DoubleGaussianAvatar: Double Gaussians for Driveable Head Avatars with Dynamic Facial Details

Fangtian Liang
Shandong University
Jinan, China
lft00@foxmail.com

Guangshun Wei
Shandong University
Jinan, China
guangshunwei@gmail.com

Pengfei Wang*
Shandong University
Jinan, China
pfwang@sdu.edu.cn

Zheng Bi
TravelSky Technology Ltd.
Beijing, China
bizheng@travelsky.com.cn

Yuanfeng Zhou*
Shandong University
Jinan, China
yfzhou@sdu.edu.cn

Abstract

With its real-time rendering performance, 3D Gaussian Splatting has established itself as a leading approach for creating animatable head avatars. However, existing methods struggle to faithfully reconstruct fine-grained facial details, posing a significant challenge in high-fidelity avatar generation. These details can be broadly categorized into static—those that remain consistent across expression variations—and dynamic—those that vary with facial movements and are inherently more challenging to model. In this paper, we propose DoubleGaussianAvatar, a novel dynamic 3D representation framework based on 3D Gaussian Splatting, designed to capture both static and dynamic facial details while retaining real-time performance. In specific, DoubleGaussianAvatar consists of two sets of 3D Gaussians: coarse Gaussians and dense Gaussians. Coarse Gaussians are assigned to the coarse human head mesh for maintaining the stability of facial expressions when rendering a new state of the human head, while dense Gaussians are bound to dense mesh for capturing static facial details. Furthermore, we propose a learnable expression-dependent Gaussian offset that dynamically adapts to facial movements, enabling accurate reconstruction of dynamic details. During avatar reconstruction, the two Gaussian sets are jointly optimized under our carefully designed regularization constraints, working synergistically to render highly realistic head avatars. Our experiments demonstrate that the proposed method achieves state-of-the-art performance in capturing fine-grained facial details while preserving real-time rendering capabilities.

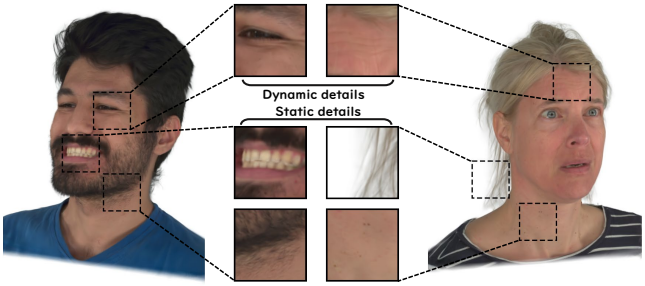

Figure 1. The results of DoubleGaussianAvatar, a novel method for creating photorealistic head avatars from multi-view videos that enables the capture of fine-grained facial details. The static details remain consistent across expressions, but the dynamic details vary with facial movements and are inherently more challenging to model.

Keywords: Parametric face models, headavatar, facial animation, facial reenactment

1. Introduction

Reconstructing photorealistic dynamic head avatars is a fundamental research topic in computer graphics and vision, enabling diverse applications in gaming, movie production, immersive telepresence, and augmented or virtual reality (AR/VR). This task involves two primary challenges: (1) achieving high-fidelity human avatars, and (2) ensuring real-time performance. Furthermore, achieving expression-controllable dynamic avatars from arbitrary viewpoints presents significant challenges, particularly in preserving fine facial details during extreme or micro-expressions.

For 3D avatar reconstruction, 3D Morphable Models (3DMM) [1, 30] represent a classical approach that en-

*Joint corresponding authors.

ables convenient facial expression control through low-dimensional PCA coefficients. However, this method suffers from limited rendering realism. To address this, some works have integrated Neural Radiance Fields (NeRF) [28] with 3DMM. While these NeRF-based 3DMM approaches [14, 51, 47] improve visual quality, their reliance on implicit neural representations and dense volumetric sampling results in computationally intensive rendering, preventing real-time performance, thus limiting the scope of their application. In contrast, 3D Gaussian Splatting (3DGS) [18] achieves both real-time rendering and high visual fidelity by optimizing explicit 3D Gaussian primitives in volumetric space. This advantage has motivated recent efforts to combine 3DGS with 3DMM, aiming to develop animatable head avatars that are simultaneously fast, realistic, and controllable.

Existing Gaussian-3DMM methods face a trade-off between reconstruction fidelity and real-time performance. While some methods [41] employ MLP-based implicit texture modeling with subsequent super-resolution to enhance detail reconstruction, this approach sacrifices real-time rendering capability. Alternative solutions [31, 32, 38] achieve animatable avatars by rigging 3D Gaussians to parametric face models, enabling control through expression transfer or manual parameter adjustment. However, these methods are inherently constrained by the limited resolution of parametric models, failing to capture subtle micro-expressions or extreme facial deformations (e.g., fine wrinkles and skin folds). Additionally, certain implementations [49] simplify the task by excluding non-facial regions (e.g., hair and neck), but this artificial constraint compromises the realism of avatar rendering.

