GGAvatar: Dynamic Facial Geometric Adjustment for Gaussian Head Avatar

Xinyang Li[†1], Jiaxin Wang[†2], Yixin Xuan[1], Gongxin Yao[1], Yu Pan[‡1].

[1]Zhejiang University, Hangzhou, China
[2]Hangzhou Dianzi University, Hangzhou, China

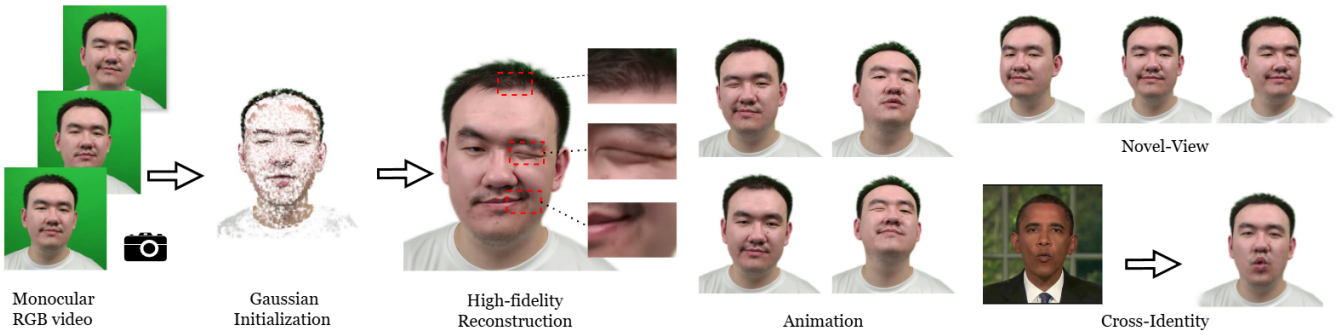

Figure 1: *Based on the Geometry Morph Adjuster, our method can learn a high-quality head avatar from an RGB video. The proposed GGAvatar method demonstrates outstanding performance in self-reconstruction, novel-view synthesis, and cross-identity reenactment tasks.*

Abstract

Reconstructing animatable 3D head avatars from target subject videos has long been a significant challenge and a hot topic in computer graphics. This paper proposes GGAvatar, a novel 3D avatar representation designed to robustly model dynamic head avatars with complex identities and deformations. GGAvatar employs a coarse-to-fine structure, featuring two core modules: a Neutral Gaussian Initialization Module and a Geometry Morph Adjuster. The Neutral Gaussian Initialization Module pairs Gaussian primitives with deformable triangular meshes, using an adaptive density control strategy to model the geometric structure of the target subject with neutral expressions. The Geometry Morph Adjuster introduces deformation bases for each Gaussian in global space, creating fine-grained low-dimensional representations of deformations to overcome the limitations of the Linear Blend Skinning formula. Extensive experiments show that GGAvatar can produce high-fidelity renderings, outperforming state-of-the-art methods in visual quality and quantitative metrics.

CCS Concepts

• ***Computing methodologies*** → ***Reconstruction; Animation; Shape modeling;***

1. Introduction

Creating high-fidelity digital avatars with real-time interaction accessible to everyone is a crucial building block for the metaverse and various applications, such as immersive telepresence and augmented or virtual reality. Although achieving photorealistic digital avatars has been a research focus in computer vision and graphics for decades, generalizing these avatars to unseen poses or expressions with low cost remains an ongoing challenge.

Given a personalized video, most prior works have primarily employed 3D Morphable Model (3DMM) techniques [LBB*17, PKA*09] or have leveraged neural implicit representations [MST*21, PFS*19] to develop animatable 3D head avatars. The former methods [GPL*22, KSLZ22, KGT*18] employ a fixed topological structure linked to a standard rasterization pipeline, en-

† Authors contributed equally to this work.
‡ Corresponding author.
This research is supported by the National Natural Science Foundation of China under Grants No. U22A20102.

DOI: 10.2312/pg.20241313

abling them to generalize to unseen deformations. However, this fixed topology lacks the structural flexibility for accessories such as hats and glasses, thereby restricting its ability to accurately model individuals with complex identities. Works based on neural implicit representations [BKY*22, HPX*22, ZAB*22] leverage the sampling of multiple points along camera rays to capture accessories. However, these approaches significantly reduce their training and rendering efficiency.

3D Gaussians Splatting (3D-GS) [KKLD23] has recently proven to be more efficient than Neural Radiance Fields (NeRF) [MST*21] in new perspective synthesis and 3D reconstruction. It represents 3D scenes explicitly by introducing discrete 3D Gaussian primitives and adapts seamlessly to the rasterization pipeline, providing real-time rendering performance. Motivated by this progress, there have been significant research efforts [CWL*23,ZBL*24,QKS*23, XGGZ23] aimed at extending these innovations to the development of 3D digital avatars. Despite substantial advancements, two major challenges persist with current Gaussian-based methods: 1) Given their discrete nature, how can one properly initialize the geometry to accelerate the convergence of training? Flashavatar [XGGZ23] addresses this by transforming the 3D head into UV space for sampling, which achieves high-fidelity digital heads with high FPS. Due to its heavy reliance on the FLAME model [LBB*17], this sampling strategy faces challenges in accurately depicting features such as long hair and detailed facial features. 2) How can a deformation strategy be designed to generalize 3D avatars to unseen poses and expressions effectively? An intuitive approach employs parameterized 3DMMs of human heads as drivers. These models provide several orthogonal bases within their parameter space, which allow for manipulating attributes such as identity, pose, and expression in the 3D model via linear blend skinning (LBS). Some approaches [QKS*23, SWL*24] leverage a strategy of anchoring Gaussians to a 3D mesh, which enables localized learning of intrinsic features and dynamically adjusts their global properties, such as position, scale, and rotation, based on changes in the 3D model's topology. However, this strategy relies on the geometric consistency of multiple views and cannot be extended to fine-grained expressions and poses through direct linear transformations due to the limitations of the LBS formula, thereby restricting the animation capabilities of 3D avatars.

