3DFaceShop: Explicitly Controllable 3D-Aware Portrait Generation

Junshu Tang, Bo Zhang, Binxin Yang, Ting Zhang, Dong Chen, Lizhuang Ma, Fang Wen

Abstract—In contrast to the traditional avatar creation pipeline which is a costly process, contemporary generative approaches directly learn the data distribution from photographs. While plenty of works extend unconditional generative models and achieve some levels of controllability, it is still challenging to ensure multi-view consistency, especially in large poses. In this work, we propose a network that generates 3D-aware portraits while being controllable according to semantic parameters regarding pose, identity, expression and illumination. Our network uses neural scene representation to model 3D-aware portraits, whose generation is guided by a parametric face model that supports explicit control. While the latent disentanglement can be further enhanced by contrasting images with partially different attributes, there still exists noticeable inconsistency in non-face areas, *e.g.*, hair and background, when animating expressions. We solve this by proposing a volume blending strategy in which we form a composite output by blending dynamic and static areas, with two parts segmented from the jointly learned semantic field. Our method outperforms prior arts in extensive experiments, producing realistic portraits with vivid expression in natural lighting when viewed from free viewpoints. It also demonstrates generalization ability to real images as well as out-of-domain data, showing great promise in real applications.

Index Terms—Controllable 3D portrait generation, 3D morphable models, Neural radiance field, 3D-aware GAN.

1 Introduction

Can you imagine synthesizing a collection of photorealistic portraits that are animatable in 3D or waking up a photo as if the character is talking in front of us? Rendering controllable portraits is of significant importance to a variety of fields like film industry, video games, extended reality or immersive telecommunication. Traditional graphics pipeline [1], [2], [3], [4], [5] involves specialized 3D model creation with texture decoration which is then illuminated with realistic lighting and rendered using a physics-based renderer. While easily controllable, it is challenging to produce a myriad of avatars with photo-realistic quality.

In recent years we have seen a surge of neural rendering approaches [6], [7] that generate highly photo-realistic faces in a data-driven manner. In particular, generative adversarial networks (GANs) [8], [9], [10], [11] hold the state of the arts, capable to synthesize high-resolution novel faces that are indistinguishable from the real ones. A prominent property of these off-the-shelf GANs is that they are often equipped with a semantic and disentangled latent space, hence one can animate the face or control specific facial attributes by traversing the latent space in the direction that correlates with the manipulated attribute [12], [13], [14], [15], [16], [16], [17], [18]. Moreover, since the whole rendering is fully differentiable, real photos can be inversely projected to the latent space [19], [20], [21] and undergo the same semantic editing process.

Due to the rise of augmented reality applications, there is an increasing demand for 3D face rendering such that faces can be rendered in different view angles while maintaining geometry consistency. While aforementioned 2D GANs [15], [17], [22], [23] allow explicit head pose control to some extent, they fail to guarantee appearance consistency, leading to inconsistent identity or facial attributes when viewed from vastly different angles. To solve this, there are a few attempts [12], [14] that leverage parametric 3D face models [24], [25] to guide the generative process, thus faces can be explicitly controlled according to a set of semantic parameters relating to head pose, face shape, expression as well as illumination in the way like the computer animation workflow [26]. Notwithstanding the improved disentanglement and controllability, the outputs yielded by 2D GANs are not truly 3D-aware in that one may still observe noticeable appearance and shape variations under distinct views of the same subject.

Meanwhile, the neural scene representations [6], [27], [28], [29], [30], [31] emerge to become an expressive 3D-structure aware representation for general scenes, objects or persons. Among them, the seminar work, neural radiance field (NeRF) [27], characterizes the complex scene as a continuous volume with radiance and density at each location, which can be rendered to 2D observations via differentiable volumetric rendering [32]. Using a partial set of 2D images as supervision, NeRF-based methods can faithfully reconstruct the scene that can be rendered at free viewpoint with photo-realistic quality. Thereafter, the research community turns to 3D-aware generative models [33], [34], [34], [35],

- *The work is done when J. Tang and B. Yang are research interns at Microsoft Research Asia.*
- *J. Tang and L. Ma are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China. E-mail: tangjs@sjtu.edu.cn, ma-lz@cs.sjtu.edu.cn.*
- *B. Yang is with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, 230026, China, and also with Zhangjiang Laboratory, Shanghai, China. E-mail: tennyson@mail.ustc.edu.cn.*
- *B. Zhang, T. Zhang, D. Chen, and F. Wen are with Microsoft Research Asia, Beijing, 100080, China. E-mail: {zhanbo, tinzhan, doch, fangwen}@microsoft.com.*
- *B. Zhang and L. Ma are the corresponding authors.*

- *Additional video results and code will be available on the project webpage.*

[36], [37], [38], [39], [40] that learn to produce the 3D neural representation from 2D imagery. Now state-of-the-art approaches can generate compelling 3D portraits, yet the results are not as vivid as real persons since they are neither animatable nor controllable.

In this work, we propose to generate 3D portraits that can be explicitly controlled by allowing the user edits upon a group of semantic parameters. To this end, we try to make the best of both the explicit parametric model and the neural face representation: the 3D Morphable Model (3DMM) [24], [25] provides the desired controllability in terms of face shape, expression and lighting; whereas the neural representation ensures multi-view consistency and offers photorealism. Specifically, the generation network is built upon the tri-plane representation proposed by [35] which can be efficiently rendered, and the generation of such proxy representation is conditioned on the control space of 3DMM. The rendered images are expected to lie on the real face distribution as guaranteed by the adversarial loss [41], with the appearance resembling the rendered face mesh. However, the 3DMM guidance does not necessarily lead to disentangled control, since the latent sub-vector that accounts for a specific facial attribute may interfere with other attributes unexpectedly. Hence, we further improve the disentanglement by forming contrastive image pairs that differ in partial latent segments and enforcing the consistency for the attributes that share identical latent codes. As such, different latent sub-vectors bring independent effects to the final output and changing a segment of latent codes will not alter uncorrelated face properties.

Nonetheless, it is still challenging to ensure ideal disentanglement, especially for non-face regions like hair, clothes and background, when controlling the facial expression. In fact, the expression animation is desired to mainly affect the facial area while bringing little influence on other parts. Given this, we propose a volume blending scheme which retains the radiance field of static area, *i.e.*, the non-face region, in the final blended output. To accomplish this, we need to precisely locate the points belonging to the face area, hence we let the generation network simultaneously learn the semantic field using the online parsing results as supervision. In this way, we can determine the semantics of continuous space and accordingly segment the portrait in 3D. During inference, the output image is rendered from the blended radiance field with edits only applied to the localized face area. As demonstrated in the experiments, such a volume blending strategy effectively benefits the disentanglement, bringing much improved temporal consistency during the facial animation.

We compare our approach with multiple state-of-the-art methods, including controllable 2D GANs and talking head works, in which our approach performs favorably in terms of perceptual quality, control diversity, and disentanglement capability. Our method can generate lifelike 3D portraits in 512×512 resolution and exhibits significant advantages for disentangled control in large poses which is particularly challenging for prior arts. We further show that our controllable portrait generation network can be personalized on real persons or even out-of-domain images, *e.g.*, cartoons, and brings the static character to the 3D world with natural lighting and vivid expressions. Moreover, the learned semantic field offers accurate portrait matting and thus can enable background replacement with remarkable quality. We summarize our contributions as follows:

- We propose a controllable 3D portrait generation network that produces neural face representation conditioned on a set of semantic control parameters, with the disentangled control achieved by the guidance of parametric face models. Our approach allows easy and disentangled control for the pose, identity, expression, illumination as well as background appearance.
- To further improve the disentanglement, we propose a volume blending scheme, which blends the dynamic and static radiance fields of animated faces, with two parts separated by the jointly learned semantic field.
- Our method outperforms the state of the arts on controllable 3D portrait generation in terms of both quantitative measure and qualitative evaluation. The network also performs well on real images and out-of-domain images like cartoon faces, showing great potential for various extended reality applications.

2 Related Work

Photo-realistic face image synthesis and animation have been the longstanding focus of computer graphics as well as computer vision. Over the past years, rapid progress has been achieved with a large volume of efforts dedicated [10], [11], [12], [13], [14], [35], [36], [41], [42], [43], [44]. Here we present a brief overview. For a more comprehensive review, please refer to the recent surveys [25], [45], [46], [47], [48].

2.1 Controllable Face Image Editing

Controlling and editing the appearance of the images, especially for face images, is an important feature demanded in many real-world applications. Early works [49], [50], [51] heavily depend on additional manual annotations for training a specific face animation model for the specific attribute. Later, a great many efforts [12], [13], [14], [15], [52], [53], [54], [55], [56] aim at learning a disentangled and meaningful latent space for face images. It is desired that different dimensions in the latent space characterize different facial attributes so that editing certain latent dimensions enables some kind of facial animation. However, it is not guaranteed that factors of interested attributes are disentangled and often facial attributes are mingled in the latent space [57], [58], which undermines the quality of face editing.

In order to enable more precise control over the generated images, many works [12], [13], [14], [15], [43], [59], [60], [61], [62], [63] have been proposed to incorporate 3D priors from parametric face models such as 3D Morphable Models (3DMMs), into GAN-based generative models. For example, DiscoFaceGAN [14] proposes an imitative-contrastive paradigm that enforces the generative network to mimic the rendering process of 3DMM. After training, it can enable precise control of the desired face properties such as pose, expression, illumination and so forth. StyleRig [12] describes a similar method to control StyleGAN via a 3DMM. GAN-control [15] also enhances GANs with an explicitly disentangled latent space and can edit the image by setting exact attributes such as age, pose, expression, etc.

These methods are still based on 2D image generators, and they are thus subject to severe 3D inconsistency issues due to the non-physical image rendering process of 2D GANs, especially under large expression and pose variations. Recent methods [44], [64], [65], [66], [67], [68] utilize neural representation to produce view-consistent and editable portrait images. FENerf [64] and IDE3D [65] utilize rendered semantic mask to edit 3D volume via GAN inversion, but they cannot produce continues animation results. Other methods [44], [66], [67], [68] incorporate 3DMM knowledge into volumetric rendering using neural scene representations to achieve consistent face editing. For example, HeadNeRF [44] integrates the neural radiance field to the parametric representation of the human head and is trained using annotated multi-view datasets. Additionally, the concurrent works cGOF [68] and GNARF [69] also leverage the parametric face model and propose a conditional generative occupancy field for face expression animation. In contrast, we focus on generative modeling that creates high-resolution and photo-realistic portraits including vivid expressions and illuminations.

2.2 3D Morphable Models

3D morphable models were first proposed in [24] as a general and statistical representation model for face shape and appearance, which are typically learned from 3D scans of multiple people [3], [70], [71], [72], [73]. In this way, faces are parameterized to a low-dimensional face space consisting of identity, expression, and illumination, which can be used to reconstruct 3D face mesh and is widely used for 3D face representation [74], [75], [76], [77]. Such a parameterized space also allows explicit control of the face synthesis with semantically interpretable parameters [25], [78]. Rather than using time-consuming optimization approaches, recent works resort to deep neural networks for the 3D face model fitting [4], [79], [80], [81], [82]. Since the faces rendered by 3DMM often lack delicate details [83], [84], recent efforts [12], [13], [14] leverage deep generative models for more photo-realistic synthesis. Along this line, this work proposes an explicitly controllable 3D-aware generative model by combining the best of both worlds. Our method bridges the emerging 3D scene neural representation with the controllability of 3DMM, and achieves controllable 3D portrait generation with much improved multi-view consistency, compared with prior works.

2.3 Neural Scene Representations

In order to generate high-quality multi-view consistent images, neural scene representation using differentiable rendering [61], [85], [86], [87], [88], [89], [90], [91] that can be optimized on a training set of only 2D multi-view images has gained popularity in the past few years. Implicit representations [27], [28], [29], [92] in particular neural radiance field (NeRF) [33], [34], [36], [42], [93], [94] have been widely used in many areas such as 3D modeling [95], [96] and face/body digitization [66], [97], [98], [99], [100], [101], [102]. They characterize the 3D scene with a continuous function via a light MLP, which is memory-efficient. On the other hand, explicit representations [103], [104] such as discrete voxels allow fast evaluation but usually suffer from huge memory consumption. Based on the complementary benefits of fully explicit and implicit representations, several methods [35], [105], [106] have explored the hybrid explicit-implicit models. Among them, EG3D [35] shows high 3D GAN image quality by designing an efficient tri-plane hybrid 3D representation for unconditional 3D face synthesis. In this work, we leverage this hybrid representation and further design a controllable framework to support precise 3D control over the generated faces such as facial expression, head pose, lighting, etc.

