Adaptive Nonlinear Latent Transformation for Conditional Face Editing

Zhizhong Huang[1] Siteng Ma[1] Junping Zhang[1] Hongming Shan[2,3]*
[1] Shanghai Key Lab of Intelligent Information Processing, School of Computer Science,
Fudan University, Shanghai 200433, China
[2] Institute of Science and Technology for Brain-inspired Intelligence and MOE Frontiers Center
for Brain Science, Fudan University, Shanghai 200433, China
[3] Shanghai Center for Brain Science and Brain-inspired Technology, Shanghai 200031, China
{zzhuang19, stma21, jpzhang, hmshan}@fudan.edu.cn

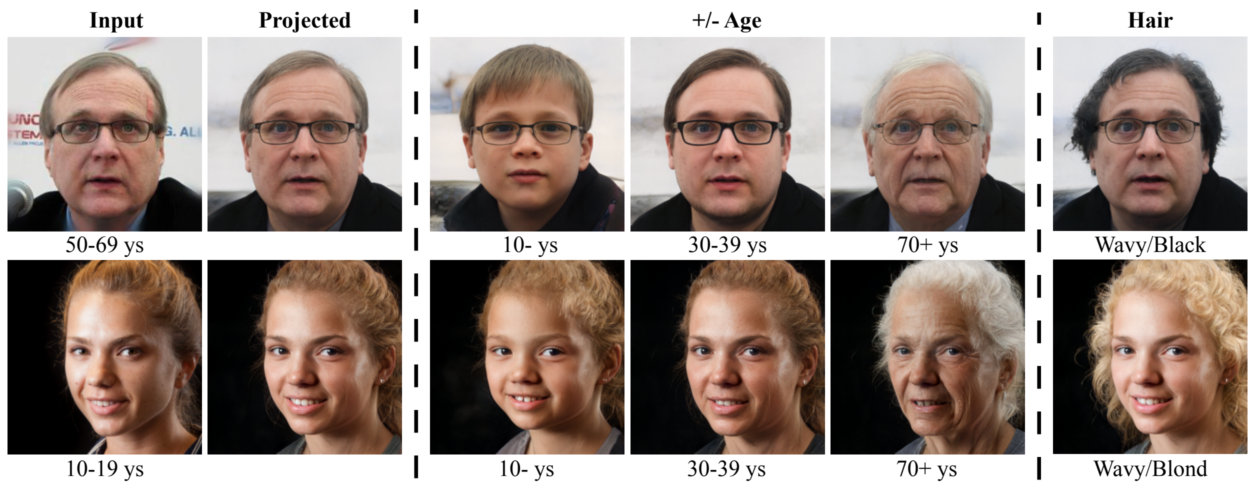


Figure 1. Sample results produced by our AdaTrans. By projecting the images into the latent space of StyleGAN, AdaTrans can achieve disentangled face editing even when the age gap is extreme large, and manipulate multiple attributes at the same time.

Abstract

Recent works for face editing usually manipulate the latent space of StyleGAN via the linear semantic directions. However, they usually suffer from the entanglement of facial attributes, need to tune the optimal editing strength, and are limited to binary attributes with strong supervision signals. This paper proposes a novel adaptive nonlinear latent transformation for disentangled and conditional face editing, termed AdaTrans. Specifically, our AdaTrans divides the manipulation process into several finer steps; i.e., *the direction and size at each step are conditioned on both the facial attributes and the latent codes. In this way, AdaTrans describes an adaptive nonlinear transformation trajectory to manipulate the faces into target attributes while keeping other attributes unchanged. Then, AdaTrans leverages a predefined density model to constrain the learned trajectory in the distribution of latent codes by maximizing the likelihood of transformed latent code. Moreover, we also propose a disentangled learning strategy under a mutual information framework to eliminate the entanglement among attributes, which can further relax the need for labeled data. Consequently, AdaTrans enables a controllable face editing with the advantages of disentanglement, flexibility with non-binary attributes, and high fidelity. Extensive experimental results on various facial attributes demonstrate the qualitative and quantitative effectiveness of the proposed AdaTrans over existing state-of-the-art methods, especially in the most challenging scenarios with a large age gap and few labeled examples. The source code is available at* https://github.com/Hzzone/AdaTrans.

1. Introduction

Face editing aims to render the faces to the target facial attributes such as aging or smiling with high fidelity while keeping other facial attributes unchanged, which has

*Corresponding author

wide applications in entertainment and forensics. Due to the intrinsic complexity of facial attributes, face editing has attracted growing research interest in recent years. Generative Adversarial Networks (GANs) [12] have shown promising results for face editing in terms of image quality. Earlier works mainly focus on network architectures [19, 39, 43] and loss functions [31]. With significant improvements in image quality, these methods usually re-train a GAN model for a specific facial attribute [17] or practical applications [27]. Unfortunately, due to the difficulties in training a good GAN, they are limited to specific tasks and fail to generalize to high-resolution images.

