3D-FM GAN: Towards 3D-Controllable Face Manipulation

Yuchen Liu[1]([✉]), Zhixin Shu[2], Yijun Li[2], Zhe Lin[2], Richard Zhang[2], and S.Y. Kung[1]

[1] Princeton University, Princeton, USA
{yl16,kung}@princeton.edu
[2] Adobe Research, San Jose, USA
{zshu,yijli,zlin,rizhang}@adobe.com

Abstract. 3D-controllable portrait synthesis has significantly advanced, thanks to breakthroughs in generative adversarial networks (GANs). However, it is still challenging to manipulate existing face images with precise 3D control. While concatenating GAN inversion and a 3D-aware, noise-to-image GAN is a straight-forward solution, it is inefficient and may lead to noticeable drop in editing quality. To fill this gap, we propose 3D-FM GAN, a novel conditional GAN framework designed specifically for **3D**-controllable **F**ace **M**anipulation, and does not require any tuning after the end-to-end learning phase. By carefully encoding both the input face image and a physically-based rendering of 3D edits into a Style-GAN's latent spaces, our image generator provides high-quality, identity-preserved, 3D-controllable face manipulation. To effectively learn such novel framework, we develop two essential training strategies and a novel multiplicative co-modulation architecture that improves significantly upon naive schemes. With extensive evaluations, we show that our method outperforms the prior arts on various tasks, with better editability, stronger identity preservation, and higher photo-realism. In addition, we demonstrate a better generalizability of our design on large pose editing and out-of-domain images.

1 Introduction

Face manipulation with precise control has long attracted attention from computer vision and computer graphics community for its application in face recognition, photo editing, visual effects, and AR/VR applications, etc. In the past, researchers developed 3D morphable face models (3DMMs) [4,26,29], which provide an explainable and disentangled parameter space to control face attributes of identity, pose, expression, and illumination. However, it is still challenging to render photo-realistic face manipulations with 3DMMs.

Y. Liu—Work done during an internship at Adobe Research.

Photo   Pose   Expression   Illumination

Fig. 1. With explicit 3D controls of pose, expression, and illumination, presented as identity-preserved rendered faces (top row), 3D-FM GAN provides controllable and disentangled face manipulations on real world images (bottom row) with strong identity preservation and high photo-realism.

In recent years, generative adversarial networks (GANs) [15] have demonstrated promising results in photo-realistic face synthesis [23,24] by mapping random noise to image domain. While latent space exploration has been attempted [1,17,39], it requires a lot of human labor to discover meaningful directions, and the editings could still be entangled. As such, the variants of conditional GANs are widely studied for identity-preserved face manipulations [3,8,47,56]. Nonetheless, they either only allow control on a single facial attribute or require reference images/human annotations for face editing.

More recently, several works introduced 3D priors into GANs [10,14,40,44] for controllable synthesis. However, most of them are learned for noise-to-image random face generation, which does not naturally fit with the image manipulation task. Hence, they require time-consuming optimization in the test time, and the inverted latent codes may not lie in the manifold for high-quality editing [45,50]. Moreover, photo-realistic synthesis remains challenging [25,40].

To this end, we propose **3D-FM GAN**, a novel framework particularly designed for high-quality **3D**-controllable **F**ace **M**anipulation. Specifically, we perform a learning process to solve the image-to-image translation/editing problem with a conditional StyleGAN [24]. Different from prior 3D GANs trained for random sampling, we train our model exactly for existing face manipulation and do not require optimization/manual tuning after the learning phase. As shown in Fig. 1, with a single input face image, our framework manages to produce photo-realistic disentangled editing on attributes of head pose, facial expression, and scene illumination, while faithfully preserving the face identity.

Our framework leverages face reconstruction networks and a physically-based renderer, where the former estimate the input 3D coefficients and the latter embeds the desired manipulations, e.g., pose rotation, into an identity-preserved rendered face. A StyleGAN [24] conditional generator then takes in both the original image and the manipulated face rendering to synthesize the edited face. The consistent identity information provided by the input and the rendered edit signals spontaneously creates a strong synergy for identity preservation in manipulation. Moreover, we develop two essential training strategies, reconstruction and disentangled training, to help our model gain abilities of identity preservation and 3D

editability. As we find an interesting trade-off between identity and editability in the network structure and the simple encoding strategy is sub-optimal, we propose a novel multiplicative co-modulation architecture for our framework. This structure stems from a comprehensive study to understand how to encode different information in the generator's latent spaces, where it achieves the best performance. We conduct extensive qualitative and quantitative evaluations on our model and demonstrate good disentangled editing ability, strong identity preservation, and high photo-realism, outperforming the prior arts in various tasks. More interestingly, our model can manipulate artistic faces which are out of our training domain, indicating its strong generalization ability.

Our contributions can be summarized as follows.

(1) We propose 3D-FM GAN, a novel conditional GAN framework that is specifically designed for precise, explicit, high-quality, 3D-controllable face manipulation. Unlike prior works, our training objective is strongly consistent with the task of existing face editing, and our model does not require any optimization/manual tuning after the end-to-end learning process.
(2) We develop two essential training strategies, reconstruction and disentangled training to effectively learn our model. We also conduct a comprehensive study of StyleGAN's latent spaces for structural design, leading to a novel multiplicative co-modulation architecture with strong identity-editability trade-off.
(3) Extensive quantitative and qualitative evaluations demonstrate the advantage of our method over prior arts. Moreover, our model also shows a strong generalizability to edit artistic faces, which are out of the training domain.

2 Related Works

3D Face Modelling. 3D morphable models (3DMMs) [4,5,26,28,29,46] have long been used for face modelling. In 3DMMs, human faces are normally parametrized by texture, shape, expression, skin reflectance and scene illumination in a disentangled manner to enable 3D-controllable face synthesis. However, 3DMMs require expensive data of 3D [29] or even 4D [26] scans of human heads to build, and the rendered images often lack photo-realism due to the low-dimensional linear representation as well as the absence of modelling in hair, mouth cavity, and fine details like wrinkles. With 3DMMs, many methods attempt to estimate the 3D parameters of 2D images [11–13,33–37,42,49]. This is normally achieved by optimization or a neural networks to extract 3D parameters from a face image and/or landmarks [11,13,37]. In our work, we use 3DMMs for 3D face representation and adopt face reconstruction networks to provide the basis of 3D editing signals from 2D images. As such, our model can be trained solely with 2D images to gain 3D controllability.

