Synthetic Prior for Few-Shot Drivable Head Avatar Inversion

Wojciech Zielonka[1, 2, 3*]   Stephan J. Garbin[3]   Alexandros Lattas[3]

George Kopanas[3]   Paulo Gotardo[3]   Thabo Beeler[3]   Justus Thies[1, 2]   Timo Bolkart[3]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]Technical University of Darmstadt   [3]Google
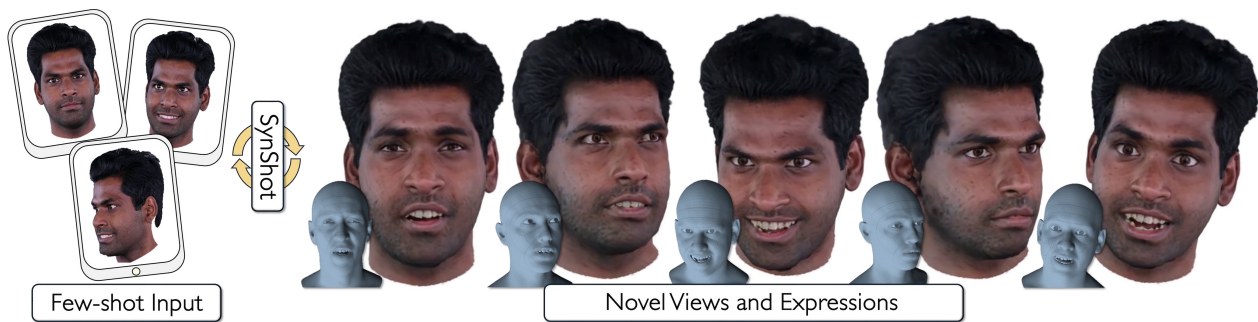https://zielon.github.io/synshot/

Figure 1. Given a few input images (left), **SynShot** generates a personalized 3D Gaussian avatar that renders from new viewpoints and unseen expressions (right). To compensate for the missing information in the input images, we leverage a generative Gaussian head avatar trained on a diverse synthetic head dataset as a 3D prior.

Abstract

*We present **SynShot**, a novel method for the few-shot inversion of a drivable head avatar based on a synthetic prior. We tackle three major challenges. First, training a controllable 3D generative network requires a large number of diverse sequences, for which pairs of images and high-quality tracked meshes are not always available. Second, the use of real data is strictly regulated (e.g., under the General Data Protection Regulation, which mandates frequent deletion of models and data to accommodate a situation when participant's consent is withdrawn). Synthetic data, free from these constraints, is an appealing alternative. Third, state-of-the-art monocular avatar models struggle to generalize to new views and expressions, lacking a strong prior and often overfitting to a specific viewpoint distribution. Inspired by machine learning models trained solely on synthetic data, we propose a method that learns a prior model from a large dataset of synthetic heads with diverse identities, expressions, and viewpoints. With few input images, SynShot fine-tunes the pretrained synthetic prior to bridge the domain gap, modeling a photorealistic head avatar that generalizes to novel expressions and viewpoints. We model the head avatar using 3D Gaussian splatting and a convolutional encoder-decoder that outputs Gaussian parameters in UV texture space. To account for the different modeling complexities over parts of the head (e.g., skin vs hair), we embed the prior with explicit control for upsampling the number of per-part primitives. Compared to SOTA monocular and GAN-based methods, SynShot significantly improves novel view and expression synthesis.*

1. Introduction

The ability to build high-fidelity drivable digital avatars is a key enabler for virtual reality (VR) and mixed reality (MR) applications. However, creating photorealistic human head models [1, 53] using traditional rendering assets requires sophisticated data capture and significant manual cleanup, which is time-consuming and expensive.

The recent advancements in learning-based methods and radiance fields [29, 43] have simplified the avatar creation process, leading to impressive progress in quality and democratization of neural head avatars [19, 42, 62]. Such progress is particularly noticeable in enhancing control

---

through lightweight animation [48, 65, 82], and reducing training time to a few minutes [79]. These methods are trained on multi-view [42, 48, 62] or single-view videos [9, 19, 65, 79], typically requiring hundreds to thousands of video frames. Processing such datasets is complex and error-prone as most methods require tracking a coarse head mesh across all frames, which is typically done by fitting a 3D morphable model [38, 45] to the image. A further limitation of existing personalized head avatars is their poor generalization to facial expressions and camera viewpoints not captured in the set of input images.

Another recent body of work addresses the problem of building 3D head avatars from one or few input images, [8, 11, 71]. However, their rendering quality and fidelity are typically lower than those of methods trained on large datasets (e.g., [48, 82]). To improve quality, some methods [69, 70, 75] first learn a multi-identity head model that is used as prior when optimizing for the personalized avatar. Training these head priors requires a large-scale multi-view image dataset that is expensive and time-consuming to capture. Moreover, managing real data under protection laws like **GDPR** is cumbersome for experimentation and maintenance, as users must periodically (e.g., every 30 days) delete all dataset derivatives and trained models, allowing dataset participants to be removed from both if needed. Alternatively, the FFHQ dataset [28] may be employed, with 4D GAN-based methods [13, 56, 74] constructing an inversion prior from it. However, these approaches tend to exhibit artifacts during novel view synthesis and struggle with preserving identity. In summary, the expressive power of this prior is strongly influenced by: the training data diversity (*e.g.*, ethnicity, age, facial features, expressions), the multi-view capture hardware setup (*i.e.*, lighting, view-density, calibration quality, frame-rate), and the quality of the data pre-processing (*e.g.*, mesh tracking, background masking).

In contrast to the previous work that focuses on expensive and cumbersome real data, we overcome these limitations and propose *SynShot*, a new method that builds a prior solely on synthetic data and adapts to a real test subject requiring only a few input images. Building on the success of ML models trained on synthetic data for tasks like 3D face regression [52], 2D landmark prediction [64], rigid face alignment [3], and few-shot head reconstruction [6, 63, 72], *SynShot* is trained solely on a large synthetic dataset generated from 3DMM samples and diverse assets. Synthetic data offers complete control over dataset creation to meet size and diversity needs for training an expressive head prior, eliminating the need for costly capture hardware and addressing privacy concerns with real subjects. The benefits brought by synthetic data come at the cost of having to handle the domain gap between the trained head prior and real images captured "in the wild". To effectively bridge this gap, we first fit the synthetic prior to real images

and then fine-tune the prior weights to the real data using the pivotal tuning strategy proposed in [49]. With as few as three input images, *SynShot* reconstructs a photorealistic head avatar that generalizes to novel expressions and camera viewpoints (Fig. 1). The results show that our method outperforms state-of-the-art personalized monocular methods [54, 65, 79] trained on thousands of images each. Our method represents head avatars using 3D Gaussian primitives [29], where Gaussian parameters are generated by a convolutional architecture in UV space [39, 50, 82].

In summary, our key contributions are:
1. A generative method based on a convolutional encoder-decoder architecture that is trained on extensive synthetic data only to produce controllable 3D head avatars.
2. A reconstruction scheme that adapts and fine-tunes a pretrained generative model on a few real images to create a personalized, photorealistic 3D head avatar.

