A Latent Transformer for Disentangled Face Editing in Images and Videos

Xu Yao[1,2], Alasdair Newson[1], Yann Gousseau[1], Pierre Hellier[2]

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France

[2] InterDigital R&I, 975 avenue des Champs Blancs, Cesson-Sévigné, France

{xu.yao,anewson,yann.gousseau}@telecom-paris.fr, Pierre.Hellier@InterDigital.com
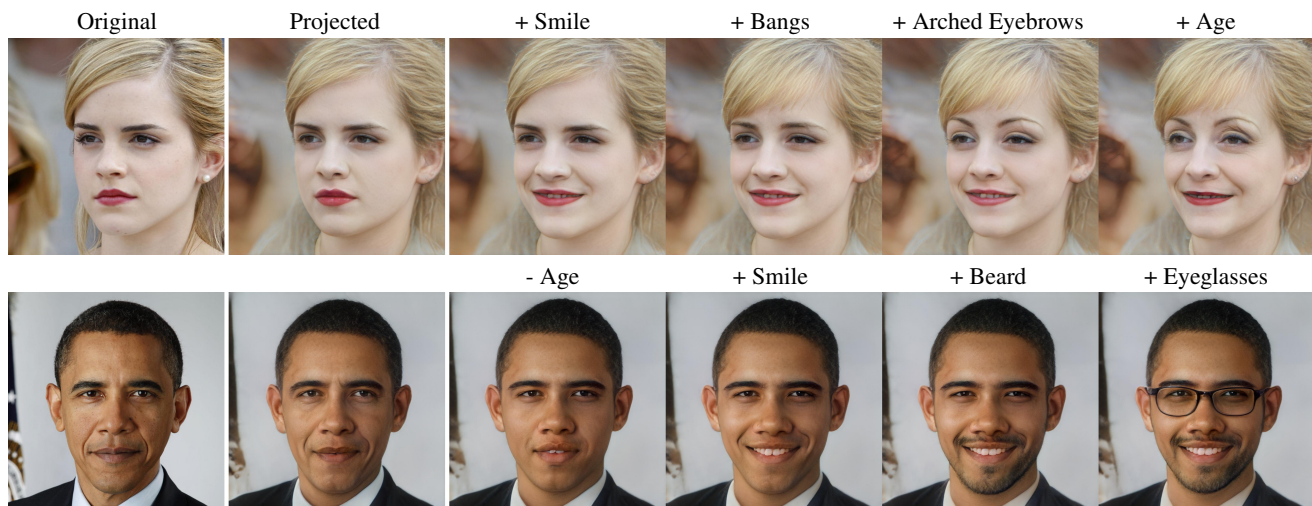
Figure 1: We project real images to the latent space of a StyleGAN generator and achieve sequential disentangled attribute editing on the encoded latent codes. From the original and the projected image, we can edit sequentially a list of attributes such as: 'smile', 'bangs', 'arched eyebrows', 'age', 'beard' and 'eyeglasses'. All results are obtained at resolution 1024×1024.

Abstract

High quality facial image editing is a challenging problem in the movie post-production industry, requiring a high degree of control and identity preservation. Previous works that attempt to tackle this problem may suffer from the entanglement of facial attributes and the loss of the person's identity. Furthermore, many algorithms are limited to a certain task. To tackle these limitations, we propose to edit facial attributes via the latent space of a StyleGAN generator, by training a dedicated latent transformation network and incorporating explicit disentanglement and identity preservation terms in the loss function. We further introduce a pipeline to generalize our face editing to videos. Our model achieves a disentangled, controllable, and identity-preserving facial attribute editing, even in the challenging case of real (i.e., non-synthetic) images and videos. We conduct extensive experiments on image and video datasets and show that our model outperforms other state-of-the-art methods in visual quality and quantitative evaluation.

Source codes are available at https://github.com/InterDigitalInc/latent-transformer.

1. Introduction

Facial attribute editing is a crucial task for photo retouching or in the film post-production industry. For example, many actors are retouched for beautification or other special makeup effects. For such tasks, it is highly desirable for artists to be able to control a facial attribute without affecting other informations. Consequently, a face editing method should rely on disentangled attributes and permit identity-preserving manipulations.

Earlier works based on deep learning focus on encoder-decoder based architectures [8, 17]. Despite the improvements in quality of recent results, these approaches are limited in resolution and generate noticeable artifacts on high resolution images. Therefore, they are not appropriate for high quality video editing. In addition, these methods are difficult to control, because the modification of one facial

attribute tends to modify other attributes.

Recently, generative networks have shown impressive progress in high quality image synthesis [5, 19, 20, 21]. Studies show that moving latent codes along certain directions in the latent space of generative models can result in variations of visual attributes in the corresponding generated images [2, 9, 34, 3, 41]. These assume that for a binary attribute, there exists a hyper-plane in the latent space which divides the data into two groups. However, this hypothesis has several limitations. Firstly, successful manipulations can only be achieved in well disentangled and linearized latent spaces. Although the latent space is disentangled compared to the image space, we show in this paper that achieving facial attribute manipulation with linear transformations is a very strong and limiting hypothesis. Furthermore, since these methods are trained on synthetic images (generated from random points in the latent space), their performance on real images (natural, "in-the-wild" photos) is less satisfying. This is an often ignored, but critical, problem.

In this work, we tackle the problem of editing facial attributes on real images. To address the aforementioned limitations, we propose a transformation network to navigate the latent space of the generative model. We project real images to the latent space of the state-of-the-art image generator StyleGAN and train our model on the projected latent codes. The transformation network generates disentangled, identity-preserving and controllable attribute editing results on real images. These key advantages allow us to extend our method to the case of videos, where stability and quality are of crucial importance. For this, we introduce a pipeline which achieves stable and realistic facial attribute editing on high resolution videos.

