# Portrait4D-v2: Pseudo Multi-View Data Creates Better 4D Head Synthesizer

Yu Deng, Duomin Wang, and Baoyuan Wang

Xiaobing.AI
https://yudeng.github.io/Portrait4D-v2/

**Abstract.** In this paper, we propose a novel learning approach for feed-forward one-shot 4D head avatar synthesis. Different from existing methods that often learn from reconstructing monocular videos guided by 3DMM, we employ pseudo multi-view videos to learn a 4D head synthesizer in a data-driven manner, avoiding reliance on inaccurate 3DMM reconstruction that could be detrimental to the synthesis performance. The key idea is to first learn a 3D head synthesizer using synthetic multi-view images to convert monocular real videos into multi-view ones, and then utilize the pseudo multi-view videos to learn a 4D head synthesizer via cross-view self-reenactment. By leveraging a simple vision transformer backbone with motion-aware cross-attentions, our method exhibits superior performance compared to previous methods in terms of reconstruction fidelity, geometry consistency, and motion control accuracy. We hope our method offers novel insights into integrating 3D priors with 2D supervisions for improved 4D head avatar creation.

**Keywords:** Head avatar · Video reenactment · NeRF

## 1 Introduction

One-shot head avatar synthesis has garnered significant attention in recent years among computer vision and graphics community. Its typical problem setting is to generate photorealistic portrait videos given a source image for appearance and a series of driving motions (from videos) for animation. Over time, the technical solution has witnessed a paradigm shift from leveraging 2D generative models [3, 11, 37, 41, 46] to animatable 3D (*i.e.*, 4D) synthesizers [24, 28, 38, 48], where the latter maintain better geometry consistency under significant head motions and support free-view rendering favored in AR/VR scenarios.

Different from 2D-based methods that directly generate 2D frames via neural networks, 3D approaches synthesize head avatars in 3D space and leverage camera projection from 3D to 2D to produce the final result. While the 2D methods are often learned in a self-supervised manner using monocular videos [3, 41], the 3D counterparts are difficult to follow a similar procedure due to the scarcity of 3D data. Learning the 3D approaches via weak supervisions from the monocular videos often turns to be highly ill-posed without resorting to proper 3D priors. Although plenty of existing methods [24, 28, 29, 32, 48] utilize head models such

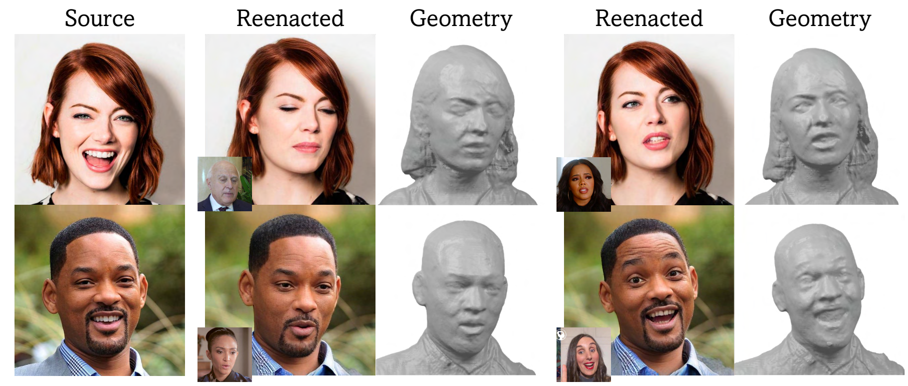


**Fig. 1:** Our method utilizes a feed-forward 4D head synthesizer to create photorealistic head avatars from a single source image. The facial expressions and neck pose of the 3D heads can be precisely controlled by another driving frame (*e.g.*, see the mouth, eye gaze, and forehead wrinkles). The synthesized results also support free-view rendering, thanks to the underlying accurate head geometries. **Best viewed with zoom-in.**

as 3D Morphable Models (3DMMs) [1, 27, 35] to help alleviate this problem, the inaccuracy of monocular 3DMM reconstructors [10, 12] and the limited expressiveness of a 3DMM itself often restrict them from synthesizing lifelike head avatars.

Ideally, if we have access to large-scale multi-view videos, it is possible to learn an image-to-4D synthesizer in a data-driven manner following recent successes in the 3D reconstruction field [17, 26, 39, 55]. However, collecting such large amount of data with enough diversity for learning a generalizable head synthesizer can be impractical. A recent method Portrait4D [9] learns a 4D GAN in advance to synthesize large-scale multi-view data for training, yet the learned 4D GAN still requires 3DMM for expression control which is less vivid than real data.

This paper suggests a more practical way for learning a one-shot 4D head synthesizer using *pseudo multi-view videos*. The key intuition is to turn existing monocular videos into multi-view ones by learning a static head novel view synthesizer (*i.e.* 3D synthesizer) in advance. This brings us several advantages: **(1)** Monocular videos exhibit high scalability for learning a generalizable head synthesizer thanks to an easy access to public datasets [34, 43, 53] and Internet videos. **(2)** Monocular head novel view synthesis has been proven achievable through knowledge distillation from a 3D GAN learned solely with in-the-wild images [39], thereby enabling the creation of multi-view videos from monocular inputs with minimal effort. **(3)** Multi-view videos would facilitate geometry learning and enable using motion representation learned from expressive data beyond 3DMM description. These factors help us avoid a heavy reliance on 3DMM for geometry reconstruction [5, 24], expression manipulation [28, 29, 45], or pose estimation [28, 32, 48], leading to a faithful 4D head synthesizer with accurate geometry prediction and vivid motion control (see Fig. 1).

Specifically, we adopt a vision-transformer-based encoder-decoder architecture for the head synthesizer following Portrait4D [9] to predict triplane-based head NeRF [4,33] from a source image. A motion embedding [40] extracted from a driving image is responsible for controlling expression of the predicted tri-planes via motion-related cross-attention layers. At the first stage, we disable all motion-related layers and enforce the synthesizer to learn static tri-plane reconstruction from source images using synthetic data from a pre-trained 3D GAN [9]. This yields a powerful 3D head synthesizer that can well generalize to real data. At the second stage, we start from the 3D synthesizer and activate all the motion-related components to learn the full model using real videos. We maintain a fixed copy of the pre-trained 3D synthesizer and use it to transform monocular video frames into multi-view ones for cross-view self-reenactment, where we extract motion embedding from a driving frame at an arbitrary viewpoint and learn to reconstruct the same frame from a different random perspective. This way, the geometry priors of the 3D synthesizer can be seamlessly distilled into the full 4D head synthesizer meanwhile detailed motion control can be learned via the self-reenacting process on real videos. Considering that the 3D synthesizer is imperfect, we focus more on reconstructing the original real driving frames and treat any other synthetic multi-view frames as regularization.