In this paper, we introduce a novel 3D avatar representation called GGAvatar, designed to robustly model dynamic head avatars with complex identities and deformations. GGAvatar consists of two core modules: a Neutral Gaussian Initialization module (defined in this article as global Gaussians without facial parameter input) and a Geometry Morph Adjuster. We initialize a 3D avatar with a neutral expression and construct a motion offset field to integrate high-frequency dynamic details and head movements. For the Neutral Gaussian Initialization module, we bind Gaussians to the FLAME mesh like the GaussianAvatar [QKS*23] approach and employ the densification strategy of 3D-GS [KKLD23] to initialize the 3D avatar with neutral expression input, which aims to focus on low-frequency deformation to avoid local optima. This initialization technique swiftly enriches areas of the avatar that lack detail with Gaussian primitives and provides a strong geometric prior for the Geometry Morph Adjuster, as shown in Figure 2.

For deformation purposes, we harness bound Gaussians that track the coarse movements of the meshes to ensure stability throughout the animation process. However, coarse deformations of the meshes cannot capture the motion of fine non-surface regions, such as wrinkles and hair. To address the limitations, we propose the Geometry Morph Adjuster. The Geometry Morph Adjuster employs a parameterized multi-resolution tri-plane, which stores the spatial information of neutral Gaussians and connects to a finely tuned Multi-Layer Perceptron (MLP) that learns the deformation bases for each Gaussian. This setup facilitates the creation of a low-dimensional representation of deformations for each Gaussian. Integrating this with pre-retrieved facial parameters allows us to predict further positional adjustments and covariance shifts, which boosts the realism and expressiveness of the avatar. By adopting this strategy, we mitigate the unpredictability of directly forecasting deformations and surpass existing works in terms of rendering novel views and reenactments from a driving video.

The contributions of our method can be summarized as:

- We propose GGAvatar, a novel representation that employs a coarse-to-fine architecture to model ultra-high-fidelity human head avatars.
- To capture low-frequency deformations, the Neutral Gaussian Initialization module employs anchoring and densification strategies to effectively model the geometric structure of the neutral Gaussian.
- To capture high-frequency dynamic details, the Geometry Morph Adjuster utilizes explicit parameterized tri-plane structures connected to a finely tuned MLP that learns deformation bases, enabling the accurate modeling of extremely complex and fine-grained facial deformations.
- Extensive testing on public datasets demonstrates that our method outperforms contemporary alternatives in visual quality and quantitative metrics.

2. Related Work

2.1. Head Avatar Reconstruction with Implicit Models

NeRF encapsulates the radiance field of a scene within a neural network, enabling photorealistic renderings of novel views through volumetric rendering techniques. [GTZN21, GSKH21, WBL*21, WGYZ21] directly manipulates NeRF using facial model parameters such as expressions and poses to create an animable head avatar. However, these methods struggle to disentangle poses and expressions effectively, and they also face challenges in generalizing to novel poses and expressions. An alternative strategy [KCG*23, YSZ23] employs the "canonical + deformation" framework to construct a head model in a standard space and generate dynamics via deformation fields. IMAvatar [ZAB*22] employs a signed distance function to represent the implicit head model and uses learnable expression blend shapes to depict the deformation field. [COBG23] provides more precise dynamic interaction by associating key points with their radiation fields. Additionally, various methods [AXS*22, BKY*22, ASS23, GTZN21, LXW*22] leverage advanced techniques to enhance training speed and rendering quality, such as TriPlane [ZWS*23], KPlane [FKMW*23], and deformable multi-layer meshes [DWS*23]. Nonetheless, these

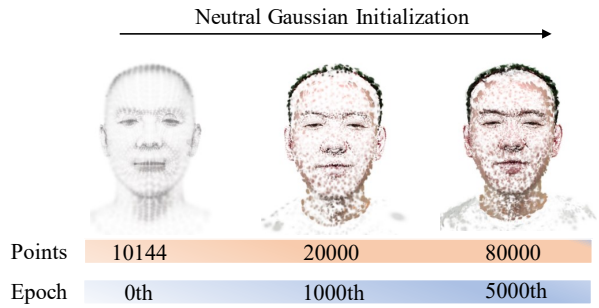

Figure 2: *The initialization process for a neutral expression (e.g., ID1) uses a densification strategy to add Gaussian primitives to non-head regions, accelerating training convergence. This method shows that even without corresponding neutral expression images, we can reconstruct the neutral Gaussian geometry using the binding Gaussian strategy.*

techniques still grapple with challenges related to efficiency in training and rendering.

2.2. Head Avatar Reconstruction with Explicit Models

3DMM [BV99] initially projects the 3D head shape into several low-dimensional Principal Component Analysis (PCA) spaces. Subsequently, numerous works have adopted 3DMM and its variants [CWZ*13, FFBB21, PLA*09, GMFB*18] to create head avatars. [FFBB21,TZN19,TET*20] employ feed-forward networks to predict vertex offsets and textures, enabling inference of unseen facial expressions. ROME [KSLZ22] encodes local photometric and geometric details to improve the quality of rendered images. 3DMM-based methods provide stable deformations but fail to model accessories such as eyeglasses. PointAvatar [ZYW*23] introduces a point-based geometric representation using differentiable point rendering, which overcomes the limitations of mesh-based models. However, it requires an excessive number of points and extensive training periods.

Recently, 3D-GS [KKLD23] has demonstrated exceptional performance, offering greater flexibility due to its anisotropic properties compared to point representations, and providing the efficiency of real-time rendering. Expanding upon PointAvatar, MonoGaussianAvatar [CWL*23] replaces Point Cloud with Gaussian primitives to improve the rendering speed and quality. PSAvatar [ZBL*24] further proposes a point-based morphable shape model to increase the flexibility of representation. Despite the impressive results achieved by these methods in rendering novel viewpoints and reenactments, the point-based initialization requiring redefinition of Gaussian features adds complexity to the training process. GaussianHeadAvatar [XCL*24] utilizes the Signed Distance Function (SDF) to depict the implicit head model and introduces 3D Gaussians to achieve more precise facial textures. FlashAvatar [XGGZ23] samples in UV space and attaches Gaussians to mesh with learnable offsets, which are represented as MLPs, enabling high-speed rendering of human head avatars. However, this initialization strategy distributes Gaussians across the FLAME model, limiting its ability to model features such as long hair and shoulders effectively. GaussianAvatars [CWL*23] associates each mesh triangle with a 3D Gaussian, incorporating densification and pruning strategies for accurate geometry representation, and uses binding inheritance to ensure seamless animation control via the parametric model. However, deformations relying on the LBS formula fail to animate intricate non-surface structures like hair and wrinkles. Therefore, building on the GaussianAvatar binding and the 3D-GS densification strategy [KKLD23], we introduce a parametric three-plane hash structure to capture more detailed deformation offsets, enabling high-fidelity animation of head avatars.

