OmniAvatar: Geometry-Guided Controllable 3D Head Synthesis

Hongyi Xu[1] Guoxian Song[1] Zihang Jiang[1,2] Jianfeng Zhang[1,2] Yichun Shi[1]
Jing Liu[1] Wanchun Ma[1] Jiashi Feng[1] Linjie Luo[1]
[1]ByteDance Inc [2]National University of Singapore
{hongyixu, guoxian.song, zihang.jiang, jianfeng.zhang, yichun.shi,
jing.liu, wanchun.ma, jshfeng, linjie.luo}@bytedance.com

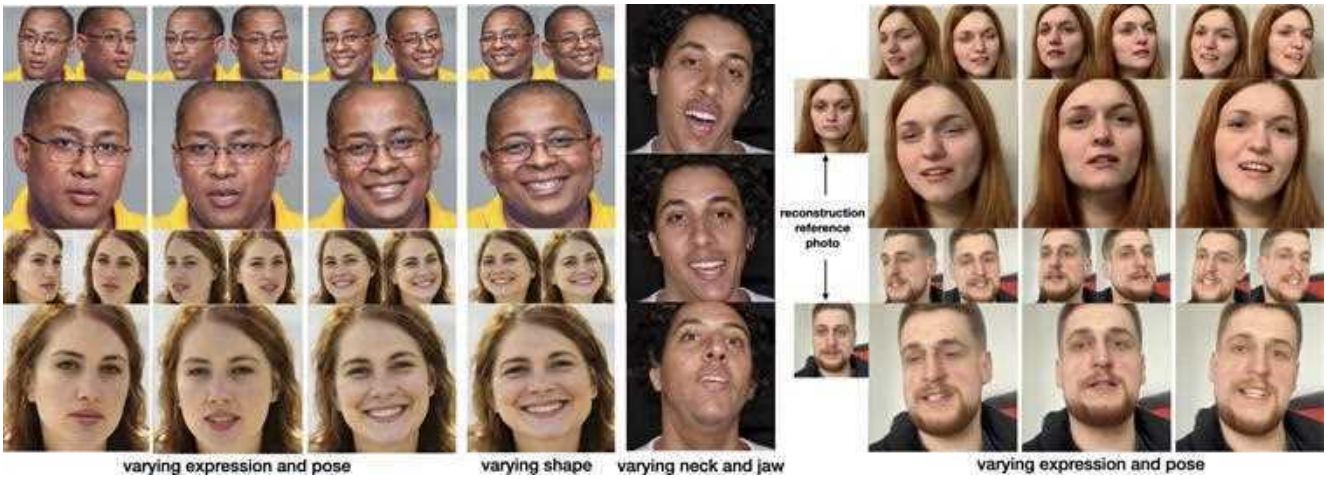

Figure 1. Our model can synthesize diverse identity-preserved 3D heads with compelling dynamic details under full disentangled control over camera poses, facial expressions, head shapes, articulated neck and jaw poses (left). Our model can also reconstruct 3D heads from a single photo reference and enable multi-view-consistent head reenactment (right).

Abstract

We present OmniAvatar, a novel geometry-guided 3D head synthesis model trained from in-the-wild unstructured images that is capable of synthesizing diverse identity-preserved 3D heads with compelling dynamic details under full disentangled control over camera poses, facial expressions, head shapes, articulated neck and jaw poses. To achieve such high level of disentangled control, we first explicitly define a novel semantic signed distance function (SDF) around a head geometry (FLAME) conditioned on the control parameters. This semantic SDF allows us to build a differentiable volumetric correspondence map from the observation space to a disentangled canonical space from all the control parameters. We then leverage the 3D-aware GAN framework (EG3D) to synthesize detailed shape and appearance of 3D full heads in the canonical space, followed by a volume rendering step guided by the volumetric correspondence map to output into the observation space. To ensure the control accuracy on the synthesized head shapes and expressions, we introduce a geometry prior loss to conform to head SDF and a control loss to conform to the expression code. Further, we enhance the temporal realism with dynamic details conditioned upon varying expressions and joint poses. Our model can synthesize more preferable identity-preserved 3D heads with compelling dynamic details compared to the state-of-the-art methods both qualitatively and quantitatively. We also provide an ablation study to justify many of our system design choices.

1. Introduction

Photo-realistic face image synthesis, editing and animation attract significant interests in computer vision and graphics, with a wide range of important downstream applications in visual effects, digital avatars, telepresence and many others. With the advent of Generative Adversarial Networks (GANs) [15], remarkable progress has been achieved in face image synthesis by StyleGAN [23–25]

as well as in semantic and style editing for face images [46, 54]. To manipulate and animate the expressions and poses in face images, many methods attempted to leverage 3D parametric face models, such as 3D Morphable Models (3DMMs) [6, 40], with StyleGAN-based synthesis models [10, 41, 53]. However, all these methods operate on 2D convolutional networks (CNNs) without explicitly enforcing the underlying 3D face structure. Therefore they cannot strictly maintain the 3D consistency when synthesizing faces under different poses and expressions.

Recently, a line of work has explored neural 3D representations by unsupervised training of 3D-aware GANs from in-the-wild unstructured images [7, 8, 11, 17, 37, 44, 45, 48, 61, 62, 67]. Among them, methods with generative Neural Radiance Fields (NeRFs) [33] have demonstrated striking quality and multi-view-consistent image synthesis [7, 11, 17, 37, 48]. The progress is largely due to the integration of the power of StyleGAN in photo-realistic image synthesis and NeRF representation in 3D scene modeling with view-consistent volumetric rendering. Nevertheless, these methods lack precise 3D control over the generated faces beyond camera pose, as well as the quality and consistency in control over other attributes, such as shape, expression, neck and jaw pose, leave much to be desired.

In this work, we present *OmniAvatar*, a novel geometry-guided 3D head synthesis model trained from in-the-wild unstructured images. Our model can synthesize a wide range of 3D human heads with full control over camera poses, facial expressions, head shapes, articulated neck and jaw poses. To achieve such high level of disentangled control for 3D human head synthesis, we devise our model learning *in two stages*. We first define a novel *semantic signed distance function* (SDF) around a head geometry (i.e. FLAME [29]) conditioned on its control parameters. This semantic SDF fully distills rich 3D geometric prior knowledge from the statistical FLAME model and allows us to build a differentiable *volumetric correspondence map* from the *observation space* to a disentangled *canonical space* from all the control parameters. In the second training stage, we then leverage the state-of-the-art 3D GAN framework (EG3D [7]) to synthesize realistic shape and appearance of 3D human heads in the canonical space, including the modeling of hair and apparels. Following that, a volume rendering step is guided by the volumetric correspondence map to output the geometry and image in the observation space.