3 Preliminaries: 3D-aware GANs

While contemporary generative adversarial networks (GANs) are capable to generate photo-realistic faces, the synthesized images are not multi-view consistent even though plenty of advanced disentangled control methods have been proposed. To ensure geometry consistency from different views, 3D-aware GANs learn to model the distribution of underlying 3D face geometry from a collection of images. The crux of these approaches is to choose an intermediate 3D representation, which can be further rendered into 2D images from different viewpoints in a differentiable manner. One of the most expressive 3D representations is the neural radiance field (NeRF) [27], which is a continuous volumetric representation. NeRF models the density $\sigma \in \mathbb{R}^+$ and view-dependent color $\boldsymbol{c} \in \mathbb{R}^3$ for each position $\boldsymbol{p} \in \mathbb{R}^3$ given the viewing direction $\boldsymbol{d} \in \mathbb{S}^2$. Hence, on the high level, the 3D-aware generator $\mathcal{G}$ parameterized by θ maps the latent code $\boldsymbol{z}$ in the latent space $\mathcal{Z}$ to the manifold of neural radiance fields:

$$\boldsymbol{c}(\boldsymbol{p}, \boldsymbol{d}), \sigma(\boldsymbol{p}) = \mathcal{G}_\theta(\boldsymbol{z}, \xi(\boldsymbol{p}), \xi(\boldsymbol{d})). \tag{1}$$

In practice, to better model the high-frequency details, the inputs $\boldsymbol{p}$ and $\boldsymbol{d}$ are represented with sinusoidal positional encoding $\xi(x) = (x, ..., \sin(2^k \pi x), \cos(2^k \pi x), ...)$ [33].

The scene represented by the above radiance field can be rendered into 2D images via volumetric rendering. Each rendered pixel in the image corresponds to a camera ray $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}$ shooting from the camera origin $\boldsymbol{o}$ and advancing in the direction $\boldsymbol{d}$. The ray traverses the radiance field, and the accumulated color along the ray is calculated by the following volume rendering integral [27],

$$\mathcal{C}(\boldsymbol{r}, \boldsymbol{z}) = \int_0^\infty T(t)\sigma(\boldsymbol{r}(t))\boldsymbol{c}(\boldsymbol{r}(t), \boldsymbol{d})dt, \tag{2}$$

where $T(t)$ denotes the accumulated transmittance along the ray:

$$T(t) = \exp\left(-\int_0^t \sigma(\boldsymbol{r}(s))ds\right). \tag{3}$$

In [27], a hierarchical sampling scheme is proposed to discretize this integral process by sampling multiple points along the ray. Parameterized with implicit multi-layer perceptrons (MLPs), NeRFs are quite expressive and can model the scene with photo-realistic quality.

4 Proposed Approach

In this work, we propose a 3D-aware GAN that allows explicit semantic control with respect to pose, identity, expression and illumination. To this end, we leverage a 3D-aware

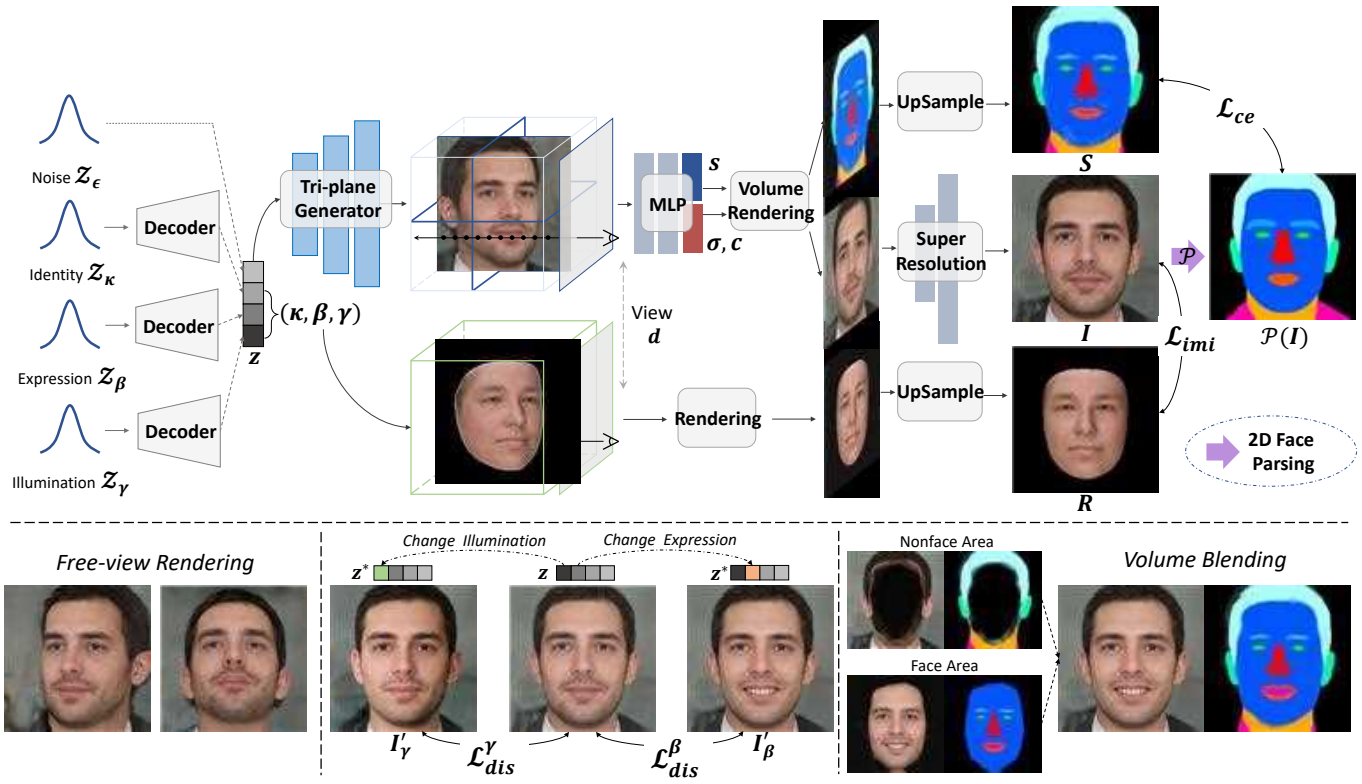


Fig. 1. **Architecture overview.** We first sample the 3DMM control parameters (κ, β, γ) by sampling from three separate decoders of pretrained VAE-GANs (Sec. 4.1). We train a generative adversarial network conditioned on these control parameters, which generates the tri-plane representation that can be further rendered into 3D-aware portrait images via volumetric rendering (Sec. 4.2). Meanwhile, we render the 3DMM face mesh using the same parameters and use it to guide the generation, thus enabling the generator semantic controllability (Sec. 4.3.1). Once such imitative learning converges, we further improve the control disentanglement by enforcing the consistency of contrastive image pairs (Sec. 4.3.2). During the radiance field generation, we simultaneously learn the portrait semantics in 3D space. With the accurately learned face parsing, we derive the final output which is a composition of the animated face region with the static non-face region(Sec. 4.4).

generator to ensure view consistency, and the generation is guided by a 3D face prior that admits semantic and interpretable control. As illustrated in Figure 1, the 3D-aware GAN samples from the latent space are formed by a set of control parameters, and the generated images are enforced to imitate the rendered face from the parametric face model. To ensure disentangled control, we compare generated image pairs in a contrastive manner — parameters associated with a certain attribute should not alter other attributes during generation. Nonetheless, inconsistency still occurs during semantic control, causing disturbing flickers in non-facial areas like hair, clothes and background. We remedy this issue by simultaneously predicting the semantic field and explicitly blending the radiance field accordingly so that only the facial area is manipulated with the non-face regions intact during the face control.

4.1 Semantic Control Space

We adopt the 3D Morphable Model (3DMM) [24] to parameterize the face attributes and use it to guide the 3D-aware generation. In 3DMM, faces can be modeled with a set of semantic parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{R}, \boldsymbol{t}) \in \mathbb{R}^{257}$. Specifically, $\boldsymbol{\alpha} \in \mathbb{R}^{80}$ describes the geometry of facial shape, $\boldsymbol{\beta} \in \mathbb{R}^{64}$ models the expression, $\boldsymbol{\delta} \in \mathbb{R}^{80}$ defines the albedo, $\boldsymbol{\gamma}$ characterizes the scene illumination, whereas $\boldsymbol{R} \in \mathrm{SO}(3)$ and $\boldsymbol{t} \in \mathbb{R}^3$ denote the head rotation and translation respectively.

The 3DMM yields a triangular mesh of $53k$ vertices. The face shape $\boldsymbol{S}$ and the appearance $\boldsymbol{A}$ can be modeled as an affine model:

$$\boldsymbol{S} = \bar{\boldsymbol{S}} + \tilde{\boldsymbol{S}}_{id}\boldsymbol{\alpha} + \tilde{\boldsymbol{S}}_{exp}\boldsymbol{\beta}, \tag{4}$$

$$\boldsymbol{A} = \bar{\boldsymbol{A}} + \tilde{\boldsymbol{A}}\boldsymbol{\delta}, \tag{5}$$

where $\bar{\boldsymbol{S}}$ and $\bar{\boldsymbol{A}}$ are the average shape and appearance, whereas $\tilde{\boldsymbol{S}}_{id}$, $\tilde{\boldsymbol{S}}_{exp}$ and $\tilde{\boldsymbol{A}}$ are two-dimensional bases that account for the variation space of face shape $\boldsymbol{\alpha}$, expression $\boldsymbol{\beta}$ and albedo $\boldsymbol{\delta}$ respectively. We choose $\tilde{\boldsymbol{S}}_{id}$ and $\tilde{\boldsymbol{A}}$ from the Basel Face Model (BFM) [75] which is computed from 200 face scans, and the expression bases $\tilde{\boldsymbol{S}}_{exp}$ are built from FaceWarehouse [74]. Jointly considering the face shape and albedo, we let $\boldsymbol{\kappa} = [\boldsymbol{\alpha}, \boldsymbol{\delta}] \in \mathbb{R}^{160}$ and use it to describe all the identity-related attributes. As for the lighting, we assume the faces to be Lambertian and approximate the scene illumination with Spherical Harmonics (SH). Specifically, for the vertex $\boldsymbol{p}_i$ with the surface normal $\boldsymbol{n}_i$ and skin color $\boldsymbol{A}_i$, its final color is computed as $\boldsymbol{A}_i \sum_{b=1}^{9} \boldsymbol{\gamma}' \boldsymbol{H}'(\boldsymbol{n}_i)$, where $\boldsymbol{H} : \mathbb{R}^3 \to \mathbb{R}$ is the SH basis function and $\boldsymbol{\gamma}' \in \mathbb{R}^3$ is the corresponding SH coefficient. Given the camera pose $[\boldsymbol{R}, \boldsymbol{t}]$, the face mesh can be rendered through Nvdiffrast [107].

In order to generate controllable 3D portraits by manipulating the attribute parameters $(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, we need to learn a disentangled $\mathcal{W}$ space which is mapped from the attribute space $\mathcal{Z}$, where the $\mathcal{Z}$ space is easy to sample from. To this

end, we first learn the latent space that accounts for the variations of each attribute. Specifically, we train separate variational auto-encoders (VAE) for identity $\boldsymbol{\kappa}$, expression $\boldsymbol{\beta}$ and illumination $\boldsymbol{\gamma}$ respectively: the encoder E_i of ith VAE maps the real 3DMM coefficients to latent code which can be used to faithfully reconstruct the original coefficients via decoder D_i. We assume Gaussian prior for the latent space, so the VAE is optimized as,

$$\begin{aligned}\mathcal{L}_{\text{VAE}_i} =& \mathbb{E}_{z\sim E_i(\boldsymbol{z}|\boldsymbol{x})}\|\boldsymbol{x} - D_i(\boldsymbol{x}|\boldsymbol{z})\|_2 && (6)\\ &+ \mu\text{KL}\big(E_i(\boldsymbol{z}|\boldsymbol{x})\|\mathcal{N}(\boldsymbol{0},\boldsymbol{I})\big) && (7)\\ &+ \mathcal{L}_{\text{VAE,GAN}}(\boldsymbol{x}), && (8)\end{aligned}$$

where the first term ℓ_2 reconstructs the input from the latent code and the second term penalizes the Kullback-Leibler (KL) divergence between the latent distribution and the normal distribution. Training VAE with ℓ_2 reconstruction objective essentially assumes Gaussian distribution for $p(\boldsymbol{x}|\boldsymbol{z})$ which is prone to produce averaged reconstruction, or mean face. Hence, we additionally introduce an adversarial loss [108], $\mathcal{L}_{\text{VAE,GAN}}$, to better distinguish the reconstructed attribute coefficients from real sampled ones. Experimental results show that this allows us to sample more diverse attributes. Besides FFHQ dataset [10], the VAE training also includes a talking face dataset [109] containing exaggerated expressions, and thus we obtain a latent space with more diverse expressions.

Once the VAEs are trained, we can derive plausible 3DMM coefficients $(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ by sampling from the VAE latent spaces. These coefficients, along with a random Gaussian noise $\boldsymbol{\epsilon} \in \mathcal{Z}_\epsilon$ accounting for the remaining variations (*e.g.*, background), define the full control space of faces.

