hand-drawn sketches. 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*. [40] resulting in images resembling hand-drawn sketches. Similar to DeepFaceDrawing, we elect to generate our sketches using the CelebA-HQ dataset [21]. Both pSp and pix2pixHD are then trained using this dataset.

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 the sketches and results published in their paper. One can see that 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. 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 9. As we are unable to directly evaluate DeepFaceDrawing using our constructed dataset, we compare our results only to those of pix2pixHD, trained and evaluated with the same data. Finally, Figure 16 illustrates additional synthesized faces using our pSp framework.

4.3.2 Face from Segmentation Map

Here, we evaluate using our framework for synthesizing photo-realistic face images from semantic segmentation

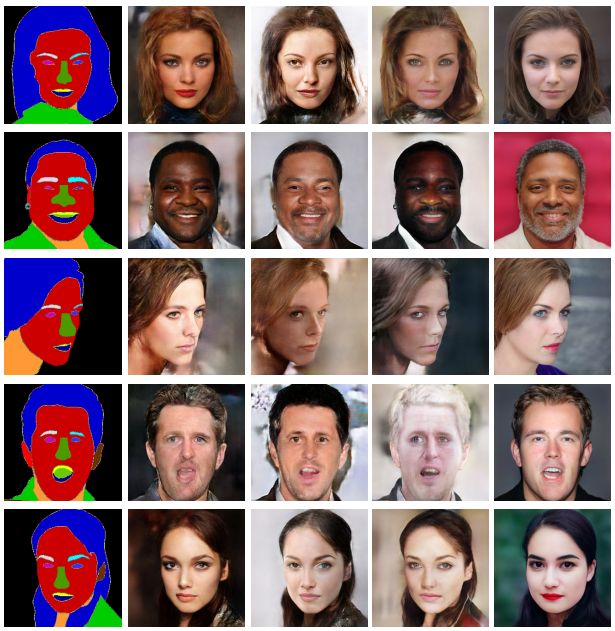

Figure 10: Comparisons to other label-to-image methods on CelebAMask-HQ.

maps. This image translation task is closely related to the generation of images from sketches. While semantic maps provide more details than abstract hand-sketches, pixel-wise approaches for solving this task are still limited in their ability to generate high-quality, diverse face images. We therefore explore pSp as an alternative to these methods. In addition to pix2pixHD, we compare our approach to two additional state-of-the-art label-to-image translation methods: SPADE [35], and CC_FPSE [32], both of which are based on pix2pixHD. We use the official implementation for each method to evaluate the approach. For a fair comparison, we train the competing methods on the same training set and report results on the same test set as our approach.

Results In Figure 10 we provide a qualitative comparison of the competing approaches on the CelebAMask-HQ dataset containing 19 semantic categories. As the three competing methods are all based on pix2pixHD, the results of all three methods are visually similar. Conversely, based on StyleGAN, our approach is able to generate high-quality outputs across a wide range of inputs including various poses and expressions. Furthermore, to show the generalization of our approach to multiple datasets, we train and evaluate our segmentation-to-image model on the Helen Faces [25] dataset, containing 11 categories, and provide qualitative results in Figure 17. Finally, using our multi-modal technique, pSp can easily generate various possible outputs for a single input. This allows us to obtain multiple outputs with the same pose and attributes but with varying fine styles. We provide examples in Figure 12.

4.4. 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 [33] approaches this task in an unsupervised manner. Specifically, for a given LR input image, PULSE traverses the HR image manifold in search of an image that *downscales* to the original LR image. Although PULSE takes an unsupervised approach to this problem, in this work we focus on applying pSp in a supervised manner for solving this task as obtaining paired data is immediate. We show that our method achieves comparable results, especially with respect to identity preservation.

