Next3D: Generative Neural Texture Rasterization for 3D-Aware Head Avatars

Jingxiang Sun[1] Xuan Wang[2] Lizhen Wang[14] Xiaoyu Li[3] Yong Zhang[3]
Hongwen Zhang[1] Yebin Liu[1]
[1]Tsinghua University [2]Ant Group [3]Tencent AI Lab [4]NNKosmos

Gaze Animation   3D-Aware Domain Adaptation

Figure 1. Our 3D GAN synthesizes generative, high-quality, and 3D-consistent facial avatars from unstructured 2D images. Unlike current animatable 3D GANs that only modify yaw-pitch head poses and facial expressions, our approach enables fine-grained control over full-head rotations, facial expressions, eye blinks, and gaze directions with strict 3D consistency and a high level of photorealism. Our approach also provides strong 3D priors for downstream tasks such as 3D-aware stylization.

Abstract

3D-aware generative adversarial networks (GANs) synthesize high-fidelity and multi-view-consistent facial images using only collections of single-view 2D imagery. Towards fine-grained control over facial attributes, recent efforts incorporate 3D Morphable Face Model (3DMM) to describe deformation in generative radiance fields either explicitly or implicitly. Explicit methods provide fine-grained expression control but cannot handle topological changes caused by hair and accessories, while implicit ones can model varied topologies but have limited generalization caused by the unconstrained deformation fields. We propose a novel 3D GAN framework for unsupervised learning of generative, high-quality and 3D-consistent facial avatars from unstructured 2D images. To achieve both deformation accuracy and topological flexibility, we propose a 3D representation called Generative Texture-Rasterized Tri-planes. The proposed representation learns Generative Neural Textures on top of parametric mesh templates and then projects them into three orthogonal-viewed feature planes through rasterization, forming a tri-plane feature representation for volume rendering. In this way, we combine both fine-grained expression control of mesh-guided explicit deformation and the flexibility of implicit volumetric representation. We further propose specific modules for modeling mouth interior which is not taken into account by 3DMM. Our method demonstrates state-of-the-art 3D-aware synthesis quality and animation ability through extensive experiments. Furthermore, serving as 3D prior, our animatable 3D representation boosts multiple applications including one-shot facial avatars and 3D-aware stylization. Project page: https://mrtornado24.github.io/Next3D/. Code: https://github.com/MrTornado24/Next3D.

1. Introduction

Animatable portrait synthesis is essential for movie post-production, visual effects, augmented reality (AR), and virtual reality (VR) telepresence applications. Efficient animatable portrait generators should be capable of synthesizing diverse high-fidelity portraits with full control of the rigid head pose, facial expressions and gaze directions at a fine-grained level. The main challenges of this task lie in how to model accurate deformation and preserve identity through animation in the generative setting, i.e. training with only unstructured corpus of 2D images.

Several 2D generative models perform image animation by incorporating the 3D Morphable Face Models (3DMM) [4] into the portrait synthesis [13, 16, 35, 52, 62, 65, 70, 73]. These 2D-based methods achieve photorealism but suffer from shape distortion during large motion due to a lack of geometry constraints. Towards better view consistency, many recent efforts incorporate 3DMM with 3D GANs, learning to synthesize animatable and 3D consistent portraits from only 2D image collections in an unsupervised manner [3, 30, 39, 44, 60, 61, 68, 74]. Bergman et al. [3] propose an explicit surface-driven deformation field for warping radiance fields. While modeling accurate facial deformation, it cannot handle topological changes caused by non-facial components, e.g. hair, glasses, and other accessories. AnifaceGAN [68] builds an implicit 3DMM-conditioned deformation field and constrains animation accuracy by imitation learning. It achieves smooth animation on interpolated expressions, however, struggles to generate reasonable extrapolation due to the under-constrained deformation field. Therefore, The key challenge of this task is modeling deformation in the 3D generative setting for animation accuracy and topological flexibility.

In this paper, we propose a novel 3D GAN framework for unsupervised learning of generative, high-quality, and 3D-consistent facial avatars from unstructured 2D images. Our model splits the whole head into dynamic and static parts, and models them respectively. For dynamic parts, the key insight is to combine both fine-grained expression control of mesh-guided explicit deformation and flexibility of implicit volumetric representation. To this end, we propose a novel representation, *Generative Texture-Rasterized Triplanes*, which learns the facial deformation through *Generative Neural Textures* on top of a parametric template mesh and samples them into three orthogonal-viewed and axis-aligned feature planes through standard rasterization, forming a tri-plane feature representation. Such texture-rasterized tri-planes re-form high-dimensional dynamic surface features in a volumetric representation for efficient volume rendering and thus inherit both the accurate control of the mesh-driven deformation and the expressiveness of volumetric representations. Furthermore, we represent static components (body, hair, background, etc.) by another tri-plane branch, and integrate both through alpha blending.

Another key insight of our method is to model the mouth interior which is not taken into account by 3DMM. Mouth interior is crucial for animation quality but often ignored by prior arts. We propose an efficient teeth synthesis module, formed as a style-modulated UNet, to complete the inner mouth features missed by the template mesh. To further regularize the deformation accuracy, we introduce a deformation-aware discriminator which takes as input synthetic renderings, encouraging the alignment of the final outputs with the 2D projection of the expected deformation.

