# GaussianHeads: End-to-End Learning of Drivable Gaussian Head Avatars from Coarse-to-fine Representations

KARTIK TEOTIA, Max Planck Institute for Informatics and Saarland Informatics Campus, Germany HYEONGWOO KIM, Imperial College London, United Kingdom PABLO GARRIDO, Flawless AI, United States

MARC HABERMANN, Max Planck Institute for Informatics and Saarland Informatics Campus, Germany MOHAMED ELGHARIB, Max Planck Institute for Informatics, Germany

CHRISTIAN THEOBALT, Max Planck Institute for Informatics and Saarland Informatics Campus, Germany



Fig. 1. GaussianHeads renders photorealistic dynamic 3D human heads in real time. In particular, it renders fine-scale details of (facial) hair and large geometric deformations in the mouth region at high fidelity. Each 3D Gaussian primitive, which is consistently tracked against large deformations over time, is visualized in colors. By providing input images as a control signal, our model can also be used for creating controllable head animations.

Real-time rendering of human head avatars is a cornerstone of many computer graphics applications, such as augmented reality, video games, and films, to name a few. Recent approaches address this challenge with computationally efficient geometry primitives in a carefully calibrated multi-view setup. Albeit producing photorealistic head renderings, they often fail to represent complex motion changes, such as the mouth interior and strongly varying head poses. We propose a new method to generate highly dynamic and deformable human head avatars from multi-view imagery in real time. At the core of our method is a hierarchical representation of head models that can capture the complex dynamics of facial expressions and head movements. First, with rich facial features extracted from raw input frames, we learn to deform the coarse facial geometry of the template mesh. We then initialize 3D Gaussians on the deformed surface and refine their positions in a fine step. We train this coarse-to-fine facial avatar model along with the head pose as learnable parameters in an end-to-end framework. This enables not only controllable facial animation via video inputs but also high-fidelity novel view synthesis of challenging facial expressions, such as tongue deformations and fine-grained teeth structure under large motion changes. Moreover, it encourages the learned head avatar to generalize towards new facial expressions and head poses at inference time. We demonstrate the performance of our method with comparisons against the related methods on different datasets, spanning challenging facial expression sequences across multiple identities. We also show the potential application of our approach

by demonstrating a cross-identity facial performance transfer application. We make the code available on our **project page**.

CCS Concepts: • Computing methodologies  $\rightarrow$  Rendering; Volumetric models.

Additional Key Words and Phrases: Volumetric Rendering, 3D Gaussian Splatting, Implicit Representations, Neural Radiance Fields, Neural Avatars, Free-viewpoint Rendering

# 1 INTRODUCTION

Photorealistic modeling and rendering of human heads is essential in applications such as virtual telepresence, video games, and movies. Achieving an immersive experience in these applications is a long-standing research problem as it necessitates representing facial expressions with a high degree of detail while also ensuring real-time performance. Current approaches struggle with a tradeoff: they either excel at rendering expressions in real time but with limited details in regions, such as mouth interior and hair [Lombardi et al. 2018, 2021; Ma et al. 2021], or offer highly detailed reconstructions at the expense of slower rendering speeds [Gao et al. 2022; Kirschstein et al. 2023; Teotia et al. 2024; Wang et al. 2023a].

Photorealistic avatars are commonly built on top of explicit meshbased representations as they offer consistent topological changes and real-time rendering capabilities [Lombardi et al. 2018; Ma et al. 2021]. However, mesh-based representations present limited capacity to model facial structures exhibiting fine details like hair and beards [Teotia et al. 2024]. To overcome limitations commonly faced by explicit representations, Mixture of Volumetric Primitives (MVP) [Lombardi et al. 2021] proposes a volumetric representation

Authors' addresses: Kartik Teotia, kteotia@mpi-inf.mpg.de, Max Planck Institute for Informatics and Saarland Informatics Campus, Germany; Hyeongwoo Kim, hyeongwoo. kim@imperial.ac.uk, Imperial College London, United Kingdom; Pablo Garrido, pablo. garrido@flawlessai.com, Flawless AI, United States; Marc Habermann, mhaberma@ mpi-inf.mpg.de, Max Planck Institute for Informatics and Saarland Informatics Campus, Germany; Mohamed Elgharib, elgharib@mpi-inf.mpg.de, Max Planck Institute for Informatics, Germany; Christian Theobalt, theobalt@mpi-inf.mpg.de, Max Planck Institute for Informatics and Saarland Informatics Campus, Germany.

based on unstructured cubic primitives. However, cube-based primitives have limited capability to represent fine structures [Teotia et al. 2024; Wang et al. 2023a]. HQ3DAvatar [Teotia et al. 2024] proposes an implicitly learned canonical representation and shows the capability of rendering HD images at interactive speeds, yet it suffers from blurring in regions undergoing large deformations. In this work, we propose a novel end-to-end framework for modeling and rendering dynamic human heads with a high level of photorealism in real time, as shown in Fig. 1. Our method leverages a deformable template face mesh, learnable global transformation parameters as part of our end-to-end learning framework, and 3D Gaussian primitives [Kerbl et al. 2023] in a coarse-to-fine strategy to represent the human head dynamics. This enables high-quality representation of facial details at real-time rates (75 FPS) on a resolution of  $960 \times 540$ . Our method is trained using multi-view images in a person-specific manner. During inference, our method takes the driving videos as input to control the digital avatar under novel expressions, global rigid head poses, and camera viewpoints. It enables controllability via an animation code and global rigid pose parameters, both parameterized by the driving RGB videos. Such parameterization drives the 3D Gaussians in the world space, together with a computationally efficient implementation of the Gaussian primitive properties, such as color and opacity predictions. Our final renderings are obtained via a tile-based rasterizer of the Gaussian primitives [Kerbl et al. 2023]. Fig. 1 shows the deforming Gaussians overlaid onto the RGB renderings of an expression sequence.

More specifically, our method first estimates the global rigid head pose and other facial deformations via an image-based encoder. It then deforms and poses a template mesh in the coarse reconstruction step, which becomes covered with 3D Gaussian primitives in the following refinement step. These primitives are allowed to move beyond regions not represented by the deformed template mesh via a learnable position decoder, and their appearance is learned via color and opacity decoders. Our model is learned from multi-view RGB video using photometric and perceptual loss terms, and newly introduced geometric and temporal constraints in an end-to-end manner. At test time, only a feed-forward pass is performed through the learned network, significantly contributing towards the fast rendering speed.

In summary, we make the following contributions:

- We present a novel method that leverages 3D Gaussian primitives to generate volumetric head avatars. Our method is trained using multi-view RGB video streams via several losses, and during test time, it can render the moving head with large deformations at real-time rates.
- We introduce a coarse-to-fine representation for head and deformation models consisting of a deformable mesh and Gaussian primitives to learn coarse and fine-grained geometry and appearance. This effectively improves the convergence of our model leading to significantly improved photorealism, especially in highly deforming regions, such as the mouth interior.
- We propose an end-to-end training pipeline allowing us to jointly supervise the head pose, geometry, and appearance in a fully differentiable manner.

We evaluate our approach visually and numerically against ground truth data. Here, we ablate our method with different design choices to illustrate their importance in the overall performance. Our approach outperforms existing state-of-the art methods [Lombardi et al. 2021; Qian et al. 2024; Teotia et al. 2024] qualitatively and quantitatively.