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

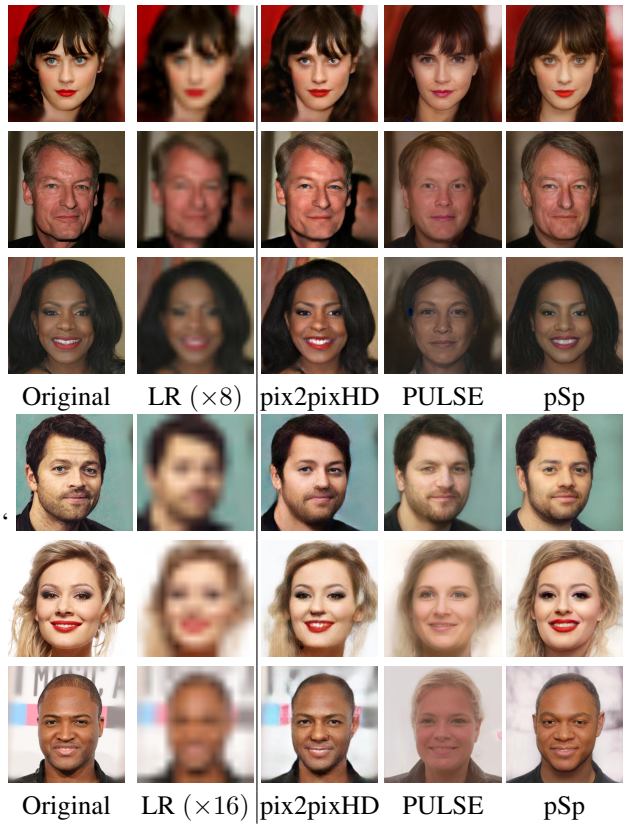

Figure 11: Visual comparison of super resolution results on CelebA-HQ.