In response to the above challenges, we introduce a novel Double-Gaussians representation that simultaneously achieves: (1) high-fidelity controllable head avatars with fine-grained facial details, as shown in Figure 1, and (2) real-time rendering performance from arbitrary viewpoints. The Double-Gaussians consists of coarse Gaussians and dense Gaussians. The dense Gaussians are anchored to a high-resolution triangular mesh, facilitating the accurate reconstruction of fine-grained static facial details. To enhance dynamic facial details that evolve with expressions, we propose a Gaussian Expression Blendshape strategy, this strategy transforms the dense Gaussians into FLAME coordinate space while applying an expression-dependent offset. To address rendering instability arising from the growing number of Gaussians and optimizable parameters, we introduce an additional set of coarse Gaussians bound to low-resolution FLAME meshes. To further enhance stability, we propose an auxiliary rotation control mechanism leveraging spherical linear interpolation (SLerp), which regularizes the orientation of dense Gaussians following expression transformation. Our method significantly outperforms existing real-time approaches in capturing fine-grained dynamic facial details.

Our contributions are as follows:

- We propose DoubleGaussianAvatar, a novel method for creating head avatars that can be driven from arbitrary viewpoints in real-time while preserving facial details.
- We combine coarse and dense Gaussians for representing dynamic head avatars, and design a joint optimization strategy that maintains rendering controllability.
- We introduce an expression-dependent offset based on expression blendshapes, with carefully designing regularization constraints, leveraging expression coefficients to capture detailed facial features.

2. Related Work

2.1. Human Face/Head Reconstruction

The 3D Morphable Model [1] and subsequent works [30, 12, 3, 43, 37] represent facial features as weighted combinations of basis vectors. Such parametric models can flexibly capture diverse face geometries and appearances by adjusting parameters for shape, expression, pose, texture, and more [7]. FLAME [22] models the entire human head, breaking through the limitation of traditional 3DMM models whose expressive range is confined to the facial region and enabling the reconstruction of detailed head avatars.

Conventional face reconstruction methods [2, 4, 21, 50] focus on obtaining more accurate model parameters from images and apply additional graphical techniques to optimize rendering. DECA [8] introduces a novel detail-consistency loss to regress 3D face shapes with animatable details that are unique to an individual and vary with expression. HRN [50] proposes a hierarchical approach, which applies deformation and displacement maps for shape optimization, coupled with an additional module to enhance the texture details.

NeRF [28] introduces a new paradigm for head reconstruction, with many methods [9, 35, 48, 40, 42, 17, 11] achieving a volumetric representation of the head by encoding head parameters along with additional facial details. HeadNeRF [14] employs NeRF as a 3D proxy to parametrically model human heads. INSTA [51] leverages Instant-NGP [29] to sample points from a voxel grid, then locates the nearest triangles on the FLAME mesh to deform query points into a normalized space. PointAvatar [47] investigates point-based representations through the use of differential point rendering. Specifically, it trains a deformation field that is conditioned on the the FLAME parameters, thereby transforming the point set in the canonical space.

3D Gaussian Splatting (3DGS) [18] has substantially advanced neural rendering by achieving real-time perfor-

mance without compromising visual fidelity. Most existing approaches [41, 36, 5, 13] employ neural deformation fields to accurately map dynamic Gaussians to their canonical space coordinates. Nevertheless, the neural network architecture continues to impose significant computational overhead, limiting rendering performance.

GaussianBlendShape [27] learns a set of Gaussians representing classical expressions to serve as a basis for blending new expressions, allowing high-frequency details to be captured through linear blending with expression coefficients. SplattingAvatar [32] disentangles the motion and appearance of a virtual human with explicit mesh geometry and implicit appearance modeling with Gaussian Splatting. GaussianAvatars [31] uses multi-view head sequences to achieve a driveable head viewable from any angle, with a binding inheritance strategy that ensures consistency and stability in multi-view rendering. TensorialAvatars [38] introduces an expressive and compact representation that encodes texture-related attributes of the 3D Gaussians in the tensorial format, this representation effectively compresses the storage of Gaussian models, but simply using opacity offset to represent dynamic appearance lacks robustness. Gaussian-Head-Avatar [41] employs MLPs to implicitly model dynamic textures, followed by super-resolution techniques to enhance details rendering, but at the cost of not being able to render in real-time. Gaussian-Eigen-Models [49] adopts an ensemble of linear eigenbases to represent head appearance, achieving a balance between computational cost and generation quality. However, this approach needs to delete areas outside the head, such as garments and neck, while simplifying the difficulty of this task, resulting in the loss of the realism of head avatar rendering.

2.2. Dynamic Gaussian Splatting

A common approach to modeling dynamic scenes is to store them in a 4D coordinate with a compressed time dimension [45, 6, 23], or simply in discrete time steps [26, 33]. These approaches can quickly render realistic scenes from arbitrary viewpoints, faithfully reproducing the stored dynamic content without altering it. Some approaches [39, 44, 25, 15] use deformation fields to control the motion of the scene but face challenges with direct object-level manipulation.