2. Related Work

Few-shot Head Avatars. 3D Morphable Models (3DMM) [4, 15, 38, 45] have long been used for creating facial avatars. When paired with generative models for textures [22, 35, 36, 41], 3DMMs can be optimized from in-the-wild images. Techniques such as inverse rendering [14], diffusion-based inpainting [44], and pivotal-tuning [37, 49] are used to disentangle appearance from identity. Neural radiance fields (NeRF) [43] and 3D Gaussian representations (3DGS) [29] have also been widely used for avatar reconstruction. EG3D [7] employs features on tri-planes, enabling consistent 3D face generation and inversion from in-the-wild images. PanoHead [2] extends EG3D through tri-grids to achieve a full 360-degree generation of static human heads. Gaussian3Diff [34] further improves quality by replacing neural features with 3D Gaussians. Rodin [63] and RodinHD [72] leverage an extensive dataset of synthetic humans to train a triplane-based avatar generator used to invert in-the-wild images; however, the results remain confined to the synthetic domain and avatars are not drivable.

Diner [46] incorporates depth information, while Preface [5] trains a volumetric prior on synthetic human data and fits it to a few input images to match a subject's likeness. Cafca [6] extends Preface to better generalize to static but arbitrary facial expressions. In contrast, our method not only bridges the domain gap from synthetic to real but also produces animatable avatars. MofaNeRF [78] and NeRFace [19] condition NeRFs on expression (and shape) codes, while Head-NeRF [25] similarly embeds NeRFs into parametric models. Portrait4D [13] introduces one-shot 4D head synthesis using a transformer-based animatable triplane reconstructor built on the EG3D [7]. Next3D [56] employs GAN-based neural textures embedded on a parametric mesh; however, it suffers from inversion problems. InvertAvatar [74] tackles
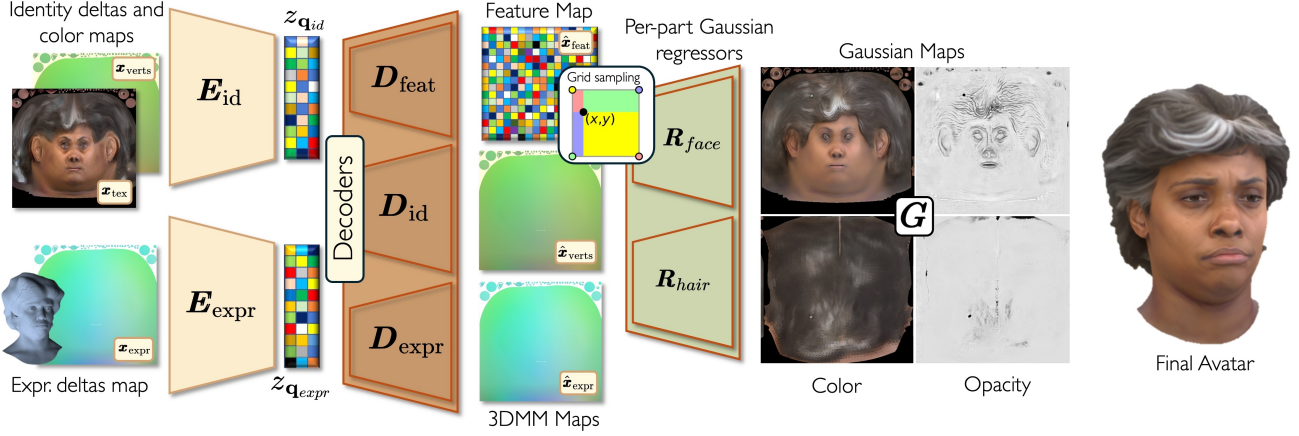
Figure 2. **Pipeline overview.** Given an extracted texture x_{tex}, rasterized position map x_{verts}, and PCA expression deltas x_{exp} our network utilized VQ-VAE to jointly optimize for two latent space z_{expr} and z_{id}. The VQ-VAE decoders predict feature map $D_{\text{feat}}(\mathbf{q}(z_{\text{id}}), \mathbf{q}(z_{\text{expr}})) \rightarrow \hat{x}_{\text{feat}}$, identity and color maps $D_{\text{id}}(\mathbf{q}(z_{\text{id}})) \rightarrow \{\hat{x}_{\text{tex}}, \hat{x}_{\text{verts}}\}$, and expression deltas $D_{\text{expr}}(\mathbf{q}(z_{\text{expr}})) \rightarrow \hat{x}_{\text{expr}}$. Finally, bilinearly sampled maps are passed to per-part regressors R_{color} and R_{gauss} to obtain primitives to rasterize.

the shortcomings of Next3D and further refines avatar inversion using few-shot images. Despite relatively good frontal performance, these GAN-based methods often exhibit artifacts, such as identity changes, in novel view synthesis.

Recent methods by Xu *et al*. [70] are conceptually similar to HeadNeRF and MofaNeRF; however, instead of embedding NeRF [43] on a mesh, they employ 3DGS [29]. GPHM [70] uses a series of MLPs to generate Gaussian primitives attached to a parametric model, enabling expression control and inversion, though it conditions only the avatar's shape. GPHMv2 [69] extends GPHM with a dynamic module for improved reenactment control and a larger dataset, further enhancing quality. HeadGAP [75] also models avatars using MLPs, utilizing part-based features and additional color conditioning to improve quality. While these methods embed primitives directly on the mesh surface, our approach explicitly learns the primitive parameters by modeling their distribution via a VQ-VAE [61], eliminating the need for a guiding mesh during the test time as the shape is captured within our latent space.

Multi-view Personalized Avatars. Volumetric primitives, combined with multi-view training, are highly effective for modeling human heads [23, 27, 31, 50, 57, 58, 68, 82] as they capture intricate details like hair and subsurface scattering [51].VolTeMorph [21] embeds a NeRF within tetrahedral cages that guide volumetric deformation. Qian *et al*. [48] attach Gaussian primitives to 3DMM triangles, whose local rotations and stretch deform the Gaussians without requiring neural networks. Xu *et al*. [68] and Giebenhain *et al*. [23] extend that work to further predict corrective fields over the Gaussians; rather than colors, they splat features that are translated into color by an image-space CNN [60]. Lombardi *et al*. [40] position 3D voxels with RGB

and opacity values at the vertices of a head mesh, using ray tracing for volumetric integration. Saito *et al*. [50] improve quality by replacing voxel primitives with 3D Gaussians and applying rasterization. Our VQ-GAN training aligns with these principles *for few-shot capture*, as we supervise the process using a hybrid mesh-primitive approach to model the generative distribution.

Monocular Personalized Avatars. Monocular methods often rely on a strong 3DMM prior, as recovering a 3D shape from a 2D image is an inherently under-constrained problem. Face2Face [59] was a seminal work that enabled real-time reconstruction and animation of a parametric model. However, it lacks detailed hair representation and relies on low-frequency PCA texture models, which significantly affects quality. This limitation has led to the rise of neural avatars based on NeRF [10, 18– 20, 62, 66, 67, 71, 76, 76, 77] and later on 3D Gaussian primitives [9, 32, 54, 65]. INSTA [79] applies triangle deformation gradients [55] to each NeRF sample based on proximity to the nearest triangle, enabling avatar animation. This approach has been adapted to 3D Gaussian Splatting (3DGS) by methods like Flash Avatar [65] or Splatting Avatar [54]. Unlike *SynShot*, these monocular methods do not generalize well to novel views and expressions. Moreover, they require three orders of magnitude more real data to create a single avatar. *SynShot* overcomes this by leveraging a synthetic prior during few shot avatar inversion, achieving high-quality results.

