MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

CoLa-SDF: Controllable Latent StyleSDF for Disentangled 3D Face Generation

Dey, Rahul; Egger, Bernhard; Boddeti, Vishnu; Wang, Ye; Marks, Tim K.

TR2024-045 May 02, 2024

Abstract

Generating 3D faces and rendering them to images has numerous practical applications in areas including AR/VR, dataset generation, and avatar creation. In recent years, there has been a significant surge in the development of high-fidelity 3D face generation techniques such as StyleSDF, which combine the benefits of 3D implicit neural representations with those of style-based 2D generative adversarial networks (GANs). Although these implicit 3D GAN approaches generate highly realistic faces using a 3D representation, the properties of the generated faces can- not easily be edited or controlled. Meanwhile, linear 3D morphable models (3DMMs) and their nonlinear extensions have also made significant strides in their expressive capacity and quality, but they have yet to match the image quality achieved by GANs. This paper proposes a new method, CoLa-SDF, which combines the controllability of nonlinear 3DMMs with the high fidelity of implicit 3D GANs. Inspired by the impressive photorealism and expressive 3D representations of StyleSDF, our model uses a similar architecture but enforces the latent space to match the interpretable and physical parameters of the nonlinear 3D morphable model MOST-GAN. We demonstrate high-fidelity image synthesis and subsequent 3D manipulation with full control over the disentangled latent parameters.

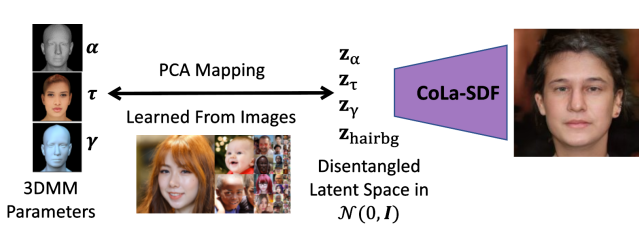
IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

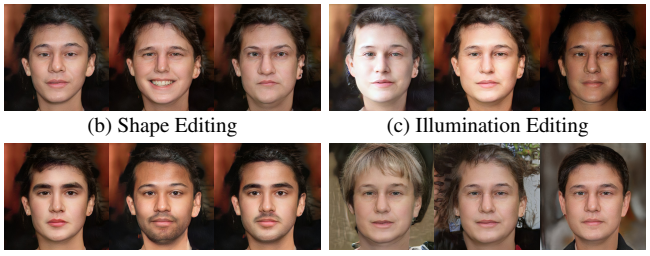
CoLa-SDF: Controllable Latent StyleSDF for Disentangled 3D Face Generation

Rahul Dey[1,2] Bernhard Egger [3] Vishnu Naresh Boddeti[2] Ye Wang[1]
Tim K. Marks[1]

[1]Mitsubishi Electric Research Laboratories (MERL)
[2]Michigan State University (MSU)
[3]Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg



(a) CoLa-SDF learns mapping between 3DMM and implicit 3D GAN.



(b) Shape Editing (c) Illumination Editing

(d) Albedo Editing (e) Hair/Background Editing

Figure 1. CoLa-SDF merges the controllability of 3DMM-based approaches with the high-quality generative capability of implicit 3D GANs, allowing independent control over shape, albedo, illumination, hairstyle, and background of the generated faces.

Abstract

Generating 3D faces and rendering them to images has numerous practical applications in areas including AR/VR, dataset generation, and avatar creation. In recent years, there has been a significant surge in the development of high-fidelity 3D face generation techniques such as StyleSDF, which combine the benefits of 3D implicit neural representations with those of style-based 2D generative adversarial networks (GANs). Although these implicit 3D GAN approaches generate highly realistic faces using a 3D representation, the properties of the generated faces cannot easily be edited or controlled. Meanwhile, linear 3D morphable models (3DMMs) and their nonlinear extensions have also made significant strides in their expressive capacity and quality, but they have yet to match the image quality achieved by GANs. This paper proposes a new method, CoLa-SDF, which combines the controllability of nonlinear 3DMMs with the high fidelity of implicit 3D GANs. Inspired by the impressive photorealism and expressive 3D representations of StyleSDF, our model uses a similar architecture but enforces the latent space to match the interpretable and physical parameters of the nonlinear 3D morphable model MOST-GAN. We demonstrate high-fidelity image synthesis and subsequent 3D manipulation with full control over the disentangled latent parameters.

1. Introduction

Face image generation has a long history in the vision and graphics communities. Regarding 3D face generation, one of the earliest methods was 3D morphable models (3DMMs) [9]. Popular linear 3DMMs such as FLAME [21] and the Basel Face Model [12, 17] are highly controllable and allow disentangled editing of shape, expression, texture, pose, and illumination. However, as they are linear models based on principal components analysis (PCA), the faces synthesized by these models lack fine details in shape and appearance. To address this, there has been a growth in nonlinear 3D face reconstruction approaches [10, 11, 22]. These nonlinear approaches have significantly improved the expressivity of 3DMM models but are still far behind the image quality of generative adversarial networks (GANs). One core limitation of 3DMMs is the strict correspondence assumption. While 3DMMs simplify modeling, they also limit the ability to model the hair, mouth, and other regions whose motion may violate pointwise correspondence.

The striking photorealism of 2D style-based GANs [18–20], as well as the ability of implicit neural representations such as neural radiance fields (NeRFs) [24] and signed distance fields (SDFs) to learn detailed 3D object representations from 2D images, have led researchers to combine the benefits of both models. The combined models [14, 26], which we refer to in this paper as *implicit 3D GANs*, can be

trained in an unsupervised way to learn and synthesize the 3D structure and high-fidelity texture of faces. Essentially, implicit 3D GANs learn to generate an implicit representation of a 3D scene that can be rendered using volumetric rendering techniques [24]. Unlike both linear and nonlinear 3DMMs, implicit 3D GANs can model highly complex structures that do not follow the correspondence assumption (such as hair). However, existing implicit 3D GANs rarely support disentangled control of generation or 3D face editing. There are some exceptions that support partial disentanglement [8, 16, 33], but they often lack photorealism.