# 2 RELATED WORK

#### 2.1 Scene Representations

In recent years, scene representations with neural components have unlocked potential applications in 3D reconstruction. NeRF [Mildenhall et al. 2022] and Deep-SDF [Park et al. 2019] are early pioneering works in modeling 3D scene properties using an MLP-based implicit learning platform. While these methods demonstrated strong results, they suffer from network capacity issues due to inefficient scene encodings. Grid-based strategies such as Instant-NGP [Müller et al. 2022] and other volumetric approaches [Fridovich-Keil et al. 2022; Lombardi et al. 2019] have been proposed for more efficient 3D scene modeling. Such advances led to increased rendering quality as well as better runtime performance. Neuro-explicit representations use point-based priors to improve reconstruction quality and runtime performance further. For instance, Point-NeRF [Xu et al. 2022] introduced a hybrid representation with point clouds and neural features to model a radiance field. Mixture of Volumetric Primitives (MVP) [Lombardi et al. 2021] used cubic primitives weakly attached to the tracked facial geometry to model dynamic 3D scene content.

3D Gaussian Splatting [Kerbl et al. 2023] presented Gaussian primitives as the base representation to model 3D scene content, with recent extensions allowing joint camera and 3D scene modeling [Fu et al. 2024]. Such advances have produced state-of-the-art static 3D scene reconstruction in terms of fidelity and runtime performance. Dynamic 3D Gaussians [Luiten et al. 2024] extended the 3D Gaussian Splatting framework to dynamic scenes but lacked user control. Hence, the scene can only be replayed back from novel camera viewpoints (i.e., no new object deformations) at inference time. Our approach introduces a controllable 3D Gaussian splatting framework for dynamic human heads. It builds on the runtime capabilities and reconstruction fidelity introduced in 3D Gaussian Splatting. Furthermore, it can render the examined head from a novel camera viewpoint under different expressions at test time.

#### 2.2 Mesh-based Head Avatars

Several avatar generation methods rely on explicit scene-tracking provided by 3D morphable models (3DMMs) to aid with the modeling of dynamic expression changes. While there are some methods that operate purely in the 2D space [Bansal et al. 2018; Siarohin et al. 2019], these methods often suffer from limited photorealism and generate talking faces with poor lip-sync. On the other hand, mesh-based scene representations have been used for generating personalized photorealistic head avatars of humans recorded in supervised multi-view studio setups [Lombardi et al. 2018]. They have also been used in neural rendering pipelines to generate portrait avatars from 2D RGB data [Kim et al. 2018; Tewari et al. 2018; Thies et al. 2019, 2016; Wang et al. 2023b]. However, the underlying template fidelity of mesh-based representations is limited. Thus, such representations usually struggle to represent fine details, such as scalp hair, beard, and mouth interior.

#### 2.3 Volumetric Head Avatars

Volumetric scene representations allow for capturing regions that are difficult to model using parametric mesh-based head trackers, e.g., the mouth interior. Hence, they have been an underpinning representation for human head avatars in recent years. Furthermore, their self-supervised training framework from 2D images makes them a very appealing approach. In this section, we further review template-free as well as template-based approaches for volumetric human head avatars.

2.3.1 Template-based Approaches. Most template-based approaches make use of a template mesh-guided canonical space to model facial expressions [Athar et al. 2022; Gao et al. 2022; Xu et al. 2023a; Zhao et al. 2024; Zheng et al. 2022, 2023; Zielonka et al. 2022] in a NeRF framework. Here, methods like INSTA [Zielonka et al. 2022], IM Avatar [Zheng et al. 2022], HAvatar [Zhao et al. 2024], and PointAvatars[Zheng et al. 2023] warp a world-space deformation field or points to canonical space using a nearest neighbor strategy to be able to deform points beyond the template representation. In contrast, our approach predicts per-frame tracking and learns to move Gaussians to regions not covered by the template in an end-to-end fashion. Other approaches, such as HeadNeRF [Hong et al. 2022] and that of Gafni et al. [Gafni et al. 2021] use 3DMM expression coefficients for conditioning the dynamic NeRF volume. Template-based NeRF avatars show impressive reenactment results but with limiting runtime performance due to the underlying coordinate-based representation. Mixture of Volumetric Primitives (MVP) [Lombardi et al. 2021] is a hybrid cube-based primitive model attached to a base mesh that renders head avatars in real-time at 1k resolution. TRAvatar [Yang et al. 2023] builds upon MVP and uses a multispectral Lighstage to reconstruct animatable relightable avatars. Unlike previous multi-view volumetric approaches [Lombardi et al. 2021; Yang et al. 2023], our 3D Gaussian-based approach works with sparser camera rigs and simpler lighting setups. It also renders dynamic heads much faster thanks to its rasterization-based rendering approach and several important design choices. Such choices include a novel efficient vertex decoder and an efficient decoder-only architecture for predicting the properties of Gaussian primitives. Due to the efficiency and rendering quality of 3D Gaussian Splats, research has quickly shifted its attention back to more explicit point representations, modeled as 3D displacements in the mesh surface space [Ma et al. 2024; Qian et al. 2024; Rivero et al. 2024; Xu et al. 2024] or UV mesh space [Lan et al. 2023; Pang et al. 2024; Saito et al. 2024a] of a base mesh or parametric model (e.g., FLAME [Feng et al. 2021]). Such representations allow for dynamic facial control via latent codes [Saito et al. 2024a] or expression parameters [Qian et al. 2024], while facial appearance can be conditioned on text-prompts [Zhou et al. 2024]. Unlike concurrent multi-view work [Qian et al. 2024; Saito et al. 2024a; Xu et al. 2024], we propose an end-to-end, coarseto-fine deformation strategy that allows us to better estimate rigid mesh alignment, which in turn improves the alignment of Gaussians over the head surface, resulting in higher-quality renderings.

2.3.2 Template-free Approaches. Template-free approaches are not limited by the representation capacity of the underlying 3DMM template. Here, methods such as Nerfies [Park et al. 2021a] and HyperNeRF [Park et al. 2021b] use an implicitly learned canonical space to model dynamic scenes captured from a moving monocular camera. LatentAvatar [Xu et al. 2023b] learns implicit latent expression codes between shared and person-specific settings for cross-actor reenactment applications, thus improving tracking and expression transfer. NeRFSemble [Kirschstein et al. 2023] propose a multi-view solution for reconstructing dynamic human heads using a learned blending of multi-resolution hash grids. While the methods above produce good results, their formulation does not extend beyond scene-replay, thus lacking controllability. GAN-based methods [Chan et al. 2022] have shown improvements in reconstruction quality and view consistency. Thus, they have been used in downstream avatar-based applications [Trevithick et al. 2023].

HQ3DAvatar [Teotia et al. 2024] is a more recent method that conditions a canonical space on expression features for photorealistic moving face synthesis. The method is constrained using optical flow in a hash-grid radiance field framework [Müller et al. 2022]. HQ3DAvatar produces high-fidelity reconstruction results and shows clear improvements over several existing methods, including MVP [Lombardi et al. 2021] and other multi-view extensions, such as HyperNeRF [Park et al. 2021b] and NeRFBlendShape [Gao et al. 2022]. However, HQ3DAvatar struggles to reproduce details in regions undergoing strong deformations. Our approach benefits from the dense template mesh as a prior, yet it can flexibly deviate away from regions not covered by the underlying template. Such trade-off results in highly accurate renderings and real-time inference speeds, as demonstrated by extensive experiments.