3. Method

This section describes *SynShot*, how we train the synthetic prior to generate drivable 3D Gaussian head avatars, and how we use it for few-shot head avatar reconstruction.

3.1. Preliminaries

We represent a base 3D mesh as $\mathbf{S} = \bar{\mathbf{S}} + \boldsymbol{\delta}\mathbf{B}_{id} + \boldsymbol{\gamma}\mathbf{B}_{expr}$, where $\bar{\mathbf{S}}$ is the average shape, $\mathbf{B}_{(id,expr)}$ are the bases for identity and expression of a 3DMM, and $\boldsymbol{\delta}$, $\boldsymbol{\gamma}$ denote the corresponding coefficients. Additionally, we use linear blend skinning (LBS) for head rotation around the neck with pose corrective offsets, and to rotate the eyeballs.

The head avatar is rendered via 3D Gaussian splatting, using the CUDA implementation of 3DGS [29]. The rasterizer is defined as $\mathcal{R}(\boldsymbol{G}, \mathbf{K}) \to \bar{\mathcal{I}}$, for a camera $\mathbf{K}$ and a set of n 3D Gaussians $\boldsymbol{G} \in \mathbb{R}^{n \times (11 + 16 \times 3)} := \{\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\alpha}, \boldsymbol{h}\}$, with position $\boldsymbol{\phi} \in \mathbb{R}^{n \times 3}$, rotation $\boldsymbol{\theta} \in \mathbb{R}^{n \times 3 \times 3}$, scale $\boldsymbol{\sigma} \in \mathbb{R}^{n \times 3}$, opacity $\boldsymbol{\alpha} \in \mathbb{R}^n$, and the (third-degree) spherical harmonics parameters $\boldsymbol{h} \in \mathbb{R}^{n \times 16 \times 3}$, where n is the number of Gaussians. See Kerbl *et al.* [29] for more details.

3.2. Gaussian Prior Model

Our prior is modeled as a generative convolutional network with additional lightweight regressors that output Gaussian 2D maps, i.e. multichannel parameter textures. To sample a flexible number of Gaussian primitives, UV positions and features are bilinearly interpolated from intermediary feature maps, before decoding the standard Gaussian attributes that are rendered using $\mathcal{R}(\cdot)$. The architecture of the prior learned by *SynShot* is illustrated in Fig. 2.

Drivable VQ-VAE. Our network has an encoder-decoder architecture based on the VQ-VAE [61]. We follow the approach of Esser *et al.* [16], and use a transformer operating in a quantized latent codebook space to better model long-range dependencies between encoded patches in images. The input to the encoder consists of an RGB texture map $\boldsymbol{x}_{\text{tex}} \in \mathbb{R}^{H \times W \times 3}$, an XYZ vertex position map $\boldsymbol{x}_{\text{verts}} = \mathcal{R}_{uv}(\boldsymbol{\delta}\mathbf{B}_{id}) \in \mathbb{R}^{H \times W \times 3}$ representing the rasterized positions of the neutral mesh, and an expression map $\boldsymbol{x}_{\text{exp}} = \mathcal{R}_{uv}(\boldsymbol{\gamma}\mathbf{B}_{expr}) \in \mathbb{R}^{H \times W \times 3}$ denoting rasterized expression offsets from the neutral mesh, where $\mathcal{R}_{uv}(\cdot)$ denotes UV space rasterization. The encoder network consists of two parallel branches, one for identity and one for expression. This way we explicitly disentangle static components, such as face shape and appearance, from dynamic ones, such as wrinkles, and self-shadowing using two separate latent spaces. We denote them as $\boldsymbol{E}_{\text{id}}(\boldsymbol{x}_{\text{tex}}, \boldsymbol{x}_{\text{verts}}) \to \boldsymbol{z}_{\text{id}}$, where $\boldsymbol{z}_{\text{id}} \in \mathbb{R}^{h \times w \times n_{\text{id}}}$ is the identity code and $\boldsymbol{E}_{\text{expr}}(\boldsymbol{x}_{\text{exp}}) \to \boldsymbol{z}_{\text{expr}}$ with $\boldsymbol{z}_{\text{expr}} \in \mathbb{R}^{h \times w \times n_{\text{expr}}}$ representing the expression code. The identity and expression latents undergo element-wise quantization $\mathbf{q}(\cdot)$. For simplicity, we omit the subscript and let $\boldsymbol{z} \in \{\boldsymbol{z}_{\text{id}}, \boldsymbol{z}_{\text{expr}}\}$ denote identity and expression latent codes, with spatial codes $z_{ij} \in \mathbb{R}^n$, which we quantize by:

$$\mathbf{q}(\boldsymbol{z}) := \left(\arg\min_{z_k \in \mathcal{Z}} \| z_{ij} - z_k \| \right), \qquad (1)$$

with a learned discrete codebook $\mathcal{Z} = \{z_k\}_{k=1}^K$, with

$z_k \in \mathbb{R}^n$. The quantized latent codes are fed into the decoder, which is implemented as three output branches: a feature map decoder, $\boldsymbol{D}_{\text{feat}}(\mathbf{q}(\boldsymbol{z}_{\text{id}}), \mathbf{q}(\boldsymbol{z}_{\text{expr}})) \to \hat{\boldsymbol{x}}_{\text{feat}} \in \mathbb{R}^{H \times W \times F}$ with F-dimensional feature vectors per texel; an identity map decoder, $\boldsymbol{D}_{\text{id}}(\mathbf{q}(\boldsymbol{z}_{\text{id}})) \to \{\hat{\boldsymbol{x}}_{\text{tex}}, \hat{\boldsymbol{x}}_{\text{verts}}\}$; and an expression decoder, $\boldsymbol{D}_{\text{expr}}(\mathbf{q}(\boldsymbol{z}_{\text{expr}})) \to \hat{\boldsymbol{x}}_{\text{expr}}$. Given the output vertex position and expression maps, $\hat{\boldsymbol{x}}_{\text{verts}}$ and $\hat{\boldsymbol{x}}_{\text{expr}}$, the positions of the Gaussian primitives are then computed as $\boldsymbol{\phi} = \hat{\boldsymbol{x}}_{\text{verts}} + \hat{\boldsymbol{x}}_{\text{expr}}$.

Gaussian Primitives Regression. A common limitation of using CNNs to regress Gaussian maps is the fixed output resolution, which ties the number of primitives to the output dimensions. This restriction can significantly limit the quality of the reconstructed avatar (see Table 1). To address this issue, we use a part-based densification mechanism. Similar to Kirschstein *et al.* [32], we use bilinear sampling, $\mathcal{B}(\cdot, u, v)$ to sample the output of the decoders at UV-positions (u, v). As different head regions $r \in \{face, hair\}$ have varying requirements for the density of Gaussian primitives, we bilinearly sample separate parameter maps for the face and scalp region, rather than a single joint map. Thus, per-part map sampling acts as adaptive primitive densification for the individual regions to improve visual quality (Table 1).

