Animatable 3D-Aware Face Image Generation for Realistic Video Avatars

Anonymous Author(s)
Affiliation
Address
email

Abstract

Face image generation and animation have been a longstanding task. Although many 2D generative models yield excellent manipulations in 2D space, they often suffer from 3D inconsistency and undesirable artifacts when rendering from different camera viewpoints, and thus are not suitable for animations in video. Recently, 3D-aware GANs extend 2D GANs by using underlying 3D representations. Although these methods can preserve the 3D consistency across different viewpoints, they cannot achieve fine-grained control over attributes, most importantly, facial expression. In this paper, we propose an animatable 3D-aware face image generation method. Our framework mainly consists of a template implicit field and a 3D deformation field. The template field represents the canonical space and is shared across the same identity. Different expressions can be generated by deforming the manifolds in the target space to the canonical space. We enforce the generation to follow a prior 3D face parametric model by incorporating 3D-level imitative learning to encourage the deformation field to follow 3D prioris. Experiments show our method can produce high-quality animatable video avatars with strong visual 3D consistency.

1 Introduction

Face image synthesis and animation have been a longstanding task in computer vision and computer graphics with a wide range of applications such as virtual avatars and video conferencing. Remarkable progress has been achieved in recent years with a large volume of methods proposed [50, 2, 19, 57, 52, 68, 58, 56, 63, 1, 67, 69, 46, 73]. This progress is hinged on a number of advances in machine learning within which generative adversarial networks (GANs) are arguably the core underpinning.

However, most existing face GANs are based on 2D convolutional neural networks (CNNs) and do not model the underlying 3D facial geometry. When synthesizing faces under different poses and expressions, the results cannot maintain strict 3D consistency. Consequently, these methods can be used in interactive face manipulation but are not suitable for high-quality face video generation and animation. Some methods incorporate 3D prioris into the generation process [11, 64, 30, 20, 51] for better 3D rigging over the 2D images. Perhaps the most relevant work to ours is DiscoFaceGAN [11], which considers an unconditional and disentangled generative modeling setup as we do. Still, exact 3D consistency cannot be guaranteed by the aforementioned approaches, and their CNN generators often lead to temporal artifacts such as flickering and texture stitching [25], which are undesirable for realistic video avatars.

Recently, a number of 3D-aware GANs are proposed with 3D representations [37, 31, 38, 10, 14, 22, 12, 9, 41]. Among them, methods that generate neural radiance fields (NeRF) [35] have demonstrated striking image synthesis results [10, 22, 12, 9]. Owning to its 3D field modeling and volumetric

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

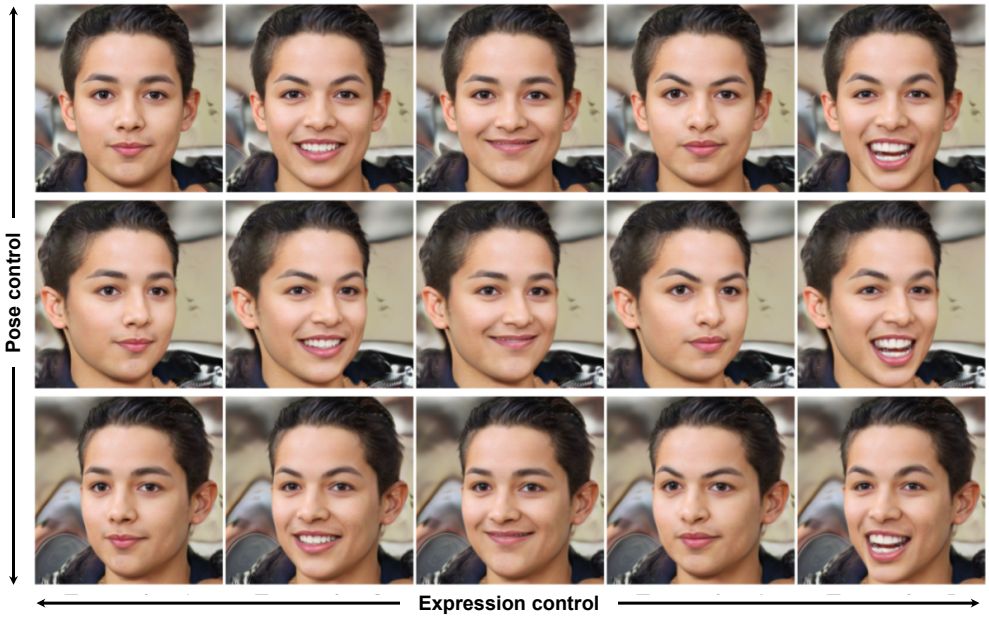

Figure 1: Multiview images with different expressions generated by our AniFaceGAN. Note the high texture consistency across poses and expressions.

rendering scheme, the NeRF representation is capable of producing realistic images with strong 3D consistency across different views, rendering it suitable for high-quality 3D scene synthesis.