It is worth emphasizing that our intention is not to design an advanced network architecture but to explore ***a novel learning paradigm*** suitable for one-shot 4D head synthesis by creating pseudo multi-view videos from existing monocular ones. This yields a fundamental difference between our method and the previous ones that learn from reconstructing monocular video frames [5, 28, 38, 45, 48]. Certainly, our learning paradigm can be incorporated into these frameworks as well. We demonstrate that by simply using a transformer backbone following [9], our method (*i.e.*, Portrait4D-v2) largely outperforms previous approaches in terms of reconstruction fidelity, geometry consistency, and motion control accuracy. We hope that our method will serve as inspiration for future works, encouraging a more seamless integration of 3D priors with in-the-wild 2D data for creating generalizable 4D head avatar synthesizers.

## 2 Related Work

We review approaches falling in the scenario of one-shot head avatar synthesis, which perform subject-agnostic head creation and animation given a single appearance image. We categorize them into 2D-based and 3D-aware ones, depending on whether or not they leverage an explicit camera model for rendering.

***2D-based talking head generation.*** The success of CNNs in image generation [19, 23] have given rise to plenty of approaches leveraging 2D networks for direct head image synthesis [3, 13, 41]. A common strategy among them is to interwine the latent features of appearance and motion within the 2D generative network to achieve photorealisitic and animatable image generation. An earlier work [49] injects appearance embedding into a U-Net backbone via

AdaIN [18] to transfer driving landmarks to reenacted images. [3] and follow-up works [40, 52] learn disentangled latent representations of appearance and motion as input to a StyleGAN-like [23] generator. A prevailing trend in recent methods [11, 15, 36, 37, 41, 50] involves enforcing a warping field onto intermediate appearance feature maps, leveraging the inherent characteristics of facial movements. However, these methods fall short in maintaining the head geometry consistency under large pose changes due to a lack of 3D understanding. While some methods [11,41] introduce 3D inductive bias into the framework, they cannot fully alleviate the inconsistency issue with their learned non-physical projection process from 3D to 2D. In addition, none of the above methods support free-view rendering owing to the absence of an explicit camera model.

***3D-aware head avatar synthesis.*** To allow free-view rendering with more strict 3D consistency, recent works incorporate 3D representations and perspective camera projection into the head image synthesis process. Earlier methods [24,44] utilize morphable meshes to represent head geometries and textures, and successive works [5, 16, 28, 29, 32, 45, 48] leverage more advanced representations such as NeRFs [4,33] to better model hairs and other accessories. Notably, introducing the 3D representations poses significant challenges for the model to learn monocular 3D reconstruction and animation from 2D videos, which is highly ill-posed without leveraging 3D priors or multi-view data. Consequently, some methods [16,54] resort to multi-view videos collected from constrained environment as supervision, yet the limited diversity of the training data restricts their generalizability to unconstrained images. Another line of works stick to in-the-wild videos as supervision and leverage monocular 3DMM reconstructors [7, 10, 12] to facilitate 3D geometry and motion learning, by using reconstructed 3DMM meshes for head modeling [24, 44], estimating camera poses for feature transformation [28] and image rendering [32, 48], or utilizing 3DMM expressions for animation synthesis [5,28,29,45]. Nevertheless, the limited reconstruction accuracy and expressiveness of 3DMM hinder them from synthesizing faithful head avatars. [38] avoids using 3DMM and achieves expression disentanglement in a data-driven manner leveraging a 3D head synthesizer [39] for pose canonicalization. Still, it uses monocular videos as supervision which differs from our learning paradigm. A recent method Portrait4D [9] learns from synthetic data of multi views. However, its data is generated by a 4D GAN with 3DMM-based expression control, which exhibits a non-negligible domain gap when compared to real videos of subtle facial expressions, and thus can limit the generalizability of the learned model.

## 3 Preliminaries: Triplane-Based 3D Representation

We follow previous successes [4, 9, 28, 39] to model 3D head via a triplane-based NeRF representation [4]. Tri-plane depicts a 3D volumetric space using three orthogonal 2D feature planes $\mathbf{T}_{xy}, \mathbf{T}_{yz}, \mathbf{T}_{zx}$, each representing projected 3D features onto corresponding axis-aligned planes $xy, yz, zx$, respectively. Given a point $\mathbf{x} \in \mathbb{R}^3$, its volume density $\sigma \in \mathbb{R}$ and color $\mathbf{c} \in \mathbb{R}^{d_c}$ can be decoded from the tri-plane via:

$$(\mathbf{c}, \sigma) = \mathrm{MLP}(\mathbf{T}_{xy}(\mathbf{x}) + \mathbf{T}_{yz}(\mathbf{x}) + \mathbf{T}_{zx}(\mathbf{x})), \tag{1}$$

where $\mathbf{T}.(\mathbf{x})$ denotes projecting $\mathbf{x}$ onto the corresponding plane for feature acquisition, and MLP decodes the summed feature to the final radiance (*i.e.*, color and density). Following volume rendering process [21] as in NeRF [33], one can obtain a 2D feature image $I_c$ (usually the first three channels represent RGB color) from the 3D space by accumulating radiance of each point along rays from a given camera viewpoint $\boldsymbol{\theta}$. The rendered image $I_c$ is then sent into a 2D super-resolution module [4, 42] yielding a final synthesized image $I$ at a higher resolution. For brevity, we use a single renderer $\mathcal{R}$ to represent the above rendering process from the tri-plane $\mathbf{T}$ to the final image $I$, leading to $I = \mathcal{R}(\mathbf{T}, \boldsymbol{\theta})$.

## 4 Method

Our goal is to synthesize an animatable 3D head represented by the tri-plane $\mathbf{T}$ from a source image $\hat{I}_s$, where $\mathbf{T}$ faithfully reconstructs the source appearance and its head motion mimics that from another driving image $\hat{I}_d$. We adopt the feed-forward backbone of Portrait4D [9] which directly predicts $\mathbf{T}$ from the input images via a transformer-based reconstructor $\Psi$ (Sec. 4.1). To learn $\Psi$ for faithful head avatar synthesis, we create pseudo multi-view videos from monocular ones as supervision, by first learning a 3D head synthesizer $\Psi_{3d}$ to generate novel views of each video frame (Sec. 4.2), and then leveraging the synthesized multi-view videos to perform cross-view self-reenactment for training $\Psi$ (Sec. 4.3). Figure 2 shows the overview and we describe each part in detail below.

### 4.1 Revisiting Portrait4D Reconstructor

We begin by revisiting the tri-plane reconstructor $\Psi$ of Portrait4D which uses a hybrid structure of CNNs and Vision Transformer (ViT) blocks. Specifically, two CNN encoders are first leveraged to extract global and detail appearance feature maps from the source image, respectively. The global feature map is then sent to several ViT blocks for pose canonicalization, and further concatenated with the detail feature map and fed into a ViT decoder with convolutions to generate the final tri-plane $\mathbf{T}$. To inject motion information for reenactment, Portrait4D introduces additional cross-attention layers into each canonicalization ViT block, where the cross-attentions in the first several blocks receive motion embedding $\boldsymbol{v}_s$ of the source image for expression neutralization and the remaining receive that of the driving image (*i.e.*, $\boldsymbol{v}_d$) for reenactment (see Fig. 2). For faithful expression reconstruction, Portrait4D leverages a pre-trained 2D motion encoder $E_{mot}$ [40] learned from video reenactment to extract the motion embedding.