Our contributions can be summarized as follows. We propose a latent transformation network for facial attribute editing, achieving disentangled and controllable manipulations on real images with good identity preservation. Our method can carry out efficient sequential attribute editing on real images. We introduce a pipeline to generalize the face editing to videos and generate realistic and stable manipulations on high resolution videos.

The rest of the paper is organized as follows: In Section 2 we summarize the related works in facial attribute editing, disentangled representation and video editing. Section 3 presents our latent transformation network and the training details. In Section 4 we present the experimental results of disentangled attribute editing on real images, and provide qualitative and quantitative comparisons with state-of-the-art methods. We further present results of sequential attribute editing on real images and give an ablation study on the choice of loss compositions. In Section 5 we introduce the pipeline to apply facial attribute editing on videos and show experimental results on video sequences. We conclude the paper in Section 6.

2. Related works

Facial Attribute Editing. Previous works regarding facial attributes are extensive and mainly focus on images of limited resolution. An optimization-based approach by Upchurch *et al.* [37] showed that it is possible to achieve semantic transformations such as aging or adding facial hair by interpolating deep features in a pre-trained feature space. Another type of approach train feed-forward models for the attribute editing task. Attribute2image [42] proposed to train a Conditional Variational Auto-Encoder to generate attribute-conditioned images. With the success of generative networks in image synthesis, a number of studies [8, 15, 24, 25, 40] explored the possibility of training auto-encoders using adversarial learning. FaderNet [24] and StarGAN [8] proposed to disentangle different attributes in the latent space of auto-encoder and generate the output image conditioned on the target attributes. AttGAN [15] and STGAN [25] enhanced the flexible translation of attributes to improve the image quality by relaxing the strict constraints on the target attributes. Several studies investigate different possibilities to tackle high resolution images. CooGAN [7] proposed a patch-based local-global framework to process HR images in patches. Observing the great progress of generative networks in high quality image synthesis, Viazovetskyi *et al.* [38] trained the pix2pixHD model [39] for single attribute editing with the synthetic images generated by StyleGAN2 [21].

Disentangled Representations. In the paper of StyleGAN [20], the authors examined the effects of mixing two latent codes on the generated image (which is referred as style mixing), and found that each subset controls meaningful high-level attributes of the image. Inspired by this, some studies have attempted to explore disentangled representations in the latent space of generative networks, especially StyleGAN. One optimization-based method, Image2StyleGAN++ [2], carried out local editing along with global semantic edits on images by applying masked interpolation on the activation features of StyleGAN. Collins *et al.* [9] performed a k-means clustering on the activations of StyleGAN and detected a disentanglement of semantic objects, which enables further local semantic editing on the generated image. For high level semantic edits, Ganalyze [13] learned a manifold in the latent space of BigGAN [5] to generate images of different memorability. InterFaceGAN [34] proposed to learn a hyper-plane for a binary classification in the latent space, which one can use to manipulate the target facial attribute by simple interpolation. Following their work, StyleSpace [41] carried out a quantitative study on the latent spaces of StyleGAN [21] and realized a highly localized and disentangled control of the visual attributes. StyleFlow [3] achieved conditional exploration of the latent space by training conditional normalizing flows. StyleRig [35] introduced a method to provide a face rig-

like control over a pretrained and fixed StyleGAN via a 3D morphable face model. Yao *et al.* [43] proposed to navigate the latent space of StyleGAN in a non-linear manner to achieve disentangled manipulations of facial attributes. To find out the disentangled directions, some researches attempted to analyze the latent space of generative networks using Principal Component Analysis (PCA). PCAAE [30] presented an PCA auto-encoder, whose latent space is progressively increased during training and results in a separation of intrinsic data attributes into different components. GANSpace [18] performed PCA in the latent space of generative networks, explored the principal directions and discovered interpretable controls. The above-mentioned methods generally focus on manipulations of synthetic images, as it remains a challenge to project real images to the latent space of StyleGAN. Image2StyleGAN used optimization method to project real images to an extended latent space of StyleGAN, but the characteristics of which differ from the original latent space thus not suitable for manipulation. Some recent works [26, 31, 33, 45] trying to train an encoder together with the StyleGAN model. Although the images cannot be perfectly reconstructed, we see the possibility of carrying out attribute editing on real images using the disentanglement characteristics of the StyleGAN latent space.

Video-based Face Editing. Recent works on face video synthesis mainly address two problems: 1) generating a face video sequence from a sketch video or a reference image (often referred as face reenactment), and 2) facial attribute editing on videos. Garrido *et al.* [12] proposed an image-based reenactment system to achieve face replacement in video. Face2Face [36] presented an approach for real-time facial reenactment of a target video using non-rigid model-based bundling. Averbuch-Elor *et al.* [4] presented a technique to animate a still portrait with a driving video, but limited to mild movements. Kim *et al.* [22] proposed to use generative neural networks for re-animation of portrait videos, which transforms not only the facial expressions but also the full upper body and background. Most of these methods handle only low-quality video shots. A popular open-source deepfake system DeepFaceLab [29] has attracted much attention. Incorporating productivity tools such as manual face detection and landmark extraction tool, their pipeline generates high fidelity face swapping results on videos. To realize facial attribute editing directly on videos, Rav-Acha *et al.* [32] suggested to convert video frames to an "unwrap mosaic", paint and re-render the mosaic to videos. Despite satisfying results, computing the mosaic for each video shot is a long process and requires a smooth variation to construct successfully. Duong *et al.* [10] proposed an approach to generate age-progressed facial images in video sequences using deep reinforcement learning. Many recent works use deep learning techniques

to realize facial attribute editing on still images. However, up to now, only a few works have addressed video-based attribute editing problem [44].