In this paper, we present *AniFaceGAN*, an animatable 3D-aware face generation method that is used to synthesize realistic face images with controllable pose and expression sequences. Our method is an unconditional generation model that can generate novel identities and is trained on unstructured 2D face image collections without any 3D or multiview data. To achieve animation, we disentangle expression in the latent representation and attain explicit controllability by incorporating prioris of 3D parametric faces models. To ensure 3D consistency under expression change, we leverage 3D deformation to derive the desired expressions. Although deformation has been used in some recent NeRF methods [53, 44, 16, 45], these works mostly focused on modeling single dynamic scenes from videos. How to learn deformation in a generative setting from unstructured 2D images and how to achieve explicit and accurate expression control through deformation with unsupervised learning remains underexplored.

Our AniFaceGAN generates two 3D fields for face image rendering: a template radiance field for the geometry and appearance of a generated identity and an expression-driven deformation field. The former is based on the recent generative radiance manifold (GRAM) approach that shows state-of-the-art 3D-aware image generation quality [12]. We incorporate a 3D face morphable model (3DMM) [5, 47] into training and enforce our expression deformation to imitate it. In contrast to previous methods such as [11] which imposes imitations on 2D images, we propose a set of 3D-space losses defined on both geometry and expression deformation. We train our method on the FFHQ dataset [26] and show that it can generate high quality and 3D consistent images across different poses and expressions (an example is shown in Fig. 1) . We believe it paves the way for realistic face video generation and animation.

The contributions of our work can be summarized as follows:

- We present an animatable 3D-aware face GAN with expression variation disentangled through our proposed expression-driven 3D deformation fields.
- We introduce a set of novel 3D-space imitative losses, which align our expression generation with a prior 3D parametric face model to gain accurate and semantically-meaningful control of facial expression.
- We show that our method can generate realistic portrait videos with high visual quality. We believe it opens a new door to face video synthesis and photorealistic virtual avatar creation.

2 Related work

Neural implicit scene representations. A plenty of works [42, 34, 60, 59, 35, 39] leverage neural implicit functions to represent 3D scenes in a continuous domain. Among them, NeRF [35, 3] shows its superiority at modeling complex scenes with detailed appearance and synthesizing multi-view images with strict 3D consistency. The original NeRF and most of its successors [43, 33, 45, 49, 48, 40, 32, 70] focus on learning scene-specific representation using a set of posed images or a video sequence of a static or dynamic scene. A few methods [55, 10, 38, 72, 22, 12, 9] have explored the generative modeling task using in-the-wild images as training data. A very recent method GRAM [12] constrains radiance field learning on a set of 2D manifolds and shows promising results for high-quality and multi-view consistent image generation. Our method is also built upon the radiance manifolds representation [12] for high-quality face image generation and animation.

3D-aware generative model for image synthesis. Generative adversarial networks [21, 26, 27] are widely used for realistic 2D image synthesis. Recent methods [36, 62, 37, 31, 38, 10, 14, 22, 12, 9, 41] extend GANs to 3D-aware image synthesis by incorporating 3D knowledge into their generators. This enables them to learn multi-view image generation given only unconstrained 2D image collections as supervisions. For example, HoloGAN [36] utilizes 3D-CNN to generate low-resolution voxel features and projects them as 2D feature maps for further neural rendering. GRAF [55] adopts a generative radiance field as scene representation and generates images via volumetric rendering [35]. GRAM [12] further sparsifies the radiance field as a set of radiance manifolds and leverages manifold rendering [74, 12] for more realistic image generation. Although 3D-aware GANs are able to control camera viewpoints, they cannot achieve fine-grained control over the generated instances such as shapes and expressions for human faces. This prevents them from being used for multi-view face animation generation tasks. In this work, we introduce imitative learning of the 3D Morphable Model into a 3D-aware GAN framework to achieve expression-controllable face image generation.

Face image synthesis with 3D morphable model guidance. 3D Morphable Models (3DMMs) [5, 47, 7] play an important role for face image and animation synthesis [66, 28, 71, 54]. Earlier works [65, 66] directly render 3DMM textured meshes using traditional rendering pipeline for face reenactment. However, these methods often produce over-smooth textures and cannot generate non-face regions due to the limitation of 3DMM. Later works [28, 18, 15, 71, 8] apply refinement networks on top of the rendered 3DMM images to generate more realistic texture details and fill the holes. Nevertheless, these methods still rely on a rendering procedure to synthesize coarse images at inference time, which complicates the whole system. Recently, DiscoFaceGAN [11] proposes an imitative-contrastive learning scheme that enforces the generative network to mimic the rendering process of 3DMM. After training, it can directly generate realistic face images and control multiple face attributes in a disentangled manner via the learned 2D GAN. Some concurrent and following works [64, 30, 20, 51] of it share a similar spirit. However, these methods often encounter 3D-inconsistency issues when controlling camera poses due to the non-physical image rendering process of 2D GANs. Some recent methods [16, 24] incorporate 3DMM knowledge into NeRF's training process to achieve animatable face image synthesis. However, they do not leverage a generative training paradigm [21] and require video sequences or multiview images for supervision.