GAN. Recently, unconditional GANs show promising results in synthesizing photo-realistic faces [22–24]. While latent space exploration [17,39,41] has proved to be effective, it requires extensive human labors to obtain meaningful control for

generation. As such, a rich set of literatures propose to use conditional GANs [3, 8, 21, 30, 31, 38, 47, 48, 51, 56] for controllable identity-preserved image synthesis by disentangling identity and non-identity factors.

Noticeably, DR-GAN [47] and TP-GAN [21] disentangles identity and pose to allow frontal view synthesis, while Zhou et al. [56] extracts spherical harmonic lighting from source image for portrait relighting. However, these works can only manipulate one attribute of the faces, whilst we are able to conduct disentangled editing on pose, lighting, and expression under a unified framework. We even outperform some of them on the tasks that they are solely trained on.

To transfer multiple attributes, Bao et al. [3] and Xiao et al. [51] extract identity from one image and facial attributes from another one for reference-based generation. StarGAN [8] leverages multiple labeled datasets to learn attributes translation. In contrast, our model provide manipulations with just a single input, and it does not require any labeled information/datasets for training.

3D Controllable GAN. In line with our work, several prior methods [6, 10, 14, 25, 27, 40, 43, 44] introduce 3D priors into GANs to achieve 3D controllability over face attributes of expression, pose, and illumination.

Deng et al. [10] enforce the input space of its GAN to bear the same disentanglement as the parameter space of a 3DMM to achieve controllable face generation. GIF [14] conditions the space of StyleGAN's layer noise on render images from FLAME [26] to control pose, expression, and lighting. Tewari et al. [44] leverage a pretrained StyleGAN and learn a RigNet to manipulate latent vectors with respect to the target editing semantics. However, these approaches are all trained for random face generation, not for existing face manipulation. Although GAN inversion can well project existing images in their latent spaces for good reconstruction, these latent codes may not fall on the manifold with good editability, leading to noticeable quality drop after manipulation. On the contrary, our model is trained exactly for the task of real face editing, which demonstrates a clear improvement in manipulation quality upon these works.

While CONFIG [25] does not need GAN inversion for real image manipulation, its parametric editing space doesn't inherit identity information from the input images, resulting in a clear identity loss. Moreover, our novel generator architecture also provides us with a larger range of editability and higher photorealism upon them. While PIE [43] proposes a specialized GAN inversion process to be later combined with StyleRig for real image manipulation, we again find our approach provides better quality and higher efficiency. Compared to a more recent VariTex [6] approach which can not synthesize background and rigid body like glasses, our method produces much more realistic outputs.

3   Methodology

3.1   Overview and Notations

In Fig. 2, we show the workflow of 3D-FM GAN which consists of: the generator **G**, the face reconstruction network **FR**, and the renderer **Rd**. Given an input face image $P \in \mathbb{R}^{H \times W \times 3}$, it first estimates the lighting and 3DMM parameters
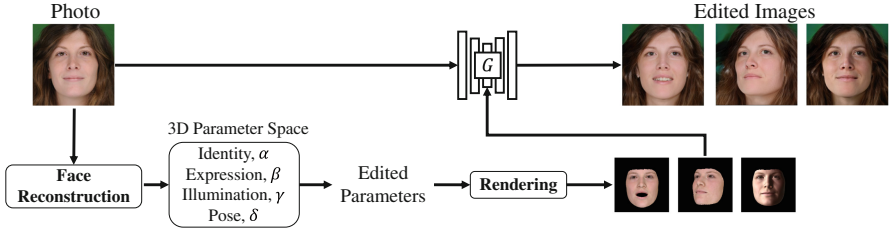
Fig. 2. Workflow of 3D-FM GAN. Given a photo input, our framework first extracts its 3D parameter by face reconstruction and then renders identity-preserved 3D-manipulated edit faces. The input photo and the edit signals are later jointly encoded into a generator to synthesize various photo editings.

of the face $p = \mathbf{FR}(P)$, $p = (\alpha, \beta, \gamma, \delta) \in \mathbb{R}^{254}$. Naturally, p has disentangled controllable components for identity $\alpha \in \mathbb{R}^{160}$, expression $\beta \in \mathbb{R}^{64}$, lighting $\gamma \in \mathbb{R}^{27}$, and pose $\delta \in \mathbb{R}^3$ [29]. The disentangled editing is then achieved in this parameter space, where we keep the identity factor α unchanged but adjust β, γ, δ to the desired semantics for the expression, lighting, and pose, which returns a manipulated parameter $\hat{p}$ to render an image $\hat{R} = \mathbf{Rd}(\hat{p})$ [4]. Finally, the manipulated photo is generated by feeding P and $\hat{R}$ through the generator as $\hat{P} = \mathbf{G}(P, \hat{R})$. In this way, the synthesized output $\hat{P}$ will preserve the identity from P, while its expression, illumination, and pose, follow the control from $\hat{p}$.

3.2 Dataset

The training data are in the form of photo and render image pairs (P, R), where P and R share the same attributes of identity, expression, illumination, and pose. We construct our dataset with both the FFHQ data and synthetic data. We show examples of the data pairs in Fig. 3.
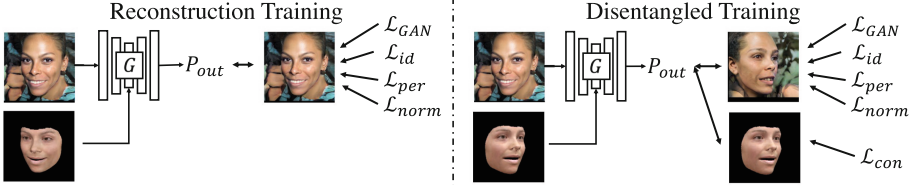

Fig. 3. Examples of photo and render image pairs (P, R). **Left:** FFHQ data. Each identity just has one corresponding image. **Right:** Synthetic data from [10]. We generate multiple images for an identity with varied expression, pose, and illumination.

