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

Jingxiang Sun1 Xuan Wang2 Lizhen Wang1 Xiaoyu Li3 Yong Zhang3
Hongwen Zhang1 Yebin Liu1
1Tsinghua University 2Ant Group 3Tencent AI Lab

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.

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 synthesiz-
ing 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 [58] 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. To our knowledge, we are the first method to incorporate Neural Textures into animatable 3D-aware synthesis.
- 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 [58]. These approaches either don’t allow for topology changes or need
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 retain 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 $G_{\text{texture}}$. 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 gain 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. 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

Since the FLAME template doesn’t contain inner mouth, we propose a teeth synthesis module, $G_{\text{teeth}}$, 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 $64 \times 64$. Then, the stacked mouth features are processed by $G_{\text{teeth}}$ which employs a style-modulated UNet [66]. The downsampling process of $G_{\text{teeth}}$ encodes $f_{\text{teeth}}$ into multi-scale feature maps which serve as content conditions for the following StyleGAN layers. The output teeth features $f_{\text{teeth}}$ 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, $T_{\text{static}}$, which is generated by a StyleGAN2 CNN generator $G_{\text{static}}$, sharing the same latent code with $G_{\text{texture}}$. The plane features of $T_{uv}$ and $T_{\text{static}}$ 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.

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 $w$.
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 \( \sigma \) 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 \( \sigma \) 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 \( 64 \times 64 \) and \( 512 \times 512 \), 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 \( D_{\text{dual}} \) 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:

\[
\mathcal{L}_{\text{D}} = \mathbb{E}_{x \sim p_{x}} \left[ f(D_{\text{dual}}(G(z, e))) \right] + \mathbb{E}_{I^{r} \sim p_{I^{r}}} \left[ f(-D_{\text{dual}}(I^{r})) + \lambda \| \nabla D_{\text{dual}}(I^{r}) \|_{2}^{2} \right],
\]

\[
\mathcal{L}_{\text{density}} = \sum_{x, e \in \mathcal{S}} \| d(x) - d(x + e) \|_{2},
\]

\[
\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{D}} + \lambda \mathcal{L}_{\text{density}},
\]

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 \( p_{I^{r}} \). 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 are under-constrained, leading to overfit the expression bias (smiling with mouth half-opened) of datasets. Compared to

\(^1\)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.

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>FID</th>
<th>AED</th>
<th>APD</th>
<th>APD*</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiscoFaceGAN (256(^2)) [13]</td>
<td>17.1</td>
<td>0.42</td>
<td>0.046</td>
<td>0.024</td>
<td>0.73</td>
</tr>
<tr>
<td>AniFaceGAN (256(^2)) [68]</td>
<td>20.1</td>
<td>0.25</td>
<td>0.041</td>
<td>0.022</td>
<td>0.82</td>
</tr>
<tr>
<td>3DFaceShop (512(^2)) [61]</td>
<td>23.7</td>
<td>0.31</td>
<td>0.045</td>
<td>0.024</td>
<td>0.75</td>
</tr>
<tr>
<td>Ours (512(^2))</td>
<td>3.9</td>
<td>0.16</td>
<td>0.023</td>
<td>0.019</td>
<td>0.84</td>
</tr>
</tbody>
</table>

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.

Quantitative evaluation. Tab. 1 demonstrates quantita-
Figure 6. Ablation study on model designs. \( T_{\text{static}} \) encourages better identity consistency; \( G_{\text{teeth}} \) benefits realistic mouth interior; the discriminator with \( I_{\text{synthetic}} \) input allows more consistent reconstruction of detailed expressions.

<table>
<thead>
<tr>
<th></th>
<th>FID ↓</th>
<th>AED ↓</th>
<th>APD ↓</th>
<th>ID ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o ( T_{\text{static}} )</td>
<td>6.6</td>
<td>0.25</td>
<td>0.026</td>
<td>0.63</td>
</tr>
<tr>
<td>w/o ( G_{\text{teeth}}(uv) )</td>
<td>7.2</td>
<td>0.37</td>
<td>0.042</td>
<td>0.71</td>
</tr>
<tr>
<td>w/o ( G_{\text{teeth}}(s) )</td>
<td>8.4</td>
<td>0.32</td>
<td>0.036</td>
<td>0.79</td>
</tr>
<tr>
<td>w/o ( I_{\text{synthetic}} )</td>
<td>3.8</td>
<td>0.18</td>
<td>0.025</td>
<td>0.74</td>
</tr>
<tr>
<td>Ours</td>
<td>3.9</td>
<td>0.16</td>
<td>0.023</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Table 2. Ablation study on model designs. Modeling static components by \( T_{\text{static}} \) improves identity consistency. The teeth synthesis \( G_{\text{teeth}} \) module benefits both animation and synthesis quality significantly while adding synthetic renderings \( I_{\text{synthetic}} \) 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 \( T_{\text{static}} \) 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 \( T_{\text{static}} \) leads to worse identity consistency.

**Mouth Synthesis.** When removing the mouth synthesis module \( G_{\text{teeth}} \), we consider two altered choices: representing mouth features by \( T_{\text{static}} \) or \( T_{uv} \), named w/o \( G_{\text{teeth}}(s) \) (red lines in Fig. 8) and w/o \( G_{\text{teeth}}(uv) \) (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 \( G_{\text{teeth}} \) show significant degradation in AED and APD.