The basic idea of our model is to combine the controllability of nonlinear 3DMMs with the photorealism of implicit 3D GANs. Previous methods [8, 22, 37] have combined the photorealism of 2D GANs with the controllability of 3DMMs with good success, but they suffer from limitations. Since both StyleRig [37] and DiscoFaceGAN [8] generate 2D faces and rely on StyleGAN, their controllability is limited by the 3D disentanglement a 2D StyleGAN can learn; e.g., the inherent 2D nature of StyleGAN hampers its disentanglement of pose from other attributes.

Inspired by the excellent 3D shape and texture modeling capabilities of MOST-GAN [22], a nonlinear 3DMM, and by the impressive photorealism and explicit 3D surface modeling of StyleSDF [26], we incorporate both methods into our approach, CoLa-SDF. MOST-GAN's rich texture modeling is supported by use of the StyleGAN2 [20] architecture, but it is unable to model the hair and inner mouth regions in full 3D as there is no pointwise correspondence across subjects in those regions. By combining the disentangled controllability of MOST-GAN [22] with the ability of StyleSDF [26] to learn 3D generation from 2D images, we can retain the best features of both nonlinear 3DMMs and implicit 3D GANs. By incorporating the nonlinear 3DMM via loss functions only, we maintain the photorealism provided by the StyleSDF architecture. The control is enforced during training of the StyleSDF architecture via loss functions that enforce consistency between a set of disentangled latent vectors mapped into MOST-GAN's parametric space ("3DMM Parameters" on the left edge of Fig. 1a), and the parameters obtained by encoding the rendered images using MOST-GAN's encoder.

To summarize, in this paper we propose CoLa-SDF, which combines a nonlinear 3DMM with an implicit 3D GAN in order to get the best of both worlds. Our proposed approach utilizes a differentiable nonlinear 3DMM to supervise the training of an implicit 3D GAN in order to learn disentangled representations for shape, texture, and illumination. In addition, we employ face parsing (semantic segmentation of face images) to further disentangle a latent code for the hair and background from the latent representation of the face. As a result, CoLa-SDF can generate high-fidelity 3D faces, which can then be edited by changing separate latent codes for shape, texture, illumination, pose, and hair and background, either independently or in various combinations. Our main contributions include:

- We propose a new 3D GAN model for faces that uniquely disentangles the latent code into 5 components—pose, shape, albedo, illumination, and hair/background—setting it apart from other controllable 3D GANs;
- The disentanglement is achieved through consistency losses that use a face parser (for hair/background disentanglement) and MOST-GAN encoders (for other attributes' disentanglement), coupled with a carefully designed training routine featuring two sub-iterations;
- We achieve this disentanglement without using any dataset that had controlled (disentangled) face capture
- We perform extensive quantitative and qualitative experiments comparing our model's performance versus other controllable 3D GANs such as [8, 16, 33, 35] and show the superiority of our approach in terms of image generation quality and attribute control.

2. Related Work

2.1. Implicit 3D GANs and Neural Representations

Implicit 3D models, unlike traditional meshes or point clouds, utilize functions (typically implemented using neural networks) to represent 3D objects and scenes. Neural radiance fields (NeRFs), pioneered by Mildenhall et al. [24], query density and radiance at 3D locations through the network, rendering scenes via volume rendering. NeuS [43] employs signed distance fields (SDFs) to represent object surfaces instead of density fields. While these methods were not inherently generative, subsequent approaches similar to neural rendering-based implicit 3D GANs [25, 28] emerged. pi-GAN [5] introduces a novel architecture leveraging periodic activation functions and FiLM [30] to enhance view consistency and generation quality. Computational costs limit these methods' ability to generate high-resolution images, but EG3D [4] proposes a tri-planar framework to boost the computational efficiency of implicit 3D GANs. Recent advancements such as StyleNeRF [14] and StyleSDF [26] combine a low-resolution volume renderer with a CNN-based super-resolution network. Though they offer direct 3D viewpoint manipulation, they lack explicit control over generated objects' shape and appearance.

2.2. Controllable 3D GANs

Several techniques aim to model controllable 3D GANs. BANMo [46] employs neural blend skinning for significant deformations. NeRF-Editing [48] uses ray-bending to edit static NeRFs. StyleRig [37], PIE [36], and GAN-Control [29] project image attributes into StyleGAN's latent space. CLIP-NeRF [42] edits low-resolution objects based on text or exemplars. HeadNeRF [16] disentangled

the latent space of an implicit 3D GAN for faces by training on data containing multiple images for each subject with the same labeled variations in expression and illumination. Disentangled3D [38] and FENeRF [33] train separate shape deformation and appearance networks, but they do not disentangle illumination and only generate low-resolution images. RigNeRF [1] and GNARF [2] learn a canonical-space 3D representation which is then deformed to the desired shape and expression. CGOF [34], 3DFaceShop [35], and Next-3D [32] utilize mesh-guidance to control the shape and expression of generated faces. IDE-3D [31] utilizes semantic-masks and 3D GAN inversion to edit generated faces. As such, these methods either lack photorealism or are restricted to controlling only a few attributes (e.g., shape and expression, but not texture, illumination, or hairstyle).

2.3. Linear and Nonlinear 3DMMs

3D morphable models (3DMMs) of faces are parametric 3D models that enable explicit control over the semantic attributes of the face such as shape, expression, and albedo. Since the first linear 3DMM of human faces [3], these models have grown to include complex pose and expression modalities [12, 17, 21]. While relatively simple and effective, these linear models often lack expressivity and detail. Several nonlinear 3DMM approaches have since been proposed that have improved the expressivity and photorealism of 3DMMs [9–11, 39–41, 45, 47]. Notably, MOST-GAN [22] trained a nonlinear 3DMM to integrate the expressiveness of style-based GANs with the physical disentanglement of 3DMMs, along with a 2D hair manipulation network. In [13], a coarse mesh refinement approach is employed to learn subject-specific head avatars that model the entire head including hair. However, compared to implicit 3D GANs, these models' generation quality is not as good, and they have limited modeling of hair and teeth since these facial features lack pointwise correspondence across subjects and are not part of the underlying 3DMM model.

3. Background

Our method builds upon StyleSDF [26] and MOST-GAN [22], which we now introduce in more detail.

