Portrait4D: Learning One-Shot 4D Head Avatar Synthesis using Synthetic Data

Yu Deng Duomin Wang Xiaohang Ren Xingyu Chen Baoyuan Wang
Xiaobing.AI
https://yudeng.github.io/Portrait4D/

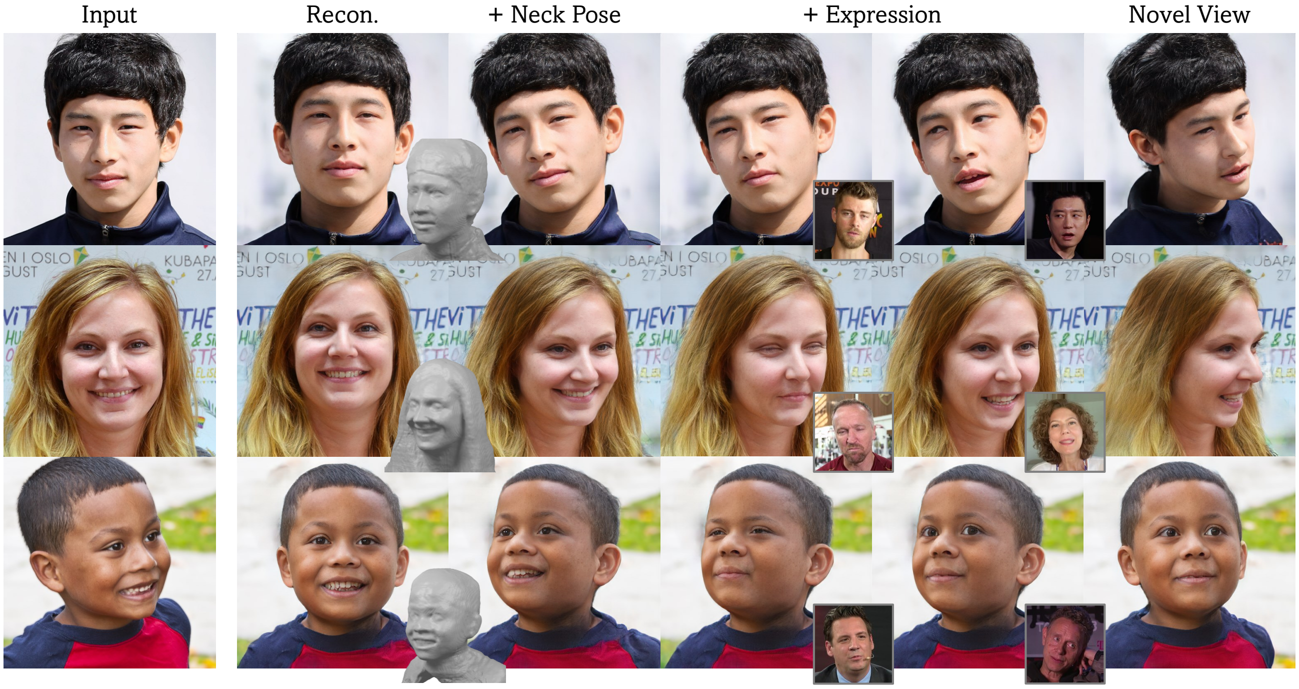

Figure 1. Our method generates a photorealistic 4D head avatar via a feed-forward process of a monocular image. It allows head reenactment given another driving image, with full motion control of face, mouth, eyes, and neck, as well as foreground-background separation.

Abstract

Existing one-shot 4D head synthesis methods usually learn from monocular videos with the aid of 3DMM reconstruction, yet the latter is evenly challenging which restricts them from reasonable 4D head synthesis. We present a method to learn one-shot 4D head synthesis via large-scale synthetic data. The key is to first learn a part-wise 4D generative model from monocular images via adversarial learning, to synthesize multi-view images of diverse identities and full motions as training data; then leverage a transformer-based animatable triplane reconstructor to learn 4D head reconstruction using the synthetic data. A novel learning strategy is enforced to enhance the generalizability to real images by disentangling the learning process of 3D reconstruction and reenactment. Experiments demonstrate our superiority over the prior art.

1. Introduction

Animating a portrait image for photorealistic video synthesis (*i.e.*, one-shot head avatar synthesis) is a crucial task in computer vision and graphics, which can benefit diverse applications like video conferencing, live streaming, and Virtual Reality. Compared to various approaches based on 2D generative models [9, 20, 53, 64, 77, 80], animatable 3D head (*i.e.*, 4D head) synthesis yields better 3D consistency during view changes, thus more favorable in scenarios that require large head pose variation and free-view rendering.

Nevertheless, reconstructing an animatable 3D head from a single image is highly challenging, and there lack of large-scale 3D data to directly learn a deep model to tackle it. Consequently, previous methods [35, 39, 43, 74, 79] resort to monocular videos and leverage differentiable renderers for training with image-level weak supervisions. Since this process is highly ill-posed, monocular 3DMM reconstructions [4, 8, 15, 21, 38, 46] are often involved to extract pose, shape, or expression to facilitate the geometry learning. However, 3DMM reconstruction from a monocular image is evenly challenging, which hinders the existing methods to obtain reasonable geometries for large-pose reenactment. What's more, they often neglect eye gaze and neck pose control to simplify the problem. Besides, the monocular video setting also makes it difficult to learn foreground and background separation.

Therefore, we wonder if it is possible to leverage synthetic data that can provide extensive multi-view images with diverse identities and expressions for training. With them, we can fully exploit the power of advanced neural networks [19, 70] to learn 4D head synthesis in a data-driven manner, avoiding errors from monocular 3DMM reconstruction. Background isolation is also feasible with separately synthesized foreground and background data. However, synthesizing such data with enough photorealism is extremely difficult. A recent method [60] resorts to multi-view images generated by a 3D GAN [11] for learning static novel view synthesis, yet it does not deal with animations. And existing 3D GANs [3, 57, 68, 69, 73] are difficult to generate head images with full motion control and separated background that meet our requirements. What's more, it is unclear if a model learned on such synthetic data can generalize well to real image reenactment.

In this paper, we present a framework for learning one-shot 4D head avatar synthesis by creating large-scale synthetic data as supervision. Learning on the synthetic data alone, our framework achieves high-fidelity 4D head reconstruction of real images via a feed-forward pass, and supports motion control of the face, mouth, eyes, and neck, given another driving image for reenactment. We also allow independent foreground animation with background separation. The core of our framework is two-fold: **1)**, a 4D generative model (GenHead) capable of synthesizing photorealistic head images with free pose and full head motion control, trained on only monocular images. **2)**, a one-shot 4D head synthesis model (Portrait4D) together with its disentangled learning strategy for photorealistic 4D head reconstruction from a real image while trained on synthetic data of GenHead.