To summarize, the contributions of our approach are:

- We present an animatable 3D-aware GAN framework for photorealistic portrait synthesis with fine-grained animation, including expressions, eye blinks, gaze direction and full head poses.
- We propose *Generative Texture-Rasterized Triplanes*, an efficient deformable 3D representation that inherits both fine-grained expression control of mesh-guided explicit deformation and flexibility of implicit volumetric representation.
- Our learned generative animatable 3D representation can serve as a strong 3D prior and boost the downstream application of 3D-aware one-shot facial avatars. Our model also pushes the frontier of 3D stylization with high-quality out-of-domain facial avatars.

2. Related Work

Generative 3D-aware Image Synthesis. Generative adversarial networks [24] have achieved photorealistic synthesis in 2D domain. Building on the success of 2D GANs, many efforts have lifted the image synthesis into 3D with explicit view control. Early voxel-based approaches [19, 28, 41, 42, 67, 82] adopt the 3D CNN generators whose heavy computational burden limits the high-resolution image synthesis. Recent works incorporate more efficient neural scene representations, such as fully implicit networks [6, 8–10, 14, 47, 54, 57, 81], sparse voxel grids [55], multiple planes [79] or a combination of low-resolution feature volume and 2D super-resolution [7, 27, 43, 46, 71, 72, 76, 78]. We leverage the tri-plane representation proposed in [7] and endow it with animation ability by the orthogonal-rasterized *Generative Neural Textures*.

Some other current works [3, 30, 36, 44, 58, 59, 61, 68, 77, 83] focus on the editability of 3D-aware generative models. FENeRF [59] and IDE-3D [58] perform semantic-guided 3D face editing by incorporating semantic-aware radiance fields and GAN inversion. However, they cannot produce continuous and stable editing on videos. Other methods [3, 30, 44, 61, 68] employ 3D priors to achieve animatable image synthesis and the main differences lie on the deformation strategies including linear blend skinning [30, 44], surface-driven deformation [3], 3DMM-guided latent decomposition [61] and neural deformation fields [68]. These approaches either don't allow for topology changes or need

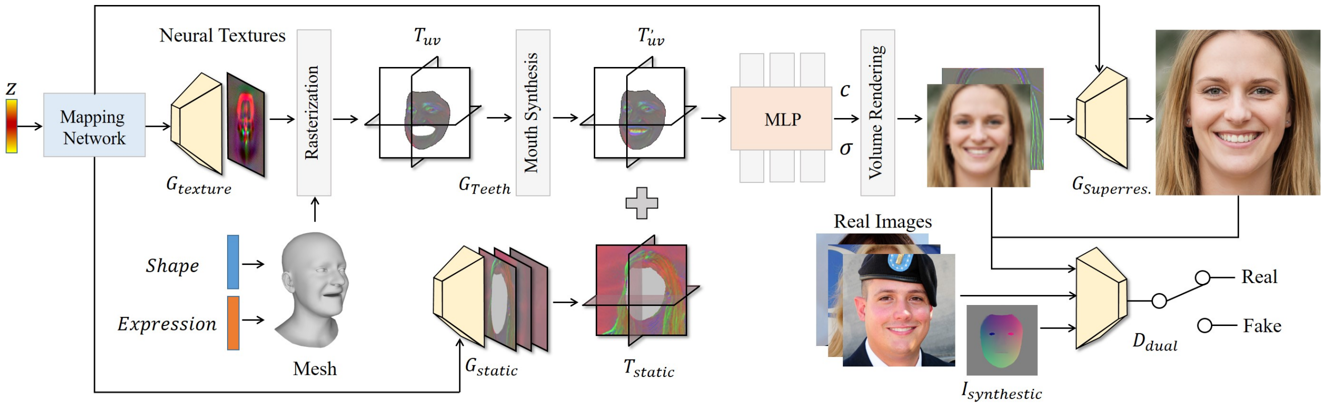

Figure 2. Our 3D GAN framework consists of two tri-plane branches T_{uv} and T_{static} modeling dynamic and static components. T_{uv} is formed by the orthogonal rasterized Generative Neural Textures which are synthesized by a StyleGAN generator, $G_{texture}$, on top of deformable template mesh. T_{static} is synthesized by another StyleGAN generator, G_{static}. The mouth synthesis module, G_{teeth}, is presented for completing mouth interior. Blended triplanes are incorporated with hybrid neural renderer consisting of volume rendering and a super-resolution module $G_{superres.}$. For discrimination, synthetic renderings $I_{synthetic}$ is taken into the dual discriminator D_{dual}.

elaborate loss design to ensure the accuracy of deformation. On the contrary, our approach naturally achieves accurate animation with explicit mesh guidance and further allow for topology changes by adapting surface deformation into a continuous volumetric representation.