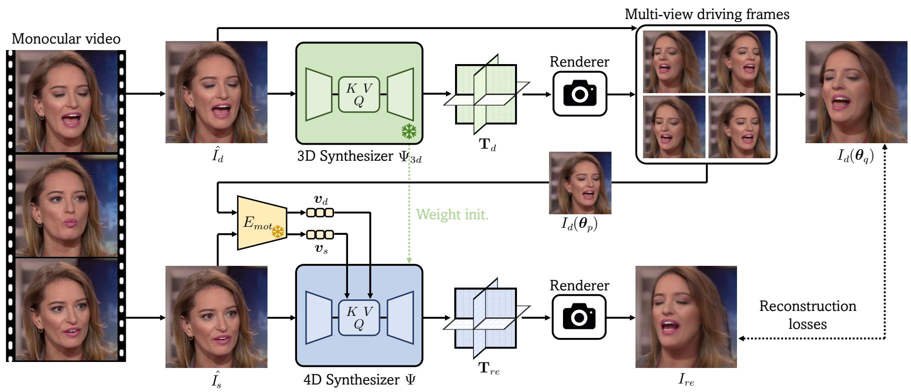


**Fig. 2:** Overview of our approach. Given a monocular training video, we first leverage a pre-trained 3D synthesizer $\Psi_{3d}$ to turn each driving frame within the video into multi-view one, and then use the pseudo multi-view driving frames and a source frame sampled from the original video to perform cross-view self-reenactment for learning a feed-forward 4D head synthesizer $\Psi$. After training, $\Psi$ can synthesize an animatable 3D head given two arbitrary images as the source and driving, respectively.

***Characteristics of the reconstructor.*** An advantage of $\Psi$ is that it is purely data-driven without being restricted by 3DMM-related inductive bias as in many previous approaches [28, 45, 48]. However, it also necessitates extensive and precise 3D data for effective training of the 4D head synthesizer. Otherwise, the network is inclined to synthesize 3D head with flattened geometries (see Fig. 7). Portrait4D addresses this challenge by training on 4D synthetic data generated by a pre-trained GAN [9], yet the synthesized expressions remain confined to the 3DMM description space which cannot fully exploit the expressive capability of the motion embedding. Besides, the synthetic identities have domain gaps with the real ones which can be detrimental to the model's generalizability.

An interesting observation is that if we remove the motion-related cross-attentions in $\Psi$ and use the same synthetic data from Portrait4D to train a static 3D head synthesizer[1], it can well generalize to real images with nuanced expressions hard to be described by 3DMM, as shown in Fig. 3a. This inspires us that instead of generating 4D data directly from scratch as supervision, it is more practical to learn a 3D head synthesizer in advance and use it to create multi-view videos from existing monocular ones for training, as described below.

### 4.2 3D Synthesizer for Multi-View Video Creation

We learn a 3D head synthesizer to turn an arbitrary frame to its corresponding tri-plane for free-view rendering. Following Sec. 4.1, we adopt the reconstructor $\Psi$ as backbone with all its cross-attention layers disabled, denoted as $\Psi_{3d}$.

[1] Under this circumstance, the problem degenerates to that of [39] for head novel view synthesis, which has been proven achievable using synthetic data from EG3D [4].

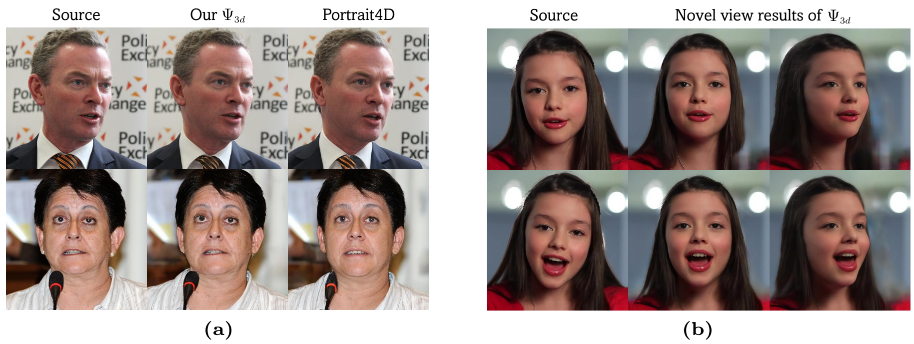


**Fig. 3:** (a) Self-reconstruction comparison between $\Psi_{3d}$ and [9] learned with the same data. $\Psi_{3d}$ yields better reconstruction fidelity. (b) $\Psi_{3d}$ learned using static images is capable of maintaining geometry consistency across different frames within a video clip.

While previous methods [4, 38, 45] have learned similar 3D head synthesizers using synthetic data from EG3D [4], we do not follow them but leverage the data generated by the 4D GAN of Portrait4D (namely GenHead [9]) as supervision. Compared to EG3D, GenHead separates foreground head with background image, and its generated data are aligned with the scale of FLAME [27], allowing further neck pose rotation via an $SO(3)$ field derived from the FLAME mesh.

***Learning the 3D synthesizer.*** Specifically, GenHead requires FLAME parameters as input condition for data synthesis. Therefore, we first apply off-the-shelf 3DMM reconstructors [2, 10] with further landmark-based optimization to in-the-wild images to extract FLAME parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ from them, where $\boldsymbol{\alpha} \in \mathbb{R}^{300}$ is a shape code, $\boldsymbol{\beta} \in \mathbb{R}^{100}$ is an expression code, and $\boldsymbol{\gamma} \in \mathbb{R}^{9}$ is a pose code. We randomly sample $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ triplet among the extracted parameter set and combine it with a random Gaussian noise $\boldsymbol{z} \in \mathbb{R}^{512}$ as input to GenHead to create arbitrary 3D heads. We then sample camera extrinsics $\boldsymbol{\theta} \in SE(3)$ from a pre-defined uniform distribution and render each generated 3D head to its multi-view images $\{\bar{I}(\boldsymbol{\theta}_i)\}$. To learn the 3D synthesizer $\Psi_{3d}$, we send an arbitrary synthetic image $\bar{I}(\boldsymbol{\theta}_p)$ of a 3D head to $\Psi_{3d}$, and enforce it to reconstruct a tri-plane $\mathbf{T}$ of the input that is consistent with another image $\bar{I}(\boldsymbol{\theta}_q)$ of the 3D head when rendered at an arbitrary view $\boldsymbol{\theta}_q$. Following [9], we apply multi-level supervisions to $\Psi_{3d}$ to ensure its faithful reconstruction of the synthetic data:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_{\text{LPIPS}} + \mathcal{L}_{id} + \mathcal{L}_{adv} + \mathcal{L}_{depth} + \mathcal{L}_{opa} + \mathcal{L}_{T}, \tag{2}$$

where $\mathcal{L}_1$ is the L1 distance between reconstructed images of $\Psi_{3d}$ and the synthetic images of GenHead, $\mathcal{L}_{\text{LPIPS}}$ is the perceptual difference [51] between them, $\mathcal{L}_{id}$ calculates negative cosine similarity between face recognition features [8] of the images, $\mathcal{L}_{adv}$ is the adversarial loss of [9], $\mathcal{L}_{depth}$ and $\mathcal{L}_{opa}$ calculate L1 difference of rendered depth and opacity images between $\Psi_{3d}$'s reconstructions and the synthetic data, respectively, and $\mathcal{L}_{T}$ computes the L1 distance of sampled tri-plane features between the reconstructions and the ground truth.