3. Method

Figure 3 presents a schematic of GGAvatar. Given a monocular portrait video of a target subject, the objective is to reconstruct an animatable head avatar. The FLAME meshes [LBB*17] feature vertices at varying positions but share the same topology. Consequently, we pair 3D Gaussian primitives with triangles of the mesh, employing a densification strategy [KKLD23] to geometrically initialize the neutral Gaussian and facilitate coarse Gaussian deformations (see section 3.2). Additionally, we learn an extra deformation basis for each Gaussian to achieve fine deformations and enhance high-frequency details (see section 3.3).

3.1. Preliminary

3D-GS utilizes discrete Gaussian primitives for the geometry representation of static scenes. A Gaussian primitive is defined by a 3D covariance matrix $\mathbf{\Sigma}$ centered at point (mean) μ:

$$G(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}. \tag{1}$$

To ensure differentiable computation of the covariance matrix, [KKLD23] defines a parametric ellipsoid with a scaling matrix $\mathbf{S}$ and a rotation matrix $\mathbf{R}$, constructing the covariance matrix by:

$$\mathbf{\Sigma} = \mathbf{RSS}^T \mathbf{R}^T. \tag{2}$$

In particular, the scaling and rotation matrices are represented by a learnable scaling vector $\mathbf{s} \in \mathbb{R}^3$ and a learnable quaternion $\mathbf{r} \in \mathbb{R}^4$, respectively. Finally, the color of a pixel can be determined through the blending of overlapped Gaussians:

$$\mathbf{C} = \sum_{i \in N} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{3}$$

where $\mathbf{c}_i$ is the color of each point represented by the spherical harmonic function, and blending weight α_i is computed by the 2D projection of the 3D Gaussian multiplied by a per-point opacity o.

3.2. Neutral Gaussian Initialization

Given a set of tracked FLAME triangular meshes with vertices $\{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2\}$ and edges defined as $\{\mathbf{a}_{10} = \mathbf{v}_1 - \mathbf{v}_0$, $\mathbf{a}_{20} = \mathbf{v}_2 - \mathbf{v}_0$, and $\mathbf{a}_{21} = \mathbf{v}_2 - \mathbf{v}_1\}$, we initialize each Gaussian at the centroid of these meshes $\mathbf{M}$. Specifically, we establish that $\mathbf{M}$ is the coordinate origin for each Gaussian in its local space. Then, We establish that the rotation matrix $\mathbf{R}$ represents the orientation of the triangle

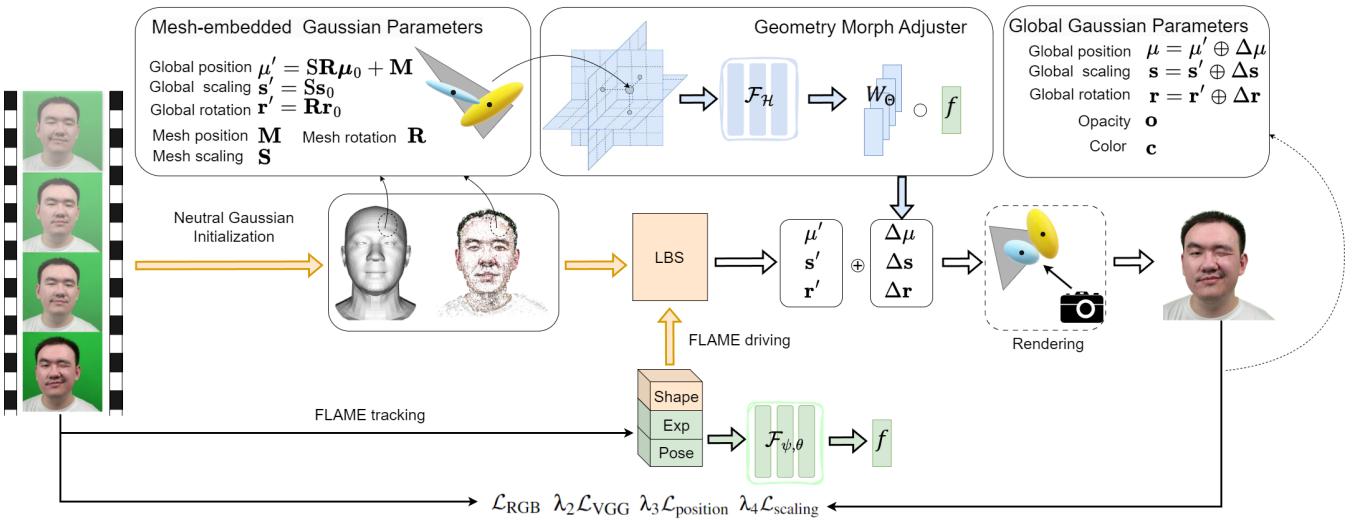

Figure 3: *Overview of GGAvatar. A Neutral Gaussian initialization strategy is proposed to model the geometry of neutral Gaussians. The neutral Gaussians are then coarsely deformed with FLAME mesh. To capture high-frequency dynamic details, we introduce the Geometry Morph Adjuster. To further enhance the representation capability of the deformation bases, we generate a latent vector from expression and pose parameters using MLPs. The deformed Gaussians are then splatted to render the image with a given pose.*

in global space, and the scaling S reflects the average size of the triangle as follows:

$$\mathbf{R} = [\mathbf{n}_0; \mathbf{n}_1; \mathbf{n}_2], \tag{4}$$

$$\mathrm{S} = (\|\mathbf{a}_{20}| + |\mathbf{n}_2 \cdot \mathbf{a}_{21}\|)/2, \tag{5}$$

where S is represented as the average length of one edge vector and its perpendicular distance, and $[\mathbf{n}_0; \mathbf{n}_1; \mathbf{n}_2]$ respectively represent the unit vector along $\mathbf{a}_{10}$, the unit normal vector of the mesh, and the cross product of $\mathbf{n}_0$ and $\mathbf{n}_1$.