We define the primitive positions in the 3DMM space using only shape and expression. Global rotation, translation, and linear blend skinning (LBS) are factored out and applied to the primitives just before splatting to place them in the correct world space. We compute initial *per-part Gaussian parameters* for our primitives. Note that we do not use a fixed canonical space [23, 32, 82], as our initialization is derived from predicted position maps. We first obtain positions by sampling $\boldsymbol{\phi}_r = \mathcal{B}(\boldsymbol{\phi}, u_r, v_r)$, for $r \in \{face, hair\}$. Next, for each $\boldsymbol{\phi}_r$, we compute nearest neighbor distance and initialize scale as $\boldsymbol{\sigma}_r = \min_{j \neq i} \| \boldsymbol{\phi}_{r_i} - \boldsymbol{\phi}_{r_j} \|_2$. Initial opacity is set to $\boldsymbol{\alpha} = 0.7$. Finally, the per-part rotations are computed as $\boldsymbol{\theta}_r = \left[\frac{\mathbf{T}}{\|\mathbf{T}\|} \quad \frac{\mathbf{B}}{\|\mathbf{B}\|} \quad \frac{\mathbf{N}}{\|\mathbf{N}\|} \right] \in \mathbb{R}^{h \times w \times 3 \times 3}$, based on the image space gradient:

$$\mathbf{T} = \frac{\partial \boldsymbol{\phi}_r}{\partial u}, \quad \mathbf{B} = \frac{\partial \boldsymbol{\phi}_r}{\partial v}, \quad \mathbf{N} = \mathbf{T} \times \mathbf{B}. \qquad (2)$$

Following common practice [23, 32, 50, 68, 75, 81, 82], we predict a neural corrective field for all Gaussian parameters. For this, we use the regressed feature map $\hat{\boldsymbol{x}}_{\text{feat}}$, sampling $\boldsymbol{s}_r = \mathcal{B}(\hat{\boldsymbol{x}}_{\text{feat}}, u_r, v_r)$, and lightweight regressors composed of four stacked convolutional blocks with skip connections. Per region, we define two regressors:

$$\boldsymbol{R}_{color}(\boldsymbol{s}_r) \to \boldsymbol{h}_r \in \mathbb{R}^{h \times w \times 16 \times 3}, \qquad (3)$$

$$\boldsymbol{R}_{gauss}(\boldsymbol{s}_r) \to \{\delta\boldsymbol{\phi}_r, \delta\boldsymbol{\theta}_r, \delta\boldsymbol{\sigma}_r, \delta\boldsymbol{\alpha}_r\}, \qquad (4)$$

where $\boldsymbol{R}_{color}$ regresses the spherical harmonics coefficients $\boldsymbol{h}_r$, and $\boldsymbol{R}_{gauss}$ regresses additive parameter offsets

$\Delta := \{\delta\boldsymbol{\phi}_r, \delta\boldsymbol{\theta}_r, \delta\boldsymbol{\sigma}_r, \delta\boldsymbol{\alpha}_r\}$ from the per-part Gaussian parameters. Finally, we apply Δ to the primitives of the individual parts, concatenate them, and splat as $\mathcal{R}(\boldsymbol{G}, \mathbf{K}) \rightarrow \bar{\mathcal{I}}$, where $\bar{\mathcal{I}}$ is the final rendered image and $\boldsymbol{G}$ represents the combined Gaussian primitives.

Training Objectives. We supervise the training of our model by minimizing the photometric loss:

$$\mathcal{L}_{\text{color}} = \alpha\mathcal{L}_{L1} + \beta\mathcal{L}_{\text{SSIM}} + \gamma\mathcal{L}_{\text{LPIPS}} \tag{5}$$

between the pairs of input and output maps $\{\boldsymbol{x}_{\text{tex}}, \hat{\boldsymbol{x}}_{\text{tex}}\}$, $\{\boldsymbol{x}_{\text{verts}}, \hat{\boldsymbol{x}}_{\text{verts}}\}$, and $\{\boldsymbol{x}_{\text{exp}}, \hat{\boldsymbol{x}}_{\text{expr}}\}$, and between the pairs of target images and the final splatted images $\{\mathcal{I}, \bar{\mathcal{I}}\}$.

Additionally, the position maps $\hat{\boldsymbol{x}}_{\text{verts}}$ and expression maps $\hat{\boldsymbol{x}}_{\text{expr}}$ are supervised by $\mathcal{L}_{\text{geom}} = \delta\mathcal{L}_{L1}$. The final loss is defined as $\mathcal{L} = \mathcal{L}_{\text{color}} + \mathcal{L}_{\text{geom}}$. Moreover, we apply L_2 regularization on position, scale, opacity, and the FC ($l >= 1$) part of the spherical harmonics coefficients: The final loss is defined as $\mathcal{L} = \mathcal{L}_{photo} + \mathcal{L}_{reg}$. We train our network end-to-end using 8 GPUs Nvidia H100 with batch size 16 (2 per GPU). We optimize the network for 500K iterations with the Adam optimizer [30] with lr=$1.3e^{-5}$ and multi-step scheduler which decays the learning rate every milestone by gamma=0.66.

3.3. Few-shot Avatar Reconstruction

To bridge the gap between in-the-wild and synthetic avatars, we carefully designed a two-stage inversion process based on pivotal fine-tuning [49]. First, we optimize the encoder $\boldsymbol{E}_{\text{id}}$ while keeping the rest of the network fixed such that we recover $\boldsymbol{z}_{\text{id}}$. Note that $\boldsymbol{E}_{expr}$ remains unchanged as it should model independent expressions. Once $\boldsymbol{E}_{id}$ is fine-tuned, we fix its predicted identity latent code $\boldsymbol{z}_{\text{id}}$, we fine-tune the decoders $\{\boldsymbol{D}_{\text{feat}}, \boldsymbol{D}_{\text{id}}, \boldsymbol{D}_{\text{expr}}\}$ and the regressors $\{\boldsymbol{R}_{color}, \boldsymbol{R}_{gauss}\}$ for the hair and face regions (Fig. 3).

To make the problem tractable, we employ a few heuristics to aid the optimization. These include early stopping with a warmup phase and an exponential moving average on the loss to determine the stopping criteria. Additionally, we scale the number of optimization steps based on the number of training frames, using a constant factor of 10 to increase the likelihood that each sample is seen at least once. As a training objective, in addition to our photometric term $\mathcal{L}_{\text{color}}$ (Eq. 5), we follow Lattas *et al.* [37] and, based on ArcFace [12], define two additional objectives: $\mathcal{L}_{\text{id}}$ and $\mathcal{L}_{\text{arc}}$. The final inversion loss is equal to $\mathcal{L} = \mathcal{L}_{color} + \mathcal{L}_{arc} + \mathcal{L}_{id}$. For a number of views, up to 20, the optimization takes less than **10** minutes on a single Nvidia H100 which is comparable to INSTA [79]. The training time increases with the number of frames as we scale the iterations accordingly.

3.4. Synthetic Dataset

Our dataset consists of approximately 2,000 unique identities, which we render with resolution 768×512 using
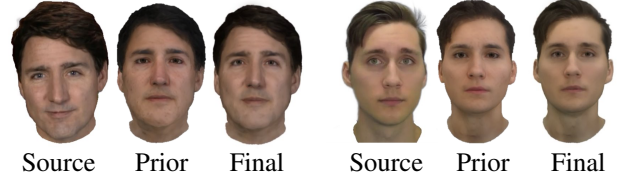

Figure 3. Result of the pivotal tuning before (Prior) and after fine-tuning the model decoders and regressors (Final).