Another effective paradigm is to use appropriate geometric proxies to precisely control the animation of the Gaussians of the scene. Li et al. [24] learn a parametric template from input videos and then parameterize this template onto two canonical Gaussian maps (front and back). The learned template adapts to the garments being worn, enabling the modeling of looser clothing, such as dresses. Jiang et al.[16] introduce physics-based rules to model realistic in-scene interactions. Gao et al.[10] distribute the Gaussians over an editable mesh and propose additional strategies, such as Face Split and Normal Guidance, to refine the geometric proxies, enabling more flexible deformation and control.

3. Proposed Method

3.1. Overview

The input to our method is a multi-view video recording of a human head with different poses and expressions, as shown in Figure 2. At each time step, we use a photometric head tracker based on [34] to fit FLAME parameters, utilizing multi-view observations and known camera parameters. The FLAME meshes share the same topology, but their vertices vary in position. Based on this, we create two sets of 3D Gaussians(Section Double Gaussian splatting), bound to the original FLAME meshes and the dense binding subregions divided by the original mesh, respectively. The two sets of 3D Gaussians are rendered together into images using a tile-based differentiable rasterization renderer [18]. The training process supervises rendered images against ground-truth references to learn photorealistic head avatars (Section Joint Optimization).

As with static scenes, the two sets of Gaussians are jointly optimized, with gradients fed back separately. Both sets of Gaussians use the adaptive density control operations [18] to densify and prune Gaussian splats for optimal quality. To maintain the connection between triangles and splats, we use a binding inheritance strategy [31], ensuring that new Gaussian points remain bound to the FLAME mesh. For the coarse Gaussians, we employ both color supervision and a regularization term on the scales of Gaussian splats [31] to maintain rendering quality under novel expressions and viewing poses. For the dense Gaussians, we impose two critical constraints: (1) position offsets modulated by expression coefficients, and (2) orthogonality regularization of the expression basis. This coupled formulation enables faithful reconstruction of fine-grained facial details during during extreme and micro-expressions.

3.2. Preliminary

3DGS. Given images and camera parameters, 3D Gaussian Splatting [18] provides a solution for reconstructing a static scene using anisotropic 3D Gaussians. Specifically, the scene is represented by a set of Gaussian splats, each defined by a covariance matrix Σ centered at the point (mean) $\boldsymbol{\mu}$:

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

The semi-definite covariance matrix Σ has actual physical meaning and can be represented as:

$$\Sigma = RSS^T R^T, \quad (2)$$

where R is a rotation matrix and S is a diagonal scaling matrix, representing the orientation and spatial scale of each Gaussian primitive respectively.

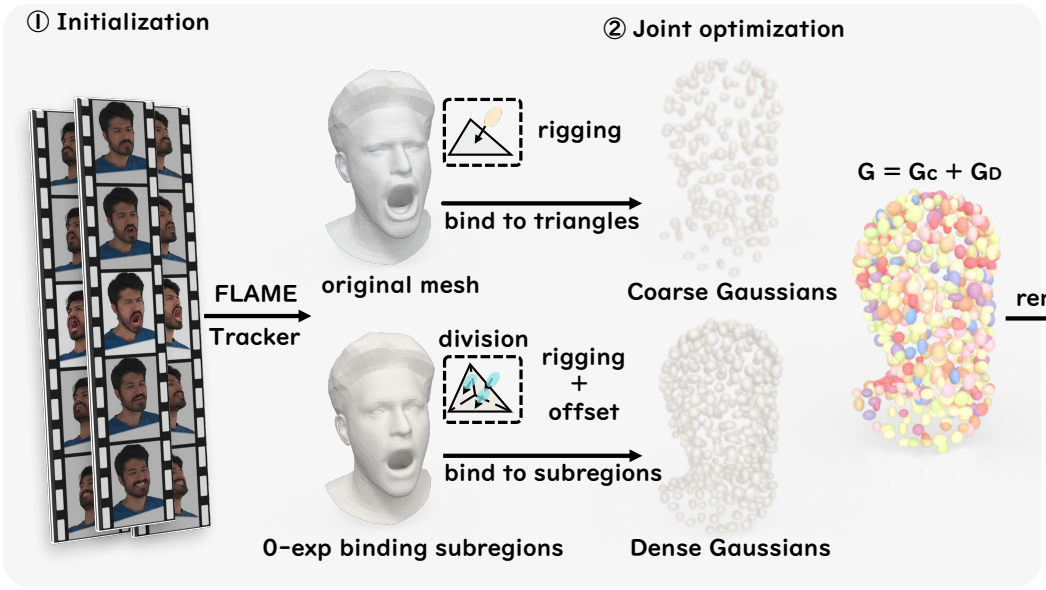

Figure 2. **Overview.** The input to our method is a multi-view video recording of a human head with varying poses and expressions. We create two 3D Gaussian sets, bound to the original FLAME mesh and the dense binding subregions divided by the original mesh, respectively. These two sets of 3D Gaussians are rendered together into images using a tile-based differentiable rasterization renderer [18]. The rendered images are then supervised by the ground-truth images to train the model, enabling the learning of a photorealistic human head avatar.