StyleSDF [26] consists of two components: a signed distance function (SDF)-based volume renderer and a 2D styled generator. Given a latent code $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I)$, the volume renderer takes in a 3D query point $\mathbf{x}$ and a viewing direction $\mathbf{v}$ and maps them into an SDF value $d(\mathbf{x}, \mathbf{z})$, a radiance $c(\mathbf{x}, \mathbf{v}, \mathbf{z})$, and a feature vector $f(\mathbf{x}, \mathbf{v}, \mathbf{z})$. A low-resolution (64×64) image $\mathbf{C}$ and feature map $\mathbf{F}$ are generated using volume rendering. Each pixel in $\mathbf{C}$ and $\mathbf{F}$ is computed by querying points along the ray $\mathbf{r} = \mathbf{o} + t\mathbf{p}$ originating from the camera position $\mathbf{o}$ and passing through the pixel location corresponding to $\mathbf{p}$ as follows: $\mathbf{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{p})dt$, where $T(t) = \exp\left(-\int_{t_n}^{t_f} \sigma(\mathbf{r}(s))ds\right)$ represents the visibility of each point along the ray. The density field $\sigma(\mathbf{x})$ is obtained from the SDF $d(\mathbf{x})$ according to $\sigma(\mathbf{x}) = \frac{1}{\alpha}\text{Sigmoid}\left(\frac{-d(\mathbf{x})}{\alpha}\right)$, where α is a learned parameter. The styled generator maps the feature map $\mathbf{F}$ into a high-resolution image $\mathbf{I}$ conditioned on the style code $\mathbf{w} = g(\mathbf{z})$. The volume renderer and the styled generator are trained separately using losses $\mathcal{L}_{\text{volren}}$ and $\mathcal{L}_{\text{gen}}$ (detailed in the supplementary material).

MOST-GAN [22] is a nonlinear 3DMM that includes a set of encoders $\mathbf{E}_\alpha, \mathbf{E}_\tau, \mathbf{E}_\gamma, \mathbf{E}_\theta$, a shape decoder $\mathbf{G}_\alpha$, and an albedo decoder $\mathbf{G}_\tau$. Given a face image, the encoders extract the shape parameters $\boldsymbol{\alpha}$, the albedo parameters $\boldsymbol{\tau}$, the spherical harmonics illumination parameters $\boldsymbol{\gamma}$ [27, 49], and a 3D pose $\boldsymbol{\theta}$. The decoders generate the full 3D shape $\mathbf{S}$ and albedo map $\mathbf{A}$, with $\mathbf{G}_\alpha : \boldsymbol{\alpha} \rightarrow \mathbf{S}$, $\mathbf{G}_\tau : \boldsymbol{\tau} \rightarrow \mathbf{A}$. Next, a differentiable renderer Φ renders the reconstructed face image $\mathbf{I}_{\text{most}}$ from the generated 3D model and lighting and pose parameters: $\mathbf{I}_{\text{most}} = \Phi(\mathbf{S}, \mathbf{A}, \boldsymbol{\gamma}, \boldsymbol{\theta})$. Further details are in the supplementary material. In this work, we use the pretrained MOST-GAN weights provided by the authors.

4. CoLa-SDF

Overview Our proposed approach is based on building a semantically disentangled latent space for an implicit 3D GAN, such that each part of the latent code corresponds to a different physical attribute. We achieve this by enforcing a correspondence between the latent codes for these factors (shape, albedo, and illumination) and the parameters of a 3DMM, which has built-in disentangled representations of these parameters. Pose control can be easily handled using 3D volume rendering and the view-dependence property of implicit 3D GANs [14, 26]. However, 3DMMs do not facilitate disentanglement of hair and background, because these attributes are not represented well in 3DMM models. To encourage part of the latent code to correspond to hair and background, we introduce a photo-consistency loss on the hair and background regions of the generated images that encourages different faces generated using the same hair and background code to have consistent hair and background appearance.

Training the latent space of an implicit 3D GAN to disentangle according to a 3DMM requires the 3DMM to be differentiable and highly expressive, so for our model we adopted the nonlinear 3DMM model MOST-GAN [22], as it matches these requirements. For our implicit 3D GAN architecture, we selected StyleSDF [26], because of its 1) high rendering quality, and 2) explicit modeling of the object's 3D shape in the form of a signed distance field (SDF). Since our proposed modifications and enhancements to StyleSDF enable disentangled control of physical attributes by modi-

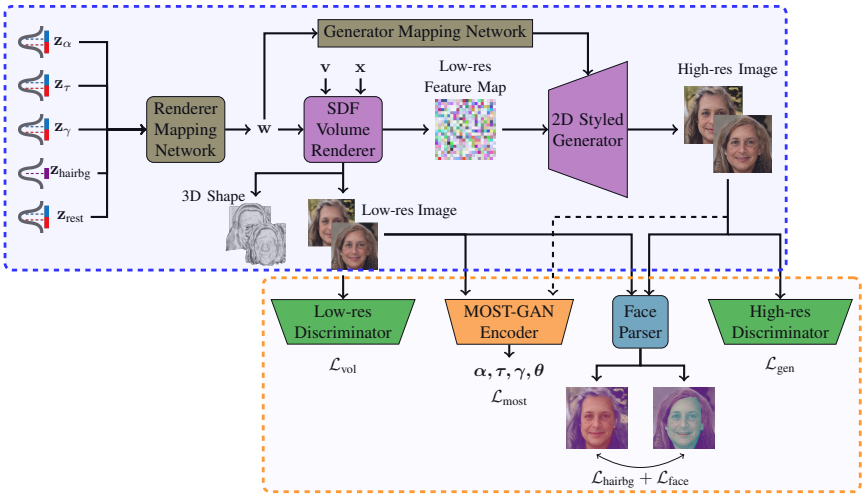


Figure 2. **Overview.** [*Top, inside blue rectangle*]: The SDF volume renderer generates the low-resolution SDF surface, image, and feature map conditioned on the latent codes $\mathbf{z}_\alpha$, $\mathbf{z}_\tau$, $\mathbf{z}_\gamma$, $\mathbf{z}_{\text{hairbg}}$, and $\mathbf{z}_{\text{rest}}$, which the styled generator decodes into a high-resolution image. [*Bottom, inside orange rectangle*]: To disentangle shape, albedo, and illumination, we enforce parametric consistency between the sampled latent codes and the MOST-GAN encodings $\boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\gamma}, \boldsymbol{\theta}$. To disentangle hair/background, we alternately (1) resample face parameters $\mathbf{z}_\alpha, \mathbf{z}_\tau$, and $\mathbf{z}_\gamma$ and enforce image-based consistency on the hair and background; and (2) resample $\mathbf{z}_{\text{hairbg}}$ and enforce consistency on the face regions.

fying disjoint segments of its latent code, we call our model Controllable Latent StyleSDF (CoLa-SDF).