Facial animation with 3D morphable face models. Blanz et al. [4] model facial texture and shape as vector spaces, known as the 3D Morphable Model (3DMM). Extensions of 3DMM, such as full-head PCA models [12, 51], blend-shape models [38], are extensively studied and widely used in facial animation tasks [23, 65]. Benefiting from 3DMM, these methods can model the deformation of facial parts accurately and continuously, nevertheless, struggle to represent non-facial areas missed by 3DMM, e.g. hair, teeth, eyes, and body. Moreover, these methods are prone to lacking facial details. To fill the missing areas and complete more realistic facial details, later works [13, 16, 18, 22, 35, 52, 62, 64, 70] apply learned approaches on top of 3DMM renderings. DiscoFaceGAN [13] maps 3DMM parameters into the latent space and decouple them by imitative-contrastive learning. Though efficient and photorealistic animation is achieved, these 2D methods don't model 3D geometry and thus cannot remain strict 3D consistency with large head pose changes. For strict 3D consistency, recent efforts [1, 20, 25, 75, 80] incorporate 3DMM into volumetric representations [31, 32, 40, 49, 56] to achieve view-consistent facial animation. Furthermore, volumetric representations enable the modeling of thin structures such as hair, and also mouth interior thanks to its spatial continuity. While 3DMM have been adopted to animate radiance fields for single-scene scenarios, it is challenging to adapt to the generative setting with the absence of groundtruth supervision.

Neural scene representations. The neural scene representations can be roughly categorized into implicit and explicit surface representations and volumetric representations [63]. The surface can be represented explicitly by point clouds [37, 50], meshes [2, 5, 45, 64], or defined implicitly as a zero level-set of a function [11, 48, 69], like signed distance function, which can be approximated by coordinate-based multi-layer perceptrons (MLPs). On the contrary, volumetric representations [32, 40, 49, 56] store volumetric properties (occupancies, radiance, colors, etc.) instead of the surface of an object. These properties can be stored in explicit voxel grids [32, 49, 56] or the weights of a neural network implicitly [40]. In this work, we propose a hybrid surface–volumetric representation. Specifically, we learn the deformable surface radiance by neural textures [26, 64] on top of the template mesh and rasterize it into three orthogonal-viewed feature planes. Then, the planes are reshaped to a tri-plane representation with decoding into neural radiance fields for volume rendering.

3. Approach

We present an animatable 3D-aware facial generator that equips with fine-grained expression and pose control, photorealistic rendering quality, and high-quality underlying geometry. The proposed method models dynamic and static components by two independent tri-plane branches. Specifically, we propose *Generative Texture-rasterized Tri-planes* for modeling dynamic facial parts (Sec. 3.1). Furthermore, we propose an efficient mouth synthesis module to complete the mouth interior that is not included in 3DMM (Sec. 3.2). We further adopt another tri-plane branch for the static components (Sec. 3.3). Both tri-planes are blended together for hybrid neural rendering (Sec. 3.4). We introduce an deformation-aware discriminator (Sec. 3.5) and illustrate the training objectives in Sec. 3.6.

3.1. Generative texture-rasterized tri-planes

EG3D [7] presents an efficient tri-plane-based hybrid 3D representation to synthesize high-resolution images with multi-view consistency. Nonetheless, EG3D lacks control over facial deformations and thus cannot be directly applied to animation tasks. To this end, we leverage Neural textures [64] to represent deformable facial parts. In general, Neural Textures are a set of learned high-dimensional feature maps that can be interpreted by a neural renderer. We extend it to our generative setting and synthesize the neural textures through a StyleGAN2 CNN generator $G_{texture}$. As shown in Fig. 2, we first sample a latent code z and map it into an intermediate latent space by the mapping network. Our texture generator architecture closely follows StyleGAN2 backbone [34], except producing a $256 \times 256 \times 32$ neural texture map, T, instead of a three-channel RGB image. Storing a high-dimensional learned feature vector per texel, T can be rasterized to a view-dependent screen-space feature map given a mesh with uv-texture parameterization and a target view as input. In our case, we use the FLAME template [38] to provide a coarse mesh that can be driven by deformation parameters. Given the pre-designed texture mapping function, we employ the standard graphics pipeline to rasterize the neural textures from the texture space into the screen space based on the template mesh. We choose *Neural Textures* as the deformation method for two reasons. First, compared with other explicit deformation (e.g. linear blend skinning and surface-driven deformation) highly dependent on the accurate underlying geometry, Neural Textures embed high-level features which compensate the imperfect geometry and thus are more suitable for our settings where template meshes are not accurate. Furthermore, unlike implicit deformation methods [13, 61, 68], our explicit mesh-guided deformation alleviates the requirement of elaborate imitation learning while gaining better expression generalization (Fig. 4).

Neural Textures encode surface deformation accurately with mesh guidance but lack generalization to 3D points far from surface. Besides, it also doesn't allow for topological changes. To this end, we propose Generative Texture-Rasterized Tri-planes, T_{uv} which reshapes the rasterized textures into a tri-plane representation. Therefore, we can adapt such surface deformation into a continuous volume, and once a 3D point out of the surface could be projected within the facial part on at least one view (e.g. glasses), its feature can be interpolated by the meaningful surface feature through grid sampling in the tri-plane. Specifically, we rasterize neural textures based on the template mesh into three orthogonal views and place them in three axis-aligned feature planes. In practice, considering the zygomorphy, the rasterization is applied at both the left and right views and the rasterized features are concatenated for one single plane by summation. In this way, Our hybrid surface-volumetric representation inherits the best of both worlds: accurate mesh-guided facial deformation of Neural Textures and the continuity and topological flexibility of volumetric representations. See Fig. 5 for the topological-aware animation of portraits with glasses.