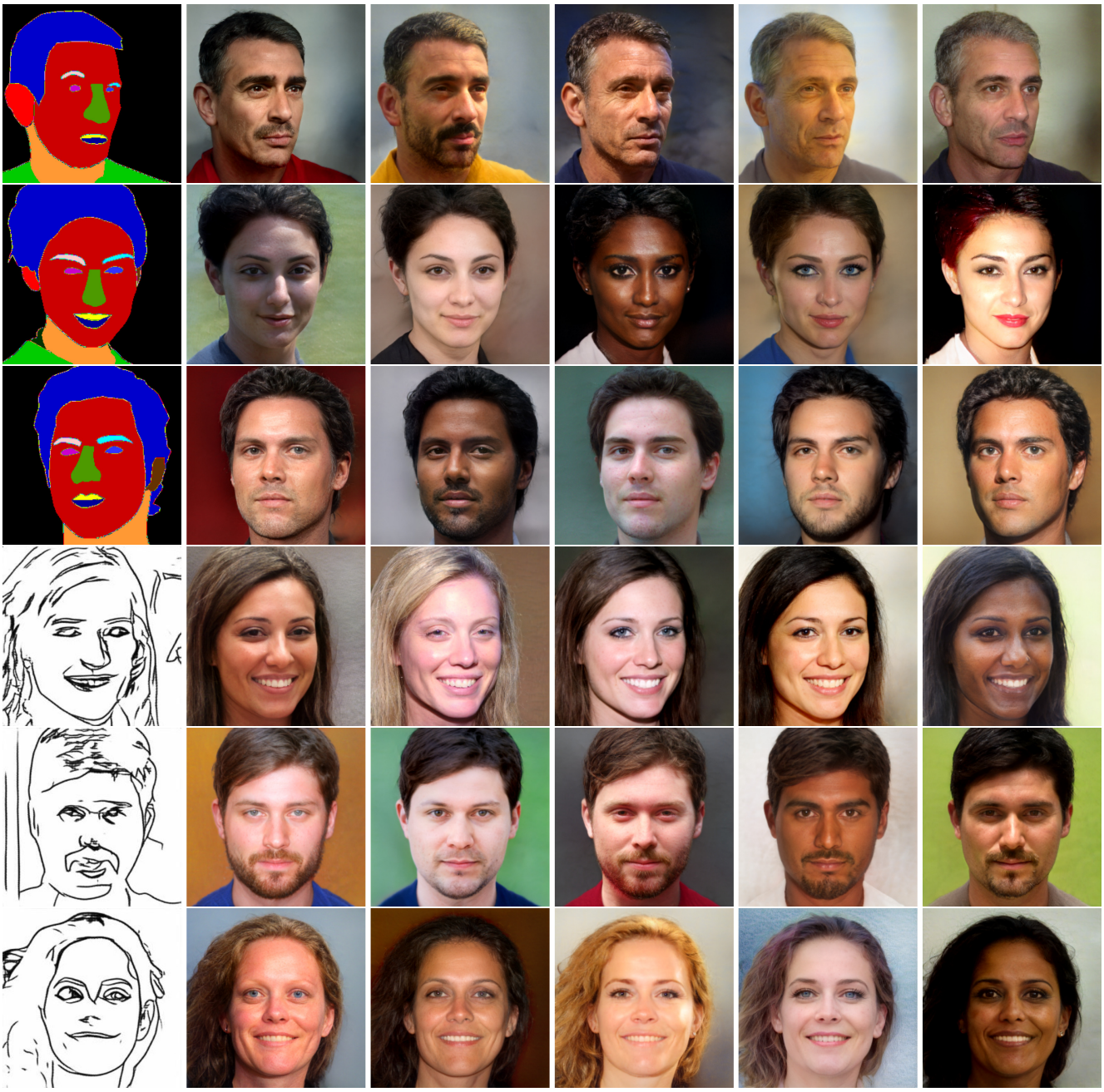

Figure 12: Conditional image synthesis results from sketches and segmentation maps displaying the multi-modality property of our approach.

Results Figure 11 demonstrates the visual quality of the resulting images from our method along with those of the previous approaches. First, by learning a pixel-wise correspondence between the LR and HR images, pix2pixHD is able to obtain good 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. 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 $\times 8$. Contrary to these previous works, we are able to obtain high-quality, photo-realistic images while successfully preserving identity, even when down-sampling by $\times 16$. Finally, as the mapping of a low-resolution image to a high-resolution image is a one-to-many mapping, there are many plausible output images for

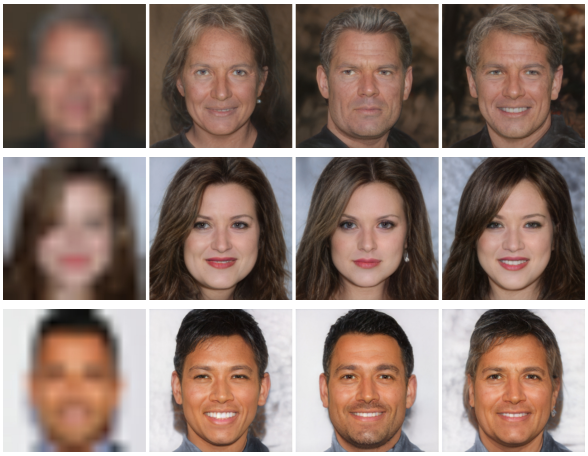

Figure 13: By performing style-mixing we are able to obtain multiple realistic high-resolution images for a single low-resolution input image.

a given low-resolution image that we generate using our multi-modal technique. In particular, we perform style-mixing on layers $(4-7)$ with an α value of 0.5 with a randomly sampled w vector, which alters medium-level styles that mainly control facial features. Figure 13 illustrates the results.

4.5. Additional Applications

To better show the flexibility of our pSp framework, We present three additional applications, which are summarized in Figure 15.

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.

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, as mentioned in previous sections, our method takes a global approach and does not utilize locality. While this is advantageous for many tasks, it does introduce a challenge in preserving fine 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 [17]. Figure 14 presents some examples of such reconstruction failures.

6. Conclusion

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.

Figure 14: Challenging cases for StyleGAN Inversion.