FFHQ. FFHQ [23] is a human face photo dataset, where most identities only have one corresponding image. For each of the training image P, we extract its render counterpart by $R = \mathbf{Rd}(\mathbf{FR}(P))$ to form the (P, R) pair.

Synthetic Dataset. We also require a dataset where each identity has multiple images with various attributes of expression, pose, and illumination. Such a dataset is crucial for model to perform learning for editing. While this kind of high-quality dataset is not publicly available, we leverage DiscoFaceGAN [10], $\mathbf{G_d}$, to synthesize one as follows.

Fig. 4. Proposed model learning strategies of reconstruction training (**Left**) and disentangled training (**Right**).

Given a parameter p of our 3D parameter space and a noise vector $n \in \mathbb{R}^{32}$, $\mathbf{G_d}$ synthesizes a photo image $P = \mathbf{G_d}(p, n)$ that resembles the identity, expression, illumination, and pose of its render counterpart $R = \mathbf{Rd}(p)$. We can thus generate multiple images of the same identity with other attributes varied following the steps below: (1) randomly sample a 3D parameter $p_1 = (\alpha_1, \beta_1, \gamma_1, \delta_1)$ and a noise n; (2) keep α_1 unchanged and re-sample M - 1 tuples of (β, γ, δ) such that we have $p_2 = (\alpha_1, \beta_2, \gamma_2, \delta_2)$, ..., $p_M = (\alpha_1, \beta_M, \gamma_M, \delta_M)$; (3) Use $\mathbf{G_d}$ and $\mathbf{Rd}$ to generate photo-render pairs of (P_1, R_1), ... (P_M, R_M), where $P_i = \mathbf{G_d}(p_i, n)$ and $R_i = \mathbf{Rd}(p_i)$. Such process is iterated for N identities to form a dataset of $N \times M$ pairs. Examples of such image pairs are in Fig. 3 (**Right**).

3.3   Training Strategy

While **FR** and **Rd** do not require further tuning, we design two strategies, reconstruction and disentangled training (Fig. 4) to train **G**. We find the former helps for identity preservation while the latter ensures editability (Fig. 17). Formally, we denote the input pair (P_{in}, R_{in}) and its output $P_{out} = \mathbf{G}(P_{in}, R_{in})$.

Reconstruction Training. We first equip **G** with the ability to reconstruct P_{in} from (P_{in}, R_{in}). In this case, we want P_{out} to be as similar as P_{in}, and set the target output $P_{tg} = P_{in}$. We first define a face identity loss with a face recognition network $\mathbf{N_f}$ [9] :

$$\mathcal{L}_{id} = ||\mathbf{N_f}(P_{out}) - \mathbf{N_f}(P_{tg})||_2^2 \tag{1}$$

We also enforce P_{out} and P_{tg} to have similar low-level features and high-level perception by imposing an $\ell 1$ loss and a perceptual loss based on LPIPS [53]:
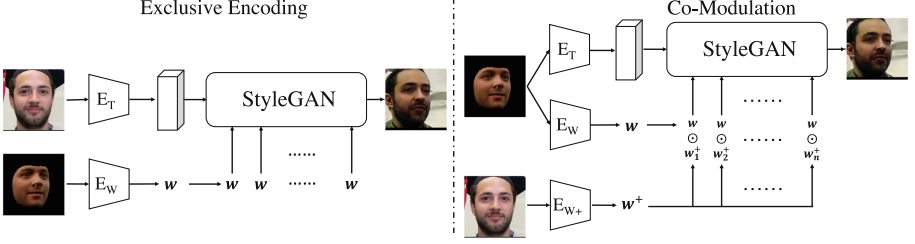
$$\mathcal{L}_{norm} = ||P_{out} - P_{tg}||_1 \tag{2}$$

$$\mathcal{L}_{per} = LPIPS(P_{out}, P_{tg}) \tag{3}$$

Finally, we adopt the GAN loss, $\mathcal{L}_{GAN}$ such that the generated images P_{out} shall match the distribution of P_{tg}. In this way, our loss is constructed as:

$$\mathcal{L}_{rec} = \mathcal{L}_{GAN} + \lambda_1 \mathcal{L}_{id} + \lambda_2 \mathcal{L}_{norm} + \lambda_3 \mathcal{L}_{per} \tag{4}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the weights for different losses. We use both synthetic and FFHQ datasets for this procedure.

Fig. 5. Our generator **G** can take forms of both exclusive modulation (**Left**) and co-modulation (**Right**) architectures.

Disentangled Training. To achieve our goal, only teaching the model how to "reconstruct" is not sufficient. Thus, we propose a disentangled training strategy to enable editing, which can only be achieved by the synthetic dataset as it has multiple images of the same identity with varying attributes.

Specifically, we first sample two pairs, (P_{in}^1, R_{in}^1) and (P_{in}^2, R_{in}^2), from the same identity. Then, given P_{in}^1 and R_{in}^2, we want our model to produce P_{in}^2, which shares the same edit signal in R_{in}^2 while the same identity of P_{in}^1. In this case, we set $P_{out}^1 = \mathbf{G}(P_{in}^1, R_{in}^2)$ and $P_{tg}^1 = P_{in}^2$, and impose the prior defined loss of $\mathcal{L}_{GAN}, \mathcal{L}_{id}, \mathcal{L}_{norm}$, and $\mathcal{L}_{per}$ between P_{out}^1 and P_{tg}^1. Different from reconstruction, we also inject a content loss to better capture the target editing signals, where we set $R_{tg}^1 = R_{in}^2$ and define the loss as:

$$\mathcal{L}_{con} = ||M \odot (P_{out}^1 - R_{tg}^1)||_2^2 \tag{5}$$

M is the face region that R_{tg}^1 has non-zero pixels and $\odot$ is the element-wise multiplication. To sum up, the loss of our disentangled training is:

$$\mathcal{L}_{dis} = \mathcal{L}_{GAN} + \lambda_1 \mathcal{L}_{id} + \lambda_2 \mathcal{L}_{norm} + \lambda_3 \mathcal{L}_{per} + \lambda_4 \mathcal{L}_{con} \tag{6}$$

with the same weights of $\lambda_1, \lambda_2, \lambda_3$ as reconstruction training. To use the loaded data more efficiently, we repeat the same procedure for $P_{out}^2 = \mathbf{G}(P_{in}^2, R_{in}^1)$.

