HeadGAP: Few-Shot 3D Head Avatar via Generalizable Gaussian Priors

Xiaozheng Zheng[1] Chao Wen[1†] Zhaohu Li[1] Weiyi Zhang[1] Zhuo Su[1] Xu Chang[1]
Yang Zhao[1] Zheng Lv[1] Xiaoyuan Zhang[1] Yongjie Zhang[1] Guidong Wang[1] Lan Xu[2]
[1]ByteDance [2]ShanghaiTech University

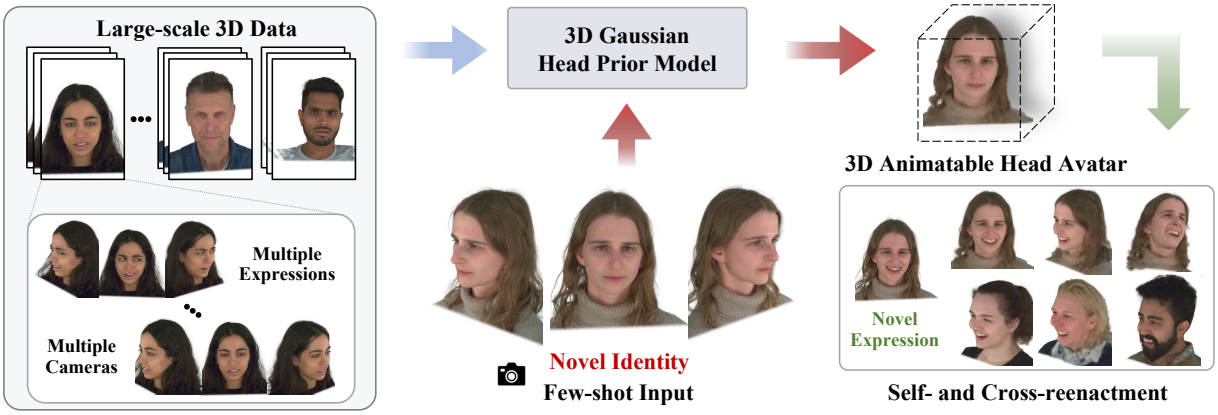

Figure 1. We present HeadGAP to create photo-realistic animatable 3D head avatars from only a few or even one image of the target person. **Firstly**, we utilize large-scale 3D data to learn 3D head prior with our designed 3D Gaussian head prior model. **Secondly**, we can use few-shot data to create 3D animatable avatars. **Finally**, we can animate the few-shot avatars with novel expressions.

Abstract

In this paper, we present a novel 3D head avatar creation approach capable of generalizing from few-shot in-the-wild data with high-fidelity and animatable robustness. Given the underconstrained nature of this problem, incorporating prior knowledge is essential. Therefore, we propose a framework comprising prior learning and avatar creation phases. The prior learning phase leverages 3D head priors derived from a large-scale multi-view dynamic dataset, and the avatar creation phase applies these priors for few-shot personalization. Our approach effectively captures these priors by utilizing a Gaussian Splatting-based auto-decoder network with part-based dynamic modeling. Our method employs identity-shared encoding with personalized latent codes for individual identities to learn the attributes of Gaussian primitives. During the avatar creation phase, we achieve fast head avatar personalization by leveraging inversion and fine-tuning strategies. Extensive experiments demonstrate that our model effectively exploits head priors and successfully generalizes them to few-shot personalization, achieving photo-realistic rendering quality, multi-view consistency, and stable animation.

Project page: https://headgap.github.io/

† Corresponding author

1. Introduction

Creating photo-realistic 3D avatars is a central challenge in computer graphics, encompassing applications such as movies, games, AR/VR, and the metaverse. There is a significant interest in generating digital avatars from real-world captures to create a precise digital copy of an actual person. These digital avatars can be animated and rendered from various viewpoints, maintaining high visual fidelity.

Recent advances [41, 46, 73, 85, 87] have achieved photo-realistic rendering quality of digital humans. In particular, 3D Gaussian Splatting (3DGS) [36] has been widely adopted for head avatars [13, 29, 47, 54, 70, 75] due to its efficient and realistic rendering capabilities. However, these advancements rely heavily on publicly available multi-view [40] or sequential datasets [24, 84], which are labor-intensive to capture and process for the average user. To address this limitation, many studies [12, 14, 19, 21, 81] aim to reduce the high data requirements for creating 3D avatars, allowing users to generate avatars from just a few images. Unfortunately, these methods often suffer from significant performance degradation compared to those [29, 54, 75] that utilize dense data for avatar creation. The challenge of generating few-shot personalized head avatars with high fidelity and stable animation remains unresolved.

To this end, we propose **HeadGAP** to facilitate high-

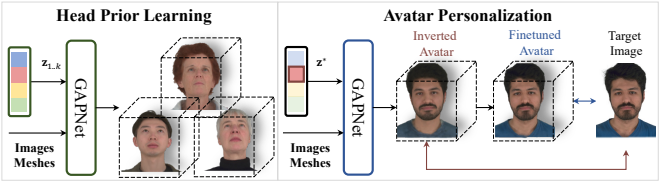

Figure 2. HeadGAP framework. The prior learning phase uses different IDs' data to embed head priors into the GAPNet. The personalization phase firstly optimizes identity codes to obtain the inverted avatar, then updates the GAPNet to get the fine-tuned avatar.

fidelity few-shot head avatar creation. As illustrated in Fig. 1, the core of HeadGAP lies in learning *generalizable* 3D Gaussian head priors from large-scale data and leveraging them to create *high-quality* personalized head avatars with few-shot input. As shown in Fig. 2, the HeadGAP framework consists of two phases: 1) the *prior learning* phase and 2) the *few-shot personalization* phase. In the *prior learning* phase, multi-view dynamic data is used to embed 3D prior knowledge into **GAPNet** (**GA**ussian **P**rior **Net**work). The *prior learning* phase is conducted only once. Subsequently, the *few-shot personalization* phase uses the learned priors to create avatars of new identities by inversion and fine-tuning. This framework focuses on two perspectives that would allow GAPNet to learn effective priors: 1) To achieve *high-quality*, we introduce a 3DGS-based head representation boosted with part-based and dynamic modeling. 2) For enhancing *generalizability*, we design GAPNet in an auto-decoder manner, which constructs continuous part-based identity spaces that can serve as powerful generative priors to guide the few-shot creation. Additionally, we leverage mesh tracking priors by predicting Gaussian attributes relative to the tracked mesh [54].

We conduct comprehensive experiments on the NeRSemble dataset [40] to substantiate our design choices and demonstrate our method's superiority over existing approaches. To illustrate the robustness and practical applicability of our method, we present numerous avatars of novel identities generated from both public datasets and images captured with consumer-grade devices.