Face editing and animation with GANs. A large amount of works adopt GANs for face image manipulation and animation synthesis [50, 2, 19, 57, 52, 68, 58, 56, 63, 1, 67, 69, 46, 17, 73]. However, these methods do not offer a generative modeling of face identities thus are difficult to generate virtual subjects. They also suffer from 3D-inconsistency and texture flickering issues when changing camera poses due to using 2D CNN as an image renderer.

3 Approach

Our AniFaceGAN is a generative model trained on a collection of unstructured single face images. The generator produces two 3D fields, namely a template radiance field and an expression-driven 3D deformation field. Figure 2 shows the overall pipeline of our method. The inputs of our model include an identity code $\boldsymbol{z}_{id} \in \mathbb{R}^{d_i}$, an expression code $\boldsymbol{z}_{exp} \in \mathbb{R}^{d_e}$, a noise $\boldsymbol{\varepsilon} \in \mathbb{R}^{d_\varepsilon}$ controlling other attributes such as lighting and background, and a camera pose $\boldsymbol{\theta} \in \mathbb{R}^{d_\theta}$. We incorporate the 3D priori

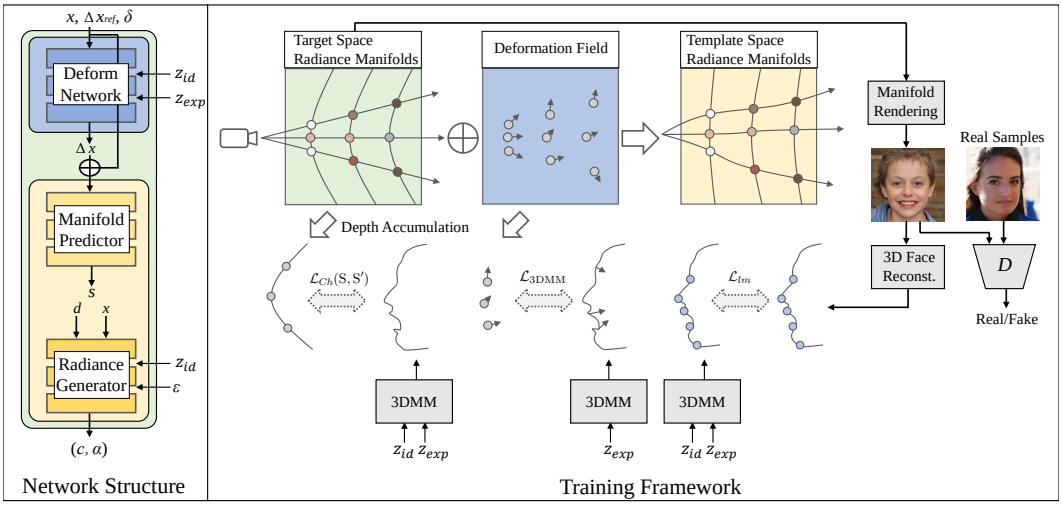

Figure 2: Our proposed framework with a template radiance field and an expression-driven deformation field for animatable 3D-aware face image generation.

of a 3DMM model [47] into our generative modeling. With 3DMM, a 3D face S is parameterized as

$$\mathrm{S} = \mathrm{S}(\alpha, \beta) = \bar{\mathrm{S}} + \mathrm{B}_{\mathrm{id}}\alpha + \mathrm{B}_{\mathrm{exp}}\beta, \tag{1}$$

where B_{id} and $\mathrm{B}_{\mathrm{exp}}$ are the PCA bases for identity and expression, α, β are the identity and expression coefficients, and $\bar{\mathrm{S}}$ is the averaged face shape. Our latent spaces of identity and expression are aligned with the 3DMM model to facilitate our 3D imitation learning (to be described later) and expression control. During GAN training, we sample $\boldsymbol{z}_{id}$ and $\boldsymbol{z}_{exp}$ from prior distributions estimated from real images, which is similar to [11].

3.1 Template radiance field

The template radiance field determines the geometry and appearance of the generated identity, and is defined as a canonical space with a neutral expression. It is modeled by a template radiance field generator G, which takes a 3D point $\boldsymbol{x}$, identity code $\boldsymbol{z}_{id}$, noise code ε, and view direction $\boldsymbol{d}$ as input and output the color $\boldsymbol{c}$ and occupancy α of $\boldsymbol{x}$:

$$G : (\boldsymbol{x}, \boldsymbol{z_{id}}, \boldsymbol{\varepsilon}, \boldsymbol{d}) \in \mathbb{R}^{d_i + d_\varepsilon + 6} \rightarrow (\boldsymbol{c}, \alpha) \in \mathbb{R}^4. \tag{2}$$

In this work, we use the generative radiance manifold (GRAM) [12] approach to generate our template radiance field. GRAM is a state-of-the-art 3D-aware GAN that can generate high-quality images with strong mutliview consistency by regularizing radiance learning and sampling on a set of learned surface manifolds. In theory, our template field can be generated by any 3D-aware generative models. We modify the generator of GRAM by changing its input to ours and train it from scratch within our whole framework.

3.2 Expression-driven 3D deformation field

The 3D deformation field is used to generate different expressions by deforming the 3D neural face in the template radiance field. A deformation network F is applied to represent it.

