Text-conditional Attribute Alignment across Latent Spaces for 3D Controllable Face Image Synthesis

Feifan Xu[1], Rui Li[2*], Si Wu[1,3,4*], Yong Xu[1,3,4] and Hau San Wong[5]
[1]School of Computer Science and Engineering, South China University of Technology
[2]Department of Computer Science, Shantou University
[3]Peng Cheng Laboratory [4]PAZHOU LAB
[5]Department of Computer Science, City University of Hong Kong
cs_feifan@mail.scut.edu.cn, ruili@stu.edu.cn, {cswusi, yxu}@scut.edu.cn, cshswong@cityu.edu.hk

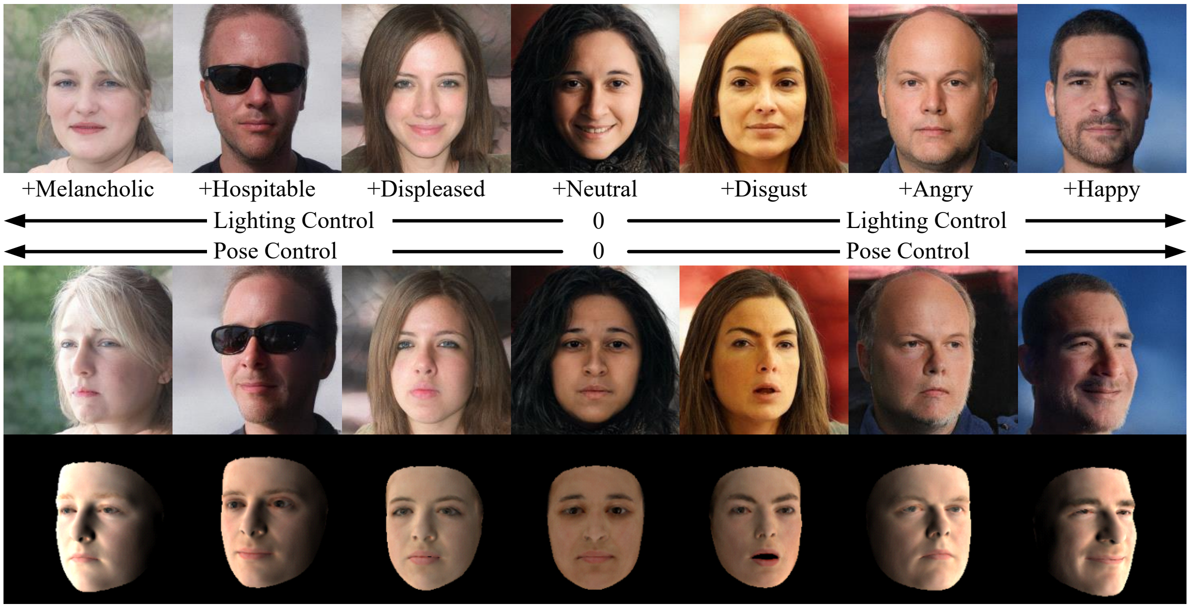

Figure 1. The examples to demonstrate that the proposed method that can explicitly control face expression, pose, and illumination in a fine-grained manner. The manipulated images are consistent with the 3D rendered images shown in the last row.

Abstract

With the advent of generative models and vision-language pretraining, significant improvement has been made in text-driven face manipulation. The text embedding can be used as target supervision for expression control. However, it is non-trivial to associate with its 3D attributes, i.e.*, pose and illumination. To address these issues, we propose a Text-conditional Attribute aLignment approach for 3D controllable face image synthesis, and our model is referred to as TcALign. Specifically, since the 3D rendered image can be precisely controlled with the 3D face representation, we first propose a Text-conditional 3D Editor to produce the target face representation to realize text-driven manipulation in the 3D space. An attribute embedding space spanned by the target-related attributes embeddings is also introduced to infer the disentangled task-specific direction. Next, we train a cross-modal latent mapping network conditioned on the derived difference of 3D representation to infer a correct vector in the latent space of StyleGAN. This correction vector learning design can accurately transfer the attribute manipulation on 3D images to 2D images. We show that the proposed method delivers more precise text-driven multi-attribute manipulation for 3D controllable face image synthesis. Extensive qualitative and quantitative experiments verify the effectiveness and supe-*

[*] Co-corresponding author.

riority of our method over the other competing methods.

1. Introduction

The automatic manipulation of face characteristics *e.g*. expression, pose, and illumination has a wide range of applications in movie production and or game design. Among various techniques, style-based generative models [22, 23] learn a disentangled latent space ($\mathcal{W}$) and synthesize high-fidelity images. Therefore, instead of training a generator from scratch, many works [2, 39, 40, 49] explore the latent space to discover semantic latent paths for the manipulation of different attributes. However, existing methods typically involve a well-trained attribute classifier or a large number of annotated data for supervision, which are cumbersome and only enable the manipulation of predefined attributes.

Recently, with the advent of the Contrastive Language-Image Pre-training (CLIP) model [34], some works [30, 32, 51] leveraged a semantic consistency constraint based on CLIP to realize text-driven image manipulation, which is more flexible and user-friendly. However, these methods, *e.g*. StyleCLIP [32] are mainly designed for appearance control with a text prompt, which is not sufficient to achieve fine-grained control, particularly on 3D physical attributes. On the other hand, there are some works targeted for 3D controllable face synthesis [10, 11, 42]. They often learn or construct a global 3D face representation as condition to synthesize a 2D face, so that the desired control can be achieved by editing the 3D representation. However, the quality of generated faces is often sub-optimal due to the difficulty in directly mapping 2D and 3D spaces. There is also no text-driven control which limits their applicability.