In summary, our contributions can be listed as follows:

- We introduce a novel framework that exploits generalizable 3D Gaussian priors for fast 3D head avatar personalization using only a few input images. These avatars exhibit high fidelity and consistent animatable quality.
- We present proper designs that effectively utilize part-based dynamic Gaussian head priors and generalize them for high-quality few-shot head avatar personalization.
- We substantiate the efficacy and robustness of our framework through comprehensive experiments. Meanwhile, we showcase its potential in real scenarios by avatar creation using images captured by consumer-grade devices.

2. Related Work

3D Animatable Head Avatar. Since the advent of 3D neural implicit representations, remarkable progress has been made in creating animatable 3D head avatars from monocular or multi-view videos with various expressions and poses. Existing works have explored varieties of avatar representation. Some previous approaches [30, 38, 49] employ 3DMM [5, 43] with neural textures. Many recent studies [1, 2, 24, 25, 34, 41, 46, 73, 84, 87] focus on creating neural volumetric avatars. 3DMM [4, 5, 27, 43, 53] is often employed in those approaches. More recently, point-based representations are widely adopted [13, 28, 29, 54, 58, 70, 75, 85]. Among those works, 3DGS [36] is the most prevalent representation due to its efficient rendering and topological flexibility. Similar to these approaches, our approach is also based on 3DGS. However, there exist significant differences between previous works and ours, including our model being designed with 1) *part-based dynamic modeling* and designed for 2) *3DGS-based generative modeling* rather than single-subject modeling.

One-shot 2D Head Avatar. One-shot 2D head avatar synthesis has attracted lots of attention in recent years. Plenty of works leverage 2D generative models for talking head synthesis at high fidelity. One part of those works [22, 26, 33, 55, 59, 66, 67, 82] learn latent deformed features and feed them to 2D generators for face reenactment. Some other studies [6, 32, 78] map images to the latent space of a pre-trained StyleGAN2 [35]. While 2D-based methods can produce photorealistic images, they struggle to preserve the 3D consistency. Therefore, several methods [14, 19, 20, 34, 37, 44, 45, 48, 50, 77, 80] pursue animatable 3D head synthesis. They often resort to monocular 3DMM [16, 23] for providing geometry or pose guidance. Another line of work for 3D-aware portrait generations is also capable of few-shot avatar animations. Lots of studies [9–11, 17, 31, 52, 57, 61] demonstrate that the combination of 3D representations and adversarial learning on monocular images makes it possible to learn a 3D-aware generator for multi-view image generation. Many works [3, 62–64, 68, 69, 72] introduce 3DMM for animation control. These methods can be combined with advanced GAN inversion techniques [18, 42, 56, 71, 79] for head avatar reconstruction. However, those 2D-based approaches are still inferior in 3D consistency due to their representations and training schemes.

Few-shot Head Avatar with Data-driven 3D Priors. We focus on the few-shot 3D avatar personalization with data-driven 3D priors from large-scale data. To solve this problem, there are also some recent works [7, 8, 12, 76, 81] designed in this manner. Morphable Diffusion [12] introduces a multi-view consistent diffusion model to create head avatars from a single image. Preface [7] trains a NeRF-based auto-decoder generative model and achieves few-shot

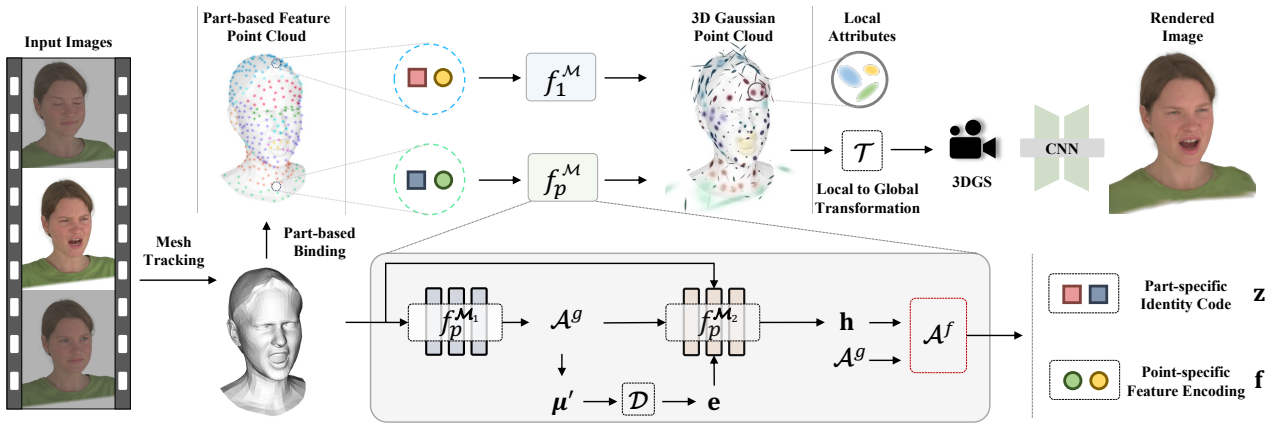

Figure 3. Illustration of the GAPNet. Given the tracked meshes of the input images, GAPNet binds part-based Gaussian primitives with initialized features to the mesh. Then, it employs part-specific modules to predict the local attributes of each primitive. The local attributes are transformed into global ones for 3DGS rendering. Finally, the renderings are fed into the CNN to obtain the final rendered images.

high-fidelity 3D static head creations. PhoneScan [8] extends MVP [46] to an auto-encoder generative model and supports novel avatar creation with the phone-captured data. VRMM [76] is an auto-decoder generative model built upon MVP [46], which supports few-shot relightable avatar creation. One2Avatar [81] adapts MonoAvatar [2] to an auto-decoder generative model based on a 3DMM-anchored neural radiance field [51]. Similar to [7, 76, 81], we also design our model in an auto-decoder manner. Different from these approaches relying on volume rendering [7, 8, 76, 81] or diffusion model [12], our approach employs 3DGS for rendering. We concentrate on designing a 3DGS-based generative model to achieve high-fidelity few-shot personalization with robust animations.

3. Method

In this section, we first introduce the preliminary (Sec. 3.1). Then, we detail our avatar representation designed for creating head avatars with learned generalizable head prior knowledge (Sec. 3.2). Finally, we present our HeadGAP framework (Fig. 2), including 1) head prior learning phase (Sec. 3.3) and 2) few-shot personalization phase (Sec. 3.4).