The key of GenHead is a combination of a tri-plane generator [11] for canonical head NeRF [44] synthesis and a deformation field derived from FLAME meshes [38] for animation. Different from previous 3D head GANs [3, 73], we leverage part-wise deformation fields and canonical tri-planes to deal with complex motions around eyes and mouth, which help with full head motion control important to vivid reenactment. GenHead is trained on in-the-wild real images via image-level adversarial learning [11, 24] and successfully "turns" these monocular data to 4D synthetic ones to enable learning the 4D head reconstruction.

Our 4D head synthesis model adopts a transformer-based encoder-decoder architecture which directly reconstructs a triplane-based head NeRF from a source image and animates it via deep motion features from a driving image. We utilize a pre-trained motion encoder [62] for expression extraction with identity information excluded, and inject it into the tri-plane reconstruction process via cross-attention with the appearance features to achieve motion control. This way, we get rid of the reliance on 3DMM estimation for pose and expression. Our pipeline is trained end-to-end via self-reenacting the synthetic identities and comparing the corresponding results at arbitrary views with their ground truth. This differentiates us with previous methods leveraging a monocular frame as both the driving and the ground truth, and yields better 3D geometries.

While our synthetic data are of high photorealism, they still have domain gaps with the real ones which influence the model's generalizability. To tackle it, we further introduce a disentangled learning strategy that isolates the 3D reconstruction and the reenactment process of the model to alleviate overfitting. The core idea is to randomly switch off the motion-related cross-attention layers and let the remaining parts focus on static 3D reconstruction only. This way, the learned model achieves faithful 3D head reconstruction and animation on unseen real images.

We train our GenHead model on FFHQ dataset [33] at 512^2 resolution and use it to generate synthetic 4D data to learn the subsequent 4D head synthesis model. Experiments show that our method achieves high-fidelity 4D head reconstruction with reasonable geometry and complete motion control (Fig. 1), outperforming previous approaches. We believe our method opens up a new way for scaling up photorealistic head avatar creation.

2. Related Work

One-shot head avatar synthesis. One-shot head avatar synthesis aims to animate a source portrait via motions from a target image. In recent years, plenty of works have taken advantage of 2D generative models for talking head synthesis at high fidelity. For example, [20, 22, 29, 48, 53, 64, 65, 80] learn latent feature deformations and send the deformed features to a 2D generator for face reenactment. [7, 26, 77] map images to the latent space of a pre-trained StyleGAN2 [34] to utilize the power of the unconditional GAN. While these methods can produce photorealistic face images, they struggle to preserve the 3D structures under large pose changes due to a lack of 3D understanding. Recently, several approaches [30, 35, 39, 40, 43, 79] pursue animatable 3D head synthesis in order for better 3D consistency. They often resort to monocular 3DMM reconstructions [15, 21] to provide geometry or pose guidance due to a lack of large-scale multi-view training data. For example, [39, 40] leverage a monocular face reconstructor [15] to

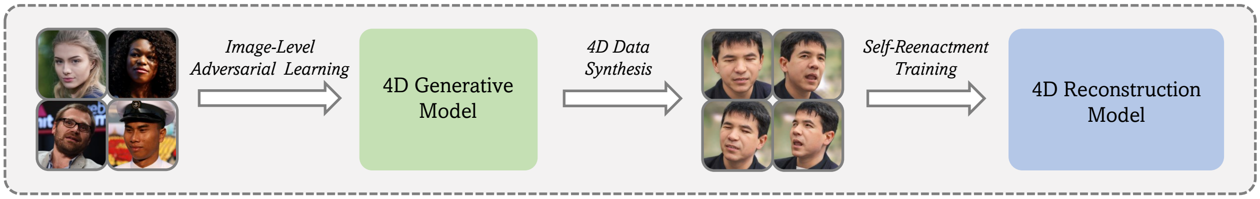

Figure 2. Overview of our method. We first learn a 4D generative head model from monocular images to synthesize large-scale 4D data. Then, we utilize the synthetic data to learn a one-shot 4D head reconstruction model in a data-driven manner.

predict 3DMM coefficients as guidance for learning 3D motions and canonical heads. [35] relies on reconstructed mesh from [21] as initialization for head rigging. [43, 79] combine 3DMM poses and expressions with a pre-trained 3D GAN [11] for reconstructing animatable head NeRFs. However, estimating 3DMMs and their poses from a monocular image can be inaccurate, thus limiting their performance in 3D head reenactment. [54] learns 3D head geometries and deformation from monocular videos in an unsupervised manner, yet its synthesis quality lags further behind.

3D-aware portrait generation. Recent studies [10–12, 17, 25, 45, 52] demonstrate that by combining 3D representations such as NeRF with adversarial learning on monocular images, it is possible to learn a 3D-aware generator for multi-view image generation of an object category, especially human portrait. While earlier works [10, 11, 17, 56] focus on static portrait synthesis, recent methods [3, 57–59, 68, 69, 73] introduce 3DMMs for extra animation control. For example, [68, 73] learn 3D deformation fields to mimic the motions of 3DMM for driving canonical NeRFs. [57] rasterizes 3DMM meshes with neural textures onto the tri-planes for animatable NeRF generation. These 3D-aware GANs serve as powerful tools for monocular head avatar reconstruction when combined with advanced GAN inversion techniques [18, 36, 50, 71, 78]. Nevertheless, the inversion process also requires camera pose or 3DMM estimation from a single image which brings inaccuracy. In addition, weight correction [50] of the pre-trained generator is often needed for out-of-domain images, which makes them less practical when applied to large-scale scenarios.

Learning with synthetic data. Synthetic data are widely used for training deep learning models in various tasks [28, 37, 47, 49, 51, 55, 63, 67, 75, 76]. Typically for human portraits, traditional graphics pipelines are often used to create synthetic data for face analysis and synthesis. For instance, [66, 67] leverage photorealistic raytracing renderer [5] to render high-quality 3D face models for landmark detection. [61, 83, 84] fit 3DMMs from images and train a CNN to regress the coefficients for monocular 3D reconstruction. [82] creates face images of various lightings via Spherical Harmonic (SH) model [2] for portrait relighting. However, for photorealistic image synthesis, these CG-style data encounter large domain gaps and often require extra adaptations [16, 23, 76]. Compared to the CG data, 3D-aware GANs [11, 57] can generate images with higher photorealism. A recent method [60] shows that images generated by a 3D-aware GAN can be used for learning one-shot novel view synthesis at high fidelity. Still, using synthetic data for one-shot 4D head synthesis is largely underexplored.