3. Method

In this section, we propose a framework to edit faces in real images and videos via the latent space of StyleGAN.

3.1. Overview

For a given real image $\mathbf{I}$, we assume that we can compute a latent representation $\mathbf{w} \in \mathcal{W}$ associated to a generator $\mathbf{G}$. An inversion method is trained so that $\mathbf{I} \approx \mathbf{G}(\mathbf{w})$. We aim to train a latent transformer $\mathbf{T}$ in the latent space to edit a single attribute of the projected image $\mathbf{G}(\mathbf{w})$. The image synthesized from $\mathbf{T}(\mathbf{w})$ is denoted by $\mathbf{G}(\mathbf{T}(\mathbf{w}))$. It shares all the attributes with $\mathbf{G}(\mathbf{w})$ except the target attribute being manipulated.

To train the latent transformer, we propose a training framework that computes all the losses solely in the latent space $\mathcal{W}$. Let $\{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_N\}$ be a set of image attributes, where N is the total number of considered attributes. For each attribute $\mathbf{a}_k$, a different $\mathbf{T}_k$ is trained. To predict the attributes from the latent codes we use a latent classifier $\mathbf{C} : \mathcal{W} \rightarrow \{0,1\}^N$. $\mathbf{C}$ is pre-trained and then its weights are frozen during the training of $\mathbf{T}_k$. We train $\mathbf{T}_k$ with the following three objectives:

- To ensure that $\mathbf{T}_k$ manipulates attribute $\mathbf{a}_k$ effectively, we minimize the binary classification loss:

$$\mathcal{L}_{\text{cls}} = -y_k \log\left(\mathbf{p}_k\right) - (1 - y_k) \log\left(1 - \mathbf{p}_k\right), \quad (1)$$

  where $\mathbf{p}_k = \mathbf{C}(\mathbf{T}_k(\mathbf{w}))[k]$ is the probability of the target attribute and $y_k \in \{0,1\}$ is the desired label.

- To ensure that other attributes $\mathbf{a}_i, i \neq k$ remain the same, we apply an attribute regularization term:

$$\mathcal{L}_{\text{attr}} = \sum_{i \neq k} (1 - \gamma_{ik}) \, \mathbb{E}_{\mathbf{w},i}[||\mathbf{p}_i - \mathbf{C}(\mathbf{w})[i]||_2], \quad (2)$$

  where γ_{ik} is the absolute correlation value between $\mathbf{a}_i$ and the target attribute $\mathbf{a}_k$, measured on the training dataset. This regularization term is weighted based on correlation to prevent the attributes which are naturally correlated with the target from being over-constrained, *i.e.* 'chubby' and 'double chin'.

- To ensure that the identity of the person is preserved, we further apply a latent code regularization:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{\mathbf{w}}[||\mathbf{T}(\mathbf{w}) - \mathbf{w}||_2]. \quad (3)$$

The full objective loss can be described as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{attr}}\mathcal{L}_{\text{attr}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}, \quad (4)$$

where λ_{attr} and λ_{rec} are weights balancing each loss.
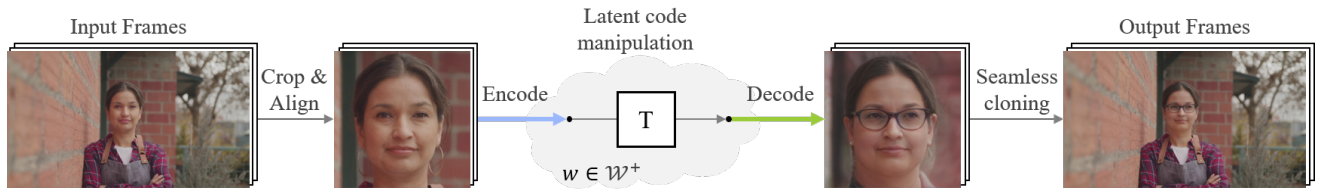
Figure 2: **Video manipulation pipeline.** Each input frame is cropped and aligned to a face image individually. A pretrained encoder [33] is used to encode the face images to the latent space $\mathcal{W}^+$ of StyleGAN [21]. The obtained latent codes are processed by the proposed latent transformer **T** to realize the attribute editing. The manipulated latent codes are further decoded by StyleGAN to generate the manipulated face images, which are blended with the original input frames to get the output frames.

3.2. Training and models

To realize attribute editing on real images, we first need to compute the corresponding latent codes in the latent space of StyleGAN. Unlike a traditional generative network which feeds random vectors as an input to the first layers, StyleGAN has a different design. The generator takes a constant tensor as input, while each convolution layer output is controlled by style codes via adaptive instance normalization layers [16]. A Gaussian random latent code $\mathbf{z} \in \mathcal{Z}$ is first passed through a mapping network to get an intermediate latent code $\mathbf{w} \in \mathcal{W}$, which is further specialized by learned affine transforms to style codes $\mathbf{y}$. Given a target image $\mathbf{x}$, finding the corresponding latent code in $\mathcal{W}$ remains difficult, and the quality of the reconstruction is not fully satisfactory. Image2StyleGAN [1] go a step further by computing the latent code in an extended latent space $\mathcal{W}^+$, where latent code $\mathbf{w}$ controlling each layer may be different, whereas the original setting requires them to be the same. The target image is thus better reconstructed from the latent code obtained in $\mathcal{W}^+$.

In our approach, we train the latent transformer **T** in the latent space $\mathcal{W}^+$, which is specifically designed for the projection of latent codes of real images. To prepare the training data, we compute the latent codes in $\mathcal{W}^+$ for each image in CelebA-HQ dataset [19], using a pre-trained StyleGAN encoder proposed by Richardson *et al.* [33]. Combined with the annotations for each image, we obtain the "latent code - label" pairs as our training data.