#### 3 METHOD

Our goal is to produce photorealistic, 3D-consistent moving heads at real-time rendering speeds. We use 3D Gaussians [Kerbl et al. 2023] as the base representation, and introduce several novel losses and design choices to ensure fast rendering and high-quality reconstructions. Our method, illustrated in Fig. 2, utilizes data from a subject's multi-view facial performance, captured with 24 cameras sparsely distributed around the entire head. Such data supervises our end-to-end learning framework, which employs a coarse-to-fine strategy to capture head movements and detailed facial expressions accurately. To train our method, we track FLAME [Li et al. 2017] parameters for each frame using a multi-view facial landmarks-based tracking implementation from Shimada et al. [2023]. At test time, our method requires a forward pass through the trained encoders and decoders to render the subject. Our method has an image encoder that separates input images into a local animation code and global transformations. The animation code drives deformations on top of a template mesh, which is then posed using the global transformation parameters. 3D Gaussians initialized on this posed mesh are refined to capture fine-scale details. The same animation code also drives the learning of opacity and RGB values. The Gaussian properties, as well as the RGB and opacity values, are learned in the 2D UV space of the template mesh, enabling the use of efficient CNN-based decoders. This, combined with the fast rasterization of

4 • Kartik Teotia, Hyeongwoo Kim, Pablo Garrido, Marc Habermann, Mohamed Elgharib, and Christian Theobalt



Fig. 2. **Method Overview**. Multi-view driving images are provided as input, and an image encoder extracts the animation code ( $Z_{exp}$ ) and global head pose parameters ( $R_{rigid} T_{rigid}$ ). The animation code is fed into a vertex deformer network to generate per-vertex displacements ( $\delta_v$ ) relative to a canonical template mesh in a rest pose with vertices  $v_t$ , resulting in an expression-dependent deformed mesh with vertices  $v_d$ . This is then globally transformed to a posed mesh with vertices  $v_p$  using the head pose parameters ( $R_{rigid} T_{rigid}$ ). 3D Gaussians with positions ( $p_{in}$ ), scales ( $s_{in}$ ), and rotations ( $r_{in}$ ) are initialized on the globally transformed deformed mesh. The deltas decoder predicts deformation for position ( $\delta_p$ ), rotation ( $\delta_r$ ), and scale ( $\delta_s$ ) to refine the initialized 3D Gaussians. We employ two decoders to predict per-Gaussian RGB color and opacity aligned to the UV map of the template. 3D Gaussian Splatting projects the deformed Gaussians into the image plane, resulting in the rendered image. The pipeline is trained end-to-end using multi-view RGB images, expression tracking data, and rigid head pose tracking data. During testing, only a feedforward pass of the input is required to drive the global rigid head pose and facial expressions.

3D Gaussian Splatting [Kerbl et al. 2023], enables real-time inference. Notably, our method requires no explicit global rigid pose as an additional input, just the driving images during testing. Thus, it is more robust to jitter, which is usually introduced by explicit global rigid pose-tracking. In this section, we first lay out the image formation process via 3D Gaussian Splatting (Sec. 3.1). Then, we describe our method in more detail, including our encoding strategy (Sec. 3.2), our coarse-to-fine learning scheme (Sec. 3.3), the proposed efficient decoders (Sec. 3.4), and finally the objective functions (Sec. 3.5).

#### 3.1 Preliminaries: 3D Gaussian Splatting

We adopt the image formation model from 3D Gaussian Splatting [Kerbl et al. 2023], which represents the scene using 3D Gaussian primitives. Each Gaussian primitive is represented by a  $3 \times 3$ covariance matrix  $\Sigma$  centered at mean position **p** as follows:

$$q(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x} - \mathbf{p})^T \Sigma^{-1}(\mathbf{x} - \mathbf{p})} .$$
(1)

The covariance matrix is parameterized by a rotation matrix **R** and a scaling matrix **S** to constrain it to be positive semi-definite:

$$\Sigma = \mathbf{R}\mathbf{S}\mathbf{S}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}} \tag{2}$$

where **R** and **S** are the matrix representations of the learnable quaternion **r** and the 3D scaling **s** vectors, respectively. Then, the 3D Gaussians are projected onto screen space using EWA Splatting [Zwicker et al. 2002] as follows:

$$\Sigma' = \mathbf{J}\mathbf{W}\Sigma\mathbf{W}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}$$
(3)

where  $\Sigma'$  is a projected 2×2 covariance matrix.  $\mathbf{J}' \in \mathbf{R}^{2\times 3}$  denotes the Jacobian of the affine-approximated projective transformation matrix and  $\mathbf{W} \in \mathbf{R}^{3\times 3}$  is the viewing transformation. Finally, the pixel-space rendered color  $C_r$  is computed using point-based rendering [Kerbl et al. 2023]:

$$C_r = \sum_{i \in \mathbb{N}} c_i a_i \prod_{j=1}^{i-1} (1 - a_j) .$$
(4)

Here, N denotes the total number of ordered points.  $c_i$  represents the per-Gaussian decoded color, and  $a_i$  represent the result of multiplying the covariance matrix  $\Sigma'$  with the the per-Gaussian decoded opacity  $o_i$ , respectively.

#### 3.2 Encoding

Unlike 3D Gaussian Splatting [Kerbl et al. 2023], which was originally proposed to model static scenes, we aim to control the 3D Gaussians based on RGB images as input. To this end, we introduce an encoder  $E_{\gamma}$  that takes in  $I_d = 3$  multi-view RGB images as input to encode the facial appearance and the global rigid head pose from varying viewpoints. The encoder is designed to encode the changes in local dynamics like facial expression changes via an animation code  $Z_{exp} \in \mathbb{R}^{256}$  and global transformations, separately. Our choice



Fig. 3. **Coarse-to-fine Learning.** We refine the FLAME shape-fitted mesh by an initial registration step (b), which results in the template mesh with vertices  $v_t$  used in our framework. (c) The template mesh is deformed based on the input and posed to the world space using the global transformation parameters. We initialize the 3D Gaussians on the rigidly translated deformed mesh with vertices and refine their properties through our Coarse-to-Fine framework, resulting in dense, head surface-aligned 3D Gaussians (d). The 3D Gaussians are splatted, resulting in the render (e), which is supervised with ground truth image (f).

of encoding multi-view RGB images is similar to related works that encode expressions via multi-view RGB images [Lombardi et al. 2018, 2021; Wang et al. 2021; Yang et al. 2023]. Additionally, the encoder outputs rigid transformation parameters, where the rotation matrix  $\mathbf{R}_{\text{rigid}}$  lies in SO(3), and the translation vector  $\mathbf{T}_{\text{rigid}}$  belongs to  $\mathbb{R}^3$ . The encoder has separate branches for the animation code and the rigid transformation parameters. The rotation output is normalized to ensure it lies on the unit sphere, representing a unit quaternion. While TRAvatar [Tran et al. 2019] shares a similar design choice of predicting the global transformation parameters, we find the need in our framework for constraining these parameters via a landmark alignment loss as weak supervision to guide the positioning of the rigid head pose, as discussed in Sec. 3.5.

#### 3.3 Coarse-to-fine Learning Framework

Our method leverages a coarse-to-fine framework for deforming a template mesh. It first initializes the 3D Gaussians and then refines the positions of the Gaussians. This hierarchical approach is imperative to our high-fidelity results, as discussed in Sec. 4.2. The coarse step handles the large vertex-level deformations that result from extreme or exaggerated facial expressions. The fine deformation step then refines these changes, allowing for capturing the more fine-scale or intricate details like the teeth. Fig. 3 provides a high-level overview of the coarse-to-fine framework.