For rendering, the color C of a pixel is computed by blending all 3D Gaussians overlapping the pixel:

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

where c_i is the color of each point, modeled using 3rd-degree spherical harmonics. The blending weight α' is given by evaluating the 2D projection of the 3D Gaussian, multiplied by a per-point opacity α. The Gaussian splats are sorted by depth before blending to ensure correct visibility order.

Dynamic Avatar. To extend 3DGS from static scene representation to dynamic avatar modeling, GaussianAvatars [31] establish a connection between the head mesh (obtained by fitting FLAME parameters to multi-view observations) and 3DGS. The core mechanism involves associating each 3D Gaussian splat with a corresponding mesh triangle, where geometric attributes (position $\boldsymbol{\mu}$, rotation $\boldsymbol{r}$, anisotropic scaling $\boldsymbol{s}$) are defined within the triangle's local coordinate. After the optimization, these attributes remain fixed, creating a persistent relative positioning between each splat and its parent triangle while enabling global motion through the triangle's deformation. The local coordinate system for each triangle is constructed as follows:

- The origin is placed at the centroid T, computed as the mean position of the three vertices.
- The rotation matrix R is derived from an orthonormal basis consisting of the normalized direction vector of a selected edge and the triangle's normal vector.
- A scale factor k is calculated as the average length of both a primary edge and its corresponding perpendicular edge.

For each paired 3D Gaussian of a triangle, this framework enables the transformation of Gaussian attributes from local to global space by:

$$\boldsymbol{r}' = \boldsymbol{R}\boldsymbol{r}, \tag{4}$$
$$\boldsymbol{\mu}' = k\boldsymbol{R}\boldsymbol{\mu} + \boldsymbol{T}, \tag{5}$$
$$\boldsymbol{s}' = k\boldsymbol{s}. \tag{6}$$

where $\boldsymbol{T}$ is centroid position, $\boldsymbol{R}$ is orientation, and $\boldsymbol{k}$ is the approximate scale of the parent triangle.

3.3. Double Gaussian splatting

There are two main factors that hinder GaussianAvatars [31] from capturing dynamically generated facial details. On one hand, the FLAME model demonstrates limited accuracy in head geometry representation from images due to its fundamental reliance on a global PCA framework and linear transformation mechanics. This architecture produces facial expressions through linear combinations of PCA basis vectors and coefficients via dot product operations, resulting in inherent constraints characteristic of

linear parametric models. On the other hand, the Gaussian rigging strategy, which solely applies affine transformations from an inaccurate mesh to the Gaussians, fails to capture the precise geometry of real images and constrains the Gaussians toward a neutral expression. Consequently, dynamically generated facial details, such as wrinkles, are lost.

To address the first issue, we introduce a dense division mesh representation where individual triangular elements serve as geometric proxies to guide the 3D Gaussian model, aiming to decouple the coarse mesh from its dual role in governing both geometric constraints and rendering properties of the 3D Gaussians, thereby enabling specialized optimization of each attribute space. To tackle the second issue, we apply an expression-dependent position offset to the affine transformation from the inaccurate mesh to the Gaussians, effectively mitigating the negative impact of imprecise geometry.

The proposed strategies introduce additional optimizable parameters, which can compromise rendering stability. To mitigate this issue, we introduce a secondary set of coarse Gaussians anchored to the original FLAME mesh, ensuring robust during facial expression rendering. Extensive ablation studies confirm the effectiveness of our approach.

To this end, we employ a Double-Gaussians representation, combining both coarse and dense Gaussian components for dynamic avatar modeling. The coarse Gaussians inherit from GaussianAvatars [31]. Building on this, we additionally introduce a dense Gaussians. Specifically, we densify the original FLAME mesh with N vertices and F faces by dividing each triangle into three binding subregions based on the centroid, as shown in Figure 2. Notably, this approach yields a denser, uniformly sampled mesh with F_d subregions on the head without altering the geometry of the original FLAME mesh where $F_d = 3F$. Next, we bind the Gaussians to the denser mesh with F_d subregions. For each triangular subregion on the dense mesh, we assign a scalar k, defined by the mean of one edge's length and its perpendicular. We use the centroid in FLAME coordinate space $\boldsymbol{T}$ as the target binding location and design a subregion-level learnable position offset basis $\mathcal{E}_d = [\mathbf{E}_d^1, \cdots, \mathbf{E}_d^{|\vec{\psi}|}] \in \mathbb{R}^{3F_d \times |\vec{\psi}|}$ given expression coefficients $\vec{\psi}$:

$$offset(\vec{\psi}; \mathcal{E}_d) = \sum_{n=1}^{|\vec{\psi}|} \mathbf{E}_d^n \vec{\psi}_n \tag{7}$$

describes the additional expression-dependent position offsets for each 3D Gaussian in a binding subregion.