In recent years, StyleGAN [22–24] has achieved significant progress in synthesizing photorealistic faces. In particular, the pre-trained StyleGAN generator presents a meaningful intermediate latent space, traversing on which the faces can be semantically manipulated [2, 3, 14, 34, 35, 38, 40, 41]. Typically, the faces before editing need to be inverted into the latent space of StyleGAN to obtain the latent codes that can be used to faithfully reconstruct the inputs [4, 33, 36]. As a result, the latent code is manipulated along certain directions, giving rise to the changes in the corresponding attribute in generated faces. The methods to obtain those directions can be roughly categorized as supervised ones and unsupervised ones. The supervised methods [3, 34, 38, 41] leverage the labeled data to compute the semantic directions, leading to better controllability in the editing process. For example, InterFaceGAN [34] trains a hyperplane in the latent space to separate the examples with binary attributes. Unsupervised methods [2, 14, 35, 40] are to discover the interpretable directions using PCA [14, 35] or texts [2]. Despite the meaningful transformations, they cannot produce precise user-desired editing without any human annotations.

In summary, most of these methods assume that the binary attributes can be well separated, so they edit the faces by linear interpolation in the latent space. Although it is sufficient to some degree, for those more complicated scenarios, *e.g.*, with large age gaps, they cannot perform disentangled editing to preserve the unrelated attributes when linear assumption does not hold. Meanwhile, the users are usually required to manually tune the editing strength for accurate manipulation. Though flexible, the optimal strength varies among different examples. Furthermore, there is a critical yet ignored problem in current literature that the latent codes could be over-manipulated, *i.e.* falling out of the latent space, which inevitably harms the quality of the edited face.

In this paper, we propose a novel framework for conditional face editing, termed AdaTrans, to address these issues in the following aspects. *First*, instead of manually manipulating the latents with fixed directions, we propose an adaptive nonlinear transformation strategy that dynamically estimates the editing direction and step size, conditioned by the target attributes and transformation trajectory. Such a strategy can handle various attributes at the same time for conditional multi-attribute editing, by only changing the target attributes while keeping others unchanged. *Second*, we propose to maximize the likelihood of edited latent codes, regularize the transformed trajectory in the distribution of latent space, and hence improve the fidelity of edited faces predicted by a pretrained generator. *Last*, we propose a disentangled learning strategy under a mutual information framework, attenuating the entanglement between attributes and relaxing the need for labeled data in supervised face editing methods. The merits of AdaTrans are disentanglement, high fidelity, controllability, and flexibility. The sample results are shown in Fig. 1.

The contributions are summarized as follows:

- We present AdaTrans, a novel face editing method that explores an adaptive nonlinear transformation for disentangled and multi-attribute face editing.
- We propose a novel density regularization term, which can encourage an in-distribution transformation in the latent space, without harming fidelity.
- We further show a disentangled learning strategy, which can eliminate the entanglement between attributes and relax the need for labeled data.
- Experimental results on various facial attributes demonstrate the effectiveness of AdaTrans both quantitatively and qualitatively. In particular, AdaTrans can produce disentangled editing, even with extremely large age gaps or few labeled data.

2. Related Work

Generative adversarial networks. Generative Adversarial Networks (GANs) describe a competition between the generator and discriminator, where the generator maps the random noise (*e.g.* Gaussian) to the complicated data distribution (*e.g.* image), and the discriminator tries to distinguish the true/generated data. Various works have made significant progress in synthesizing photorealistic faces from different aspects such as loss functions [5, 29] and architectures [6, 21, 23]. In particular, StyleGAN [22–24] presents a meaningful intermediate latent space $\mathcal{W}$ which is better disentangled than a standard Gaussian latent space $\mathcal{Z}$. To utilize a well pre-trained StyleGAN generator, the faces are first inverted to $\mathcal{W}$ to obtain the latent codes that can be used to faithfully reconstruct the input images [4, 16, 33, 36, 42]. Consequently, semantical face editing can be achieved by manipulating the latent space of GANs, which is fed into the generator to obtain the manipulated faces.

Face editing in GANs. The prior literature on face editing can be roughly split into supervised and unsupervised methods. The supervised methods [3, 25, 34, 38, 40, 41, 44] typically employ the human annotations with particular facial attributes or a pre-trained classifier to identify how to