*3.3.1 Coarse Step.* Before training, we first register the FLAMEfitted mesh on the first frame of the sequence to multi-view RGB images of the subject. Sec. A.1 presents the details of this template mesh generation step. To deform the template mesh to match the subject's expression closely, we utilize a per-vertex offset prediction network  $D_v$ . This network takes the animation code  $Z_{exp}$  as input and decodes per-vertex offsets  $\delta_v$  on top of a neutral expression template mesh  $M_{template}$  with vertices  $v_t$ . This results in deformed vertices  $v_d$  as follows:

$$\mathbf{v}_{\mathrm{d}} = \mathbf{v}_{\mathrm{d}} + \delta_{v} \,. \tag{5}$$

The deformed mesh  $M_d$  with vertices  $v_d$  shares the same topology as FLAME [Li et al. 2017], with a total of 5023 vertices. Since the number of vertices is fairly small, we employ a lightweight MLP-based architecture to learn the per-vertex offsets on top of the template vertices  $v_t$ . The vertices of the deformed mesh are then rigidly posed to the world space using the estimated global rigid rotation  $R_{rigid}$  and translation  $R_{rigid}$  as follows:

$$\mathbf{v}_{p} = \mathbf{R}_{rigid}(\mathbf{v}_{d}) + \mathbf{T}_{rigid} .$$
 (6)

This global transformation results in the posed mesh  $M_{\rm p},$  defined by vertices  $v_{\rm p}.$ 

3.3.2 3D Gaussian Initialization. We initialize N<sub>G</sub> 3D Gaussians on the posed mesh  $M_p$ . To do so, we first uniformly sample on the UV space of the template mesh at a resolution of 512 × 512. This allows for sampling N<sub>G</sub> = N<sub>g</sub><sup>2</sup> 3D Gaussians in 2D space, where N<sub>g</sub> corresponds to the spatial resolution in the x and y direction of the UV map. Once sampled, each Gaussian is assigned a position  $p_{in}$ , rotation  $r_{in}$ , and scale  $s_{in}$ . The position  $p_{in}$  is the interpolated vertex position on the deformed mesh, computed using barycentric interpolation on the UV map. For the first frame, the scales are initialized according to a distance transform-based technique akin to 3D Gaussian Splatting [Kerbl et al. 2023], which serves as the initial scale reference  $s_{in}$ . The initial rotations  $r_{in}$  are set to zero. We apply a binary mask that filters out the samples that do not lie on the UV parameterization of the template mesh.

3.3.3 Fine Step. A 2D CNN-based decoder  $D_m$  is used to decode the animation code  $Z_{exp}$  into deltas  $(\delta_p, \delta_r, \delta_s)$ , which represent the offsets in positions, rotations, and scales with respect to the initialized Gaussian primitives. Thus, we predict the positions  $\mathbf{p} = \mathbf{p}_{in} + \delta_p$ , rotations  $\mathbf{r} = \mathbf{r}_{in} + \delta_r$ , and scales  $\mathbf{s} = \mathbf{s}_{in} + \delta_s$  for the 3D Gaussians. We use a 2D CNN to generate the deltas in a grid of size  $N_G \times N_G \times 10$ . This grid corresponds to the primitives sampled in the UV map. The position offsets  $\delta_p$  represent the fine-level deformation of facial structures showing fine-grained features, such as facial hair and mouth interior, as these offsets allow for 3D Gaussians to be placed in these regions.

#### 3.4 Gaussian Opacity and Color Decoding

To achieve real-time and high-fidelity renderings, it is crucial to decode the opacity and appearance information of each 3D Gaussian efficiently. This is achieved through computationally efficient decoders that translate the animation code into per-Gaussian properties, capturing both the opacity and RGB color information. Our approach leverages 2D CNNs for efficient decoding of both RGB and opacity values. This design choice is similar to several concurrent works [Jiang et al. 2024; Pang et al. 2023; Saito et al. 2024b]. 3.4.1 Opacity Decoder. The animation code  $Z_{exp}$  is decoded into a 2D opacity map O of size  $N_G \times N_G \times 1$  using  $D_0$ . Each value on the grid represents the per-Gaussian opacity  $o_i$ .

3.4.2 RGB Decoder. The animation code  $Z_{exp}$ , along with the objectcentric view direction v, is decoded into a 2D color map  $N_G \times N_G \times 3$ using  $D_{RGB}$ . Each value on the decoded grid represents the per-Gaussian RGB color  $c_i$ . Note that the view direction v is introduced to represent view-dependent appearance effects.

#### 3.5 Objective Function

Given the above representation, we learn the encoder  $E_{\rm gamma}$  and the decoders  $D_v$ ,  $D_o$ ,  $D_{\rm RGB}$ , and  $D_m$  in an end-to-end manner using multi-view image supervision. For this, we optimize for the following objective function:

$$\mathbf{L} = \lambda_1 \mathbf{L}_1 + \lambda_2 \mathbf{L}_{\text{SSIM}} + \lambda_3 \mathbf{L}_{\text{geo}} + \lambda_4 \mathbf{L}_{\text{perc}} + \lambda_5 \mathbf{L}_{\text{temp}} + \lambda_6 \mathbf{L}_{\text{lmk}} + \lambda_7 \mathbf{L}_{\text{reg}} .$$
(7)

Here,  $L_1$  denotes the difference between the rendered RGB image  $I_r$  and ground truth RGB image  $I_g$ . The perceptual distance,  $L_{\rm perc}$ , is based on a pretrained VGG network [Simonyan and Zisserman 2015]. Similar to [Cao et al. 2022], we take 4 random patches of size 500×500 pixels from the ground truth and rendered image pair to compute the perceptual distance.  $L_{\rm SSIM}$  denotes the structural similarity measure [Wang et al. 2004a]. To supervise the deformed mesh  $M_d$  with vertices  $v_d$ , we employ an  $L_2$  loss between the deformed vertices  $v_d$  and the tracked vertices  $v_s$  as follows: the space before the

$$\mathbf{L}_{\text{geo}}(\mathbf{v}_{d}, \mathbf{v}_{s}) = \frac{1}{N_{v}} \sum_{i=1}^{N_{v}} (\mathbf{v}_{d}[i] - \mathbf{v}_{s}[i])^{2}, \qquad (8)$$

where N<sub>v</sub> represents the number of vertices. L<sub>geo</sub> allows the vertices to move freely towards more deformable regions like the mouth interior while remaining close to the tracked vertices  $v_t$ . This acts as a soft prior to preserve the structure of the face. We further use a spatial regularization term L<sub>reg</sub> that ensures stable training. L<sub>reg</sub> is implemented as the summation of L1 penalties for scales s, Gaussian mean position offsets  $\delta p$ , and Gaussian rotations offsets  $\delta r$ : the space before the

$$L_{\text{reg}} = \frac{1}{N_{\text{G}}} \sum_{i=1}^{N_{\text{G}}} (|\mathbf{s}_{i}| + |\delta p_{i}| + |\delta r_{i}|)$$
(9)