3.1. Preliminary

3D Gaussian Splatting (3DGS) [36] proposes a point-based scene representation, where each point represents a Gaussian primitive that is described by a global space position $\boldsymbol{\mu}$, rotation $\mathbf{r}$, scale $\mathbf{S}$, opacity α and color $\mathbf{c}$. In the following, we let the following notation:

$$\mathcal{A} = \{\boldsymbol{\mu}, \mathbf{r}, \mathbf{S}, \alpha, \mathbf{c}\}, \quad \mathbf{I} = \mathcal{R}(\mathcal{A}, \pi_{\mathbf{K},\mathbf{E}}) \tag{1}$$

denote the set of attributes $\mathcal{A}$ composing the Gaussian point cloud, and its tile-based differentiable rasterization $\mathcal{R}$ into an image $\mathbf{I}$ under the camera projection π described by intrinsic and extrinsic parameters $\mathbf{K}$ and $\mathbf{E}$ respectively.

GaussianAvatars [54] connects Gaussian primitives to the mesh faces. For each primitive, the position $\boldsymbol{\mu}'$, rotation $\mathbf{r}'$, and scaling $\mathbf{S}'$ are initialized in the local space. During rendering, these properties are converted into the global space:

$$\mathbf{r} = \mathbf{R}\mathbf{r}', \quad \boldsymbol{\mu} = s\mathbf{R}\boldsymbol{\mu}' + \mathbf{T}, \quad \mathbf{S} = s\mathbf{S}', \tag{2}$$

where $\mathbf{R}$ describes the orientation of the triangle face in the global space, s describes the scaling, and $\mathbf{T}$ describes the mean position of three vertices of a triangle face. For simplicity, we define $\mathcal{T}$ as the operation that transforms the local Gaussian attributes $\mathcal{A}'$ to global ones:

$$\mathcal{A} = \mathcal{T}(\mathcal{A}', \mathcal{M}), \quad \mathcal{M} = \{\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\delta}\}, \tag{3}$$

where $\mathcal{M}$ denotes the input FLAME parameters, consisting of shape parameters $\boldsymbol{\beta}$, pose parameters $\boldsymbol{\theta}$, expression parameters $\boldsymbol{\phi}$, and static vertex offsets $\boldsymbol{\delta}$.

3.2. Avatar Representation

Our avatar representation is based on an auto-decoder prior model [7, 76, 81, 83] that can learn head prior knowledge from multiple identities and be used for head avatar creation from few-shot images. As illustrated in Fig. 3, our representation builds upon a point-based representation with *part-based modeling*, where each point is only responsible for one semantic part. Firstly, we initialize the *part-based feature point cloud* consisting of 1) *part-based identity code* and 2) *point-specific feature encoding*, based on the tracked mesh. Then, we conduct *dynamic modeling* by feeding the feature point cloud to the *part-based multi-layer perceptions* (MLPs) to regress the Gaussian attributes of all the points for 3DGS rendering. Finally, we utilize a *convolutional neural network* (CNN) module to refine the 3DGS renderings to obtain the final rendered image. In the following, we will describe those key components.

Part-based Feature Point Cloud. For initializing the *part-based feature point cloud* based on the tracked mesh $\mathcal{M}$, we first utilize UV-based initialization [70] to obtain the point cloud with n Gaussian primitives, with each pixel in the UV

map bound to one triangle of the mesh. The initialization contributes to more uniform primitives distributed on the head region than face-based initialization [54].

Then, we set up the initial features for the point cloud. The features contains two types, including the 1) *point-specific feature encodings* $\mathbf{f} = \{\mathbf{f}_i \in \mathbb{R}^{c_1}\}_{i=1}^{n}$ and 2) *part-specific identity codes* $\mathbf{z} = \{\{\mathbf{z}_j^l \in \mathbb{R}^{c_2}\}_{l=1}^{p}\}_{j=1}^{k}$, where p and k denote the part and identity number respectively. The point encodings $\mathbf{f}$ embeds identity-shared priors and the identity codes $\mathbf{z}$ serve as the identity codebook for the auto-decoder model. All the encodings are randomly initialized learnable parameters. The part of a primitive is determined by its parent triangle and the identity codes are the same for all the primitives belonging to the same part.

Part-based Dynamic Gaussian Attributes Modeling. For simplicity, we use $\mathbf{f}$ and $\mathbf{z}$ to denote the per-point features belonging to a specific part p and also omit the part notation for other notations, unless otherwise stated. Given $\mathbf{f}$ and $\mathbf{z}$, we regress dynamic local Gaussian attributes by:

$$\mathcal{A}^g = f_p^{\mathcal{M}_1}(\mathbf{f}, \mathbf{z}), \quad \mathbf{h} = f_p^{\mathcal{M}_2}(\mathbf{f}, \mathbf{z}, \mathbf{e}, \mathcal{A}^g), \tag{4}$$

where $\mathbf{h}$ is point appearance attribute, $\mathcal{A}^g = \{\boldsymbol{\mu}', \mathbf{r}', \mathbf{S}', \alpha\}$ denotes other attributes, and $\mathbf{e} := \mathcal{D}(\boldsymbol{\mu}') = \mathcal{T}(\boldsymbol{\mu}') - \mathcal{T}(\boldsymbol{\mu}'_{neutral})$ is the point-specific dynamic signals obtained by subtracting the global neutral point position $\mathcal{T}(\boldsymbol{\mu}'_{neutral})$ from the global posed point position $\mathcal{T}(\boldsymbol{\mu}')$. Both $f_p^{\mathcal{M}_1}$ and $f_p^{\mathcal{M}_2}$ are part-specific MLPs. We define the overall dynamic modeling as: $\mathcal{A}^f = f_p^{\mathcal{M}}(\mathbf{f}, \mathbf{z})$, where $\mathcal{A}^f = \mathcal{A}^g \cup \{\mathbf{h}\}$ denotes the final Gaussian attributes used for splatting.

Part-based and dynamic modeling contribute to better few-shot performance, as shown in Sec. 4.5. The part-based modeling allows the specialized module to learn the particular part's priors, resulting in easier optimization and more powerful priors. The dynamic modeling employs point-specific expression signals $\mathbf{e}$ for predicting dynamic local attributes, which is better at capturing dynamic details than GaussianAvatars [54] using static local attributes.

Gaussian Splatting with CNN refinement. Inspired by recent works [74, 75], we apply a screen-space CNN $f^{\mathcal{C}}$ to refine the rendered results:

$$[\mathbf{I}_{rgb}, \mathbf{I}_h] = \mathcal{R}(\mathcal{T}(\mathcal{A}_f, \mathcal{M}), \pi_{\mathbf{K},\mathbf{E}}), \tag{5}$$

$$\mathbf{I} = f^{\mathcal{C}}([\mathbf{I}_{rgb}, \mathbf{I}_h]), \tag{6}$$

where $\mathcal{A}^f = \{\mathcal{A}_i^f\}_{i=1}^{n}$ denotes the final Gaussian attributes for all the points, $\mathbf{I}_{rgb}$ denotes rendered RGB images, and $\mathbf{I}_h$ is a latent feature image used for the CNN refinement.