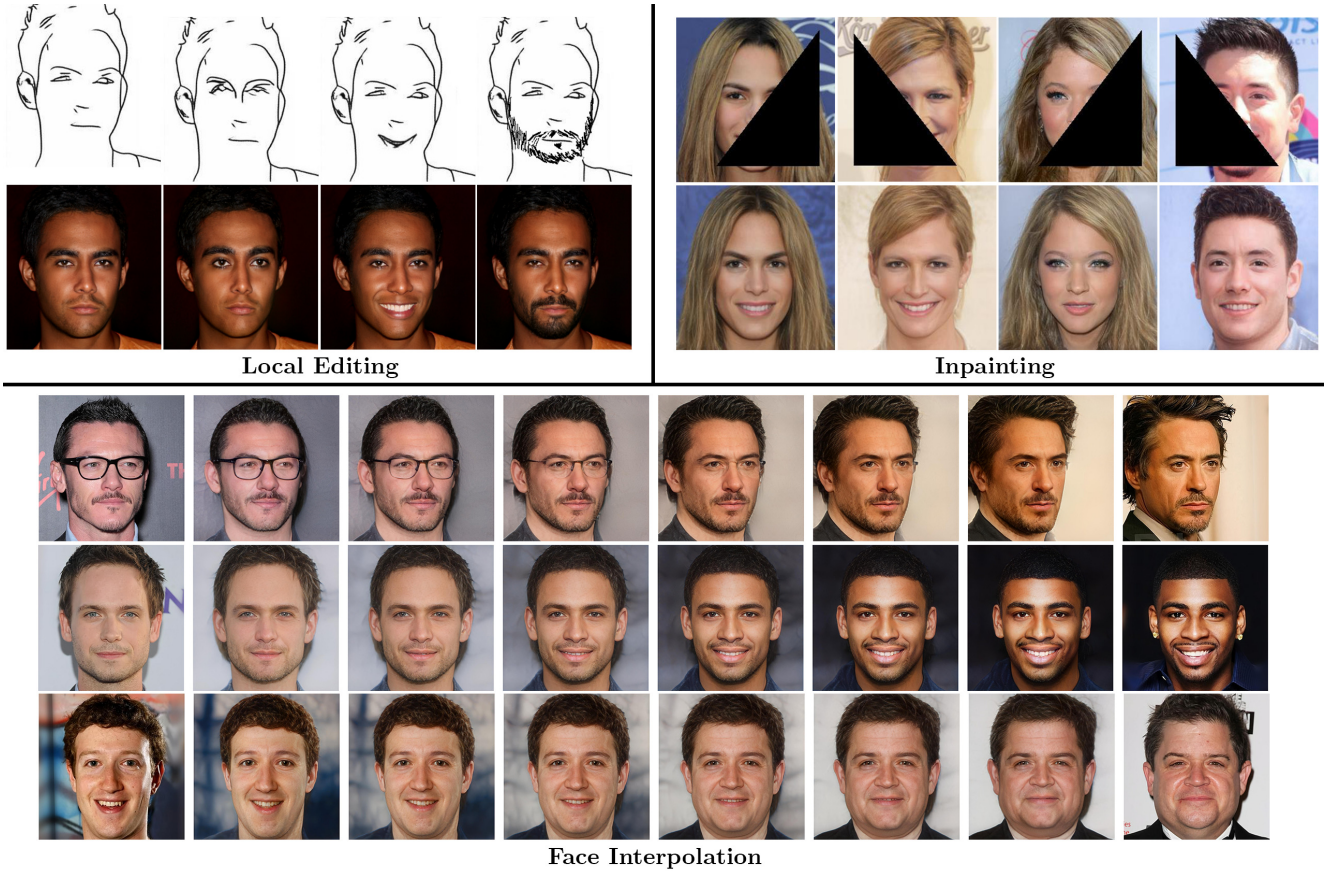

Figure 15: Additional applications for the pSp framework.