3.2. Mouth synthesis module

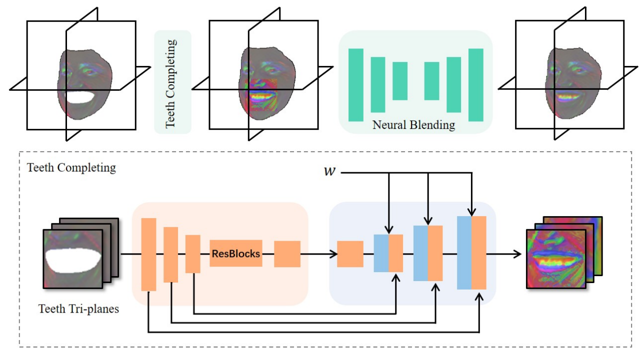

Figure 3. The teeth synthesis module consists of a teeth completing module and a neural blending module. The teeth competing module produces mouth interior conditioned on multi-scale teeth exterior features and latent codes w.

Since the FLAME template doesn't contain inner mouth, we propose a teeth synthesis module, G_{teeth}, to complete the missing teeth features in T_{uv}. As shown in Fig. 3, for each feature plane of T_{uv}, we crop the teeth area by the expanded landmarks and resize it into 64×64. Then, the stacked mouth features are processed by G_{teeth} which employs a style-modulated UNet [66]. The downsampling process of G_{teeth} encodes f_{teeth} into multi-scale feature maps which serve as content conditions for the following StyleGAN layers. The output teeth features $f^{'}_{teeth}$ are transformed inversely and concatenated with the feature planes of T_{uv}. To eliminate the texture flickering of mouth boundary, we further feed T_{uv} into a shallow-UNet-based neural blending module and obtain $T^{'}_{uv}$. We conduct a series of ablation studies and prove that the proposed teeth synthesis module brings a remarkable improvement on both the animation accuracy and synthesis quality (Sec. 4.2).

3.3. Modeling static components

The generative texture-rasterized tri-planes manage to model dynamic faces varying expressions and shapes, though, it is challenging to synthesize static parts like diverse haircut, background and upper body which are not included in the FLAME template. To this end, we model these parts by another tri-plane branch, T_{static}, which is generated by a StyleGAN2 CNN generator G_{static} sharing the same latent code with $G_{texture}$. The plane features of $T^{'}_{uv}$ and T_{static} are blended on each plane by the alpha masks rendered by rasterization. Such design not only benefits the modeling of various-styled static components but also enforces their consistency during facial animations.

3.4. Neural rendering

Given the blended tri-planes, for any point in the 3D space, we project it into each plane and sample the features bi-linearly. Then, the sampled features are aggregated by summation and decoded into volume density σ and feature f by a lightweight decoder. Similar to [7,58], the decoder is a single hidden layered multi-layer perceptron (MLP) with softplus activation. The volume rendering is employed to accumulate σ and f along the rays cast through each pixel to compute a 2D feature image I_f. Similar to [7,27,46], we leverage a 2D super-resolution module $G_{superres.}$ to interpret the feature image into RGB image I_{RGB} with higher resolution. The super-resolution module consists of three StyleGAN2 synthesis blocks and the noise input is removed for alleviating texture flickers. In our case, I_f and I_{RGB} are set to 64×64 and 512×512, respectively.

3.5. Deformation-aware discriminator

To learn the unsupervised 3D representations, we adopt a 2D convolutional discriminator D to critique the renderings. Inspired by [7], we regularize the first three channels of I_f as low-resolution RGB image, which is concatenated with I_{RGB} as the input for the discriminator. However, the discrimination with only image input can only ensure that the deformed images are always in a correct distribution instead of matching the expected deformations. Therefore, we make the discriminator aware of the expression and shape under which the generated image are deformed. For the same purpose, GNARF [3] conditions the discriminator on the FLAME parameters by concatenating them as the input to the mapping network. However, we find empirically that such a conditioning method leads to training instability, consistent with [3]. Instead, we re-render the template mesh under the rendered pose to get the synthetic rendering $I_{synthetic}$ and feed it into D_{dual} along with image pairs. Here, we adopt correspondence images inspired by [35]. Such concatenation encourages the final output to align with the synthetic rendering and learn the expected deformation.

3.6. Training objectives

During training, we use the non-saturating GAN loss with R1 regularization. Moreover, we adopt the density regularization proposed in EG3D [7]. Therefore, the total learning objective is:

$$\begin{aligned}\mathcal{L}_{D_{dual},G} =&\mathbb{E}_{z\sim p_z,\epsilon\sim p_\epsilon}[f(D_{dual}(G(z,\epsilon)))]+\\ &\mathbb{E}_{I^r\sim p_{I^r}}[f(-D_{dual}(I^r)+\\ &\lambda\|\bigtriangledown D_{dual}(I^r)\|^2],\end{aligned} \tag{1}$$

$$\mathcal{L}_{density} =\sum\nolimits_{x_s\in\mathcal{S}}\|d(x_s)-d(x_s+\epsilon)\|_2, \tag{2}$$

$$\mathcal{L}_{total} =\mathcal{L}_{D_{dual},G}+\lambda_{density}\mathcal{L}_{density}, \tag{3}$$

where I^r is the combination of real images, blurred real images, and the corresponding synthetic renderings, which are sampled from the training set with distribution p_I. We adopt many training hyperparameters from EG3D and StyleGAN2 (learning rates of generator and discriminator, batch size, R1 regularization, etc.). We train our model based on the pretrained model of EG3D [7] and continue to train on 4 3090 GPUs for roughly 4 days. Please refer to the supplemental material for the implementation details.