Different from previous methods [74, 75], we do not conduct super-resolution, but keep the input and output with the same resolution for refinement. We aim to use large-scale data to enable CNN to capture generalizable structured appearance priors that are challenging to exploit by our 3DGS-based representation. As indicated in Sec. 4.5, using CNN-based refinement can indeed capture those priors to render more photo-realistic results for few-shot personalization.

Overall Representation. The overall head avatar representation $\mathcal{H}$ is defined formally:

$$\mathcal{H} : (\mathcal{M}; f^{\mathcal{M}}, f^{\mathcal{C}}, \mathbf{f}, \mathbf{z}) \mapsto \mathbf{I}\ , \tag{7}$$

where $f^{\mathcal{M}} = \{f_l^{\mathcal{M}}\}_{l=1}^{p}$ denotes the MLPs for all parts.

3.3. Head Prior Learning

We highly rely on the head's prior knowledge to achieve high-fidelity avatar creation for the unconstrained problem with only a few input images. Among various priors, we aim at learning high-quality, animatable, and 3D-consistent head priors from available multi-view dynamic head data with multiple identities. Therefore, the goal for the prior learning stage is to learn k head avatars within the GAPNet, with the respective identity codes $\mathbf{z}_{1\ldots k}$ and other network parameters optimized. Before starting model training, we first conduct FLAME tracking for the training data to obtain $\mathcal{M}$. Then, we use those data to jointly optimize $\mathcal{M}$, $f^{\mathcal{M}}$, $f^{\mathcal{C}}$, $\mathbf{f}$, and $\mathbf{z}$ with our total loss term:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{rec}(\mathbf{I}, \mathbf{I}^*) + \mathcal{L}_{rec}(\mathbf{I}_{rgb}, \mathbf{I}^*) + \\ & \lambda_m \mathcal{L}_{rec}(\mathbf{I}_m, \mathbf{I}_m^*) + \mathcal{L}_{reg},\end{aligned} \tag{8}$$

where $\mathcal{L}_{rec}$ and $\mathcal{L}_{reg}$ denote the image reconstruction loss and training regularization loss respectively. The ground truth image is denoted as $\mathbf{I}^*$. To improve the fidelity of the mouth region, we further supervise mouth region $\mathbf{I}_m$ with masked ground truth mouth region $\mathbf{I}_m^*$, inspired by FlashAvatar [70]. Specifically, the image reconstruction loss:

$$\mathcal{L}_{rec} = \lambda_{l1}\mathcal{L}_{l1} + \lambda_{ssim}\mathcal{L}_{ssim} + \lambda_{lpips}\mathcal{L}_{lpips} \tag{9}$$

consists of L1 loss $\mathcal{L}_{l1}$, SSIM loss $\mathcal{L}_{ssim}$, and perceptual loss $\mathcal{L}_{lpips}$ with the VGG as the backbone. Meanwhile, the training regularization loss:

$$\mathcal{L}_{reg} = \lambda_\alpha \mathcal{L}_\alpha + \lambda_s \mathcal{L}_s + \lambda_\mu \mathcal{L}_\mu + \lambda_{arap}\mathcal{L}_{arap}, \tag{10}$$

includes opacity regularization $\mathcal{L}_\alpha = \|\mathbf{I}_\alpha - \tilde{\mathbf{I}}_{mask}\|_1$, primitive local scaling regularization $\mathcal{L}_s = \|\max(\mathbf{s}, \epsilon_s)\|_2$, primitive local position regularization $\mathcal{L}_\mu = \|\max(\boldsymbol{\mu}, \epsilon_\mu)\|_2$, and ARAP (As-Rigid-As-Possible) regularization $\mathcal{L}_{arap}$ [60]. The opacity regularization $\mathcal{L}_\alpha$ is used to constrain the Gaussian primitives to stay within the head region and their opacity accumulated to 1, which is computed between accumulated opacity image $\mathbf{I}_\alpha$ and the head mask $\tilde{\mathbf{I}}_{mask}$. We utilize the same thresholds $\epsilon_s = 0.6$ and $\epsilon_\mu = 1$ for $\mathcal{L}_s$ and $\mathcal{L}_\mu$ respectively as [54] to constraint the local scaling and position of the Gaussian primitives. The ARAP regularization $\mathcal{L}_{arap}$ is employed for regularizing the optimization of static offset. All the λs described above are used for balancing different loss terms.

3.4. Few-shot Personalization

After the prior learning phase, we encode dynamic head prior knowledge within GAPNet. Consequently, all the

Method	Reference	Input	Frontal view			All views			
			LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	ID↑
ROME [37]	ECCV'22	1 image	0.237	17.56	0.813	0.286	15.45	0.796	0.658±0.119
GOHA [45]	NeurIPS'23	1 image	0.224	15.85	0.760	0.308	13.13	0.736	0.588±0.146
VOODOO3D [65]	CVPR'24	1 image	0.294	17.07	0.773	0.310	14.59	0.752	0.626±0.111
HiDe-NeRF [44]	CVPR'23	1 image	0.290	15.63	0.789	0.368	14.51	0.784	0.640±0.156
Portrait4Dv1 [19]	CVPR'24	1 image	0.180	16.59	0.796	0.273	14.52	0.752	0.674±0.143
Portrait4Dv2 [20]	ECCV'24	1 image	0.155	17.77	0.810	0.269	14.53	0.757	0.694±0.141
GPAvatar [14]	ICLR'24	1 image	0.180	18.42	0.827	0.294	13.83	0.775	0.631±0.169
Ours-SV		1 image	0.142	17.91	0.829	0.217	14.75	0.792	0.768±0.113
DiffusionRig [21]	CVPR'23	20 images	0.220	16.94	0.811	0.298	14.88	0.786	0.817±0.112
NHA† [30]	CVPR'22	mono video	0.161	17.42	0.850	0.266	14.77	0.807	0.577±0.138
FlashAvatar† [70]	CVPR'24	mono video	0.146	18.89	0.854	0.286	16.56	0.791	–
GaussianAvatars‡ [54]	CVPR'24	3 images	0.320	16.19	0.723	0.337	15.80	0.705	–
GaussianAvatars♦ [54]	CVPR'24	3 videos	0.147	21.05	0.852	0.301	16.62	0.728	–
Ours		3 images	0.138	21.67	0.866	0.144	20.90	0.868	0.821±0.094

Table 1. Quantitative comparisons with state-of-the-art methods on NeRSemble [40] dataset. We use colors to denote the first , second and third places respectively. The results are averaged across novel-view and novel-pose.

learned parameters of GAPNet can serve as powerful priors to aid few-shot or even one-shot personalization.