Learning Schedule. In practice, we alternate between reconstruction and disentangled training: for every S iterations, we do 1 step of disentangled training and S - 1 steps of reconstruction. Moreover, as reconstruction can be performed by both synthetic and FFHQ datasets, we carry out our learning in two phases. In phase-1, we takes synthetic data for both training strategies. In phase-2, we switch to FFHQ for reconstruction while still uses synthetic data for disentangled training. Figure 17 shows the advantages of this 2-phase learning.

3.4 Architecture

Our conditional generator **G** is composed of a set of encoders **E** and a Style-GAN [24] generator $\mathbf{G_s}$. We utilize three latent spaces of $\mathbf{G_s}$ for information encoding, namely, the input tensor space $\mathcal{T} \in \mathbb{R}^{512 \times 4 \times 4}$, the modulation space

$\mathcal{W} \in \mathbb{R}^{512}$, and the extended modulation space $\mathcal{W}^+ \in \mathbb{R}^{512 \times L}$ (L: number of layers in $\mathbf{G_s}$). We denote the encoder to each of these spaces as $\mathbf{E_T}$, $\mathbf{E_W}$, and $\mathbf{E_{W^+}}$ and we conduct the study of what "information" (photo P or render R) to be encoded into which "space" ($\mathcal{T}$, $\mathcal{W}$, and $\mathcal{W}^+$), where we experiment both exclusive modulation and co-modulation architectures.

Exclusive Modulation. Naively, we can exclusively encode P or R into the modulation space ($\mathcal{W}/\mathcal{W}^+$). While one is encoded into the modulation space, the other one can only be encoded into $\mathcal{T}$. Figure 5 (**Left**) shows an example of such architectures, where R is encoded into $\mathcal{W}$ and P is encoded into $\mathcal{T}$. This structure is denoted as Render-$\mathcal{W}$. Whether R or P is used for modulation and whether the space is $\mathcal{W}$ or $\mathcal{W}^+$ provides us with 4 variants of exclusive modulation in total, and we investigate all of them.

Co-Modulation. We further investigate to encode both P and R into $\mathcal{W}$ and $\mathcal{W}^+$ and combine their embeddings for final modulation. A representative of such a architecture is shown in Fig. 5 (**Right**), where R is encoded into $\mathcal{W}$, and P is encoded into $\mathcal{W}^+$. In particular, the modulation signal for layer l is obtained by $\mathcal{W}_l^+ \odot \mathcal{W}$ where $\mathcal{W}_l^+ \in \mathbb{R}^{512}$ is the l-th column of $\mathcal{W}^+$ and $\odot$ is the element-wise multiplication. Unlike prior works that use concatenation or a recent approach of tensor transform plus concatenation scheme [55] to combine $\mathcal{W}_l^+$ and $\mathcal{W}$, we find our multiplicative co-modulation owns the best effectiveness. Moreover, in Fig. 5, we also encode R into $\mathcal{T}$ to further improve the identity-editability trade-off.

4   Experiment

4.1   Experimental Setup

We adopt the face reconstruction network **FR** [11], the 3DMM [29], and the renderer **Rd** [4]. Our conditional generator **G** consists of the StyleGAN [24] generator $\mathbf{G_s}$ and ResNet [18] encoders **E**. Specifically, $\mathbf{E_T}$ and $\mathbf{E_W}$ are based on ResNet-18 structure, where $\mathbf{E_T}$ outputs the feature prior to final pooling and $\mathbf{E_W}$ outputs the layer after that. We use a 18-layer PSP encoder [32] as $\mathbf{E_{W^+}}$. The discriminator architecture is the same as [24]. The synthetic data are generated by DiscoFaceGAN [10], where we set N and M to 10000 and 7. We use the first 65k FFHQ images (sorted by file names) for training and the rest 5k images as held-out testing set. All images (render and photo input, model output) are of 256 px resolution. We set S to 2 and use a batch size of 16 for both reconstruction and disentangled training. We set λ_1, λ_2, λ_3, and λ_4 to be 3, 3, 30, and 20. The model is learned in 2-phase, where phase-1 takes 140k iterations followed by 280k updates of phase-2.

4.2   Evaluation Metrics

We develop several quantitative metrics with the held-out 5k FFHQ images (denoted as $\mathbb{P}$) to evaluate identity preservation (**Identity**), editing controllability (**Face Content Similarity and Landmark Similarity**), and photo-realism (**FID**) of our model **G** for image manipulation.

Manipulated Images. For each $P \in \mathbb{P}$, we first get $p = \mathbf{FR}(P) = (\alpha, \beta, \gamma, \delta)$ and then re-sample (β, γ, δ) to form edited control parameters $\hat{p}$. The editing signals and the manipulated images are thus $\hat{R} = \mathbf{Rd}(\hat{p})$ and $\hat{P} = \mathbf{G}(P, \hat{R})$. We generate 4 $\hat{P}$ for each P.

Identity. For each $(P, \hat{P})$ pair, we measure the identity preservation by computing the cosine similarity of $< \mathbf{N_f}(P), \mathbf{N_f}(\hat{P}) >$.

Landmark Similarity. For each $(\hat{P}, \hat{R})$ pair, we use a landmark detection network $\mathbf{N_l}$ [7], to extracts both of their 68 2D landmarks. The similarity metric is defined as $||\mathbf{N_l}(\hat{P}) - \mathbf{N_l}(\hat{R})||_2^2$.

Face Content Similarity. For each $(\hat{P}, \hat{R})$ pair, we follow Eq. 5 to measure the face content similarity.

FID. We denote all edited images as $\hat{\mathbb{P}}$ and measure FID [19] between $\mathbb{P}$ and $\hat{\mathbb{P}}$ to evaluate $\mathbf{G}$'s photo-realism.