4.2 3DMM Conditioned 3D-Aware GAN

We expect the 3D-aware GAN that conditions on $\boldsymbol{z} := (\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\epsilon})$ can generate faces that accurately resemble the identity, expression, and illumination of the 3DMM renderings. Hence, as illustrated in Figure 1, we train a 3D-aware GAN that starts from the sampling space $\mathcal{Z} = (\mathcal{Z}_\kappa, \mathcal{Z}_\beta, \mathcal{Z}_\gamma, \mathcal{Z}_\epsilon)$ and generates 3D faces that demonstrate the desired properties. Formally, the radiance field is generated by,

$$\boldsymbol{c}(\boldsymbol{p}, \boldsymbol{d}), \sigma(\boldsymbol{p}) = \mathcal{G}_\theta(\boldsymbol{z}, \xi(\boldsymbol{p}), \xi(\boldsymbol{d})). \quad (9)$$

In this work, instead of using MLPs to directly regress the continuous radiance field, we adopt the tri-plane representation recently proposed by [35] which factorizes the 3D space using three orthogonal feature planes, denoted as $\boldsymbol{F}_{xy}, \boldsymbol{F}_{xz}, \boldsymbol{F}_{yz} \in \mathcal{R}^{H\times W\times C}$. Since much of the scene information is memorized explicitly while the decoder computed in the ray tracing is lightweight, it is computationally efficient to render the tri-plane representation while not compromising the expressivity. Specifically, for the 3D point $\boldsymbol{p}$, we can project it onto the feature planes and obtain the aggregated feature by summing the retrieved feature from each feature plane. Such position-wise features are thereafter decoded to color and density as required by the volumetric rendering with a shallow MLP decoder, *i.e.*,

$$\boldsymbol{c}(\boldsymbol{p}, \boldsymbol{d}), \sigma(\boldsymbol{p}) = \mathcal{G}_\theta^{\text{MLP}}\big(\boldsymbol{F}_{xy}(\boldsymbol{p}_{xy}) + \boldsymbol{F}_{xz}(\boldsymbol{p}_{xz}) + \boldsymbol{F}_{yz}(\boldsymbol{p}_{yz})\big). \quad (10)$$

Another significant advantage of using tri-plane representation is that one can directly take advantage of a powerful 2D CNN-based generator, *e.g.*, StyleGAN [10], [11], to generate each feature plane. Hence the 3D-aware generator could enjoy many effective training strategies built for 2D GANs. Therefore, we map the control parameters to $\boldsymbol{w} \in \mathbb{R}^{18\times 512}$ code in the $\mathcal{W}$ space which is known to be disentangled, and the non-linear mapping $f_\theta : Z \to \mathcal{W}$ is implemented with eight fully connected layers. The different layers of the generator are modulated by the $\boldsymbol{w}$ code and output the tri-plane feature maps,

$$\{\boldsymbol{F}_{xy}, \boldsymbol{F}_{xz}, \boldsymbol{F}_{yz}\} = \mathcal{G}_\theta^{\text{CNN}}(\boldsymbol{w}, \boldsymbol{d}). \quad (11)$$

Here, the tri-plane features are pose-dependent in that some facial attributes, *e.g.*, expression, often correlate with the head pose in the face image capturing.

4.2.1 Generative Tri-plane Training Details

While such an explicit-implicit representation greatly reduces the memory footprint, it is still challenging to scale the method for high-resolution generation. One workaround is to rely on an image super-resolution network to enhance the fine details, yet this will sacrifice the view consistency and cause annoying flickers when changing the viewpoint. To alleviate this, following [35] we introduce a dual discriminator $\mathcal{D}_\theta$ which does not only discriminate the realism of the generation outputs but also examines the consistency between the 3D-aware low-resolution outputs and the resolution-enhanced counterpart. In this work, our network first generates coarse 3D-aware portraits at 128×128 resolution and ultimately yields 512×512 outputs by the super-resolution module.

At training time, the real images are sampled from the data distribution p_{data}. To render generated images, we randomly sample the camera position from a unit sphere with the camera pointing to the sphere center [33]. The sampled yaw and pitch distributions p_{cam} are pre-computed from the training dataset. The tri-plane generation network is trained until an adversarially trained discriminator cannot distinguish the rendered generated images from the real ones. Both the generator and the discriminator are trained using the hinge loss [11]:

$$\begin{aligned}\mathcal{L}_{\text{GAN}}^{\mathcal{D}} =& -\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\big[\min(0, -1 + \mathcal{D}_\theta(\boldsymbol{x}))\big]\\ &-\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}, \boldsymbol{d}\sim p_{\text{cam}}}\big[\min\big(0, -1 - \mathcal{D}_\theta(\mathcal{C}(\mathcal{G}_\theta(\boldsymbol{z}, \boldsymbol{d})))\big)\big]\\ &+\lambda_{R1}\mathcal{L}_{\text{GP}},\\ \mathcal{L}_{\text{GAN}}^{\mathcal{G}} =& -\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}, \boldsymbol{d}\sim p_{\text{cam}}}\big[\mathcal{D}_\theta(\mathcal{C}(\mathcal{G}_\theta(\boldsymbol{z}, \boldsymbol{d})))\big],\end{aligned} \quad (12)$$

where, $\mathcal{C}$ denotes the volumetric renderer given by Equation 2 and $\mathcal{L}_{\text{GP}} = \|\nabla_\theta[\mathcal{D}_\theta(\boldsymbol{x})]\|^2$ is the R_1 gradient penalty loss [110] that helps stabilize the adversarial training.

4.2.2 Simultaneously Learned Semantic Field

On top of obtaining the image renderings, we also train a semantic radiance field so that we can parse the portrait in 3D and better enforce the disentanglement. Specifically, we propose a multi-head decoder which consists of an apparent head and a semantic head: the appearance head outputs the RGB feature and volume density, whereas the semantic head translates the feature to the semantic prediction at

each point. Both heads are two-layer MLPs with the first layer shared. Let $\boldsymbol{s}(\boldsymbol{p})$ denote the probability of K semantic classes for point $\boldsymbol{p}$, *i.e.*, $\boldsymbol{s}(\boldsymbol{p}) = \mathcal{G}_\theta^{\text{MLP}_s}(F(\boldsymbol{p})))$, the semantic field is rendered in a similar form as Equation 2 except that the semantic field $\boldsymbol{s}$ does not depend on the viewing direction, which is,

$$\mathcal{S}(\boldsymbol{r}) = \int_0^\infty T(t)\sigma(\boldsymbol{r}(t))\boldsymbol{s}(\boldsymbol{r}(t))dt, \tag{13}$$

where $\mathcal{S}(\boldsymbol{r}) \in \mathbb{R}^{HW\times K}$ is the rendered semantic mask. Likewise, we train this semantic field using image-level supervision. Specifically, we leverage a pretrained face parsing model $\mathcal{P}$ [111] to online extract the semantic mask for the generated image $\boldsymbol{I}$ and use this result as the 2D supervision. We minimize the categorical cross-entropy between the rendered semantic mask and the ground truth:

$$\mathcal{L}_{\text{ce}} = -\frac{1}{H\times W}\sum_{i=1}^{H\times W}\sum_{k=1}^{K}\mathcal{P}(\boldsymbol{I})_{i,k}\log(\boldsymbol{S}_{i,k}), \tag{14}$$

where i indexes the $H\times W$ image and k is the semantic label index. In the following, we will show how the accurately learned semantic field facilitates disentangled control.

4.3 Learning Controllable and Disentangled Radiance Field

Given the camera pose, our method can generate the image $\boldsymbol{I}$ along with the semantic mask $\boldsymbol{S}$ conditioned on the control parameters $(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\epsilon})$. We expect the generated faces can accurately resemble the identity, expression, and illumination of the 3DMM face $\boldsymbol{R}$ rendered from the same parameters. We thus enable explicit controllability by feeding different semantic parameters, which is achieved in the first imitative learning stage. Moreover, the control is desired to be disentangled — editing one specific attribute should not cause changes in other attributes. We enforce such independent attribute editing in the disentanglement learning stage. We subsequently describe the two training stages.

4.3.1 *Imitative Learning*

In this stage, we encourage the similarity of $\boldsymbol{I}$ and $\boldsymbol{R}$. Since the 3DMM model only renders the face area without hair and background, the training is actually guided using the blended face rendering $\boldsymbol{R}' = \boldsymbol{R} + \boldsymbol{I}\odot(\boldsymbol{1}-\boldsymbol{B})$, where $\boldsymbol{B}$ is the face mask. The texture difference is penalized with ℓ_2 loss, *i.e.*, $\mathcal{L}_{\text{tex}} = \|\boldsymbol{I}-\boldsymbol{R}'\|_2$, and the identity similarity is encouraged by enforcing the identity loss , *i.e.*, $\mathcal{L}_{\text{id}} = 1 - <\mathcal{F}(\boldsymbol{I}), \mathcal{F}(\boldsymbol{R}')>$, where $\mathcal{F}$ is the embedding extracted from a pretrained face recognition model [112] and $<\cdot>$ denotes feature cosine similarity.

To further encourage the expression similarity, we introduce the 3D landmark loss. We denote the 68 3D landmarks of the guided 3DMM image as $\boldsymbol{l}^R \in \mathbb{R}^{68\times 3}$. For the generated portrait $\boldsymbol{I}$, we adopt the differentiable face reconstruction method [113] and reconstruct a new face mesh with reconstructed landmarks $\boldsymbol{l}^I \in \mathbb{R}^{68\times 3}$. To better capture the subtle expressions, we use the re-weighted landmark loss:

$$\mathcal{L}_{\text{lmk}} = \sum_i^{68} w_i\|\boldsymbol{l}_i^I - \boldsymbol{l}_i^R\|_2, \tag{15}$$

(a) Ours w/o $\mathcal{L}_{\text{lip}}$ (b) Ours

Fig. 2. **The effect of lip style loss.** We visualize results of the same identity using the model trained with (left) or without (right) lip loss, where we highlight the lip color. Training without the lip loss leads to obvious lip appearance change as shown in (a).

where w_i denotes the weight for different landmarks. By default, we set $w = 1$ but apply $w = 100$ for landmarks relating to eyebrows and mouth.

Similarly, to accurately constrain the illumination, we enforce the similarity between the reconstructed illumination coefficients $\dot{\boldsymbol{\gamma}}$ and the input illumination condition $\boldsymbol{\gamma}$:

$$\mathcal{L}_{\text{ill}} = \|\boldsymbol{\gamma} - \dot{\boldsymbol{\gamma}}\|_2. \tag{16}$$

Therefore, imitative learning optimizes the following loss:

$$\mathcal{L}_{\text{imi}} = \lambda_{\text{tex}}\mathcal{L}_{\text{tex}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{lmk}}\mathcal{L}_{\text{lmk}} + \lambda_{\text{ill}}\mathcal{L}_{\text{ill}} + \lambda_{\text{ce}}\mathcal{L}_{\text{ce}}, \tag{17}$$

where $\lambda_{(\cdot)}$ denotes the hyper-parameter that balances the terms.

4.3.2 *Disentanglement Learning*

To learn a more disentangled latent space, the network is further regularized such that the image pairs that differ in one type of control parameter will remain consistent in other properties. Next, we modify one attribute each time and introduce the loss functions when contrasting the identity, expression, and illumination respectively.

When contrasting the image pairs $\boldsymbol{I}$ and $\boldsymbol{I}^{\boldsymbol{\kappa}}$ that differ in the identity parameter $\boldsymbol{\kappa}$, inconsistency is likely to happen in non-face areas like hair, clothes, and backgrounds. To address this issue, we render the semantic mask $\boldsymbol{S}$ and convert it to a binary mask $\tilde{\boldsymbol{S}}_{\text{bg}}$ with ones indicating the non-face parts. We then enforce the consistency for non-face parts by imposing the identity-aware disentanglement loss :

$$\mathcal{L}_{\text{dis}}^{\kappa} = \|\boldsymbol{I}^{\kappa}\odot\tilde{\boldsymbol{S}}_{bg}^{\kappa} - \boldsymbol{I}\odot\tilde{\boldsymbol{S}}_{bg}\|_2. \tag{18}$$

Similarly, the image pairs $\boldsymbol{I}$ and $\boldsymbol{I}^{\gamma}$ with different illumination coefficients $\boldsymbol{\gamma}$ should keep the same identity and expression. Hence, the illumination disentanglement is enforced through:

$$\mathcal{L}_{\text{dis}}^{\gamma} = \mathcal{L}_{\text{id}}^{\gamma}(\boldsymbol{I}^{\gamma}, \boldsymbol{I}) + \mathcal{L}_{\text{lmk}}^{\gamma}(\boldsymbol{I}^{\gamma}, \boldsymbol{I}). \tag{19}$$

Disentangled expression control, on the other hand, is the most challenging. We introduce multiple losses to ensure consistency in various aspects. For the contrastive image pair $\boldsymbol{I}$ and $\boldsymbol{I}^{\beta}$ with conditioned on different $\boldsymbol{\beta}$, we first

apply the identity loss $\mathcal{L}_{\text{id}}^{\beta}(\boldsymbol{I}^{\beta}, \boldsymbol{I})$ to ensure identity consistency. To promote the fine-grained appearance similarity, the two images are expected to share the same appearance for corresponding areas. To encourage this, following [14] we compute the 2D flow of rendered 3DMM faces $\boldsymbol{R}$ and $\boldsymbol{R}^{\beta}$, and use it to warp the portrait $\boldsymbol{I}$ which results to $\mathcal{W}(\boldsymbol{I})$, where $\mathcal{W}$ denotes the warping operator. The warped image should be similar to $\boldsymbol{I}^{\beta}$, so the texture loss now becomes:

$$\mathcal{L}_{\text{tex}}^{\beta} = \|\boldsymbol{I}^{\beta} - \mathcal{W}(\boldsymbol{I})\|_2. \tag{20}$$

Note that the flow computed from the 3DMM faces fails to consider the occlusions by hair and face wearings. Hence we zero the flow of non-face regions according to the rendered semantic mask when computing $\mathcal{L}_{\text{tex}}^{\beta}$.