References

[1] R. Abdal, Y. Qin, and P. 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, 3

[2] R. Abdal, Y. Qin, and P. 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. 2, 3

[3] R. Adbal, P. Zhu, N. J. Mitra, and P. Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *arXiv preprint arXiv:*, 2020. 2, 3

[4] Baylies. stylegan-encoder. https://github.com/pbaylies/stylegan-encoder, 2019. Accessed: April 2020. 2

[5] S.-Y. Chen, W. Su, L. Gao, S. Xia, and H. 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. 3, 7, 8

[6] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 3

[7] Y. Choi, Y. Uh, J. Yoo, and J.-W. 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. 3

[8] E. Collins, R. Bala, B. Price, and S. 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. 2, 3, 4

[9] A. Creswell and A. A. Bharath. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 30(7):1967–1974, 2018. 2

[10] J. Deng, J. Guo, N. Xue, and S. 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. 4, 5

[11] E. Denton, B. Hutchinson, M. Mitchell, and T. Gebru. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439*, 2019. 3

[12] L. Goetschalckx, A. Andonian, A. Oliva, and P. Isola. Ganalyze: Toward visual definitions of cognitive image properties, 2019. 3

[13] S. Guan, Y. Tai, B. Ni, F. Zhu, F. Huang, and X. Yang. Collaborative learning for faster stylegan embedding. *arXiv preprint arXiv:2007.01758*, 2020. 2, 4

[14] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020. 2, 3

[15] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 3, 4

[16] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. 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. 5

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. 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. 2, 11

[18] A. Jahanian, L. Chai, and P. Isola. On the "steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019. 2, 3

[19] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 4

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5, 15

[21] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 8

[22] T. Karras, S. Laine, and T. 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. 1, 2, 4, 5

[23] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. 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. 1, 2

[24] O. Katzir, D. Lischinski, and D. Cohen-Or. Cross-domain cascaded deep feature translation. *arXiv*, pages arXiv–1906, 2019. 3

[25] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision – ECCV 2012*, pages 679–692, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 9, 16

[26] Y. Li, X. Chen, F. Wu, and Z.-J. 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. 3

[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2, 3

[28] Z. C. Lipton and S. Tripathi. Precise recovery of latent vectors from generative adversarial networks. *arXiv preprint arXiv:1702.04782*, 2017. 2

[29] W. Lira, J. Merz, D. Ritchie, D. Cohen-Or, and H. Zhang. Ganhopper: Multi-hop gan for unsupervised image-to-image translation. *arXiv preprint arXiv:2002.10102*, 2020. 3

[30] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019. 5

[31] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017. 3

[32] X. Liu, G. Yin, J. Shao, X. Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Advances in Neural Information Processing Systems*, pages 570–580, 2019. 3, 9

[33] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 9

[34] Y. Nitzan, A. Bermano, Y. Li, and D. Cohen-Or. Disentangling in latent space by harnessing a pretrained generator. *arXiv preprint arXiv:2005.07728*, 2020. 2

[35] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 3, 9

[36] G. Perarnau, J. Van De Weijer, B. Raducanu, and J. M. Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016. 2

[37] S. Pidhorskyi, D. A. Adjeroh, and G. Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14104–14113, 2020. 5, 6

[38] E. Richardson and Y. Weiss. The surprising effectiveness of linear unsupervised image-to-image translation. *ArXiv*, abs/2007.12568, 2020. 2, 4

[39] Y. Shen, J. Gu, X. Tang, and B. 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. 2, 3

[40] E. Simo-Serra, S. Iizuka, K. Sasaki, and H. Ishikawa. Learning to simplify: fully convolutional networks for rough sketch cleanup. *ACM Transactions on Graphics*, 35:1–11, 07 2016. 8

[41] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Pérez, M. Zollhöfer, and C. Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. *arXiv preprint arXiv:2004.00121*, 2020. 2, 3

[42] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2, 3, 5

[43] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2009. 8

[44] C. Yang, Y. Shen, and B. Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis, 2019. 2

[45] M. Zhang, J. Lucas, J. Ba, and G. E. Hinton. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, pages 9597–9608, 2019. 5

[46] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. 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. 4

[47] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR 2011*, pages 513–520, 2011. 8

[48] H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5911–5920, 2020. 6

[49] J. Zhu, Y. Shen, D. Zhao, and B. Zhou. In-domain gan inversion for real image editing. *arXiv preprint arXiv:2004.00049*, 2020. 2, 3, 5, 6

[50] J.-Y. Zhu, T. Park, P. Isola, and A. 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. 3