Since we have divided the dense binding regions, the rotation $\boldsymbol{R}$ of the original mesh face can only roughly reflect the directional properties of each subregion. To achieve more precise Gaussian rotation while maintaining controllability, we designed a learnable interpolation coefficient $\boldsymbol{t} \in \mathbb{R}^{F_d}$ for all binding subregions. The rotation of the subregions is represented by spherical linear interpolation (SLerp) of quaternions between the identity matrix and the rotation matrix $\boldsymbol{R}$ of their corresponding mesh faces(the faces on the original mesh that partition them out) under a specific expression. The rotation of the subregions is defined as:

$$\boldsymbol{R}' = Slerp(\boldsymbol{I}, \boldsymbol{R}, \boldsymbol{t}). \tag{8}$$

Dense Gaussians with location $\boldsymbol{\mu}$, rotation $\boldsymbol{r}$, and anisotropic scaling $\boldsymbol{s}$ are transformed into FLAME coordinate space through an affine transformation that includes rotation, scaling and translation, with an added expression-dependent position offset based on the expression blendshapes:

$$\boldsymbol{r}' = \boldsymbol{R}'\boldsymbol{r}, \tag{9}$$

$$\boldsymbol{\mu}' = k\boldsymbol{\mu} + \boldsymbol{T}^0 + offset(\vec{\psi}; \mathcal{E}_d), \tag{10}$$

$$\boldsymbol{s}' = k\boldsymbol{s}. \tag{11}$$

We align Gaussians to the FLAME coordinate space using triangle centroids from the zero-expression state $\boldsymbol{T}^0$, rather than centroids under a certain expression, since we utilize expression coefficients as an indicator of the expression state. The learnable positional offset basis $\mathcal{E}_d \in \mathbb{R}^{3F_d \times |\vec{\psi}|}$ is initialized as the mean of the vertex expression basis $\mathcal{E} \in \mathbb{R}^{3N \times |\vec{\psi}|}$ from FLAME[22], since each triangular subregion corresponds to three original FLAME mesh vertices, while rotation interpolation coefficients $\boldsymbol{t}$ are set to 1 (yielding $\boldsymbol{R}' = \boldsymbol{R}$). During initialization, Gaussians are positioned within their parent triangular regions by the sum of $\boldsymbol{T}^0$ and $offset$. Following optimization, $\boldsymbol{T}^0$ maintains the canonical zero-expression positions, while the learned offset enables expression-adaptive fine-tuning of Gaussian locations.

3.4. Joint optimization

The Double-Gaussians model consists of two sets of Gaussians, defined as $\boldsymbol{G_C}$ and $\boldsymbol{G_D}$ respectively. Before being sent to the differentiable rasterizer, we transform the geometric properties of each set, which are then concatenated into an overall Gaussian model $\boldsymbol{G}$ for rendering:

$$\boldsymbol{G} = \boldsymbol{G_C} + \boldsymbol{G_D} \tag{12}$$

We supervise the rendered images with a combination of $\mathcal{L}_1$ term and a D-SSIM term following [18]:

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

with $\lambda = 0.2$.

To constrain $\boldsymbol{G_C}$, we use a regularization loss [31] to prevent the Gaussian splats from excessive shrinkage:

$$\mathcal{L}_{\text{scaling}} = \| \max(\boldsymbol{s}, \epsilon_{\text{scaling}}) \|_2, \tag{14}$$

with threshold $\epsilon_{\text{scaling}} = 0.6$.

Contrary to GaussianAvatars [31], we have eliminated the position loss applied to the Gaussian offset relative to the triangle. Our main consideration is that introducing the position loss significantly restricts the freedom.

To constrain the transformation of $\boldsymbol{G_D}$, we propose an orthogonality loss of the learnable expression-dependent position offset basis:

$$\mathcal{L}_{\text{ortho}} = \|triu(\mathcal{E}_d^T \mathcal{E}_d)\|_2, \tag{15}$$

where the upper triangular region elements of the matrix $triu(\mathcal{E}_d^T \mathcal{E}_d)$ represent the dot product of all the matches of the expression basis vectors, which we use to constrain the position offset basis and maintain its orthogonality property.

Our final loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{scaling}} \mathcal{L}_{\text{scaling}} + \lambda_{\text{ortho}} \mathcal{L}_{\text{ortho}}, \tag{16}$$

where $\lambda_{\text{scaling}} = 1$ and $\lambda_{\text{ortho}} = 100$.