In practice, F models an "inverse" deformation, which deforms a 3D point in target space to the template space. The input to F includes a target-space 3D point $\boldsymbol{x}$ and expression code $\boldsymbol{z}_{exp}$, and the output is a displacement vector representing the 3D offset from the target space to template space. Since identity and expression are correlated in 3DMM (see their linear combination in Eq. 1), we also add $\boldsymbol{z}_{id}$ to the input of G. To further leverage the guidance from 3DMM, we compute a "reference" offset using the corresponding 3DMM face. Specifically, for each 3D point $\boldsymbol{x}$, we first find its nearest point on the 3DMM mesh $\boldsymbol{x}_{ref}$, and then compute its deformation to the neural face according to the 3DMM model:

$$\Delta\boldsymbol{x}_{ref} = \mathrm{3DMMDeform}(\boldsymbol{x}_{ref}, \boldsymbol{z}_{id}, \boldsymbol{z}_{exp}) = -\big(\mathrm{S}(\boldsymbol{z}_{id}, \boldsymbol{z}_{exp})(\boldsymbol{x}_{ref}) - \mathrm{S}(\boldsymbol{z}_{id}, \boldsymbol{0})(\boldsymbol{x}_{ref})\big). \tag{3}$$

In addition, we compute the distance between $\boldsymbol{x}$ and $\boldsymbol{x}_{ref}$ as $\delta = \|\boldsymbol{x} - \boldsymbol{x}_{ref}\|_2$, which serves as a measure of confidence. We feed Δx_{ref} and $\boldsymbol{\delta}$ into F to provide guidance for 3D deformation reasoning. In summary, F can be written as:

$$F : (\boldsymbol{x}, \boldsymbol{z}_{id}, \boldsymbol{z}_{exp}, \Delta\boldsymbol{x}_{ref}, \delta) \in \mathbb{R}^{d_i+d_e+7} \rightarrow \Delta\boldsymbol{x} \in \mathbb{R}^3. \tag{4}$$

3.3 Image rendering

Our rendering follows a volume rendering paradigm tailored to the radiance manifold rendering of [12]. For each viewing ray $\boldsymbol{r}$ in the target space, we uniformly sample N points $\{\boldsymbol{x}_i\}$ along the ray, then deform them to the template space by F as $\{\boldsymbol{x}'_i\}$. Then we calculate the M intersection points between the (deformed) ray and template radiance manifold of GRAM, denoted as $\{\boldsymbol{x}^*_j\}$. A differentiable intersection calculation method is used as in [12]. Finally, we obtain the color and occupancy of $\{\boldsymbol{x}^*_j\}$ with G, and composite the output color via the following rendering equation [40, 74, 12]:

$$C(\boldsymbol{r}) = \sum_{j=1}^{M} T(\boldsymbol{x}^*_j)\alpha(\boldsymbol{x}^*_j)\boldsymbol{c}(\boldsymbol{x}^*_j) = \sum_{j=1}^{M} \prod_{k<j} (1 - \alpha(\boldsymbol{x}^*_k))\alpha(\boldsymbol{x}^*_j)\boldsymbol{c}(\boldsymbol{x}^*_j). \tag{5}$$

3.4 3D-space imitation learning

We employ a set of 3D-space losses to enforce the geometry and expression deformation to imitate the 3DMM model.

Dense geometry imitation. First, we enforce the similarity between the underlying 3D geometry of our generated instance and its corresponding 3DMM face. To achieve this, we first extract a depth map in the target space. For each viewing ray $\boldsymbol{r}$ with point sample $\boldsymbol{x}_i$, its depth value is computed via

$$z(\boldsymbol{r}) = \sum_{i=1}^{N} T(\boldsymbol{x}_i) z(\boldsymbol{x}_i) \tag{6}$$

where $T(\cdot)$ is defined in Eq. 5, and the occupancy values of $\{\boldsymbol{x}_i\}$ are obtained by interpolating occupancy of $\{\boldsymbol{x}^*_j\}$. Then we reproject each pixel to 3D according to its depth, generating a point cloud S'. Since 3DMM only represents a small facial region, we employ a directed Chamfer distance [61] to align the 3DMM shape to S':

$$\mathcal{L}_{Ch}(\mathrm{S}, \mathrm{S}') = \frac{1}{|\mathrm{S}|} \sum_{\mathrm{x} \in \mathrm{S}} \min_{\mathrm{y} \in \mathrm{S}'} \|\mathrm{x} - \mathrm{y}\|_2^2 . \tag{7}$$

3D landmark imitation. We also incorporate a 3D landmark loss function to enforce the deformation field to generate desired expressions. Let I be the image generated by our model with identity code $\boldsymbol{z}_{id}$ and expression code $\boldsymbol{z}_{id}$. We use a face reconstruction network [13] to extract the 3DMM identity and expression coefficients on I, denoted as $\hat{\boldsymbol{z}}_{id}$, $\hat{\boldsymbol{z}}_{exp}$. Then the 3DMM shape can be reconstructed using $\hat{\boldsymbol{z}}_{id}$, $\hat{\boldsymbol{z}}_{exp}$. The 3D landmark loss is defined as

$$\mathcal{L}_{lm} = \|f_{lm}(\mathrm{S}(\boldsymbol{z}_{id}, \boldsymbol{z}_{exp})), f_{lm}(\mathrm{S}(\hat{\boldsymbol{z}}_{id}, \hat{\boldsymbol{z}}_{exp}))\|_2^2 , \tag{8}$$

where f_{lm} represents a function of extracting the landmarks. Here we use the 68 landmark points as in [13].