***Synthesizing multi-view videos.*** We use the learned $\Psi_{3d}$ to turn a monocular video $\hat{V} = \{\hat{I}_t\}_{t=1:T}$ to its multi-view version $V_{mv} = \{\{I_t(\boldsymbol{\theta}_i)\}_{i=1:N}\}_{t=1:T}$ via

$$I_t(\boldsymbol{\theta}_i) = \mathcal{R}(\mathbf{T}_t, \boldsymbol{\theta}_i) = \mathcal{R}(\Psi_{3d}(\hat{I}_t), \boldsymbol{\theta}_i), \tag{3}$$

where $\mathcal{R}$ is the renderer defined in Sec. 3. Although $\Psi_{3d}$ is learned using static multi-view images, we empirically find that it can maintain the geometry consistency between different frames in a video clip to some extent (see Fig. 3b). What's more, it better preserves detailed shape and expression of a real frame compared to Portrait4D (as in Fig. 3a), even though they are trained with the same synthetic data generated by the 3DMM-conditioned 4D GAN. We conjecture this is due to that: (1) the static reconstruction process utilizes less 3DMM inductive bias within the data as it does not tackle expression change constrained by 3DMM deformtation; (2) the synthetic data have certain degrees of freedom beyond 3DMM thanks to the additional noise $\boldsymbol{z}$. With the pseudo multi-view videos created via Eq. (3), we learn our final 4D head synthesizer $\Psi$ as follows.

### 4.3 Cross-View Self-Reenactment Learning

We learn the 4D synthesizer $\Psi$ via cross-view self-reenactment, leveraging real videos $\hat{V}$ and their corresponding multi-view versions $V_{mv}$. As shown in Fig. 2, during each iteration, we randomly select a source frame $\hat{I}_s$ from $\hat{V}$ and two free-view driving images $I_d(\boldsymbol{\theta}_p)$ and $I_d(\boldsymbol{\theta}_q)$ of the same motion from $V_{mv}$. We let the synthesizer $\Psi$ to take $\hat{I}_s$ and $I_d(\boldsymbol{\theta}_p)$ as input and synthesize a reenacted image $I_{re}$ at $I_d(\boldsymbol{\theta}_q)$'s camera viewpoint $\boldsymbol{\theta}_q$:

$$I_{re} = \mathcal{R}(\mathbf{T}, \boldsymbol{\theta}_q) = \mathcal{R}(\Psi(\hat{I}_s, \boldsymbol{v}_s, \boldsymbol{v}_d), \boldsymbol{\theta}_q), \tag{4}$$

where $\boldsymbol{v}_s$ and $\boldsymbol{v}_d$ are the motion embeddings of $\hat{I}_s$ and $I_d(\boldsymbol{\theta}_p)$ extracted via the motion encoder $E_{mot}$ [40], respectively. We impose $I_{re}$ to match the content of $I_d(\boldsymbol{\theta}_q)$ through a set of losses:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_{\text{LPIPS}} + \mathcal{L}_{id} + \mathcal{L}_{adv}, \tag{5}$$

where the four terms of $\mathcal{L}$ are similarly defined in Eq. (2) and calculated between $I_{re}$ and $I_d(\boldsymbol{\theta}_q)$. We initialize $\Psi$ using the pre-trained weights of $\Psi_{3d}$ except the motion-related components which are initialized randomly.

Considering that $I_d(\boldsymbol{\theta}_p)$ and $I_d(\boldsymbol{\theta}_q)$ are generated by an imperfect 3D synthesizer, we replace either of them with the original driving frame $\hat{I}_d$ from $V$ with certain possibilities to let the 4D synthesizer $\Psi$ be supervised by reenacting the real videos as well. We substitute $I_d(\boldsymbol{\theta}_p)$ with a probability of 10% and $I_d(\boldsymbol{\theta}_q)$ with a probability of 80%. Therefore, $\Psi$ concentrates more on reconstructing the original real frame $\hat{I}_d$ while the pseudo free-view frame $I_d(\boldsymbol{\theta}_q)$ serves more like a geometry regularization. We apply a lower substitution rate for $I_d(\boldsymbol{\theta}_p)$ and a higher one for $I_d(\boldsymbol{\theta}_q)$, this is because the former is only used for extracting high-level motion embedding where the imperfect result has less influence compared to that of the latter used for supervising pixel-wise image reconstruction. Besides, we use all the four losses in Eq. (5) if $\hat{I}_d$ is used as the ground truth and otherwise (*i.e.*, $I_d(\boldsymbol{\theta}_q)$ is used) only $\mathcal{L}_{\text{LPIPS}}$.