Latent Classifier. To predict attributes on the manipulated latent codes, we train an attribute classifier **C** on the "latent code - label" pairs. The classifier consists of three fully connected layers with ReLU activations in between. **C** is *fixed* during the training of the latent transformer.

Latent Transformer. Given a latent code $\mathbf{w} \in \mathcal{W}^+$, the latent transformer **T** generates the direction for a single attribute modification, where the amount of changes is controlled by a scaling factor α. The network is expressed with a single layer of linear transformation f:

$$\mathbf{T}(\mathbf{w}, \alpha) = \mathbf{w} + \alpha \cdot f(\mathbf{w}). \tag{5}$$

During training the scaling factor α is set according to the probability p of the target attribute of the input latent code ($1 - p$ for $p < 0.5$, $-p$ for $p > 0.5$). At test time, α can be sampled from $[-1, 1]$ or set beyond this range based on the desired amount of changes.

3.3. Video manipulation

In this section, we propose a pipeline which applies the image editing method to the case of videos. The encoding process ensures that the encoded latent codes of two consecutive frames are similar to each other. Therefore, we can reconstruct a face video using the frames projected to the latent space of StyleGAN, which provides the basics for the next manipulation step. Thanks to the stability of our proposed latent transformer, the manipulation does not affect the consistency between the latent codes and generates stable edits on the projected frames. An overview of our proposed pipeline is presented in Figure 2. The pipeline consists of three steps: pre-processing, image editing and seamless cloning.

Pre-processing. In order to edit the video in the latent space of StyleGAN, we first extract face images from the frames, according to the StyleGAN setting. We crop and align each frame around the face, following the pre-processing step of FFHQ dataset [20], on which the StyleGAN is pretrained. For face alignment we detect landmarks independently on each frame using a state-of-the-art method [6]. To avoid jitter, we further process the landmarks using optical flow between two consecutive frames and a Gaussian filtering along the whole sequence. All frames are cropped and aligned to have eyes at the center and resized to 1024×1024.

Image editing. In this step, we apply our manipulation method on the processed face images. Each frame is encoded to the latent space of StyleGAN using the pre-trained encoder [33]. The encoded latent codes are processed by the proposed latent transformer to realize the attribute editing. The manipulated latent codes are further decoded by StyleGAN to generate the manipulated face images.

Seamless cloning. We use Poisson image editing method [28] to blend the modified faces with the original input frames. In order to blend only the face area, we use the segmentation mask obtained from the detected facial land-
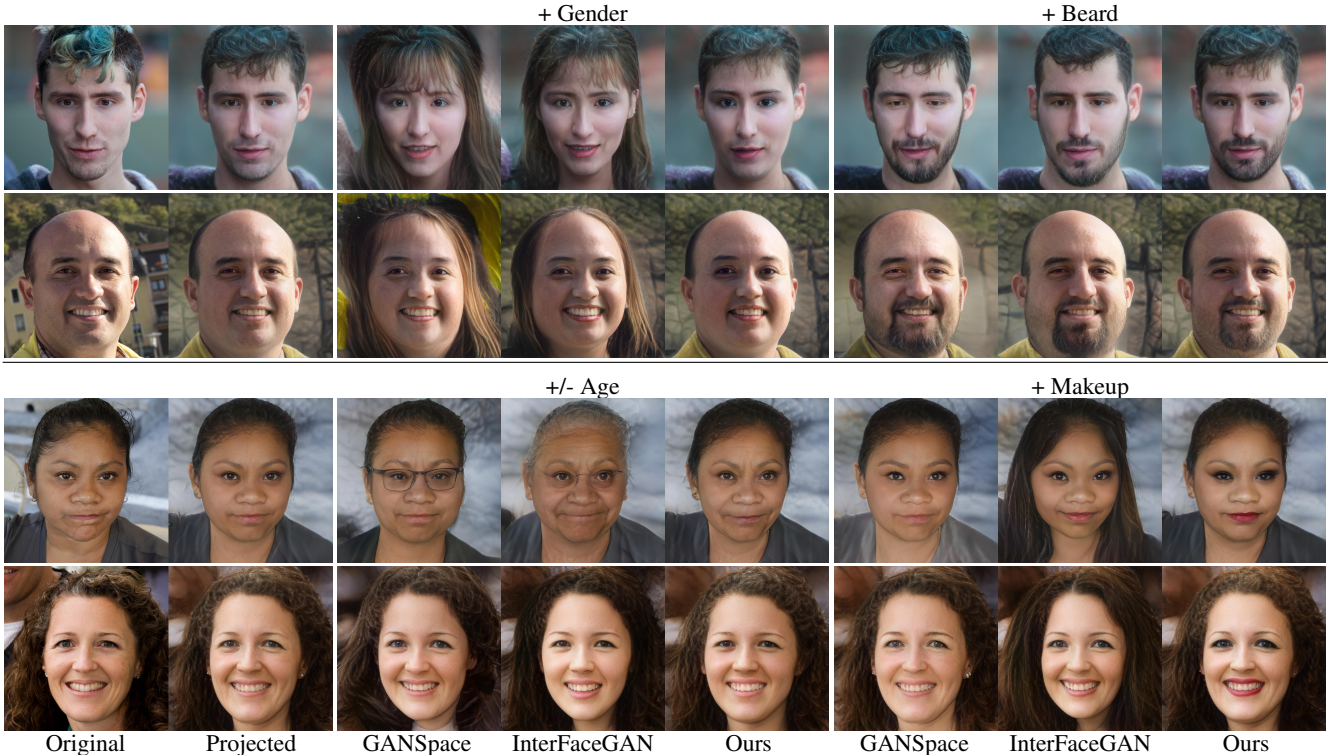
Figure 3: **Disentangled facial attribute editing on real images.** The first two columns show the original image and the projected image reconstructed with the encoded latent code in StyleGAN. From the 3rd column in each subfigure, from left to right are the manipulation result of GANSpace [18], that of InterFaceGAN [34] and ours. Compared to recent approaches, our method achieves a controllable, disentangled and realistic editing, where the person's identity is preserved.

marks.