To ensure the consistency of synthesized 3D head shape with controlling head geometry, we introduce a *geometry prior loss* to minimize the difference between the synthesized neural density field and the FLAME head SDF in observation space. Furthermore, to improve the control accuracy, we pre-train an image encoder of the control parameters and formulate a *control loss* to ensure synthesized images matching the input control code upon encoding. Another key aspect of synthesis realism is dynamic details such as wrinkles and varying shading as subjects change expressions and poses. To synthesize dynamic details, we propose to condition EG3D's triplane feature decoding with noised controlling expression.

Compare to state-of-the-art methods, our method achieves superior synthesized image quality in terms of Frechet Inception Distance (FID) and Kernel Inception Distance (KID). Our method can consistently preserve the identity of synthesized subjects with compelling dynamic details while changing expressions and poses, outperforming prior methods both quantitatively and qualitatively.

The contributions of our work can be summarized as:

- A novel geometry-guided 3D GAN framework for high-quality 3D head synthesis with full control on camera poses, facial expressions, head shapes, articulated neck and jaw poses.
- A novel semantic SDF formulation that defines the volumetric correspondence map from observation space to canonical space and allows full disentanglement of control parameters in 3D GAN training.
- A geometric prior loss and a control loss to ensure the head shape and expression synthesis accuracy.
- A robust noised expression conditioning scheme to enable dynamic detail synthesis.

2. Related Work

3D-Aware Generative Image Synthesis. Generative adversarial networks [15, 24, 25] gained popularity over the last decade due to their remarkable ability in photo-realistic image synthesis. Building on the success of 2D image-based GANs, recent works have extended the capabilities to view-consistent image synthesis with unsupervised learning from 2D single-view images. The key idea is to combine differential rendering with 3D scene representations, such as meshes [31, 50], point clouds [1, 28], voxels [34, 35, 57], and recently implicit neural representation [7, 8, 11, 17, 36, 37, 44, 48, 67]. We build our work on recent 3D GAN model by Chan et al [7] that uses an efficient triplane-based NeRF generation, combined with 2D CNN-based super-resolution. Even though 3D-aware GANs are able to control camera viewpoints, they lack precise 3D control over the other attributes such as shapes and expressions. In this work, we empower 3D-aware GANs with disentangled precise control over shapes and expressions.

Controllable Face Image Synthesis. Considerable work [9, 10, 14, 27, 41, 52, 53] has been devoted to incorporate 3D priors of statistical face models, such as 3D Morphable Models (3DMMs) [6, 40], in controllable face synthesis and animation. Among them, DiscoFaceGAN [10] proposed imitative-contrastive learning to mimic the 3DMM rendering process by the generative model.

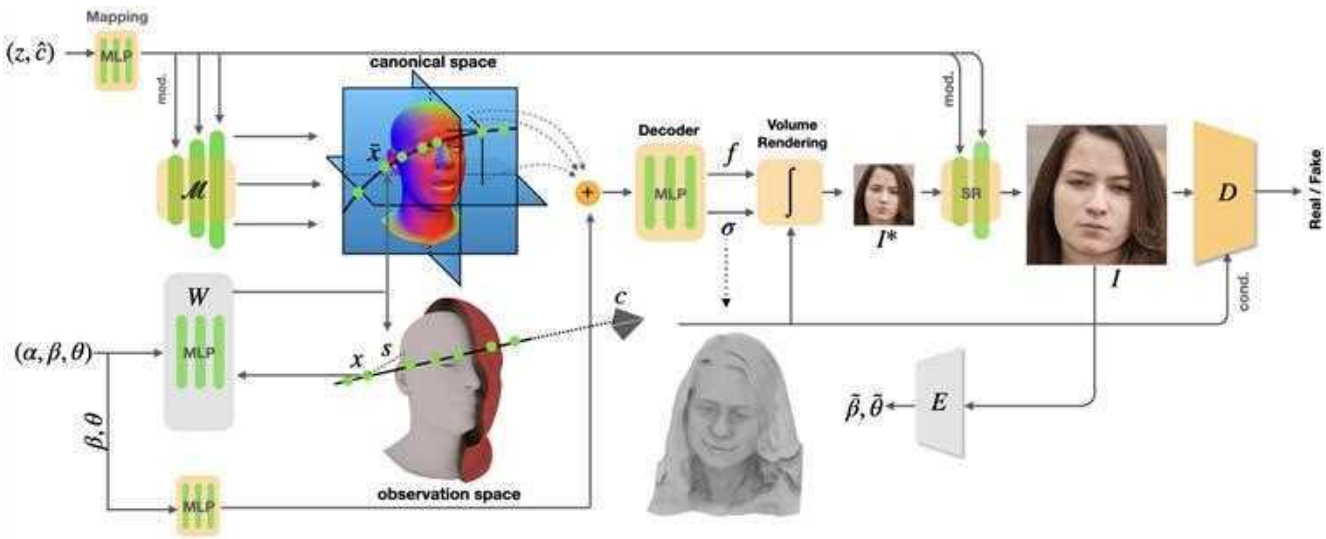

Figure 2. **Overview of our training framework. Stage 1**: Trained from parameterized FLAME [29] mesh collections, a MLP-network W maps a shape $\boldsymbol{\alpha}$, expression $\boldsymbol{\theta}$ and articulated jaw and neck pose $\boldsymbol{\theta}$ into 3D point-to-point volumetric correspondences from observation to canonical space, together with a signed distance function of the corresponding FLAME head. **Stage 2**: Given a Gaussian latent code $\mathbf{z}$, our model generates a tri-plane represented 3D feature space of a canonical head, disentangled with shape and expression controls. The volume rendering is then guided by the volumetric correspondence field to map the decoded neural radiance field from the canonical to observation space. We condition the NeRF decoding with expression and joint pose for modeling dynamic details. A super-resolution module synthesizes the final high-resolution RGB image from the volume-rendered feature map. For fine-grained shape and expression control, we apply the FLAME SDF as geometric prior to the synthesized NeRF density, and self-supervise the image synthesis to commit to the target expression $\boldsymbol{\beta}$ and joint pose $\boldsymbol{\theta}$ by comparing the input code against the re-estimated values $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}$ from synthesized images.