Deformation imitation. We also encourage the deformation field to follow 3DMM deformation, and employ a loss fucntion $\mathcal{L}_{\mathrm{3DMM}}$:

$$\mathcal{L}_{\mathrm{3DMM}} = \|\Delta\boldsymbol{x} - \Delta\boldsymbol{x}_{ref}\|_2^2 . \tag{9}$$

Deformation regularizations. Also, we expect the deformation of most points in the target space to be small except for regions influenced by expression change. So we employ a minimal deformation constraint via

$$\mathcal{L}_{reg} = \|\Delta x\|_2^2 . \tag{10}$$

CONFIG DiscoFaceGAN DiscoFaceGRAM Ours

Figure 3: Visual comparison between our method and others. The quality of the images synthesized by CONFIG and DiscoFaceGRAM is clearly lower than ours. Both DiscoFaceGAN and our method can generate photorealistic images. However, DiscoFaceGAN suffers from inconsistency for different expressions and poses. (**See the accompanying video for better visualization and comparison**.)

Table 1: Quantitative evaluation of different models on the FFHQ dataset. FID and KID ($\times 100$) are computed with 5K synthesized images and 5K real images. DiscoFaceGRAM is our own implementation. †: results obtained using the pre-trained models provided by the authors.

	GRAM [12]	CONFIG [30]	DiscofaceGAN [11]	DiscofaceGRAM	Ours
FID↓	19.4	52.6†	17.9	34.9	21.3
KID↓	0.64	3.38†	0.79	2.04	0.86

Moreover, to encourage the deformation fields to be smooth and avoid abrupt deformation changes, we impose a simple smoothness loss on the deformation field:

$$\mathcal{L}_{smooth} = \|\Delta \boldsymbol{x} - \Delta(\boldsymbol{x} + \xi)\|_2^2 \,, \tag{11}$$

where ξ is a small random perturbation.

Adversarial learning. Following [12], real images are randomly sampled from the training set during training. We apply a discriminator D to differentiate synthesized images from real images and apply the same adversarial loss used in [12].

4 Experiments

We evaluate our method on the FFHQ [26] dataset[1], which contains 70K high-resolution face images. Following [12], we randomly sample latent codes $\boldsymbol{z}_{id}$, $\boldsymbol{z}_{exp}$, and camera pose $\boldsymbol{\theta}$ from estimated prior distributions, and sample ε from a normal distribution. In our experiments, the Adam optimizer [29] with $\beta_1 = 0$ and $\beta_2 = 0.9$ is applied for training our model. We set the learning rate to $2e{-}5$ for the deformation network and generative radiance manifolds. For the discriminator, the learning rate is $2e{-}4$. We train our models on 8 Nvidia Tesla V100 GPUs with a batch size of 32 at the resolution of 128×128. Our model takes approximately three days for training. More implementation details can be found in the supplement.

4.1 Comparison with prior works

We compare our method with two previous face generation models, CONFIG [30] and DiscoFaceGAN [11] that achieve disentangled control over camera views and expressions. We further compare with a baseline method we call DiscoFaceGRAM, which naively combines GRAM [12] with DiscoFaceGAN's imitative-contrastive learning scheme. For CONFIG and DiscoFaceGAN, we test

[1]FFHQ is released under the Creative Commons BY-NC-SA 4.0 license; the individual human face images therein were published on Flickr by their respective authors under licenses that all allow free use for non-commercial purposes.