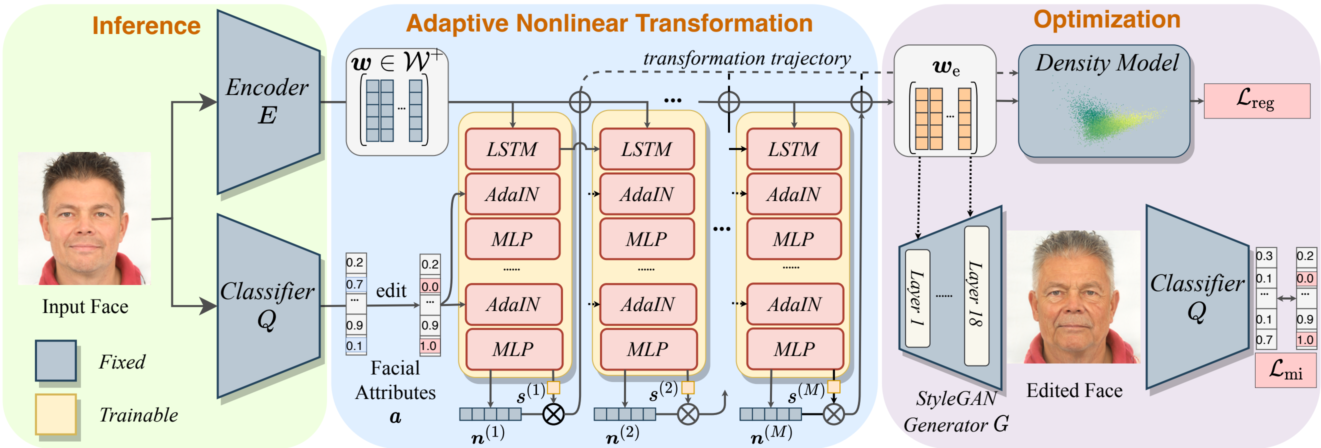


Figure 2. Illustration of the proposed AdaTrans. Given an input face, the pre-trained classifier predicts its attributes and the encoder inverts it to the latent space of StyleGAN generator. By changing the facial attributes, our proposed AdaTrans can manipulate the faces in the latent space for adaptive nonlinear transformation without tuning the strength. The proposed density regularization term can encourage the in-distribution edited latent codes and the disentangled learning strategy can attenuate the entanglement between attributes. The merits of AdaTrans are flexible, disentangled, controllable, and nonlinear.

manipulate the faces in the latent space. InterFaceGAN [34] uses the hyperplane separating latent codes with binary attributes to carry out the linear editing. StyleFlow [3] remaps the latent codes to Gaussian noise with normalizing flows and then samples the new latent codes conditioned by the target attributes and original noise. Latent Transformer [41] trains a transformation network for specified attributes under the supervision of other attributes. Differently, the unsupervised methods [1,2,9,14,20,35,37] do not need the valuable annotated data. GANSpace [14] and SeFa [35] find editing directions from the principal components of the latent space. Clip2Stylegan [2] links the latent transformation with the text descriptions under the guidance of CLIP [32]. However, the discovered editing directions still need to be manually labeled for meaningful editing, and it is hard to produce the user-desired directions.

In this paper, we study an adaptive nonlinear transformation rather than linear interpolation in [2,14,25,34,35,40,41], and investigate a density regularization to encourage in-distribution latent transformation.

3. The Proposed AdaTrans

In this section, we first formulate the problem of face editing and present our motivation in Sec. 3.1, then describe the proposed adaptive nonlinear transformation in Sec. 3.2 and latent density regularization in Sec. 3.3, followed by the training and inference of the proposed AdaTrans in Sec. 3.4. Fig. 2 illustrates the framework of the proposed AdaTrans.

3.1. Problem Formulation and Motivation

As presented in Fig. 2, we are now given a classifer network Q to estimate the facial attributes $\boldsymbol{a} =$

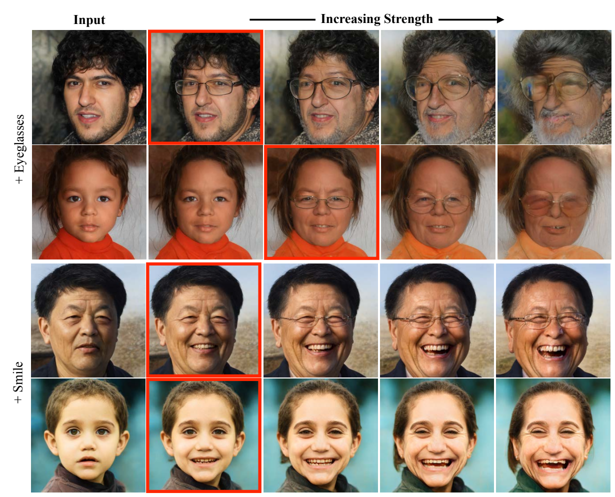


Figure 3. Sample results of linear interpolation of InterFaceGAN [34] varying the strength. The leftmost text describes the changed attributes, and the red boxes indicate that faces have been successfully manipulated to target attributes. With the increase of strength, the unrelated attributes are changed and fidelity is harmed.

$\{a_1, a_2, \dots, a_N\}$, where N is the number of attribute annotations, and a_i can be binary or one-hot value. Face editing focuses on manipulating one or multiple attributes in $\boldsymbol{a}$ without changing others. Another pre-trained StyleGAN generator G is employed to produce high-resolution photo-realistic faces $\mathcal{I} \subset \mathbb{R}^{H\times W}$ from the latent space $\mathcal{W} \subset \mathbb{R}^{512}$, where $H \times W$ represents the image size.

To perform face editing with G, the face $\boldsymbol{I} \in \mathcal{I}$ should be inverted into the latent space with an encoder E: $\mathcal{I} \to \mathcal{W}$ to obtain the latent code $\boldsymbol{w} = E(\boldsymbol{I})$. The practical choices [4, 33, 36] tend to adopt $\mathcal{W}^+ \subset \mathbb{R}^{18\times 512}$ with layer-wise latent codes for faithful reconstruction. Then, the inverted latent codes are manipulated to change the target attributes. The resulting edited latents $\boldsymbol{w}_{\mathrm{e}}$ can be fed into G to obtain the generated face $G(\boldsymbol{w}_{\mathrm{e}})$.