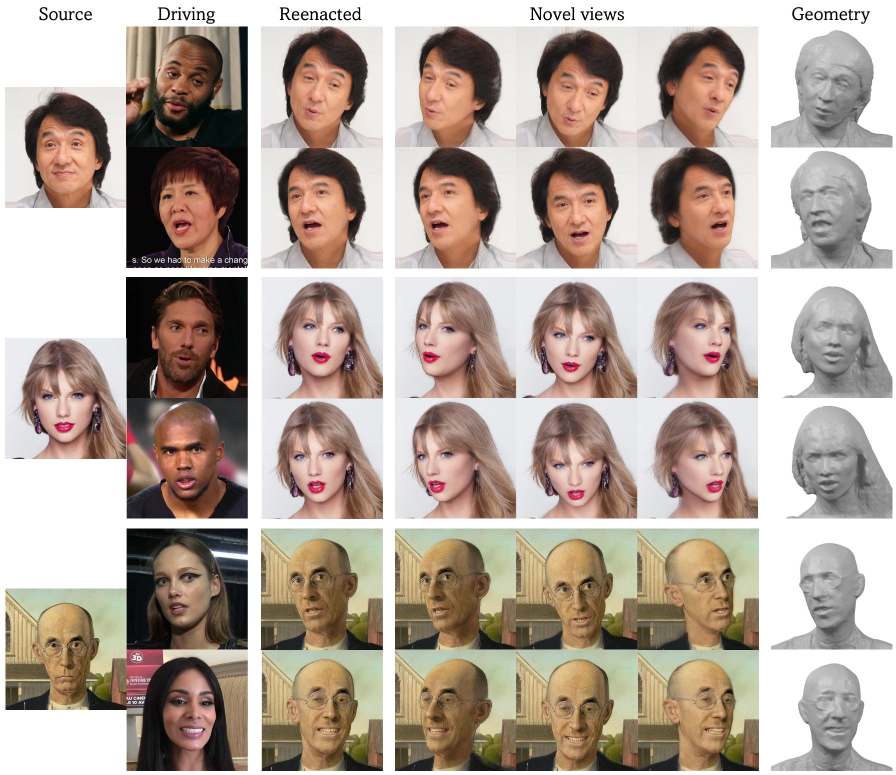


**Fig. 4:** One-shot head synthesis results of our method on in-the-wild images.

***Effect of cross-view learning.*** Our learning process distills the geometry and appearance priors within the 3D synthesizer $\Psi_{3d}$ into the 4D synthesizer $\Psi$, as the latter is enforced to mimic the reconstruction of an arbitrary image $I_d(\boldsymbol{\theta}_q)$ by the former at any camera pose $\boldsymbol{\theta}_q$. This yields more plausible reconstruction result for unseen regions compared to learning with only monocular supervision, as we will show in Sec. 5.3. What's more, our strategy enables using a non-3D-aware motion embedding learned from 2D reenactment [40]. By leverage $I_d(\boldsymbol{\theta}_p)$ at an arbitrary view $\boldsymbol{\theta}_p$ to extract the motion embedding $\boldsymbol{v}_d$ for reenactment, any pose-related information within $\boldsymbol{v}_d$ can be squeezed out. As a result, $\Psi$ can well utilize the detailed expressive ability of the 2D motion embedding meanwhile ignore its pose-related information to achieve strong 3D consistency.

## 5 Experiment

***Implementation.*** To learn $\Psi_{3d}$, we use GenHead [9] trained on FFHQ [22] at $512^2$ to synthesize multi-view supervisions. We adopt an Adam optimizer [25] with a learning rate of $1e-4$ and a batch size of 32, and train $\Psi_{3d}$ to see 10M

**Table 1:** Self- and cross-ID reenactment results on VFHQ at $512^2$. LPIPS$_h$ is for head region only. ●, ●, and ● denote the 1st, 2nd and 3rd places, respectively.

| Method | Self reenactment | | | | | | Cross-ID reenactment | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | LPIPS$_h$↓ | LPIPS ↓ | FID ↓ | ID ↑ | AED ↓ | APD ↓ | FID ↓ | ID ↑ | AED ↓ | APD ↓ |
| PIRenderer [36] | 0.213 | 0.323 | 56.2 | 0.801 | 0.034 | 1.402 | 78.6 | 0.582 | 0.146 | 2.833 |
| Face-vid2vid [41] | 0.187 | 0.289 | 52.0 | 0.826 | 0.028 | 1.603 | 79.2 | 0.606 | 0.178 | 5.188 |
| StyleHEAT [46] | 0.224 | 0.320 | 66.9 | 0.633 | 0.053 | 1.543 | 81.7 | 0.499 | 0.165 | 5.864 |
| ROME [24] | 0.326 | 0.594 | 99.8 | 0.660 | 0.036 | 1.007 | 108 | 0.532 | 0.148 | 2.294 |
| OTAvatar [32] | 0.291 | 0.556 | 110 | 0.436 | 0.080 | 3.650 | 107 | 0.350 | 0.194 | 5.779 |
| HideNeRF [28] | 0.282 | 0.417 | 74.5 | 0.800 | 0.035 | 2.182 | 105 | 0.491 | 0.166 | 4.242 |
| Real3DPortrait [45] | 0.190 | 0.283 | 44.6 | 0.802 | 0.029 | 1.180 | 61.3 | 0.721 | 0.183 | 3.951 |
| GPAvatar [5] | 0.191 | 0.418 | 63.6 | 0.786 | 0.028 | 0.747 | 80.6 | 0.570 | 0.154 | 2.409 |
| GOHA [29] | 0.179 | 0.373 | 52.0 | 0.677 | 0.048 | 0.697 | 68.7 | 0.493 | 0.169 | 1.152 |
| Portrait4D [9] | 0.181 | 0.320 | 43.0 | 0.773 | 0.033 | 0.612 | 54.8 | 0.620 | 0.164 | 1.020 |
| Ours | 0.139 | 0.224 | 26.3 | 0.873 | 0.019 | 0.580 | 48.5 | 0.736 | 0.161 | 0.858 |

**Table 2:** (a) Evaluation on cross-ID expression similarity. (b) A user study, where the participants are asked to choose one best method among all.

**(a)** Cross-ID expression accuracy

| Method | AED$_f$ ↓ | SyncNet ↑ |
|---|---|---|
| PIRenderer [36] | 0.209 | 3.11 |
| ROME [24] | 0.212 | 2.99 |
| Real3DPortrait [45] | 0.252 | 3.49 |
| GPAvatar [5] | 0.200 | 2.95 |
| Portrait4D [9] | 0.205 | 3.23 |
| Ours | **0.185** | **3.66** |

**(b)** User study

| Method | Image quality ↑ | Expression Acc. ↑ |
|---|---|---|
| PIRenderer [36] | 0.47% | 1.40% |
| ROME [24] | 0.47% | 3.72% |
| Real3DPortrait [45] | 0.47% | 2.33% |
| GPAvatar [5] | 6.28% | 14.7% |
| Portrait4D [9] | 11.6% | 6.28% |
| Ours | **80.7%** | **71.6%** |

images in total. Then, we learn the 4D synthesizer $\Psi$ using a training split of 10K video clips from VFHQ [43]. The model is trained to see 6M images, using Adam with a learning rate of $2.5e-4$ for the motion-related layers and $1e-4$ for the remaining, and a batch size of 32. See the *suppl. material* for more details.

### 5.1 One-Shot Head Synthesis Results

Figure 1 and 4 demonstrate our one-shot head reenactment results on in-the-wild images. Our method faithfully reconstructs the source appearance meanwhile mimics the nuanced expressions in different driving images. It can also model the dynamic wrinkles brought by exaggerated facial motions. What's more, our method supports free view rendering of the reconstructed heads thanks to the underlying tri-plane representation. The accurately reconstructed head geometries (visualized via Marching Cubes [31]) guarantee strong 3D consistency across different views. While our method is trained on real videos, it can also handle artistic portraits as depicted by the last case in Fig. 4.

***Inference speed.*** Our method runs at 10 FPS with a batch size of 1 on an A100 GPU using the naive Pytorch framework without specialized acceleration. Further caching the neutralized feature map of the source image leads to 15 FPS.