**Deformation-aware discriminator.** Tab. 2 demonstrates that the deformation-aware discriminator with \( I_{\text{synthetic}} \) in-
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.

put 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 exploration of rare 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 animatable 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 transfer 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 (e.g. eye blinks, pouting, etc.), it struggles to model some other challenging expressions with full consistency, such as one-side mouth up, frown, sticking tongue out, etc. We could use high-quality video clips with more abundant expressions for training as well as a more powerful face model for better extrapolation. We leave it for future work. Furthermore, our model has the potential to provide a strong 3D prior for accelerating person-specific avatar reconstruction. Besides, extending our methods into full-body settings is also a promising direction.

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 animatable 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.
References


A. Additional experiments

A.1. Deformation-aware discriminator

We propose a deformation-aware discriminator which additionally takes the synthetic renderings as input. Furthermore, we also take experiments on the parameter conditioning method proposed in GNARF [3]. Specifically, we first train our model without either synthetic renderings or FLAME parameters conditioning for about two days. Then, we test two methods based on the same checkpoint and report the changing trend of FID scores for two methods in Fig. 9. The discriminator with synthetic rendering input converges to a better FID score, while the one conditioned on FLAME parameters incurs divergency. Note that we have added random noise to the FLAME parameters for better convergency following GNARF.

A.2. Training strategy of 3D-aware stylization

We conduct an ablation study on two strategies for freezing layers of the generator during 3D-aware stylization. The first one is the default setting following StyleGAN-NADA [21] that freezes all toRGB layers in the synthesis network. Though it works in 2D space, we found it leads to degraded image quality and dissymmetry. To this end, we adopt another strategy which optimizes the last toRGB layer for each synthesis network. In our case, there are three StyleGAN-based synthesis network including a neural texture generator $G_{uv}$, a static tri-plane generator $G_{static}$, and a teeth completing module $G_{teeth}$ so we add the last toRGB layers of these three synthesis networks into optimization. As can be seen in Fig. 10, the second strategy improves the synthesis quality.

B. Implementation details

We implemented our 3D GAN framework on top of the official PyTorch implementation of EG3D [7] ². We adopt several hyperparameters and training strategies of EG3D including blurred real images at the beginning, pose-conditioned generator, density regularization, learning rates of the generator and discriminator. Due to the limitation of computing material, we drop the two-stage training strategy and fix the neural rendering resolution to 64 and the final resolution to 512 instead.

B.1. Data preprocessing

We use FLAME template model to drive the facial deformation and use DECA [17] to extract FLAME parameters. Since there is no suitable model to accurately extract eye poses, we optimize eye poses with an off-the-shelf landmark detector ³. Specifically, the detector extracts five landmarks around the eyes, as shown in Fig. 11. Accordingly, we select five vertices on the template mesh and the optimizable variables of eye poses are yaw and pitch. To optimize eye poses of a given portrait image, we minimize the re-projection errors of the vertices and detected landmarks by the PyTorch-implemented gradient descent. Since the FLAME template mesh has a different scale to the pre-trained EG3D model, we initially rescale the template by 2.5 for a coarse visual alignment and fine-tune the translation and scale during training.

B.2. Generator

Our generator introduces a style-unet-based teeth completing module $G_{teeth}$ whose architecture is illustrated in Fig. 12. The left part encodes the concatenated tri-plane teeth textures with dimensions of 768 (256 × 3) into multiscale feature maps ranging from $64^2$ to $8^2$. Then the feature map with a resolution of $8^2$ is processed into the residual blocks and fed into the right generator as the input feature map. Finally, the generator outputs a $64 \times 64 \times 768$ feature map.

C. Experiment details

Inversion-based one-shot facial avatars. We use an off-the-shelf face detector [15] to extract camera poses and crop the portraits in the wild to be consistent with the trainingset.

²https://github.com/NVlabs/eg3d
³https://mediapipe.dev/
We further extract the FLAME parameters and obtain the template mesh for each image by DECA [17]. Following Pivotal Tuning Inversion (PTI) [53], we first optimize the latent code for 450 iterations and then fine-tune the generator weights for an additional 500 iterations.

**3D-aware stylization.** Following StyleGAN-NADA [21], we optimize partial generator weights with others fixed. In practice, we fixed all toRGB layers of the synthesis blocks except for the last ones for the texture generator and static generator. We also fix the NeRF decoders for preventing the 3D consistency from degeneration.

### D. Additional visual results

In this section, we provide additional visual results as a supplement to the main paper. Fig. 13 provides selected examples of four certain expressions and poses, highlighting the image quality, expression controllability (e.g. gaze animation), and the diversity of outputs produced by our method. Fig. 14 provides a qualitative comparison against baselines on facial animation.

Fig. 15 provides more results of animated virtual avatars with high-quality shapes. Note that the motions of eyelids can be reflected on the extracted meshes. Furthermore, the eyes are modeled as convex, suggesting that “hollow face illusion” is alleviated. This is because while the gaze directions are highly pose-related, the rotated eyeballs in the template mesh provide an explicit gaze direction signal and thus helps to model such pose-related attribute and decouple them during inference.

Finally, we show additional results of the applications of our methods including one-shot avatars for real portraits and 3D-aware stylization in Fig. 16. We encourage readers to view the accompanying supplemental video for the dynamic results.
Figure 13. Generated examples with selected expressions and poses.
Figure 14. Qualitative comparison against baselines.
Figure 15. Animated virtual avatars with high-quality shapes.
Figure 16. Visual results of one-shot avatars for real portraits and 3D-aware stylization.