Previous literature [2, 14, 34, 35, 41] usually edits the facial attribute a_i by linear interpolation in the latent space with certain editing directions $\boldsymbol{n}_{a_i} \in \mathbb{R}^{512}$, which can be formulated as follows:

$$\boldsymbol{w}_{\mathrm{e}} = \boldsymbol{w} + \alpha \boldsymbol{n}_{a_i}. \tag{1}$$

In such a setting, $\alpha \in \mathbb{R}$ is a scalar to control the editing strength of manipulating the face to a_i. $\boldsymbol{n}_{a_i}$ can be learned by training a hyperplane/fully-connected layers in the latent space [34, 41], or discovered from the principal components [14, 35] and text information [2].

However, as observed in Fig. 3, there are several limitations in such simple linear interpolation. *First*, linearly manipulating the latent code would change other unrelated attributes due to the entanglement of the latent space. *Second*, the optimal strength for accurate editing is hard to tune. Small strength may not change the desired attributes while a large one would harm the face quality as the latent codes are out of the latent space. Importantly, the strength varies from input faces. *Last*, it is limited to binary attributes such as `gender` and `young`, and cannot handle those more complicated attributes, *e.g.*, finer age that cannot be controlled by a linear transformation. To verify this claim, supplementary Fig. A3 visualizes the latent codes with fine age labels.

Consequently, we need an adaptive transformation for nonlinear editing to address the above issues, which will be detailed in the following sections.

3.2. Adaptive Nonlinear Transformer

To avoid tuning the optimal strength for face editing with linear interpolation, we propose an adaptive nonlinear transformer for traversing the latent space of GANs. Specifically, instead of directly learning an editing direction, we opt to divide the whole transformation process into several fine steps, where the size and direction at each step are conditioned on the target attributes. As a result, a nonlinear transformation trajectory can be obtained by these fine linear steps and the endpoint is the edited latent code that we are desired.

As shown in Fig. 2, our adaptive transformer takes the latent code $\boldsymbol{w}$ as the input. An LSTM [13] is also employed at the beginning to smooth the transformation trajectory. At step t, learnable affine transformations are adopted to inject the changed/unchanged attributes to modulate the size and the direction of manipulation, which are parameterized by a two-layer multilayer perceptron (MLP) following the AdaIN [18] operation:

$$\mathrm{AdaIN}(\boldsymbol{h}_j^{(t)}, \boldsymbol{a}) = \boldsymbol{y}_{j,s}^{(t)}(\boldsymbol{a})\boldsymbol{h}_j^{(t)} + \boldsymbol{y}_{j,b}^{(t)}(\boldsymbol{a}), \tag{2}$$

where $\boldsymbol{h}_j$ is the input feature, and $\boldsymbol{y}_{j,s}(\cdot)$ and $\boldsymbol{y}_{j,b}(\cdot)$ output the learned scale and bias, respectively.

The intermediate edited latent code at t step is formed as:

$$\boldsymbol{w}_{\mathrm{e}}^{(t)} = \boldsymbol{w}_{\mathrm{e}}^{(t-1)} + s^{(t)}\boldsymbol{n}^{(t)}, \tag{3}$$

where $\boldsymbol{w}_{\mathrm{e}}^{(0)} = \boldsymbol{w}$, and the adaptive transformer f outputs $\{\boldsymbol{n}_t, s_t\} = f(\boldsymbol{w}_{\mathrm{e}}^{(t-1)}, \boldsymbol{a})$ conditioned on the target attributes $\boldsymbol{a}$. Here, $s_t \in (0, 1)$ activated by a sigmoid function is a scalar to control the step size, and $\boldsymbol{n}$ is a unit vector, *i.e.*, $\|\boldsymbol{n}\|_2 = 1$. Consequently, our adaptive transformer can adaptively adjust the latent code in terms of previously manipulated results.

In summary, the manipulation process can be described as the combination of the intermediate states as follows:

$$\boldsymbol{w}_{\mathrm{e}} = \boldsymbol{w} + \sum_{t=1}^{M} s^{(t)}\boldsymbol{n}^{(t)}, \tag{4}$$

where M is the maximum steps of the trajectory.

3.3. Latent Density Regularization

Although benefitting from the well-trained StyleGAN generator, the latent space of StyleGAN generator is fixed during editing and hence over-manipulating the latent codes to the target attributes would harm the image quality since the codes are out of the distribution $\mathcal{W}$. It is natural to observe this problem. For example, in Fig. 3, when the strength of manipulation is set to a large value, the faces have been changed a lot to target attributes at the cost of unnatural generated faces, since they are far away from decision boundaries [34].

The simple solution is to employ an additional loss to restrict the Euclidean distance between the original and edited latent code [41], which can be written as:

$$\mathcal{L}_{\mathrm{dist}} = \|\boldsymbol{w}_{\mathrm{e}} - \boldsymbol{w}\|_2. \tag{5}$$

However, it cannot deal with the essential nature of this problem with the following problems: (1) simply restricting the editing strength may fail to change the desired attributes, and (2) this problem still remains when increasing the strength. Another solution is to train a discriminator again to improve face quality, which, however, will significantly increase the computational cost and the training difficulty.

To address this issue, we propose to use a regularization term to encourage the edited latent codes to fall into the distribution of the latent space. We employ a simple yet effective method by maximizing the likelihood of the transformation trajectory with a pre-trained density model. Here, the density model is trained in advance to model the latent space. In this paper, we adopt the real NVP [11] due to its simplicity and effectiveness in modeling data distribution.

Figure 4. Disentangled and controllable face editing on real faces with binary facial attributes. Given an input face, we gradually manipulate it into the target attributes underneath, while keeping the previous manipulating effects.