In this paper, we propose TcALign which incorporates 3D priors [4, 12] based on StyleCLIP to enable text- and 3D-control, simultaneously. The key is to ensure that the synthesized image is consistent with its 3D face representation manipulated by a text prompt in all physical characteristics as shown in Figure 1 (last row). However, there involve three different modalities, *i.e*., text, 2D, and 3D. Learning precise attribute alignment among them within one model is challenging, and we observe that naively using a target text embedding to infer the corresponding 3D or latent transformation often results in inaccurate face control or non-target attribute change.

To address the above issues, we first learn a text-conditional 3D Editor to enable 3D editing with textual prompts (shown in Figure 2) in 3D space. To precisely build relationships between text and 3D modalities for accurate control, we introduce an Attribute Embedding Space (AES) to infer the disentangled target direction for 3D image manipulation, and the text-conditional 3D image embedding can reflect the target attributes which will be used to produce the corresponding 3D face representation with 3D Editor, thus enables text-conditional synthesis for 3D images. To achieve high-quality 2D face synthesis, we further adopt a cross-modal latent mapping network based on the previous 3D transformation information to infer a correction vector within the $\mathcal{W}$ space. The final target image is generated from the latent code of the source image added by the correction vector via the pretrained StyleGAN. Compared to the global latent code prediction, this correction vector learning design leads to more precise multi-attribute manipulation (Figure 1). We have performed both qualitative and quantitative experiments to verify the effectiveness and superiority of our method over recent competing methods. To summarize, our contributions include the followings:

- The proposed approach enables both 3D-aware and text-driven face control on expressions, poses, and illuminations, which is important in many real-world face synthesis applications.
- We incorporate AES in 3D Editor to achieve accurate 3D image control with text prompts via the discovery of disentangled manipulation direction.
- We propose a cross-modal latent mapping network conditioned on the difference of 3D representation to infer a correction vector within the latent space of StyleGAN, which transfers the precise attribute rendering on 3D images to 2D images.

2. Related work

Generic face editing. Previous methods heavily adopted Conditional-GAN [31] for face editing. Pix2Pix [20] is one of the pioneering works for image translation which requires paired data. To improve the data efficiency, cycle-consistency regularization is proposed in CycleGAN [56] which realizes unpaired image translation. Afterwards, many works have been proposed for more diverse types of translation, *e.g*., MUNIT [19] aims for multi-modal translation by introducing content- and style-encoder; StarGAN [6] enables multi-domain translation with a single generator, *etc*. Except for image translation, the generator can be conditioned on some *predefined* factors [16, 43] for controllable editing. Recently, some studies used text as the condition to guide face synthesis for flexible control. A pioneering work [37] trained a CGAN with the text embedding from a language model. Meanwhile, various techniques were proposed to improve the generation quality, such as the stacked architecture [52], visual-content disentanglement [29], cross-modality fusion [27], contrastive learning [53], *etc*. However, these methods trained the generator from scratch on a specific dataset which is less effective for real-world applications. As a result, some recent works adopted a pretrained generator for high-quality and generalizable synthesis [24, 51].

Generative prior-based face manipulation. The latent spaces of a pretrained StyleGAN [22, 23] demonstrate promising disentangled properties. Therefore, many works [1, 7, 39, 40, 44, 49] used StyleGAN as the backbone and investigated its latent spaces for performing a wide variety of image manipulations. One direction is to train a network to discover the corresponding path in the latent spaces for manipulating a desired attribute [15, 40, 49]. Another direction is to encode a reference image into the latent spaces of StyleGAN to realize exemplar-based manipulation [1, 3, 38, 46]. Meanwhile, with the advent of CLIP [34] with the semantic visual and textual embeddings, there are increasing number of works combining StyleGAN and CLIP for image manipulation with text descriptions [32]. For example, TediGAN [50] aligns the text and image embeddings in the latent spaces of StyleGAN for generation. StyleCLIP [32] proposed a linguistic-visual semantic loss based on CLIP to guide image synthesis with text input. DeltaEdit [30] focused on improving the distribution alignment between the textual and visual embedding by identifying a delta image and text space. For multiple-attribute manipulation, StyleFlow [2] used a conditional flow model to learn sequential transformations in which each transformation corresponds to one attribute edit, while TUSLT [47] introduced an auxiliary attribute classifier together with CLIP to manipulate multiple attributes with a single pass. Despite the significant progress, it is still difficult to enable precise 3D physical control, and undesired manipulation may be incurred due to the large gap between 3D and text spaces.

3D-aware face generation. Our method is related to 3D-aware manipulation methods which incorporate 3D priors for face generation. Shi *et al.* [41] directly disentangled the latent space of StyleGAN into 3D components *i.e.*, texture, shape, viewpoint, lighting. Kim *et al.* [25] leveraged 3DMM to generate faces with the 3D coefficients estimated from input images. DiscoFaceGAN [10] trained a generator to imitate the rendered 3D faces via 3DMM. DiffusionRig [11] used DECA [13] to extract 3D coefficients from a single image which serve as the condition for a diffusion model [18], these coefficients can be easily edited to achieve the desired 3D control. However, most 3D-aware face generation methods do not support text-driven manipulation. In contrast, our method is the first to use an AES to infer the disentangled direction for a 3D image, and propose a 3D Editor conditioned on the manipulated embedding to realize text manipulation in 3D space. Next, the resulting 3D difference information is used as the condition for a cross-modal latent mapper, which enables 3D-aware manipulations in the latent space. This framework allows 3D controllable face synthesis with text, and the manipulated image is consistent with its 3D face representation.