While the above expression disentanglement losses suffice for most cases, we observe severe lip color change when animating the portraits to open-mouth expressions, as shown in Figure 2(a). We conjecture that this arises from the training data bias that faces with open-mouth expressions are rare, and the learned model inevitably shows some appearance tendency. To solve this, we propose a lip style loss that encourages the appearance consistency of the lip region. Specifically, we obtain the lip mask $\boldsymbol{S}_{\text{lip}}$ from the learned semantic field and calculate the style loss [114] with the pretrained VGG model [115]:

$$\begin{aligned}\mathcal{L}_{\text{lip}} = &\sum_{i=1}^{2} ||\mu(\phi_i(\boldsymbol{I}^{\beta}) \odot \boldsymbol{S}_{\text{lip}}^{\beta}) - \mu(\phi_i(\boldsymbol{I}) \odot \boldsymbol{S}_{\text{lip}})||_2 + \\ &\sum_{i=1}^{2} ||\nu(\phi_i(\boldsymbol{I}^{\beta}) \odot \boldsymbol{S}_{\text{lip}}^{\beta}) - \nu(\phi_i(\boldsymbol{I}) \odot \boldsymbol{S}_{\text{lip}})||_2,\end{aligned} \tag{21}$$

where ϕ_i denotes the activation of ith layer whereas $\mu(\cdot)$ and $\nu(\cdot)$ calculate the feature mean and variance respectively. As shown in Figure 2, this style loss effectively preserves the lip appearance during animation. Therefore, overall expression disentangle loss is:

$$\mathcal{L}_{\text{dis}}^{\beta} = \lambda_{\text{id}}^{\beta}\mathcal{L}_{\text{id}}^{\beta} + \lambda_{\text{tex}}^{\beta}\mathcal{L}_{\text{tex}}^{\beta} + \lambda_{\text{lip}}^{\beta}\mathcal{L}_{\text{lip}}^{\beta}, \tag{22}$$

where λ are the weights for different terms.

Once imitative learning converges, we embark on the disentanglement learning. To achieve the disentanglement of different attributes, we additionally enforce the following losses to the generator:

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{dis}}^{\kappa} + \mathcal{L}_{\text{dis}}^{\gamma} + \mathcal{L}_{\text{dis}}^{\beta}. \tag{23}$$

4.3.3 Training Details

In the two-stage adversarial training, we use Adam optimizer [116] with $\beta_1 = 0, \beta_2 = 0.99$, and initialize the learning rate to 0.002 for the generator and 0.0025 for the discriminator, respectively. As for the volumetric rendering, 96 points are sampled for each ray where 48 points for uniform sampling and 48 for importance sampling.

We train and evaluate our method at 512×512 resolution on the FFHQ dataset. We utilize an exponential moving average (EMA) of model weights for inference. The imitative learning is trained for 300,000 iterations with a batch size of 4 and the disentanglement learning takes another 100,000 iterations with a batch size of 2. The loss weights in Equation 17 are set to $\lambda_{\text{tex}} = 10, \lambda_{\text{id}} = 10, \lambda_{\text{lmk}} = 10, \lambda_{\text{ill}} = 1e3$.

Fig. 3. **The error map during the facial animation.** The error map is computed as the per-pixel error between the original pixel color and the edited image. The volume blending significantly reduces the error map in the non-face region, such as hair, shoulder and background, when changing the facial expression.

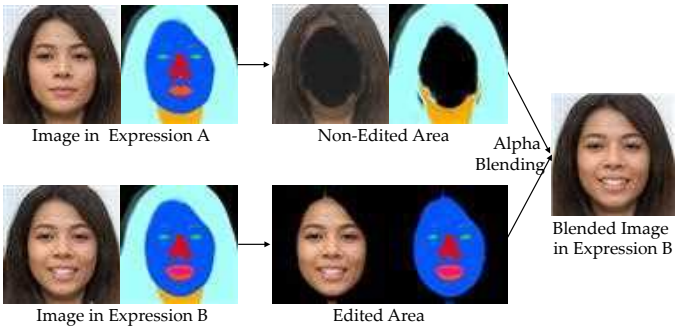


Fig. 4. **The volume blending.** During the expression animation, the appearances of non-face regions such as hair and background possibly change. To amend this, we remain the radiance field of these regions the same as the original neutral inputs.

The weights in Equation 22 are set to $\lambda_{\text{id}}^{\beta} = 100, \lambda_{\text{flow}} = 50, \lambda_{\text{lip}} = 100$. The overall training process takes 14 days on 8 NVIDIA Tesla 32GB V100 GPUs.

4.4 Explicit Volume Blending

So far, we can achieve disentangled control with the above two-stage training. Nonetheless, the disentanglement learning somehow compromises generation diversity. Reducing the strength of disentanglement learning, however, leads to obvious appearance inconsistency. As shown in Figure 3, such inconsistency mainly happens in the non-face areas such as hair, clothes and backgrounds. Motivated by this, we propose a volume blending scheme during inference that allows a smaller strength of disentanglement learning while preserving consistency during the expression animation.

Figure 4 illustrates the volume blending process. Given the latent code $\boldsymbol{z} = (\boldsymbol{\epsilon}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ and $\boldsymbol{z}' = (\boldsymbol{\epsilon}, \boldsymbol{\kappa}, \boldsymbol{\beta}', \boldsymbol{\gamma})$ with animated expression coefficients, they yield two separate radiance fields — $(\boldsymbol{c}, \boldsymbol{\sigma}, \boldsymbol{s})$ and $(\boldsymbol{c}', \boldsymbol{\sigma}', \boldsymbol{s}')$, where $\boldsymbol{c}$, $\boldsymbol{\sigma}$ and $\boldsymbol{s}$ denote the view-dependent color, density and semantics of the field respectively. To remain the consistency during animation, we produce a composite radiance field $(\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{\sigma}})$, which is blended from the controlled output and the original radiance field, *i.e.*,

$$(\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{\sigma}}) = \big(\boldsymbol{w}\boldsymbol{c} + (1-\boldsymbol{w})\boldsymbol{c}', \boldsymbol{w}\boldsymbol{\sigma} + (1-\boldsymbol{w})\boldsymbol{\sigma}'\big), \tag{24}$$

where $\boldsymbol{w}_p$ indicates the probability of belonging to the face part for the location $\boldsymbol{p}$, which is inferred from the learned semantic field $\boldsymbol{s}'$. We opt for volume blending over image-level blending since the former guarantees view consistency and brings fewer blending artifacts. In Figure 3, we show that the temporal appearance variation of the non-edited

area during animation is significantly reduced using the blending strategy. Both the qualitative and quantitative studies show much improved consistency in the disentangled control.

5 Experiments

5.1 Dataset

We adopt 70,000 wild images with 512×512 resolutions in FFHQ [10], [11] as the training dataset. We utilize a face reconstruction method [113] to extract the 3DMM coefficients for each real image, and leverage these extracted coefficients as the training set to train VAE-GANs for control parameters.

Training on FFHQ, however, gives dull expressions. To improve the expression diversity, we additionally leverage the expression coefficients collected from the emotional video dataset RAVDESS [109]. This dataset contains 24 professional actors including 12 females and 12 males who act with exaggerated expressions. In practice, for each actor, we select 10 videos and sample 400 images from each video, resulting in 96,000 images in total. We extract expression coefficients from these images and combine these coefficients together with the expressions extracted from FFHQ to train the expression-related VAE-GAN.

5.2 Image Preprocessing

In order to align the neural radiance space and the 3DMM face, we extract 5-point landmarks using MTCNN [117] for the original face image and meanwhile determine the 3D landmarks of a canonical 3DMM model, which we use for face alignment. Then we reconstruct the 3DMM parameters for the face image. After that, for each 3DMM, we set all the translation coefficients as a pre-defined value so as to place the reconstructed face in a proper image location, which results in a new 3DMM mesh along with corresponding 3D landmarks. We compute the affine transformation that aligns the original five landmarks and the new 3DMM landmarks, and apply the computed affine matrix to produce aligned faces for training. As a result, images are aligned with their corresponding 3DMM reconstruction in the canonical space. The image alignment process is illustrated in Figure 5. We show the importance of using aligned training data in Sec. 5.5.2.

5.3 Qualitative Comparison

5.3.1 Portrait Image Animation

We compare our method with prior controllable portrait works including DiscoFaceGAN [14], GAN-Control [15], HeadNeRF [44] and PIRenderer [43]. The first two methods are the 2D generative methods that also support 3DMM control. HeadNeRF [44] is a NeRF-based parametric head model. PIRenderer [43] is a state-of-the-art face reenactment method that animates the source image according to a driving video. To adapt it to controllable portrait animation, we use neutral faces generated by our method as the source images and use a pretrained PIRenderer model to animate the face into target expressions and poses.

The visual comparison is shown in Figure 6. PIRenderer has difficulties in producing high-quality results and

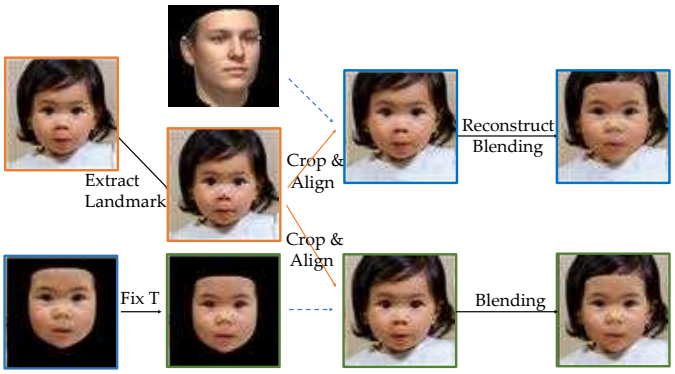


Fig. 5. **The data alignment process.** The Image in orange is the original input, while the image in blue is the corresponding 3D reconstruction. "Fix T" refers to setting the translation to a predefined value. We use the aligned face (with green box) for training. Here, we blend the 3DMM rendering into the aligned image to validate the alignment.

preserving the identity on large poses. DiscoFaceGAN can accurately control the expression, but still fails to maintain the consistency of non-face areas (*e.g.*, glasses, hair and background) when varying expressions and poses. Similarly, GAN-Control can always generate high-quality results but may demonstrates severe inconsistency, especially during the pose control. Compared with these image-based methods, HeadNeRF adopts the 3D neural representation and thus ensures perfect multi-view consistency. Yet the generated images lack fine details, and hence exhibit limited perceptual quality. In contrast, our method produces the most compelling images with consistent appearance when viewed from different angles.

We further present more visual results of our method in terms of varied expressions, illuminations and camera poses in Figure 7, where our method animates portrait images in a disentangled manner. We also change the poses and expressions simultaneously, and our method shows impressive control capability. The non-face areas (*e.g.*, glasses, hairs, background) that should not be affected by the face control parameters are temporally consistent during the semantic control.

5.3.2 Multi-view Consistency and Temporal Consistency

Following [93], we visualize the consistency when varying expressions and poses using different methods. Figure 8 shows the results when modifying the face poses or camera positions with other face attributes fixed. We check the texture of straight lines on the teeth and hair area. Figure 8 shows that DiscoFaceGAN, PIRenderer and GAN-Control suffer from inconsistency when smoothly changing the camera location, while HeadNeRF leads to many blurry results. In comparison, the proposed approach produces 3D-consistent images of high fidelity.

We also vary expression continuously and the results are shown in Figure 9. For each frame, we illustrate the texture along the depicted lines to show the temporal consistency. GAN-Control clearly produces noise and distortion patterns. DiscoFaceGAN and our method produce more consistent results but still suffer tiny distortion. The volume blending, in comparison, further improves the temporal consistency.