In doing so, we can always restrict the transformation trajectory in the latent space, so that the fidelity of generated faces would not be distorted. The resultant regularization term can be thus written as:

$$\mathcal{L}_{\text{reg}} = -\frac{1}{M}\sum_{t=1}^{M} \log p_\phi(\boldsymbol{w}_{\text{e}}^{(t)}). \tag{6}$$

Here, the density model is parameterized by ϕ and M is the maximum step of the trajectory.

3.4. Training and Inference

Disentangled learning. Given the attributes $\boldsymbol{a} = \{a_1, \hat{a}_2, \dots, a_N\}$ to be changed or kept, we propose to achieve disentangled face editing by maximizing the mutual information $\mathbb{I}(\boldsymbol{a}; G(\boldsymbol{w}_{\text{e}}))$ between conditions and generated faces, following [8]. Formally, the mutual information loss function can be written as:

$$\mathcal{L}_{\text{mi}} = -\sum_{a\in\boldsymbol{a}} \log p_\theta(a|G(\boldsymbol{w}_{\text{e}})), \tag{7}$$

where the true posterior $p_\theta(a|G(\boldsymbol{w}_{\text{e}}))$ is approximated by the classifier Q. In the supervised setting, Q is pre-trained on the labeled dataset and fixed during training. We would like to highlight that AdaTrans puts all attributes in $\boldsymbol{a}$, which enables AdaTrans to attenuate the entanglement between attributes [8]. The most related work to AdaTrans is [3] which models the conditional distributions of latent codes and attributes using continuous normalizing flows [7]. However, the sampling procedure in [3] makes it difficult to handle the scenario with few labeled examples, as studied in Sec. 4.

Training and inference. Combining Eqs. (5), (6), and (7), the overall training objective for AdaTrans is to minimize the sum of all losses:

$$\mathcal{L} = \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{mi}}\mathcal{L}_{\text{mi}}, \tag{8}$$

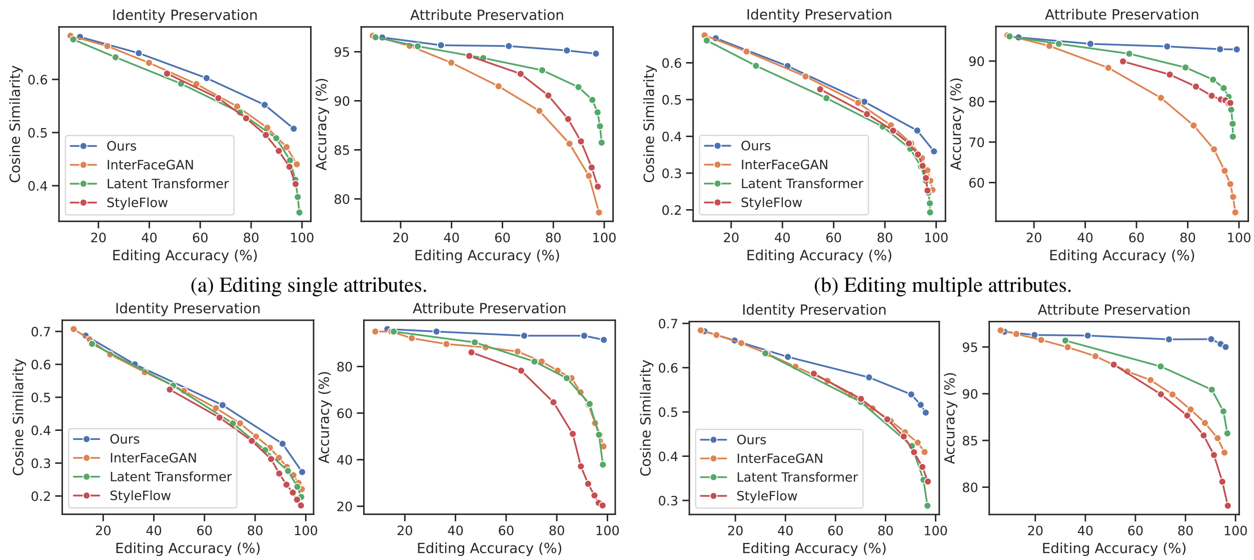
where λ_* controls the importance between different loss components. Here, $\mathcal{L}_{\text{reg}}$ is employed to compress those unnecessary changes in the latent codes. During training, the binary attributes of the given face are randomly manipulated to $\{0, 1\}$. As a result, AdaTrans can manipulate the latent code to desired attributes while keeping others unchanged.

4. Experiments

4.1. Implementation Details

In this paper, we perform face editing in the latent space of the StyleGAN2 generator [24] pre-trained on FFHQ [23] and employ e4e encoder [36] to project the faces into $\mathcal{W}^+$ space. For the fixed attribute classifier, we trained the last linear layer of ResNet-50 [15] on the binary attributes of CelebA dataset [28] and the discrete age labels of FFHQ from [30]. We train another classifier from scratch for better classification performance and obtain the attribute labels of FFHQ dataset. The first 69k faces of FFHQ are left as training data with an image size of 256×256. The rest 1k faces and CelebA-HQ [21] are used as the testing data.

AdaTrans is trained for 10k iterations with a batch size of 16, Adam optimizer [26] with a fixed learning rate of



(a) Editing single attributes.