4.3   Architectures Evaluation

Exclusive Modulation. We first evaluate exclusive modulation architectures in Table 1, where we observe a trade-off between identity preservation and editability. For example, Photo-$\mathcal{W}^+$ shows the best identity preservation (**Id**), while its editability of landmark (**LM**) and face content similarity (**FC**) is the worst. On the other hand, Photo-$\mathcal{W}$ and Render-$\mathcal{W}^+$ owns strong **LM** and **FC**, yet their **Id** are much poorer and they even have issues on photo-realism. Moreover, we see that Render-$\mathcal{W}$ provides us with a decent editability, while improves a lot on **Id** compared to Render-$\mathcal{W}^+$ and Photo-$\mathcal{W}$. For good identity preservation, we design a co-modulation architecture based on Render-$\mathcal{W}$ and Photo-$\mathcal{W}^+$.

Co-Modulation. Based on the study above, we find that encoding P into $\mathcal{W}^+$ produces the best identity preservation, while encoding R into $\mathcal{W}$ provides good editability. Thus, we investigate three 2-encoder co-modulation architectures where $\mathbf{E_W}$ encodes R and $\mathbf{E_{W^+}}$ encodes P. Combining $\mathcal{W}^+$ and $\mathcal{W}$ are achieved via multiplication, concatenation, and a variant of concatenation named tensor transform in [55]. From Tab. 1, we find the multiplicative co-modulation achieves
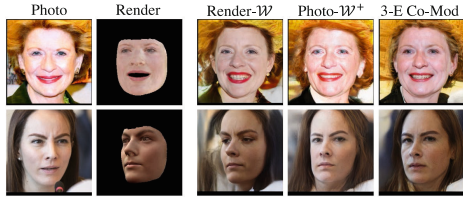
Table 1. Quantitative measurement of identity preservation (**Id**), editing control (**LM** & **FC**), and photo-realism (**FID**) for different architectures. $\uparrow$ means the higher the better, and vice versa for $\downarrow$.

Model		Metrics			
Mod. scheme	Type	Id.$^\uparrow$	LM.$^\downarrow$	FC.$^\downarrow$	FID$^\downarrow$
Exclusive modulation	Render-$\mathcal{W}$	0.57	17.8	0.021	14.5
	Render-$\mathcal{W}^+$	0.46	16.2	0.019	25.8
	Photo-$\mathcal{W}$	0.50	15.6	0.018	13.7
	Photo-$\mathcal{W}^+$	0.66	27.3	0.033	12.3
2-Encoder Co-Mod	Concat.	0.64	59.3	0.031	17.8
	Tensor	0.62	24.9	0.028	18.6
	Multi.	0.66	22.2	0.025	12.4
3-E Co-Mod	Multi.	0.66	17.2	0.020	12.2

the best results from all perspectives. This could be accounted by the fact that modulation itself is a multiplicative operation and thus merging signals together multiplicatively would provide the best synergy. We further propose a

3-encoder multiplicative co-modulation architecture (bottom of Fig. 5) to boost the editability, which achieves the best trade-off from our observation.

Visualization. We show a visual comparison among Render-$\mathcal{W}$ (**Col. 3**), Photo-$\mathcal{W}^+$ (**Col. 4**), and the 3-encoder co-modulation scheme (**Col. 5**) with the same set of inputs (**Col. 1 & 2**) in Fig. 6. In the first row, we find that Render-$\mathcal{W}$ has a clear identity loss, while Photo-$\mathcal{W}^+$ can hardly manipulate the light intensity, showing inferior editability. Moreover, Render-$\mathcal{W}$ and Photo-$\mathcal{W}^+$ both generate artifacts in the second row. On the

Fig. 6. Visual comparison among architectures. The co-modulation scheme takes advantages from both: good editability from Render-$\mathcal{W}$ and strong identity preservation from Photo-$\mathcal{W}^+$.

contrary, the co-modulation scheme improves the identity-editability by combining the merits from both schemes: good editability from Render-$\mathcal{W}$ and strong identity preservation from Photo-$\mathcal{W}^+$.

4.4   Controllable Image Synthesis

We apply our 3-encoder co-modulation architecture to several image manipulation tasks where it all shows good editing controllability, strong identity preservation, and high photo-realism. More samples are in Supplementary.
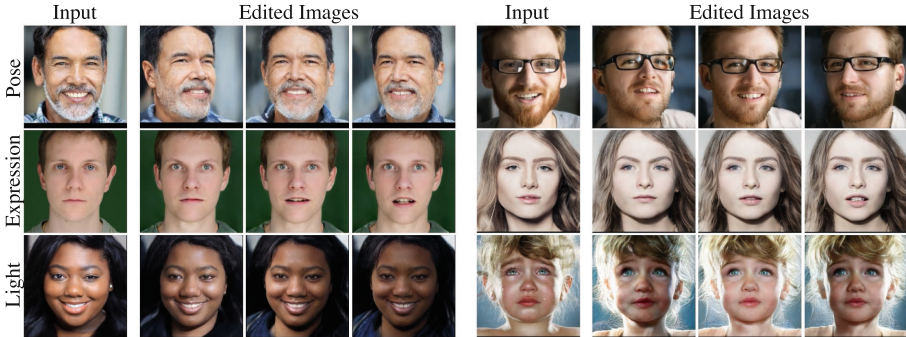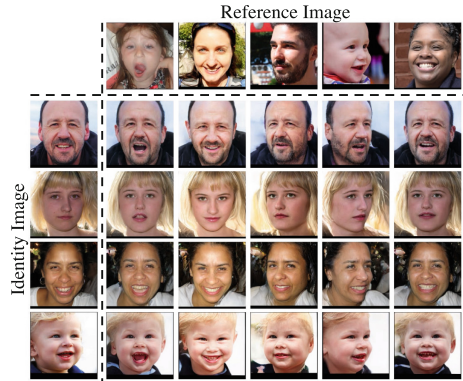

Fig. 7. Our model provides a variety of disentangled controls for pose (**Row 1**), expression (**Row 2**), and illumination (**Row 3**). It shows strong preservation across diverse identities and for facial details like glasses.