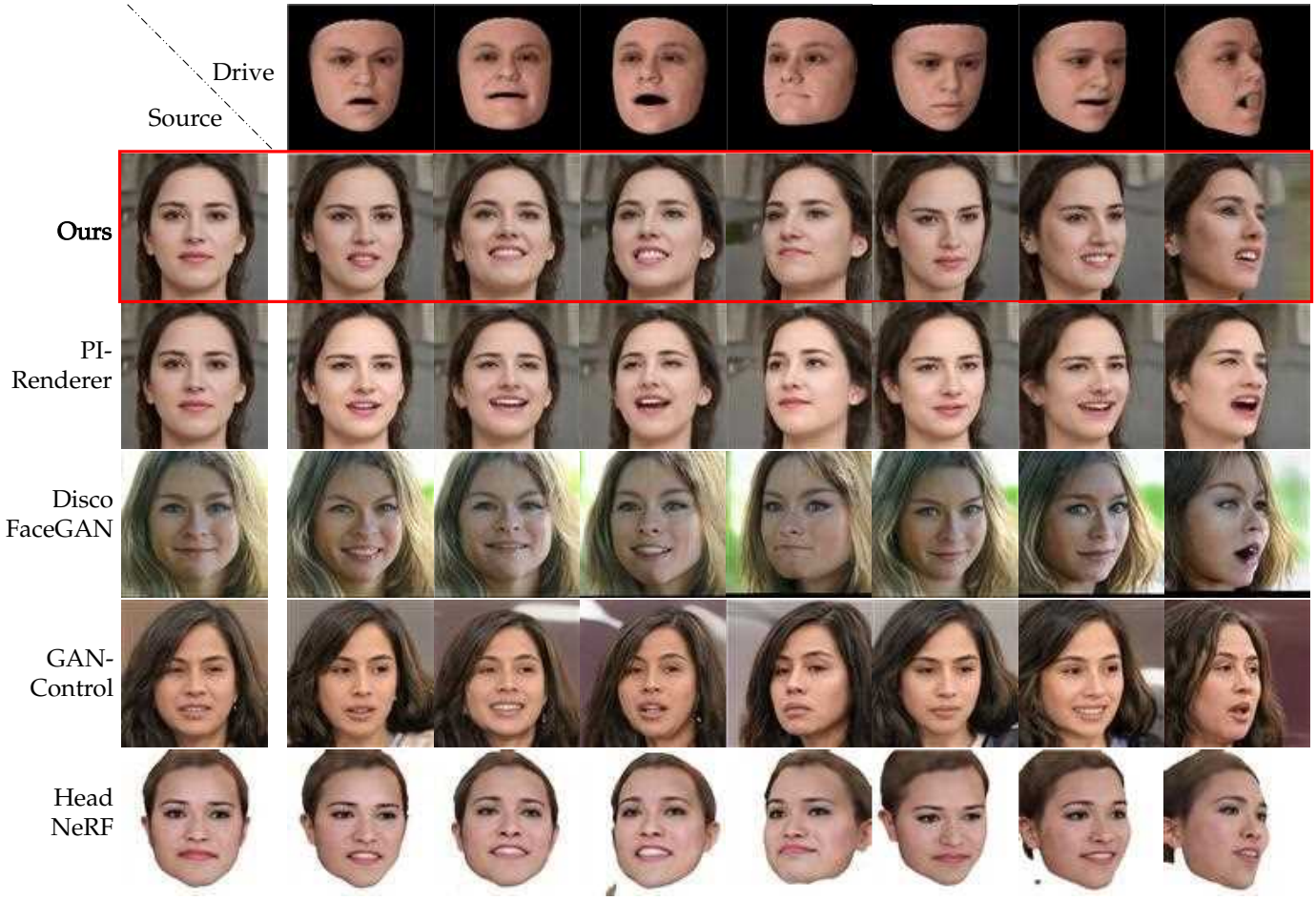


Fig. 6. **Visual comparison with PIRenderer [43], DiscoFaceGAN [14], GAN-Control [15] and HeadNeRF [15].** Given a source image and control parameters, we use different approaches to generate faces with animated expressions in different head poses. For the face reenactment method PIRenderer, its input source image is produced by our method.

5.4 Quantitative Comparison with Prior Methods

5.4.1 *Image Quality*

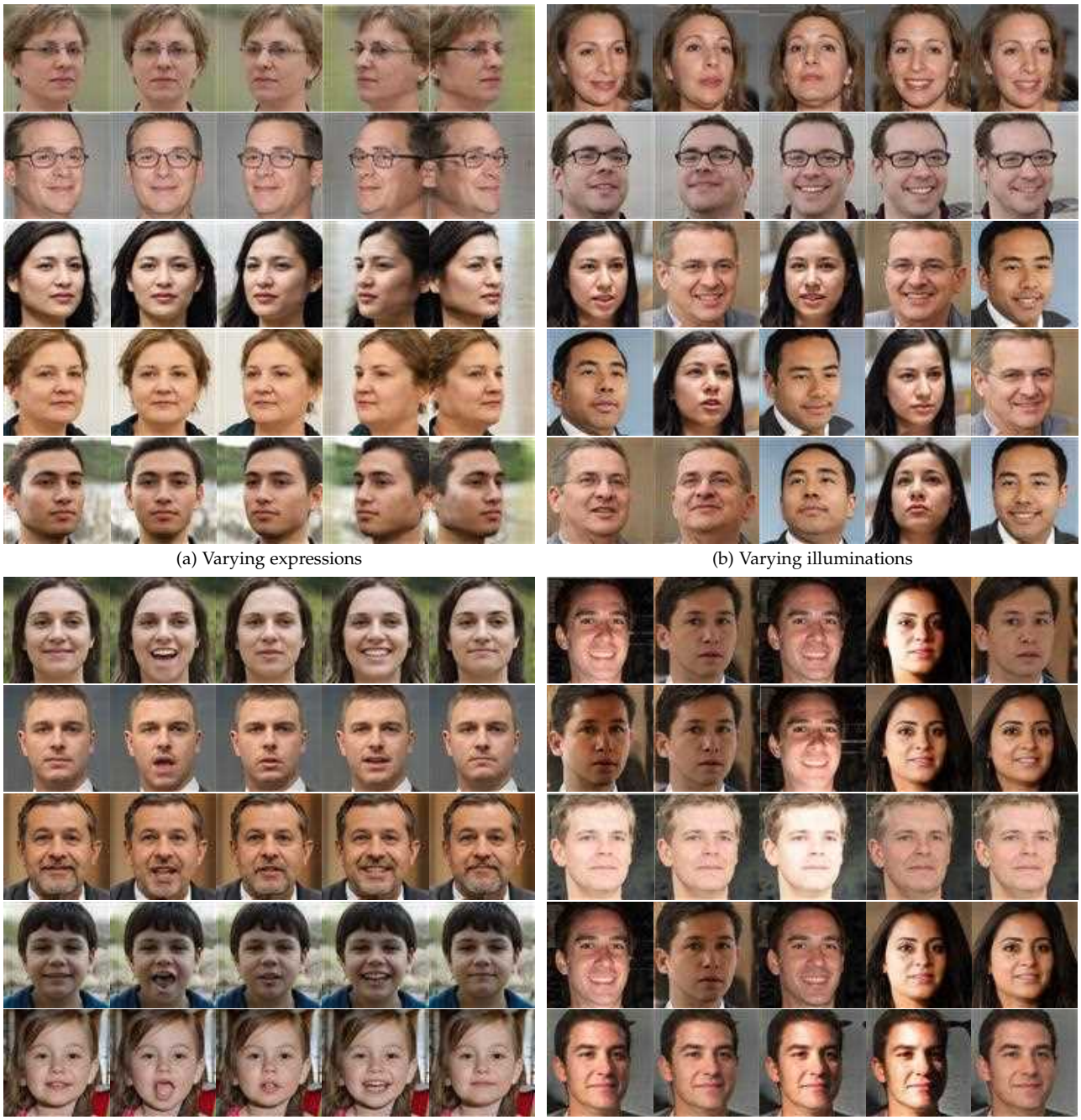
We measure the image quality with Frechet Inception Distance (FID) [118] using the ImageNet-pretrained Inception-V3 model [119] and the vision-language pretrained CLIP model [120], and denote the latter score as "FID-clip". We compute the FID between the 5,000 real images and 5,000 generated images. Since StyleRig [12] and PIE [13] do not release codes, we use images containing 168 identities with diverse expressions showcased on the project website[1] for comparison. When comparing DiscoFaceGAN, we run its model pretrained on FFHQ and generate 5,000 images with random poses, expressions and illuminations. Since this work is trained on 256×256 resolution, we utilize a state-of-the-art super-resolution method, SwinIR [121], to upsample their results to 512×512 resolution for a fair comparison. PIRenderer uses 5,000 frontal face images produced by our method as inputs, and performs the face reenactment using randomly sampled expressions and poses. We finetune the released model of HeadNeRF and generate 5,000 images with randomly sampled 3DMM coefficients.

We align images using the same preprocessing procedure as Sec. 5.2 for all the compared methods. As shown in Table 1, our method significantly outperforms StyleRig, PIE, DiscofaceGAN, PIRenderer and HeadNeRF in terms of both FID and FID-clip. While GAN-Control achieves a better FID score, its disentanglement ability is much worse than our work as shown in the qualitative comparison and the following quantitative evaluations.

1. https://vcai.mpi-inf.mpg.de/projects/PIE/

TABLE 1
Quantitative comparison of image quality. The star symbol (*) denotes we only use available data for evaluation. "UP" denotes the bicubic upsampling. "SR" denotes the super-resolution results using [121]. We highlight the best score and underline the second best.

	FID (512)↓	FID-clip (512)↓
StyleRig* [12]	60.3	32.2
PIE* [13]	60.9	24.4
DiscoFaceGAN-UP [14]	56.6	18.6
DiscoFaceGAN-SR [14]	39.1	17.3
PIRenderer-UP [43]	77.4	23.8
PIRenderer-SR [43]	68.9	24.0
GAN-Control [15]	**13.5**	**8.5**
HeadNeRF [44]	142.6	64.7
Ours	<u>24.1</u>	<u>14.7</u>



(a) Varying expressions (b) Varying illuminations

(c) Varying poses (d) Varying expressions and poses

Fig. 7. **The controllable generation results.** We vary expressions, illuminations and poses, independently, which are shown in (a), (b), and (c). We also vary both expressions and poses as shown in (d). The proposed approach can precisely control expression, illumination and pose while preserving the identity and other attributes.

5.4.2 Control Accuracy

We evaluate the control accuracy of all methods using Average Expression Distance (AED), Average Pose Distance (APD) and Average Illumination Distance (AID), which have been used in [43]. For StyleRig and PIE, we also use the animation results from their project webpages. For HeadNeRF, DiscoFaceGAN, GAN-Control and our method, we randomly generate 1,000 identities with neutral expression and randomly apply 10 expressions, poses and illuminations. For PIRenderer, we use the same source images and driving coefficients as our method. Our method gets 10,000 generated images, for which we use the face reconstruction network to reconstruct 3DMM coefficients. Finally, we calculate the average distance between the input control parameters and the reconstructed coefficients. Table 2 shows that our method achieves the best AED, APD and AID scores.

5.4.3 Disentanglement

We evaluate the disentanglement using the Disentanglement Score (DS), which is first proposed in [14]. We denote the DS for expression, pose and illumination as DS_β, DS_r, DS_γ respectively. Specifically, we randomly sample images

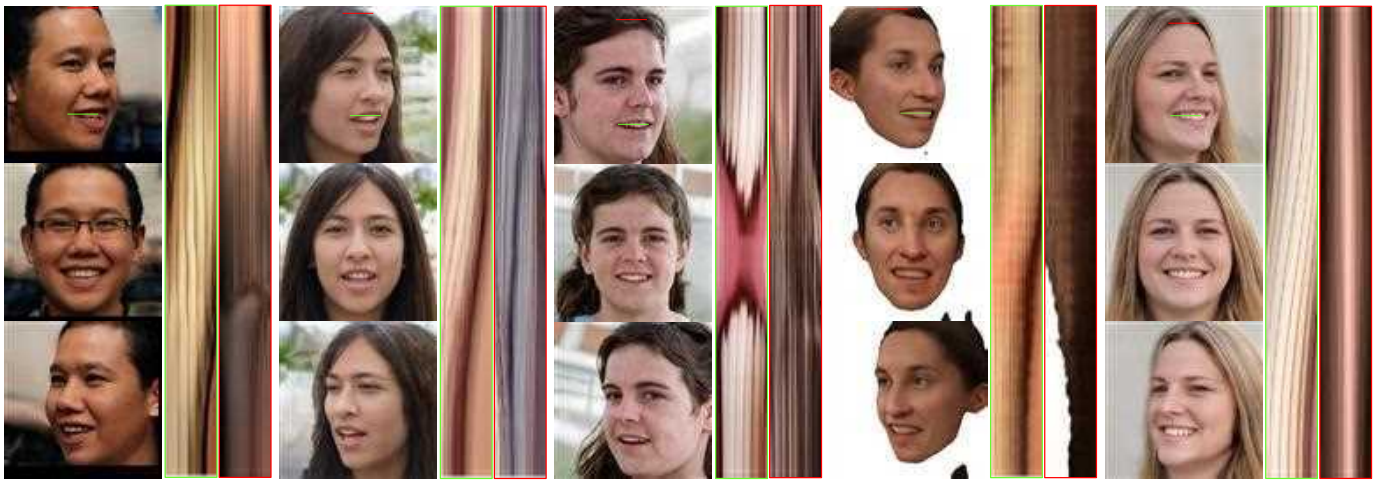


Fig. 8. **Visualization of multi-view consistency on teeth (green) and hair (red) area.** We visualize the texture along a straight line when changing poses. The view-consistent results should demonstrate smooth pattern in the visualization.