3. Approach

Given a source image I_s and a driving image I_d, we learn a model to synthesize a 3D head with the appearance of I_s and the motion of I_d, allowing free-view rendering. We adopt triplane-based NeRF [11] as the underlying 3D representation for its fidelity and efficiency. To effectively learn monocular NeRF reconstruction, we introduce a part-wise generative model G (GenHead) to synthesize multi-view head images of diverse identities and motions as training data (Sec. 3.1 and 3.2). With the large-scale synthetic data, we learn an animatable tri-plane reconstructor Ψ to directly reconstruct 4D head NeRFs from monocular images via a feed-forward pass (Sec. 3.3). A disentangled learning strategy is introduced to isolate the reconstruction and the reenactment process of Ψ to improve its generalizability to real images (Sec. 3.4). Figure 2 shows an illustration of the overall framework. We describe each part in detail below.

3.1. Part-wise Generative Head Model

The goal of GenHead is to “turn” accessible monocular data into 4D ones to enable learning the subsequent 4D NeRF reconstruction in a data-driven manner. To this end, we adopt the recent 3D-aware GAN framework [3, 57, 73] which effectively learns NeRF synthesis via adversarial learning on monocular images. Nonetheless, existing head GANs are not designed for full motion control of the face, eyes, mouth, and neck. Therefore, we introduce a part-wise generative model to deal with the complex head animation.

Specifically, the GenHead model G consists of a part-wise tri-plane generator G_{ca} for canonical head NeRF synthesis and a part-wise deformation field $\mathcal{D}$ for morphing the canonical head. G_{ca} receives a random noise $\boldsymbol{z} \in \mathbb{R}^{512}$ and a FLAME [38] shape code $\boldsymbol{\alpha} \in \mathbb{R}^{300}$, and generates two tri-planes for the head region as well as eyes and mouth:

$$\mathrm{G}_{ca} : (\boldsymbol{z}, \boldsymbol{\alpha}) \rightarrow [T_h, T_p] \in \mathbb{R}^{256\times256\times96\times2}, \qquad (1)$$

where T_h and T_p are tri-planes of a canonical 3D head and its eyes and mouth regions, respectively. Following [11], we leverage a StyleGAN2 [34] backbone for G_{ca}. A 3D point can be projected onto the tri-planes to obtain features $\boldsymbol{f}_h$ and $\boldsymbol{f}_p \in \mathbb{R}^{32}$ for radiance decoding and volume rendering (see [11] for details). Note that the head synthesized by G_{ca} is aligned with FLAME's average face shape, and $\boldsymbol{\alpha}$ in Eq. (1) is only for shape-related appearance but not shape variations, without which we found quality drops due to random combination of deformation and appearance.

The deformation field $\mathcal{D}$ receives the shape code $\boldsymbol{\alpha}$, together with FLAME's expression code $\boldsymbol{\beta} \in \mathbb{R}^{100}$ and pose code $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_{eye}, \boldsymbol{\gamma}_{jaw}, \boldsymbol{\gamma}_{neck}] \in \mathbb{R}^9$, and outputs part-wise 3D deformations for a point $\boldsymbol{x}$ in the observation space:

$$\mathcal{D} : (\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \rightarrow [\Delta\boldsymbol{x}_h, \Delta\boldsymbol{x}_p] \in \mathbb{R}^{3\times 2}. \tag{2}$$

By adding $\Delta\boldsymbol{x}_h$ or $\Delta\boldsymbol{x}_p$ to $\boldsymbol{x}$, it is deformed into the canonical spaces of T_h or T_p for feature acquisition, respectively. We utilize a FLAME mesh $\boldsymbol{m}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ to directly derive $\mathcal{D}$. Specifically, we first obtain deformation between $\boldsymbol{m}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ and an average face mesh $\boldsymbol{m}(\mathbf{0}, \mathbf{0}, \boldsymbol{\gamma}_{ca})$ for each vertex on $\boldsymbol{m}$, where $\boldsymbol{\gamma}_{ca} = [\mathbf{0}, \boldsymbol{\gamma}_{jaw,ca}, \mathbf{0}]$ denotes a canonical pose with mouth open. Then, for a free-space point, we derive its deformation to the canonical spaces via a weighted average of that of its nearest vertices on $\boldsymbol{m}$. For $\Delta\boldsymbol{x}_h$, we do not consider the eye gaze $\boldsymbol{\gamma}_{eye}$. For $\Delta\boldsymbol{x}_p$, we adopt two different derivations around eyes and mouth: for the eye region, we use the deformation between $\boldsymbol{m}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ and $\boldsymbol{m}(\mathbf{0}, \mathbf{0}, \boldsymbol{\gamma}_{ca})$ as described above; for the mouth region, we use the deformation between $\boldsymbol{m}(\boldsymbol{\alpha}, \mathbf{0}, \boldsymbol{\gamma})$ and $\boldsymbol{m}(\mathbf{0}, \mathbf{0}, \boldsymbol{\gamma}_{ca})$ without expression $\boldsymbol{\beta}$, to handle relative movements between lips and teeth caused by expression. See the *suppl. material* for more details and illustrations.

After obtaining features from each tri-plane, we perform volume rendering [32, 44] to synthesize two feature maps from $\boldsymbol{f}_h$ and $\boldsymbol{f}_p$, respectively, and blend them via a rasterized mask of $\boldsymbol{m}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ at the corresponding viewpoint $\boldsymbol{\theta}$. We leverage another StyleGAN2 to generate 2D background I_{bg}, fuse it with the rendered foreground I_f, and send the result to a 2D super-resolution module for final image synthesis, as in [1]. The whole model is trained end-to-end via adversarial loss [24] using a dual discriminator [57].

After training, GenHead is free to synthesize 4D head data by altering the combination of $(\boldsymbol{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta})$.

3.2. 4D Data Synthesis with GenHead

We generate two types of data for learning the 4D head synthesis model, namely "dynamic" data and "static" data. The dynamic data are responsible for head reenactment which contain synthetic identities $(\boldsymbol{z}, \boldsymbol{\alpha})$ with multiple motions $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ per subject and different camera poses $\boldsymbol{\theta}$ per subject and motion. The static data are used to enhance the 3D reconstruction generalizability, which contains only a single motion per identity with different camera views.