Blender (Cycles) following Wood *et al.* [64], see Fig. 4. We positioned fourteen cameras in front of the subject and an additional fourteen cameras sampled from the upper hemisphere, centered on the scene. We randomly assign assets such as hairstyles and beards to these avatars. Additionally, we utilize high-quality face textures which are randomly distributed among the samples. By combining different shapes and appearances, we augment the set of identities, following practices in synthetic data [64] and 3D face reconstruction [14, 37, 44]. To incorporate tracked expressions from multi-view setups, we propagate them to the avatars during sequence rendering. We additionally compute a hair proxy from strands by voxelizing and fitting it to the scalp region; we apply the same approach for beards. Using a neutral mesh and its hair proxy, we backproject the images onto the texture map. During test time, we use a 3DMM regressor and the input images to extract a texture, which is then used as an initialization for our method. In total, our dataset comprises 14 million images.

Figure 4. Random samples of our synthetic dataset show a diverse range of identities, expressions, and hairstyles that would be challenging to capture in an in-house studio with real subjects.

4. Results

We compare *SynShot* to two different types of methods, state-of-the-art personalized monocular methods, and inversion-based general methods. The personalized monocular methods are controlled by FLAME [38] meshes and include INSTA [79], Flash Avatar [65], and SplattingAvatar [54]. For monocular methods, we used an ensemble of four datasets [19, 24, 76, 82] processed using the face tracker from Zielonka *et al.* [80]. SplattingAvatar follows the approach of Zheng *et al.* [76] and uses the monocular 3D face regressor DECA [17] for tracking. In our experiments, we adopted a similar approach, employing an in-house regressor, similar to DECA, to estimate 3DMM expression and pose parameters. While these methods produce photorealis-
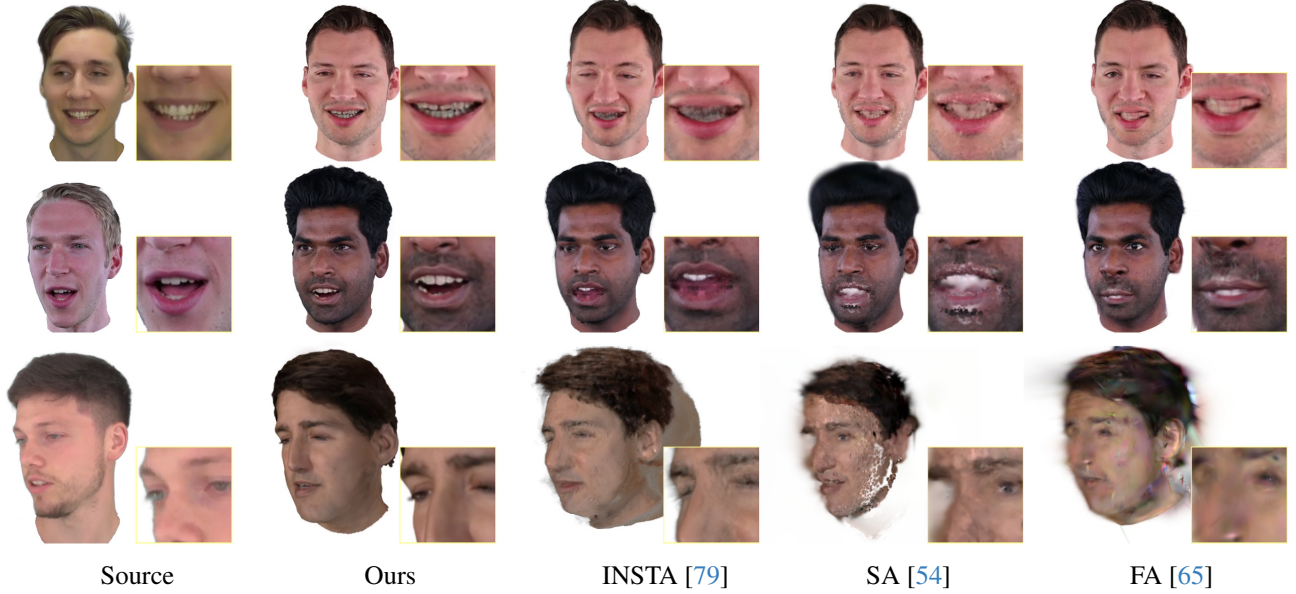
Figure 5. Cross-reenactment comparison of *SynShot* inversion using only **3** views to state-of-the-art (SOTA) methods: **INSTA** [79], Flash Avatar (**FA**) [65], and Splatting Avatar (**SA**) [54], each of which was trained on an average of **3000** frames. It is evident that without a strong prior, these methods fail to generalize to novel expressions and views. Inversion input images are in the supplemental materials.

Figure 6. Novel view evaluation of long hair and beard inversion using only three input images demonstrates the strong generalization capability of SynShot.
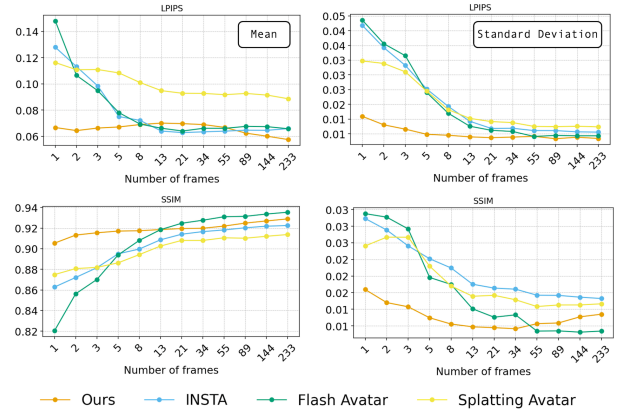

Figure 7. We evaluated the reconstruction error with respect to the number of frames using LPIPS and SSIM metrics. For each frame count, we report the average error (left) and standard deviation (right) over 600 frames across 11 subjects, highlighting the importance of our synthetic prior.

tic avatars, they struggle with generalization to novel views and poses (see Figure 5). For inversion-based methods, we compare PanoHead [2], HeadNeRF [25], and MofaNeRF [78]. We also compare to concurrent works including Portrait4D [13], Next3D [56], and InvertAvatar [74] (Figures 8 and 9). We use three images for all inversion experiments see supp. material. Figure 6 presents a novel view evaluation of challenging long hair and beard inversion, demonstrating the generalization capabilities of *SynShot*.

Evaluation. To measure the performance of *SynShot* without introducing bias, we selected training frames from $\{F_n\}_{n=1}^{16} = \{1, \ldots, 987\}$, where F_n denotes the Fibonacci sequence. For all experiments, we use progressive farthest point sampling [47] in the 3DMM expression space to select a specified number of frames from the training set. The self-reenactment sequences were evaluated using LPIPS and SSIM on the last 600 frames from the INSTA dataset [79].