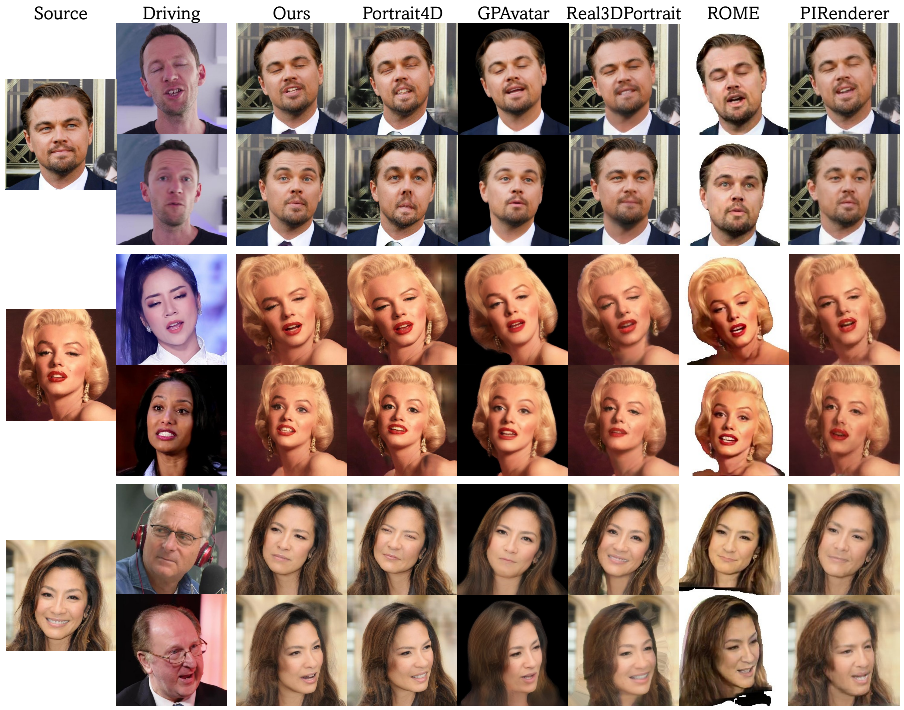


**Fig. 5:** Qualitative comparison using in-the-wild sources and VFHQ drivings.

### 5.2 Comparisons

***Baselines.*** We compare with existing one-shot video-based head reenactment methods, including 2D-based approaches: PIRenderer [36], Face-vid2vid [41], and StyleHEAT [46]; and 3D-aware ones: ROME [24], OTAvatar [32], HideNeRF [28], Real3DPortrait [45], GPAvatar [5], GOHA [29], and Portrait4D [9].

***Quantitative results.*** We first conduct self- and cross-ID reenactment on a test split of 100 video clips in VFHQ, each with around 200 frames. To measure image synthesis quality, we use LPIPS [51] and Fréchet Inception Distances (FID) [14] between the synthesized results and the ground truth. For identity similarity, we calculate cosine distance of the face recognition features [8] between the synthesis results and the source appearance. For expression, we use Average Expression Distance (AED) [30] measured by a 3DMM estimator [10]. We also compute Average Pose Distance (APD, $\times$1000 by default) [4] to measure the pose control accuracy. As shown in Tab. 1, our method largely outperforms previous methods in terms of reconstruction fidelity (*i.e.*, LPIPS and FID), identity similarity (*i.e.*, ID), and expression and pose control accuracy (*i.e.*, AED and APD).

Our cross-ID AED is slightly higher than a few approaches. We argue that this commonly used metric, reliant on a 3DMM reconstructor [10], may not


**Fig. 6:** Reconstructed head geometries using different 3D-aware methods.

adequately capture the similarity of expressions across identities with distinct shapes. To better evaluate cross-ID expression similarity, we report an $\mathrm{AED}_f$ metric based on the feature of an emotion classifier [7] in Tab. 2a. Moreover, to fairly evaluate the cross-ID lip motion accuracy, we utilize driving sequences from CelebV-Text [47] and measure the SyncNet [6] score between the reenacted results and the corresponding audios of the driving frames. As shown in Tab. 2a, our method achieves the best expression and lip motion accuracy, compared to other methods with good cross-ID performance (indicated by Tab. 1).

***Qualitative results.*** Figure 5 shows visual comparisons between different methods. Compared to the alternatives, our method excels at reconstructing the appearance details of the source images while capturing subtle facial movements of mouth, eyes, eyebrows, etc., in the driving frames. Besides, our method can well maintain the identity consistency under large pose variations.

Figure 6 visualizes the reconstructed head geometries (via Marching Cubes) of a given source image using different 3D-aware methods. Our method best preserves the overall head structure, which is the key for a good 3D consistency. Note that for HideNeRF [28], we failed to obtain a reasonable geometery.

***User study.*** For a more comprehensive evaluation, we ask 22 participants to select the best method in terms of image synthesis quality and expression similarity while inspecting 20 cross-ID reenactment results synthesized by different methods, respectively. As shown in Tab. 2b, our method demonstrates significant advantages compared to the others. Raw results are in the *suppl. material*.

### 5.3 Ablation Study

We conduct ablation studies to validate our proposed strategies. Table 3 reports the quantitative results following the experiment setting in Sec. 5.2. We introduce an extra metric (*i.e.*, Cons.) to evaluate identity consistency of the synthesized results given different drivings. Specifically, for each source, we use 50 drivings with different identities and poses for reenactment and compute the average cosine similarity between identity features [8] extracted from the synthesized results at frontal view. A higher Cons. indicates better disentanglement of motion with other information (*e.g.*, pose and ID) in the drivings. In Fig. 7, we visualize the reconstructed source geometries and *frontal-view* reenacted results under two different drivings. All configurations are trained to see 5M images by default.

**Table 3:** Ablation study of our method conducted on the test split of VFHQ.

| | Configuration | | | | | | Self reenactment | | | | Cross-ID reenactment | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Init. | $I_d(\boldsymbol{\theta}_p)$ | $I_d(\boldsymbol{\theta}_q)$ | Mot. | 3D Syn. | Data | LPIPS ↓ | ID ↑ | AED ↓ | APD ↓ | ID ↑ | AED ↓ | $\text{AED}_f$ ↓ | APD ↓ | Cons. ↑ |
| A | Rand. | | | $E_{mot}$ | $\Psi_{3d}$ | 10K | 0.216 | 0.874 | 0.021 | 0.752 | 0.776 | 0.167 | 0.213 | 1.887 | 0.822 |
| B | $\Psi_{3d}$ | | | $E_{mot}$ | $\Psi_{3d}$ | 10K | 0.213 | 0.879 | 0.018 | 0.755 | 0.718 | 0.161 | 0.185 | 0.796 | 0.813 |
| C | $\Psi_{3d}$ | ✓ | | $E_{mot}$ | $\Psi_{3d}$ | 10K | 0.223 | 0.873 | 0.020 | 1.251 | 0.757 | 0.165 | 0.190 | 1.436 | 0.913 |
| D | $\Psi_{3d}$ | | ✓ | $E_{mot}$ | $\Psi_{3d}$ | 10K | 0.219 | 0.875 | 0.019 | 0.739 | 0.712 | 0.156 | 0.185 | 0.688 | 0.805 |
| E | Rand. | ✓ | ✓ | $E_{mot}$ | $\Psi_{3d}$ | 10K | 0.232 | 0.865 | 0.023 | 0.967 | 0.777 | 0.172 | 0.207 | 1.257 | 0.929 |
| F | $\Psi_{3d}$ | ✓ | ✓ | 3DMM | $\Psi_{3d}$ | 10K | 0.221 | 0.873 | 0.018 | 1.074 | 0.704 | 0.139 | 0.202 | 0.993 | 0.811 |
| G | $\Psi_{3d}$ | ✓ | ✓ | $E_{mot}$ | [9] | 10K | 0.225 | 0.867 | 0.021 | 1.476 | 0.744 | 0.163 | 0.188 | 1.067 | 0.912 |
| H | $\Psi_{3d}$ | ✓ | ✓ | $E_{mot}$ | $\Psi_{3d}$ | 1K | 0.273 | 0.580 | 0.050 | 0.603 | 0.470 | 0.141 | 0.181 | 0.700 | 0.864 |
| I | $\Psi_{3d}$ | ✓ | ✓ | $E_{mot}$ | $\Psi_{3d}$ | 2K | 0.257 | 0.707 | 0.038 | 0.586 | 0.567 | 0.144 | 0.178 | 0.810 | 0.879 |
| J | $\Psi_{3d}$ | ✓ | ✓ | $E_{mot}$ | $\Psi_{3d}$ | 5K | 0.236 | 0.818 | 0.026 | 1.527 | 0.678 | 0.154 | 0.187 | 1.039 | 0.904 |
| K(Ours) | $\Psi_{3d}$ | ✓ | ✓ | $E_{mot}$ | $\Psi_{3d}$ | 10K | 0.224 | 0.873 | 0.019 | 0.758 | 0.743 | 0.162 | 0.186 | 0.943 | 0.912 |
| L(Ours*) | $\Psi_{3d}$ | ✓ | ✓ | $E_{mot}$ | $\Psi_{3d}$ | 10K | 0.224 | 0.873 | 0.019 | 0.580 | 0.736 | 0.161 | 0.185 | 0.858 | 0.908 |