3. Proposed Method

Given a 2D face image, our goal is to precisely control its expression, pose, and illumination with text descriptions. Recent works explored the $\mathcal{W}$ space of StyleGAN for high-quality generation, and leveraged CLIP linguistic-visual semantic embeddings for supervisions. We further consider incorporating 3D priors which provide a reference 3D space for fine-grained face control. However, discrepancies among different modalities often result in incorrect controls or undesired artifacts. Therefore, improving alignments among different spaces is essential for enhancing the performance of fine-grained text-driven face manipulation.

Toward this end, we propose a two-stage training pipeline. First, since a 3D face image has less variations and can be easily edited with its 3D representation, we propose a Text-conditional 3D Editor (Γ) to infer the target 3D face representation θ_{target} based on the target text t and original 3D face representation θ_0, which realizes text manipulation in 3D space. Second, to transfer the manipulation in 3D space to the latent space for high-quality 2D face image synthesis, a cross-modal latent mapping network (Φ) conditioned on the 3D difference information $\Delta\theta = \theta_{target} - \theta_0$ aims to predict the correction vector Δw in the $\mathcal{W}$ space of StyleGAN (G_{sty}). Compared with global latent code prediction, two residual transformations are easily aligned for cross-modal mapping. As a result, text-driven 3D controllable face image synthesis can be accurately performed. The overall framework is shown in Figure 2.

3.1. 3D Face Representation

In this work, we incorporate 3D priors for fine-grained face attribute control. Based on [9], 3D priors include the 3DMM, illumination model, and camera model.

Specifically, in 3DMM, the Basel Face Model [33] is used for face shape and texture, and face expression is based on [14]. Therefore, the coefficients in 3DMM include $\alpha \in \mathbb{R}^{80}$, $\delta \in \mathbb{R}^{80}$ and $\beta \in \mathbb{R}^{64}$ for identity, texture and expression control, respectively. The illumination model is based on Spherical Harmonics (SH) [35, 36], where $\gamma \in \mathbb{R}^{27}$ is used for illumination control. The coefficients $p \in \mathbb{R}^{6}$ for the perspective camera model control the face poses, including pitch and yaw rotations.

Therefore, the 3D face representation is $\theta = (\alpha, \delta, \beta, \gamma, p) \in \mathbb{R}^{257}$, which can be used to render a 3D image with differentiable mathematical operations $\mathcal{R}$ via the 3DMM database. The rendering process shown in Figure 2 includes $\theta = \mathcal{R}_{inv}(x)$ and $x' = \mathcal{R}(\theta)$. $\mathcal{R}_{inv}$ is a pretrained 3D Predictor [9], which can be viewed as an *inverse rendering* process. x' denotes the rendered 3D image.

3.2. Text-conditional 3D Editing

We consider using the rendered 3D images for CLIP image embedding extraction which enhances 3D physical control, and the corresponding embedding without irrelevant background is more disentangled for text control. Given the original x'_0 and a target attribute text t, Γ aims to produce

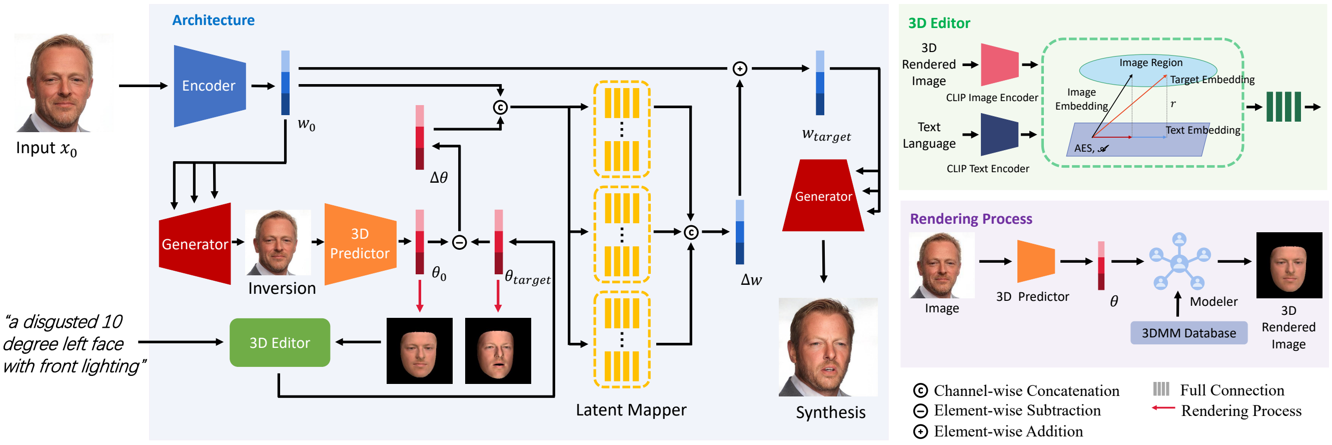

Figure 2. Overview of TcALign which consists of two learnable modules: The Text-conditional 3D Editor infers the target 3D face representation to realize text-driven control on 3D images. The Cross-modal Latent Mapping network predicts the latent transformation based on 3D difference information to achieve 3D controllable face synthesis. Encoder (e4e), Generator (StyleGAN), 3D Predictor (Deep3DFR), CLIP Encoders are fixed during training.

the corresponding target 3D face representation θ_{target}. In this case, the text can precisely describe the manipulations in 3D space.