At the beginning of the training, which we refer to as the coarse stage, we disable the $offset$ and rotation control for $\boldsymbol{G_D}$, as well as $\mathcal{L}_{\text{ortho}}$. The main consideration is $\boldsymbol{G_D}$ perform a similar translation to $\boldsymbol{G_C}$, and $\boldsymbol{G_C}$ and $\boldsymbol{G_D}$ will together describe the appearance of the same region. Subsequently, once $\boldsymbol{G_D}$ stabilized, we synchronously add the position offset basis $\mathcal{E}_d$ and the interpolation coefficient $\boldsymbol{t}$ to the optimisation process and turn on $\mathcal{L}_{\text{ortho}}$. When the expression-related position offset is directly applied to the rendering primitives, it allows $\boldsymbol{G_D}$ to more closely match the shape of the head in a specific expression state. At this stage—termed the fine stage— $\boldsymbol{G_C}$ represents a more consistent head geometry, constraining $\boldsymbol{G_D}$ to produce a more geometrically sound rendered result, thus avoiding distortions when faced with extreme expressions or poses.

4. Experiments

4.1. Settings

Dataset. We use multi-view data derived from the NeRSemble dataset [20], which contains dynamic sequences of real human heads captured from 16 different viewpoints. Specifically, we select 9 subjects with 11 video sequences showcasing complex expressions, and all images are downsampled to a resolution of 802 × 550. For training, we reserve 9 of the 10 prescribed sequences and 15 out of the 16 available cameras, while one sequence is used for testing. The remaining free performance sequence is employed for cross-identity reenactment.

Metric and Tasks. We evaluate performance using PSNR, SSIM, and LPIPS [46] on the following multi-view tasks: 1) *Dynamic Novel-View Synthesis*: driving an avatar with head poses and expressions from training sequences, and rendering from a new viewpoint. 2) *Multi-View Self-Reenactment*: driving an avatar with unseen poses and expressions, and rendering from all 16 camera views.

Implementation Details. We use the Adam optimizer [19] for parameter optimization, with consistent hyperparameters applied across all subjects. The multi-view head tracker from GaussianAvatars [31] is used to obtain FLAME parameters for each time step. Each subject is trained for 120,000 iterations. The learning rates are set to 1e-6 for the learnable position offset basis and 1e-2 for the rotation interpolation coefficients, with fine stage starting at iteration 20,000.The face-dimensional position offset basis vectors are initialized by linearly interpolating FLAME's vertex-dimensional expression basis vectors. Adaptive density control with binding inheritance is enabled every 2,000 iterations, starting from iteration 4,000 until the end. All other settings are kept consistent with GaussianAvatars, with the same learning rate applied for both $\boldsymbol{G_C}$ and $\boldsymbol{G_D}$. All experiments are conducted on an NVIDIA RTX 4090 GPU with 24GB of VRAM.

4.2. Results and Comparisons

We highlight the superior performance of our approach in real-time multi-view rendering tasks. To demonstrate the advantages of our method, we conduct both qualitative and quantitative comparisons with three leading real-time mesh-driven head avatar creation methods: GaussianAvatars [31], SplattingAvatar [32] and TensorialAvatars [38].

Qualitative comparison. As shown in Figure 3 and Figure 4, our method produces photorealistic results in both novel-view synthesis and self-reenactment, surpassing existing approaches in visual fidelity and detail preservation. By leveraging the dense mesh as a geometric proxy, our approach faithfully reconstructs fine facial details such as beard density and the shape of skin moles. In contrast to GaussianAvatars, we introduce expression-dependent positional offsets that capture more dynamic facial deformations such as skin folds from facial squeezing. Moreover, our dense binding strategy and positional offsets enable the Gaussian representation to go beyond the template mesh, accurately modeling structures such as teeth, hair, and even parts of the collar area. On the other hand, SplattingAvatar lacks a fixed binding mechanism and simplifies spherical harmonics, leading to unstable rendering. Similarly, TensorialAvatars relies solely on opacity modulation for texture updates, which fails to reproduce dynamic wrinkles in regions such as the eye corners. Finally, cross-identity reenactment results in Figure 5 confirm that our method retains these advantages while avoiding visual artifacts.

Figure 3. Qualitative comparison on novel-view synthesis of head avatars. Our method allows for the representation of more fine-grained head details, such as the shape of teeth, hair and mole spots, etc.