Prior to personalization, we employ a tracker to acquire the FLAME [43] parameters of the input image. Given input images with FLAME trackings, we first find the most similar avatar from the identity codebook through inversion. Specifically, we optimize part-specific linear combination weights $\mathbf{w} \in \mathbb{R}^{k\times p\times 1}$ to obtain the identity code $\mathbf{z}^* = \text{softmax}(\mathbf{w}) \odot \mathbf{z} \in \mathbb{R}^{k\times p\times c_2}$ used for rendering an avatar similar to the input. During the inversion optimization, we keep all the parameters of the network frozen except for $\mathbf{w}$. Formally, given an input image $\mathbf{I}^*$ of the target identity, we optimize to render an image $\mathbf{I}$ that resembles the target identity. This procedure is optimized with the loss function in Eq. (8) with respect to $\mathbf{w}$.

Then, we start fine-tuning to update the network's parameters so that the avatar can capture the details of the target identity from the inputs. We leverage prior knowledge in this procedure through three strategies. First, we use small learning rates for all parameters except $\mathbf{f}$. Next, we exploit extracted part-based priors by excluding the fine-tuning for the mouth region, as modeling the highly flexible mouth region with few inputs is challenging. Finally, we apply view regularization to prevent overfitting to the target view, inspired by previous methods [56, 83]. Specifically, we constraint the fine-tuning results of some reference views with neutral face $\{\mathbf{R}_i\}_{i=1}^m$ to be close to the rendering results before fine-tuning $\{\tilde{\mathbf{R}}_i\}_{i=1}^m$, where m is the number of the generated reference views. With the prior knowledge, our personalized avatar achieves stable reenactment while preserving the details of the target identity. The fine-tuning is conducted by minimizing the loss function in Eq. (8):

$$\arg\min_{\xi} \mathcal{L}_f = \mathcal{L}(\mathbf{I}, \mathbf{I}^*) + \lambda_{ref} \sum_{i=1}^{m} (\mathcal{L}(\mathbf{R}_i, \tilde{\mathbf{R}}_i)), \quad (11)$$

where ξ denotes all the learnable parameters, and λ_{ref} is used to balance different loss terms.

4. Experiments

4.1. Setup

Dataset. We utilize facial images of 164 subjects with 16 camera viewpoints in the NeRSemble [40] dataset for experiments. We separated the data into training and testing sets, comprising 119 and 45 subjects respectively. The training sets are used for prior learning, while the testing sets are employed to quantitatively and qualitatively evaluate few-shot personalization performance. We also constructed an in-house dataset as part of the testing data. In addition to the leave-out testing data, we conduct experiments on data captured by consumer-grade devices to evaluate the performance towards in-the-wild inputs.

FLAME Tracking. Inspired by previous studies [54, 86], we designed a tracking algorithm that can optimize the FLAME [43] parameters with different numbers of input views. In the following experimental section, unless otherwise stated, the training data uses this tracking algorithm to obtain the ground truth 3DMM parameters. For the test data, we simplify the tracking to use sparse view inputs and provide the corresponding 3DMM parameters for testing.

4.2. Implementation Details

Model Detail. We divide the primitives into $p = 11$ parts according to the face masks from FLAME [43]. We utilize $k = 119$ identities for prior learning. All the MLPs $f^{\mathcal{M}}$ consist of 4 layers and the CNN $f^{\mathcal{C}}$ contains 6 layers.

Training Detail. We adopt Adam [39] optimizer for the training. For prior learning, we set the batch size to 32. All parameters start with a learning rate of $1e^{-3}$, which decreases using a cosine scheduler. The prior model is trained

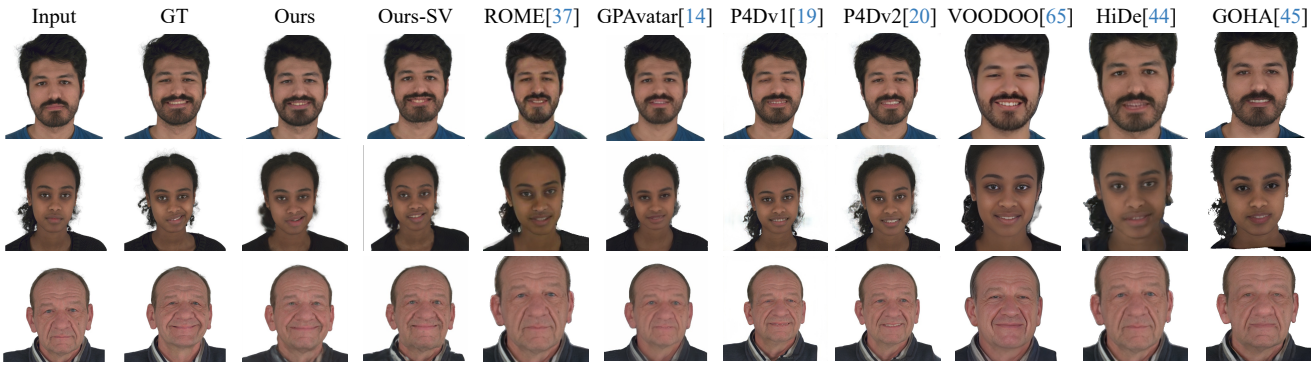

Figure 4. Qualitative comparisons of our approach against state-of-the-art methods using a single image as input.

on 8 A100 GPUs for 100K steps, taking around 2 days. For few-shot personalization, we use a batch size of 1. Both inversion and fine-tuning take 500 steps, totaling about 5 minutes on an A100 GPU. Please refer to supplementary materials for more details.

4.3. Baselines and Metrics

Baselines. We classify baselines into two types based on their training approaches. The Type-I, like ours, uses multi-ID datasets to train prior features of the head, which can generalize to novel IDs, termed *prior-based methods*. The Type-II requires individual training for each person, termed *per-subject optimization methods*. For Type-I, due to the lack of available code for most head avatar generation methods using a few views [8, 76, 81], we compare our method with approaches using a single image, including mesh-based methods [37] and the state-of-the-art tri-plane based methods [14, 19, 20, 44, 45, 65]. We also compare with 3D-aware diffusion models [21] equipped with multi-view inputs. For Type-II, we compare our approach with methods using Gaussian Splatting [54, 70] or explicit mesh [30] as 3D representations. Per-subject methods require more training data, so we provide these baselines with monocular video†, multi-view images‡, or multi-view videos$^{\blacklozenge}$.