According to section 3.1, we parameterize all its properties in local space, defined as $G = \{\mu_0, \mathbf{r}_0, \mathbf{s}_0, o, \mathbf{c}\}$. In the rough rendering stage, we convert these properties into the global space by:

$$\mathbf{r}' = \mathbf{R}\mathbf{r}_0, \tag{6}$$

$$\mu' = \mathbf{S}\mathbf{R}\mu_0 + \mathbf{M}, \tag{7}$$

$$\mathbf{s}' = \mathbf{S}\mathbf{s}_0. \tag{8}$$

The motion of head avatars can be divided into rigid movements related to head pose and non-rigid transformations associated with facial expressions. Head pose, including rotation and translation, can be manipulated through camera parameters. Based on this observation, we input zero expression parameters during the initialization phase to facilitate learning low-frequency geometric details by neutral Gaussians. By computing images of various expressions from the training data and using backpropagation to adjust the properties of neutral Gaussians, we can perform coarse modeling of the geometric structure of neutral Gaussians even without corresponding ground-truth inputs for neutral expressions.

To better capture the geometric shape of human head avatars, we employ an adaptive density control strategy [KKLD23], adding and removing splats based on the gradient of the view-space position and the opacity of each Gaussian. Then, we bind new Gaussians to the old ones during the optimization process to restore fidelity in local areas.

3.3. Geometry Morph Adjuster

Gaussians moving with the meshes is sufficient to capture coarse head movements; however, constrained by the direct linear representation of LBS, they are insufficient for capturing all the fine and intricate dynamic textures, such as hair and wrinkles. To enhance representational capacity while minimizing resource usage, we utilize a multi-resolution tri-plane to store high-frequency spatial information surrounding the head, denoted as $\mathcal{H}^3$. Specifically, we initially transfer the neutral Gaussians to global space and denote their positions as $\mathbf{x}$ with the tri-plane outputting $\mathcal{H}^3(\mathbf{x})$. Inspired by the dynamic representations of PointAvatar [ZYW*23], we learn an additional basis $\mathcal{W}_\Theta$ for each neutral Gaussian, aimed at achieving precise deformations for each Gaussian given conditional parameter inputs as follows:

$$\mathcal{W}_\Theta = \mathcal{F}_\mathcal{H}(\mathcal{H}^3(\mathbf{x})), \tag{9}$$

where $\mathcal{F}_\mathcal{H}$ is a basis prediction network represented as MLPs.

To reduce computational costs and enhance the accuracy of deformations, we employ another MLP $\mathcal{F}_{\psi,\theta}$ to reduce the dimensionality of the expression ψ and pose θ parameters, representing them as a latent vector f:

$$f = \mathcal{F}_{\psi,\theta}(\psi, \theta). \tag{10}$$

Figure 4: *Qualitative Comparisons with State-of-the-Art Methods. From top to bottom are ID1, ID2, ID3, ID4, and ID6. GGAvatar generates more realistic face reconstructions, especially in capturing high-frequency dynamic details and reconstructing extreme expressions.*

Additional deformations of the Gaussian can be represented as the matrix product of the base and the latent vector:

$$\Delta\mu, \Delta\mathbf{r}, \Delta\mathbf{s} = \mathcal{W}_\Theta \cdot f. \tag{11}$$

The final refined spatial features can be calculated as:

$$\mu, \mathbf{r}, \mathbf{s} = \left(\mu' \oplus \Delta\mu, \mathbf{r}' \oplus \Delta\mathbf{r}, \mathbf{s}' \oplus \Delta\mathbf{s}\right). \tag{12}$$

It is important to note that we address the limitations of mesh motion by transforming Gaussians into global space and learning a set of additional offsets to deform each Gaussian, as detailed in section 4.5.

3.4. Training Objectives

For the loss function, we utilize L1 loss $\mathcal{L}_1$ to monitor the pixel-wise difference between the ground truth and the rendered images. Additionally, we incorporate a D-SSIM term $\mathcal{L}_{\text{D-SSIM}}$ to further ensure structural similarity between the two images:

$$\mathcal{L}_{\text{RGB}} = (1-\lambda_1)\mathcal{L}_1 + \lambda_1\mathcal{L}_{\text{D-SSIM}}, \tag{13}$$

where λ_1 is taken as 0.2. Thanks to the powerful rendering technique [KKLD23] that can process all the images during every training step, allowing us to apply perceptual losses $\mathcal{L}_{\text{VGG}}$ [JAFF16] on the whole image.

Besides, our rendering quality partially relies on the proper alignment between Gaussian splats and triangles; otherwise, unnat-

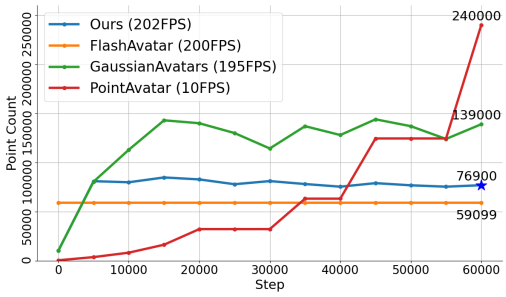

Figure 5: *The trend of point changes every 5000 steps and the final FPS in ID4 case. Our method combines speed and compactness.*

ural jitter and artifacts may occur during video synthesis. To alleviate this issue, we regularize the local position and scaling of each Gaussian to ensure they remain within reasonable limits:

$$\mathcal{L}_{\text{position}} = \left\| \max\left(\mu - \epsilon_{\text{position},0}\right) \right\|_2, \tag{14}$$

$$\mathcal{L}_{\text{scaling}} = \left\| \max\left(\mathbf{s} - \epsilon_{\text{scaling},0}\right) \right\|_2, \tag{15}$$

where $\epsilon_{\text{position}} = 1$ and $\epsilon_{\text{scaling}} = 0.6$ are the thresholds for the maximum allowable position and scaling, respectively. When below these thresholds, we disable their corresponding loss terms.

Our final loss function is thus:

$$\mathcal{L} = \mathcal{L}_{\text{RGB}} + \lambda_2 \mathcal{L}_{\text{VGG}} + \lambda_3 \mathcal{L}_{\text{position}} + \lambda_4 \mathcal{L}_{\text{scaling}}, \tag{16}$$

where $\lambda_2 = 0.02$, $\lambda_3 = 0.01$, and $\lambda_4 = 1$.