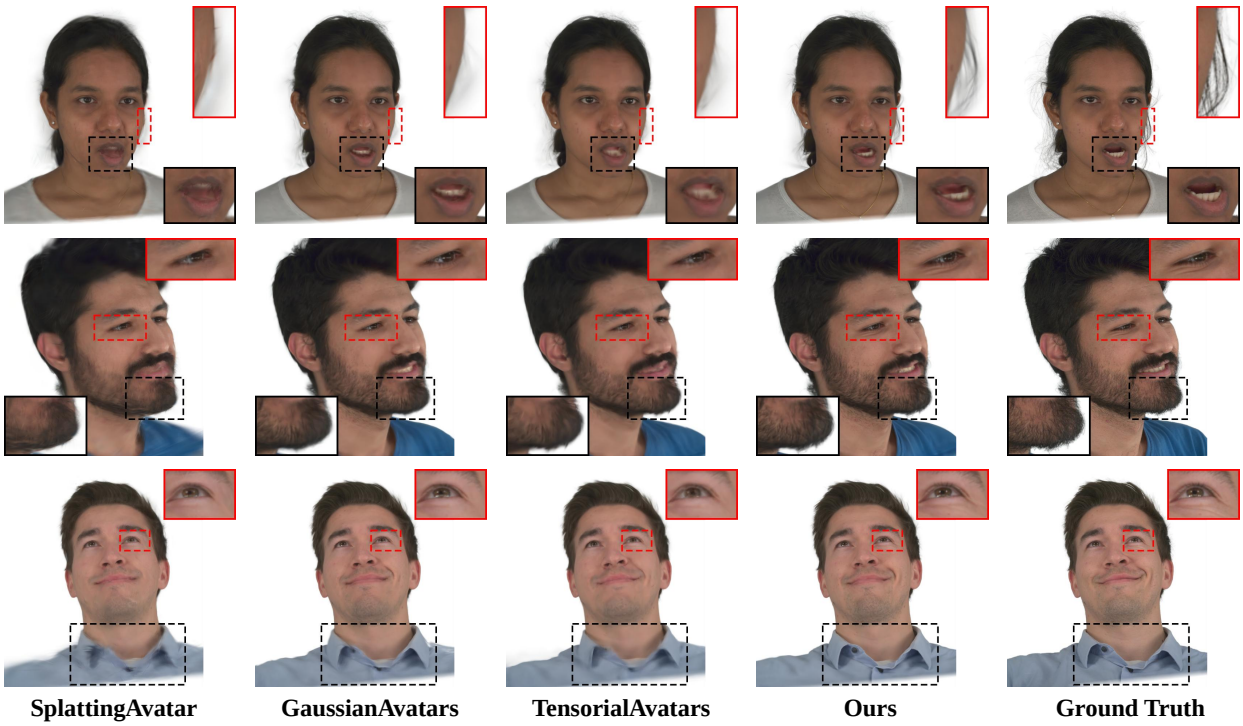

Figure 4. Qualitative comparison on self-reenactment of head avatars. Our method allows for the representation of more fine-grained head details, such as skin folds due to micro-expressions or extreme expressions, etc.

Figure 5. Qualitative comparison on cross-identity reenactment of head avatars. Our method restores the details of the head appearance while keeping the rendering stable and controllable.

	Novel-View Synthesis			Self-Reenactment			FPS↑
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
GaussianAvatars [31]	30.9	0.936	0.064	26.3	**0.915**	0.077	212
SplattingAvatar [32]	25.4	0.886	0.142	25.6	0.896	0.131	98
TensorialAvatars [38]	27.4	0.915	0.102	26.1	0.912	0.102	192
Ours (w/o coarse stage)	**34.8**	**0.955**	**0.041**	26.5	0.912	0.059	208
Ours (w coarse&fine stage)	34.7	**0.955**	**0.041**	**26.7**	0.914	**0.058**	

Table 1. Quantitative comparison with state-of-the-art real-time methods. Darker green shades indicate superior performance. **Bolded number** indicates the best.

Quantitative comparison. The quantitative evaluation, shown in Table 1, demonstrates a significant improvement over other methods. In the novel-view synthesis task, our method outperforms others across all metrics that assess image quality and expression accuracy, highlighting its superiority in reconstructing dynamic scenes. It is worth noting that our approach maintains rendering stability in the reenactment task, which can be attributed to the coarse Gaussians that strictly follow the original FLAME mesh. While GaussianAvatars maintains a more stable structure during self-reenactment and achieves slightly higher SSIM, our method outperforms it in terms of PSNR and LPIPS. In addition, the last two rows of Table 1 demonstrate that the warm-up stage is beneficial for the stability of optimization results, and it can yield a stable yet modest improvement in evaluation metrics, especially when rendering views under new expressions.

A comparison of rendering speeds is also provided in Table 1. Our method achieves full real-time performance at over 200 FPS without compromising rendering quality. Furthermore, as evidenced in Figure 6, our dense binding strategy enables superior rendering quality and expression fidelity while maintaining a comparable number of Gaussians after training.

4.3. Ablation Study

To evaluate the effectiveness of the proposed modules and loss functions, we deactivate each one individually and report both quantitative and qualitative results.

Joint optimization strategy. It is important to emphasize that although coarse Gaussians ultimately account for only a small proportion of all Gaussians (as shown in Figure 6), both sets of Gaussians in our method are indispensable components—and the joint optimization of these two sets of Gaussians is the key design that enables our method to maintain the stability of head shape when rendering facial details.