Monocular Avatar Self-Reenactment. Our combined dataset consists of eleven monocular sequences (512×512 resolution), many of which are in-the-wild videos with very limited head motion, resulting in a low error as the test sequences closely resemble the training data, leaving limited room to assess diversity. To address this and accurately measure the effective error, we trained each method on a varying number of frames, corresponding to frames used in our inversion pipeline. The reconstruction error is evaluated on 600 test frames. Figure 7 demonstrates the effective-
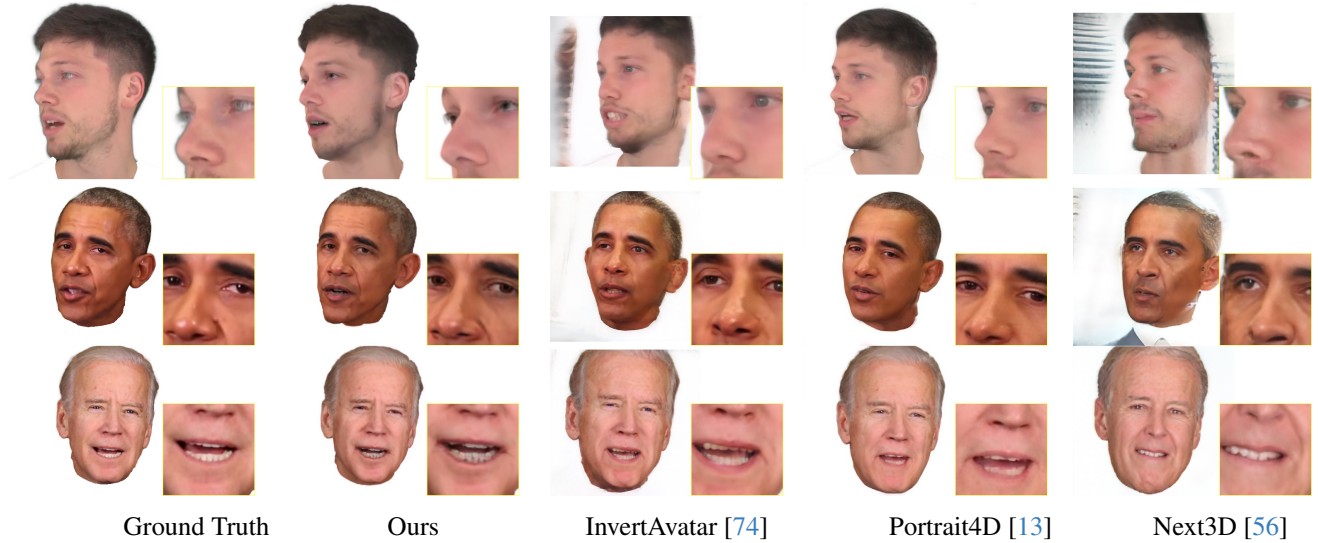
Figure 8. The GAN-based self-reenactment comparison again shows that SynShot better captures identity and synthesizes novel views, proving its usefulness as a synthetic prior and for pivotal fine-tuning in inversion. LPIPS scores: Ours (**0.0236**), InvertAvatar (0.0962), Portrait4D (0.0843), and Next3D (0.2274). Inversion input images can be found in supplemental materials.
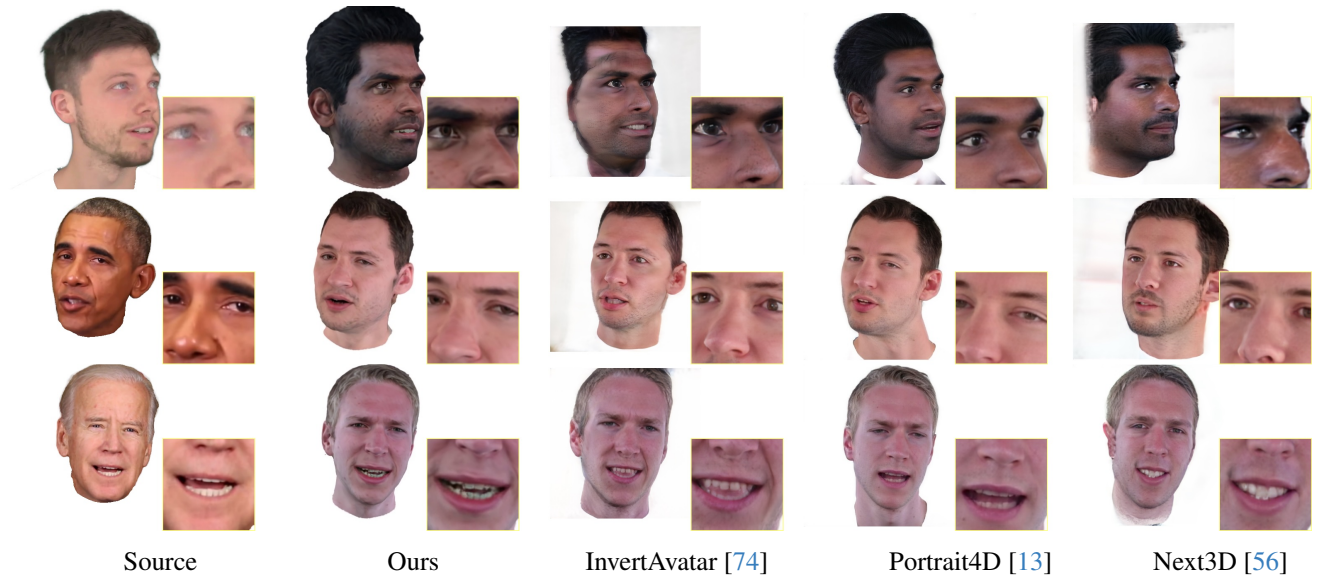
Ground Truth   Ours   InvertAvatar [74]   Portrait4D [13]   Next3D [56]

Source   Ours   InvertAvatar [74]   Portrait4D [13]   Next3D [56]

Figure 9. The GAN-based cross-reenactment comparison shows that our method better reconstructs the target subject's appearance (identity) and remains faithful to the source subject's head poses and expressions, whereas the other methods suffer from artifacts.

ness of our inversion, particularly with up to 233 training frames. Due to the lack of a strong prior, monocular methods fail in low training frame regimes, and, even with larger training datasets, they do not perform well and produce artifacts. Please note that to benefit from an increased number of input frames, i.e., to ground the avatar reconstruction more on the input than the synthetic prior, it requires an increased number of optimization iterations during pivotal fine-tuning. The number of iteration steps affects the metrics, causing LPIPS to vary non-monotonically.

Monocular Avatar Cross-Reenactment. We would like to emphasize the importance of evaluating cross-reenactment, which often reveals issues with generalization and overfitting; however, these aspects are frequently underemphasized, as evaluation sequences are commonly not sufficiently challenging. For instance, Figure 7 indicates that 13 frames may be sufficient for monocular methods to perform well on the test set. Despite achieving high-quality results, most monocular methods [19, 24, 54, 65, 79] struggle with cross-reenactment involving novel expressions and

views. In the supp. mat. we present a full evaluation. Without a strong prior, these methods frequently exhibit artifacts when driven by out-of-distribution sequences. In contrast, our method, leveraging only three images and a synthetic prior with effective shape-expression disentanglement, is able to invert an avatar that significantly outperforms state-of-the-art models trained on thousands of frames. Figure 5 demonstrates cross-reenactment, with the leftmost column serving as the source for expression and view. This shows that incorporating a strong prior enhances the visual quality.

Architecture	L1 $\downarrow$	LPIPS $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$
$F = 128$	*0.0356*	**0.2686**	0.8189	*20.1536*
Tex. up-sampling	**0.0352**	*0.2695*	**0.8196**	**20.1909**
Single Layer	0.0369	0.2702	0.8177	19.8871
$F = 32$	0.0375	0.2732	0.8146	19.7002
w/o VQ	0.0396	0.2747	0.8122	19.2861
$F = 64$	0.0400	0.2765	0.8104	19.2731
No Sampling	0.0403	0.2853	0.8158	19.9787
256×256	0.0365	0.2865	*0.8194*	20.4010

Table 1. We evaluated various configurations of our VQ-VAE. Each configuration uses the final textures of 512×512, unless stated otherwise. As our final model ($F = 128$) we selected the one which produces sharpest results in terms of LPIPS.