4.1. Architecture

Starting with StyleSDF [26] at the core, we make two important changes to successfully disentangle the latent code, as shown in Fig. 2. First, we partition the 256-dimensional latent code $\mathbf{z}$ into separate latent codes, each denoted as $\mathbf{z}$ with a subscript indicating the attribute to which the code will correspond: the face shape $\mathbf{z}_\alpha$, albedo $\mathbf{z}_\tau$, illumination $\mathbf{z}_\gamma$, and hair and background $\mathbf{z}_{\text{hairbg}}$. We also introduce a final segment of the latent code, $\mathbf{z}_{\text{rest}}$, to which the model is free to assign to any facial appearance factors not explained by MOST-GAN [22]. Second, we modify the training method for StyleSDF to encourage each latent code to control only the desired attribute by incorporating novel consistency loss functions. One set of consistency loss functions enforces consistency between the latent codes that generate a face and the parameters that MOST-GAN extracts from the generated face image. A second set of consistency loss functions minimizes the impact that changes in $\mathbf{z}_{\text{hairbg}}$ can have on the face appearance, and it similarly minimizes the effect that the face-specific latent codes can have on the hair and background appearance. Careful design of both the latent code factorization and the consistency losses during training are crucial to attain the desired disentanglement. We now describe these in detail.

4.2. Latent Code Factorization

We partition the 256 dimensions of the latent code $\mathbf{z}$ into disjoint subsets: 128 dimensions corresponding to the MOST-GAN [22] attributes, further partitioned into $\mathbf{z}_\alpha$, $\mathbf{z}_\tau$, and $\mathbf{z}_\gamma$; 64 dimensions $\mathbf{z}_{\text{hairbg}}$ corresponding to hair and background appearance, and 64 dimensions $\mathbf{z}_{\text{rest}}$ to account for any remaining details in and around the face. To determine the dimensionality to allot to each of the MOST-GAN factors $\mathbf{z}_\alpha$, $\mathbf{z}_\tau$, and $\mathbf{z}_\gamma$, we perform eigen-decomposition over the corresponding data covariance matrices $\boldsymbol{\Sigma}_\alpha$, $\boldsymbol{\Sigma}_\tau$, and $\boldsymbol{\Sigma}_\gamma$, respectively, that we obtain by encoding images in the FFHQ [19] dataset to the MOST-GAN [22] shape $\boldsymbol{\alpha}$, albedo $\boldsymbol{\tau}$, and illumination $\boldsymbol{\gamma}$ parameters using the pretrained MOST-GAN encoders. Based on this analysis, we chose a dimensionality of $d_\alpha = 37$ for $\mathbf{z}_\alpha$ and $d_\tau = 64$ for $\mathbf{z}_\tau$, which accounted for well over 95% of the variance in their respective distributions. In order to enable full explicit control over the 27 spherical harmonics lighting parameters used in MOST-GAN, we chose $d_\gamma = 27$ for $\mathbf{z}_\gamma$. Since we desire $\mathbf{z}_\omega \sim \mathcal{N}(\mathbf{0}, I)$ for $\boldsymbol{\omega} \in (\boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\gamma})$, we use Principal Components Analysis (PCA) to create a mapping between the parameter encoding of MOST-GAN [22] and the corresponding latent codes in our model:

$$\boldsymbol{\omega}_{\text{sample}} = \boldsymbol{U}'_\omega \boldsymbol{\Lambda}'_\omega \mathbf{z}_\omega + \boldsymbol{\mu}_\omega, \tag{1}$$

where $\boldsymbol{U}'_\omega$ and $\boldsymbol{\Lambda}'_\omega$ are the top d_ω eigenvectors and eigenvalues of $\boldsymbol{\Sigma}_\omega$, and $\boldsymbol{\mu}_\omega$ is the data mean.

4.3. Training

StyleSDF Losses: As in [26], we train the model in two stages. In the first stage, we train the volume renderer, then we freeze its weights in the second stage and train the 2D styled generator. In addition to the original StyleSDF losses (described below), in both stages we introduce new consistency losses that we will describe in Section 4.3.1.

In stage 1, we train the volume renderer using loss $\mathcal{L}_{\text{vol}}$

consisting of non-saturating GAN loss with R1 regularization [23] $\mathcal{L}_{\text{adv}}$, pose alignment loss $\mathcal{L}_{\text{view}}$, eikonal loss $\mathcal{L}_{\text{eik}}$, and minimal surface loss $\mathcal{L}_{\text{surf}}$, as defined in [26]. In stage 2, we train the styled 2D generator using loss $\mathcal{L}_{\text{gen}}$ consisting of a path regularization loss $\mathcal{L}_{\text{path}}$ along with $\mathcal{L}_{\text{adv}}$:

$$\mathcal{L}_{\text{vol}} = \mathcal{L}_{\text{adv}} + \lambda_{\text{view}}\mathcal{L}_{\text{view}} + \lambda_{\text{eik}}\mathcal{L}_{\text{eik}} + \lambda_{\text{surf}}\mathcal{L}_{\text{surf}}, \quad (2)$$

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{adv}} + \lambda_{\text{path}}\mathcal{L}_{\text{path}}, \quad (3)$$

where $\lambda_{\text{view}} = 15$, $\lambda_{\text{eik}} = 1$, $\lambda_{\text{surf}} = 1$ and $\lambda_{\text{path}} = 2$.

4.3.1 Disentanglement-Enforcing Consistency Losses

We enforce disentanglement by introducing the MOST-GAN and hair/background consistency losses in both stages of training, in addition to the StyleSDF losses (3). In stage 1, our new losses are applied to the low-resolution images; in stage 2, they are applied to the high-resolution images.