Disentangled Editing. Figure 1 and Fig. 7 show the results of single factor editing, where we only change one factor of pose, expression, and illumination at a time. Our model provides highly disentangled editing for the edited factors, while all others remain the same. Moreover, it shows strong preservation for the identity across people with diverse ages, genders, etc., and subtle facial details like the glasses and the teeth.

Fig. 8. Reference based face generation. The facial attributes of pose, expression, and illumination are extracted from the reference images to manipulate the identity images.
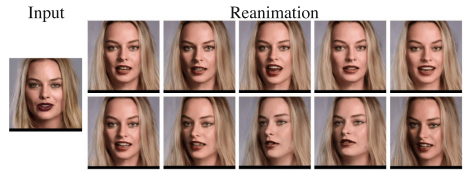
Reference-Based Synthesis. Our model can also perform image manipulation based on reference images shown in Fig. 8. With the pose, expression, and illumination extracted from the reference images, we re-synthesize our identity images to bear these editing facial attributes while the identities are still well preserved.

Fig. 9. Image reanimation. Our model again well preserves the identity and some subtle facial attributes like the dark lipstick.

Face Reanimation. Our model can also be applied for face reanimation, as shown in Fig. 9. With a single input photo image, we provide a series of editing render signals to make it animated, where the identity of the person is well preserved across frames. Moreover, our model again well preserves facial details like the dark lipstick.

Artistic Images Manipulation. We further perform manipulation on artistic faces [52] in Fig. 10. Surprisingly, although our model is only trained on photography faces, it can still provide controllable and identity-preserved editing on artistic images that are out of the training domain. This well indicates the strong generalizability of our model.
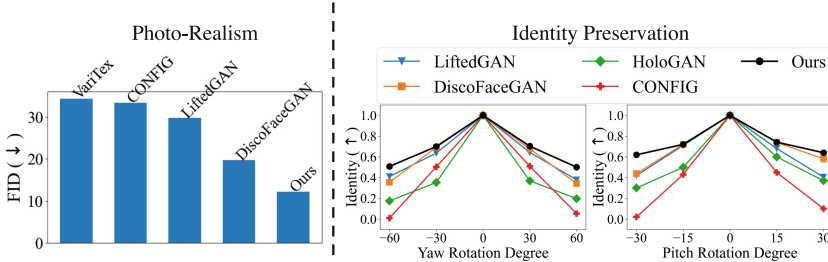
5   Comparison to State of the Arts

We compare with prior 3D-controllable GANs [6, 10, 25, 27, 40, 43, 44], and show more results in Supplementary.

Fig. 10. Although our model is solely trained on photo faces, it demonstrates a strong generalizability to manipulate artistic faces.

5.1   Quantitative Comparison

From Fig. 11 (**Left**), we clearly find that our model produces the most photo-realistic images with the lowest FID. We also follow the similar strategy in [40] to measure identity preservation, where we use all frontal images from the held-out FFHQ set and perform pose editing at different angles to compute the identity cosine similarity between the edited faces and the original ones. While prior methods evaluates the preservation between their generated images that naturally fit with their latent manifolds, we are assessing the identity preservation with real world images, which represents a more challenging task. Surprisingly, as shown in Fig. 11 (**Right**), our model still outperforms prior arts in all rotation angles on a harder task, and it can well preserve identity even at large angles.

Fig. 11. Quantitative comparison with prior arts. Our method achieves the best photo realism (**Left**) and better identity preservation (**Right**) at different rotation angles.

5.2   Visual Comparison

DiscoFaceGAN. We first compare 3D-FM GAN with the direct combination of GAN inversion [2] + noise-to-image, 3D GAN, here DiscoFaceGAN (DFG) [10] for image manipulation in Fig. 12. Although GAN inversion successfully retrieves latent codes that well project the image in DFG, manipulating these codes for high-quality editing is still challenging. On the contrary, our approach provides both good image reconstruction and high-quality disentangled editing.

We further notice DFG is primarily trained to disentangle its λ space where 3DMM parameter p lies in, while its image embedding is conducted in $\mathcal{W}+$
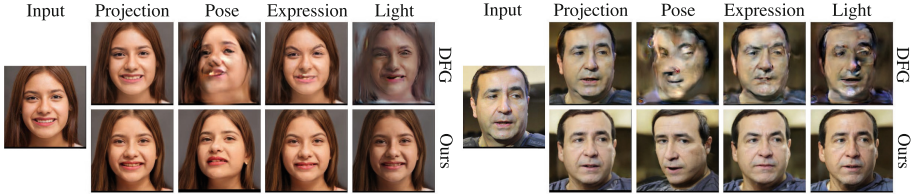
Fig. 12. Comparing 3D-FM GAN with GAN inversion + DiscoFaceGAN (DFG) [10] for face editing. Although GAN inversion allows good projection with DFG, providing faithful manipulation with the inverted codes remains challenging. On the contrary, our method achieves good reconstruction and high-quality disentangled editing.


Fig. 13. Analysis of $\mathcal{W}+$ and λ space of DiscoFaceGAN (DFG) [10]. While DFG's image embedding is performed in $\mathcal{W}+$ space (**Row 1**) for editing, we also extract the input's 3DMM parameter p by **FR** and conduct the same editing in λ space (**Row 2**). While its λ space provides realistic synthesis, its inverted code in $\mathcal{W}+$ falls off the manifold of good editability trained in λ. In contrast, 3D-FM GAN (**Row 3**) uses the same editing spaces for training and testing, which easily leads to high-quality editing.

space[1]. This already creates an obvious disparity between training and test time tasks as different latent spaces are used. We thus analyze how DFG behaves in these two spaces in Fig. 13, where we retrieve p from photo input by **FR** and perform the same editing in both λ and $\mathcal{W}+$ space. We clearly see that DFG's λ space is well trained for realistic disentangled synthesis, yet its $\mathcal{W}+$ space is not. This suggests that despite the inverted code in $\mathcal{W}+$ can well embed the image, it may not lie on the manifold with good editability trained from λ. On the contrary, our method utilizes the same editing space for both training and testing, and this consistency guarantees the high-quality manipulation.