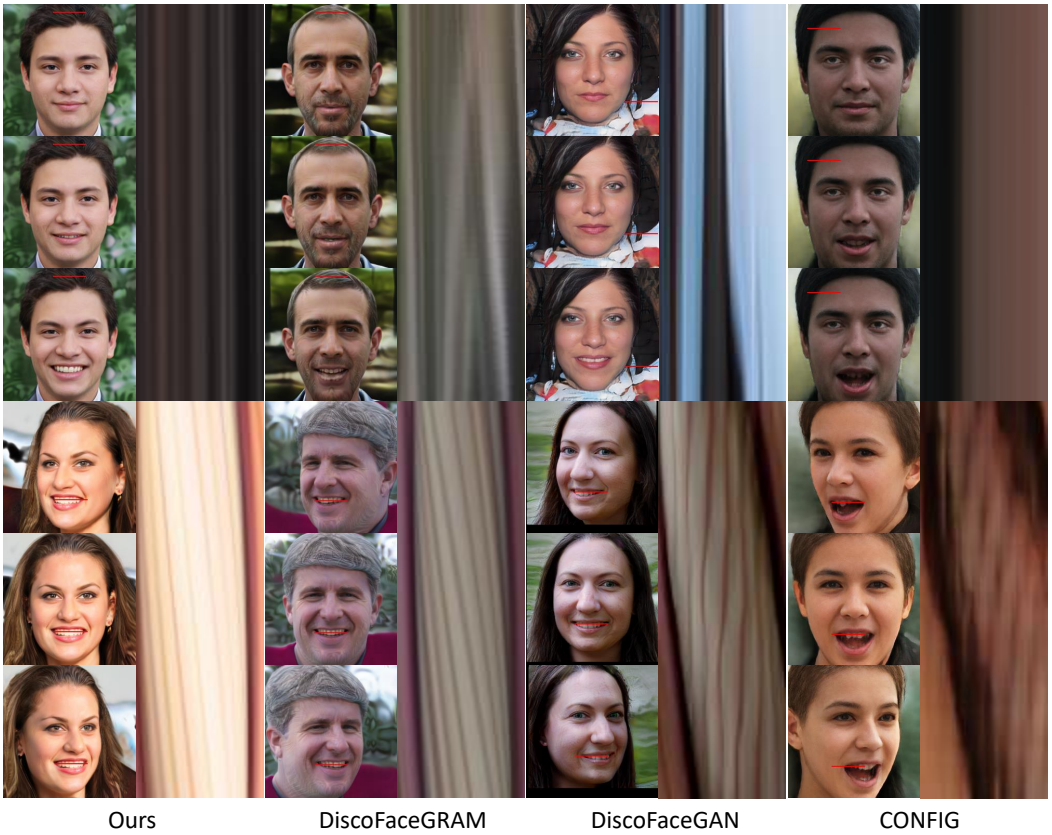

Figure 4: Comparison of consistency using a visualization method similar to [6]. For the first group, we fix the camera location and smoothly change the expression. For the second group, we use the same expression and rotate the camera smoothly. We check the texture of a line segment at a fixed 2D position. Twisted lines in the first group indicate the flickering degree of other details. The distortion/noise pattern in the second group indicates 3D-inconsistency of the rendered images.

them with the official implementations and released pretrained models. For DiscoFaceGRAM, we discard the deformation field and directly send $\boldsymbol{z}_{id}$, $\boldsymbol{z}_{exp}$ to the template field represented by GRAM. Since we do not control textures and illuminations as in [11], we only apply the imitative loss for 2D landmarks and the contrastive loss for expressions to train this baseline method.

The visual comparison is shown in Fig. 3. Both CONFIG and DiscoFaceGAN suffer from 3D-inconsistency issue (e.g. see bangs of the hairs) when varying the camera pose as these methods are based on black-box CNN renderer. By contrast, our method shows strong multiview consistency between generated images of different poses and expressions, thanks to the manifold rendering of our inherent 3D representation. A visual comparison in terms of consistency is shown in Fig. 4. The baseline method DiscoFaceGRAM cannot guarantee a fully disentanglement between identity and expression using only image-space imitative and contrastive constrains during training. Instead, our proposed AniFaceGAN leverages the expression-driven deformation field and 3D-space imitations to physically disentangle the two factors and thus demonstrates better generation quality and more reasonable expression control.

For quantitative evaluation, we evaluate the quality of generated images using the Fréchet inception distance (FID) [23] and Kernel Inception Distances (KID) [4] among 5K randomly synthesized images and 5K randomly sampled real images [12], as shown in Table 1. Compared with GRAM, our method introduces strong controllability over expression with only a slight degradation of the image quality. DiscoFaceGAN shows better image quality in terms of FID and KID, but sacrifices the 3D-consistency which is an important factor for realistic video generation. The large FID&KID gap between DiscofaceGRAM and our method further validates the effectiveness of our model and loss design.

Figure 5: Face animation with expression interpolation. (**See our video for continuous animation**.)

Table 2: Ablation study on different components on the FFHQ dataset using FID between 5K generated images and 5K real images.

	w/o F	w/o $\mathcal{L}_{Ch}$	w/o $\mathcal{L}_{lm}$	w/o $\mathcal{L}_{\text{3DMM}}$	with $\mathcal{L}_{lm}^{2d}$	Full Model
FID↓	28.3	23.9	22.4	21.8	25.5	21.3

4.2 Ablation study

We further conduct ablation studies to evaluate the efficacy of each component of our framework. Results are shown in Table 2 and Fig 6. *w/o* F represents controlling expressions by directly sending $\boldsymbol{z}_{id}$ and $\boldsymbol{z}_{exp}$ into the template GRAM without using a 3D deformation field. *w/o* $\mathcal{L}_{Ch}$ and *w/o* $\mathcal{L}_{lm}$ denote removing the Chamfer distance loss and 3D landmark loss defined in Eq. 7 and Eq. 8, respectively. *w/o* $\mathcal{L}_{\text{3DMM}}$ indicates that the deformation network F is trained without the guidance of 3DMM deformation field (Eq. 9). *with* $\mathcal{L}_{lm}^{2d}$ denotes replacing the 3D landmark loss in Eq. 8 with the 2D landmark regression loss between generated images and projected 3DMM as in [11].