(b) Editing multiple attributes.

(c) Editing with large age gap.

(d) Editing with limited labeled data.

Figure 5. Quantitative comparisons with recent methods for face editing under different experimental settings. We employ attribute/identity preservation during editing to evaluate different methods. We desire higher editing accuracy while keeping the original identity and unrelated attributes. Therefore, a higher curve indicates better performance.

10^{-4}, $\beta_1 = 0.9$ and $\beta_2 = 0.99$. A one-layer LSTM and 10 two-layer MLP with the dimension 512 and ReLU activation are stacked. The direction and size of each step are output by a separate full-connected layer. The loss weights λ_* and the maximum step M are set to 1 and 5, respectively. Real NVP [11] is employed as the density model trained on the latent codes without any supervision for 10k iterations.

4.2. Disentangled and Controllable Face Editing

4.2.1 Qualitative Results

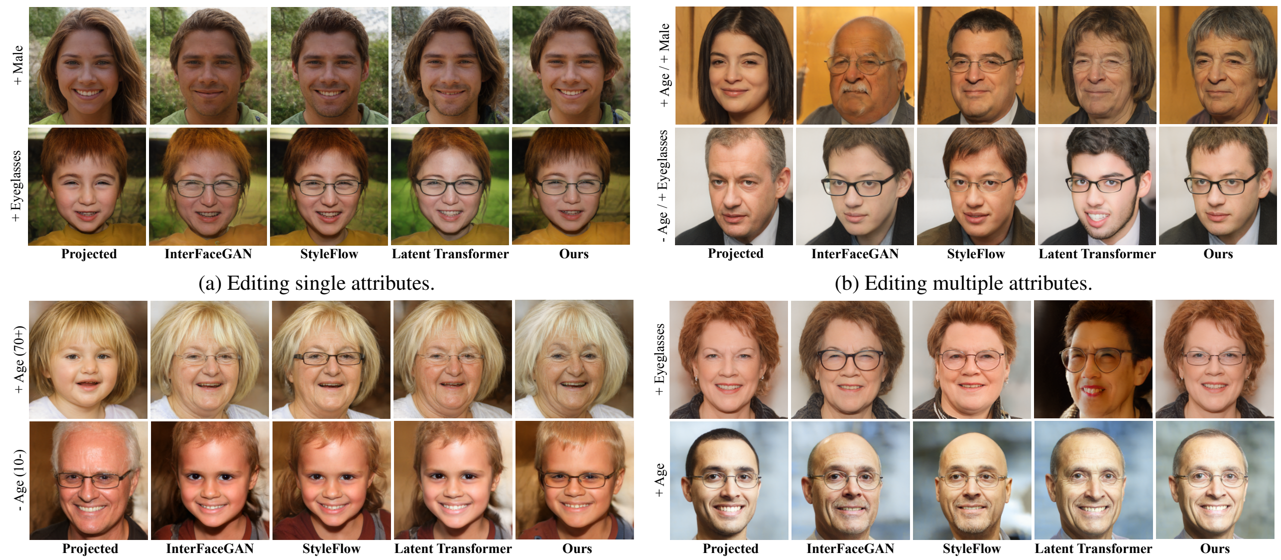
Fig. 4 showcases example results on manipulating the faces into target attributes. We project the faces into the $\mathcal{W}^+$ space of StyleGAN2 using [36], and then gradually manipulate the latent codes with a sequential attribute list. Obviously, AdaTrans achieves photorealistic and disentangled modifications on the resultant faces, *i.e.*, the identity and other attributes are well preserved during the manipulation process. Note that we train all attributes presented in Fig. 4 in a unified model. Therefore, AdaTrans has achieved great flexibility and controllability simultaneously and is not limited to a defined order of attributes. We provide additional examples for sequential editing in supplementary Fig. A5 and the results for editing a single attribute in supplementary Fig. A6.

4.2.2 Comparisons with State-of-the-arts

Competitors. To validate the effectiveness of AdaTrans, we perform comparisons with the recent state-of-art methods including InterFaceGAN [34], StyleFlow [3] and Latent Transformer [41]. We implement all methods on the latent codes of FFHQ dataset projected by [36], strictly following their experimental settings. We highlight that editing on real faces is much more difficult than synthetic ones. Only three binary attributes for `gender`, `eyeglasses`, and `young` are adopted to evaluate different methods, which are definitely shown to be entangled together [34] with each other in the latent space of StyleGAN. For fair comparisons, the same attribute classifier is used for Latent Transformer.

Evaluation metrics. We employ three widely-used metrics to compare different methods quantitatively, including editing accuracy, attribute preservation accuracy, and identity preservation. A ResNet-50 is trained from scratch to predict the attributes of manipulated faces, which can be used to measure the editing accuracy and attribute preservation for different facial attributes. The editing accuracy indicates the proportion of the samples that have been successfully manipulated into target attributes, *i.e.*, the probability is greater than 0.5. On the other hand, attribute preservation accuracy is the proportion of the rest attributes being kept. Identity preservation is the cosine similarity between the facial embeddings of the original input and manipulated faces extracted by [10]. In a sense, we desire a higher editing accuracy (*i.e.*, manipulating the faces into target attributes) with fewer drops in identity and attribute preservation (*i.e.*, identity and other unrelated attributes are unchanged).