In Figure 2 (3D Editor module), CLIP contains an image encoder E_{img} and a text encoder E_{txt}, which map x'_0 and t into a shared 512-d space, *i.e.*, $e_{x'_0} = E_{img}(x'_0)$ and $e_t = E_{txt}(t)$. Although CLIP demonstrates strong semantic representations, recent studies [30, 54] observed that the corresponding image embedding and text embedding are not well aligned owing to the modality gap. Therefore, instead of using e_t as the target embedding for supervision, we introduce an attribute embedding space (AES) $\mathscr{A}$ to incorporate text information e_t by discovering the relevant disentangled transformation for $e_{x'_0}$. Specifically, we first project $e_{x'_0}$ into $\mathscr{A}$, which is spanned with a set of basis attribute embeddings related to t, and the residual vector r of the projection is recalled as follows:

$$e_{p'_x} = \mathcal{P}_{\mathscr{A}}(e_{x'_0}) \qquad r = e_{x'_0} - e_{p'_x}, \tag{1}$$

where $\mathcal{P}_{\mathscr{A}}$ is the projection operation on $\mathscr{A}$. Then, $e_{p'_x}$ is in the text embedding space which can be directly manipulated with e_t, and the manipulated embedding is finally projected back to the CLIP image embedding space by adding r as shown in Figure 2. The final text-manipulated 3D image embedding is:

$$e_{tx'} = \mathcal{M}(e_{p'_x}, e_t) + r, \tag{2}$$

where $\mathcal{M}$ is the manipulation operation guided by e_t, which is realized by weakening the other irrelevant basis attributes of e_t as in [54]

In summary, since $\mathscr{A}$ is constructed by basis attribute embeddings, the projection ensures that image manipulation operations are disentangled, which alleviates undesired changes. Therefore, $e_{tx'}$ should respect the attributes described by t. Toward this end, we train Γ to generate θ_{target} conditioned on $e_{tx'}$. The CLIP embedding of the 3D image rendered by θ_{target} should be close to $e_{tx'}$. The semantic correspondence constraint is formulated as follows:

$$\mathcal{L}^{corr} = 1 - \cos(E_{img}(\mathcal{R}(\Gamma(e_{tx'}))), e_{tx'}), \tag{3}$$

where $\cos(\cdot,\cdot)$ denotes the cosine distance between two image embeddings. The rendered target image $x'_{target} = \mathcal{R}(G_{3d}(e_{xt}))$ is precisely associated with $\theta_{target} = \Gamma(e_{xt})$. As a result, we can use text to manipulate the 3D image in a fine-grained manner.

3.3. Cross-modal Latent Mapping

StyleGAN enables high-quality 2D face generation, and its latent space $\mathcal{W}$ demonstrates meaningful and disentangled properties. A pretrained StyleGAN inversion Encoder e4e [45] E_{sty} extends to $\mathcal{W}+$ space for better generation quality. In our experiment, we use e4e to derive the latent code $w_0 = E_{sty}(x_0)$ of the source image x_0. Therefore, the goal is to find a latent transformation which can respect the target attributes and preserve the original attributes as well in the latent space.

However, owing to the modality gap, it is challenging to precisely infer the latent transformations for producing physically meaningful 3D rigging with text. To address this issue, we consider incorporating the above 3D difference information ($\Delta\theta = \theta_{target} - \theta_0$) into the latent mapping network (Φ) to transfer the manipulation in 3D space to latent space, where $\Delta\theta$ exactly represents the precise 3D rigging for controlling the generation.

Specifically, Φ is conditioned on both w_0 and $\Delta\theta$, which produces latent correction vector Δw. Compared to global latent code prediction, the correction learning strategy is input-aware and more faithful to the original image, and can preserve the original ID and attributes better. In addition, Δw and $\Delta\theta$ can be easily associated across two different modalities. Following [32], our Φ consists of 3 parallel groups which correspond to coarse-, medium-, and fine-level generation. The formulation is expressed as:

$$\Delta w = \Phi(w_0, \Delta\theta). \tag{4}$$

As a result, the target latent code w_{target} can be obtained by adding the original latent code to the correction vector, which will be decoded by the pretrained StyleGAN G_{sty} for generating the target image:

$$x_{target} = G_{sty}(w_0 + \Phi(w_0, \Delta\theta)). \tag{5}$$

Due to the unique design of Φ which is conditioned on 3D transformation to estimate Δw for 2D generation, we can achieve 3D-$\mathcal{W}+$ alignment via the proposed 2D-3D self-consistency constraints as follows.

3D consistency. Since x_{target} represents the target text-manipulated image, its 3D face representation should be the same as θ_{target}. Therefore, we can use $\mathcal{R}_{inv}$ to inverse render x_{target}, and the 3D consistency constraint is expressed as follows:

$$\mathcal{L}^{3d} = ||\theta_{target} - \mathcal{R}_{inv}(x_{target})||_2. \tag{6}$$

By minimizing $\mathcal{L}^{3d}$, Φ seeks suitable latent transformations such that the attributes of the target image are well represented in 3D space.

2D landmark consistency. Although 3D attributes *i.e.*, pose, expression, illumination are well represented with 3D priors, 2D details associated with mouth, eyes, eyebrows, *etc*. should be subtly processed. Instead of using pixel-wise $\mathcal{L}_2$ loss, we adopt a pretrained landmark detection model to produce the key parts of a face in 2D space, and propose a 2D landmark consistency constraint as follows:

$$\mathcal{L}^{2d} = ||\mathcal{R}(\theta_{target}) - F(x_{target})||_2, \tag{7}$$

where F is the pretrained landmark detection model [5], and the rendering process $\mathcal{R}$ can automatically produce the landmark with θ_{target}. By minimizing $\mathcal{L}^{2d}$, x_{target} can preserve 2D details to achieve high-quality image generation.

3.4. Model Optimization

By integrating the above 3D Editor and 3D-aware latent mapper, the overall optimization process can be expressed as follows:

$$\begin{aligned} &\min_{\Gamma} \mathcal{L}^{corr}, \\ &\min_{\Phi} \mathcal{L}^{3d} + \mathcal{L}^{2d}, \end{aligned} \tag{8}$$

We only train $\{\Gamma, \Phi\}$ and randomly specify the target text for manipulation. During inference, we can use text to manipulate face images in a fine-grained manner.