Here,  $N_{\rm G}$  represents the total number of Gaussians. In addition, we employ a landmark constraint  $L_{\rm lmk}$  to supervise the global transformation. This constraint ensures that the rigid landmarks of the face align with the rigid landmarks of the rigidly tracked head pose mesh  $M_{\rm r},$  which is at neutral expression.

$$\mathbf{L}_{\rm lmk} = \frac{1}{N_{\rm L}} \sum_{i=1}^{N_{\rm L}} \left| \mathbf{M}_{\rm p,i} - \mathbf{M}_{\rm r,i} \right|$$
(10)

Here,  $N_L$  is the total number of rigid landmarks.  $M_{p,i}$  represents the *i*-th rigid landmark of the posed mesh  $M_p$  and  $M_{r,i}$  denotes the corresponding *i*-th rigid landmark of the rigidly tracked mesh  $M_r$  at a neutral expression. Finally, we employ a temporal smoothness loss  $L_{temp}$  between two consecutive frames at time *t* and *t*+1, defined on the Gaussian offsets predicted by the deltas decoder  $\mathbf{D}_{\mathrm{m}}$  as follows:

$$\mathbf{L}_{\text{temp}} = \frac{1}{N_{\text{G}}} \sum_{i=1}^{N_{\text{G}}} \left( \left| \delta p_{i}^{(t+1)} - \delta p_{i}^{(t)} \right| + \left| \delta \mathbf{s}_{i}^{(t+1)} - \delta \mathbf{s}_{i}^{(t)} \right| + \left| \delta r_{i}^{(t+1)} - \delta r_{i}^{(t)} \right| \right)$$
(11)

In our experiments, we set  $\lambda_1 = 0.8$ ,  $\lambda_2 = 0.2$ ,  $\lambda_3 = 0.1$ ,  $\lambda_4 = 0.01$ ,  $\lambda_5 = 0.1$ ,  $\lambda_6 = 0.8$ , and  $\lambda_7 = 0.1$ .

## 3.6 Implementation Details

We train our method for 300k iterations on a single NVIDIA A40 GPU, which takes around 24 hours. We adopt the Adam optimizer [Kingma and Ba 2015] for estimating the optimal parameters of our learning-based framework. We set the learning rate to 1*e*-4 for D<sub>v</sub>. The Opacity decoder D<sub>o</sub> and RGB decoder D<sub>RGB</sub> are trained with a learning rate of 6*e*-4, while the image encoder E<sub>γ</sub> is trained with a learning rate of 5*e*-4. The delta decoder D<sub>m</sub> is trained with a learning rate of 1*e*-4. We use N<sub>L</sub> = 4 rigid landmarks (2 for each eye) to supervise the landmark alignment. In our framework, the view direction v is calculated as the mean position of the subject's head defined by the posed mesh M<sub>p</sub> with vertices v<sub>p</sub> subtracted by the camera viewpoint center.

#### 4 EXPERIMENTS



Fig. 4. **Qualitative Results.** Novel view synthesis at virtual camera paths for different subjects. Our method excels at representing fine details, such as facial hair and fine scalp hair strands.

Our method produces photorealistic moving heads under novel viewpoints and with unseen expressions. This section discusses the various experiments we have performed to examine our method. We show results on multiple participants recorded in a multi-view facial performance capture rig. Unless stated otherwise, all our experiments used for evaluations are trained with 23 cameras, sparsely distributed over the entire head. We hold out 1 camera for evaluation. Furthermore, the default resolution used in our experiments is  $960 \times 540$ , unless stated otherwise. We show qualitative and quantitative comparisons on 4 subjects performing challenging facial expressions, which includes 1 subject from the NeRSemble dataset



Fig. 5. Qualitative Results. Expression synthesis from a novel camera viewpoint. Our approach can synthesize challenging expressions and motions, such as the tongue sticking out.

[Kirschstein et al. 2023]. We hold out 300 frames for both quantitative and qualitative evaluation unless stated otherwise. We first show qualitative results for a wide variety of sequences (Sec. 4.1). Through quantitative and qualitative analysis, we highlight the important design choices of our method (Sec. 4.2) and perform comparisons against related work, namely MVP [Lombardi et al. 2021], GaussianAvatars [Qian et al. 2024], and HQ3DAvatar [Teotia et al. 2024]. For image sequence results, please refer to the supplemental video.

#### 4.1 **Qualitative Results**



Fig. 6. Ablation Study. Left to right: Without coarse deformations, without fine deformations, without rigid landmark alignment loss  $\ensuremath{L_{\text{lmk}}}\xspace$  , ours, and Ground Truth. The highlighted regions show expression alignment in the mouth region. We observe that our design choices result in high-quality details in the mouth interior. Note: Best viewed if zoomed in or in the supplementary video.

Fig. 4 shows turntable renderings generated by our approach. Fig. 5 shows novel expressions synthesized for 2 different subjects at a holdout camera viewpoint. Our results show high-quality view synthesis, as well as novel expression synthesis. We also show an

Table 1. Analysis of our core design choices. We outperform the baseline approaches in quantitative metrics.

Metrics	No Coarse Def.	No Fine Def.	No L <sub>lmk</sub>	Ours
PSNR ↑	32.95	32.91	33.08	33.91
$L1\downarrow$	7.44	7.56	7.41	4.93
SSIM ↑	0.80	0.80	0.80	0.85
LPIPS $\downarrow$	0.13	0.13	0.11	0.11

application that demonstrates that our model can work with inputs from a parametric face tracker [Li et al. 2017] in the supplementary video.

#### 4.2 Ablative Analysis



Loss loss

Fig. 7. Ablation Study. Left to right: Without perceptual loss, without vertex loss  $L_{geo},$  and ours. The highlighted regions show expression alignment in the mouth region. Note that using a perceptual loss term leads to sharper facial hair and tongue reconstructions, while Lgeo helps with detailed expression alignment.



Without Temporal Loss

Ground Truth

Fig. 8. Ablation Study. Left to right: Without Ltemp, ours, and Ground Truth. Ltemp helps reproduce details in structures undergoing large motion changes like teeth.

Kartik Teotia, Hyeongwoo Kim, Pablo Garrido, Marc Habermann, Mohamed Elgharib, and Christian Theobalt



1 Driving view 2 Driving views Ours Ground Truth (3 Driving views)

Fig. 9. Ablation Study. Left to right: 1 input driving view, 2 input driving views, ours (3 input driving views), and Ground Truth. We observe that our choice of driving the avatar with an image-triplet leads to overall better expression alignment.



Without scale init.

Combined Decoders Ground Truth

Fig. 10. Ablation Study. Left to Right: Without scale initialization, RGCA (--), using combined decoders for 3D Gaussian geometry properties, ours, and Ground Truth. Our design choice generates sharper results in deformable regions like the mouth.

Table 2. Ablation Study: Quantitative evaluation of temporal loss term Ltemp, perceptual loss term Lperc, and vertex loss Lgeo. Our method attains the best numerical performance on almost all metrics (see bold numbers).

Metrics	Without L <sub>temp</sub>	Without Lperc	Without Lgeo	Ours
PSNR ↑	33.64	33.08	32.98	33.91
L1 ↓	4.88	7.37	7.35	4.93
SSIM ↑	0.85	0.81	0.80	0.85
$\Box PIPS \downarrow$	0.11	0.15	0.11	0.11

Table 3. Analysis of the number of input driving views. We observe that driving with more input views increasingly improves head avatar synthesis.