Metrics. We employ standard image quality metrics for our quantitative evaluations: 1) Peak Signal-to-Noise Ratio (PSNR), 2) Structure Similarity Index (SSIM), and 3) Learned Perceptual Image Patch Similarity (LPIPS), following previous works [24, 54, 81]. Furthermore, we also report 4) ID that measures the identity similarity [15] between the predictions and ground truth ones.

4.4. Fast Avatar Personalization

We compare avatar creation with the state-of-the-art on 3 subjects ("074", "175", and "210') of NeRSemble [40]. To conduct thorough comparisons, we categorize the baselines into methods utilizing: 1) a single image and 2) multiple images. We analyze the performance of various approaches for both frontal view and all views. We also present our results using 1-shot and 3-shot inputs. The quantitative comparisons are listed in Tab. 1. We also illustrate qualitative comparisons in Fig. 4 and Fig. 5.

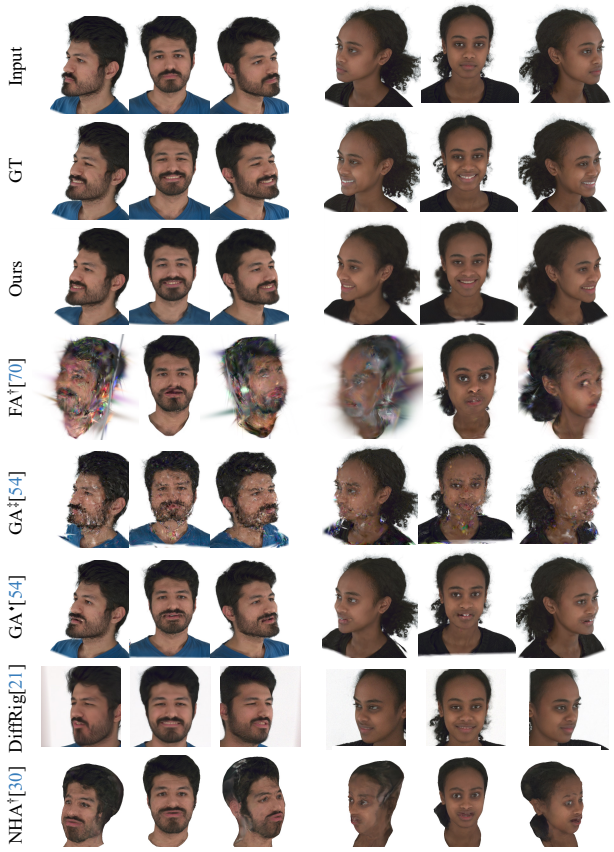

Figure 5. Qualitative comparisons of our approach against state-of-the-art methods using few-shot input.

One-shot personalization. For fair comparisons, our 1-shot results use personalized tracking data from a monocular tracker MICA [86], referred to as "Ours-SV." Since we rely on the neck pose from FLAME [43], and MICA lacks this, our performance degrades a lot. Despite this, our approach achieves the best or the second best results across all metrics, demonstrating its robustness.

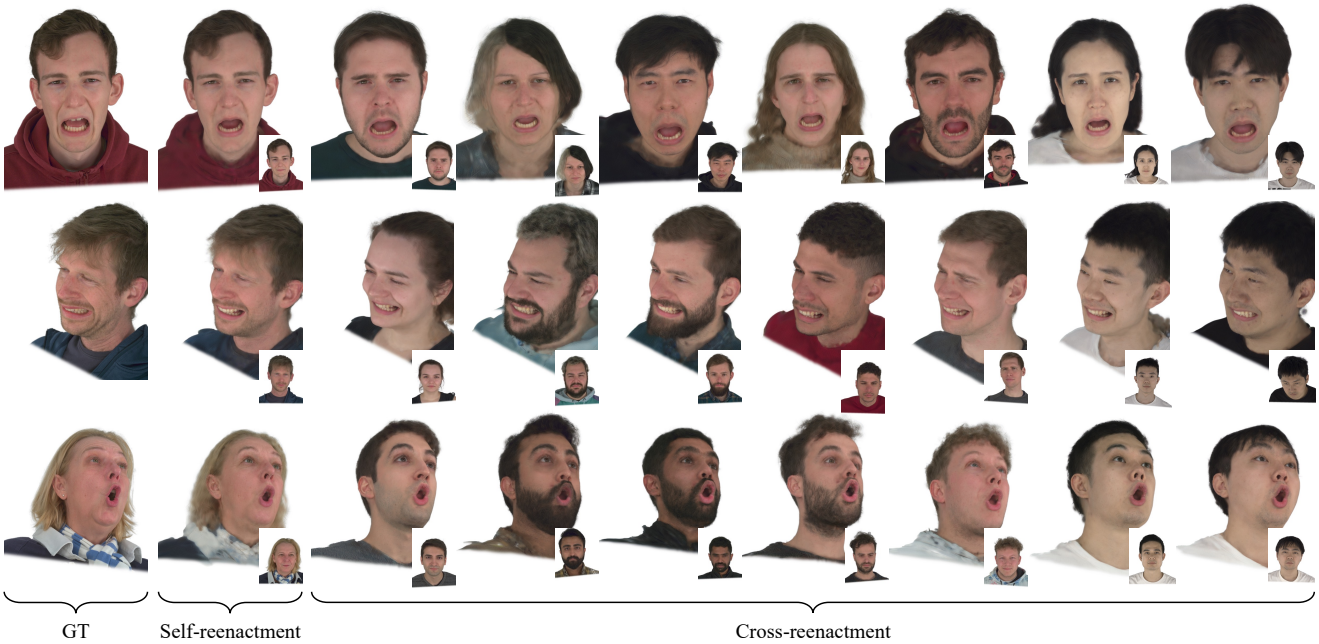

Figure 6. Our 3-shot (3-view of the neutral face) results on NeRSemble and our in-house data. From left to right, we show the ground truth, self-reenactment, and cross-reenactment. The lower right of the reenactment results presents one of the three inputs.

Figure 7. Our 3-shot results on in-the-wild data captured by iPad. An example of the input is shown on the left. The lower right of the reenactment results presents one of the three inputs.

Few-shot personalization. Increasing the number of images significantly enhances the authenticity of the head avatar, so we focus on few-shot personalization. We compare our 3-shot results, using a more precise input mesh from our tracker, with other baseline methods. Under this setting, our approach achieves the best results across all metrics (Please note that for [54, 70], the presence of excessive artifacts makes it impossible to evaluate the ID metrics). To better show our robustness, we illustrate more of our few-shot results in Fig. 6. To show the applicability of our approach in real-world scenarios, we also present avatars created from 3 images captured by an iPad in Fig. 7. We use these 3 RGB-D images for FLAME fitting. These results further demonstrate the generalization capability of our approach.