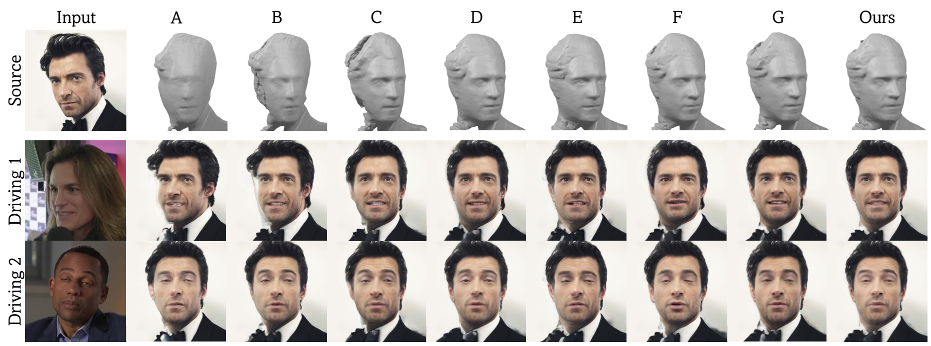


**Fig. 7:** Ablation study on different configurations. The first row visualizes the reconstructed geometries of the source. The bottom two rows depict cross-ID reenactment results rendered at a frontal view given two different drivings.

***Initialization.*** We first validate the efficacy of weight initialization using the pre-trained $\Psi_{3d}$ (*i.e.*, A *vs.* B and E *vs.* Ours). As shown in Tab. 3, using $\Psi_{3d}$ as initialization yields better self-reenactment performance as well as more accurate expression and pose control in the cross-ID setting. Moreover, Fig. 7 depicts that $\Psi_{3d}$ helps with more reasonable geometry learning.

***Pseudo multi-view data.*** We evaluate the influence of the pseudo multi-view videos by removing $I_d(\boldsymbol{\theta}_p)$ or $I_d(\boldsymbol{\theta}_q)$ during training (B *vs.* C *vs.* D *vs.* Ours). Leveraging $I_d(\boldsymbol{\theta}_p)$ for extracting the motion embedding largely improves the identity consistency (*i.e.* Cons.) with only a slight drop of the overall reenactment performance. This validates that $I_d(\boldsymbol{\theta}_p)$ can help eliminate the non-expression information (*e.g.*, pose) within the pre-trained 2D motion embedding [40], which is also in accordance with the higher ID metric in the cross-ID setting. Using $I_d(\boldsymbol{\theta}_q)$ as multi-view supervision facilitates geometry learning and reconstruction of unseen regions, as indicated by the lower APD metrics and shown in Fig. 7. Utilizing both of them yields our method with faithful 3D reconstruction ability and competitive reenactment performance. More analysis are in the *suppl.*

***Motion representation.*** We further replace our motion embedding from [40] with motion-related 3DMM parameters extracted following Sec. 4.2 (*i.e.*, F *vs.* Ours). As shown in Tab. 3, using 3DMM parameters for motion control is detrimental to identity preservation in the cross-ID setting, due to a known identity leakage issue of linear 3DMMs [9, 20]. While it has a lower AED metric (which is also a 3DMM-based metric without considering identity entanglement), it performs worse in terms of $\text{AED}_f$ and cannot faithfully capture detailed facial expressions as shown in Fig. 7. Notably, while leveraging the pre-trained 2D motion embedding can yield better cross-ID reenactment result, it is non-trivial to obtain reasonable 3D head geometries without our pseudo multi-view data for training (as indicated by config. B). See the *suppl.* for more visual analysis.

***3D Synthesizer.*** We also compare with a baseline which adopts portrait4D [9] for synthesizing multi-view training data (G *vs.* Ours). Our 3D head synthesizer $\Psi_{3d}$ facilitates expression control and geometry learning as indicated by the lower AED, $\text{AED}_f$, and APD in Tab. 3, respectively, due to its better structure preservation ability during novel-view synthesis (see Fig. 3a).

***Number of training data.*** We evaluate our method using different numbers of training videos (H *vs.* I *vs.* J *vs.* Ours). Table 3 shows that increasing the number of training data largely improves the reconstruction fidelity and identity preservation ability. We believe a even higher performance is reachable by further scaling up the training data. However, we observe that the models learned with fewer data have lower AED and APD in the cross-ID setting. We speculate that this may be due to a more severe information-leakage issue of the motion embedding when trained with limited data, which misleads the corresponding metrics (as indicated by the lower Cons. metric).

***Hyper parameters.*** We slightly tune the hyper parameters for a better performance (Ours *vs.* Ours*). By default, all parameters within the model share the same learning rate of $1e-4$. We scale up that of the motion-related layers by a factor of 2.5, and train the model to see 6M images, which empirically lowers the APD metric with minor influence to the remaining, yielding our final model.

## 6 Conclusion

In this paper, we presented a novel learning approach for synthesizing lifelike 4D head avatars from a single appearance image. The key intuition is to turn existing monocular videos into pseudo multi-view ones to enable learning an image-to-4D-head synthesizer in a data-driven manner instead of leveraging 3DMM-based inductive bias with limited expressiveness. Extensive experiments have demonstrated the superiority of our method over the prior arts in terms of reconstruction fidelity, geometry consistency, and reenactment accuracy. We hope our method will inspire future works towards exploring a better incorporation of 3D priors with 2D data for more realistic head avatar synthesis.