We enforce consistency of the rendered image with respect to the sampled MOST-GAN [22] parameters using the MOST-GAN consistency loss $\mathcal{L}_{\text{most}}$:

$$\mathcal{L}_{\text{most}} = \lambda_\alpha\mathcal{L}_\alpha + \lambda_\tau\mathcal{L}_\tau + \lambda_\gamma\mathcal{L}_\gamma, \quad (4)$$

where $\mathcal{L}_\alpha = ||\mathbf{E}_\alpha(\mathbf{I}) - \boldsymbol{\alpha}_{\text{sample}}||_2^2$ enforces that MOST-GAN's shape encoding of rendered image $\mathbf{E}_\alpha(\mathbf{I})$ is the same as the sampled shape parameters $\boldsymbol{\alpha}_{\text{sample}}$ obtained using Eq. 1. Similarly, we define the albedo consistency loss $\mathcal{L}_\tau$ and the illumination consistency loss $\mathcal{L}_\gamma$. We set $\lambda_\alpha = 3000$, $\lambda_\tau = 100$, and $\lambda_\gamma = 60$.

Existing 3DMM-based approaches do not model hair and background. Hence, to disentangle hair/background from other physical attributes, we adopt a novel approach that restricts the hair/background code $\mathbf{z}_{\text{hairbg}}$ to only model the hair and background. Specifically, we perform a second sub-iteration of the generator with latent code resampling, where, during even iterations, we re-sample $\mathbf{z}_\alpha$, $\mathbf{z}_\tau$ and $\mathbf{z}_\gamma$, and enforce hair/background consistency using $\mathcal{L}_{\text{hairbg}}$. In the odd iterations, we re-sample $\mathbf{z}_{\text{hairbg}}$ and enforce face consistency using $\mathcal{L}_{\text{face}}$. Both $\mathcal{L}_{\text{hairbg}}$ and $\mathcal{L}_{\text{face}}$ are composed of photometric and perceptual components as follows:

$$\mathcal{L}_{\text{hairbg}} = \mathcal{L}_{\text{photo}}(\mathbf{I}_{s1}, \mathbf{I}_{s2}, \mathbf{M}_h) + \mathcal{L}_{\text{vgg}}(\mathbf{I}_{s1}, \mathbf{I}_{s2}, \mathbf{M}_h), \quad (5)$$

$$\mathcal{L}_{\text{face}} = \mathcal{L}_{\text{photo}}(\mathbf{I}_{s1}, \mathbf{I}_{s2}, \mathbf{M}_f) + \mathcal{L}_{\text{vgg}}(\mathbf{I}_{s1}, \mathbf{I}_{s2}, \mathbf{M}_f), \quad (6)$$

where $\mathbf{I}_{s1}$ and $\mathbf{I}_{s2}$ are the images rendered in sub-iterations 1 and 2, respectively. Here, $\mathbf{M}_h = \mathbf{M}_{\text{hairbg},s1} \cup \mathbf{M}_{\text{hairbg},s2}$ is the union of the hair masks from the two sub-iterations, and $\mathbf{M}_f = \mathbf{M}_{\text{face},s1} \cup \mathbf{M}_{\text{face},s2}$ is the union of the face masks from the two sub-iterations, where we have used a pre-trained face parser [6] to segment the rendered face images into a face mask, and a hair and background mask. We define the masked photometric loss as $\mathcal{L}_{\text{photo}}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{m}) = ||(\mathbf{x}_1 - \mathbf{x}_2) \odot \mathbf{m}||_1$, where $\odot$ is the element-wise product operator. Similarly, we define the masked perceptual loss as $\mathcal{L}_{\text{vgg}}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{m}) = ||\phi(\mathbf{x}_1 \odot \mathbf{m}) - \phi(\mathbf{x}_2 \odot \mathbf{m})||_2^2$.

Thus, the overall loss for stage 1, the volume renderer training, is given by:

$$\mathcal{L}_{\text{vol}}^{\text{cola}} = \mathcal{L}_{\text{vol}} + \mathcal{L}_{\text{most}} + \lambda_{\text{hairbg}}\mathcal{L}_{\text{hairbg}} + \lambda_{\text{face}}\mathcal{L}_{\text{face}}. \quad (7)$$

Similarly, the overall loss for stage 2, the training of the 2D styled generator, is given by:

$$\mathcal{L}_{\text{gen}}^{\text{cola}} = \mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{most}} + \lambda_{\text{hairbg}}\mathcal{L}_{\text{hairbg}} + \lambda_{\text{face}}\mathcal{L}_{\text{face}}. \quad (8)$$

We set $\lambda_{\text{hairbg}} = 5$ in even iterations but $= 0$ in odd iterations, and $\lambda_{\text{face}} = 5$ in odd iterations but $= 0$ in even iterations, for both Eqs. (7) and (8).

Initialization of Each Stage: To obtain meaningful MOST-GAN encodings and face parsing, we need the generated images to look like faces. Hence, we initialize each stage by training with only StyleSDF based losses (no consistency losses) for up to 5000 iterations, following which $\mathcal{L}_{\text{most}}$, $\mathcal{L}_{\text{hairbg}}$ and $\mathcal{L}_{\text{face}}$ are introduced. Failing to do so may result in longer training time and poor convergence.

5. Experiments

We train our model on the FFHQ dataset [19], which consists of 70,000 high-resolution images of portrait faces of varying age, ethnicity, and image conditions. We evaluate our model in terms of both its 3D face generation and attribute disentanglement capabilities versus other controllable 3D GANs [2, 8, 16, 33–35]. To evaluate the generation quality and diversity, we compute the Fréchet Inception Distance (FID) [15] for each method. To evaluate attribute control quantitatively, we evaluate the Disentanglement Score (DS) as described in [8] and study the effect of changing various attributes on face identity. Qualitatively, we demonstrate our model's capability to disentangle the latent space for shape, albedo, illumination, and hair/background.

5.1. Face Generation and Multiview Consistency

CoLa-SDF generates high-fidelity 3D faces which can be rendered into photorealistic and view-consistent faces up to at least ± 0.45 radians azimuth and ± 0.225 radians elevation (see Fig. 3). To demonstrate the quality of the underlying 3D surface, we also show the corresponding marching cubes mesh obtained from the signed distance field. Further, we map the shape code $\mathbf{z}_\alpha$ to the MOST-GAN parameter $\boldsymbol{\alpha}$ using Eq. (1) and generate the corresponding MOST-GAN mesh using its shape decoder $\mathbf{S} = \mathbf{G}_\alpha(\boldsymbol{\alpha})$. As shown in Fig. 3, the MOST-GAN meshes correspond well with both the images and SDF surfaces generated by CoLa-SDF. This shows that CoLa-SDF has learned a good correspondence with MOST-GAN in addition to learning a high-quality underlying 3D representation.