4. Experiments

In this section, we first show qualitative and quantitative comparisons to state-of-the-art 2D / 3D animatable generative facial models (Sec. 4.1), and then discuss the conducted ablation studies of our design choices (Sec. 4.2). Furthermore, we show various applications combining our efficient 3D representation with GAN technologies such as GAN inversion and style transfer (Sec. 4.3).

Datasets. We train and test our methods on FFHQ [33]. We augment FFHQ with horizontal flips and use an off-the-shelf pose estimator [15] to label images with the approximated camera extrinsic parameters and constant intrinsics. To support full pose animation, in-plane (roll) rotation is also considered. Furthermore, we use DECA [17] to estimate the FLAME parameters of facial identity $\beta \in \mathbb{R}^{100}$, jaw pose $\theta_{jaw} \in \mathbb{R}^3$ and expression $\psi \in \mathbb{R}^{50}$. Since DECA doesn't account for eyeball movement, we additionally adopt an efficient facial detector [1] to detect 2D landmarks of irises and optimize eye poses $\theta_{eye} \in \mathbb{R}^6$ by minimizing the re-projection errors. Based on these FLAME parameters, we produce a template mesh with 5023 vertices and 9976 faces to drive facial deformations.

4.1. Comparisons

Baselines. We compare our method against two state-of-the-art methods for animatable 3D-aware image synthesis: 3DFaceShop [61], and AniFaceGAN [68]. Besides, we also select DiscoFaceGAN [13] as a baseline, which generates animatable 2D portraits conditioned on 3DMM parameters.

Qualitative comparison. Fig. 4 provides a qualitative comparison against baselines. Overall, as can be seen, our method outperforms all baselines by some margin on both synthesis quality and animation accuracy. Specifically, DiscoFaceGAN [13] suffers from inconsistent identity during animation. Moreover, it cannot generate reasonable mouth interior, e.g. stretched teeth. 3DFaceshop and AnifaceGAN synthesize 3D-consistent images, nevertheless, still struggle to model consistent mouth interior with the driving images. This is because their implicit deformation approaches

[1] https://mediapipe.dev/

Figure 4. Comparison with the state-of-the-art animatable 3D & 2D image synthesis methods. We extract several frames from a video clip and use the interpreted face model parameters to animate random virtual avatars.

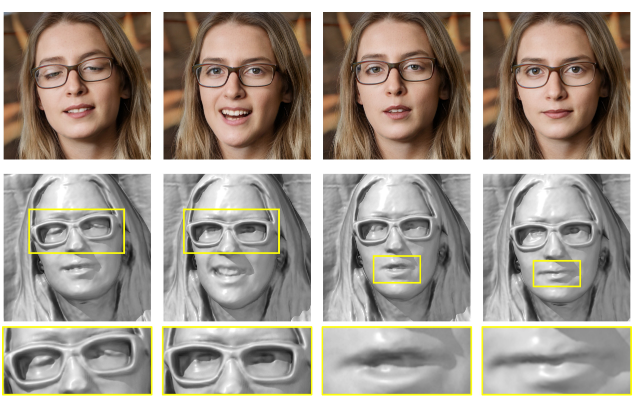

Figure 5. High-quality dynamic shapes with topological changes. As can be seen, we model detailed dynamic shapes of eyelids and lips, while keeping glasses unchanged.

are under-constrained, leading to overfit the expression bias (smiling with mouth half-opened) of datasets. Compared to the other methods, our approach not only synthesizes images with higher quality, but also preserves more detailed expressions of the driver images, including mouth interior, eye blinks and eye movements. Furthermore, we are the only method that supports in-plane head rotations. Fig. 5 provides visual examples of the synthesized high-quality geometry. Our approach can model detailed shape deformations (see zoomed eyelids and lips in Fig. 5) with topological awareness, i.e. glasses are kept unchanged.

	FID↓	AED↓	APD↓	APD*↓	ID↑
DiscoFaceGAN (256^2) [13]	17.1	0.42	0.046	0.024	0.73
AniFaceGAN (256^2) [68]	20.1	0.25	0.041	0.022	0.82
3DFaceShop (512^2) [61]	23.7	0.31	0.045	0.024	0.75
Ours (256^2)	3.7	0.16	0.018	0.013	0.85
Ours (512^2)	3.9	0.16	0.023	0.019	0.84

Table 1. Quantitative comparison using FID, average expression distance (AED), average pose distance (APD), and identity consistency (ID) for FFHQ. APD* means calculating pose distance with roll fixed.