3.4. Implementation details

We explore disentangled manipulations in the latent space of StyleGAN2 [21] pretrained on FFHQ dataset [20]. In this paper, we conduct all the experiments with the latest StyleGAN2. For simplicity, when we mention StyleGAN, it refers to the latest version - StyleGAN2. To prepare the training data, we project images of CelebA-HQ [19] to the latent space $\mathcal{W}^+$ of StyleGAN using a pre-trained encoder [33] and obtain the corresponding latent codes. CelebA-HQ contains $30K$ face images at 1024^2 resolution, each annotated by 40 facial attributes. We train a separate latent transformer for each facial attribute.

To predict the attributes on the latent codes, we train a latent classifier on the 'latent code - label' pairs, which is fixed during the training of the latent transformer. The model is designed to predict all the 40 attributes together, and trained with binary cross entropy loss.

For the training of the latent transformer, we use 90% of the prepared data for training set and train the model for $100K$ iterations, with a batch size of 32. The weights balancing each loss are set to $\lambda_{\mathrm{attr}} = 1$ and $\lambda_{\mathrm{rec}} = 10$. We use Adam optimizer [23] with a learning rate of 0.001, $\beta_1 = 0.9$

and $\beta_2 = 0.999$.

4. Experiments

4.1. Disentangled manipulation on real images

We compare our results with two state-of-the-art methods: InterFaceGAN [34] and GANSpace [18]. For a fair comparison, we follow the methodology of InterFaceGAN and train their model on StyleGAN2 for the attributes of CelebA-HQ using their official code. The official implementation of GANSpace on StyleGAN2 is available. For the evaluation data we use FFHQ, independent from the training of all methods. We project the real images of FFHQ to the latent space $\mathcal{W}^+$ of StyleGAN using the pretrained encoder [33], and manipulate the latent codes using each method with the suggested magnitude of edits (3 for InterFaceGAN, specified range based on attributes for GANSpace and 1 for our method). Figure 3 compares the manipulation results on the attributes which are available for all methods ('gender', 'age', 'beard' and 'makeup'). Our method achieves better disentangled manipulations. For example, when changing 'gender', both GANSpace and InterFaceGAN modify the hairstyle, and when changing 'age', GANSpace adds eyeglasses and InterFaceGAN af-
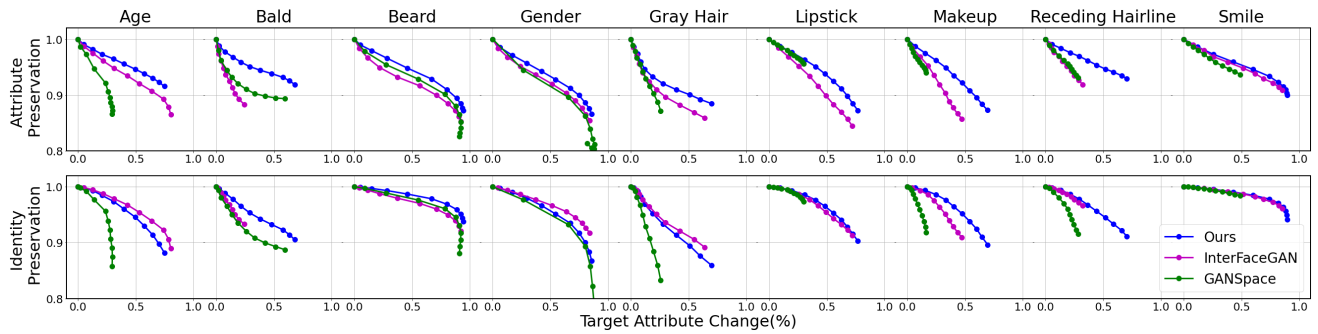
Figure 4: **Attribute and identity preservation vs. target attribute change.** For each method, we edit each target attribute with 10 different scaling factors ($\{0.2 \cdot d, 0.4 \cdot d, ..., 2 \cdot d\}$, d is the magnitude of change suggested in each method), and generate the modified images. Attribute preservation rate and identity preservation score are measured on the output images. In the figure, each point corresponds to a scaling factor, where the position x indicates the target attribute change rate (the fraction of the samples with target attribute successfully changed among all the manipulations). In the upper sub-figure, the position y indicates the average attribute preservation rate on the other attributes. In the bottom sub-figure, the position y indicates the average identity preservation score. Ideally, we want higher attribute and identity preservation for the same amount of change on the target attribute (higher curve is better).
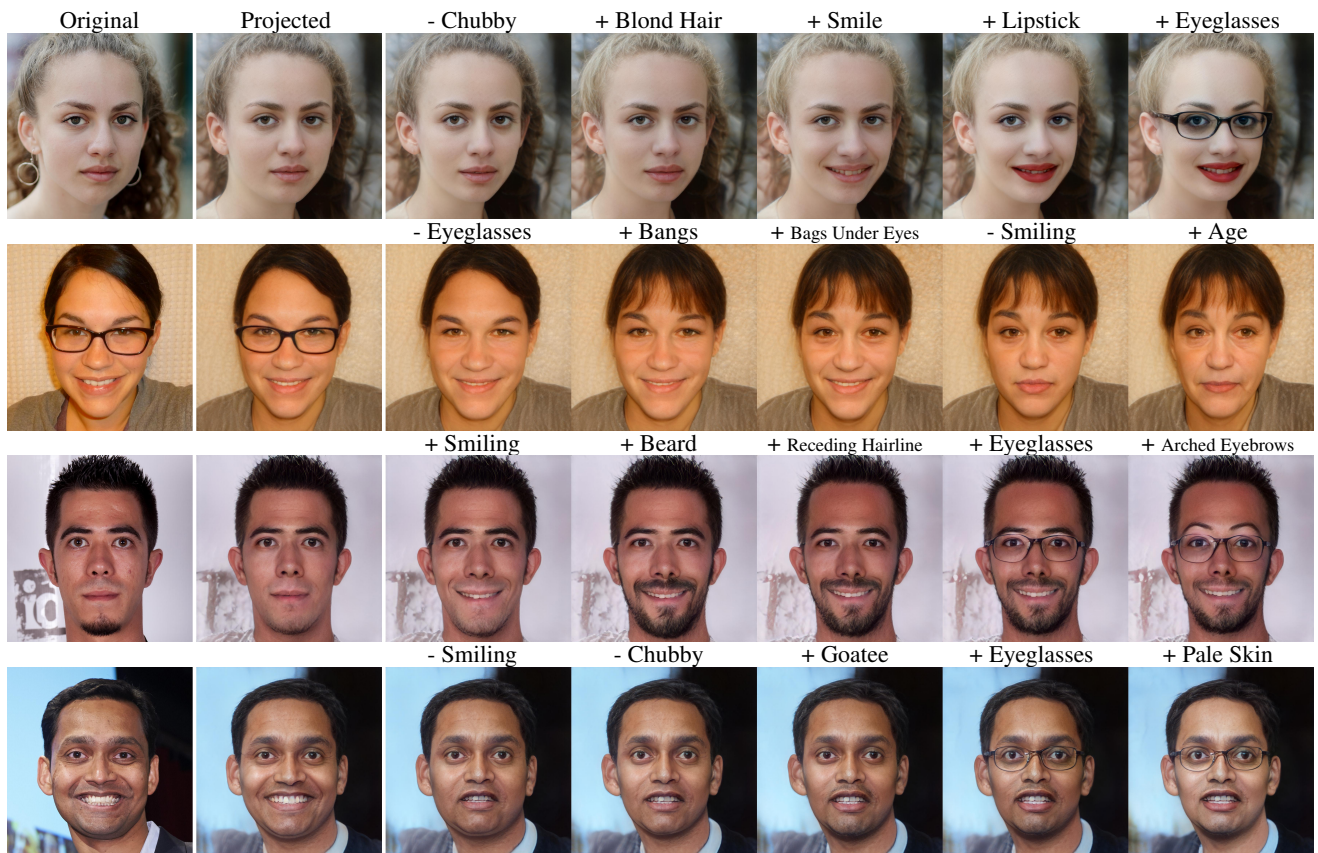