To quantitatively evaluate the face generation capability of various approaches, we compute FID metrics [15] at a resolution of 256×256 (see Tab. 1). One can observe that our method reports the second best FID among other con-

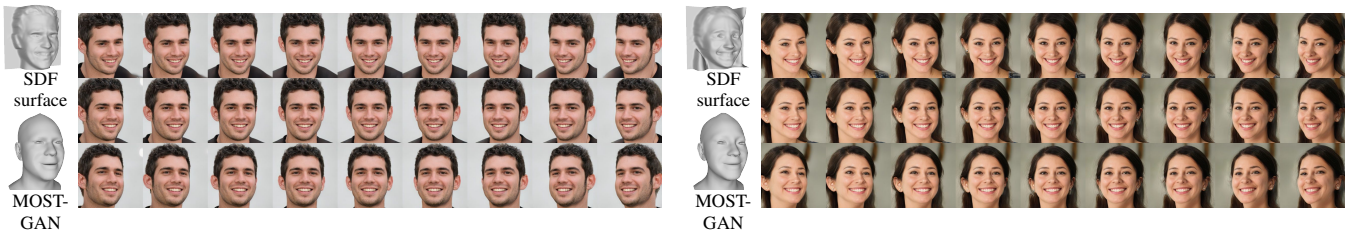


Figure 3. Multiview image renderings and 3D shapes using marching cubes from CoLa-SDF as well as MOST-GAN reconstructions. Please zoom-in for details.

Table 1. Quantitative evaluation of 3D GAN approaches based on FID and Disentanglement Scores (DS) / Controllability (✓). Best results are marked in **bold**, while second best is marked in red.

Method	3D Representation	Disentanglement Score (↑) / Controllability (✓) Iden	Expr	Alb	Illu	Pose	Hair/Bg	FID (↓)
GAN-Control [29]	3DMM	7.07	7.51			9.33		83.6
DiscoFaceGAN [8]	3DMM	5.97	15.70		✓	5.23		77.5
StyleRig [37]	3DMM	1.64	13.03			2.01		56.7
HeadNeRF [16]	Volume	6.39	5.99	✓	✓	10.26		159.6
FENeRF [33]	Volume	✓		✓		✓		150.1
PIE [36]	Volume	1.66	15.24			2.65		59.6
CGOF [34]	Volume	21.72	27.47			**22.82**		31.8
3DFaceShop [35]	Triplane	✓	✓		✓	✓		24.1
GNARF [2]	Triplane	✓	✓			✓		**17.9**
CoLa-SDF (Ours)	Volume	**28.15**		**23.60**	**20.72**	16.72	✓	19.4

trollable implicit 3D GANs. Both FENeRF [33] and HeadNeRF [16] only generate the face area and not background, as can be seen in Figs. 5f and 5j, which could partially explain their poor FIDs.

5.2. Disentanglement of the Latent Space

We evaluate the disentaglement capability of 3D GANs in terms of Attribute Disentanglement Score (DS) [8]. We report the disentanglement scores of our method and other approaches in Tab. 1. For some approaches, the DS is not available. In those cases, we use a ✓to denote which attributes that method can disentangle. CoLa-SDF's disentanglement scores are the highest for shape, albedo and illumination, and second highest for pose. In addition, our method is the only one that disentangles hair and background. In the following paragraphs, we further qualititatively evaluate CoLa-SDF's latent space disentanglement in terms of shape, albedo, illumination, and hair/background.

Shape, Albedo, Lighting, and Hairstyle Manipulation: To demonstrate the disentanglement capability of our model, we manipulate the shape, albedo, lighting, and hair and background of generated faces and show their variations (see Fig. 1). To alter the attributes of a face image generated using some latent code $\mathbf{z}$, we independently resample one or more of $\mathbf{z}_\alpha$, $\mathbf{z}_\tau$, $\mathbf{z}_\gamma$ and $\mathbf{z}_{\text{hairbg}}$ from the latent space. Subsequently, we replace the original values in the chosen segments of $\mathbf{z}$ with the newly sampled values, and generate a modified image. Changing the shape code allows us to explore various expression and structural modalities in faces. Altering the albedo code results in changes to properties such as lip color, skin tone, facial hair, and eyebrow density, while leaving the face shape virtually unchanged. Similarly, varying the illumination and hair/background latent codes only affect those factors, while maintaining the face's shape and albedo.

Illumination Editing using Spherical Harmonics: Since MOST-GAN's illumination code is based on the spherical harmonics coefficients [27], we can perform controlled manipulation of illumination. We can directly configure the values of these coefficients and use Eq. (1) to map these values into the space of $\boldsymbol{z}_\gamma$. We traverse through the first two spherical harmonics bases for each channel and show the illumination variations in Fig. 4a. Traversing through the first basis results in global illumination change, while traversing through the second basis results in the illumination direction changing from right to left. Notice that as the magnitude and direction of light changes, it affects not only the face but also the hair and background. This is in contrast to 3DMM-based approaches, which apply illumination only to the face region. This makes illumination editing using CoLa-SDF more natural than using MOST-GAN.

5.2.1 Attribute Transfer

To further demonstrate the attribute disentanglement of our method, we transfer attributes such as shape, albedo, lighting, and hair and background from a source image (left column) to a target image (top row), as illustrated in Fig. 5.

Illumination Transfer (Fig. 4b): CoLa-SDF can transfer

(a) Illumination editing using spherical harmonics.

(b) Illumination transfer.

Figure 4. CoLa-SDF disentangles illumination and maps it to the spherical harmonics space [27], enabling us to edit and transfer illumination. (a) For three randomly generated faces, we alter the lighting by directly modifying the spherical harmonics coefficients. Varying the first coefficient (*left*) controls the level of global (ambient) illumination, while the second coefficient (*right*) controls the illumination's horizontal directionality. (b) Illumination transfer from source to target.

the tone, hue, brightness, and direction of the illumination across the image, including illumination of the hair and background. This includes some of the rare illumination conditions in the training dataset like rows 2 and 3. On the other hand, 3DFaceShop [35] can be ineffective in illumination transfer (see Fig. 5i, where illumination is not effectively transferred from the source image to the destination). In DiscoFaceGAN [8], editing the illumination can incorrectly alter the background and clothing (see Fig. 5e).