As shown in Table 2, all variants degrade the quality of generated images compared to our full model. Figure 6 further demonstrates the influences on expression control when removing each of our designed components. In particular, removing the deformation field leads to a loss of disentanglement between identity and expression. Further enforcing disentanglement by adding contrastive loss (DiscoGRAM) decreases the image generation quality. Replacing our 3D space landmark loss with 2D landmark loss introduces obvious artifacts for certain expressions. We conjecture that it is due to the misalignment between the 2D position distributions of the projected 3DMMs and the faces in real images, which causes conflicts between the 2D landmark imitative loss and the adversarial loss. By contrast, our full model shows high-quality image generation results with precise control over expressions.

4.3 Expression control

Our method can precisely control facial expressions while preserving the identity and facial appearance details. As shown in Fig. 5, our method can generate consistent image sequences with interpolated expressions and changing camera views. We also compare our method with other alternative designs in terms of the ability of controlling expression. To this end, we extract expression codes from real face images using the face reconstruction network and use them to render virtual identities. As shown in Fig. 6, the DiscoFaceGRAM fails to preserve the facial appearance (e.g. hair and face color) and background, although it applies the contrastive and imitative losses in [11] to disentangle the identity and expression. For *Without* F, the 3D geometry still varies a lot under different expressions although the same identity code is given. For *With* $\mathcal{L}_{lm}^{2D}$, the performance is not as good as our full model as the error of landmark detection on 2D space and other 2D-space supervision does not necessarily constraint the 3D geometry. By contrast, our method can control expression precisely while retaining the identity.

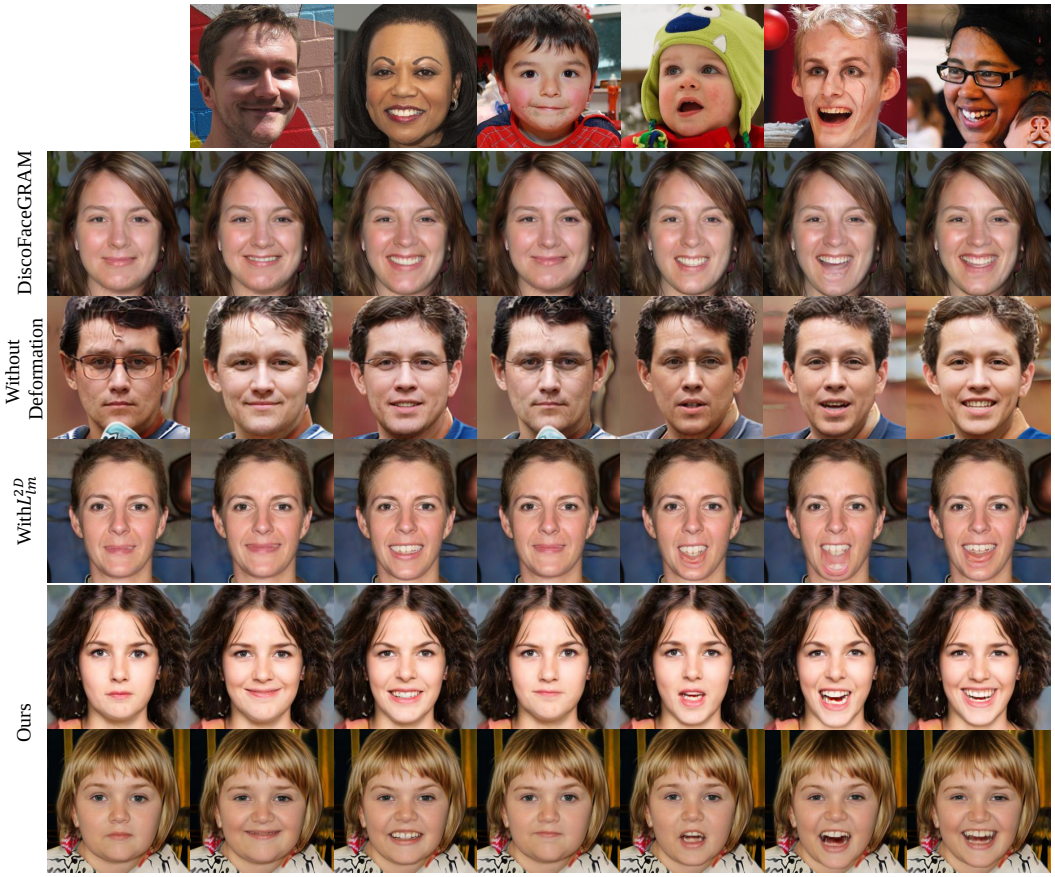

Figure 6: Visual comparison on expression control. The first row shows the real images that provide reference expressions, and other rows are the results from different methods. See text for details.

5 Conclusions

We have presented an animatable 3D-aware face image generation method named AniFaceGAN, which aims to synthesize realistic face images with controllable poses and expressions. To maintain 3D consistency across different expressions, we leverage 3D deformation to derive the desired expressions. For face image rendering, we propose to generate two 3D fields namely a template radiance field and an expression-driven deformation field. A set of novel 3D-space imitation losses are proposed to effectively train our model along with adversarial learning. The proposed AniFaceGAN is an unconditional generation model that is trained on unstructured 2D face images without any dependency on 3D or multiview data. Experimentally, our method trained on FFHQ dataset can produce high-quality animatable faces with strong visual 3D consistency across different poses and expressions. We hope our method could serve as a strong baseline for realistic face video generation and animation.