GAN-based baselines. We compared SynShot to three animatable GAN-based methods. For our method and InvertAvatar [74], we used three input images, whereas Portrait4D [13] and Next3D [56] are single-shot. Figure 8 presents qualitative self-reenactment results, with additional quantitative LPIPS scores: Ours (**0.0236**), InvertAvatar (0.0962), Portrait4D (0.0843), and Next3D (0.2274). Both results show that SynShot significantly outperforms the baselines. Moreover, Figure 9 presents expression transfer, where our method best captures the subject's identity and is more stable for novel views and expressions, whereas GAN-based methods tend to introduce artifacts in side views.

VQ-VAE Architecture Ablation. Table 1 presents an ablation study of our VQ-VAE architecture. Each model was evaluated on 50 test actors excluded from the training set. Our best model, in terms of sharpness and quality, regresses a feature map $\hat{x}_{\text{feat}} \in \mathbb{R}^{H \times W \times F}$, where $F = 128$, at a resolution of 512×512. Regressing Gaussian primitives directly (*No Sampling*) suffers from lack of quality. Using a *Single Layer* instead of two (for hair + face) results in a lower number of Gaussians, which also decreases the final quality. A key feature of our network is densification through texture sampling. In the (*Tex. up-sampling*) experiment, we predict feature maps at 256×256 resolution compared to 512×512 and apply bilinear sampling to upscale the per-region sampled feature maps to 512×512. This approach achieves results that are almost on par while saving VQ-VAE computation and memory. Finally, using codebook quantization

of latent space improves the final image quality (*w/o VQ*).

Loss	L1 $\downarrow$	LPIPS $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$
$\mathcal{L}_{photo} + \mathcal{L}_{VGG} + \mathcal{L}_{Id} + \mathcal{L}_{ArcFeat}$	0.0229	0.0776	0.9073	23.7474
$\mathcal{L}_{photo} + \mathcal{L}_{VGG}$	0.0244	0.0839	0.9058	23.1191
$\mathcal{L}_{photo} + \mathcal{L}_{VGG} + \mathcal{L}_{Id}$	0.0246	0.0848	0.9048	23.1949
$\mathcal{L}_{photo} = \mathcal{L}_{L1} + \mathcal{L}_{SSIM}$	0.0217	0.0904	0.9094	23.7331

Table 2. Ablation for our inversion losses.

Inversion Ablation. Our inversion pipeline consists of several losses that help bridge the gap between synthetic and real images. This is an important step in our pipeline, as real subjects often have appearance and illumination conditions that differ significantly from our distribution. To address this, we rely on pixel-wise losses and, most importantly, on perceptual losses, which have been shown to aid in effectively matching two distributions [5, 6, 26, 37, 73]. Table 2 shows the inversion reconstruction error using different combinations of losses. As can be seen, using only $\mathcal{L}_{photo}$ is insufficient. The combination of $\mathcal{L}_{VGG}$, based on AlexNet [33], $\mathcal{L}_{ID}$, and $\mathcal{L}_{ArcFeat}$ provides the best results.

5. Discussion

While significantly outperforming monocular methods, *SynShot* has certain limitations that we identify. A key challenge is bridging the domain gap between synthetic and real data. There is considerable room for improvement in the generation of synthetic data. For example, all our synthetic subjects share the same teeth geometry and texture. As a consequence, teeth in our inverted head avatars often closely follow the prior and do not adapt easily. Furthermore, our synthetic data lacks diverse expression-dependent wrinkles, affecting its overall visual quality. Additionally, our dataset was ray-traced with a single environment map, limiting generalization to varied lighting conditions.

6. Conclusion

We have proposed *SynShot*, a method for reconstructing a personalized 3D Gaussian head avatar from just a few images. *SynShot* builds a generative head avatar purely from synthetic data and then utilizes this model as a prior in an inversion pipeline. This inversion pipeline follows a pivotal tuning strategy that successfully bridges the domain gap between the prior and the real input images. We demonstrate that our personalized head avatars generalize better to unseen expressions and viewpoints than SOTA head avatars.

References

[1] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. The digital emily project: photoreal facial modeling and animation. In *ACM SIGGRAPH 2009 Courses*. Association for Computing Machinery, 2009. 1

[2] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. PanoHead: Geometry-aware 3D full-head synthesis in 360deg. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20950–20959, 2023. 2, 6

[3] Jan Bednarík, Erroll Wood, Vasileios Choutas, Timo Bolkart, Daoye Wang, Chenglei Wu, and Thabo Beeler. Learning to stabilize faces. *Computer Graphics Forum (CGF)*, 43(2), 2024. 2

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999. 2

[5] Marcel C. Buehler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, and Abhimitra Meka. Preface: A data-driven volumetric prior for few-shot ultra high-resolution face synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 8

[6] Marcel C. Buehler, Gengyan Li, Erroll Wood, Leonhard Helminger, Xu Chen, Tanmay Shah, Daoye Wang, Stephan Garbin, Sergio Orts-Escolano, Otmar Hilliges, Dmitry Lagun, Jérémy Riviere, Paulo Gotardo, Thabo Beeler, Abhimitra Meka, and Kripasindhu Sarkar. Cafca: High-quality novel view synthesis of expressive faces from casual few-shot captures. In *SIGGRAPH Asia Conference Papers (SA)*, 2024. 2, 8

[7] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, 2022. 2

[8] Xiyi Chen, Marko Mihajlovic, Shaofei Wang, Sergey Prokudin, and Siyu Tang. Morphable diffusion: 3D-consistent diffusion for single-image avatar creation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10359–10370. IEEE, 2024. 2

[9] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Mono-GaussianAvatar: Monocular gaussian point-based head avatar. In *SIGGRAPH Conference Papers (SA)*, page 58. ACM, 2024. 2, 3

[10] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[11] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar, 2024. 2

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 5

[13] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 6, 7, 8

[14] Abdallah Dib, Luiz Gustavo Hafemann, Emeline Got, Trevor Anderson, Amin Fadaeinejad, Rafael MO Cruz, and Marc-André Carbonneau. MoSAR: Monocular semi-supervised model for avatar reconstruction using differentiable shading. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1770–1780, 2024. 2, 5

[15] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3D morphable face models–past, present and future. *Transactions on Graphics (TOG)*, 39(5):157:1–157:38, 2020. 2

[16] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 4

[17] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *Transactions on Graphics, (Proc. SIGGRAPH)*, 40(4):88:1–88:13, 2021. 5

[18] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J. Black. Learning disentangled avatars with hybrid 3D representations, 2023. 3

[19] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2021. 1, 2, 5, 7

[20] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. *Transactions on Graphics (TOG)*, 41(6):1–12, 2022. 3

[21] Stephan Garbin, Marek Kowalski, Virginia Estellers, Stanislaw Szymanowicz, Shideh Rezaeifar, Jingjing Shen, Matthew Johnson, and Julien Valentin. VolTeMorph: Real-time, controllable and generalizable animation of volumetric representations. *Computer Graphics Forum*, 43, 2024. 3