4.5. Model Analysis

We concentrate on few-shot personalization, thus, the experiments primarily examine how our designs affect final few-shot performance.

Ablation. We use subjects "256" and "270" for conducting the ablation study. The ablation is divided into two parts, with the first part to validate our model designs and the second part to justify our few-shot strategies. Tab. 2 presents the quantitative results for the testing sequence. Moreover, we illustrate the qualitative results in Fig. 8.

For model design ablations, we remove one component at a time to demonstrate their effectiveness. Excluding part-based modeling ("w/o Part") causes a significant performance drop, as it hampers effective prior learning and part-based few-shot strategies. Not modeling dynamic information ("w/o Dynamic") also degrades performance, making it challenging to capture details in highly dynamic regions (*e.g.*, the mouth). Lastly, using CNN for refinement reduces artifacts and enhances detail realism ("w/o CNN").

3DGS-based methods have strong input-fitting capabilities, leading to overfitting with limited training data and poor generalization to novel views and expressions ("Base"). Thus, 3DGS-based methods heavily rely on prior knowledge for few-shot personalization. Using our prior model for inversion ("+ Inversion") produces avatars simi-

Method	LPIPS↓	PSNR↑	SSIM↑
Prior model design:			
Full-model	**0.140**	**22.87**	**0.854**
w/o Part	0.154	22.21	0.850
w/o Dynamic	0.148	22.56	0.853
w/o CNN	0.156	22.01	0.849
Few-shot strategy:			
Base (w/o Prior)	0.237	19.25	0.811
+ Inversion	0.171	19.67	0.829
+ Finetune	0.143	22.08	0.842
+ View Reg.	**0.140**	**22.87**	**0.854**

Table 2. Quantitative ablation study. The highlights denote the full-model.

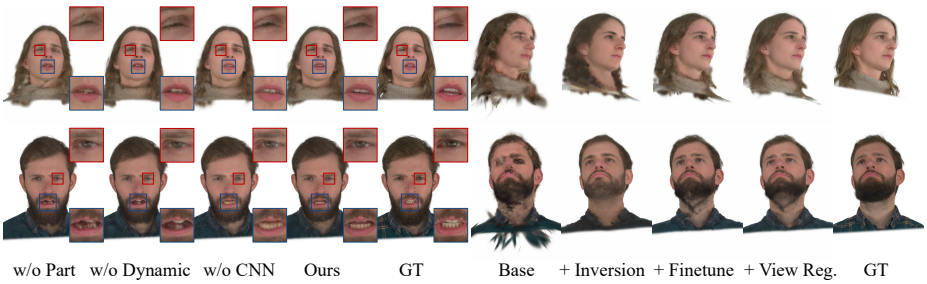

Figure 8. Qualitative ablation study. Please zoom in for more details.

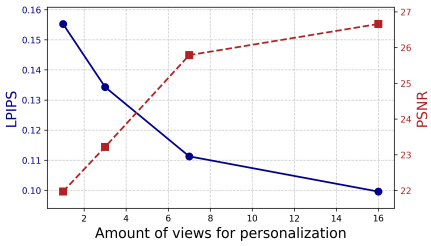

Fig.9a Quality w.r.t #personalizing data.

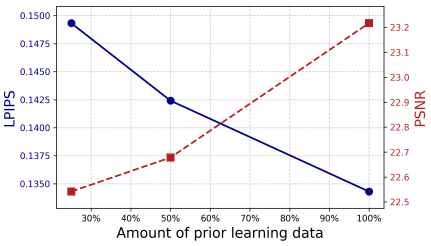

Fig.9b Quality w.r.t #prior learning data.

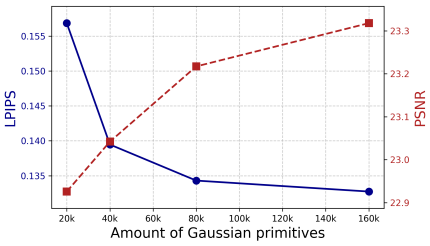

Fig.9c Quality w.r.t #primitives.

lar to the input data with robust animation. Additional finetuning ("+ Finetune") enhances realism with personalized details. View regularization ("+ View Reg.") helps reduce novel-view artifacts.

Quality w.r.t quantity of personalizing data. HeadGAP supports various numbers of inputs, allowing us to analyze few-shot performance with different data amounts. The analysis has two parts: 1. Quantitative results with different numbers of views of the neutral face (Fig. 9a) show that performance improves as the number of views increases. 2. Performance with 8 additional images of different frontal view expressions (Fig. 10) demonstrates that more inputs help the model capture personalized dynamic details.

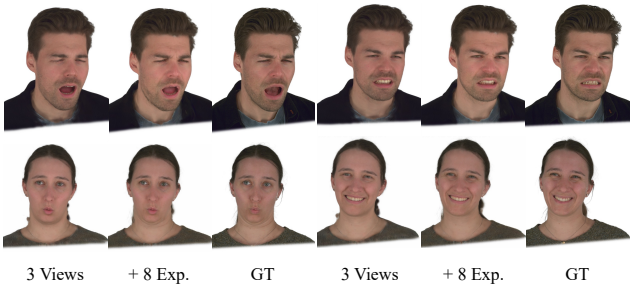

Figure 10. Comparisons between using 3-view data and 3-view with 8 additional expressions' data of the frontal view.

Quality w.r.t quantity of prior learning data. The large-scale prior learning data is the core of the high-fidelity few-shot personalization. Therefore, we conduct analysis to show our performance with varying amounts (IDs) of data for prior learning. As depicted in Fig. 9b, more prior learning data makes the final avatar more realistic. Moreover, we do not observe significant saturation, indicating our model can benefit from more available 3D data. It is worth noting that our model is robust for different amounts of prior learning data due to our complete pipeline.

Quality w.r.t quantity of primitives. We also evaluate our performance with varying numbers of Gaussian primitives. Fig. 9c proves our method can create more realistic avatars with more primitives. To balance the fidelity and efficiency, we utilize 80k primitives for our default model. We do not focus on finding the optimal method to control the number of primitives. We believe our approach could also benefit from other adaptive density control approaches [36].

Network comparison. GAPNet is capable of adapting to different numbers of IDs for training. To demonstrate the network capability, we compare its performance for a single person against [54]. GAPNet achieves better performance in single-person modeling, with an LPIPS/PSNR of 0.091/25.48 compared to GaussianAvatars' 0.119/25.32.