6 Broader Impact

This work aims to design an animatable 3D-aware face image generation method. There may be unpredictable failures, similar as most face image synthesis and animation methods. The consequences of failures by this algorithm are determined on the down-stream applications, and please do not use it for scenarios where failures will lead to serious consequences. The method is data driven, and the performance may be affected by the biases in the data. So please also be careful about the data collection process when using it.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40(3):1–21, 2021.

[2] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: Fine-grained image generation through asymmetric training. In *IEEE international conference on computer vision*, pages 2745–2754, 2017.

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[4] Sai Bi, Kalyan Sunkavalli, Federico Perazzi, Eli Shechtman, Vladimir G Kim, and Ravi Ramamoorthi. Deep cg2real: Synthetic-to-real translation via image disentanglement. In *International Conference on Computer Vision*, pages 2730–2739, 2019.

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999.

[6] Robert C Bolles, H Harlyn Baker, and David H Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International journal of computer vision*, 1(1):7–55, 1987.

[7] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3D morphable model learnt from 10,000 faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5543–5552, 2016.

[8] Marcel C Bühler, Abhimitra Meka, Gengyan Li, Thabo Beeler, and Otmar Hilliges. Varitex: Variational neural face textures. In *IEEE/CVF International Conference on Computer Vision*, pages 13890–13899, 2021.

[9] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. *arXiv preprint arXiv:2112.07945*, 2021.

[10] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5799–5809, 2021.

[11] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3D imitative-contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5154–5163, 2020.

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

[13] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[14] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W Taylor, and Joshua M Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[15] 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.

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

[17] Yue Gao, Fangyun Wei, Jianmin Bao, Shuyang Gu, Dong Chen, Fang Wen, and Zhouhui Lian. High-fidelity and arbitrary face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16115–16124, 2021.

[18] 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 *European conference on computer vision (ECCV)*, pages 217–234, 2018.

[19] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018.

[20] Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J Black, and Timo Bolkart. Gif: Generative interpretable faces. In *2020 International Conference on 3D Vision (3DV)*, pages 868–878. IEEE, 2020.

[21] 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.

[22] 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.

[23] 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*, pages 6626–6637, 2017.

[24] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. *arXiv preprint arXiv:2112.05637*, 2021.

[25] 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, 2021.

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[28] 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*, 37(4):1–14, 2018.

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[30] Marek Kowalski, Stephan J Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. Config: Controllable neural face image generation. In *European Conference on Computer Vision*, pages 299–315. Springer, 2020.

[31] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2020.

[32] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)*, 40(6):1–16, 2021.

[33] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021.

[34] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019.

[35] 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*, pages 405–421. Springer, 2020.

[36] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3D representations from natural images. In *IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019.

[37] Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy Mitra. BlockGAN: Learning 3d object-aware scene representations from unlabelled images. In *Advances in Neural Information Processing Systems*, 2020.

[38] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021.

[39] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020.

[40] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[41] 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. *arXiv e-prints*, pages arXiv–2112, 2021.

[42] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019.

[43] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Deformable neural radiance fields. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[44] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.

[45] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021.

[46] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021.

[47] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009.

[48] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021.

[49] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021.

[50] Guim Perarnau, Joost Van de Weijer, Bogdan Raducanu, and Jose M. Alvarez. Invertible conditional gans for image editing. In *Advances in Neural Information Processing Systems Workshop on Adversarial Training*, 2016.

[51] Jingtan Piao, Keqiang Sun, Quan Wang, Kwan-Yee Lin, and Hongsheng Li. Inverting generative adversarial renderer for face reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15628, 2021.

[52] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *European conference on computer vision*, pages 818–833, 2018.

[53] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

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

[55] 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*, 2020.

[56] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020.

[57] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *IEEE conference on computer vision and pattern recognition*, pages 821–830, 2018.

[58] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019.

[59] 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, 2020.

[60] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, pages 1121–1132, 2019.

[61] Dmitriy Smirnov, Matthew Fisher, Vladimir G. Kim, Richard Zhang, and Justin Solomon. Deep parametric shape predictions using distance fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[62] Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3d shape learning from natural images. *arXiv preprint arXiv:1910.00287*, 2019.

[63] 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.

[64] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020.

[65] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Transactions on Graphics (TOG)*, 34(6):183–1, 2015.

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

[67] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

[68] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *Advances in Neural Information Processing Systems*, 2019.

[69] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2021.

[70] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. *arXiv preprint arXiv:2201.04127*, 2022.

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

[72] Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. Generative occupancy fields for 3d surface-aware image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

[73] Yanbo Xu, Yueqin Yin, Liming Jiang, Qianyi Wu, Chengyao Zheng, Chen Change Loy, Bo Dai, and Wayne Wu. Transeditor: Transformer-based dual-space gan for highly controllable facial editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[74] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4):1–12, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Appendix.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]