[51] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017. 3, 4

[52] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation, 2017. 4

[53] P. Zhu, R. Abdal, Y. Qin, and P. 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. 3

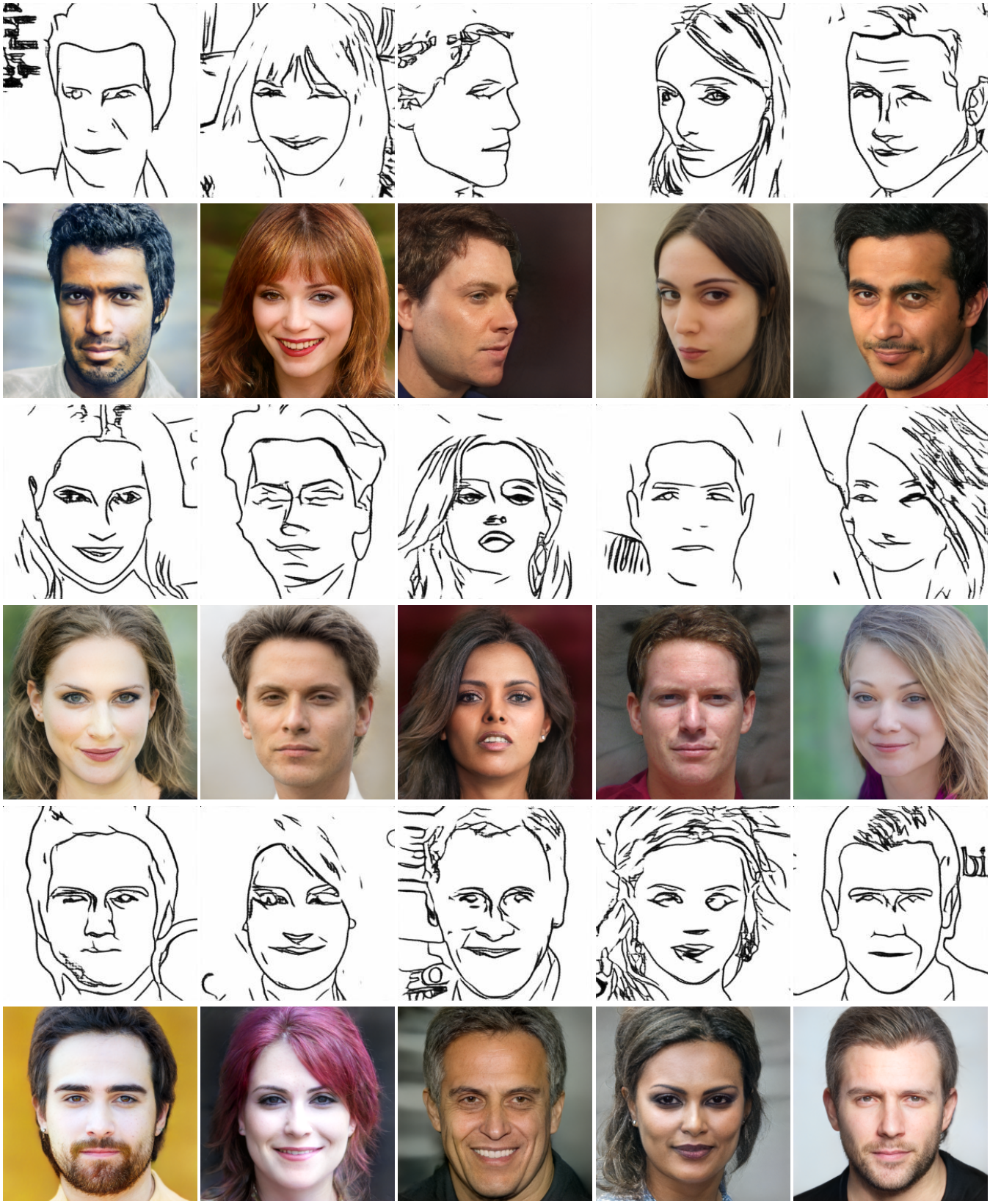

Figure 16: Additional results using pSp for the generation of face images from sketches on the CelebA-HQ [20] test dataset.