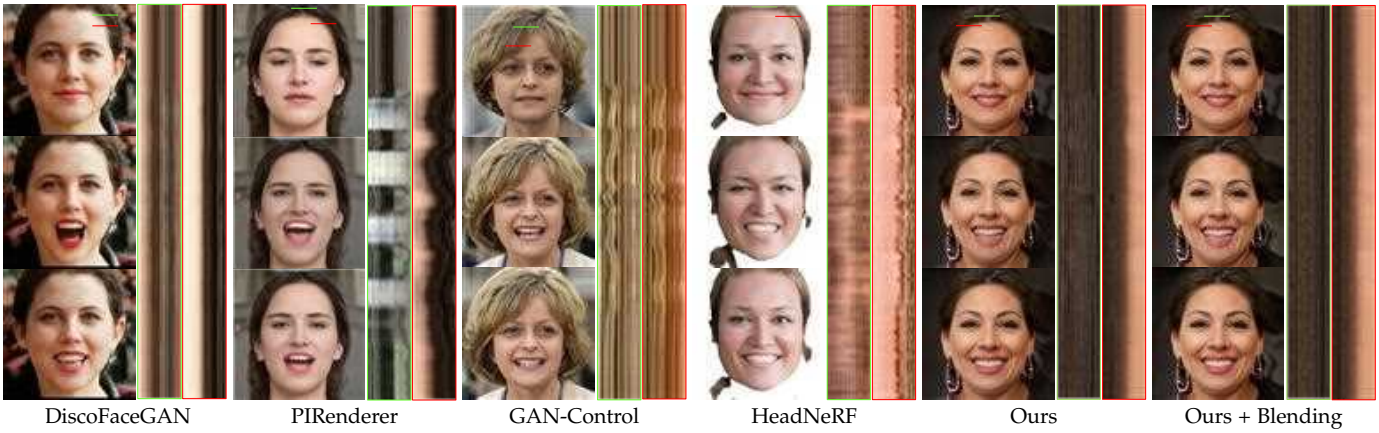


Fig. 9. **Visualization of temporal consistency on hair (green) and hairline (red) area during continuous expression animation.** We visualize the texture along a straight line during animation. The temporally consistent results should show smooth pattern in the visualization.

TABLE 2
Quantitative comparison of the control accuracy. For PIRenderer we only compute AED and APD, since it cannot control illumination. We highlight the best score and underline the second best.

	AED↓	APD↓	AID↓
DiscoFaceGAN [14]	<u>0.2207</u>	0.0016	0.0295
PIRenderer [43]	0.2360	0.0030	-
GAN-Control [15]	0.2221	0.0022	0.0217
HeadNeRF [44]	0.2732	<u>0.0015</u>	<u>0.0129</u>
Ours	**0.2026**	**0.0010**	**0.0018**

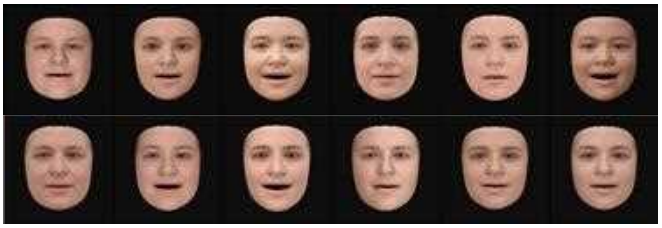
TABLE 3
Quantitative comparison of disentanglement. We highlight the best score and underline the second best. DS denotes the disentanglement score and IS denotes the identity similarity score.

	$DS_\beta \uparrow$	$DS_r \uparrow$	$DS_\gamma \uparrow$	IS ↑
StyleRig* [12]	24.3	18.4	3.0	0.66
PIE* [13]	27.4	31.3	2.0	0.74
DiscoFaceGAN [14]	32.5	77.3	18.2	0.60
PIRenderer [43]	7.83	65.9	-	0.27
GAN-Control [15]	<u>37.6</u>	67.5	<u>25.3</u>	0.72
HeadNeRF [44]	37.0	<u>82.6</u>	3.68	<u>0.80</u>
Ours	**46.1**	**117.5**	**33.2**	**0.87**

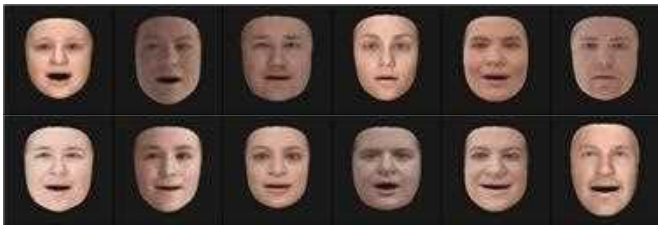
with a single modified attribute. Then, we re-estimate the 3DMM parameters from the generated images and calculate the variance of the estimated coefficients $(\boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{\gamma})$. To discount the influence of coefficient magnitude, we normalize the score according to the coefficient variance computed from the FFHQ dataset. Thus, DS_i is calculated as:

$$\mathrm{DS}_i = \prod_{\forall j \neq i} \frac{\sigma_i}{\sigma_j}, \sigma_i = \mathrm{var}(i), i, j \in \{\boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{\gamma}\}. \quad (25)$$

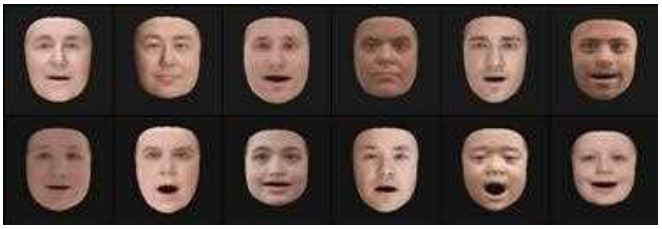
A higher DS indicates that other attributes will remain the same when editing one specified attribute. Besides, we also measure identity similarity during portrait editing by calculating the cosine similarity of the face embedding using a pretrained face recognition [122]. We randomly generate 1,000 source images and randomly apply two poses, expressions or illuminations to the model. Numerical results in Table 3 show that our method achieves the best disentanglement ability in terms of all the evaluation metrics.



(a) VAE on FFHQ



(b) VAE-GAN on FFHQ



(c) VAE-GAN on FFHQ & RAVDESS

Fig. 10. **Ablation study of learning VAE-GANs.** (a) VAE tends to give mean expressions. (b) VAE-GAN improves expression diversity. (c) The supplement with RAVDESS training images further helps to sample more vivid expressions.

5.5 Ablation Study

In this section, we perform extensive ablation studies of the proposed technical components to investigate their effects on the final output.

5.5.1 The Effect of VAE-GANs and Auxiliary Expression Data

We compare the vanilla VAE model trained on FFHQ, the VAE-GAN trained on FFHQ and the VAE-GAN model trained on combined FFHQ and RAVDESS datasets. Figure 10 shows that the vanilla VAE tends to give averaged faces with dull expressions. Adding adversarial training (VAE-GAN) improves the expression diversity, yet the expression diversity is still limited due to the data distribution of FFHQ. To address this issue, we add expression coefficients from RAVDESS, which helps to sample diverse and vivid expressions. Thus we can sample expressions like angry, surprised or fearful faces with delicate eyebrow movements.

5.5.2 The Effect of Face Alignment

We further show the necessity of data preprocessing as mentioned in Sec. 5.2. The experimental results shown in Table 4 and Figure 11 prove that the alignment in data preprocesses significantly improves the performance of the 3DMM guidance. Without this preprocessing, the generated neural radiance fields fail to align with the 3DMM guidance, leading to obvious artifacts on the eyebrows.

TABLE 4
The influence of proposed components to image quality and controllability.

	FID↓	FID-clip↓	AED↓	APD↓	AID↓
Ours	24.1	<u>14.7</u>	**0.2026**	**0.0010**	**0.0018**
w/o SF	**19.0**	**10.8**	0.2203	<u>0.0013</u>	0.0035
w/o align	38.3	37.2	0.2798	0.0049	0.0027
w/o $\mathcal{L}_{id}$	28.0	21.4	0.2146	0.0036	0.0029
w/o $\mathcal{L}_{tex}$	24.2	17.4	<u>0.2102</u>	0.0058	0.0021
w/o $\mathcal{L}_{lm}$	41.6	40.0	0.2730	0.0211	0.0029
$-\mathcal{L}_{ce}$ $+\mathcal{D}_{sf}$	<u>21.0</u>	16.3	0.2179	0.0026	<u>0.0019</u>

TABLE 5
The influence of proposed components to disentanglement.

	DS_e ↑	DS_p ↑	DS_{il} ↑	IS↑
Ours	**46.1**	**117.5**	**33.2**	**0.87**
w/o SF	42.4	101.7	13.2	<u>0.85</u>
w/o $\mathcal{L}_{\text{dis}}$	21.6	105.2	11.3	0.54
w/o $\mathcal{L}^{\kappa}_{\text{dis}}$	44.9	110.3	25.8	0.82
w/o $\mathcal{L}^{\beta}_{\text{dis}}$	30.5	108.5	<u>29.8</u>	0.68
w/o $\mathcal{L}^{\gamma}_{\text{dis}}$	39.4	<u>111.3</u>	15.6	0.80
w/o blend	<u>45.3</u>	**117.5**	**33.2**	<u>0.85</u>

5.5.3 The Effect of Loss Functions

We conduct ablation studies to validate the benefits of different loss functions used in imitation learning and disentanglement learning, respectively. Table 4 and Table 5 show the quantitative ablation studies of different loss functions. Figure 11 illustrates how they affect image quality and controllability. These results show that without the identity loss, the identity similarity between 3DMM renderings and the generated images drops and image quality also deteriorates. Removing the landmark loss greatly impairs the generation quality and the control accuracy. The ablation of losses in the disentanglement learning, *i.e.*, $\mathcal{L}^{\kappa}_{\text{dis}}$, $\mathcal{L}^{\beta}_{\text{dis}}$, and $\mathcal{L}^{\gamma}_{\text{dis}}$, degrades the control disentanglement and causes inconsistency for other facial attributes.

5.5.4 The Benefit of Learned Semantic Field

While involving semantic learning may introduce additional burden to the generator and slightly worsens the image quality (see results in Table 2), it significantly improves the control disentanglement and slightly benefits the control accuracy, as shown in Table 2 and Table 3. This is because we can accurately parse the 3D portrait and let the generator focus on the face part for disentangled editing. Besides, the explicit volume blending also leverages the learned semantic field, and makes the non-face regions perfectly consistent.

Moreover, we compare the way to train a good semantic field. Rather than relying on the online parsing result as the ground truth, we may introduce a semantic discriminator $\mathcal{D}_{sf}$ that examines the realism of rendered parsing result, which we denote as "- $\mathcal{L}_{ce}$ + $\mathcal{D}_{sf}$" Figure 11 shows that the discriminator cannot reliably provide the supervision for the semantic filed learning.

We also visualize more rendered results for the radiance field and the semantic field in Figure 12. Both the image

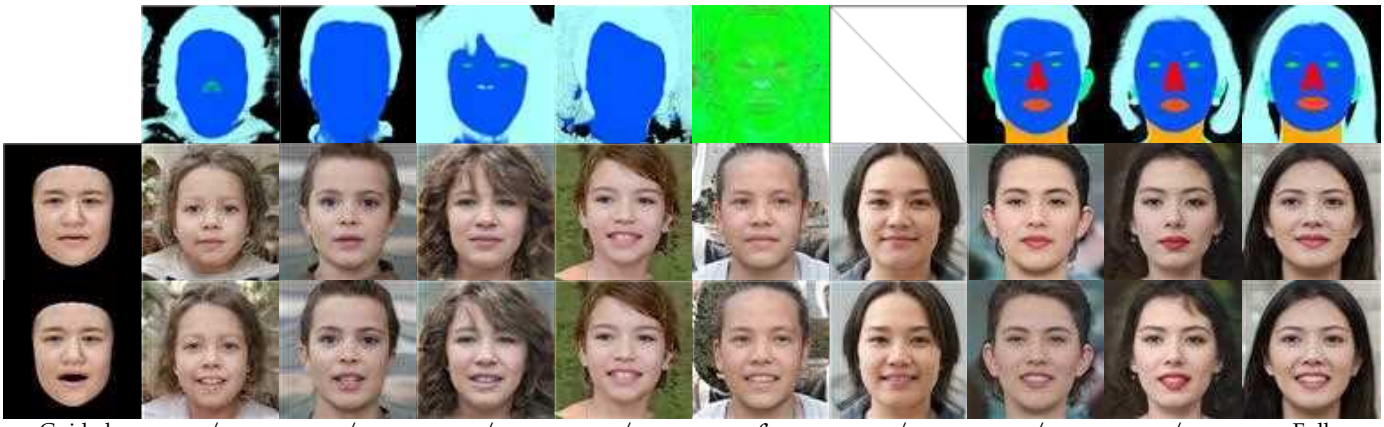


Fig. 11. **Ablation studies of loss functions.** Given two guided 3DMM coefficients, we visualize the generated image as well as face parsing results from the models in different settings.