Figure 5: **Sequential facial attribute editing on real images**. Given an input image, we manipulate a list of attributes sequentially, where each time a single attribute is modified from the previous latent representation.

fects smile. In contrast, our method succeeds to separate hairstyle from 'gender' and disentangle 'eyeglasses' from 'age', thanks to the attribute and latent code regularization terms. The directions of GANSpace are discovered from PCA so that they may control several attributes simultane-ously. For InterFaceGAN, no attribute preservation is ap-plied when searching the semantic boundary. Compared with their methods, our editing results are of better visual quality and preserve the original facial identities better.

4.2. Quantitative evaluation

We compare our method quantitatively with GANSpace and InterFaceGAN using three metrics: target attribute change rate, attribute preservation rate and identity preservation score. Given a set of manipulated samples, the target attribute change rate refers to the percentage of the samples with target attribute varied among all the samples. The attribute preservation rate indicates the proportion of unchanged samples on the other attributes apart from the target. The identity preservation score refers to the average cosine similarity between the VGG-Face [27] embeddings of the original projected images and the manipulated results.

For the evaluation data, we project the first $1K$ images of FFHQ into the the latent space $\mathcal{W}^+$ of StyleGAN using the pre-trained encoder [33]. For each input image and each method, we edit each attribute with 10 different scaling factors ($\{0.2 \cdot d, 0.4 \cdot d, ..., 2 \cdot d\}$, d is the magnitude of change suggested by each method) and generate the corresponding images. To predict the attributes on the modified images, we use a state-of-the-art facial attribute classifier [14], independent from all methods. Based on the classification result, we consider an attribute active if its probability is greater than 0.5, otherwise inactive (*i.e.*, w/ bangs versus w/o bangs). For each scaling factor, we compute the target attribute change rate, and the attribute preservation rate averaged on the other attributes. To check the identity preservation, we compute the average identity preservation score. Figure 4 presents the attribute and identity preservation w.r.t. the target attribute change on the attributes detected by all methods. For attributes like 'beard', 'gender' and 'smile', all the methods handle well. For other attributes, we observe that for the same amount of change on the target attribute, our approach has a higher attribute preservation rate while achieving a comparable or better identity preservation score. Overall our method achieves better disentanglement and better identity preservation than existing methods.

4.3. Sequential editing

Thanks to the disentanglement property of our approach, it achieves sequential modifications of several attributes on real images. We project real images of FFHQ to the latent space $\mathcal{W}^+$ of StyleGAN using the pre-trained encoder [33], and apply manipulations on a list of attributes sequentially. As shown in Figure 5, our method achieves disentangled and realistic modifications, and is not limited to a defined order of attributes.