A similar strategy has also been adopted with concurrent and follow-up works [14, 27, 41, 53]. However all of these approaches suffer from 3D inconsistency due to the use of 2D CNNs as image renderer. HeadNeRF [21] combines 3DMM with 3D NeRF representation, and is able to synthesize 3D heads conditioned on 3DMM attributes. However, the training relies on annotated multiview datasets whereas our approach learns the disentangled 3D head synthesis with only single-view images. There has been concurrent work [4,49,51,58,64] to ours in 3D-aware controllable face or full-body GANs. Differently from these approaches, we use a full-head parametric model FLAME [29], and fully exploit the spatial geometric prior knowledge beyond the surface deformation and skinning. We have also achieved fine-grained control with our novel losses and enhanced face animation with rich dynamic details.

Controllable Neural Implicit Field of Face. Neural implicit functions [59], have emerged as a powerful continuous and differential representations of 3D scenes. Among them, Neural Radiance Field [3, 33] has been widely adopted due to its superiority in modeling complex scene details and synthesizing multiview images with inherited 3D consistency. While initial proposals have focused on static scene modeling, recent work have successfully demonstrated application of NeRF in modeling dynamic scenes [32,38,42,55,60]. In particular, dynamic neural radiance fields of human heads [13,18,22,38,39,56,66,68] have enabled photo-realistic head animation, often by conditioning with pose parameters or deforming the radiance field with 3D morphable models. However, they do not leverage a generative training paradigm and focus on scene-specific learning from video sequences or multiview images. In contrast, our model learns generative and controllable neural radiance fields from widely accessible single-view images.

3. Method

Our goal is to build a geometry-guided 3D head synthesis model with full control of camera poses, face shapes and expressions, trained from in-the-wild unstructured image collections.

3.1. Overview

Problem. To achieve our goal, we leverage a 3D-aware generator (EG3D [7]) for photo-realistic, multiview consistent image synthesis, while disentangle control of head geometric attributes from image generation with a 3D statistical head model (FLAME [29]). Specifically, given a random Gaussian-sampled latent code $\mathbf{z}$, a camera pose $\mathbf{c}$ and a FLAME parameter $\mathbf{p} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta})$ consisting of shape $\boldsymbol{\alpha}$, expression $\boldsymbol{\beta}$, jaw and neck pose $\boldsymbol{\theta}$, the generator G synthesizes a photo-realistic human head image $I_{RGB}(\mathbf{z}|\mathbf{c}, \mathbf{p})$ with corresponding attributes as defined in $\mathbf{p}$.

Framework. As illustrated in our pipeline 2, our controllable 3D-aware GAN is trained in two stages. From a large collection of 3D deformed FLAME meshes, we first pre-train a deformable semantic SDF around the FLAME geometry that builds a differential volumetric correspondence map from the observation to a predefined canonical space (Section 3.2.1). In the second stage, guided by the pre-trained volumetric mapping, we then deform the detailed 3D full heads synthesized in the disentangled canonical space to the desired shapes and expressions (Section 3.2.2). Fine-grained expression control is achieved by supervising image synthesis such that expressions estimated from the generated images is consistent with the input control(Section 3.2.3). Our approach further enhances temporal realism with dynamic details, such as dimples and wrinkles, synthesizing realistic shading and geometric variations as expression changes (Section 3.2.4).

3D-Aware GAN Background. To ensure appearance consistency from different views, we choose EG3D [7] as our backbone for 3D-aware image synthesis. The generator G takes a random latent code $\mathbf{z}$ and conditioning camera label $\hat{\mathbf{c}}$, and maps to a manifold of triplane features $\mathcal{M}(\mathbf{z}, \hat{\mathbf{c}})$. For presentation clarity, we absorb $\hat{\mathbf{c}}$ to $\mathbf{z}$ and simply denote triplane as $\mathcal{M}(\mathbf{z})$. A low-resolution feature map $I^*(\mathbf{z}|\mathbf{c})$ is then rendered from a desired camera pose $\mathbf{c}$ by sampling the triplane features and integrating MLP-decoded neural radiance $(\sigma, \mathbf{f})$ along camera rays. A super-resolution module is followed to modulate the feature map and synthesize the final RGB images at high resolution. We train G and a dual discriminator D with adversarial training.

3.2. Controllable 3D-Aware Image Synthesis

To synthesize an image $\mathbf{I}(\mathbf{z}|\mathbf{c}, \mathbf{p})$ with desired FLAME control $\mathbf{p}$, we leverage the EG3D framework to generate a triplane-based 3D volumetric feature space $\mathcal{M}(\mathbf{z})$ of a synthesized identity with canonical shape and expression. Guided by our pretrained volumetric correspondence map, we deform the synthesized feature volume into our target observation space, which is further decoded and volume rendered into high-fidelity head appearance and geometry with the target shape and expression. Our design explicitly disentangles the underlying geometric variations of changing shape and expression from canonical geometry and appearance synthesis. Following EG3D [7], we associate each training image with a set of camera parameters $\mathbf{c}$ and control parameters $\mathbf{p}$, which are obtained from a nonlinear 2D landmarks-based optimization.