4. Experiments

4.1. Experitmenal Setup

Training data. We use Flickr-Face-HQ (FFHQ) [23] during training, which contains 70,000 1024×1024 face images crawled from Flickr. Images have no annotated attributes, and we randomly sample the target text attributes and pre-process them as in [32].

Implementation details. Our framework includes five pretrained models: StyleGAN [22] (G_{sty}) for generation based on the latent codes, e4e encoder (E_{sty}) for $\mathcal{W}+$ codes extraction, 3D Predictor ($\mathcal{R}_{inv}$) for inverse rendering, CLIP text and image encoders (E_{txt}, E_{img}) for semantic linguistic-visual embedding extraction. There are two trainable models Γ and Φ for 3D Editing and latent correction vector prediction, respectively. We use Adam [26] to optimize the trainable model with a learning rate of 0.5. We observe that the resulting performance is robust.

Baseline models. We compare our model with recent representative and state-of-the-art methods: StyleClip [32] and DeltaEdit [30] are text-driven image manipulation models which use CLIP and improved CLIP delta similarity to guide image editing with text semantics. DiscoFaceGAN [10], GANControl [42] and DiffusionRig [11] are 3D-aware face editing models which incorporate 3D priors into GAN and Diffusion model training, respectively. We use official implementations or source pretrained models of these methods during evaluations.

Evaluation metrics. We use Fréchet Inception Distance (FID) [17] for evaluation of the image quality, which is widely adopted in image generation tasks [48]. To assess the performance of identity preservation, the IDentity Distance (IDD) between original and manipulated images is computed with a well-trained face recognition network [8]. We also adopt the Target Attribute Recognition Rate (TARR) to measure the attribute correctness. Specifically, we use the off-the-shelf expression classifier [28], light spherical harmonics predictor [55] and face pose estimator [21] to compute similarity or distance scores for the expression, light and pose attributes, respectively.

4.2. Ablation Study

We conduct several ablation studies to validate the effectiveness of each proposed component. The qualitative and quantitative results are presented in Figure 3 and Figure 4, respectively.

(a) w/o $\mathscr{A}$. We first conduct an experiment to demonstrate the effectiveness of text-conditional 3D Editing via $\mathscr{A}$. Specifically, we remove the alignment operations in $\mathscr{A}$ for disentangled text manipulation, and directly use the target text embedding for supervision. As shown in Figure 3(a), the resulting variant cannot control the 3D attributes (*i.e.*, a large score of TARR:Pose in Figure 4) due to

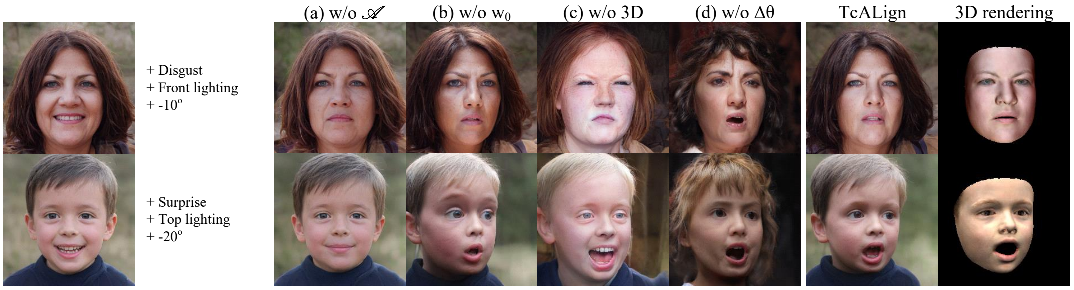

Figure 3. Ablation studies of different components in our framework. (a) removes $\mathscr{A}$ and directly uses the text embedding as the target for supervision. (b) removes w_0 in Φ; (c) removes 3D priors and uses the pretrained networks for 3D supervision; (d) removes correction vector learning and adopts global latent code prediction.

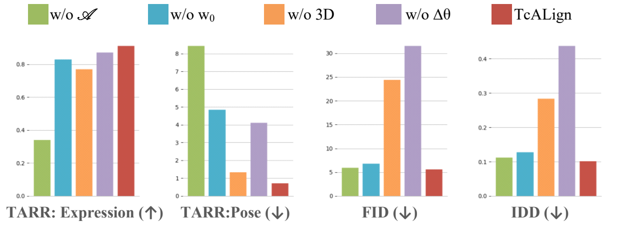

Figure 4. Comparison between TcALign and its variants in terms of different metrics.

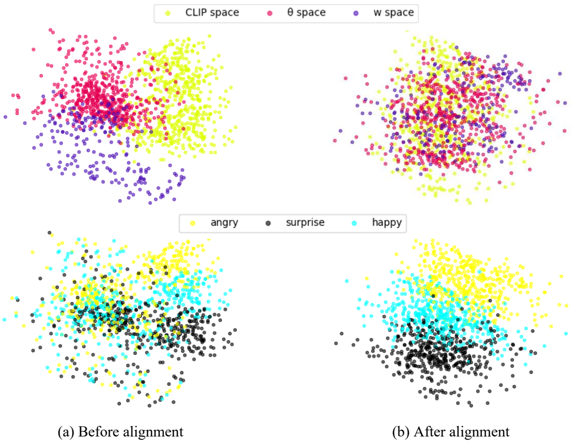

Figure 5. Visualization of CLIP image embeddings from different modalities (a) before alignment and (b) after alignment. Different colors denote different modalities (top) and attributes (bottom).

the discrepancy between different modalities. On the other hand, our model can better infer the transformation with the discovery of target-specific direction in $\mathscr{A}$ for precise manipulation in 3D space.