Shape Transfer (Fig. 5a): Our method can transfer extreme identity- and expression-related shape variations from the source to the target, while keeping other physical attributes intact. We show transfer of face attributes including face width and height, shape of nose, jawline, as well as expression changes. In contrast, 3DFaceShop [35] hardly transfers any identity or expression from the source to the target, as shown in Figs. 5g and 5h (see columns/rows boxed in red). In FENeRF [33], shape transfer also causes unintentional transfer of some appearance information, such as face texture and hair texture—the identity matches the source image *much* more closely than the target face, which would not be the case with shape-only transfer.

Albedo Transfer (Fig. 5b): As reported in Tab. 1, our method is one of the only methods that can transfer albedo, including attributes such as skin tone, thickness of eyebrows, and lip color. Interestingly, our model can also transfer eyeglasses (row 4), which are external to the face and hence not accounted for by any 3DMM model. In our trained model, we were surprised to observe that hair color is somewhat affected by the albedo code as well as the hair/background code.

Hair and Background Transfer (Fig. 5c): Notice that transferring the hair and background does not change the identity or other attributes of the face. In this figure, we again observe that while the hair/background code determines the hair geometry/hairstyle, its color is also partly controlled by the albedo code.

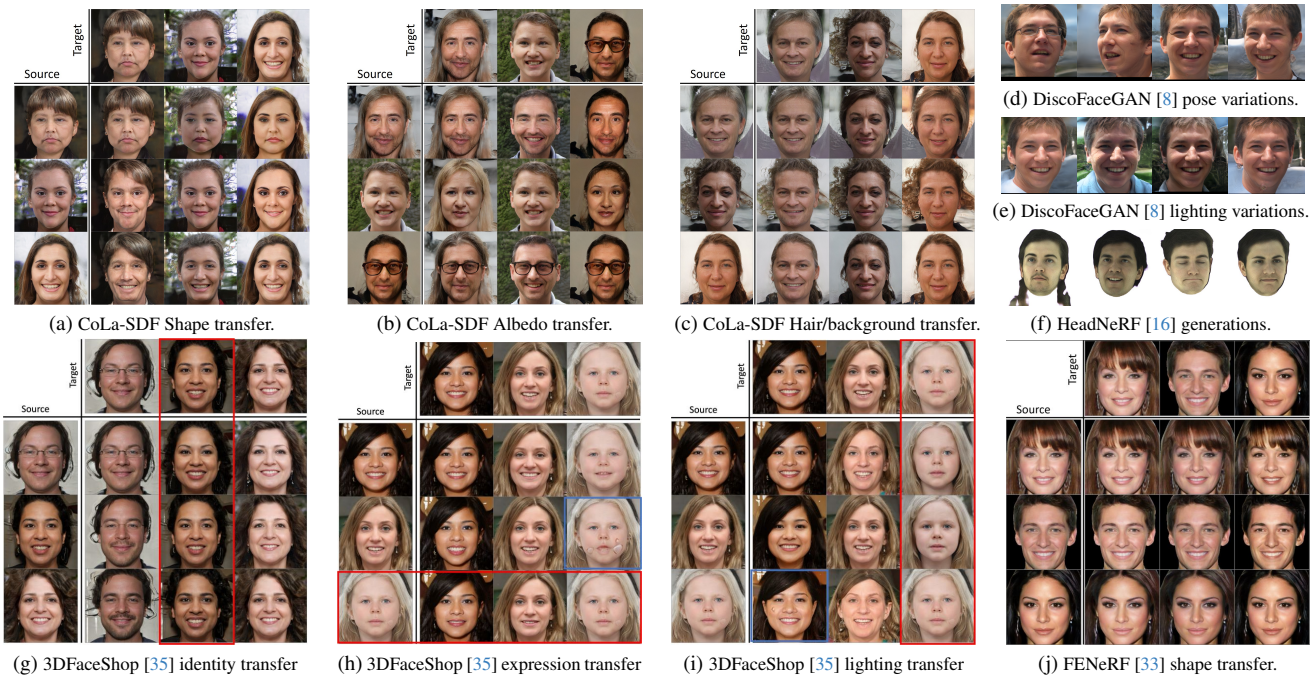
Table 2. Identity consistency between pairs of faces generated by CoLa-SDF that differ only in their pose, illumination, hair/bg, or shape and/or albedo latents. For each attribute, we report % of pairs with $< 70^\circ$ distance, as measured by ArcFace [7]. CoLa-SDF retains facial identity when changing non-identity-related attributes (columns 1–3), but changes identity otherwise.

Pose	Illu	Hair/Bg	Shape	Albedo	Shape + Albedo
97.7	99.7	98.2	75.2	65.7	4.7

Additional Observations: HeadNeRF [16] can disentangle identity, expression, albedo, and illumination well, but it has poor photorealism (see Fig. 5f). In DiscoFaceGAN [8], modifying the pose can unintentionally add artifacts such as glasses and expression changes (see Fig. 5d). In contrast, CoLa-SDF can simultaneously control pose and transfer other attributes as shown in the supplementary.

5.2.2 Identity Consistency across Unrelated Attributes

We further evaluate the disentanglement of CoLa-SDF by studying the effect on the identity of the generated face when we change identity-related attributes such as shape and albedo versus non-identity-related attributes such as pose, illumination, and hair/background. We randomly generated 1000 faces from our model and edited their view, illumination, hair/background, shape, and albedo by resampling the corresponding latent codes. We extract face identity features using ArcFace [7], and measure the identity match between the original and edited faces. Based on the analysis in [7], we set 70° as the threshold for a reasonable match. The results, in Tab. 2, show that as desired, changes in viewpoint, illumination, and hair/background have minimal impact on the generated face's identity. In contrast, changing shape and albedo individually cause partial but not complete identity alterations (this corresponds well with human perception of identity changes in Figs. 5a and 5b), while simultaneously changing both shape and



(a) CoLa-SDF Shape transfer. (b) CoLa-SDF Albedo transfer. (c) CoLa-SDF Hair/background transfer. (d) DiscoFaceGAN [8] pose variations. (e) DiscoFaceGAN [8] lighting variations. (f) HeadNeRF [16] generations. (g) 3DFaceShop [35] identity transfer (h) 3DFaceShop [35] expression transfer (i) 3DFaceShop [35] lighting transfer (j) FENeRF [33] shape transfer.