## References


1. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1999). pp. 187–194. ACM Press (1999)
2. Bolkart, T.: Bfm to flame. `https://github.com/TimoBolkart/BFM_to_FLAME` (2020)
3. Burkov, E., Pasechnik, I., Grigorev, A., Lempitsky, V.: Neural head reenactment with latent pose descriptors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13786–13795 (2020)
4. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
5. Chu, X., Li, Y., Zeng, A., Yang, T., Lin, L., Liu, Y., Harada, T.: Gpavatar: Generalizable and precise head avatar from image (s). arXiv preprint arXiv:2401.10215 (2024)
6. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. pp. 251–263. Springer (2017)
7. Daněček, R., Black, M.J., Bolkart, T.: Emoca: Emotion driven monocular face capture and animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20311–20322 (2022)
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
9. Deng, Y., Wang, D., Ren, X., Chen, X., Wang, B.: Learning one-shot 4d head avatar synthesis using synthetic data. arXiv preprint arXiv:2311.18729 (2023)
10. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: 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. pp. 0–0 (2019)
11. Drobyshev, N., Chelishev, J., Khakhulin, T., Ivakhnenko, A., Lempitsky, V., Zakharov, E.: Megaportraits: One-shot megapixel neural head avatars. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 2663–2671 (2022)
12. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics (ToG) **40**(4), 1–13 (2021)
13. He, T., Guo, J., Yu, R., Wang, Y., Zhu, J., An, K., Li, L., Tan, X., Wang, C., Hu, H., et al.: Gaia: Zero-shot talking avatar generation. arXiv preprint arXiv:2311.15230 (2023)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. pp. 6626–6637 (2017)
15. Hong, F.T., Zhang, L., Shen, L., Xu, D.: Depth-aware generative adversarial network for talking head video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3397–3406 (2022)

16. Hong, Y., Peng, B., Xiao, H., Liu, L., Zhang, J.: Headnerf: A real-time nerf-based parametric head model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20374–20384 (2022)
17. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023)
18. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
19. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134 (2017)
20. Jiang, Z.H., Wu, Q., Chen, K., Zhang, J.: Disentangled representation learning for 3d face shape. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11957–11966 (2019)
21. Kajiya, J.T., Von Herzen, B.P.: Ray tracing volume densities. ACM SIGGRAPH **18**(3), 165–174 (1984)
22. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
23. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
24. Khakhulin, T., Sklyarova, V., Lempitsky, V., Zakharov, E.: Realistic one-shot mesh-based head avatars. In: European Conference on Computer Vision. pp. 345–362. Springer (2022)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015)
26. Li, J., Tan, H., Zhang, K., Xu, Z., Luan, F., Xu, Y., Hong, Y., Sunkavalli, K., Shakhnarovich, G., Bi, S.: Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. arXiv preprint arXiv:2311.06214 (2023)
27. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017)
28. Li, W., Zhang, L., Wang, D., Zhao, B., Wang, Z., Chen, M., Zhang, B., Wang, Z., Bo, L., Li, X.: 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. pp. 17969–17978 (2023)
29. Li, X., De Mello, S., Liu, S., Nagano, K., Iqbal, U., Kautz, J.: Generalizable one-shot neural head avatar. arXiv preprint arXiv:2306.08768 (2023)
30. Lin, C.Z., Lindell, D.B., Chan, E.R., Wetzstein, G.: 3d gan inversion for controllable portrait image animation. arXiv preprint arXiv:2203.13441 (2022)
31. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. ACM SIGGRAPH **21**(4), 163–169 (1987)
32. Ma, Z., Zhu, X., Qi, G.J., Lei, Z., Zhang, L.: Otavatar: One-shot talking face avatar with controllable tri-plane rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16901–16910 (2023)
33. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision. pp. 405–421. Springer (2020)
34. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612 (2017)

35. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3D face model for pose and illumination invariant face recognition. In: IEEE International Conference on Advanced Video and Signal Based Surveillance. pp. 296–301 (2009)
36. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13759–13768 (2021)
37. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in Neural Information Processing Systems **32** (2019)
38. Tran, P., Zakharov, E., Ho, L.N., Tran, A.T., Hu, L., Li, H.: Voodoo 3d: Volumetric portrait disentanglement for one-shot 3d head reenactment. arXiv preprint arXiv:2312.04651 (2023)
39. Trevithick, A., Chan, M., Stengel, M., Chan, E., Liu, C., Yu, Z., Khamis, S., Chandraker, M., Ramamoorthi, R., Nagano, K.: Real-time radiance fields for single-image portrait view synthesis. ACM Transactions on Graphics (TOG) **42**(4), 1–15 (2023)
40. Wang, D., Deng, Y., Yin, Z., Shum, H.Y., Wang, B.: Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17979–17989 (2023)
41. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10039–10049 (2021)
42. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9168–9178 (2021)
43. Xie, L., Wang, X., Zhang, H., Dong, C., Shan, Y.: 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. pp. 657–666 (2022)
44. Xu, S., Yang, J., Chen, D., Wen, F., Deng, Y., Jia, Y., Tong, X.: Deep 3d portrait from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7710–7720 (2020)
45. Ye, Z., Zhong, T., Ren, Y., Yang, J., Li, W., Huang, J., Jiang, Z., He, J., Huang, R., Liu, J., et al.: Real3d-portrait: One-shot realistic 3d talking portrait synthesis. arXiv preprint arXiv:2401.08503 (2024)
46. Yin, F., Zhang, Y., Cun, X., Cao, M., Fan, Y., Wang, X., Bai, Q., Wu, B., Wang, J., Yang, Y.: Styleheat: One-shot high-resolution editable talking face generation via pretrained stylegan. arXiv preprint arXiv:2203.04036 (2022)
47. Yu, J., Zhu, H., Jiang, L., Loy, C.C., Cai, W., Wu, W.: Celebv-text: A large-scale facial text-video dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14805–14814 (2023)
48. Yu, W., Fan, Y., Zhang, Y., Wang, X., Yin, F., Bai, Y., Cao, Y.P., Shan, Y., Wu, Y., Sun, Z., et al.: Nofa: Nerf-based one-shot facial avatar reconstruction. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–12 (2023)
49. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9459–9468 (2019)
50. Zhang, B., Qi, C., Zhang, P., Zhang, B., Wu, H., Chen, D., Chen, Q., Wang, Y., Wen, F.: Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22096–22105 (2023)

51. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
52. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4176–4186 (2021)
53. Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., Loy, C.C.: Celebv-hq: A large-scale video facial attributes dataset. In: European conference on computer vision. pp. 650–667. Springer (2022)
54. Zhuang, Y., Zhu, H., Sun, X., Cao, X.: Mofanerf: Morphable facial neural radiance field. In: European Conference on Computer Vision. pp. 268–285. Springer (2022)
55. Zou, Z.X., Yu, Z., Guo, Y.C., Li, Y., Liang, D., Cao, Y.P., Zhang, S.H.: Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. arXiv preprint arXiv:2312.09147 (2023)