(b) w/o w_0. We remove w_0 in Φ to verify the importance of input-aware generation. As shown in Figure 3(b), there exist some inaccurate controls on the pose and illumination attributes and some non-target attributes are changed. Figure 4 also reports that the score of TARR:pose is largely degraded. Input-aware generation indicates deriving a specific transformation for each input, thus more precise latent correction transformation can be inferred by considering the original information within w_0.

(c) w/o 3D. We further remove the 3D priors in our framework, and use pretrained models to realize 3D control. In Figures 3 and 4, we can observe that although the TARR scores are not seriously degraded which indicates effective guidance with pretrained models, there exist obvious undesired controls and artifacts (large FID score), and the IDD score is also increased. These results confirm the effectiveness of 3D priors for better performance.

(d) w/o $\Delta\theta$. Instead of directly inferring the global latent code w_{target} with θ_{target}, our model adopts a correction vector learning scheme which infers Δw with $\Delta\theta$. This experiment shows that global prediction across two modalities is quite difficult, and the corresponding Δ transformations in the two spaces can be easily aligned. As shown in Figure 3(d), the variant results are significantly degraded, with obvious changes on undesired attributes and ID, which is also reported in Figure 4. This strategy enables w_{target} to be close to w_0, which avoids degraded performance. As a result, our results are more faithful to the original input and demonstrate accurate control.

4.3. Modality Alignment Analysis

In this section, we conduct an experiment to validate the capability of our method to align all involved modalities in the CLIP image embedding space. Specifically, there exist three representations that correspond to the same target manipulated image in our framework, which are text-manipulated image embedding $e_{tx'}$, target 3D face representation θ_{target}, and target latent code w_{target}. Therefore, $\mathcal{R}(\theta_{target})$ and $G_{sty}(w_{target})$ should be aligned in CLIP image embedding space together with $e_{tx'}$.

We randomly sample 100 face images, and manipulate them with 3 expressions (angry, surprise, happy) to obtain the text-manipulated image embedding $e_{tx'}$. Next, the corresponding CLIP image embedding of the generated target

Method	FID ↓	IDD ↓	TARR		
			Expression ↑	Illumination ↑	Pose ↓
DiffusionRig [11]	27.634	0.224	N/A	0.846	1.736
DiscoFaceGAN [10]	63.363	0.591	N/A	0.808	2.106
GANControl [42]	34.782	0.38	N/A	0.834	1.571
StyleCLIP [32]	7.595	0.124	0.728	N/A	N/A
DeltaEdit [30]	7.658	0.119	0.547	N/A	N/A
TcALign (Ours)	7.104	0.107	0.913	0.859	0.714

Table 1. Quantitative comparison of TcALign and competing methods on the image quality, ID preservation and attribute correctness.

images $e_{x_{target}}$ and the rendered target images $e_{\mathcal{R}(\theta_{target})}$ are produced. As shown in Figure 5, we use PCA to project all embeddings from three modalities into a 2D space. It can be observed that before alignment, there exists a large modality gap among different spaces, and embeddings with different expressions are randomly scattered. On the other hand, the three embedding spaces are well aligned in our model after training, and three attributes are clearly clustered. These observations suggest that TcALign can align these three modalities and achieve accurate disentangled image manipulation in both 3D and latent spaces.

4.4. Visualization Analysis

In this section, we present visualization results of our method and competing methods. Figure 6 shows the manipulated images with different expressions, illuminations and poses from our model and competing methods. It is clear that our results exhibit higher quality with the multiple attributes precisely controlled. Meanwhile, DiscofaceGAN without pretrained generator is likely to change the face ID, and GANControl cannot precisely control the expressions. Two representative text-driven methods StyleCLIP and DeltaEdit fail to manipulate 3D attributes, *i.e.*, pose and illumination. DiffusionRig [11] shows promising 3D-aware control. However, it has no text control for expressions and requires carefully personalized finetuning with around 20 images. In Figure 6, the corresponding generation quality is degraded with few finetuning samples. In contrast, TcALign achieves high-quality generation without further finetuning and precise control without changing the other attributes. The poses and illuminations are also consistent with the rendered 3D images.

4.5. Quantitative Comparison

We further perform the quantitative comparison between our method and competing methods. We randomly sample 3,000-ID real face images and manipulate each of them with 5 different attributes. All the metrics are computed under the same setting for fair comparison.

Table 1 shows the quantitative comparison results in terms of FID, IDD and TARR between TcALign and competing methods. First, TcALign can significantly outperform previous 3D-aware generation methods (*i.e.*, DiffusionRig, DiscoFaceGAN and GANControl) in terms of FID which are generally conditioned on the global 3D representation. In particular, DiscoFaceGAN delivers an unsatisfactory FID score of 63.363, while the FID of our model is significantly improved to 7.104 which implies high-quality generation. In addition, DiffusionRig and GANControl demonstrate improved performance on the manipulations of pose and illumination with the adoption of a pretrained generator. On the other hand, TcALign obtains much better performance in terms of TARR than both of them, which verifies its capability of performing fine-grained 3D control with the correction vector learning scheme.

Second, we compare our approach with the SOTA text-driven image manipulation methods (StyleCLIP and DeltaEdit). Both methods are based on StyleGAN which does not enable 3D control on poses and illuminations. StyleCLIP shows better performance on expression controls and DeltaEdit achieves better IDD score. TcAlign is the only model that can manipulate all these attributes simultaneously, and achieves the significant improvement on TARR:Expression score of 0.913 with the disentangled text manipulation.