3.2.1 Semantic Signed Distance Function

For disentangled geometric modeling, we formulate an implicit semantic SDF representation $W(\mathbf{x}|\mathbf{p} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta})) = (s, \bar{\mathbf{x}})$, where $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are the linear shape and expression

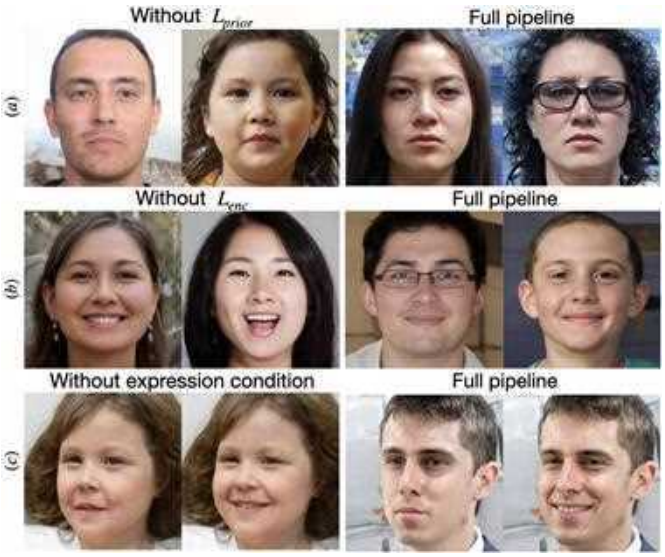

Figure 3. We synthesize different identities with the same shape and expression code respectively in (a) and (b). We observe less shape variation with our geometric prior loss (a) and more consistent expression with our control loss (b). In (c), we show our expression-conditioned NeRF modeling builds rich dynamic details as the subject varies expressions.

blendshape coefficients, and $\boldsymbol{\theta}$ controls the rotation of a 3-DoF jaw and neck joint. Specifically, given a spatial point $\mathbf{x}$ in observation space $\mathcal{O}(\mathbf{p})$, W returns its 3D correspondence point $\bar{\mathbf{x}}$ (i.e., semantics) in canonical space $\mathcal{C}(\bar{\mathbf{p}})$, with which we project and query the triplane features $\mathcal{M}(\mathbf{z})$. Additionally it also computes the closest signed distance $s(\mathbf{x}|\mathbf{p})$ to the FLAME mesh surface $\mathbf{S}(\mathbf{p})$. We illustrate the function in Figure 4. We co-learn the highly-correlated volumetric correspondence and SDF with the property that the signed distance is preserved between canonical and observation correspondence points as $s(\bar{\mathbf{x}}|\bar{\mathbf{p}}) = s(\mathbf{x}|\mathbf{p})$.

We learn $W(\mathbf{x}|\mathbf{p})$ with a large corpus of 3D FLAME meshes $\mathbf{S}(\mathbf{p})$ sampled from its parametric control space. Similar to IGR [16] and imGHUM [2], we model our implicit field as an MLP and optimize $W(\mathbf{x}|\mathbf{p})$ with losses,

$$L_{iso} = \frac{1}{|N|} \sum_{\mathbf{x} \in N} (|s(\mathbf{x}|\mathbf{p})| + |\nabla s_{\mathbf{x}}(\mathbf{x}|\mathbf{p}) - \mathbf{n}(\mathbf{x}|\mathbf{p})|), \quad (1)$$

$$L_{eik} = \frac{1}{|F|} \sum_{\mathbf{x} \in F} \|\nabla s_{\mathbf{x}}(\mathbf{x}|\mathbf{p}) - 1\|_2, \quad (2)$$

$$L_{sem} = \frac{1}{|N|} \sum_{\mathbf{x} \in N} (|\bar{\mathbf{x}}(\mathbf{x}|\mathbf{p}) - \bar{\mathbf{x}}^*(\mathbf{x}|\bar{\mathbf{p}})|) \quad (3)$$

where N, F are a batch of on and off surface samples. For the surface samples, the L_{iso} encourages the signed distance values to be on the zero-level-set and the SDF gradient to be equal to the given surface normals $\mathbf{n}$. The Eikonal loss L_{eik} is derived from [16] where the SDF is differen-

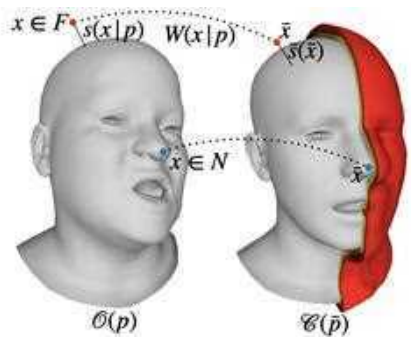

Figure 4. Illustration to semantic SDF learning.

tiable everywhere with gradient norm 1. The semantic loss L_{sem} supervises the mapping of surface samples $\mathbf{x} \in F$ to the ground-truth correspondence points $\bar{\mathbf{x}}^*$ on the canonical FLAME surface, where $\bar{\mathbf{x}}^*$ and $\mathbf{x}$ share the same barycentric coordinates. The L_{eik} provides a geometric regularization over the volumetric SDF, whereas the L_{iso} and L_{sem} act as the boundary condition to the SDF and volumetric correspondence field respectively.

To guide the canonical correspondence learning for off-the-surface points, one could try to minimize the signed distance difference between $s(\bar{\mathbf{x}}|\bar{\mathbf{p}})$ and $s(\mathbf{x}|\mathbf{p})$. However, we note that the volume correspondence between the observation and canonical space is still ill-regularized, considering infinite number of points exist with the same signed distance value. We therefore reformulate volumetric correspondence field as $W(\mathbf{x}|\mathbf{p}) = (\bar{s}(\bar{\mathbf{x}}), \bar{\mathbf{x}})$, where the signed distance of an observation-space point is obtained by mapping it into its canonical correspondence $\bar{\mathbf{x}}$ and querying the pre-computed canonical SDF $\bar{s}$. Thus we only learn a correspondence field with which we can deform the canonical SDF to different FLAME configurations, and then supervise the deformed SDFs with the FLAME surface boundary, normals (L_{iso}) and Eikonal regularization (L_{eik}). As such, even for off-the-surface samples, their canonical correspondences are well regularized in space with the geometric properties of signed distance functions via L_{iso} and L_{eik}. In contrast to explicit mesh deformation [4], our implicit volumetric correspondence is more accurate, differentiable, smooth everywhere and semantically consistent with the properties of SDF.

3.2.2 Canonical Generation with Geometric Prior

With our pretrained semantic SDF $W(\mathbf{x}|\mathbf{p})$ modeling the shape and expression variation, we leverage the triplane for 3d-aware generation of human heads with canonical shape and expression. In particular, to generate a neural radiance feature $(\sigma, \mathbf{f})$ for a point $\mathbf{x}$ in observation space $\mathcal{O}(\mathbf{p})$, we use our correspondence function W to back warp $\mathbf{x}$ into $\bar{\mathbf{x}}$, with which we project and sample the canonical triplane features followed with a tiny MLP decoding.