3.5. Implementation Details

We implement our network with Pytorch and use Adam for parameter optimization. We use the analysis-by-synthesis-based face tracker from MICA [ZBT22] for FLAME tracking. We conducted 120,000 training iterations for all target subjects and set all experiments to be performed on a single NVIDIA GTX 4090.

We set the learning rate to 5e-3 for the position and the scaling of 3D Gaussians and the remaining parameter learning rate is the same as vanilla 3D-GS [KKLD23]. Particularly, the learning rate for the positions decays exponentially, reaching 0.01 times the initial value by the 90,000 iterations. From the 500 iterations until the 60000 iterations, we activate the densification strategy with binding inheritance every 100 iterations and reset the opacity every 3,000 iterations.

After 5,000 iterations, we refine the Geometry Morph Adjuster using an Adam optimizer with $\beta = (0.9, 0.999)$ and add perceptual loss $\mathcal{L}_{\text{VGG}}$. The learning rates for both the basis network $\mathcal{F}_{\mathcal{H}}$ and the latent vector generation network $\mathcal{F}_{\psi,\theta}$ are set at 1e-4. The learning rate for the three-plane hash table is set at 0.005.

4. Experiments

4.1. Datasets

All our experiments are conducted on publicly available datasets released by previous work [GZX*22] to ensure fair comparisons, with all data resized to a resolution of 512x512 in advance. The training data consists of approximately 2,500 to 3,000 frames. Furthermore, we select each video's last 10% segment from the original dataset to serve as test data. The binary mask was created using RVM [LYSS22] and employed for foreground segmentation.

4.2. Baselines

For comparison purposes, we select three state-of-the-art head avatar reconstruction methods: 1) PointAvatar [ZYW*23], which utilizes explicit point clouds with upsampling to construct head geometry; 2) GaussianAvatars [QKS*23], which binds Gaussians to the FLAME mesh, incorporating a mesh-driven approach to guide the deformation of Gaussian; and 3) Flashavatar [XGGZ23], which employs Gaussian initialization based on the UV plane and learns additional deformations to achieve high-fidelity digital human driving. For PointAvatar, we follow the author's suggestions to train it on a 32GB V100 GPU and use 240,000 points for all subject data. For a fair comparison, GaussianAvatars is trained under the same experimental conditions as ours, with an additional perceptual loss applied for supervision. Additionally, we set the UV resolution of FlashAvatar to 256 and added the boundary of the FLAME mesh.

4.3. Qualitative and Quantitative Comparison in Reconstruction

Figure 4 shows the results of our visual comparison with the baseline rendering. PointAvatar reconstructs avatars using explicit point clouds and learns the LBS basis of each point to control point movement. However, this approach often fails to fit extreme expressions as illustrated in the first row. Furthermore, the primitives adopted are points with fixed shapes, inhibiting clear facial structure modeling. GaussianAvatars employs a standard LBS formula to control facial movements, which fails to model extreme expressions and dynamic high-frequency details. Evidence of these issues can be observed in the avatars generated in the 1st and 4th rows, depicting eye closure and high-frequency mouth details, respectively. Additionally, extracting facial and camera parameters from monocular video is more prone to inaccuracies compared to multi-view video.GaussianAvatars are highly sensitive to noise and minor variations, which often result in unnatural artifacts and erroneous deformations, as shown in the shoulder area in the 3rd row. FlashAvatar employs Gaussian primitives embedded at fixed positions within the FLAME mesh to initialize, which lacks adaptive control over the density of these primitives, resulting in visual artifacts such as visible cracks or seams, as seen in the mouth area in the 1st and the 2nd rows. This issue persists even when the head resolution is set to 256. Additionally, this sampling strategy faces challenges in accurately depicting features such as long hair and detailed facial features, as shown in the hair region in the 5th row. Compared to the methods above, our rendering images more closely approximate the ground truth, achieving more accurate deformations and capturing nearly all dynamic facial details. This improvement is primarily due to the Geometry Morph Adjuster, which can adaptively learn and accurately model the subtle deformations in the scene.

Quantitative result comparisons are recorded in Table 1. We calculate metrics including PSNR, SSIM, and LPIPS [ZIE*18] sepa-

Table 1: *Quantitative comparison with state-of-the-art methods. Red indicates the best and yellow indicates the second.*

Methods	PointAvatar			GaussianAvatars			FlashAvatar			Ours		
Metrics	PSNR↑	SSIM ↑	LPIPS ↓	PSNR↑	SSIM ↑	LPIPS ↓	PSNR↑	SSIM ↑	LPIPS ↓	PSNR↑	SSIM ↑	LPIPS ↓
ID 1	24.01	0.903	0.125	29.01	0.927	0.075	31.35	0.936	0.120	32.00	0.944	0.068
ID 2	21.38	0.886	0.135	20.52	0.911	0.098	23.48	0.920	0.125	24.45	0.930	0.079
ID 3	18.69	0.878	0.156	17.91	0.896	0.113	22.94	0.927	0.093	24.68	0.941	0.064
ID 4	22.56	0.917	0.148	23.41	0.945	0.064	25.40	0.958	0.067	26.24	0.965	0.060
ID 5	19.36	0.864	0.164	18.47	0.829	0.199	21.70	0.882	0.149	24.91	0.909	0.067
ID 6	21.61	0.852	0.149	22.28	0.902	0.105	25.91	0.919	0.099	26.31	0.922	0.068

Figure 6: *Novel view synthesis results of GGAvatar. We demonstrate multi-view geometric consistency across both x- and y-axis rotations.*

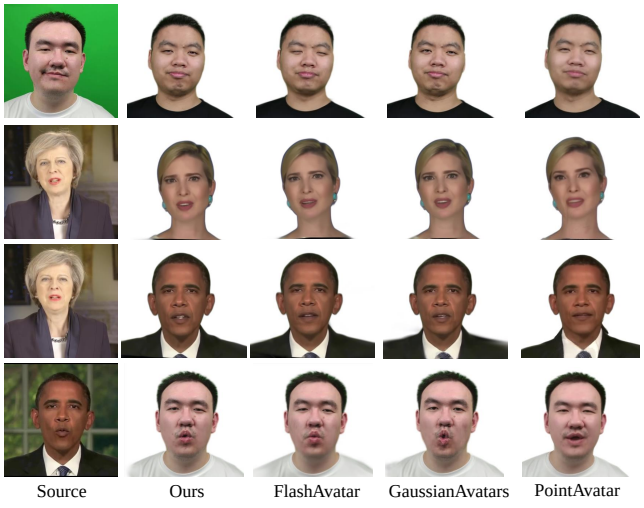

Figure 7: *Cross-Identity Reenactment results comparison. Our method achieves personalized expression driving and synthesizes more natural results.*

rately for each subject. It is seen that our approach outperforms others by a significant margin. At the same time, we show the results of points' numbers during the training process and the rendering FPS in Figure 5. Our method achieves better results than GaussianAvatars when using fewer points, further proving the effectiveness of the Geometry Morph Adjuster.