In particular, we extract $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta})$ from monocular images and videos via off-the-shelf 3DMM reconstruction methods [6, 15] with further landmark-based optimization to form a sample set. During training the 4D head synthesizer, we randomly construct online pairs of $(\boldsymbol{z}, \boldsymbol{\alpha})$ out of the sample set as the dynamic identities. For each identity, we assign random head motions and random camera poses per motion. Note that for each dynamic identity, we sample expression $\boldsymbol{\beta}$ extracted from the same video clip to alleviate the identity leakage issue born in the linear 3DMM model [31]. For the static data, we follow a similar procedure to construct random identities with an arbitrary motion per identity and random camera poses. We also keep intermediate outputs of GenHead as extra labels, including sampled tri-plane features $\bar{T}(\boldsymbol{x})$, rendered low-resolution feature maps and backgrounds $\bar{I}_f$ and $\bar{I}_{bg}$, depth images $\bar{I}_{depth}$, and opacity images $\bar{I}_{opa}$, to facilitate learning in Sec. 3.4. More details and visualizations of the data are in the *suppl.*

Notably, although we leverage monocular 3DMM reconstruction in building the synthetic data, we do not use the reconstructed 3DMM codes as input to the one-shot synthesis model as in [39, 43, 79]. This way, we avoid inheriting the 3DMM reconstruction errors at inference time.

3.3. Animatable Triplane Reconstructor

Our 4D head synthesis model builds upon an animatable tri-plane reconstructor Ψ, which takes a source image I_s and a driving image I_d as input, and reconstructs a tri-plane $T \in \mathbb{R}^{256\times 256\times 96}$ for synthesizing reenacted image I_{re} at a camera pose $\boldsymbol{\theta}$ with a renderer $\mathcal{R}$:

$$\Psi : (I_s, I_d) \rightarrow T, \ \mathcal{R} : (T, \boldsymbol{\theta}) \rightarrow I_{re}, \tag{3}$$

where $\mathcal{R}$ consists of radiance decoding, volume rendering, and 2D super-resolution, similar to Sec. 3.1. The tri-plane T in Eq. (3) represents a 3D head with the identity of I_s and the motion of I_d (with zero neck pose) instead of an average head as in GenHead. To faithfully obtain T, we construct Ψ with two appearance encoders E_{global} and E_{detail}, a motion encoder E_{mot}, a canonicalization and reenactment module Φ, and a tri-plane decoder G_T, as shown in Fig. 3.

E_{global} and E_{detail} are used to extract appearance feature maps from I_s for the subsequent 3D reconstruction and reenactment, which adopt the CNN structures in [60].

Then, we introduce Φ to canonicalize the feature map F_{global} from E_{global} as well as reenact it with the motion from I_d. Φ consists of two transformer-based modules Φ_{de} and Φ_{re} sharing the same structure, one for neutralizing the expression of F_{global} and the other for injecting motions of I_d for reenactment, meanwhile they both serve for 3D pose canonicalization. Specifically, they contain multiple transformer blocks each with a cross-attention layer, a self-attention layer, and an MLP. The self-attention and

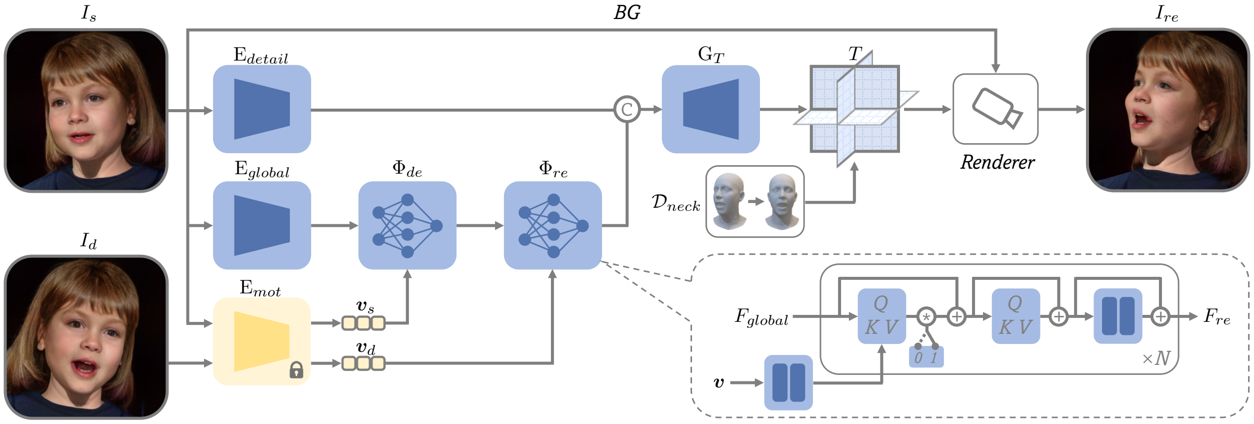

Figure 3. Architecture of the animatable triplane reconstructor Ψ. An encoder E_{global} first extracts the appearance feature map of I_s. The feature is then sent to a canonicalization and reenactment module Φ consisting of a de-expression module Φ_{de} and a reenactment module Φ_{re} sharing the same structure, which receives motion features from either I_s or I_d for expression neutralization or motion injection accordingly. The reenacted feature is then concatenated with a detail feature map from another encoder E_{detail}, and sent to a decoder G_T to synthesize a tri-plane T, bearing the appearance of I_s and the motion of I_d. With a FLAME-derived 3D deformation field $\mathcal{D}_{neck}$ to handle neck pose and a volumetric renderer with 2D super-resolution, T can be rendered to a reenacted image I_{re} at an arbitrary view.

the MLP layers are for pose canonicalization. The cross-attention layers receive additional motion features of either I_s or I_d, and add them onto the global feature map of I_s for de-expression or reenactment accordingly, where the global feature map provides queries while the motion features provide keys and values, as depicted in Fig. 3. Notably, introducing Φ_{de} for expression neutralization is important. Otherwise, we find the self-attention and the MLP layers take over the de-expression job which easily leads to overfitting. What's more, this architecture enables us to disentangle the learning of 3D reconstruction and reenactment, by simply multiplying the output of all cross-attention layers by zeros or not. This strategy is crucial to the model's generalizability on real images and will be further discussed in Sec. 3.4.

Besides, the motion features provided to Φ are equally important as they are required to capture identity-irrelevant animations. To this end, we utilize a pre-trained motion encoder [62] as E_{mot}, which is learned via reconstructing large-scale monocular videos for faithful cross-identity reenactment. E_{mot} predicts a motion vector $\boldsymbol{v} \in \mathbb{R}^{548}$ from an image, which is then sent to Φ to compute the cross-attentions. Compared to FLAME's expression $\boldsymbol{\beta}$, $\boldsymbol{v}$ is more identity-agnostic, yielding better results when combined with our 4D synthetic data, as we will show in Sec. 4.3.

Finally, the reenacted feature map F_{re} after Φ is concatenated with the detail feature map F_{detail} from E_{detail}, and sent into a decoder G_T [60] to obtain the tri-plane T. Then, T is rendered to the reenacted image I_{re} via Eq. (3). Before the volume rendering process, we apply a FLAME-derived deformation field $\mathcal{D}_{neck}$ to handle the neck pose rotation similarly as in Sec. 3.1. Since the neck pose change yields almost uniform rigid transformations, this strategy is acceptable as it does not require highly accurate 3DMM reconstructions (see Sec. 4.3). Besides, we use a shallow U-Net to predict a 2D background feature map I_{bg} and blend it with the foreground I_f for rendering.