In spite of the control disentanglement, there is no explicit loss that constrains the triplane-generated neural radiance density to conform to the shape and expression as defined in $\mathbf{p}$. Therefore, we guide the generation of neural radiance density field by minimizing its difference to the underlying FLAME head geometry represented with SDF, as

$$L_{prior} = \frac{1}{|R|} \sum_{\mathbf{x} \in R} e^{-\gamma |s(\mathbf{x}|p)|} |\sigma(\mathbf{x}|\mathbf{z}, \mathbf{p}) - \sigma^*(\mathbf{x}|\mathbf{p})|, \quad (4)$$

$$\sigma^*(\mathbf{x}|\mathbf{p}) = \frac{1}{\kappa} \cdot \text{Sigmoid}(\frac{-s(\mathbf{x}|\mathbf{p})}{\kappa}) \quad (5)$$

where R is the stratified ray samples for volume rendering and κ is a learnable scalar controlling the density tightness around the SDF boundary. Following [37, 63], we convert SDF value $s(\mathbf{x}|\mathbf{p})$ to proxy 3D density $\sigma^*(\mathbf{x}|\mathbf{p})$ assuming non-hollow surfaces. We decay the weights for our geometric prior loss L_{prior} as the point moving away from the SDF boundary, allowing higher degrees of freedom in generation of residual geometries, such as hair and glasses. The geometric prior loss effectively guides the 3D head geometry learning but should not be overpowered which otherwise might lead to loss of geometric details.

3.2.3 Fine-Grained Expression Control

Our geometric prior loss L_{prior} provides local 3D point-wise guidance, and is able to well regularize the shape generation and achieve coarse-level expression control. However, for delicate expressions, such as eye blinks, L_{prior} provides little supervision as the geometric variation is subtle. Moreover, for regions with complex correspondences, such as around the lips, it is challenging to guide the formation of correct expressions globally, just with point-wise geometric losses. To improve the control granularity, we propose an image-level supervision loss that requires a synthesized image $I_{RGB}(\mathbf{z}|\mathbf{c}, \mathbf{p})$ matching the target expression as defined in the input $\mathbf{p}$. Using our training images with estimated control labels $\mathbf{p}$, we first pretrain an image encoder $E(I_{RGB}) = (\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})$ that regresses the expression coefficients $\tilde{\boldsymbol{\beta}}$ and joint poses $\tilde{\boldsymbol{\theta}}$. During our 3D GAN training, we then apply our image-level control supervision as

$$L_{enc} = |\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}| + |\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}| + |\mathbf{S}(\boldsymbol{\alpha}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}}) - \mathbf{S}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta})| + \\ + |\mathbf{J}\mathbf{S}(\boldsymbol{\alpha}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}}) - \mathbf{J}\mathbf{S}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta})|, \quad (6)$$

where $\mathbf{S}$, $\mathbf{J}$ are the FLAME mesh vertices and 3D landmarks regressor respectively. While being straightforward for the first 2 terms, the last two terms in L_{enc} penalize deviation of 3D vertex coordinates and surface landmarks after mesh decoding. We note that we do not supervise shape $\boldsymbol{\alpha}$ in L_{enc} since our geometric prior loss L_{prior} suffices in shape control already and also due to the ambiguity of shape scaling estimated from monocular images.

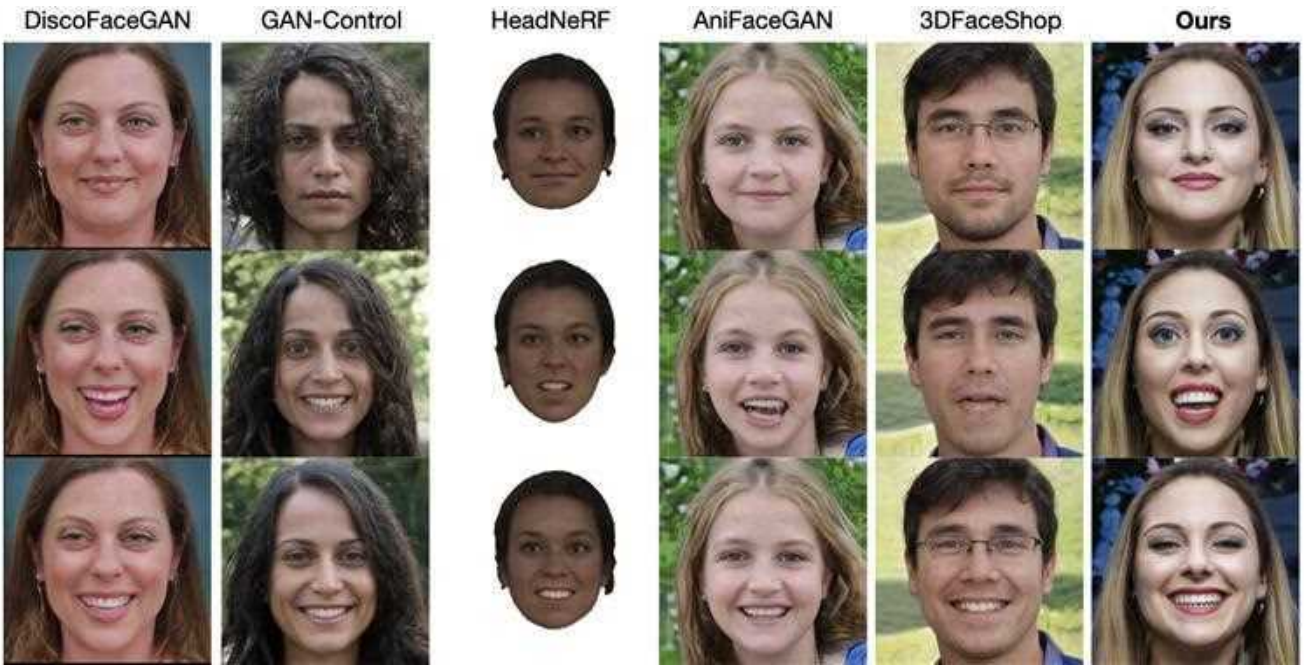

Figure 5. **Qualitative comparisons in expression control**. Under a similar expression control, our method achieves the best perceptual quality with better identity-preservation (compared to GAN-Control [47] and 3DFaceShop [51]), more realistic dynamic details (compared to DiscoFaceGAN [10]) and higher expression control accuracy (compared to AniFaceGAN [58], 3DFaceShop [51], HeadNeRF [21]).