4.4. Novel View Synthesis and Cross-Identity Reenactment

To verify the robustness of the model, we designed a multi-view synthesis experiment for the GGAvatar model and a cross-identity reenactment experiment, comparing our results with the baseline methods.

The results of the multi-view synthesis experiment are shown in Figure 6. Our model can render images from different angles, producing images similar to the ground truth. Our model accurately synthesized new perspectives among hundreds of pictures in the test set. This demonstrates the model's ability to maintain geometric consistency and high-fidelity detail reproduction across various viewing angles.

The comparison results with the baseline in cross-identity reenactment are shown in Figure 7. Using Gaussian for expression parameter decoding, our model better replicates the original expressions onto new faces, capturing most details, including the corners of the eyes and mouth.

The baseline methods, such as FlashAvatar [XGGZ23], GaussianAvatars [QKS*23], and PointAvatar [ZYW*23], show limitations in accurately preserving these fine details. FlashAvatar tends to over-smooth facial features, losing critical expression nuances. GaussianAvatars, while preserving some facial contours, often distort subtle expression elements, leading to less natural results. PointAvatar, despite its geometric approach, struggles with maintaining expression fidelity across different identities.

Our model's ability to accurately transfer expressions while retaining fine details significantly enhances the realism of cross-

identity reenactment. This superior performance is attributed to our advanced encoding and decoding algorithm, which precisely maps the expression parameters onto related 3D Gaussian structures. The results demonstrate the robustness and versatility of our approach, making it a promising solution for applications requiring high-fidelity facial reenactment.

4.5. Ablation Study

To validate the effectiveness of our method components, we deactivate each of them and report results in Table 2. In addition, we perform a visual comparison in Figure 8. For the ablation of the loss term, please refer to the supplementary material.

4.5.1. Geometry Morph Adjuster

Without the Geometry Morph Adjuster, the 3D avatar animations rely only on the linear transformations defined by the LBS formula. Although LBS-based deformations can robustly model coarse geometric changes, they fail to accurately capture high-frequency facial details, as shown in the 1st row of Table 2, and the 3rd column of Figure 8. It is worth noting that without the Geometry Morph Adjuster, our approach differs slightly from GaussianAvatars. To reduce training time, we do not employ the strategy of refining FLAME parameters throughout GGAvatar's training process.

To further verify the effectiveness of the Geometry Morph Adjuster, we incorporated this adjuster into the vanilla GaussianAvatars method, called "GaussianAvatars+GMA". We conducted comparative experiments using subject 074 from the NeRSemble [KQG*23] dataset. Keeping all other experimental conditions the same as the vanilla GaussianAvatars method, we trained both models for 100,000 iterations. As shown in Table 3, quantitative metrics demonstrate that GaussianAvatars+GMA captures high-frequency motion details in both novel view synthesis and self-reenactment tasks, which is reflected in the significant reduction of the LPIPS metric. Visually, this is evidenced by clearer facial details such as beards and hair, as illustrated in Figure 9. As shown in Table 3 and Figure 9, the Geometry Morph Adjuster effectively captures more complex spatial features in the novel view synthesis task, reducing distortion and artifacts during viewpoint transitions, which leads to improved PSNR and SSIM metrics. However, in the self-reenactment task, the facial parameters extracted from multi-view images are relatively accurate. As a result, the Geometry Morph Adjuster relies on the topological changes of the LBS formula, leading to less noticeable improvements in PSNR and SSIM.

4.5.2. Parametric Tri-plane

In the Geometry Morph Adjuster, we utilize a parameterized tri-plane hash table to store the geometric information of the neutral Gaussians, where its multi-resolution grid encoding mechanism facilitates high-quality feature representation. For comparison, we replace the tri-plane with the positional encoding introduced by [MST*21], as shown in the 2nd row of Table 2.

4.5.3. Linear Blend Skinning

Learning complex facial deformations from scratch is challenging. We use the Geometry Morph Adjuster to improve fitting efficiency and enhance the LBS-based deformations rather than directly learning the complete deformations. This allows our Geometry Morph Adjuster to focus only on high-frequency dynamic details. To demonstrate the effectiveness of our strategy, we performed an ablation study by removing the LBS and using the Geometry Morph Adjuster only to predict deformations. As shown in the 3rd row of Table 2, removing the LBS formula negatively impacts image quality and deformation results. Visually, the removal of the LBS formula makes it difficult for the Geometry Morph Adjuster to learn complex deformations. This is evident in the 4th column of Figure 8, where issues such as incomplete eye closures and incorrect mouth shapes are observed.

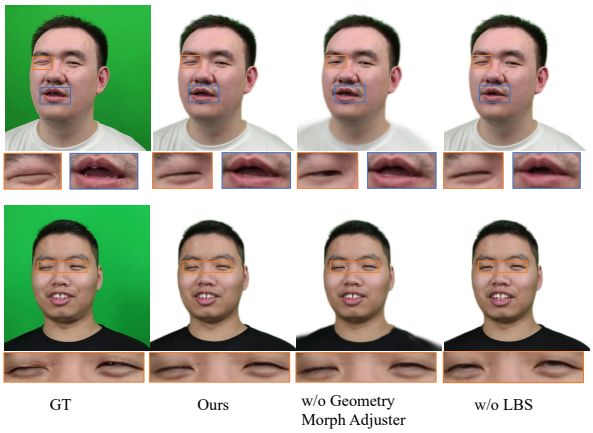

Figure 8: *Visual comparison results of various components in the paper. The results establish the important role each component plays in the overall performance of the system.*

Table 2: *Ablation study on ID 1 and ID 2. Red indicates the best and yellow indicates the second.*

Methods	PSNR↑	SSIM ↑	LPIPS ↓
w/o Geometry Morph Adjuster	25.81	0.925	0.097
w/o tri-plane	27.99	0.936	0.074
w/o LBS	27.41	0.933	0.077
w/o initialization	28.04	0.936	0.077
Ours	28.23	0.937	0.074

4.5.4. Neutral Gaussian Initialization

We apply the Neutral Gaussian Initialization strategy before training our Geometry Morph Adjuster to model neutral Gaussian geometric shapes and accelerate convergence. We conducted an ablation study by removing the Neutral Gaussian Initialization. Specifically, we trained the Geometry Morph Adjuster from the beginning without the initialization phase of the neutral Gaussians. As shown in the 4th row of Table 2, without the Neutral Gaussian Initialization phase, the Geometry Morph Adjuster struggles to capture fine-grained deformation details. Therefore, an effective Neutral Gaussian Initialization strategy is crucial for achieving higher deformation accuracy and fidelity.