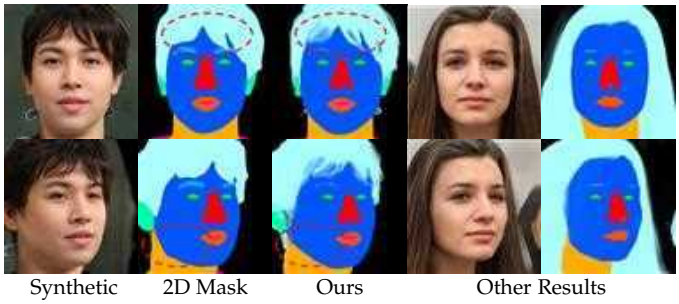


Fig. 12. **Visualization of 2D semantic mask and our rendered semantic mask.** Compared with a 2D semantic mask, the rendered semantic mask from the learned semantic field can be even more accurate than the 2D parsing results while being more view consistent.

rendering and the face parsing result are view consistent. It is interesting to see that the learned semantic field can even more accurately parse the face than the 2D ground truth.

6 Applications

In this section, we showcase several applications of our method. Our method could enable video-driven talking head generation with the editable background. We can also achieve real portrait image editing by projecting the real image into the latent space of the network. We also investigate the out-of-domain generalization ability of our method on two stylized portrait datasets.

6.1 Talking Face Generation

Specifically, for an input talking face video, we can capture the expression coefficients and head pose information from a specific frame using a face reconstruction network [113]. Then we can explicitly feed each expression coefficient to our pipeline and render the image with the corresponding head pose frame by frame. Given a driving video, we can randomly generate a talking face video with the same expressions and head movements for different identities.

We visualize the video-driven talking face results in Figure 13. Our method can capture expressions and mouth movements. We also show examples of background replacement results in Figure 14. To achieve this, we make all background volume transparent, and save the alpha value for each pixel to generate an alpha mask. We achieve background replacement by blending the opaque background image with a semi-transparent rendered image using the alpha mask. Thanks to the learned semantic field, we can achieve delicate portrait matting and high-quality background replacement effects.

6.2 Real Portrait Image Editing

As illustrated in Figure 15, our method allows precise editing of input portrait images in various expression, illumination, and pose while preserving identity. Inspired by the disentanglement capability of StyleGAN's latent space [10], we project the input portrait image into $\boldsymbol{Z}+$ space of the pretrained model and explicitly manipulate the latent codes for real image editing.

To obtain the corresponding latent codes, we use the off-the-shelf inversion technique introduced in StyleGAN [10]. Specifically, the latent code $\boldsymbol{z} = (\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\epsilon})$ in semantic control space is initialized with the predicted identity $\boldsymbol{\kappa}$, expression $\boldsymbol{\beta}$, illumination $\boldsymbol{\gamma}$ coefficients from a pretrained face reconstruction network [113] and a random vector $\boldsymbol{\epsilon}$. We keep the parameters of the generator fixed and directly optimize the latent codes by measuring the similarity between generated images and real images with the LPIPS [123] loss, a pixel-wise $\mathcal{L}_2$ loss and the ID loss [122].

Moreover, in StyleGAN, we can map the latent code $\boldsymbol{z}$ on each iteration to $\mathcal{W}$ space, which is an intermediate latent space after the fully connected mapping, and get the latent code $\mathbf{w}$. To improve the image quality and photo-realism, we introduce the regularization loss $\mathcal{L}_{\text{reg}}$ on $\mathcal{W}$ space as:

$$\mathcal{L}_{\text{reg}}(\mathbf{w}, \overline{\mathbf{w}}) = ||\mathbf{w} - \overline{\mathbf{w}}||_2, \tag{26}$$

where $\overline{\mathbf{w}}$ is the average latent code in $\mathcal{W}$ space of the pretrained model. The overall training objective function can be written as:

$$\mathbf{z}^* = \arg\min_{\boldsymbol{z}} (\lambda_1 \mathcal{L}_{\text{LPIPS}} + \lambda_2 \mathcal{L}_2 + \lambda_3 \mathcal{L}_{\text{id}} + \lambda_4 \mathcal{L}_{\text{reg}}), \tag{27}$$

where $\lambda_{(\cdot)}$ are the weights for different terms.

Fig. 13. **Example results produced by our method on video-driven talking face generation**. Our method can generate photo-realistic portrait images (bottom) according to the drive motion (top).

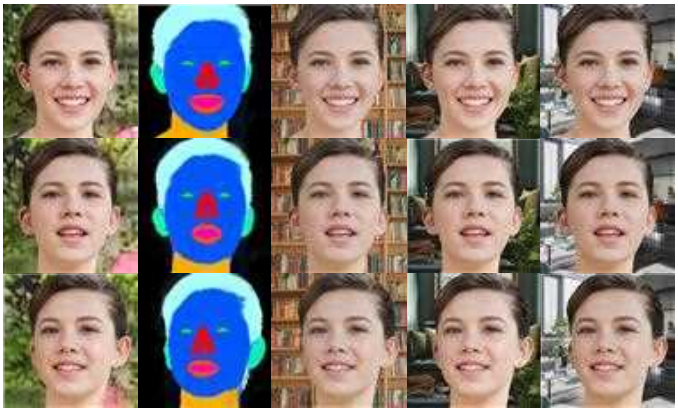


Fig. 14. **Visualization of varying different background images.** From left to right, we show the generated portrait image using our method, rendered semantic mask, and images with diverse background images.

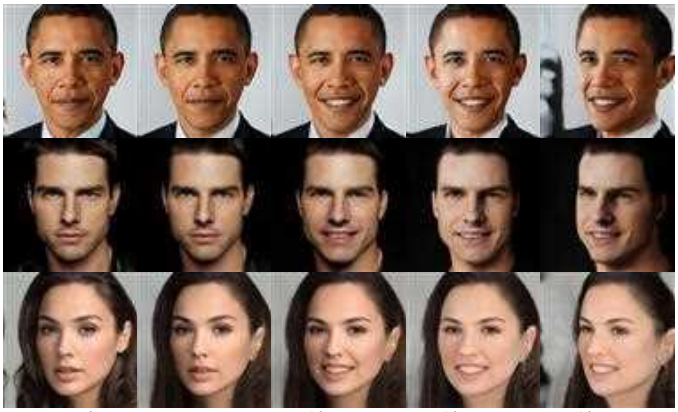


Fig. 15. **Visualization of real portrait image inversion and editing results.** Given a real image, we project the real image into the latent space and achieve portrait image editing sequentially.

To improve the reconstruction quality, we slightly alter the parameters of generator $\mathcal{G}_\theta$ to $\mathcal{G}_\theta^*$ while keeping the optimized latent codes $\mathbf{z}^*$ fixed. The inversion process takes only 5min for one image on a single NVIDIA Tesla V100 GPU. After obtaining the latent codes $\mathbf{z}^*$ and the corresponding generator $\mathcal{G}_\theta^*$, the expressions and illuminations can be edited by manipulating their corresponding latent codes, and the pose can be changed by controlling the render pose while keeping the identity. Results show that our trained disentangled semantic control space has generalization ability.

6.3 Out-of-domain Image Editing

In addition to realistic faces, our method also supports out-of-domain image editing, which is shown in Figure 16. We make use of two stylized portrait datasets: Metface [124], including 1,336 images from the collection of the Metropolitan Museum of Art, and Disney cartoon face, involving 400 online images of Disney cartoon characters collected by [125]. We preprocess the training data using the same preprocess method introduced in Sec. 5.2. We freeze the first several layers of the discriminator as in Freeze-D [126], and fine-tune the network with adversarial loss. The results show that our method has the out-of-domain generalization ability.

7 Limitations

Our method focuses on building a generative and controllable 3D-aware neural radiance field that can be rendered to a high-quality portrait image. However, there might be some limitations. First, our method leverages the 3DMM face mesh as guidance for manipulating portrait images, but 3DMM tends to represent smooth textures and limited human identities. In order to generate diverse results, the identity of the generated image might not be the same as the guided face. Second, although our method can handle background replacement and can split background and foreground, we cannot disjoint the control for head and body, which is still a challenge in this area. Besides, when generating a talking face, our method tends to represent smiling and laughing expressions due to the data bias of the FFHQ dataset, which can possibly be addressed in the future by combining talking face data for training.

8 Conclusion

In this work, we propose a 3D-aware portrait generation network that produces 3D consistent portraits while being controllable according to semantic parameters regarding pose, identity, expression and illumination. We can explicitly control the generated neural scene representation using a parametric face model and achieve latent disentanglement. In order to enforce consistency in non-face areas, *e.g.*, hair

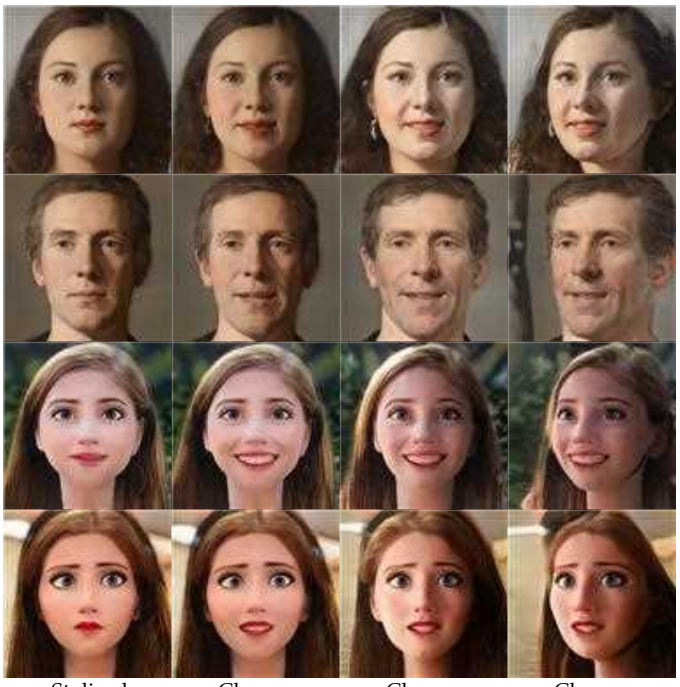


Fig. 16. **Visualization of stylized portrait image editing results.** Results show we can implement progressive editing by changing one specific attribute sequentially while keeping other attributes and identities unchanged.

and background, when animating expressions, we simultaneously train a semantic radiance field to separate dynamic and static areas. We propose a blending strategy in which we form a composite output by blending the dynamic and static radiance fields, with two parts segmented from the jointly learned semantic field. Experimental results show that our method outperforms prior controllable arts. We also investigate multiple applications to demonstrate the generalization ability to real images as well as out-of-domain cartoon faces. The proposed approach opens doors for various extended reality applications that demand 3D consistent avatars with explicit control.

REFERENCES

[1] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel, "Free-viewpoint video of human actors," *ACM transactions on graphics (TOG)*, vol. 22, no. 3, pp. 569–577, 2003. 1

[2] D. Casas, M. Volino, J. Collomosse, and A. Hilton, "4d video textures for interactive character appearance," in *Computer Graphics Forum*, vol. 33, no. 2. Wiley Online Library, 2014, pp. 371–380. 1

[3] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017. 1, 3

[4] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "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*, 2019, pp. 7763–7772. 1, 3

[5] A. E. Ichim, S. Bouaziz, and M. Pauly, "Dynamic 3d avatar creation from hand-held video input," *ACM Transactions on Graphics (ToG)*, vol. 34, no. 4, pp. 1–14, 2015. 1

[6] A. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, W. Yifan, C. Lassner, V. Sitzmann, R. Martin-Brualla, S. Lombardi *et al.*, "Advances in neural rendering," in *Computer Graphics Forum*, vol. 41, no. 2. Wiley Online Library, 2022, pp. 703–735. 1

[7] H. Kato, D. Beker, M. Morariu, T. Ando, T. Matsuoka, W. Kehl, and A. Gaidon, "Differentiable rendering: A survey," *arXiv preprint arXiv:2006.12057*, 2020. 1

[8] M.-Y. Liu, X. Huang, J. Yu, T.-C. Wang, and A. Mallya, "Generative adversarial networks for image and video synthesis: Algorithms and applications," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 839–862, 2021. 1

[9] A. Aggarwal, M. Mittal, and G. Battineni, "Generative adversarial network: An overview of theory and applications," *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100004, 2021. 1

[10] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410. 1, 2, 5, 8, 13

[11] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. CVPR*, 2020. 1, 2, 5, 8

[12] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Pérez, M. Zöllhofer, and C. Theobalt, "Stylerig: Rigging stylegan for 3d control over portrait images, cvpr 2020," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, june 2020. 1, 2, 3, 9, 11

[13] A. Tewari, M. Elgharib, M. BR, F. Bernard, H.-P. Seidel, P. Pérez, M. Zöllhofer, and C. Theobalt, "Pie: Portrait image embedding for semantic control," vol. 39, no. 6, December 2020. 1, 2, 3, 9, 11

[14] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong, "Disentangled and controllable face image generation via 3d imitative-contrastive learning," in *IEEE Computer Vision and Pattern Recognition*, 2020. 1, 2, 3, 7, 8, 9, 10, 11

[15] A. Shoshan, N. Bhonker, I. Kviatkovsky, and G. Medioni, "Gan-control: Explicitly controllable gans," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 14 083–14 093. 1, 2, 8, 9, 11

[16] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2085–2094. 1

[17] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable gan controls," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9841–9850, 2020. 1

[18] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in gans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1532–1540. 1

[19] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang, "Gan inversion: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[20] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for stylegan image manipulation," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–14, 2021. 1