Metrics	1 input view	2 input views	Ours (3 input views)
PSNR ↑	30.94	31.78	33.11
L1 ↓	15.52	9.38	5.58
SSIM ↑	0.67	0.75	0.83

We ablate the main components as well as loss terms employed in our approach. The ablation study in Fig. 6 demonstrates the importance of our core components. Without coarse deformation, the model struggles to recover the structure in the mouth interior. Without fine deformation, the model still lacks fine-grained detail of

Table 4. Ablation Study: Quantitative evaluation with combined decoders for predicting Gaussian properties, RGCA (- -), and without using scale initialization. Our method attains the best numerical performance (see bold numbers).

Metrics	Combined Decoders	RGCA ()	No Scale Init.	Ours
PSNR $\uparrow$	34.25	34.11	33.80	34.79
$L1\downarrow$	4.80	4.95	5.29	3.91
SSIM ↑	0.86	0.86	0.85	0.87
LPIPS $\downarrow$	0.10	0.12	0.11	0.10

facial structures, such as the teeth. Without landmark alignment loss L<sub>lmk</sub>, the model misaligns expressions in the mouth. Our model successfully captures the details in the mouth interior at a high quality. Fig. 7 shows the impact of the perceptual loss term L<sub>perc</sub> and the per-vertex loss term  $L_{\text{geo}}.$  The absence of  $L_{\text{perc}}$  results in overall blurry face renderings.  $\mathbf{L}_{geo}$  helps with expression alignment in regions undergoing large deformations, e.g., the mouth. Fig. 8 highlights the impact of temporal loss  $L_{temp}$ . We find that  $L_{temp}$ helps preserve fine-scale details of structures undergoing motion across time, such as the teeth. Table 1 shows the average imagebased quality assessment metric scores. We use PSNR, SSIM [Wang et al. 2004b], L1 distance, and LPIPS distance [Zhang et al. 2018] as the image-based quality assessment metrics. The scores are averaged over the holdout frames for 2 subjects. Numerical results show that our design choices are well-validated. Tab. 2 shows a quantitative evaluation for different design choices, including the use of the temporal loss term Ltemp, the perceptual loss term Lperc, and the vertex loss term  $\mathbf{L}_{\text{geo}}.$  We observe that our final method, and our method without Ltemp performs on par on quantitative metrics as its effects are more apparent during abrupt facial deformations. Quantitative evaluations focusing solely on the mouth region show improvements in the L1 metric (7.27 with the temporal loss and 7.41 without). To quantify the temporal stability for  $L_{temp}$ , we run the JOD metric [Mantiuk et al. 2021] with and without Ltemp in place, which yields superior results when using the proposed loss term (7.10 with the temporal loss vs 7.04 without), where a higher number suggests better quality. Fig. 9 provides a qualitative comparison on the design choice of using 3 input driving views on Subject 1's data. We observe that our design choice of using an image-triplet to encode the facial expressions leads to overall better results. Fig. 10 shows a qualitative comparison for different baselines. They include no initialization for Gaussian scales, our implementation of RGCA decoders [Saito et al. 2024a], combined decoders for Gaussian opacity and offset values, and ours. We remark that, unlike our method, RGCA utilizes combined decoders for Gaussian opacity and offset values. We observe that our results reproduce sharper details around the mouth interior than those obtained by the other approaches. Tab. 4 compares the same baselines against our approach on Subject S<sub>NeRSemble1</sub> and supports our qualitative results.

#### 4.3 Comparisons

We compare our real-time multi-view dynamic head generation method against three multi-view techniques: HQ3DAvatar [Teotia et al. 2024], GaussianAvatars [Qian et al. 2024], and Mixture of GaussianHeads: End-to-End Learning of Drivable Gaussian Head Avatars from Coarse-to-fine Representations • 9



Fig. 11. **Qualitative comparisons** with the related methods. *Top to bottom*: Subject 1, Subject 2, Subject S<sub>NeRSemble1</sub>, and Subject 3. *Left to right*: HQ3DAvatar [Teotia et al. 2024], GaussianAvatrars [Qian et al. 2024], MVP [Lombardi et al. 2021], and Ours. Note that yellow boxes highlight the detail reproduction with respect to the ground truth. Our method produces crisper details in the mouth interior and hair.



Fig. 12. **Qualitative comparison on a speech sequence**. Our method results in overall better mouth-region synthesis.

Table 5. **Quantitative comparison with related approaches**. We outperform the related methods numerically on several image-quality metrics (see **bold text**).

Metrics	HQ3DAvatar	GaussianAvatars	MVP	Ours
PSNR ↑	30.78	32.22	32.50	32.89
L1 ↓	13.68	8.90	9.03	7.33
SSIM ↑	0.77	0.79	0.78	0.81
LPIPS $\downarrow$	0.14	0.15	0.15	0.11

Volumetric Primitives (MVP) [Lombardi et al. 2021]. Notably, GaussianAvatars [Qian et al. 2024], MVP [Lombardi et al. 2021], and our method operate in real time. HQ3DAvatar employs an implicitly learned canonical radiance field, while GaussianAvatars registers 3D Gaussians on a tracked mesh. MVP uses ray-marching through cubic volumetric primitives on a template mesh. MVP [Lombardi et al. 2021] and our approach can work with RGB images as input, whereas Gaussian Avatars [Qian et al. 2024] is designed to work with a parametric face model. Fig. 11 presents qualitative reconstruction comparisons. HQ3DAvatar [Teotia et al. 2024], GaussianAvatars [Oian et al. 2024], and MVP [Lombardi et al. 2021] struggle with topological changes, fine-scale hair details, dental features, and beard textures, often producing blurry and less defined results. In contrast, our method captures these details more accurately and sharply, as evidenced by the sharper representation of the tongue (top row), fine hair-strand structure (second row), clear dental features (third row), and distinct beard textures (bottom row). Table 5 shows the quantitative performance of our approach against MVP [Lombardi et al. 2021] and HQ3DAvatar [Teotia et al. 2024] averaged over holdout frames for 4 subjects. We observe that for the image-quality assessment metrics, our method shows the best numerical performance.

Fig. 12 shows a qualitative comparison of our method against the different baselines for a speech sequence trained on a subject speaking phonetically balanced sentences. We observe that our approach outperforms baseline methods in terms of image synthesis quality, especially for the mouth region.

# 5 LIMITATIONS AND FUTURE WORK



Our Rendering

Ground Truth

Fig. 13. Failure Case. Left to right: Our rendering and Ground Truth. Our model produces ghosting artifacts for subtle movements in the eye region.

While our method clearly sets the state of the art for human head avatar rendering in terms of speed and photorealism, it still has certain limitations. First, the image encoder takes the entire face region as input, which prevents independent animation of local face regions, such as the cheeks. It also struggles with subtle movements of the eyes, as highlighted in Fig. 13. Future approaches could integrate localized animation models for fine-grained control. Second, the choice of sampling could be made more efficient. Future directions could involve sampling a low number of Gaussians and learning their initial positioning on the deformed template surface rather than uniformly sampling a dense number of Gaussians. Third, we rely on a multi-view camera rig during training. In the future, we plan to investigate learning photoreal head avatars using more lightweight setups. For driving the avatar from outside of the multiview capture setup, an augmentation strategy similar to [Xiang et al. 2023] could help reduce the domain gap between the capture and driving environments by color-augmenting the driving images. At the same time, future research should investigate robustness to lighting changes or even truthful disentanglement of lighting and material properties. Additionally, audio-based appearance synthesis is also a promising future direction that our current work can enable.