In addition, the ID preservation analysis results are presented in Figure 7. We can observe that DiscoFaceGAN fails in preserving the original ID (higher IDD score) in most cases in the absence of the pretrained generator, while our model is the most powerful in preserving ID under different manipulation factors. This further verifies the effectiveness of our 3D controllable synthesis which maintains the original information as well.

4.6. Further Analysis

In this section, we further analyse our model by manipulating images with infrequent open-vocabulary expressions. The results are shown in Figure 8 where StyleCLIP often fails, and some infrequent expressions *i.e.* ‘claustrophobic’ are not understandable for DeltaEdit. In contrast, TcALign can accurately understand these open-vocabulary expressions via the disentangled text manipulation in 3D space.

We also present the generation results based on interpolation with respect to θ_{target}. As shown in Figure 9, DiffusionRig cannot precisely control the expressions and illuminations, and there exist generation artifacts. Meanwhile, the attributes of our results are smoothly transferred with high-quality generation, which demonstrates that TcAlign successfully achieves fine-grained control in a 3D setting.

5. Conclusion

Text-conditional face manipulation is an important research topic which allows more flexible controls of face attributes. Due to the complex variations of 2D images and the domain

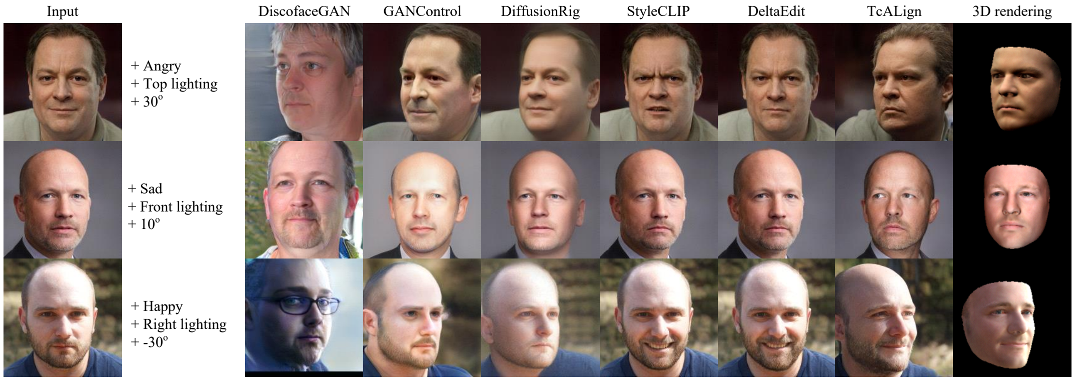

Figure 6. Multiple-attribute (including expression, lighting and pose) manipulation results of TcALign and competing methods.

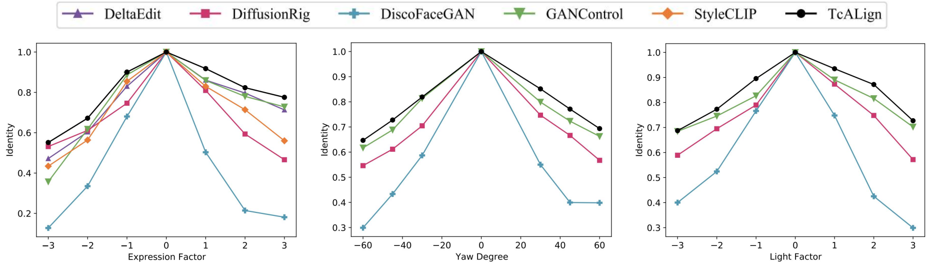

Figure 7. ID preservation analysis with different attribute variations. 0 indicates no manipulation with the original ID.

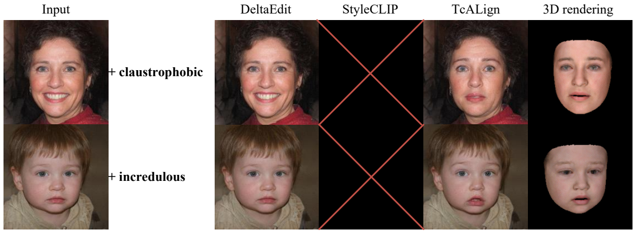

Figure 8. Comparison of our model and competing methods in manipulating infrequent open-vocabulary expressions.

gap between texts and images, it is very challenging to infer the disentangled manipulation direction to avoid incurring undesired artifacts. We present a new Text-conditional attribute alignment method for fine-grained face manipulation, which is a new task and not addressed by previous methods. Specifically, we use a 3D rendered image which can be precisely manipulated with a 3D face representation, and propose a 3D Editor to achieve text manipulation in 3D space. We demonstrate that by leveraging the 3D difference information as conditions, our cross-modal latent mapping network can generate disentangled and controllable latent transformations in the latent space of StyleGAN, which enables 3D control with text. Extensive experiments demonstrate the superiority of our model and its capability in performing precise text-driven face manipulation, which is important for a broad range of real-world applications.

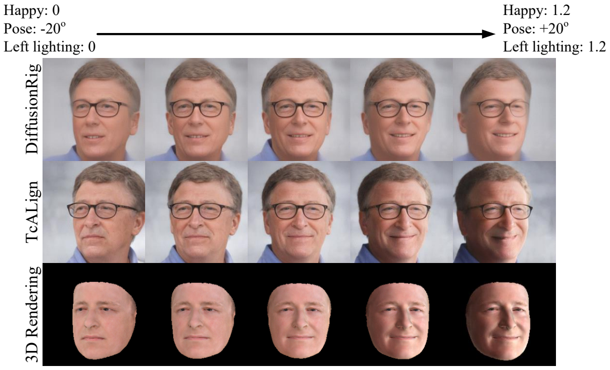

Figure 9. Multi-attribute manipulation of our model with interpolation on θ_{target}. Each row shares the same face, and the corresponding θ_{target} is linearly interpolated between 0 and 1.2.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Project No. 62072189), in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11206622), in part by the GuangDong Basic and Applied Basic Research Foundation (Project No. 2020A1515010484, 2022A1515011160), and in part by TCL Science and Technology Innovation Fund (Project No. 20231752).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, pages 8293–8302, 2020. 3