[21] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2287–2296. 1

[22] X. Pan, B. Dai, Z. Liu, C. C. Loy, and P. Luo, "Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans," *arXiv preprint arXiv:2011.00844*, 2020. 1

[23] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Advances in Neural Information Processing Systems*, vol. 32, 2019. 1

[24] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194. 1, 2, 3, 4

[25] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani *et al.*, "3d morphable face models–past, present, and future," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 5, pp. 1–38, 2020. 1, 2, 3

[26] R. Parent, *Computer animation: algorithms and techniques*. Newnes, 2012. 1

[27] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European conference on computer vision*. Springer, 2020, pp. 405–421. 1, 3

[28] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for

shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174. 1, 3

[29] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4460–4470. 1, 3

[30] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3504–3515. 1

[31] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, "Neural fields in visual computing and beyond," in *Computer Graphics Forum*, vol. 41, no. 2. Wiley Online Library, 2022, pp. 641–676. 1

[32] N. Max, "Optical models for direct volume rendering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 1, no. 2, pp. 99–108, 1995. 1

[33] E. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis," in *Proc. CVPR*, 2021. 1, 3, 5

[34] J. Gu, L. Liu, P. Wang, and C. Theobalt, "Stylenerf: A style-based 3d aware generator for high-resolution image synthesis," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=iUuzzTMUw9K 1, 3

[35] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. D. Mello, O. Gallo, L. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, "Efficient geometry-aware 3D generative adversarial networks," in *CVPR*, 2022. 1, 2, 3, 5

[36] M. Niemeyer and A. Geiger, "Giraffe: Representing scenes as compositional generative neural feature fields," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3

[37] P. Zhou, L. Xie, B. Ni, and Q. Tian, "Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis," *arXiv preprint arXiv:2110.09788*, 2021. 1

[38] R. Or-El, X. Luo, M. Shan, E. Shechtman, J. J. Park, and I. Kemelmacher-Shlizerman, "Stylesdf: High-resolution 3d-consistent image and geometry generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 503–13 513. 1

[39] X. Zhao, F. Ma, D. Güera, Z. Ren, A. G. Schwing, and A. Colburn, "Generative multiplane images: Making a 2d gan 3d-aware," *arXiv preprint arXiv:2207.10642*, 2022. 1

[40] D. Rebain, M. Matthews, K. M. Yi, D. Lagun, and A. Tagliasacchi, "Lolnerf: Learn from one look," pp. 1558–1567, 2022. 1

[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014. 2

[42] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger, "Graf: Generative radiance fields for 3d-aware image synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3

[43] Y. Ren, G. Li, Y. Chen, T. H. Li, and S. Liu, "Pirenderer: Controllable portrait image generation via semantic neural rendering," 2021. 2, 8, 9, 10, 11

[44] Y. Hong, B. Peng, H. Xiao, L. Liu, and J. Zhang, "Headnerf: A real-time nerf-based parametric head model," 2022. 2, 3, 8, 9, 11

[45] A. Nickabadi, M. S. Fard, N. M. Farid, and N. Mohammadbagheri, "A comprehensive survey on semantic facial attribute editing using generative adversarial networks," *arXiv preprint arXiv:2205.10587*, 2022. 2

[46] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt, "State of the art on monocular 3d face reconstruction, tracking, and applications," in *Computer graphics forum*, vol. 37, no. 2. Wiley Online Library, 2018, pp. 523–550. 2

[47] T. Zhang, L. Deng, L. Zhang, and X. Dang, "Deep learning in face synthesis: A survey on deepfakes," in *2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology (CCET)*. IEEE, 2020, pp. 67–70. 2

[48] Z. Lu, Z. Li, J. Cao, R. He, and Z. Sun, "Recent progress of face image synthesis," in *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2017, pp. 7–12. 2

[49] Y. Lu, Y.-W. Tai, and C.-K. Tang, "Attribute-guided face generation using conditional cyclegan," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 282–297. 2

[50] X. Di and V. M. Patel, "Face synthesis from visual attributes via sketch using conditional vaes and gans," *arXiv preprint arXiv:1801.00077*, 2017. 2

[51] Y. Wang, A. Dantcheva, and F. Bremond, "From attribute-labels to faces: face generation using a conditional generative adversarial network," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0. 2

[52] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 818–833. 2

[53] J. Tang, Z. Shao, and L. Ma, "Fine-grained expression manipulation via structured latent space," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6. 2

[54] ——, "Eggan: Learning latent space for fine-grained expression manipulation," *IEEE MultiMedia*, vol. 28, no. 3, pp. 42–51, 2021. 2

[55] T. Wang, Y. Zhang, Y. Fan, J. Wang, and Q. Chen, "High-fidelity gan inversion for image attribute editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[56] B. Zhang, S. Gu, B. Zhang, J. Bao, D. Chen, F. Wen, Y. Wang, and B. Guo, "Styleswin: Transformer-based gan for high-resolution image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 304–11 314. 2

[57] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *international conference on machine learning*. PMLR, 2019, pp. 4114–4124. 2

[58] B. Dai and D. Wipf, "Diagnosing and enhancing vae models," *arXiv preprint arXiv:1903.05789*, 2019. 2

[59] J. Piao, K. Sun, Q. Wang, K.-Y. Lin, and H. Li, "Inverting generative adversarial renderer for face reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 619–15 628. 2

[60] Z. Geng, C. Cao, and S. Tulyakov, "3d guided fine-grained face manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9821–9830. 2

[61] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y.-L. Yang, "Hologan: Unsupervised learning of 3d representations from natural images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7588–7597. 2, 3

[62] J. Piao, C. Qian, and H. Li, "Semi-supervised monocular 3d face reconstruction with end-to-end shape-preserved domain transfer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9398–9407. 2

[63] S. Xu, J. Yang, D. Chen, F. Wen, Y. Deng, Y. Jia, and X. Tong, "Deep 3d portrait from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7710–7720. 2

[64] J. Sun, X. Wang, Y. Zhang, X. Li, Q. Zhang, Y. Liu, and J. Wang, "Fenerf: Face editing in neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7672–7682. 3

[65] J. Sun, X. Wang, Y. Shi, L. Wang, J. Wang, and Y. Liu, "Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis," *arXiv preprint arXiv:2205.15517*, 2022. 3

[66] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner, "Dynamic neural radiance fields for monocular 4d facial avatar reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8649–8658. 3

[67] S. Athar, Z. Xu, K. Sunkavalli, E. Shechtman, and Z. Shu, "Rignerf: Fully controllable neural 3d portraits," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 364–20 373. 3

[68] K. Sun, S. Wu, Z. Huang, N. Zhang, Q. Wang, and H. Li, "Controllable 3d face synthesis with conditional generative occupancy fields," *arXiv preprint arXiv:2206.08361*, 2022. 3

[69] A. W. Bergman, P. Kellnhofer, Y. Wang, E. R. Chan, D. B. Lindell, and G. Wetzstein, "Generative neural articulated radiance fields," *arXiv preprint arXiv:2206.14314*, 2022. 3

[70] V. Blanz, C. Basso, T. Poggio, and T. Vetter, "Reanimating faces in images and video," in *Computer graphics forum*, vol. 22, no. 3. Wiley Online Library, 2003, pp. 641–650. 3

[71] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou, "3d face morphable models" in-the-wild"," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 48–57. 3

[72] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, 2017. [Online]. Available: https://doi.org/10.1145/3130800.3130813 3

[73] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3d morphable models," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 233–254, 2018. 3

[74] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2013. 3, 4

[75] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 2009, pp. 296–301. 3, 4

[76] L. Tran and X. Liu, "Nonlinear 3d face morphable model," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7346–7355. 3

[77] L. Tran, F. Liu, and X. Liu, "Towards high-fidelity nonlinear 3d face morphable model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1126–1135. 3

[78] C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou, "Real-time facial animation with image-based dynamic avatars," *ACM Transactions on Graphics*, vol. 35, no. 4, 2016. 3

[79] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt, "Reconstruction of personalized 3d face rigs from monocular video," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 3, pp. 1–15, 2016. 3

[80] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman, "Unsupervised training for 3d morphable model regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8377–8386. 3

[81] E. Richardson, M. Sela, R. Or-El, and R. Kimmel, "Learning detailed face reconstruction from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1259–1268. 3

[82] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt, "Fml: Face model learning from videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 812–10 822. 3

[83] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018. 3

[84] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019. 3

[85] A. Szabó, G. Meishvili, and P. Favaro, "Unsupervised generative 3d shape learning from natural images," *arXiv preprint arXiv:1910.00287*, 2019. 3

[86] Y. Shi, D. Aggarwal, and A. K. Jain, "Lifting 2d stylegan for 3d-aware face generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6258–6266. 3

[87] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," *Advances in neural information processing systems*, vol. 29, 2016. 3

[88] J.-Y. Zhu, Z. Zhang, C. Zhang, J. Wu, A. Torralba, J. Tenenbaum, and B. Freeman, "Visual object networks: Image generation with disentangled 3d representations," *Advances in neural information processing systems*, vol. 31, 2018. 3

[89] M. Gadelha, S. Maji, and R. Wang, "3d shape induction from 2d views of multiple objects," in *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017, pp. 402–411. 3

[90] T. H. Nguyen-Phuoc, C. Richardt, L. Mai, Y. Yang, and N. Mitra, "Blockgan: Learning 3d object-aware scene representations from unlabelled images," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6767–6778, 2020. 3

[91] C. Xie, C. Wang, B. Zhang, H. Yang, D. Chen, and F. Wen, "Style-based point generator with adversarial rendering for point cloud completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4619–4628. 3

[92] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7462–7473, 2020. 3

[93] Y. Deng, J. Yang, J. Xiang, and X. Tong, "Gram: Generative radiance manifolds for 3d-aware image generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3, 8

[94] J. Xiang, J. Yang, Y. Deng, and X. Tong, "Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds," *arXiv preprint arXiv:2206.07255*, 2022. 3

[95] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *arXiv preprint arXiv:2106.10689*, 2021. 3

[96] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, "Volume rendering of neural implicit surfaces," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4805–4815, 2021. 3

[97] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5784–5794. 3

[98] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou, "Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9054–9063. 3

[99] A. Raj, M. Zollhofer, T. Simon, J. Saragih, S. Saito, J. Hays, and S. Lombardi, "Pixel-aligned volumetric avatars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 733–11 742. 3

[100] S.-Y. Su, F. Yu, M. Zollhoefer, and H. Rhodin, "A-nerf: Surface-free human 3d pose refinement via neural rendering," *arXiv preprint arXiv:2102.06199*, 2021. 3

[101] Z. Wang, T. Bagautdinov, S. Lombardi, T. Simon, J. Saragih, J. Hodgins, and M. Zollhofer, "Learning compositional radiance fields of dynamic human heads," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5704–5713. 3

[102] H. Ouyang, B. Zhang, P. Zhang, H. Yang, J. Yang, D. Chen, Q. Chen, and F. Wen, "Real-time neural character rendering with pose-guided multiplane images," *arXiv preprint arXiv:2204.11820*, 2022. 3

[103] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh, "Neural volumes: Learning dynamic renderable volumes from images," *arXiv preprint arXiv:1906.07751*, 2019. 3

[104] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhofer, "Deepvoxels: Learning persistent 3d feature embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2437–2446. 3

[105] T. DeVries, M. A. Bautista, N. Srivastava, G. W. Taylor, and J. M. Susskind, "Unconstrained scene generation with locally conditioned radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 304–14 313. 3

[106] J. N. Martel, D. B. Lindell, C. Z. Lin, E. R. Chan, M. Monteiro, and G. Wetzstein, "Acorn: Adaptive coordinate networks for neural scene representation," *arXiv preprint arXiv:2105.02788*, 2021. 3

[107] S. Laine, J. Hellsten, T. Karras, Y. Seol, J. Lehtinen, and T. Aila, "Modular primitives for high-performance differentiable rendering," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–14, 2020. 4

[108] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*. PMLR, 2016, pp. 1558–1566. 5

[109] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018. 5, 8

[110] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?" in *International conference on machine learning*. PMLR, 2018, pp. 3481–3490. 5

[111] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341. 6

[112] S. I. Serengil and A. Ozpinar, "Hyperextended lightface: A facial attribute analysis framework," in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 2021, pp. 1–4. [Online]. Available: https://doi.org/10.1109/ICEET53442.2021.9659697 6

[113] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "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, pp. 0–0. 6, 8, 13

[114] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510. 7

[115] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 7

[116] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 7

[117] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016. 8

[118] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017. 9

[119] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826. 9

[120] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763. 9

[121] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844. 9

[122] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699. 11, 13

[123] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595. 13

[124] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 104–12 114, 2020. 14

[125] C. Wang, M. Chai, M. He, D. Chen, and J. Liao, "Cross-domain and disentangled face manipulation with 3d guidance," *arXiv preprint arXiv:2104.11228*, 2021. 14

[126] S. Mo, M. Cho, and J. Shin, "Freeze the discriminator: a simple baseline for fine-tuning gans," *arXiv preprint arXiv:2002.10964*, 2020. 14