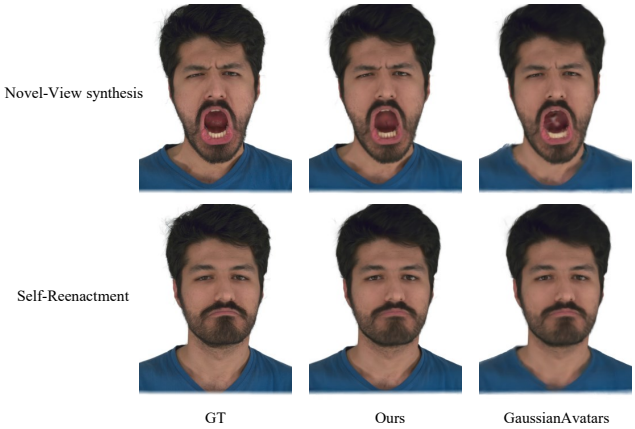

Figure 9: *Visual comparison results on subject #074.*

Table 3: *Quantitative comparison results on subject #074. Red indicates the best.*

	Novel-View		
Metrics	PSNR↑	SSIM ↑	LPIPS ↓
GaussianAvatars	27.47	0.905	0.144
GaussianAvatars+GMA	33.78	0.936	0.084
	Self-Reenactment		
Metrics	PSNR↑	SSIM ↑	LPIPS ↓
GaussianAvatars	26.64	0.899	0.140
GaussianAvatars+GMA	27.05	0.900	0.093

5. Limitations and Future Work

Despite the promising results, our model has several limitations. The reliance on the FLAME model means that the accuracy of FLAME parameter tracking is crucial. In this paper, we did not improve the tracker, resulting in some entanglement between identity and facial parameters. Furthermore, the tracker struggles with inaccuracies in camera parameters from monocular videos, causing jitter in the synthesized video.

In some extreme facial expression scenarios, the Gaussians transformed by LBS may cluster together, leading to a significant overlap of Gaussian features. This overlap can result in unexpected outcomes that cannot be adequately adjusted even with our Geometry Morph Adjuster. Additionally, our method cannot model dynamic appearance as the color remains consistent throughout the video, requiring further improvements to handle such input data. In future work, we aim to develop a more adaptive method to handle extreme scenarios and intricate details, ultimately achieving a more robust 3D digital human reconstruction technique.

6. Conclusion and Discussion

In this paper, we introduce GGAvatar, a coarse-to-fine framework that integrates the strengths of LBS formulation and deformable basis learning. Our approach consists of two core modules: the Neutral Gaussian Initialization module and the Geometry Morph Adjuster. The Neutral Gaussian Initialization module quickly enhances avatar details using Gaussian primitives. At the same time, the Geometry Morph Adjuster employs a parameterized multi-resolution tri-plane and a finely tuned MLP to adaptively predict and optimize modeling and deformation areas that linear LBS cannot accurately control. It is worth noting that our method, to some extent, overcomes the limitations of FLAME by using only its changes based on topology (LBS) as a prior for the deformation process.

Extensive testing on public datasets demonstrates that GGAvatar outperforms existing methods in visual quality and quantitative metrics. Our method ensures realistic and expressive avatar reenactment by maintaining high fidelity in dynamic head movements and complex expressions.

References

[ASS23] ATHAR S., SHU Z., SAMARAS D.: Flame-in-nerf: Neural control of radiance fields for free view face animation. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)* (2023), IEEE, pp. 1–8. 2

[AXS*22] ATHAR S., XU Z., SUNKAVALLI K., SHECHTMAN E., SHU Z.: Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition* (2022), pp. 20364–20373. 2

[BKY*22] BERGMAN A., KELLNHOFER P., YIFAN W., CHAN E., LINDELL D., WETZSTEIN G.: Generative neural articulated radiance fields. *Advances in Neural Information Processing Systems 35* (2022), 19900–19916. 2

[BV99] BLANZ V., VETTER T.: A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques - SIGGRAPH '99* (Jan 1999). URL: http://dx.doi.org/10.1145/311535.311556, doi:10.1145/311535.311556. 3

[COBG23] CHEN C., O'TOOLE M., BHARAJ G., GARRIDO P.: Implicit neural head synthesis via controllable local deformation fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 416–426. 2

[CWL*23] CHEN Y., WANG L., LI Q., XIAO H., ZHANG S., YAO H., LIU Y.: Monogaussianavatar: Monocular gaussian point-based head avatar. *arXiv preprint arXiv:2312.04558* (2023). 2, 3

[CWZ*13] CAO C., WENG Y., ZHOU S., TONG Y., ZHOU K.: Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics 20*, 3 (2013), 413–425. 3

[DWS*23] DUAN H.-B., WANG M., SHI J.-C., CHEN X.-C., CAO Y.-P.: Bakedavatar: Baking neural fields for real-time head avatar synthesis. *ACM Transactions on Graphics (TOG) 42*, 6 (2023), 1–17. 2

[FFBB21] FENG Y., FENG H., BLACK M. J., BOLKART T.: Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG) 40*, 4 (2021), 1–13. 3

[FKMW*23] FRIDOVICH-KEIL S., MEANTI G., WARBURG F. R., RECHT B., KANAZAWA A.: K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 12479–12488. 2

[GMFB*18] GERIG T., MOREL-FORSTER A., BLUMER C., EGGER B., LUTHI M., SCHÖNBORN S., VETTER T.: Morphable face models-an open framework. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)* (2018), IEEE, pp. 75–82. 3