3.2.4 Dynamic Details Modeling

To this end, we have achieved controllable static image synthesis. However, one should observe the variation of shading and geometric details in a dynamic head motion, such as the appearance of dimples and eye wrinkles in smiling. We consider the appearance of dynamic details is highly correlated with the driving expression $(\boldsymbol{\beta}, \boldsymbol{\theta})$. To model such temporal effects, one could try to condition the triplane generation on expression as $\mathcal{M}(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\theta})$. However, such design results in entanglement of expression control with the generative latent code $\mathbf{z}$, inducing identity or appearance changes when varying expressions. It is also hard to synthesize images out of training distribution of expressions $(\boldsymbol{\beta}, \boldsymbol{\theta})$. We therefore leave our triplane generation disentangled but when decoding neural feature field $(\sigma, \mathbf{f})$ from sampled triplane features, we additionally condition the MLP-decoder on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Specifically, we have

$$(\sigma(\mathbf{x}), \mathbf{f}(\mathbf{x}) | \mathbf{z}, \mathbf{p}) = \Phi(\mathcal{M}(\bar{\mathbf{x}}(\mathbf{x}|\mathbf{p})|\mathbf{z}), \phi(\boldsymbol{\beta}, \boldsymbol{\theta})), \qquad (7)$$

where both Φ and ϕ are tiny MLPs and ϕ regresses an expression-dependent feature vector from $(\boldsymbol{\beta}, \boldsymbol{\theta})$ after positional encoding. For better extrapolation to novel expressions and jaw poses, we add Gaussian noise to the conditioning parameters to prevent MLP overfitting.

4. Experiments

Training and Dataset. Our training is devised into two stages. In the pretraining stage, we build our semantic SDF W using a four 192-dimensional MLP with a collection of 150K FLAME meshes by Gaussian sampling of the control parameters $\mathbf{p}$. For each mesh, we sample 4K surface samples N with surface normals $\mathbf{n}$ and ground-truth correspondence point $\bar{\mathbf{x}}^*$, and 4K off surface samples F distributed uniformly inside the box-bounded volume. Our canonical mesh has a neutral shape $\boldsymbol{\alpha} = 0$, expression $\boldsymbol{\beta} = 0$ and neck pose, but with an opening jaw at 10 degrees. We close the face connectivity between the lips for a watertight geometry and also as a proxy geometry for the mouth cavity modeling. We open the jaw at the canonical space such that the spatial geometry and appearance inside the mouth can be modeled with distinguishable triplane features. In the second training stage, we freeze the weights of W for best efficiency and train our model on the FFHQ [24], a human face dataset with $70,000$ in-the-wild images. For each image, we estimate the camera pose $\mathbf{c}$ and FLAME parameters $\mathbf{p}$ with a nonlinear optimization in fitting 2D landmarks, assuming zero root and neck rotation and camera located 1.0m away from the world origin. We refer to the supplementary materials for more details. We rebalance the FFHQ dataset by duplicating images with large head poses and jaw opening expressions. We note that we also use pairs of images and $(\boldsymbol{\beta}, \boldsymbol{\theta})$ for fine-tuning of a light-weight image encoder

Model	FID ↓	KID ↓	$DS_{\boldsymbol{\alpha}}$↑	$DS_{\boldsymbol{\beta},\boldsymbol{\theta}}$↑	$DS_{\mathbf{c}}$ ↑	IS↑
DiscoFaceGAN-SR [10]	42.6	0.04	-	27.2	0.11	0.59
GAN-Control [47]	9.8	0.005	-	43.7	0.28	0.62
HeadNeRF [21]	163	0.168	1.33	8.9	0.25	0.678
AniFaceGAN-SR [58]	30.5	0.024	2.0	11.1	0.04	0.73
3DFaceShop [51]	24.8	0.018	0.45	64.2	0.029	0.594
Ours	**5.7**	**0.0016**	**3.1**	**71.9**	**0.35**	**0.80**

(a) Baseline comparisons.

Model	Ours wo L_{enc}	Ours wo L_{prior}	Ours
FID ↓	6.0	6.1	**5.7**
KID ↓	0.00172	0.00174	**0.0016**
ASD (cm) ↓	**0.124**	0.135	**0.124**
$var(\boldsymbol{\alpha})$ ↓	0.000525	0.000577	**0.000515**
AED (cm) ↓	0.1233	0.0813	**0.0797**
$var(\boldsymbol{\beta},\boldsymbol{\theta})$ ↓	0.004707	0.00406	**0.00404**

(b) Ablation study.

Table 1. (a) Our method outperforms the prior 2D and 3D controllable image synthesis methods in both image quality and control disentanglement. (b) Our method demonstrates the best control accuracy over shape and expression with geometric prior L_{prior} and self-supervised reconstruction loss L_{enc}.

E, using a ResNet50 [19] backbone followed with a single-layer MLP.

4.1. Qualitative Comparisons

Controlled Portrait Synthesis. Our framework enables high-fidelity head synthesis with disentangled control over camera pose, shape and expression, as shown in Figure 1. We compare our method with prior controllable portrait synthesis works include DiscoFaceGAN [10], GAN-Control [47], HeadNeRF [21], AniFaceGAN [58] and 3DFaceShop [51]. The first two methods integrates 3DMM controls into 2D face synthesis. HeadNeRF [21] is a 3D parametric NeRF-based head model built from both indoor multiview image capture datasets and in-the-wild monocular image collections. AniFaceGAN [58] and 3DFaceShop [51] are concurrent 3D-aware controllable GANs in face synthesis.