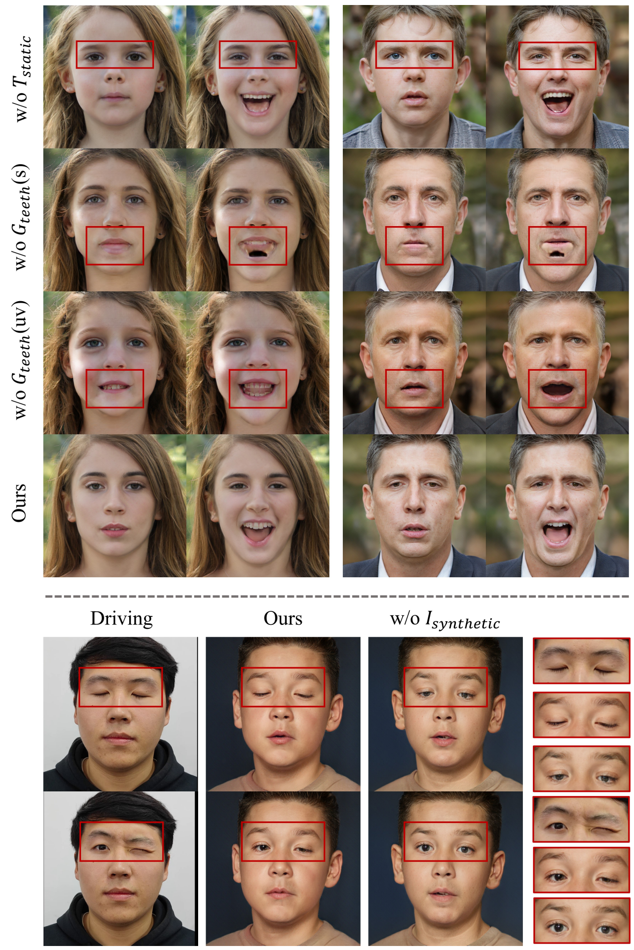

Figure 6. Ablation study on model designs. T_{static} encourages better identity consistency; G_{teeth} benefits realistic mouth interior; the discriminator with $I_{synthetic}$ input allows more consistent reconstruction of detailed expressions.

Quantitative evaluation. Tab. 1 demonstrates quantitative results comparing our method against baselines evaluated on several metrics. We measure image quality with Frechet Inception Distance (FID) [29] between the entire FFHQ dataset and 50k generated images using randomly sampled latent codes, camera poses, and FLAME parameters. Since AniFaceGAN and DiscoFaceGAN synthesize images on the resolution of 256^2, we test FID of them on FFHQ256^2. Following [3, 39], we evaluate the faithfulness of the animation with the Average Expression Distance (AED), the Average Pose Distance (APD), and identity consistency (ID). For each method, we randomly sample 500 identities and animate each with randomly sampled 20 FLAME parameters of expressions and poses. Then, we estimate the FLAME parameters for these 10000 generated images and the average distances between the driving FLAME parameters and the reconstructed ones. For identity consistency, we randomly sample 2000 poses, 2000 sets of FLAME parameters, and 1000 identities. Then we randomly select two poses and two sets of FLAME parameters for each identity, generating a total of 1000 image pairs. We calculate consistency metric using a pre-trained Arcface model [15] for each image pair and report the average result. Since the other baselines don't support in-plane (roll) head rotation, we further report APD* which only accounts for poses on yaw and pitch. Our method achieves the best performance on all metrics. Note that our model demonstrates significant improvements in FID, bringing animatable 3D GAN to the same level as unconditional 3D GANs (4.7 for EG3D [7]). For AED and APD, we also show superiority against baselines. Note that we still achieve the best pose consistency (0.019) when only considering yaw and pitch.

	FID↓	AED↓	APD↓	ID↑
w/o T_{static}	6.6	0.25	0.026	0.63
w/o $G_{teeth}(uv)$	7.2	0.37	0.042	0.71
w/o $G_{teeth}(s)$	8.4	0.32	0.036	0.79
w/o $I_{synthetic}$	**3.8**	0.18	0.025	0.74
Full	3.9	**0.16**	**0.023**	**0.84**

Table 2. Ablation study on model designs. Modeling static components by T_{static} improves identity consistency. The teeth synthesis G_{teeth} module benefits both animation and synthesis quality significantly while adding synthetic renderings $I_{synthetic}$ into discrimination takes both a step further.

4.2. Ablation study

Static tri-planes. As suggested by the grey lines in Fig. 8, this baseline removes the static tri-planes T_{static} and entangles both dynamic and static components in T_{uv}. As illustrated in the first row of Fig. 6, the identities change when varying expressions. Since there are no explicit constraints for identity consistency, the model would be prone to unexpected entanglement between expression and identity. Tab. 2 shows a similar trend where removing T_{static} leads to worse identity consistency.

Mouth Synthesis. When removing the mouth synthesis module G_{teeth}, we consider two altered choices: representing mouth features by T_{static} or T_{uv}, named w/o $G_{teeth}(s)$ (red lines in Fig. 8) and w/o $G_{teeth}(uv)$ (blue lines in Fig. 8), respectively. The first baseline, illustrated in the second row of Fig. 6, suffers from 'hole' artifacts of mouth. This is because inferior teeth move along with the jaw rotations and thus cannot be modeled by static features. The second baseline modeling teeth with T_{uv} also leads to an unreasonable mouth interior since the FLAME template doesn't account for teeth area and the neural textures for teeth would be sampled from other unrelated areas leading to artifacts. Quantitatively, both baselines without G_{teeth} show significant degradation in AED and APD.