The input faces in testing data are manipulated into the opposite attribute with different strengths to control the degree of manipulation. We gradually increase the strengths



(a) Editing single attributes.

(b) Editing multiple attributes.

(c) Editing with large age gap.

(d) Editing with limited labeled data.

Figure 6. Qualitative comparisons with recent methods for face editing under different experimental settings. The editing strengths of different methods are manually increased until the faces are manipulated into target attributes. The competitors produce unexpected changes in unrelated attributes or fail to handle the scenarios with extreme conditions like age gaps and few labeled samples.

for different methods as suggested by their official code until the editing accuracy reaches 99%. More specifically, we interpolate the label variables of StyleFlow [3], scale the learned directions for InterFaceGAN [34] and Latent Transformer [41], and modifies the maximum steps M of AdaTrans. Consequently, we can draw the curves of identity/attribute preservation w.r.t. editing accuracy.

Figs. 5 and 6 show the comparisons between different methods under different settings. Additional qualitative comparisons are provided in supplementary Fig. A7.

Editing single/multiple attributes. In this setting, each single or any two (*i.e.*, multiple) attributes are manipulated at the same time. The quantitative results are shown in Figs. 5a and 5b. Obviously, AdaTrans outperforms the other three competitors by a large margin. We showcase the qualitative results in Figs. 6a and 6b. For attributes like becoming old, the glasses appear for other methods, indicating that the entangled attributes are not handled well. Overall, the results demonstrate that AdaTrans achieves better disentangled and accurate face editing both quantitatively and qualitatively.

Editing with large age gaps. As studied in the previous setting, the most challenging task is becoming old as the glasses and age are entangled together in the latent space. We further validate the effectiveness of AdaTrans in a much more rigorous setting, *i.e.*, manipulating the children below 10 years old to the old above 70 years old. We replace the `young` label with the age labels of FFHQ dataset provided by [30], which consists of 7 discrete age groups. The quantitative results in Fig. 5c show that AdaTrans performs significantly better that the competitors. We note that it is natural to see a severe drop in identity preservation since identity would be changed a lot during the facial aging process. In terms of qualitative results in Fig. 6c, AdaTrans can still preserve the eyeglasses from 70+ to 10-, and eliminate the appearance of eyeglasses from 10- to 70+. In summary, in the most challenging experimental settings, AdaTrans can still achieve disentangled face editing over state-of-the-art methods.

Editing with limited labeled data. An important problem of supervised face editing methods [3, 34, 41] is that they require a large number of labeled samples, which is difficult to collect in practical scenarios. In this experimental setting, only 128 labeled samples for each attribute are used to challenge different editing methods. The quantitative and qualitative results for single attribute manipulation are shown in Fig. 5d and Fig. 6d, respectively. The proposed AdaTrans can still handle well and achieve photorealistic face editing with limited labeled samples, benefitting from the disentangled learning strategy. Therefore, AdaTrans is more flexible to practical applications than the competitors. An ablation study on the number of labeled samples is conducted in supplementary section A.

4.3. Ablation Study

We conduct ablation studies to validate the effectiveness of different components in AdaTrans.

Ablation study on the proposed components. We start

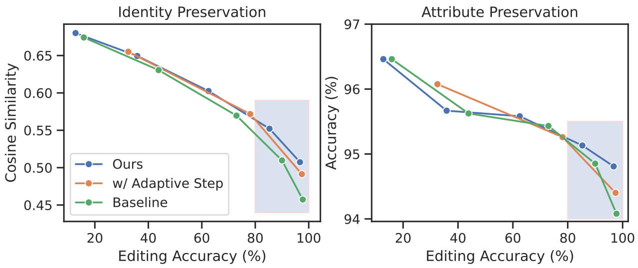


Figure 7. Quantitative results of the proposed components. The color box is the region of interest for comparisons.

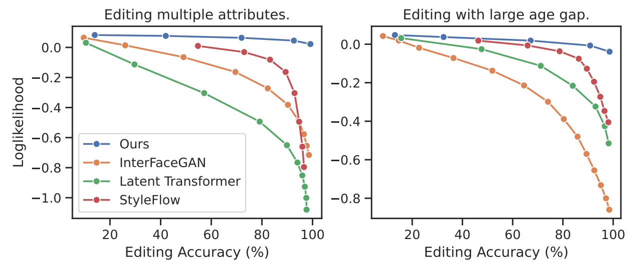


Figure 8. Quantitative comparisons of the loglikelihood.

from the baseline method without the proposed adaptive nonlinear transformer and density regularization, directly optimized by $\mathcal{L}_{\mathrm{mi}}$ and $\mathcal{L}_{\mathrm{dist}}$. The baseline still performs nonlinear editing, however, with a fixed step size at each step, compared to the full AdaTrans. Then, we gradually apply the proposed adaptive strategy and latent density regularization to the baseline, leading to two variants of our AdaTrans.

In Fig. 7, significant drops in identity/attribute preservation have arisen for the baseline; see the color box. We argue that the baseline cannot compress the unnecessary changes in the latent codes and does not adjust the manipulation process with the fixed step size.