We qualitatively compare with prior work in Figure 5 for expression control. GAN-Control shows high-quality expression manipulation but with noticeable identity variation with differences in hair, head shape contour and background, whereas DiscoFaceGAN better maintains identity but with lower perceptual quality, and lacks dynamic details as expression varies. The expression control space of 3DFaceShop [51] is sensitive, suffering from appearance changes even with minor expression variation. AniFaceGAN demonstrates visually-consistent expression editing but with limited resolution and image quality, e.g., with blurry artifacts in hair and teeth. HeadNeRF ensures decent consistency in image generation by rendering the conditional NeRF but lacks fine details with very limited perceptual quality. In contrast, our approach produces the most compelling images with consistent appearance and dynamic details under expression changes. Moreover, controllable neck pose is a key factor towards realistic video avatar in applications like talking head synthesis. As shown in Figure. 1, our method achieves explicit control of neck and jaw poses which are seldom explored in prior work. Please refer to supplementary material for qualitative comparison in view consistency and shape editing.

Dynamic details. Our method presents highly-consistent image synthesis in control of shape, expression and camera poses, but also depicts temporal realism in portrait animation, credited to the modeling of expression-dependent dynamic details. As shown in the last row of Figure 3, the appearance of wrinkles around the mouth and eyes when transiting from a neural expression to smiling largely enhances the animation realism. In comparison, without explicitly modeling the dynamic details, the wrinkles are embedded in the appearance and do not vary with expressions.

Expressive Head Synthesis with Extrapolated Controls. AniFaceGAN [58] and 3DFaceShop [51] depict dynamic details as well since their generation of the neural fields is conditioned on the input expression latent code. However, their designs result in shape and expression entanglement with the appearance generation, as reflected in the identity changes as shown in Figure. 5. By embedding the controls in a Gaussian latent space with such as Variation AutoEncoder (VAE), their approaches sacrifice expressiveness. The synthesized image quality is also highly correlated to the distribution of training images with target expressions.

In contrast, our tri-plane generation is explicitly disentangled from shape and expression control. Moreover, the volume deformation is independently learnt from the deformed FLAME mesh collections which offer abundant 3D geometric knowledge with largely augmented control space. Therefore we are much less dependent on the distribution of the training images and support better extrapolation to unseen novel expressions. In Figure. 1, we show high-quality synthesized head with extreme jaw and neck articulated movements which do not exist in our training images. Our expression control is also more expressive, supporting subtle expressions like eye blinks (Figure. 1 3).

4.2. Quantitative Comparisons

Image Quality We measure the image quality with Frechet Inception Distance (FID) [20] and Kernel Inception Distance [5] between 50K randomly synthesized images and 50K randomly sampled real images at the resolution of 512×512. Since DiscoFaceGAN [10] and AniFaceGAN [58] only synthesize images at 256×256 reso-

lution, we utilize a state-of-the-art super-resolution model, SwinIR [30] to upsample into 512×512 for a fair comparison. As shown in Table. 1a, our method is superior in both FID and KID than all prior work, demonstrating the most compelling image quality. We note that the original EG3D has a slightly lower FID at 4.8 which is expected since we introduce controllability with additional loss regularization.

Disentanglement To evaluate the disentangled controllability of our model over shape, expression and camera pose, we measure the disentanglement score [10] of synthesized images as $\text{DS}_{\boldsymbol{\alpha}}$, $\text{DS}_{\boldsymbol{\beta},\boldsymbol{\theta}}$ and $\text{DS}_{\mathbf{c}}$ respectively. DS measures the stability of other factors when a single attribute is modified in image synthesis. We employ DECA [12] for estimation of FLAME parameters from generated images and calculate the variance of the estimated parameters $(\boldsymbol{\alpha}, \{\boldsymbol{\beta}, \boldsymbol{\theta}\}, \mathbf{c})$. Specifically the DS_i is calculated as

$$DS_i = \prod_{j \neq i} \frac{var(i)}{var(j)}, \ i, j \in \{\boldsymbol{\alpha}, \{\boldsymbol{\beta}, \boldsymbol{\theta}\}, \mathbf{c}\}. \tag{8}$$

Higher value of DS indicates better disentanglement. Additionally, we evaluate the identity similarity between pairs of synthesized images with random camera poses and expressions by calculating the cosine similarity of the face embeddings with a pre-trained face recognition module [26]. Our approach demonstrates the best disentanglement numerically over all prior work as indicated in Table. 1a.

4.3. Ablation Studies

We ablate the efficacy of the individual component by removing it from our full pipeline. As shown in Figure. 3, we show the loss of control accuracy over shape and expression respectively when removing the geometric prior loss L_{prior} and control loss L_{enc}. Conditioning the neural radiance field on expression is also critical to the modeling of dynamic details. Numerically we validate the efficacy of L_{prior} and L_{enc} with Average Shape Distance (ASD) and Average Expression Distance (AED). From the prior distribution, we randomly sample 500 shapes and expressions, with which we synthesize images with 10 different identities. We then reconstruct FLAME parameters from the synthesized images and compare against the input control. We compute 3D per-vertex L_1 distance of FLAME meshes for ASD while we use 3D landmarks L_1 distance for AED calculation. Additionally we compute the variance of estimated shapes and expressions within each control group, with lower value indicating more precise control. The efficacy of L_{prior} and L_{enc} is well evidenced in Table. 1b, with even slight image quality improvements.

4.4. Applications

Talking Head Video Generation. In Figure. 1, we showcase talking head video generation in controllable views driven with animation sequences of FLAME. Thanks to the high control accuracy, our method is able to synthesize various talking head videos with the same head movements and expressions performed by different identities. Our method is expressive in depicting both large articulated neck and jaw movements and subtle facial expressions like eye blinks, with rich dynamic details. Shape manipulation is easily achievable as well by modifying the shape parameters. Please refer to our supplementary materials for more results in high resolution.

Portrait Image Manipulation and Animation. As illustrated in Figure. 1, our model also supports 3D-aware face reenactment of a single-view portrait to a video sequence. To achieve that, we perform an optimization in the latent Z+ space [24] to find the corresponding latent embedding, with FLAME parameter and camera pose estimated from the input portrait. With a frozen generator, the optimization is performed by measuring the similarity between generated image and real image using the L_2 loss and LPIPS loss [65]. For better reconstruction quality, we alter the parameters of the tri-plane synthesis module with a fixed optimized latent code [43]. After that one can explicitly manipulate the portrait with a preserved identity and in a different camera pose and expression. With the expression codes $(\boldsymbol{\beta}, \boldsymbol{\theta})$ reconstructed from a video sequence, we are also able to reenact the portrait to the video motion.