Deformation-aware discriminator. Tab. 2 demonstrates

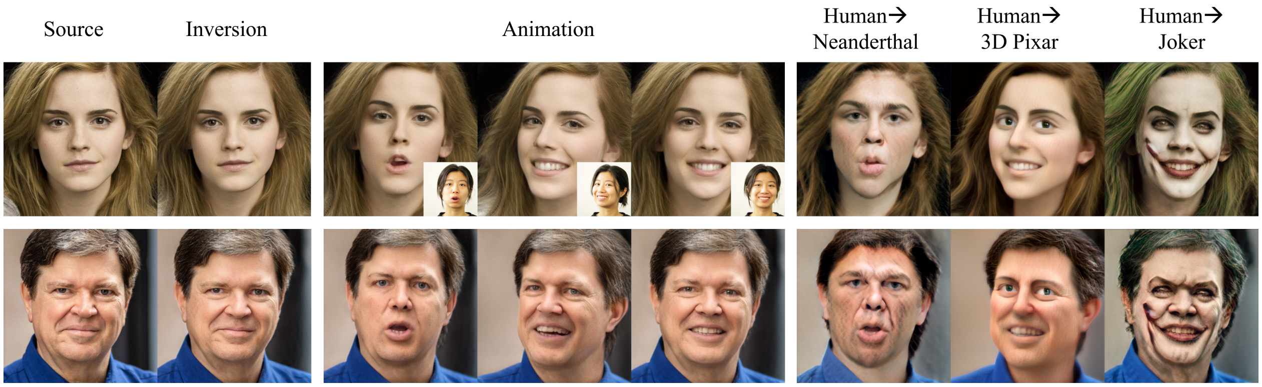

Figure 7. Applications of our model. We use PTI [53] to fit 3D-aware avatars for real portraits and animate them with sampled video clips. Furthermore, we leverage StyleGAN-NADA [21] into 3D settings and adapt these avatars into textually-prescribed domains.

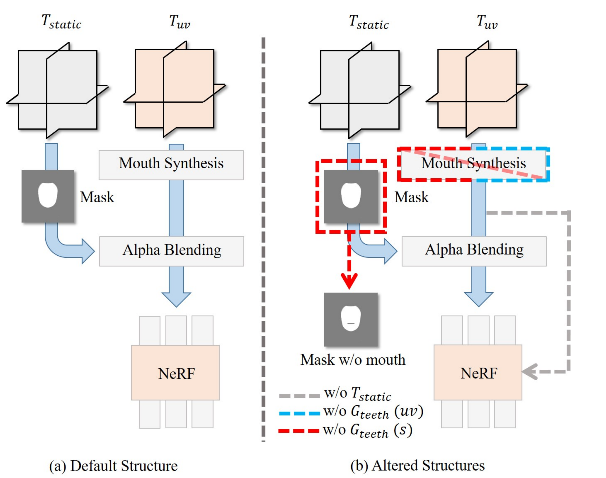

Figure 8. Illustrations of the proposed three variants of our model for ablation study.

that the deformation-aware discriminator with $I_{synthetic}$ input improves both animation accuracy and identity consistency, at the negligible expense of slightly reduced image quality. In Fig. 6, we see that this design shows a better reconstruction of detailed expressions, e.g. eye blinks.

4.3. Applications

One-shot portrait animation. Fig. 7 shows the application of our model for one-shot head avatars. The learned generative animatbale 3D representation with expressive latent space can serve as a strong 3D prior for high-fidelity single-view 3D reconstruction and animation. Note that we can generate natural and consistent animations without video data training.

Animatable 3D-aware stylization. Inspired by IDE-3D [58], we incorporate 2D CLIP-guided style tranfer methods [21] with our animatable 3D representation for 3D-aware portrait stylization. The right three columns of Fig. 7 show examples of text-driven, stylized portrait animation. Specifically, to adapt a pre-trained model through only a textual prompt, we optimize the generator with two kinds of CLIP-based guidance [21]. However, leveraging text-guided 2D methods directly into the 3D setting is challenging as it tends to break 3D awareness and deformation awareness inherited in the generator parameters. To this end, we make some necessary modifications to the training framework. Please refer to the supplemental material for details. As shown in Fig. 7, we achieve high-quality 3D-aware stylized portrait synthesis with preserving well properties (i.e. 3D consistency and accurate animation).

5. Limitations and future work

Though our approach enables reasonable extrapolation on some rare expressions, it struggles to model some others with full consistency, such as one-side mouth up, sticking tongue out, etc. Besides, the rendered portraits still suffer from slight texture flickering caused by 2D superresolution. We leave both for future work. Furthermore, our model has the potential to provide a strong 3D prior to accelerating person-specific avatar reconstruction.

6. Conclusion

We have presented Next3D, a novel animatable 3D representation for unsupervised learning of high-quality and 3D-consistent virtual facial avatars from unstructured 2D images. Our approach has pushed the frontier of photorealistic animatbale 3D-aware image synthesis. Serving as a strong 3D prior, we believe our learned 3D representation will boost a series of downstream applications including 3D-aware one-shot facial avatars and animatable out-of-domain avatar generation.

Acknowledgements

This paper is supported by National Key RD Program of China (2022YFF0902200) and the NSFC project No.62125107.