Other 3D GANs. We further compare 3D-FM GAN with CONFIG [25], StyleRig [44], and PIE [44] on disentangled image editing and reference-based synthesis tasks in Fig. 14. Our model clearly shows a larger range of pose editability and better identity preservation over CONFIG. Compared to GAN inversion [2] + StyleRig, 3D-FM GAN again provides more realistic synthesis with much less

---

[1] In [10], it claims that DFG's λ space is not feasible for image embedding.
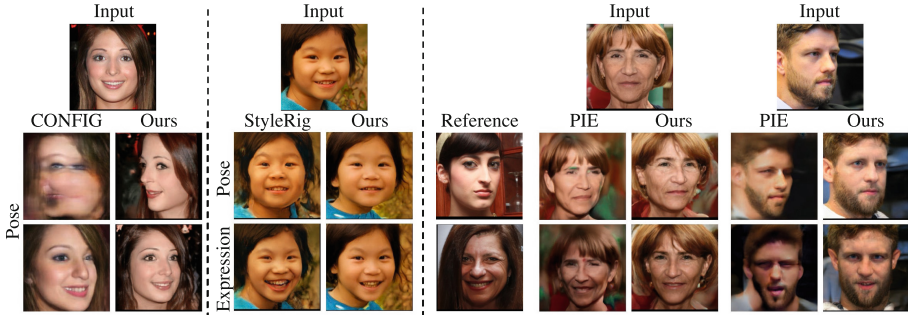
Fig. 14. Comparing 3D-FM GAN with other 3D-aware GANs for image manipulation. Specifically, we compare with CONFIG [25] on pose editing and StyleRig [44] on both pose and expression editing. We compare with PIE [43] on a reference-based synthesis task, where the pose, expression, and light are extracted from the reference images. Our method again shows the best editing results over all prior arts.

artifacts around the face. While PIE could not provide high-quality manipulation when the input is at large pose rotation angle (**2nd Example**), 3D-FM GAN still achieves faithful editing, indicating its advantage in better generalizability.

Frontalization. In Fig. 15, we compare our model with prior methods [21,31, 47,54] on the tasks of face frontalization on LFW [20], where our method best preserves the face identity and produce more photo-realistic images.

Relighting. We show portrait relighting comparison with [10,37] on Multi-PIE [16] in Fig. 16. While Sfs-Net and DFG do not synthesize realistic manipulation with artifacts in background and around the face, our method shows higher photo-realism and can preserve background pattern like the clothes around the neck. Moreover, DFG completely changes the skin tone of the person, whilst our method meshes the extreme indoor light with the skin tone more naturally.

Fig. 15. Face frontalization on LFW [20] images. Our model preserves the best identity with a higher photo realism.
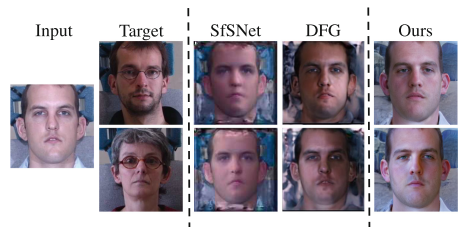

6   Ablation Study

6.1   Training Strategy
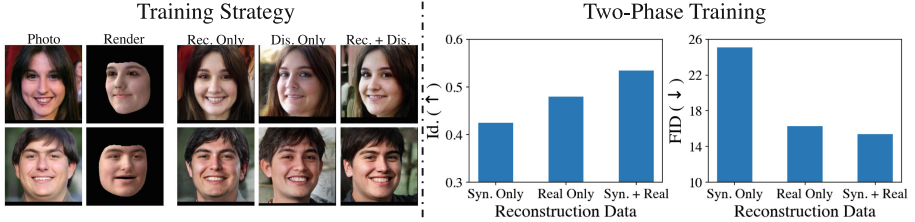
In Sect. 3.3, we propose to do alternate training between reconstruction and disentanglement. To understand

Fig. 16. Comparing prior arts on portrait relighting with Multi-PIE [16] images. Our method provides a higher photo-realism, and merges the indoor light with the person's skin tone more naturally.

Fig. 17. Ablation study. **Left:** Effectiveness of alternate training. Compared to reconstruction or disentangled only training, alternate training scheme acquire information for both editing and identity preservation. **Right:** Effectiveness of two-phase training. Using FFHQ for reconstruction significantly improves the photo-realism. The two-phase scheme, fine-tuning with synthetic reconstruction first and then switch to FFHQ, further improves identity preservation.

its effectiveness, we conduct a study and find both of them are essential to learn high-quality identity-preserved editing with results shown in Fig. 17 (**Left**). Specifically, we perform 140k training iterations with synthetic data on a Render-$\mathcal{W}$ with the following variants: (1) reconstruction only training (**Col. 3**); (2) disentangled only training (**Col. 4**); (3) alternate training (70k iterations for each) between reconstruction and disetanglement (**Col. 5**). While reconstruction only training enables good identity preservation, it's hard for the model to respond to the editing signals. On the other hand, disentangled only training provides good editability, but fails to preserve the identity like face shapes, ages, etc. Different from them, alternating between these two strategies helps the model achieve a much better performance as it picks up information from both sides.

6.2 Two-Phase Training

To study our two-phase training scheme, where different data are used for reconstruction training, we adopt a Render-$\mathcal{W}$ architecture and train for 280k iterations with the following schedules: (1) synthetic data only reconstruction; (2) real data only reconstruction; (3) 140K iterations of synthetic reconstruction followed by 140K iterations of real reconstruction. In Fig. 17 (**Right**), we see that incorporating real data for reconstruction training is crucial for achieving high photo-realism. Moreover, the two-phase training scheme, (3), yields the best identity preservation.

7 Conclusion

In this work, we propose 3D-FM GAN, a novel framework for high-quality, 3D-controllable, existing face manipulation. Unlike prior works, our model is trained exactly for the task of face manipulation, and does not require any manual tuning after the learning phase. We design two training strategies that are both essential for the model to gain abilities of high-quality, identity-preserved editing. We also study the information encoding scheme on StyleGAN's latent spaces, which leads

us to a novel multiplicative co-modulation architecture. We carry out qualitative and quantitative evaluations on our model, where it all demonstrates good editability, strong identity preservation and high photo-realism, outperforming the state of the arts. More surprisingly, our model shows a strong generalizability, where it can perform controllable editing on out-of-domain artistic faces.