On the other hand, the adaptive strategy can address this problem by adaptively controlling the manipulation according to the previous trajectory and target attributes. Furthermore, we have also found the fidelity of generated faces has been unexpectedly harmed. We show the qualitative results in supplementary Fig. A2. To be manipulated into ideal target attributes, the latent codes may fall out of the latent space, which, as a result, harms the image quality produced by a pre-trained StyleGAN generator. This phenomenon motivates us to employ latent regularization to encourage the edited latent codes to be in-distribution.

Analysis of regularization term. The pre-trained density model is employed to estimate the loglikelihood of edited latent codes (normalized by dimension) for different methods. Fig. 8 presents the results for editing multiple attributes and editing with a large age gap. The higher loglikelihood indicates more in-distribution manipulation, indicating better fidelity of the pre-trained StyleGAN generator. StyleFlow performs better than InterFaceGAN and Latent Transformer as it samples the edited codes from the latent distribution. AdaTrans is still able to encourage the latent codes in-distribution while achieving better editing accuracy, which is the key idea of latent regularization.

Ablation study on the maximum steps M. In AdaTrans, M defines the maximum steps of transformation trajectory. During training, M is empirically fixed and AdaTrans is trained to learn each step size and direction. During inference, the user can flexibly adjust M for desired visual results (*e.g.*, becoming younger or older). Fig. 9 demonstrates that increasing M during training can further improve editing performance. Although M can be set large enough since AdaTrans can adaptively adjust the step size, more computational costs may be introduced and effects of each step may be minor. Supplementary Fig. A8 visualizes the transformed trajectory and shows that AdaTrans has produced smooth transformations.

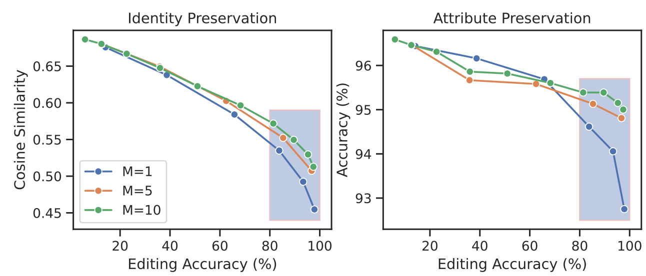


Figure 9. Quantitative results for the ablation of M.

5. Conclusion

In this paper, we introduce AdaTrans, a novel nonlinear transformation for face editing. AdaTrans divides the manipulation process into several finer steps to achieve nonlinear manipulation in the latent space of GANs. A predefined density real NVP model regularizes the trajectory of the transformed latent codes to be constrained in the distribution of latent space. A disentangled learning strategy is employed to eliminate the entanglement among attributes. Extensive experiments have been conducted to validate the effectiveness of AdaTrans, especially with the large age gap and few labeled examples. In the future, we will explore preserving the background for face editing with the help of the intermediate features of StyleGAN generator.

Acknowledgement This work was supported in part by Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01), ZJ Lab, National Natural Science Foundation of China (Nos. 62176059 and 62101136), Shanghai Municipal of Science and Technology Project (No. 20JC1419500), and Shanghai Center for Brain Science and Brain-inspired Technology.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8293–8302, 2020.

[2] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022.

[3] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021.

[4] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proc. Int. Conf. Mach. Learn.*, pages 214–223, 2017.

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[7] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, volume 31, 2018.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

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

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4690–4699, 2019.

[11] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 2672–2680, 2014.

[13] Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[14] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[16] Xueqi Hu, Qiusheng Huang, Zhengyi Shi, Siyuan Li, Changxin Gao, Li Sun, and Qingli Li. Style transformer for image inversion and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11337–11346, 2022.

[17] H. Huang, S. Chen, J. Zhang, and H. Shan. PFA-GAN: Progressive face aging with generative adversarial network. *TIFS*, 2020.

[18] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1125–1134, 2017.

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

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proc. Int. Conf. Learn Represent.*, pages 1–26, 2018.

[22] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

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

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[25] Valentin Khrulkov, Leyla Mirvakhabova, Ivan Oseledets, and Artem Babenko. Latent transformations via neuralodes for gan-based image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14428–14437, 2021.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3673–3682, 2019.

[28] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 3730–3738, 2015.

[29] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative

adversarial networks. In *Int. Conf. Comput. Vis.*, pages 2794–2802, 2017.

[30] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 739–755. Springer, 2020.

[31] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[33] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.

[34] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020.

[35] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1532–1540, 2021.

[36] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

[37] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning*, pages 9786–9796. PMLR, 2020.

[38] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021.

[39] Yanbo Xu, Yueqin Yin, Liming Jiang, Qianyi Wu, Chengyao Zheng, Chen Change Loy, Bo Dai, and Wayne Wu. Transeditor: transformer-based dual-space gan for highly controllable facial editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7683–7692, 2022.

[40] Huiting Yang, Liangyu Chai, Qiang Wen, Shuang Zhao, Zixun Sun, and Shengfeng He. Discovering interpretable latent space directions of gans beyond binary attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12177–12185, 2021.

[41] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13789–13798, 2021.

[42] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pages 592–608. Springer, 2020.

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Int. Conf. Comput. Vis.*, pages 2223–2232, 2017.

[44] Peiye Zhuang, Oluwasanmi Koyejo, and Alexander G Schwing. Enjoy your editing: Controllable gans for image editing via latent space navigation. *arXiv preprint arXiv:2102.01187*, 2021.