The whole reconstructor Ψ, except the fixed motion encoder E_{mot}, is trained end-to-end using the synthetic data from GenHead, as described below.

3.4. Disentangled Learning via Synthetic 4D Data

We learn the synthesis model Ψ in a self-reenacting manner, where we randomly choose two images of the same identity as I_s and I_d, respectively, and choose another image with I_d's identity and motion but a different camera view as the ground truth reenacted image $\bar{I}_{re}$. During training, Ψ is enforced to synthesize an image I_{re} via Eq. (3) to match the content of $\bar{I}_{re}$ at its viewing point, via a serial of losses which will be described later.

While the training process relies solely on synthetic data, we introduce a disentangled learning strategy to generalize the learned model to real data. Specifically, we find that the model's generalizability highly depends on the function of all non-cross-attention layers in Φ. By default, Φ is inclined to use the self-attention and MLP layers in Φ_{de} to handle both the expression neutralization and pose canonicalization of I_s (see the *suppl.* for an illustration), causing the network to quickly overfit to the synthetic identities.

To tackle this problem, our intuition is to disentangle the self-attentions and MLPs to focus on pose canonicalization only and let the cross-attention layers to deal with all the motion-related processes. We achieve this by simply multiplying zeros by the outputs of all the cross-attention layers at a fixed probability, and using the static data described in Sec. 3.2 to execute the self-reenacting process at the same time. This way, the training process randomly degenerates

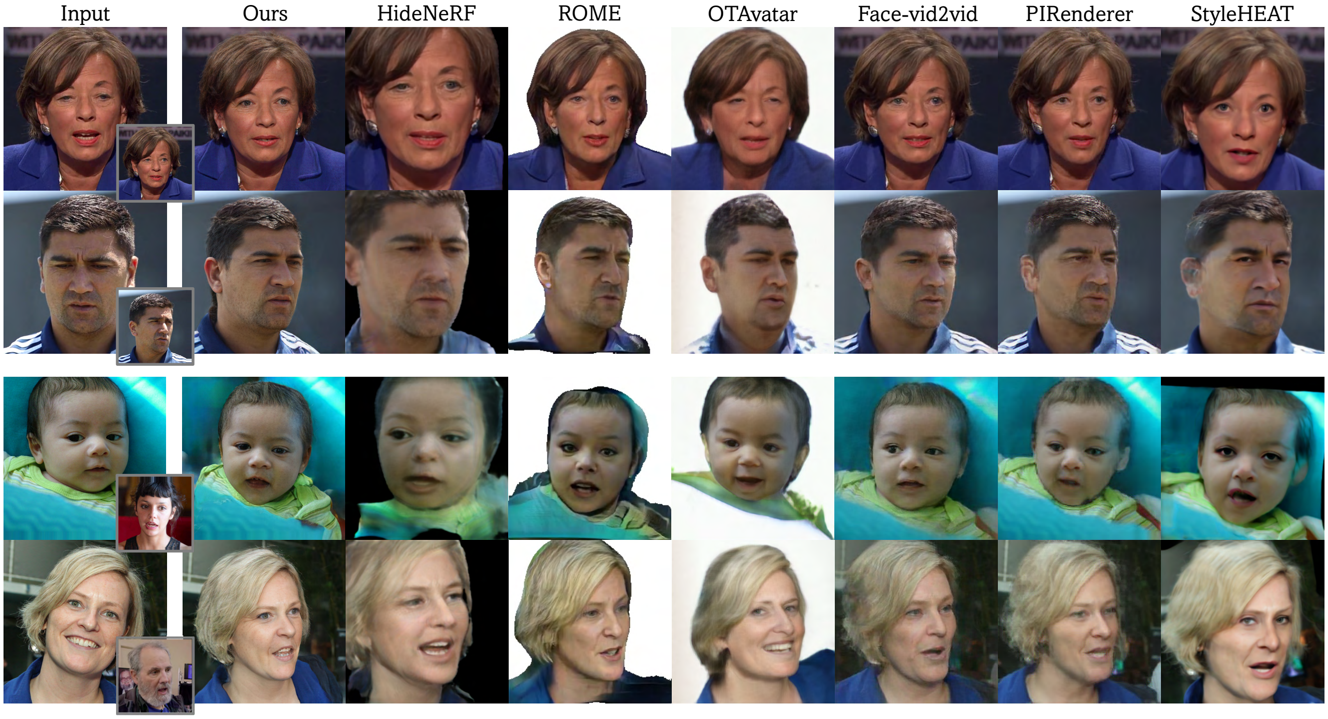

Figure 4. Qualitative comparison on one-shot head reenactment with previous methods. **Best viewed with zoom-in.**

to a static 3D reconstruction process for the remaining layers in Φ, meanwhile, all the cross-attention layers will be trained normally in a default reenactment process leveraging the dynamic data. Empirically, this strategy largely improves the reconstruction fidelity on real images (Sec. 4.3).

The overall training objective is as follows:

$$\mathcal{L} = \mathcal{L}_{re} + \mathcal{L}_f + \mathcal{L}_{tri} + \mathcal{L}_{depth} + \mathcal{L}_{opa} + \mathcal{L}_{id} + \mathcal{L}_{adv}, \quad (4)$$

where $\mathcal{L}_{re}$ calculates the perceptual difference [81] and L1 distance between I_{re} and its ground truth $\bar{I}_{re}$. $\mathcal{L}_f$ calculates the L1 distance between I_f, I_{bg} and their corresponding ground truth. $\mathcal{L}_{tri}$ computes the L1 difference between sampled tri-plane features $T(\boldsymbol{x})$ and $\bar{T}(\boldsymbol{x})$. Note that we do not calculate the difference between T and $[T_h, T_p]$ in Eq. (1) directly as in [60] because the tri-planes represent different geometries in Ψ and in GenHead. $\mathcal{L}_{depth}$ computes the L1 difference between I_{depth} and $\bar{I}_{depth}$. $\mathcal{L}_{opa}$ is the L1 difference between I_{opa} and $\bar{I}_{opa}$. $\mathcal{L}_{id}$ calculates the negative cosine similarity between face recognition features [14] of I_{re} and $\bar{I}_{re}$. $\mathcal{L}_{adv}$ is the adversarial loss between I_{re} and $\bar{I}_{re}$ leveraging the discriminator of GenHead.

4. Experiments

Implementation details. We train GenHead with a re-aligned version of FFHQ [33] at 512^2 resolution. For 4D data synthesis, we use 3DMM parameters extracted from FFHQ and VFHQ [72]. The camera poses are sampled from a pre-defined distribution. See the *suppl.* for more details.