[2] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Trans. Graph.*, 40(3):21:1–21:21, 2021. 2, 3

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *ICCV*, pages 6691–6700, 2021. 3

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, pages 187–194. ACM, 1999. 2

[5] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *CCBR*, pages 428–438, 2018. 5

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, pages 8185–8194, 2020. 2

[7] Edo Collins, Raja Bala, Bob Price, and Sabine Süsstrunk. Editing in style: Uncovering the local semantics of gans. In *CVPR*, pages 5770–5779, 2020. 3

[8] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):5962–5979, 2022. 5

[9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPRW*, pages 285–295, 2019. 3

[10] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *CVPR*, pages 5153–5162, 2020. 2, 3, 5, 7

[11] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *CVPR*, pages 12736–12746, 2023. 2, 3, 5, 7

[12] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhöfer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models - past, present, and future. *ACM Trans. Graph.*, 39(5):157:1–157:38, 2020. 2

[13] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Trans. Graph.*, 40(4):88:1–88:13, 2021. 3

[14] Yudong Guo, Juyong Zhang, Jianfei Cai, Boyi Jiang, and Jianmin Zheng. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(6):1294–1307, 2019. 3

[15] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable GAN controls. In *NeurIPS*, 2020. 3

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

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017. 5

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[19] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 179–196, 2018. 2

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976, 2017. 2

[21] We Jie. Facial-expression-recognition.pytorch, 2018. 5

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8107–8116, 2020. 2, 5

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228, 2021. 2, 5

[24] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, pages 2416–2425, 2022. 2

[25] 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 Trans. Graph.*, 37(4):163, 2018. 3

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[27] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. Manigan: Text-guided image manipulation. In *CVPR*, pages 7877–7886, 2020. 2

[28] 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 *CVPR*, pages 3673–3682, 2019. 5

[29] Yahui Liu, Marco De Nadai, Deng Cai, Huayang Li, Xavier Alameda-Pineda, Nicu Sebe, and Bruno Lepri. Describe what to change: A text-guided unsupervised image-to-image translation approach. In *ACMMM*, pages 1357–1365, 2020. 2

[30] Yueming Lyu, Tianwei Lin, Fu Li, Dongliang He, Jing Dong, and Tieniu Tan. Deltaedit: Exploring text-free training for text-driven image manipulation. In *CVPR*, pages 6894–6903, 2023. 2, 3, 4, 5, 7

[31] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 2

[32] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, pages 2065–2074, 2021. 2, 3, 5, 7

[33] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *AVSS*, pages 296–301, 2009. 3

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2, 3

[35] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *SIGGRAPH*, pages 497–500. ACM, 2001. 3

[36] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *SIGGRAPH*, pages 117–128. ACM, 2001. 3

[37] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, pages 1060–1069, 2016. 2

[38] 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 *CVPR*, pages 2287–2296, 2021. 3

[39] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, pages 9240–9249, 2020. 2, 3

[40] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44 (4):2004–2018, 2022. 2, 3

[41] Yichun Shi, Divyansh Aggarwal, and Anil K. Jain. Lifting 2d stylegan for 3d-aware face generation. In *CVPR*, pages 6258–6266, 2021. 3

[42] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gérard G. Medioni. Gan-control: Explicitly controllable gans. In *ICCV*, pages 14063–14073, 2021. 2, 5, 7

[43] Quanpeng Song, Jiaxin Li, Si Wu, and Hau-San Wong. A graph-based discriminator architecture for multi-attribute facial image editing. *IEEE Transactions on Multimedia*, pages 1–11, 2023. 2

[44] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *CVPR*, pages 6141–6150, 2020. 3

[45] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Trans. Graph.*, 40(4):133:1–133:14, 2021. 4

[46] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity GAN inversion for image attribute editing. In *CVPR*, pages 11369–11378, 2022. 3

[47] Xiwen Wei, Zhen Xu, Cheng Liu, Si Wu, Zhiwen Yu, and Hau-San Wong. Text-guided unsupervised latent transformation for multi-attribute image manipulation. In *CVPR*, pages 19285–19294, 2023. 3

[48] Yue Wu, Sicheng Xu, Jianfeng Xiang, Fangyun Wei, Qifeng Chen, Jiaolong Yang, and Xin Tong. Aniportraitgan: Animatable 3d portrait generation from 2d image collections. In *SIGGRAPH*, 2023. 5

[49] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *CVPR*, pages 12863–12872, 2021. 2, 3

[50] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*, pages 2256–2265, 2021. 3

[51] Zipeng Xu, Tianwei Lin, Hao Tang, Fu Li, Dongliang He, Nicu Sebe, Radu Timofte, Luc Van Gool, and Errui Ding. Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model. In *CVPR*, pages 18208–18217, 2022. 2

[52] Han Zhang, Tao Xu, and Hongsheng Li. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, pages 5908–5916, 2017. 2

[53] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, pages 833–842, 2021. 2

[54] Chenliang Zhou, Fangcheng Zhong, and Cengiz Öztireli. CLIP-PAE: projection-augmentation embedding to extract relevant features for a disentangled, interpretable and controllable text-guided face manipulation. In *ACM SIGGRAPH*, pages 57:1–57:9. ACM, 2023. 4

[55] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David Jacobs. Deep single-image portrait relighting. In *ICCV*, pages 7193–7201, 2019. 5

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2242–2251, 2017. 2