#### 6 CONCLUSION

We have presented a novel end-to-end framework to render highquality moving human heads with large deformations at real-time speed. By representing the head model in a coarse-to-fine manner, our method first deforms a template mesh based on an animation code extracted from input images. This has been shown to be effective in capturing coarse-level facial deformations, such as wide mouth opening and large head pose changes. We then sample the 3D Gaussians, which are parameterized by the animation code, on the deformed mesh via 2D UV space to further refine the deformation and the geometry into detailed facial expressions, such as lip rolling and fine-grained teeth shape. Utilizing a 3D splatting rasterization technique enables not only accelerating end-to-end training into 24 hours, but also real-time inference in up to 75 FPS.

With the global transformations as a learnable parameter, our method does not require an accurate, explicit global rigid pose, as opposed to competing radiance field rendering approaches. We demonstrate that the above significantly helps our model outperform state-of-the-art methods visually and numerically using various metrics. We also present a comprehensive ablative analysis of the proposed loss terms and component-level design decisions. Finally, we show the robustness of our method with some challenging scenes with highly dynamic head poses and deformations.

#### REFERENCES

- ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. 2022. RigNeRF: Fully Controllable Neural 3D Portraits. In *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, 20332–20341.
- Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. 2018. Recycle-GAN: Unsupervised Video Retargeting. In ECCV.
- Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhöfer, Shunsuke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason M. Saragih. 2022. Authentic volumetric avatars from a phone scan. ACM Trans. Graph. 41, 4 (2022), 163:1–163:19.
- 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, Tero Karras, and Gordon Wetzstein. 2022. Efficient Geometry-aware 3D Generative Adversarial Networks. In *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, 16102–16112.

- Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from in-the-wild images. ACM Trans. Graph. 40, 4 (2021), 88:1–88:13.
- Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance Fields without Neural Networks. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE, 5491–5500.
- Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. 2024. COLMAP-Free 3D Gaussian Splatting. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE.
- Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE, 8649–8658.
- Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. 2022. Reconstructing Personalized Semantic Facial NeRF Models from Monocular Video. ACM Trans. Graph. 41, 6 (2022), 200:1–200:12.
- Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. HeadNeRF: A Realtime NeRF-based Parametric Head Model. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE, 20342–20352.
- Yujiao Jiang, Qingmin Liao, Xiaoyu Li, Li Ma, Qi Zhang, Chaopeng Zhang, Zongqing Lu, and Ying Shan. 2024. UV Gaussians: Joint Learning of Mesh Deformation and Gaussian Textures for Human Avatar Modeling. arXiv preprint arXiv:2403.11589 (2024).
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graph. 42, 4, Article 139 (jul 2023), 14 pages. https://doi.org/10.1145/3592433
- Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep video portraits. ACM Trans. Graph. 37, 4 (2018), 163.
- Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In International Conference on Learning Representations (ICLR). San Diega, CA, USA.
- Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. 2023. NeRSemble: Multi-View Radiance Field Reconstruction of Human Heads. ACM Trans. Graph. 42, 4, Article 161 (jul 2023), 14 pages. https://doi.org/10.1145/3592455
- Yushi Lan, Feitong Tan, Di Qiu, Qiangeng Xu, Kyle Genova, Zeng Huang, Sean Fanello, Rohit Pandey, Thomas A. Funkhouser, Chen Change Loy, and Yinda Zhang. 2023. Gaussian3Diff: 3D Gaussian Diffusion for 3D Full Head Synthesis and Editing. *CoRR* abs/2312.03763 (2023).
- Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. ACM Trans. Graph. 36, 6 (2017), 194:1–194:17.
- Stephen Lombardi, Jason M. Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. ACM Trans. Graph. 37, 4 (2018), 68.
- Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. ACM Trans. Graph. 38, 4 (2019), 65:1–65:14.
- Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhöfer, Yaser Sheikh, and Jason M. Saragih. 2021. Mixture of volumetric primitives for efficient neural rendering. ACM Trans. Graph. 40, 4 (2021), 59:1–59:13.
- Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. 2024. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. In *3DV*.
- Shugao Ma, Tomas Simon, Jason M. Saragih, Dawei Wang, Yuecheng Li, Fernando De la Torre, and Yaser Sheikh. 2021. Pixel Codec Avatars. In *IEEE Conf. Comput. Vis. Pattern Recog.* Computer Vision Foundation / IEEE Computer Society, 64–73.
- Shengjie Ma, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2024. 3D Gaussian Blendshapes for Head Avatar Animation. In SIGGRAPH. ACM.
- Rafał K. Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. 2021. FovVideoVDP: a visible difference predictor for wide field-of-view video. ACM Trans. Graph. 40, 4, Article 49 (jul 2021), 19 pages. https://doi.org/10.1145/3450626.3459831
- Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2022. NeRF: representing scenes as neural radiance fields for view synthesis. Commun. ACM 65, 1 (2022), 99–106.
- Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. 41, 4 (2022), 102:1–102:15.
- Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. 2023. ASH: Animatable Gaussian Splats for Efficient and Photoreal Human Rendering. (2023). arXiv:2312.05941 [cs.CV]
- Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. 2024. ASH: Animatable Gaussian Splats for Efficient and Photoreal Human Rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 1165–1175.
- Jeong Joon Park, Peter Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *IEEE Conf. Comput. Vis. Pattern Recog.* Computer Vision Foundation / IEEE Computer Society, 165–174.