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: how to embed images into the stylegan latent space? In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4432–4441 (2019)
2. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: how to edit the embedded images? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8296–8305 (2020)
3. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: Towards open-set identity preserving face synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6713–6722 (2018)
4. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: Proceedings of the 26th Annual Conference on Computer Graphics And Interactive Techniques, pp. 187–194 (1999)
5. Booth, J., Roussos, A., Ponniah, A., Dunaway, D., Zafeiriou, S.: Large scale 3D morphable models. Int. J. Comput. Vis. **126**(2), 233–254 (2018)
6. Bühler, M.C., Meka, A., Li, G., Beeler, T., Hilliges, O.: Varitex: variational neural face textures. arXiv preprint arXiv:2104.05988 (2021)
7. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: International Conference on Computer Vision (2017)
8. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: unified generative adversarial networks for multi-domain image-to-image translation, pp. 8789–8797 (2018)
9. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
10. Deng, Y., Yang, J., Chen, D., Wen, F., Tong, X.: Disentangled and controllable face image generation via 3D imitative-contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5154–5163 (2020)
11. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3D face reconstruction with weakly-supervised learning: from single image to image set. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
12. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: Ganfit: generative adversarial network fitting for high fidelity 3D face reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1155–1164 (2019)
13. Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., Freeman, W.T.: Unsupervised training for 3D morphable model regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8377–8386 (2018)

14. Ghosh, P., Gupta, P.S., Uziel, R., Ranjan, A., Black, M.J., Bolkart, T.: Gif: generative interpretable faces. In: 2020 International Conference on 3D Vision (3DV), pp. 868–878. IEEE (2020)
15. Goodfellow, I.,et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
16. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image Vis. Comput. **28**(5), 807–813 (2010)
17. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: discovering interpretable GAN controls. arXiv preprint arXiv:2004.02546 (2020)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems, vol. 30 (2017)
20. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report. 07–49, University of Massachusetts, Amherst (2007)
21. Huang, R., Zhang, S., Li, T., He, R.: Beyond face rotation: global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2439–2448 (2017)
22. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANS for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
23. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
24. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119 (2020)
25. Kowalski, M., Garbin, S.J., Estellers, V., Baltrušaitis, T., Johnson, M., Shotton, J.: CONFIG: controllable neural face image generation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12356, pp. 299–315. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_18
26. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017)
27. Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., Yang, Y.L.: HoloGAN: unsupervised learning of 3D representations from natural images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7588–7597 (2019)
28. Parke, F.I.: A parametric model for human faces. The University of Utah (1974)
29. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3D face model for pose and illumination invariant face recognition. In: 2009 sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 296–301. IEEE (2009)
30. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: GANimation: anatomically-aware facial animation from a single image. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 818–833 (2018)
31. Qian, Y., Deng, W., Hu, J.: Unsupervised face normalization with extreme pose and expression in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9851–9858 (2019)

32. Richardson, E., et al.: Encoding in style: a StyleGAN encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2287–2296 (2021)
33. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1259–1268 (2017)
34. Saito, S., Wei, L., Hu, L., Nagano, K., Li, H.: Photorealistic facial texture inference using deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5144–5153 (2017)
35. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3D face shape and expression from an image without 3D supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7763–7772 (2019)
36. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1576–1585 (2017)
37. Sengupta, S., Kanazawa, A., Castillo, C.D., Jacobs, D.W.: SfSNeT: learning shape, reflectance and illuminance of faces in the wild'. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6296–6305 (2018)
38. Shen, Y., Luo, P., Yan, J., Wang, X., Tang, X.: FaceID-GAN: learning a symmetry three-player GAN for identity-preserving face synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 821–830 (2018)
39. Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in GANS. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1532–1540 (2021)
40. Shi, Y., Aggarwal, D., Jain, A.K.: Lifting 2D StyleGAN for 3D-aware face generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6258–6266 (2021)
41. Shoshan, A., Bhonker, N., Kviatkovsky, I., Medioni, G.: Gan-control: explicitly controllable GANs. arXiv preprint arXiv:2101.02477 (2021)
42. Slossberg, R., Shamai, G., Kimmel, R.: High quality facial surface and texture synthesis via generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)
43. Tewari, A., et al.: PIE: portrait image embedding for semantic control. ACM Trans. Graph. (TOG) **39**(6), 1–14 (2020)
44. Tewari, A., et al.: StyleRig: rigging StyleGAN for 3D control over portrait images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6142–6151 (2020)
45. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for StyleGAN image manipulation. ACM Trans. Graph. (TOG) **40**(4), 1–14 (2021)
46. Tran, L., Liu, F., Liu, X.: Towards high-fidelity nonlinear 3D face morphable model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1126–1135 (2019)
47. Tran, L., Yin, X., Liu, X.: Disentangled representation learning GAN for pose-invariant face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1415–1424 (2017)
48. Usman, B., Dufour, N., Saenko, K., Bregler, C.: PuppetGAN: cross-domain image manipulation by demonstration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9450–9458 (2019)
49. Ververas, E., Zafeiriou, S.: SliderGAN: synthesizing expressive face images by sliding 3D blendshape parameters. Int. J. Comput. Vis. **128**(10), 2629–2650 (2020)

50. Xia, W., Zhang, Y., Yang, Y., Xue, J.H., Zhou, B., Yang, M.H.: Gan inversion: a survey. arXiv preprint arXiv:2101.05278 (2021)
51. Xiao, T., Hong, J., Ma, J.: ELEGANT: exchanging latent encodings with GAN for transferring multiple face attributes. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 168–184 (2018)
52. Yaniv, J., Newman, Y., Shamir, A.: The face of art: landmark detection and geometric style in portraits. ACM Trans. Graph. (TOG) **38**(4), 1–15 (2019)
53. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
54. Zhao, J., et al.: Towards pose invariant face recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2207–2216 (2018)
55. Zhao, S., et al.: Large scale image completion via co-modulated generative adversarial networks. arXiv preprint arXiv:2103.10428 (2021)
56. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep single-image portrait relighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7194–7202 (2019)