Figure 5. Qualitative comparison of face generation and attribute disentanglement by various controllable 3D GANs. Attribute transfer results from 3DFaceShop [35] (g)–(i) show lack of disentanglement, as highlighted by red boxes. (j) Shape transfer in FENeRF [33] changes the target almost completely, instead of just transferring the facial structure (poor disentanglement). DiscoFaceGAN [8] (d) adds artifacts such as glasses and expression changes when rendering from different viewpoints, and it also (e) changes background and clothes when editing lighting. (f) HeadNeRF [16] does not generate photorealistic faces. On the other hand, CoLa-SDF effectively transfers the desired attribute while leaving the other facial attributes intact (a)-(c). Please zoom in for details.

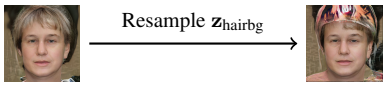


Figure 6. Spurious artifacts during hairstyle editing.

albedo codes results in a clear change of identity. This demonstrates that our method has successfully disentangled the identity-related attributes of the face from its non-identity-related attributes.

6. Discussion

Limitations: CoLa-SDF may exhibit some artifacts in hair/background editing (see Fig. 6). This might stem from the model's challenge in discerning between hair and other headwear, leading to blending between them. Additionally, as one may observe in Figs. 5b and 5c, although CoLa-SDF accurately captures hair geometry, hair color may be influenced by a mix of albedo and hair/background codes, likely due to dataset correlations. One potential remedy could involve incorporating an explicit 3D hair model during training [44, 50]. Furthermore, as MOST-GAN's shape parameter governs both identity- and expression-related shape changes, our method lacks disentanglement of these factors. One approach could be to replace the MOST-GAN encoder with a feature encoder designed for this disentanglement.

Ethics discussion: CoLa-SDF carries the potential for misuse in fabricating fake content. Our work is intended for research purposes, and we strongly denounce any improper use of our work to disseminate misinformation, harm reputations, or violate rights.

Conclusion: We propose a controllable 3D GAN, dubbed CoLa-SDF, that combines the disentangled controllability of nonlinear 3DMM approaches with the high fidelity of implicit 3D GANs for generating 3D faces and rendering them to images. Building upon the architecture of StyleSDF, we enforce the latent space to match the physical parameters of the nonlinear 3D morphable model MOST-GAN by encouraging the latent parameters to match a statistical decomposition of the MOST-GAN parameters during model training. In addition, we disentangle the control of hair and background using novel consistency losses applied over multiple sub-iterations of training. We demonstrate high-fidelity face synthesis and subsequent 3D manipulation with full control over the disentangled latent parameters. Our model presents a promising solution for generating high-quality 3D faces with controllable properties, which has practical applications in many areas including AR/VR, dataset synthesis and augmentation, media, and avatar creation.

References

[1] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20364–20373, 2022. 3

[2] Alexander W. Bergman, Petr Kellnhofer, Wang Yifan, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. In *NeurIPS*, 2022. 3, 5, 6

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 3

[4] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2

[5] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 2

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 5

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 7

[8] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5154–5163, 2020. 2, 5, 6, 7, 8

[9] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 1, 3

[10] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. volume 40, 2021. 1

[11] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 3

[12] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. 1, 3

[13] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 3

[14] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 1, 2, 3

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[16] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 2, 5, 6, 7, 8

[17] IEEE. *A 3D Face Model for Pose and Illumination Invariant Face Recognition*, Genova, Italy, 2009. 1, 3

[18] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 1

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 4, 5

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1, 2

[21] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 1, 3

[22] Safa C Medin, Bernhard Egger, Anoop Cherian, Ye Wang, Joshua B Tenenbaum, Xiaoming Liu, and Tim K Marks. Most-gan: 3d morphable stylegan for disentangled face image manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1962–1971, 2022. 1, 2, 3, 4, 5

[23] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018. 5

[24] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[25] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[26] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shecht-

man, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 1, 2, 3, 4, 5

[27] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500, 2001. 3, 6, 7

[28] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 2

[29] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. Gan-control: Explicitly controllable gans. In *ICCV*, pages 14083–14093, 2021. 2, 6

[30] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 2

[31] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. 2022. 3

[32] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *CVPR*, 2023. 3

[33] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7672–7682, 2022. 2, 3, 5, 6, 7, 8

[34] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and HongSheng Li. Controllable 3d face synthesis with conditional generative occupancy fields. *NeurIPS*, 2022. 3, 6

[35] Junshu Tang, Bo Zhang, Binxin Yang, Ting Zhang, Dong Chen, Lizhuang Ma, and Fang Wen. 3dfaceshop: Explicitly controllable 3d-aware portrait generation. *TVCG*, 2023. 2, 3, 5, 6, 7, 8

[36] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 2, 6

[37] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020. 2, 6

[38] Ayush Tewari, Xingang Pan, Ohad Fried, Maneesh Agrawala, Christian Theobalt, et al. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1516–1525, 2022. 3

[39] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017. 3

[40] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1126–1135, 2019.

[41] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 3

[42] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022. 2

[43] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[44] Keyu Wu, Yifan Ye, Lingchen Yang, Hongbo Fu, Kun Zhou, and Youyi Zheng. Neuralhdhair: Automatic high-fidelity hair modeling from a single image using implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1526–1535, 2022. 8

[45] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2020. 3

[46] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2873, 2022. 2

[47] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12803–12813, 2021. 3

[48] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. Nerf-editing: geometry editing of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18353–18364, 2022. 2

[49] Lei Zhang and Dimitris Samaras. Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):351–363, 2006. 3

[50] Meng Zhang and Youyi Zheng. Hair-gan: Recovering 3d hair structure from a single image using generative adversarial networks. *Visual Informatics*, 3(2):102–112, 2019. 8