Societal Impact. Our work focuses on improving the controllability of 3D-aware GANs in technical aspects and is not specifically designed for any malicious uses. This being said, we do see that the method could be potentially extended into controversial applications such as generating fake videos. Therefore, we believe that the synthesized images and videos should present themselves as synthetic.

5. Conclusion

In this work, we introduce OmniAvatar, a novel 3D-aware generative model for synthesis of controllable high-fidelity human head. Our model achieves disentangled semantic control by factoring the generative process into 3D-aware canonical head synthesis and implicit volume deformation to target shapes and expressions. By learning a deformable signed distance function with canonical volumetric correspondence, we instill the geometric prior knowledge of a 3D statistical head model into the generation of neural radiance fields, enabling superior generative capability to novel shapes and expressions. Our approach demonstrates expressive and compelling talking head generation and portrait image animation with fine-grained control accuracy and temporal realism. We believe the proposed method presents an interesting direction for 3D avatar creation and animation, which sheds light on many potential downstream tasks.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *ICML*, 2018. 2

[2] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *ICCV*, pages 5461–5470, 2021. 4

[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 *iccv*, pages 5855–5864, 2021. 3

[4] Alexander W Bergman, Petr Kellnhofer, Yifan Wang, Eric R Chan, David B Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. *arXiv preprint arXiv:2206.14314*, 2022. 3, 5

[5] Sai Bi, Kalyan Sunkavalli, Federico Perazzi, Eli Shechtman, Vladimir G Kim, and Ravi Ramamoorthi. Deep cg2real: Synthetic-to-real translation via image disentanglement. In *iccv*, pages 2730–2739, 2019. 7

[6] 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*, 1999. 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, et al. Efficient geometry-aware 3d generative adversarial networks. *CVPR*, 2022. 2, 3, 4

[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/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 2

[9] Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Jingyi Yu. Sofgan: A portrait image generator with dynamic styling. *sigg*, 41(1):1–26, 2022. 2

[10] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *cvpr*, pages 5154–5163, 2020. 2, 6, 7, 8

[11] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. *CVPR*, 2022. 2

[12] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. In *ACM Transactions on Graphics*, volume 40, 2021. 8

[13] 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 Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 3

[14] Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J Black, and Timo Bolkart. Gif: Generative interpretable faces. In *3dv*, pages 868–878. IEEE, 2020. 2, 3

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1, 2

[16] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *icml*, pages 3789–3799, 2020. 4

[17] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *CVPR*, 2022. 2

[18] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *iccv*, pages 5784–5794, 2021. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *cvpr*, pages 770–778, 2016. 7

[20] 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. *NeurIPS*, 2017. 7

[21] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *cvpr*, pages 20374–20384, 2022. 3, 6, 7

[22] Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzciński, and Andrea Tagliasacchi. Conerf: Controllable neural radiance fields. In *cvpr*, pages 18623–18632, 2022. 3

[23] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *NeurIPS*, 2021. 1

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 6, 8

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 1, 2

[26] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *cvpr*, pages 18750–18759, 2022. 8

[27] Marek Kowalski, Stephan J Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. Config: Controllable neural face image generation. In *eccv*, pages 299–315. Springer, 2020. 2, 3

[28] Ruihui Li, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-gan: a point cloud upsampling adversarial network. In *ICCV*, 2019. 2

[29] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6):194:1–194:17, 2017. 2, 3

[30] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *iccv*, pages 1833–1844, 2021. 8

[31] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3D controllable image synthesis. In *CVPR*, 2020. 2

[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 Trans. on Graphics*, 2021. 3

[33] 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 *ECCV*, 2020. 2, 3

[34] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3D representations from natural images. In *ICCV*, 2019. 2

[35] Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy Mitra. BlockGAN: Learning 3D object-aware scene representations from unlabelled images. In *NeurIPS*, 2020. 2

[36] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 2

[37] 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. *CVPR*, 2022. 2, 5

[38] 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 *ICCV*, 2021. 3

[39] 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*, 2022. 3

[40] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 2

[41] Jingtan Piao, Keqiang Sun, Quan Wang, Kwan-Yee Lin, and Hongsheng Li. Inverting generative adversarial renderer for face reconstruction. In *cvpr*, pages 15619–15628, 2021. 2, 3

[42] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 2021. 3

[43] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 8

[44] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *NeurIPS*, 2020. 2

[45] Yichun Shi, Divyansh Aggarwal, and Anil K Jain. Lifting 2d stylegan for 3d-aware face generation. In *cvpr*, pages 6258–6266, 2021. 2

[46] Yichun Shi, Xiao Yang, Yangyue Wan, and Xiaohui Shen. Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing. In *CVPR*, 2022. 2

[47] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. Gan-control: Explicitly controllable gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14083–14093, 2021. 6, 7

[48] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. *arXiv preprint arXiv:2206.10535*, 2022. 2

[49] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and HongSheng Li. Controllable 3d face synthesis with conditional generative occupancy fields. *arXiv preprint arXiv:2206.08361*, 2022. 3

[50] Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3D shape learning from natural images. *arXiv*, 2019. 2

[51] 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. 3, 6, 7

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

[53] 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 *cvpr*, pages 6142–6151, 2020. 2, 3

[54] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *sigg*, 40(4):1–14, 2021. 2

[55] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *ICCV*, 2021. 3

[56] Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. Learning compositional radiance fields of dynamic human heads. In *cvpr*, pages 5704–5713, 2021. 3

[57] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T. Freeman, and Joshua B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *NeurIPS*, 2016. 2

[58] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. *arXiv preprint arXiv:2210.06465*, 2022. 3, 6, 7

[59] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, volume 41, pages 641–676. Wiley Online Library, 2022. 3

[60] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *NeurIPS*, 2021. 3

[61] Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. Generative occupancy fields for 3d surface-aware image synthesis. *NeurIPS*, 34:20683–20695, 2021. 2

[62] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe hd: A high-resolution 3d-aware generative model. In *CVPR*, 2022. 2

[63] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 5

[64] 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. 3

[65] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *cvpr*, pages 586–595, 2018. 8

[66] 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 *cvpr*, pages 13545–13555, 2022. 3

[67] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis. *arXiv*, 2021. 2

[68] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. *arXiv preprint arXiv:2112.02308*, 2021. 3