[22] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1164, 2019. 2

[23] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. NPGA: Neural parametric gaussian avatars. In *SIGGRAPH Conference Papers (SA)*, 2024. 3, 4

[24] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular RGB videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18653–18664, 2022. 5, 7

[25] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. HeadNeRF: A real-time nerf-based parametric

head model. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6

[26] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016. 8

[27] Berna Kabadayi, Wojciech Zielonka, Bharat Lal Bhatnagar, Gerard Pons-Moll, and Justus Thies. GAN-Avatar: Controllable personalized GAN-based human head avatar. In *International Conference on 3D Vision (3DV)*, 2024. 3

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *Transactions on Graphics (TOG)*, 42(4), 2023. 1, 2, 3, 4

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015. 5

[31] Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. DiffusionAvatars: Deferred diffusion for high-fidelity 3D head avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5492, 2024. 3

[32] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads. *arXiv preprint arXiv:2406.09377*, 2024. 3, 4

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012. 8

[34] Yushi Lan, Feitong Tan, Di Qiu, Qiangeng Xu, Kyle Genova, Zeng Huang, Sean Fanello, Rohit Pandey, Thomas Funkhouser, Chen Change Loy, et al. Gaussian3diff: 3d gaussian diffusion for 3d full head synthesis and editing. *arXiv preprint arXiv:2312.03763*, 2023. 2

[35] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. AvatarMe: Realistically renderable 3D facial reconstruction "in-the-wild". In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 757–766, 2020. 2

[36] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P Zafeiriou. AvatarMe++: Facial shape and BRDF inference with photorealistic rendering-aware GANs. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 2

[37] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. FitMe: Deep photorealistic 3D morphable model avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8629–8640, 2023. 2, 5, 8

[38] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and ex-

pression from 4D scans. *Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2, 5

[39] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable Gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[40] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *Transactions on Graphics (TOG)*, 40(4), 2021. 3

[41] Huiwen Luo, Koki Nagano, Han-Wei Kung, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, and Hao Li. Normalized avatar synthesis using stylegan and perceptual refinement. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11662–11672, 2021. 2

[42] Shugao Ma, Tomas Simon, Jason M. Saragih, Dawei Wang, Yuecheng Li, Fernando De la Torre, and Yaser Sheikh. Pixel codec avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 64–73. Computer Vision Foundation / IEEE, 2021. 1, 2

[43] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020. 1, 2, 3

[44] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, and Stefanos Zafeiriou. Relightify: Relightable 3D faces from a single image via diffusion models. In *International Conference on Computer Vision (ICCV)*, pages 8806–8817, 2023. 2, 5

[45] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. *IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, 2009. 2

[46] Malte Prinzler, Otmar Hilliges, and Justus Thies. Diner: Depth-aware image-based neural radiance fields. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[47] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, 2017. 6

[48] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic head avatars with rigged 3D gaussians. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20299–20309, 2024. 2, 3

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

[50] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 130–141, 2024. 2, 3, 4

[51] Kripasindhu Sarkar, Marcel C. Bühler, Gengyan Li, Daoye Wang, Delio Vicini, Jérémy Riviere, Yinda Zhang, Sergio Orts-Escolano, Paulo F. U. Gotardo, Thabo Beeler, and Abhimitra Meka. LitNeRF: Intrinsic radiance decomposition for high-quality view synthesis and relighting of faces.

In *SIGGRAPH Asia Conference Papers (SA)*, pages 42:1–42:11. ACM, 2023. 3

[52] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *International Conference on Computer Vision (ICCV)*, pages 1585–1594. IEEE Computer Society, 2017. 2

[53] Mike Seymour, Chris Evans, and Kim Libreri. Meet mike: epic avatars. In *ACM SIGGRAPH 2017 VR Village*. Association for Computing Machinery, 2017. 1

[54] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 5, 6, 7

[55] Robert W. Sumner and Jovan Popović. Deformation transfer for triangle meshes. *ACM SIGGRAPH*, 2004. 3

[56] 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 *CVPR*, 2023. 2, 6, 7, 8

[57] Kartik Teotia, Hyeongwoo Kim, Pablo Garrido, Marc Habermann, Mohamed Elgharib, and Christian Theobalt. GaussianHeads: End-to-end learning of drivable gaussian head avatars from coarse-to-fine representations, 2024. 3

[58] Kartik Teotia, Mallikarjun B R, Xingang Pan, Hyeongwoo Kim, Pablo Garrido, Mohamed Elgharib, and Christian Theobalt. HQ3DAvatar: High-quality implicit 3D head avatar. *Transactions on Graphics (TOG)*, 43(3):27:1–27:24, 2024. 3

[59] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[60] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Transactions on Graphics (TOG)*, 38(4):66:1–66:12, 2019. 3

[61] Aaron Van Den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3, 4

[62] Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. StyleAvatar: Real-time photo-realistic portrait avatar from a single video. In *SIGGRAPH Conference Papers (SA)*, 2023. 1, 2, 3

[63] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4563–4573, 2023. 2

[64] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *International Conference on Computer Vision (ICCV)*, pages 3681–3691, 2021. 2, 5

[65] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. FlashAvatar: High-fidelity head avatar with efficient gaussian embedding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 5, 6, 7

[66] Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. AvatarMAV: Fast 3D head avatar reconstruction using motion-aware neural voxels. In *SIGGRAPH Conference Papers (SA)*, pages 1–10, 2023. 3

[67] Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Huang Han, Qi Guojun, and Yebin Liu. LatentAvatar: Learning latent expression code for expressive neural head avatar. In *SIGGRAPH Conference Papers (SA)*, 2023. 3

[68] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 4

[69] Yuelang Xu, Zhaoqi Su, Qingyao Wu, and Yebin Liu. Gphm: Gaussian parametric head model for monocular head avatar reconstruction. *arXiv preprint arXiv:2407.15070*, 2024. 2, 3

[70] Yuelang Xu, Lizhen Wang, Zerong Zheng, Zhaoqi Su, and Yebin Liu. 3D gaussian parametric head model. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3

[71] 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. *CoRR*, abs/2402.11909, 2024. 2, 3

[72] Bowen Zhang, Yiji Cheng, Chunyu Wang, Ting Zhang, Jiaolong Yang, Yansong Tang, Feng Zhao, Dong Chen, and Baining Guo. Rodinhd: High-fidelity 3d avatar generation with diffusion models. *arXiv preprint arXiv:2407.06938*, 2024. 2

[73] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 8

[74] Xiaochen Zhao, Jingxiang Sun, Lizhen Wang, Jinli Suo, and Yebin Liu. Invertavatar: Incremental gan inversion for generalized head avatars. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 2, 6, 7, 8

[75] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, Guidong Wang, and Xu Lan. HeadGAP: Few-shot 3D head avatar via generalizable gaussian priors. *arXiv preprint arXiv:2408.06019*, 2024. 2, 3, 4

[76] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 5

[77] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. PointAvatar: Deformable point-based head avatars from videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21057–21067, 2023. 3

[78] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. MoFaN-eRF: Morphable facial neural radiance field. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 6

[79] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4574–4584, 2022. 2, 3, 5, 6, 7

[80] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)*, 2022. 5

[81] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3D gaussian avatars. In *International Conference on 3D Vision (3DV)*, 2025. 4

[82] Wojciech Zielonka, Timo Bolkart, Thabo Beeler, and Justus Thies. Gaussian eigen models for human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3, 4, 5