[GPL*22] GRASSAL P.-W., PRINZLER M., LEISTNER T., ROTHER C., NIESSNER M., THIES J.: Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18653–18664. 1

[GSKH21] GAO C., SARAF A., KOPF J., HUANG J.-B.: Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 5712–5721. 2

[GTZN21] GAFNI G., THIES J., ZOLLHOFER M., NIESSNER M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 8649–8658. 2

[GZX*22] GAO X., ZHONG C., XIANG J., HONG Y., GUO Y., ZHANG J.: Reconstructing personalized semantic facial nerf models from monocular video. *ACM Transactions on Graphics (TOG) 41*, 6 (2022), 1–12. 6

[HPX*22] HONG Y., PENG B., XIAO H., LIU L., ZHANG J.: Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 20374–20384. 2

[JAFF16] JOHNSON J., ALAHI A., FEI-FEI L.: Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14* (2016), Springer, pp. 694–711. 5

[KCG*23] KOCABAS M., CHANG J.-H. R., GABRIEL J., TUZEL O., RANJAN A.: Hugs: Human gaussian splats. *arXiv preprint arXiv:2311.17910* (2023). 2

[KGT*18] KIM H., GARRIDO P., TEWARI A., XU W., THIES J., NIESSNER M., PÉREZ P., RICHARDT C., ZOLLHÖFER M., THEOBALT C.: Deep video portraits. *ACM transactions on graphics (TOG) 37*, 4 (2018), 1–14. 1

[KKLD23] KERBL B., KOPANAS G., LEIMKÜHLER T., DRETTAKIS G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics 42*, 4 (2023), 1–14. 2, 3, 4, 5, 6

[KQG*23] KIRSCHSTEIN T., QIAN S., GIEBENHAIN S., WALTER T., NIESSNER M.: Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG) 42*, 4 (2023), 1–14. 8

[KSLZ22] KHAKHULIN T., SKLYAROVA V., LEMPITSKY V., ZAKHAROV E.: Realistic one-shot mesh-based head avatars. In *European Conference on Computer Vision* (2022), Springer, pp. 345–362. 1, 3

[LBB*17] LI T., BOLKART T., BLACK M. J., LI H., ROMERO J.: Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph. 36*, 6 (2017), 194–1. 1, 2, 3

[LXW*22] LIU X., XU Y., WU Q., ZHOU H., WU W., ZHOU B.: Semantic-aware implicit neural audio-driven video portrait generation. In *European conference on computer vision* (2022), Springer, pp. 106–125. 2

[LYSS22] LIN S., YANG L., SALEEMI I., SENGUPTA S.: Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2022), pp. 238–247. 6

[MST*21] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM 65*, 1 (2021), 99–106. 1, 2, 8

[PFS*19] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 165–174. 1

[PKA*09] PAYSAN P., KNOTHE R., AMBERG B., ROMDHANI S., VETTER T.: A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance* (2009), Ieee, pp. 296–301. 1

[PLA*09] PAYSAN P., LÜTHI M., ALBRECHT T., LERCH A., AMBERG B., SANTINI F., VETTER T.: Face reconstruction from skull shapes and physical attributes. In *Pattern Recognition: 31st DAGM Symposium, Jena, Germany, September 9-11, 2009. Proceedings 31* (2009), Springer, pp. 232–241. 3

[QKS*23] QIAN S., KIRSCHSTEIN T., SCHONEVELD L., DAVOLI D., GIEBENHAIN S., NIESSNER M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. *arXiv preprint arXiv:2312.02069* (2023). 2, 6, 7

[SWL*24] SHAO Z., WANG Z., LI Z., WANG D., LIN X., ZHANG Y., FAN M., WANG Z.: Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. *arXiv preprint arXiv:2403.05087* (2024). 2

[TET*20] THIES J., ELGHARIB M., TEWARI A., THEOBALT C., NIESSNER M.: Neural voice puppetry: Audio-driven facial reenactment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16* (2020), Springer, pp. 716–731. 3

[TZN19] THIES J., ZOLLHÖFER M., NIESSNER M.: Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG) 38*, 4 (2019), 1–12. 3

[WBL*21] WANG Z., BAGAUTDINOV T., LOMBARDI S., SIMON T., SARAGIH J., HODGINS J., ZOLLHOFER M.: Learning compositional radiance fields of dynamic human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 5704–5713. 2

[WGYZ21] WANG X., GUO Y., YANG Z., ZHANG J.: Prior-guided multi-view 3d head reconstruction. *IEEE Transactions on Multimedia 24* (2021), 4028–4040. 2

[XCL*24] XU Y., CHEN B., LI Z., ZHANG H., WANG L., ZHENG Z., LIU Y.: Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 1931–1941. 3

[XGGZ23] XIANG J., GAO X., GUO Y., ZHANG J.: Flashavatar: High-fidelity digital avatar rendering at 300fps. *arXiv preprint arXiv:2312.02214* (2023). 2, 3, 6, 7

[YSZ23] YE K., SHAO T., ZHOU K.: Animatable 3d gaussians for high-fidelity synthesis of human motions. *arXiv preprint arXiv:2311.13404* (2023). 2

[ZAB*22] ZHENG Y., ABREVAYA V. F., BÜHLER M. C., CHEN X., BLACK M. J., HILLIGES O.: Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 13545–13555. 2

[ZBL*24] ZHAO Z., BAO Z., LI Q., QIU G., LIU K.: Psavatar: A point-based morphable shape model for real-time head avatar creation with 3d gaussian splatting. *arXiv preprint arXiv:2401.12900* (2024). 2, 3

[ZBT22] ZIELONKA W., BOLKART T., THIES J.: Towards metrical reconstruction of human faces. In *European Conference on Computer Vision* (2022), Springer, pp. 250–269. 6

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Jun 2018). URL: http://dx.doi.org/10.1109/cvpr.2018.00068, doi:10.1109/cvpr.2018.00068. 6

[ZWS*23] ZHAO X., WANG L., SUN J., ZHANG H., SUO J., LIU Y.: Havatar: High-fidelity head avatar via facial model conditioned neural radiance field. *ACM Transactions on Graphics 43*, 1 (2023), 1–16. 2

[ZYW*23] ZHENG Y., YIFAN W., WETZSTEIN G., BLACK M. J., HILLIGES O.: Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2023), pp. 21057–21067. 3, 4, 6, 7