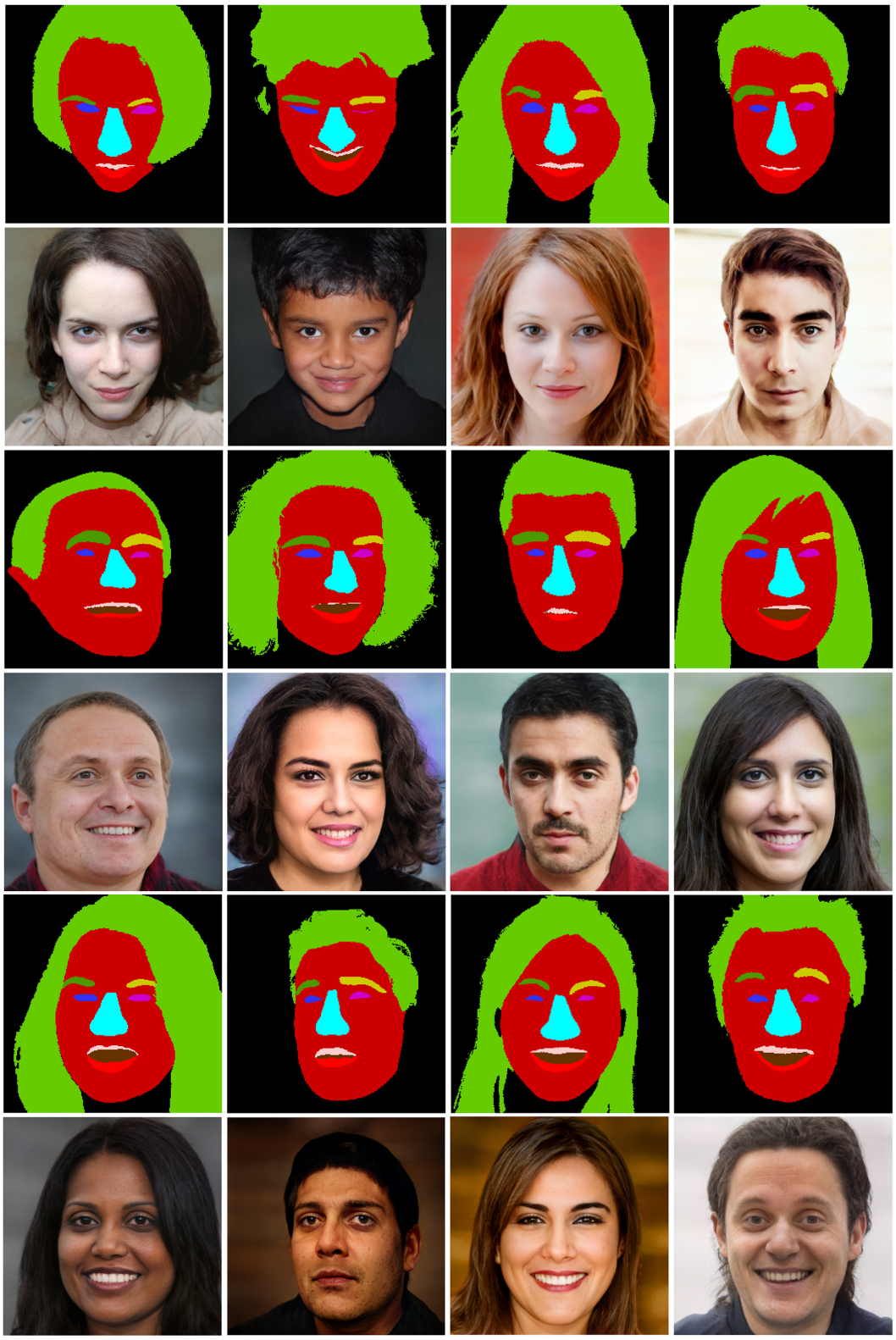

Figure 17: Additional results on the Helen Faces [25] dataset using our proposed label-to-image method.