Evaluation of GenHead. We leave the evaluation of GenHead in the *suppl. material*.

4.1. One-Shot 4D Head Synthesis Results

Figure 1 shows our one-shot head synthesis results on real images in FFHQ (more in the *suppl.*). Our method faithfully reconstructs a 3D head with rich details from a monocular image. The proxy 3D shapes extracted via Marching Cubes [42] depict reasonable 3D geometry recoveries, which is the key for 3D-consistent reenactment. We can freely change the neck pose of the reconstructed 3D heads, as well as varying their expressions given different driving images. We also support novel view synthesis thanks to the underlying 3D representation. Besides, backgrounds can be well separated during foreground animations. Our inference pipeline runs at 10 FPS on a single A100 GPU.

4.2. Comparison with Prior Arts

Baselines. We compare our method with existing one-shot head avatar synthesis methods, including 2D-based ones: PIRenderer [48], Face-vid2vid [64], and StyleHEAT [77]; and 3D-based ones: ROME [35], OTAvatar [43], and HideNeRF [39].

Metrics. We conduct self-reenactment and cross-identity reenactment on 100 test video clips in VFHQ. We use LPIPS and Fréchet Inception Distances (FID) [27] to measure the image synthesis quality, ID to measure the identity similarity [14] between the reenacted results and the ap-

Table 1. Self-reenactment results on VFHQ at 512^2. LPIPS$_h$ is for head region. 🟡, ⚪, and 🟤 denote the 1st, 2nd and 3rd places, respectively.

Method	All			Yaw Diff. $< 15°$			Yaw Diff. $15° \sim 30°$			Yaw Diff. $> 30°$		
	LPIPS$_h$↓	LPIPS ↓	FID ↓	ID ↑	AED ↓	APD ↓	ID ↑	AED ↓	APD ↓	ID ↑	AED ↓	APD ↓
PIRenderer [48]	0.214🟤	0.325	61.0🟤	0.820⚪	0.031🟤	1.245	0.647🟤	0.052	2.483🟤	0.536⚪	0.077🟤	9.326
Face-vid2vid [64]	0.187⚪	0.290🟡	52.6⚪	0.848🟡	0.025🟡	0.778⚪	0.665⚪	0.048⚪	4.035	0.502	0.098	24.99
StyleHEAT [77]	0.224	0.321🟤	67.3	0.649	0.050	1.392	0.524	0.068	2.727	0.420	0.106	10.89
ROME [35]	0.327	0.598	110	0.671	0.034	0.960🟤	0.589	0.046🟡	1.279⚪	0.497	0.076⚪	1.671⚪
OTAvatar [43]	0.291	0.557	112	0.448	0.077	3.306	0.339	0.089	7.396	0.251	0.177	8.157
HideNeRF [39]	0.281	0.418	80.5	0.820🟤	0.032	1.762	0.642	0.060	6.001	0.516🟤	0.084	7.879🟤
Ours	0.181🟡	0.320⚪	43.0🟡	0.790	0.030⚪	0.560🟡	0.668🟡	0.051🟤	0.811🟡	0.580🟡	0.074🟡	1.340🟡

Table 2. Cross-reenactment results on VFHQ dataset at 512^2.

Method	FID ↓	ID ↑	AED ↓	APD ↓
PIRenderer [48]	78.0⚪	0.582🟤	0.146🟡	3.241🟤
Face-vid2vid [64]	78.6🟤	0.606⚪	0.179	7.771
StyleHEAT [77]	81.1	0.499	0.165	5.932
ROME [35]	107	0.532	0.148⚪	2.936⚪
OTAvatar [43]	107	0.350	0.194	6.509
HideNeRF [39]	104	0.491	0.165	4.208
Ours	54.8🟡	0.620🟡	0.164🟤	1.020🟡

pearance images, Average Expression Distance (AED) [41] for expression control accuracy, and Average Pose Distance (ADP, ×1000 by default) [11] for pose control accuracy.

Qualitative results. Figure 4 shows visual comparisons on samples from VFHQ and FFHQ. Our method synthesizes head images with higher fidelity and definition compared to the alternatives. What's more, our method consistently yields better results compared to the others when the head pose differences between the sources and the drivings are significant, where we largely preserve the identities and head shapes. By contrast, the 2D-based methods experience large shape distortions. In addition, although HideNeRF, ROME, and OTAvatar also adopt 3D representations, their learning on monocular real data with imperfect 3D estimation impairs the geometry accuracy. More results can be found in the *suppl. material*.

Quantitative results. Table 1 and 2 show the quantitative comparisons on self-reenactment and cross-reenactment, respectively. In the self-reenactment, we divide the source and driving pairs into three groups based on their yaw angle differences. From Tab. 1, our reconstruction fidelity largely outperforms other 3D-based methods, and even surpasses two 2D-based approaches: PIRenderer and StyleHEAT. We also obtain comparable results with Face-vid2vid. Besides, our method yields the best pose accuracy, and achieves the highest identity similarities under large pose variations. This is inline with the visual results that our reconstructed geometries are more reasonable. We also demonstrate competitive expression control ability in terms of AED, whereas we solely learn on synthetic data without seeing real videos.

In the cross-reenactment in Tab. 2, we obtain the best image quality, identity similarity, and pose accuracy, with

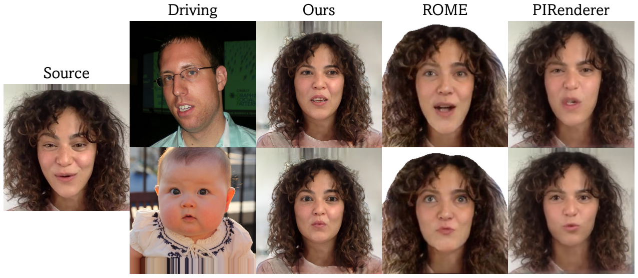

Figure 5. Reenactment results with different driving targets.

competitive expression accuracy. Note that although PIRenderer and ROME achieve lower AED, they tend to stretch the source geometry to match the absolute shapes of the drivings, leading to undesired identity changes, as shown in Fig. 5. This is a known identity leakage issue of the linear 3DMMs [38, 46], and can influence the AED metric which is also based on 3DMM coefficients. By contrast, our method is more robust to shape variations of the drivings, thanks to the identity-agnostic motion feature of [62].