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] Hendrik Baatz, Jonathan Granskog, Marios Papas, Fabrice Rousselle, and Jan Novák. Nerf-tex: Neural reflectance field textures. In *Computer Graphics Forum*. Wiley Online Library, 2021. 3

[3] Alexander W. Bergman, Petr Kellnhofer, Wang Yifan, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. In *NeurIPS*, 2022. 2, 5, 7

[4] 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*, SIGGRAPH '99, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 2, 3

[5] Andrei Burov, Matthias Nießner, and Justus Thies. Dynamic surface function networks for clothed human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10754–10764, 2021. 3

[6] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2nerf: Unsupervised conditional p-gan for single image to neural radiance fields translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3981–3990, 2022. 2

[7] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2021. 2, 4, 5, 7

[8] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[9] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20427–20437, 2022. 2

[10] Yuedong Chen, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Sem2nerf: Converting single-view semantic masks to neural radiance fields. In *ECCV*, 2022. 2

[11] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 3

[12] Hang Dai, Nick Pears, William Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision*, 128(2):547–571, 2020. 3

[13] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Tong Xin. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 4, 5, 6

[14] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[15] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 5, 7

[16] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14398–14407, 2021. 2, 3

[17] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 5

[18] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 3

[19] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *International Conference on 3D Vision (3DV)*, 2017. 2

[20] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conferfence on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 3

[21] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 8

[22] Baris Gecer, Binod Bhattarai, Josef Kittler, and Tae-Kyun Kim. Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model. In *Proceedings of the European conference on computer vision (ECCV)*, pages 217–234, 2018. 3

[23] 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*, pages 1155–1164, 2019. 3

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[25] 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

[26] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5151–5160, 2021. 3

[27] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*, 2022. 2, 5

[28] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2019. 2

[29] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 7

[30] Fangzhou Hong, Zhaoxi Chen, Yushi LAN, Liang Pan, and Ziwei Liu. EVA3d: Compositional 3d human generation from 2d image collections. In *International Conference on Learning Representations*, 2023. 2

[31] 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. 3

[32] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 3

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

[34] 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 Conference on Computer Vision and Pattern Recognition*, 2020. 4

[35] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2, 3, 5

[36] Jeong-gi Kwak, Yuanming Li, Dongsik Yoon, Donghyeon Kim, David Han, and Hanseok Ko. Injecting 3d perception of controllable nerf-gan into stylegan for editable portrait image synthesis. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 2

[37] Henning Lange and J Nathan Kutz. Fc2t2: The fast continuous convolutional taylor transform with applications in vision and graphics. *arXiv preprint arXiv:2111.00110*, 2021. 3

[38] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 3, 4

[39] Connor Z Lin, David B Lindell, Eric R Chan, and Gordon Wetzstein. 3d gan inversion for controllable portrait image animation. *arXiv preprint arXiv:2203.13441*, 2022. 2, 7

[40] 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. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[41] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[42] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

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

[44] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 2

[45] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4531–4540, 2019. 3

[46] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, 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 (CVPR)*, 2022. 2, 5

[47] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 34:20002–20013, 2021. 2

[48] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[49] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020. 3

[50] Hanspeter Pfister, Matthias Zwicker, Jeroen Van Baar, and Markus Gross. Surfels: Surface elements as rendering primitives. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 335–342, 2000. 3

[51] Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William AP Smith, Baris Gecer, and Stefanos Zafeiriou. Towards a complete 3d morphable model of the human head. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4142–4160, 2020. 3

[52] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021. 2, 3

[53] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 8

[54] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[55] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[56] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 3

[57] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[58] 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. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022. 2, 5, 8

[59] 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*, 2022. 2

[60] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and HongSheng Li. Controllable 3d face synthesis with conditional generative occupancy fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[61] Junshu Tang, Bo Zhang, Binxin Yang, Ting Zhang, Dong Chen, Lizhuang Ma, and Fang Wen. Explicitly controllable 3d-aware portrait generation. *arXiv preprint arXiv:2209.05434*, 2022. 2, 4, 5, 6

[62] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[63] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. *Computer Graphics Forum*, 41(2):703–735, 2022. 3

[64] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3, 4

[65] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2, 3

[66] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20333–20342, 2022. 4

[67] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 2

[68] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Chen Qifeng, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. In *Advances in Neural Information Processing Systems*, 2022. 2, 4, 5, 6

[69] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[70] Sicheng Xu, Jiaolong Yang, Dong Chen, Fang Wen, Yu Deng, Yunde Jia, and Xin Tong. Deep 3d portrait from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7710–7720, 2020. 2, 3

[71] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18430–18439, 2022. 2

[72] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe hd: A high-resolution 3d-aware generative model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[73] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pretrained stylegan. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 2

[74] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. Avatargen: a 3d generative model for animatable human avatars. *arXiv preprint arXiv:2208.00561*, 2022. 2

[75] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Fdnerf: Few-shot dynamic neural radiance fields for face reconstruction and expression editing. *arXiv preprint arXiv:2208.05751*, 2022. 3

[76] Jichao Zhang, E. Sangineto, Hao Tang, Aliaksandr Siarohin, Zhun Zhong, N. Sebe, and Wei Wang. 3d-aware semantic-guided generative model for human synthesis. In *ECCV*, 2022. 2

[77] Jichao Zhang, Aliaksandr Siarohin, Yahui Liu, Hao Tang, Nicu Sebe, and Wei Wang. Training and tuning generative neural radiance fields for attribute-conditional 3d-aware face generation. *arXiv preprint arXiv:2208.12550*, 2022. 2

[78] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Pan Pan, and Yi Yang. Multi-view consistent generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[79] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G. Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 2

[80] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13545–13555, 2022. 3

[81] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv:2110.09788*, 2021. 2

[82] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[83] Peiye Zhuang, Liqian Ma, Oluwasanmi Koyejo, and Alexander Schwing. Controllable radiance fields for dynamic face synthesis. In *Proc. 3DV*, 2022. 2