#### 12 • Kartik Teotia, Hyeongwoo Kim, Pablo Garrido, Marc Habermann, Mohamed Elgharib, and Christian Theobalt

- Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2021a. Nerfies: Deformable Neural Radiance Fields. In Int. Conf. Comput. Vis. IEEE, 5845-5854.
- Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. 2021b. HyperNeRF: a higher-dimensional representation for topologically varying neural radiance fields. ACM Trans. Graph. 40, 6 (2021), 238:1-238:12.
- Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. 2024. GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE.
- Alfredo Rivero, ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. 2024. Rig3DGS: Creating Controllable Portraits from Casual Monocular Videos. CoRR abs/2402.03723 (2024).
- Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. 2024a. Relightable Gaussian Codec Avatars. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE.
- Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. 2024b. Relightable Gaussian Codec Avatars. In CVPR.
- Soshi Shimada, Vladislav Golyanik, Patrick Pérez, and Christian Theobalt. 2023. Decaf: Monocular Deformation Capture for Face and Hand Interactions. ACM Transactions on Graphics (TOG) 42, 6, Article 264 (dec 2023).
- Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In Conference on Neural Information Processing Systems (NeurIPS).
- Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]
- Kartik Teotia, Mallikarjun B R, Xingang Pan, Hyeongwoo Kim, Pablo Garrido, Mohamed Elgharib, and Christian Theobalt. 2024. HQ3DAvatar: High-quality Implicit 3D Head Avatar. ACM Trans. Graph. 43, 3, Article 27 (apr 2024), 24 pages. https: //doi.org/10.1145/3649889
- Avush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. 2018. Self-Supervised Multi-Level Face Model Learning for Monocular Reconstruction at Over 250 Hz. In IEEE Conf. Comput. Vis. Pattern Recog. Computer Vision Foundation / IEEE Computer Society, 2549-2559. Justus Thies, Michael Zollhöfer, and Matthias Nießner, 2019. Deferred neural rendering:
- image synthesis using neural textures. ACM Trans. Graph. 38, 4 (2019), 66:1-66:12.
- J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. 2016. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In Proc. Computer Vision and Pattern Recognition (CVPR), IEEE.
- Luan Tran, Feng Liu, and Xiaoming Liu. 2019. Towards High-Fidelity Nonlinear 3D Face Morphable Model. In IEEE Conf. Comput. Vis. Pattern Recog. Computer Vision Foundation / IEEE Computer Society, 1126-1135.
- Alex Trevithick, Matthew Chan, Michael Stengel, Eric R. Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. 2023. Real-Time Radiance Fields for Single-Image Portrait View Synthesis. In ACM Transactions on Graphics (SIGGRAPH).
- Cong Wang, Di Kang, Yan-Pei Cao, Linchao Bao, Ying Shan, and Song-Hai Zhang. 2023a. Neural Point-Based Volumetric Avatar: Surface-Guided Neural Points for Efficient and Photorealistic Volumetric Head Avatar. In SIGGRAPH Asia 2023 Conference Papers (<conf-loc>, <city>Sydney</city>, <state>NSW</state>, <country>Australia</country>, </conf-loc>) (SA '23). Association for Computing Machinery, New York, NY, USA, Article 50, 12 pages. https://doi.org/10.1145/3610548. 3618204
- Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. 2023b. StyleAvatar: Real-time Photo-realistic Portrait Avatar from a Single Video. In ACM SIGGRAPH 2023 Conference Proceedings.
- Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhöfer. 2021. Learning Compositional Radiance Fields of Dynamic Human Heads. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 5700-5709. https://doi.org/10.1109/CVPR46437. 2021.00565
- Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004a. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13, 4 (2004), 600-612. https://doi.org/10.1109/TIP.2003.819861
- Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004b. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process, 13, 4 (2004), 600-612.
- Donglai Xiang, Fabian Prada, Zhe Cao, Kaiwen Guo, Chenglei Wu, Jessica Hodgins, and Timur Bagautdinov. 2023. Drivable Avatar Clothing: Faithful Full-Body Telepresence with Dynamic Clothing Driven by Sparse RGB-D Input. In SIGGRAPH Asia 2023 Conference Papers (Sydney, NSW, Australia) (SA '23). Association for Computing Machinery, New York, NY, USA, Article 24, 11 pages. https://doi.org/10.1145/ 3610548.3618136
- Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. 2022. Point-nerf: Point-based neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5438-5448.

- Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. 2024. Gaussian Head Avatar: Ultra High-fidelity Head Avatar via Dynamic Gaussians. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE
- Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. 2023a. AvatarMAV: Fast 3D Head Avatar Reconstruction Using Motion-Aware Neural Voxels. In ACM SIGGRAPH 2023 Conference Proceedings.
- Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Han Huang, Guojun Qi, and Yebin Liu. 2023b. LatentAvatar: Learning Latent Expression Code for Expressive Neural Head Avatar. In SIGGRAPH, Erik Brunvand, Alla Sheffer, and Michael Wimmer (Eds.). ACM, 86:1-86:10.
- Haotian Yang, Mingwu Zheng, Wanquan Feng, Haibin Huang, Yu-Kun Lai, Pengfei Wan, Zhongyuan Wang, and Chongyang Ma. 2023. Towards Practical Capture of High-Fidelity Relightable Avatars. In SIGGRAPH, June Kim, Ming C. Lin, and Bernd Bickel (Eds.). ACM, 23:1-23:11.
- Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In IEEE Conf. Comput. Vis. Pattern Recog. Computer Vision Foundation / IEEE Computer Society, 586-595.
- Xiaochen Zhao, Lizhen Wang, Jingxiang Sun, Hongwen Zhang, Jinli Suo, and Yebin Liu. 2024. HAvatar: High-fidelity Head Avatar via Facial Model Conditioned Neural Radiance Field. ACM Trans. Graph. 43, 1 (2024), 6:1-6:16.
- Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. 2022. I M Avatar: Implicit Morphable Head Avatars from Videos. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE, 13535-13545.
- Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. 2023. PointAvatar: Deformable Point-Based Head Avatars from Videos. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE, 21057-21067.
- Zhenglin Zhou, Fan Ma, Hehe Fan, and Yi Yang. 2024. HeadStudio: Text to Animatable Head Avatars with 3D Gaussian Splatting. CoRR abs/2402.06149 (2024).
- Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2022. Instant Volumetric Head Avatars. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022), 4574-4584. https://api.semanticscholar.org/CorpusID:253761096
- M. Zwicker, H. Pfister, J. van Baar, and M. Gross. 2002. EWA Splatting. IEEE Transactions on Visualization and Computer Graphics 8, 3 (07/2002-09/2002 2002), 223-238.

#### IMPLEMENTATION DETAILS



(a) FLAME Shape Fitting

Fig. 14. Left to right: We deform the initial FLAME shape mesh (a) to fit the photometric data, resulting in a deformed template mesh used for training our model (b). (c) Shows the captured identity.

We provide further details of our model implementation in the following section.

# A.1 Template registration

We deform the FLAME [Li et al. 2017] template to match the photometric data of multiview RGB images by using our method as a renderer. This step is highlighted in Fig. 14. We use the FLAME [Li et al. 2017] template as the base template mesh and learn per-vertex offsets for a static frame. In addition to the photometric loss terms used in our approach, we use a landmark-based loss term and a Laplacian-smoothness term to obtain a smooth template mesh. The details of our network architecture, such as network operators and feature dimensions, are illustrated in Fig. 16.



Fig. 15. Qualitative results: Choice of sampling resolution. We highlight that a sampling resolution of  $512 \times 512$  leads to comparative results to  $1024 \times 1024$ , i.e., the maximum sampling resolution. A  $256 \times 256$  sampling resolution leads to blurry details inside the mouth. Rendering with sampling at full  $1024 \times 1024$  is twice as slow as ours.

## A.2 Baselines for Comparisons

In Sec. 4.3, we compare our approach against HQ3DAvatar [Teotia et al. 2024], MVP [Lombardi et al. 2021], and GaussianAvatars [Qian et al. 2024]. As both MVP and HQ3DAvatar can take an image as

input, we provide both methods with the same multiview triplet input as ours. In addition, GaussianAvatars [Lombardi et al. 2021] and MVP [Lombardi et al. 2021] are provided with the same template mesh used by our approach. For comparison against RGCA [Saito et al. 2024a], we only implemented the decoding framework of RGCA [Saito et al. 2024a], which uses shared decoders for Gaussian geometry property prediction, tracked expression meshes for initializing the Gaussians, and no scale initialization.

# **B** EXPERIMENTS

Table 6. **Runtime comparison for various rendering components**. The table lists the components needed for rendering the final image and their associated computational times (in seconds). The total runtime is 0.0132 seconds or 75 FPS.

Component	Ey	$D_v$	D <sub>RGB</sub>	Do	$D_m$	Sampling	Rasterization
Time	0.0012	0.0002	0.0008	0.0008	0.0007	0.0001	0.0094

Table 6 shows the runtime breakdown of our method. The runtime is evaluated on the holdout frames of 3 subjects. Rasterization is the most expensive operation, which still runs over 100 fps. Fig. 15 shows a qualitative comparison that initializes 3D Gaussians with different UV map sampling resolutions. 14 • Kartik Teotia, Hyeongwoo Kim, Pablo Garrido, Marc Habermann, Mohamed Elgharib, and Christian Theobalt



Fig. 16. Architecture details of the different components employed in our approach.