4.4. Loss analysis

We carry out an ablation study to analyze the effects of each regularization terms in Eq. (4). In our proposed baseline, the weights balancing each regularization term are set

Figure 6: Ablation study on the loss composition in Eq. (4). Each row corresponds to a single attribute editing. From left to right: original image, manipulation results of different scenarios. When $\lambda_{\text{attr}} = 0$, editing target attribute affects the others. When $\lambda_{\text{rec}} = 0$ it fails to preserve the facial identity. Our proposed baseline yields better disentangled results with identity preserved.
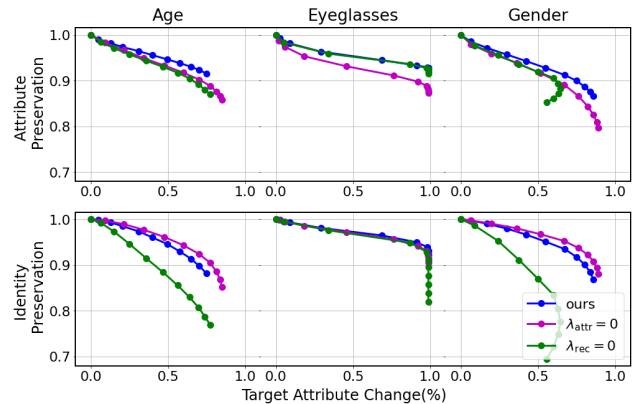

Figure 7: Quantitative comparison of different loss compositions. For each scenario, we edit each attribute with 10 scaling factors ($\{0.2, 0.4, ..., 2\}$) and measure the attribute preservation rate and identity preservation score on the modified images. Each point marker represents a scaling factor. The upper sub-figure presents the average attribute preservation rate w.r.t. target attribute change rate. The bottom sub-figure presents the average identity preservation score w.r.t. target attribute change rate. Our chosen baseline has a better trade-off between attribute and identity preservation.

to $\lambda_{\text{attr}} = 1$ and $\lambda_{\text{rec}} = 10$. We compare with two different scenarios: $\lambda_{\text{attr}} = 0$ (w/o attribute regularization) and $\lambda_{\text{rec}} = 0$ (w/o latent regularization). As shown in Figure 6, when $\lambda_{\text{attr}} = 0$ the output is not only manipulated on the target attribute but also affected on the other attributes, *e.g.*, beard added when changing gender, mouth affected when changing age. When $\lambda_{\text{rec}} = 0$, the manipulated images fail to preserve the original facial identity. Balancing each
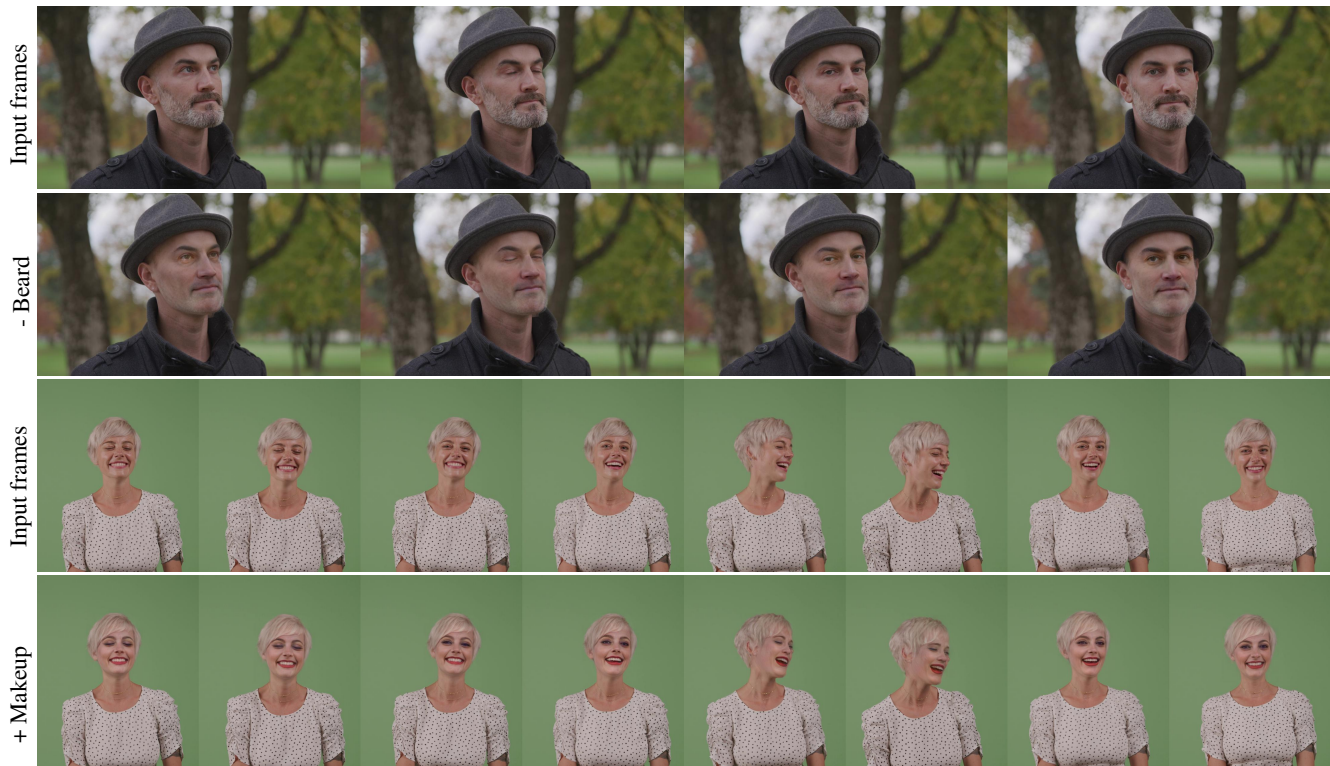
Figure 8: **Facial attribute editing on videos.** In each sub-figure, the top row shows the input frames, the bottom row shows the output frames obtained from our proposed video manipulation pipeline. A face image is cropped and aligned from each frame, and encoded to latent space of StyleGAN. The encoded latent code is passed into the latent transformer to get the target attribute varied, then decoded to an output face and blended with the input frame.

term, our proposed baseline achieves attribute editing with better disentanglement and identity preservation. Figure 7 provides a quantitative comparison of the three scenarios. As can be observed, our chosen baseline preserves the other attributes best, with the same amount of attribute change, without sacrificing the identity preservation.