4.3. Ablation Study

We conduct ablation studies on self-reconstruction and cross-reenactment using the first 1K identities in FFHQ to validate the efficacy of different components. ***A*** is a baseline that removes the cross-attention layers in Φ_{de}, the disentangled learning strategy, and the static data. ***B*** adds the learning strategy back by randomly selecting 50% identities in the dynamic data for static 3D reconstruction. ***C*** uses the complete structure of Φ_{de}. ***D*** combines *B* and *C* with both the above two components. ***E*** is our final configuration, with extra static data for the 3D reconstruction process during disentangled learning. ***F*** replaces the motion features of [62] with the FLAME codes $[\boldsymbol{\beta}, \gamma_{eye}, \gamma_{jaw}]$. ***G*** uses opacity images extracted via [13] instead of the rendered ones from GenHead for $\mathcal{L}_{opa}$. ***H*** utilizes monocular real data in FFHQ and VFHQ for training instead of the multi-view synthetic data. For efficiency, all configurations are trained to see 5000K images in total.

Module structure and learning strategy. From Tab. 3 and Fig. 6, the baseline setting leads to poor identity recon-

Table 3. Ablation study of our framework on FFHQ.

	Configurations	Recon.		Cross-Reenact.	
		LPIPS$_h$ ↓	ID ↑	AED ↓	APD ↓
A	Baseline	0.244	0.239	0.179	0.963
B	+ Disen. learning	0.238	0.281	0.182	0.994
C	+ Φ_{de}	0.247	0.227	0.179	0.963
D	B + C	0.212	0.441	0.183	0.947
E	D + Static (Ours)	0.188	0.703	0.193	1.022
F	with FLAME mot	0.213	0.451	0.150	0.879
G	with Detected $\bar{I}_{opa}$	0.198	0.552	0.189	1.080
H	with Real data	0.123	0.960	0.263	7.537

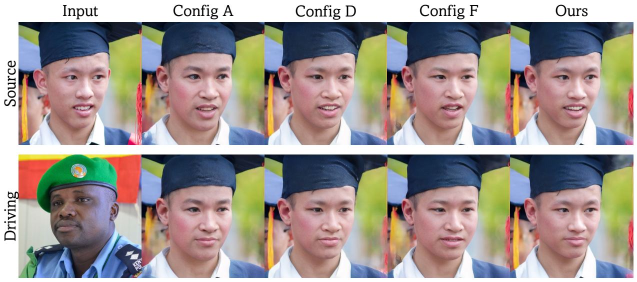

Figure 6. Reconstruction and driving results of different settings.

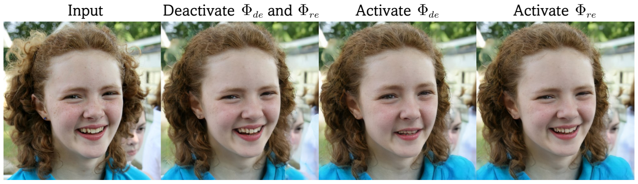

Figure 7. Function of learned Φ_{de} and Φ_{re}.

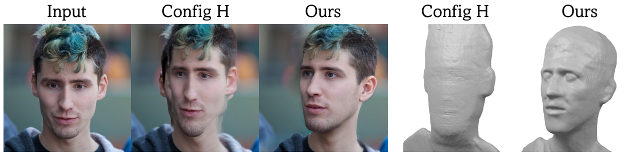

Figure 8. Comparison between models learned on different data.

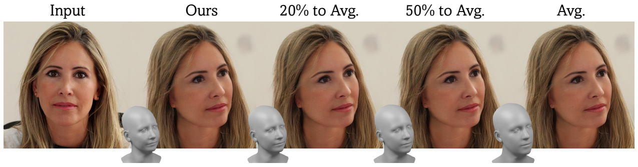

Figure 9. Neck pose control given different FLAME shapes.

struction results. Directly adding the disentangled learning strategy on top of the baseline slightly improves the reconstruction fidelity. Further introducing the de-expression module ensures a better disentanglement between 3D reconstruction and reenactment and effectively improves model's generalizability to real images. Finally, adding static data with more diverse identities for the 3D reconstruction process yields further improvement, with only minor influence to the reenactment accuracy. Figure 7 illustrates the function of the learned Φ_{de}, which turns the face into a canonical expression with mouth slightly open. Further sending the source motion to the reenactment module Φ_{re} after the de-expression step can faithfully recover the input image.

Influence of motion feature. As shown in Tab. 3, using the FLAME codes as motion feature leads to better AED and APD, yet the reconstruction fidelity largely decreases. This alternative encounters a similar issue as other 3DMM-based methods in Fig. 5. If the driving image has very different face features with those of the source, this setting can lead to a semantic change of the given expression or even identity change, as shown in Fig. 6. We conjecture that such identity leakage issue confuses the model during training, leading to inferior generalizability for reconstruction.

Synthetic data *vs.* real data. From Tab. 3, using detected opacity images is detrimental to the reconstruction fidelity. We conjecture that these detected results are not 3D-consistent across different views compared to the rendered synthetic ones, and thus increase the burden of the model to learn reasonable 3D reconstruction. What's more, if we replace the mult-view synthetic data with the monocular real ones, the learned geometry dramatically degenerates as indicated by the large APD and as shown in Fig. 8. This validates the importance of multi-view data with precise labels for learning reasonable 3D geometries.

Robustness of neck pose. Figure 9 demonstrates the robustness of our neck pose control given different FLAME shapes to derive $\mathcal{D}_{neck}$. We gradually deform an accurate FLAME mesh to an average shape. As shown, this process has negligible impact to the final synthesis quality.

5. Conclusions

We presented a framework for learning high-fidelity one-shot 4D head synthesis via large-scale synthetic data. The core idea is to first learn a 4D generative model on monocular images via adversarial learning for multi-view data synthesis of diverse identities and full motion control; then utilize a transformer-based animatable tri-plane reconstructor to learn 4D head reconstruction via the synthetic data. A disentangled learning strategy is also introduced to enhance model's generalizability to real images. Experiments have demonstrated our superiority over previous works, and revealed our potential for large-scale head avatar creation.

Limitations and ethics consideration. We leave the discussions in the *suppl. material* due to space limitation.

Acknowledgements

We thank Zixin Yin and Deyu Zhou for helping with the early exploration of GenHead.