5. Conclusion

In this paper, we present a novel approach for creating high-fidelity 3D head avatars with few-shot images. We first learn 3D Gaussian priors from large-scale 3D head data, then create avatars of the novel identities with the aid of the priors. To facilitate the learning of powerful and generalizable priors, we develop GAPNet which can exploit 3D part-based dynamic head priors and 2D structured head priors for creating high-fidelity avatars with robust animations. The comprehensive experiments justify our designs and superiority. We also showcase our robustness by creating avatars of plentiful identities from the public dataset and images captured by consumer-grade devices.

References

[1] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 20364–20373, 2022. 2

[2] Ziqian Bai, Feitong Tan, Zeng Huang, Kripasindhu Sarkar, Danhang Tang, Di Qiu, Abhimitra Meka, Ruofei Du, Mingsong Dou, Sergio Orts-Escolano, et al. Learning personalized high quality volumetric head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16890–16900, 2023. 2, 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

[4] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1063–1074, 2003. 2

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023. 2

[6] 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

[7] Marcel C Bühler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, et al. Preface: A data-driven volumetric prior for few-shot ultra high-resolution face synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3402–3413, 2023. 2, 3

[8] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabriel Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 2, 3, 6

[9] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 2

[10] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022.

[11] Xingyu Chen, Yu Deng, and Baoyuan Wang. Mimic3d: Thriving 3d-aware gans via 3d-to-2d imitation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2338–2348. IEEE Computer Society, 2023. 2

[12] Xiyi Chen, Marko Mihajlovic, Shaofei Wang, Sergey Prokudin, and Siyu Tang. Morphable diffusion: 3d-consistent diffusion for single-image avatar creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10359–10370, 2024. 1, 2, 3

[13] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. *arXiv preprint arXiv:2312.04558*, 2023. 1, 2

[14] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. GPAvatar: Generalizable and precise head avatar from image(s). In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 5, 6

[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6

[16] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2

[17] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10673–10683, 2022. 2

[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. 2

[19] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7119–7130, 2024. 1, 2, 5, 6

[20] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. *arXiv preprint arXiv:2403.13570*, 2024. 2, 5, 6

[21] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12736–12746, 2023. 1, 5, 6

[22] 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. 2

[23] 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

[24] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 1, 2, 6

[25] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. *ACM Transactions on Graphics (TOG)*, 41(6):1–12, 2022. 2

[26] 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

[27] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21003–21012, 2023. 2

[28] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Mononphm: Dynamic head reconstruction from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10747–10758, 2024. 2

[29] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Npga: Neural parametric gaussian avatars. *arXiv preprint arXiv:2405.19331*, 2024. 1, 2

[30] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 2, 5, 6

[31] 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. 2

[32] 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

[33] 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

[34] 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

[35] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[36] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):1–14, 2023. 1, 2, 3, 8

[37] 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. 2, 5, 6

[38] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM transactions on graphics (TOG)*, 37(4): 1–14, 2018. 2

[39] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[40] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 1, 2, 5, 6

[41] Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5481–5492, 2024. 1, 2

[42] 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. 2

[43] 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, 5, 6

[44] 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. 2, 5, 6

[45] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot 3d neural head avatar. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 5, 6

[46] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 1, 2, 3

[47] Haimin Luo, Min Ouyang, Zijun Zhao, Suyi Jiang, Longwen Zhang, Qixuan Zhang, Wei Yang, Lan Xu, and Jingyi Yu. Gaussianhair: Hair modeling and rendering with light-aware gaussians. *arXiv preprint arXiv:2402.10483*, 2024. 1

[48] Haoyu Ma, Tong Zhang, Shanlin Sun, Xiangyi Yan, Kun Han, and Xiaohui Xie. Cvthead: One-shot controllable head avatar with vertex-feature transformer. In *Proceedings of the*

IEEE/CVF Winter Conference on Applications of Computer Vision, pages 6131–6141, 2024. 2

[49] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 64–73, 2021. 2

[50] 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. 2

[51] 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. *Communications of the ACM*, 65(1):99–106, 2021. 3

[52] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 2

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

[54] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[55] 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

[56] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1):1–13, 2022. 2, 5

[57] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 2

[58] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1606–1616, 2024. 2

[59] 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. 2

[60] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, pages 109–116. Citeseer, 2007. 4

[61] 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. *ACM Transactions on Graphics (ToG)*, 41(6):1–10, 2022. 2

[62] 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

[63] 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.

[64] Junshu Tang, Bo Zhang, Binxin Yang, Ting Zhang, Dong Chen, Lizhuang Ma, and Fang Wen. 3dfaceshop: Explicitly controllable 3d-aware portrait generation. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 2

[65] Phong Tran, Egor Zakharov, Long-Nhat Ho, Anh Tuan Tran, Liwen Hu, and Hao Li. Voodoo 3d: Volumetric portrait disentanglement for one-shot 3d head reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10336–10348, 2024. 5, 6

[66] 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. 2

[67] 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

[68] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. *Advances in Neural Information Processing Systems*, 35:36188–36201, 2022. 2

[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. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–9, 2023. 2

[70] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1802–1812, 2024. 1, 2, 3, 4, 5, 6, 7

[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. 2

[72] 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

[73] Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. Avatarmav: Fast 3d head avatar reconstruction using motion-aware neural voxels. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–10, 2023. 1, 2

[74] Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Han Huang, Guojun Qi, and Yebin Liu. Latentavatar: Learning latent expression code for expressive neural head avatar. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–10, 2023. 4

[75] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2024. 1, 2, 4

[76] Haotian Yang, Mingwu Zheng, Chongyang Ma, Yu-Kun Lai, Pengfei Wan, and Haibin Huang. Vrmm: A volumetric relightable morphable head model. *arXiv preprint arXiv:2402.04101*, 2024. 2, 3, 6

[77] Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiawei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, et al. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. *arXiv preprint arXiv:2401.08503*, 2024. 2

[78] 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 pre-trained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022. 2

[79] 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. 2

[80] 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. 2

[81] Zhixuan Yu, Ziqian Bai, Abhimitra Meka, Feitong Tan, Qiangeng Xu, Rohit Pandey, Sean Fanello, Hyun Soo Park, and Yinda Zhang. One2avatar: Generative implicit head avatar for few-shot user adaptation. *arXiv preprint arXiv:2402.11909*, 2024. 1, 2, 3, 6

[82] Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, HsiangTao 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. 2

[83] Xiaozheng Zheng, Chao Wen, Zhuo Su, Zeran Xu, Zhaohu Li, Yang Zhao, and Zhou Xue. Ohta: One-shot hand avatar via data-driven implicit priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2024. 3, 5

[84] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13545–13555, 2022. 1, 2

[85] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21057–21067, 2023. 1, 2

[86] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022. 5, 6

[87] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4574–4584, 2023. 1, 2