4.5. Video editing

We apply our manipulation method on real-world videos collected from FILMPAC library [11]. Figure 8 shows the qualitative results of facial attribute editing on videos obtained from our proposed pipeline. From each input frame, we crop and align a face image and encode it to the latent space of StyleGAN with a pre-trained encoder [33]. The encoded latent code is processed by the latent transformer to vary the target attribute, then decoded to an output face image and blended with the input frame. As can be seen from the results, our proposed method succeeds in removing the facial hair or adding the makeup, without influencing the consistency between the frames. Nevertheless, we also observe that the proposed method has more difficulty handling extreme pose (side face), which may be due to the limitation of the generation capacity of the StyleGAN model. Please

refer to the supplementary material to check the videos and see more video results.

5. Conclusion and future work

In this paper, we have presented a latent transformation network to perform facial attribute editing in real images and videos via the latent space of StyleGAN. Our method generates realistic manipulations with better disentanglement and identity preservation than other approaches. We have extended our method to the case of videos, achieving stable and consistent modifications. To the best of our knowledge, this is the first work to present stable facial attribute editing on high resolution videos. Some future work could be dedicated to improve both the applicability of the method and the performance on videos. In particular, the method has difficulty handling side poses due to the fact that StyleGAN has difficulties in generating faces in side poses. This could be potentially addressed by jointly training the StyleGAN encoder with the generator, or training an improved StyleGAN generator using images with better diversity of poses.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 4

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

[3] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *arXiv e-prints*, pages arXiv–2008, 2020. 2

[4] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. Bringing portraits to life. *ACM Transactions on Graphics (TOG)*, 36(6):1–13, 2017. 3

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 2

[6] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 4

[7] Xuanhong Chen, Bingbing Ni, Naiyuan Liu, Ziang Liu, Yiliu Jiang, Loc Truong, and Qi Tian. Coogan: A memory-efficient framework for high-resolution facial attribute editing. In *European Conference on Computer Vision*, pages 670–686. Springer, 2020. 2

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul 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. 1, 2

[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. 2

[10] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, Nghia Nguyen, Eric Patterson, Tien D Bui, and Ngan Le. Automatic face aging in videos via deep reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10013–10022, 2019. 3

[11] FILMPAC. Filmpac footage boutique library. https://filmpac.com, 2017. 8

[12] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. Automatic face reenactment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4217–4224, 2014. 3

[13] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5744–5753, 2019. 2

[14] Keke He, Yanwei Fu, Wuhao Zhang, Chengjie Wang, Yu-Gang Jiang, Feiyue Huang, and Xiangyang Xue. Harnessing synthesized abstraction images to improve facial attribute recognition. In *IJCAI*, pages 733–740, 2018. 7

[15] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019. 2

[16] 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. 4

[17] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 1

[18] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Proc. NeurIPS*, 2020. 3, 5

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 2, 4, 5

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2, 4, 5

[21] 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. 2, 4, 5

[22] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 5

[24] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5967–5976, 2017. 2

[25] 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 Conference on Computer Vision and Pattern Recognition*, pages 3673–3682, 2019. 2

[26] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7198–7211. Curran Associates, Inc., 2020. 3

[27] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. *British Machine Vision Conference*, 2015. 7

[28] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003. 4

[29] Ivan Petrov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Jian Jiang, Luis RP, Sheng Zhang, Pingyu Wu, et al. Deepfacelab: A simple, flexible and extensible face swapping framework. *arXiv preprint arXiv:2005.05535*, 2020. 3

[30] Chi-Hieu Pham, Saïd Ladjal, and Alasdair Newson. Pcaae: Principal component analysis autoencoder for organising the latent space of generative networks. *arXiv preprint arXiv:2006.07827*, 2020. 3

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

[32] Alex Rav-Acha, Pushmeet Kohli, Carsten Rother, and Andrew Fitzgibbon. Unwrap mosaics: A new representation for video editing. In *ACM SIGGRAPH 2008 papers*, pages 1–11. 2008. 3

[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. 3, 4, 5, 7, 8

[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. 2, 5

[35] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020. 2

[36] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 3

[37] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7064–7073, 2017. 2

[38] Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. Stylegan2 distillation for feed-forward image manipulation. In *European Conference on Computer Vision*, pages 170–186. Springer, 2020. 2

[39] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 2

[40] Rongliang Wu and Shijian Lu. Leed: Label-free expression editing via disentanglement. In *European Conference on Computer Vision*, pages 781–798. Springer, 2020. 2

[41] 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. 2

[42] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016. 2

[43] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. Learning non-linear disentangled editing for stylegan. In *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021. 3

[44] Xin Zheng, Yanqing Guo, Huaibo Huang, Yi Li, and Ran He. A survey of deep facial attribute analysis. *International Journal of Computer Vision*, pages 1–33, 2020. 3

[45] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. Indomain gan inversion for real image editing. In *European Conference on Computer Vision*, pages 592–608. Springer, 2020. 3