References

[1] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20950–20959, 2023. 4

[2] Ronen Basri and David W Jacobs. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence*, 25(2):218–233, 2003. 3

[3] Alexander Bergman, Petr Kellnhofer, Wang Yifan, Eric Chan, David Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. *Advances in Neural Information Processing Systems*, 35:19900–19916, 2022. 2, 3

[4] V Blanz and T Vetter. A morphable model for the synthesis of 3d faces. In *26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1999)*, pages 187–194. ACM Press, 1999. 2

[5] Blender Foundation. Cycles renderer. https://www.cycles-renderer.org/, 2021. 3

[6] Timo Bolkart. Bfm to flame. https://github.com/TimoBolkart/BFM_to_FLAME, 2020. 4

[7] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Hyperreenact: one-shot reenactment via jointly learning to refine and retarget faces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7149–7159, 2023. 2

[8] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 2

[9] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020. 1

[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. 3

[11] 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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3, 4, 7

[12] Xingyu Chen, Yu Deng, and Baoyuan Wang. Mimic3d: Thriving 3d-aware gans via 3d-to-2d imitation. *arXiv preprint arXiv:2303.09036*, 2023. 3

[13] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 7

[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 6

[15] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 4

[16] 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. 3

[17] 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. 3

[18] Yu Deng, Baoyuan Wang, and Heung-Yeung Shum. Learning detailed radiance manifolds for high-fidelity and 3d-consistent portrait synthesis from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4423–4433, 2023. 3

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[20] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022. 1, 2

[21] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13, 2021. 2, 3

[22] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5609–5619, 2023. 2

[23] Stephan J Garbin, Marek Kowalski, Matthew Johnson, and Jamie Shotton. High resolution zero-shot domain adaptation of synthetically rendered face images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 220–236. Springer, 2020. 3

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 2, 4

[25] 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. 3

[26] Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo

Liu, Errui Ding, Ziwei Liu, et al. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1515, 2023. 2

[27] 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. 6

[28] Tomáš Hodaň, Vibhav Vineet, Ran Gal, Emanuel Shalev, Jon Hanzelka, Treb Connell, Pedro Urbina, Sudipta N Sinha, and Brian Guenter. Photorealistic image synthesis for object instance detection. In *2019 IEEE international conference on image processing (ICIP)*, pages 66–70. IEEE, 2019. 3

[29] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3397–3406, 2022. 2

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

[31] Zi-Hang Jiang, Qianyi Wu, Keyu Chen, and Juyong Zhang. Disentangled representation learning for 3d face shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11957–11966, 2019. 4

[32] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH*, 18(3):165–174, 1984. 4

[33] 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. 2, 6

[34] 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. 2, 4

[35] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *European Conference on Computer Vision*, pages 345–362. Springer, 2022. 1, 2, 3, 6, 7

[36] Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 3d gan inversion with pose optimization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2967–2976, 2023. 3

[37] Marek Kowalski, Stephan J Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. Config: Controllable neural face image generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 299–315. Springer, 2020. 3

[38] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2, 3, 7

[39] Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, and Xuelong Li. One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17969–17978, 2023. 1, 2, 4, 6, 7

[40] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot neural head avatar. *arXiv preprint arXiv:2306.08768*, 2023. 2

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

[42] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM SIGGRAPH*, 21(4):163–169, 1987. 6

[43] Zhiyuan Ma, Xiangyu Zhu, Guo-Jun Qi, Zhen Lei, and Lei Zhang. Otavatar: One-shot talking face avatar with controllable tri-plane rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16910, 2023. 1, 2, 3, 4, 6, 7

[44] 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. 2, 4

[45] 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. 3

[46] 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. 2, 7

[47] Weichao Qiu, Fangwei Zhong, Yi Zhang, Siyuan Qiao, Zihao Xiao, Tae Soo Kim, and Yizhou Wang. Unrealcv: Virtual worlds for computer vision. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1221–1224, 2017. 3

[48] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021. 2, 6, 7

[49] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 3

[50] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021. 3

[51] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned

implicit function for high-resolution clothed human digitization. In *IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 3

[52] 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. 3

[53] 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. 1, 2

[54] Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Kyle Olszewski, Jian Ren, Hsin-Ying Lee, Menglei Chai, and Sergey Tulyakov. Unsupervised volumetric animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4658–4669, 2023. 3

[55] 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. 3

[56] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *arXiv preprint arXiv:2205.15517*, 2022. 3

[57] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20991–21002, 2023. 2, 3, 4

[58] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and HongSheng Li. Controllable 3d face synthesis with conditional generative occupancy fields. *Advances in Neural Information Processing Systems*, 35: 16331–16343, 2022.

[59] 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

[60] Alex Trevithick, Matthew Chan, Michael Stengel, Eric Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. *ACM Transactions on Graphics (TOG)*, 42(4):1–15, 2023. 2, 3, 4, 5, 6

[61] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017. 3

[62] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17979–17989, 2023. 2, 5, 7

[63] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4563–4573, 2023. 3

[64] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 1, 2, 6, 7

[65] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018. 2

[66] E Wood, T Baltrusaitis, C Hewitt, M Johnson, J Shen, N Milosavljevic, D Wilde, S Garbin, T Sharp, I Stojiljkovic, et al. 3d face reconstruction with dense landmarks. arxiv 2022. *arXiv preprint arXiv:2204.02776*. 3

[67] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. 3

[68] 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. 2, 3

[69] Yue Wu, Sicheng Xu, Jianfeng Xiang, Fangyun Wei, Qifeng Chen, Jiaolong Yang, and Xin Tong. Aniportraitgan: Animatable 3d portrait generation from 2d image collections. *arXiv preprint arXiv:2309.02186*, 2023. 2, 3

[70] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2

[71] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 321–331, 2023. 3

[72] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 6

[73] Hongyi Xu, Guoxian Song, Zihang Jiang, Jianfeng Zhang, Yichun Shi, Jing Liu, Wanchun Ma, Jiashi Feng, and Linjie Luo. Omniavatar: Geometry-guided controllable 3d head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12814–12824, 2023. 2, 3

[74] Sicheng Xu, Jiaolong Yang, Dong Chen, Fang Wen, Yu Deng, Yunde Jia, and Xin Tong. Deep 3d portrait from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7710–7720, 2020. 1

[75] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 775–791. Springer, 2020. 3

[76] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*, 41(6): 1–21, 2022. 3

[77] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pretrained stylegan. *arXiv preprint arXiv:2203.04036*, 2022. 1, 2, 6, 7

[78] Yu Yin, Kamran Ghasedi, HsiangTao Wu, Jiaolong Yang, Xin Tong, and Yun Fu. Nerfinvertor: High fidelity nerf-gan inversion for single-shot real image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8539–8548, 2023. 3

[79] Wangbo Yu, Yanbo Fan, Yong Zhang, Xuan Wang, Fei Yin, Yunpeng Bai, Yan-Pei Cao, Ying Shan, Yang Wu, Zhongqian Sun, et al. Nofa: Nerf-based one-shot facial avatar reconstruction. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12, 2023. 1, 2, 3, 4

[80] Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, Hsiang-Tao Wu, Dong Chen, Qifeng Chen, Yong Wang, and Fang Wen. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22096–22105, 2023. 1, 2

[81] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[82] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7194–7202, 2019. 3

[83] Wenbin Zhu, HsiangTao Wu, Zeyu Chen, Noranart Vesdapunt, and Baoyuan Wang. Reda:reinforced differentiable attribute for 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[84] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 41(1): 78–92, 2017. 3