It is challenging to capture detail with just coarse Gaussians (like GaussianAvatars [31]). Similarly, without the coarse Gaussians to guide the joint optimization, the dense

Components						Novel View Synthesis			Self-Reenactment		
G_C	$offset$	Division	Slerp.R	$\mathcal{L}_{\text{ortho}}$	$\mathcal{L}_{\text{scaling}}$	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
✓	✓	**1-3**	✓	✓	✓	**37.4**	**0.980**	**0.018**	**31.1**	**0.959**	**0.025**
	✓	**1-3**	✓	✓	✓	37.3	0.979	0.020	27.9	0.947	0.037
✓		**1-3**	✓	✓	✓	33.0	0.966	0.036	**31.1**	0.958	0.036
✓	✓	**1-2**	✓	✓	✓	**37.4**	0.978	0.020	30.8	0.956	0.028
✓	✓	**1-4**	✓	✓	✓	37.2	0.979	0.020	29.1	0.949	0.035
✓	✓	**1-3**		✓	✓	37.3	0.979	**0.018**	30.9	0.957	0.026
✓	✓	**1-3**	✓		✓	37.1	0.979	0.019	30.5	0.953	0.029
✓	✓	**1-3**	✓	✓		37.2	0.979	0.020	30.2	0.954	0.028
✓	✓	**1-3**	✓	✓	$+\mathcal{L}_{\text{scaling}_d}$	37.2	**0.980**	**0.018**	30.9	0.957	0.026

Table 2. Ablation study on subject #306. **Green shades** indicate the optimal results.

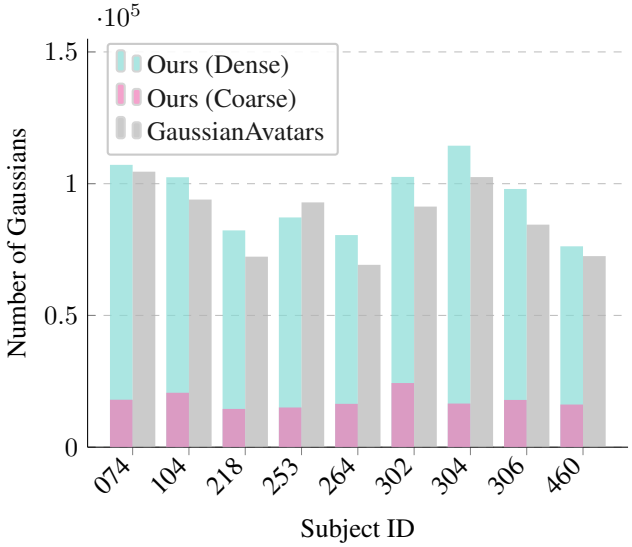

Figure 6. The comparison of the number of Gaussians between Ours (Coarse+Dense) and GaussianAvatars [31]. The red/cyan bars in the chart represent the combined count of both our coarse and dense Gaussians.

Gaussians fail to maintain structure in the head region, as shown in Figure 7 second row. When employing both dense and coarse Gaussians, our method successfully captures fine-grained facial details while preserving structural integrity in the head region. While using only dense Gaussians yields satisfactory results for training sequences, it leads to distorted avatar renderings when handling novel poses and expressions, primarily due to transformation mismatches.

Expression-dependent position offset. The expression-dependent position offset is the key for our method to capture dynamic facial details. When it is disabled, the Gaussians are strictly constrained by the FLAME mesh, as shown in the third row of Table 2. While this maintains stability during self-reenactment, it prevents the recovery of fine-grained details, as demonstrated in Figure 8.

Division strategy. To evaluate the effectiveness of our proposed mesh division strategy (1-3 division, i.e., dividing one triangle into 3 subregions.), we conduct a comparison with both simpler (1-2 division) and more complex (1-4 division) approaches. Both qualitative results in Figure 7 and quantitative metrics (rows 4-5 in Table 2) demonstrate that while simpler division models exhibit limited expressive power, more complex variants with additional initial parameters present optimization difficulties, ultimately resulting in inferior rendering quality during self-reenactment.

Regularization on orthogonality of offset basis. The orthogonality of FLAME's expression basis ensures that the parameterized model remains free of ambiguity during expression blendshapes, which we aim to inherit to enhance the stability of our expression-dependent offset.

When the ortho loss is disabled, the head avatar exhibits unexpected colors. The results in the seventh row of Table 2 and the qualitative comparison in Figure 7 clearly demonstrate the effectiveness of our orthogonality regularization in producing a more consistent and realistic head shape. Moreover, thanks to this loss function and the rotation control, our method can significantly reduce erroneous colors and distortions (see Figure 7, column 4, 5, 6).

Regularization on the scaling of Gaussian splats. Without the scaling loss, the reconstruction quality deteriorates in the self-reenactment task. This occurs because the coarse Gaussians, with uneven scales, struggle to be distributed structurally across the template mesh, preventing them from effectively constraining the coarse Gaussians to their parent triangles. Furthermore, the last row of Table 2 confirms that adding this regularization to the dense Gaussians does not enhance the reconstruction results, which aligns with previ-

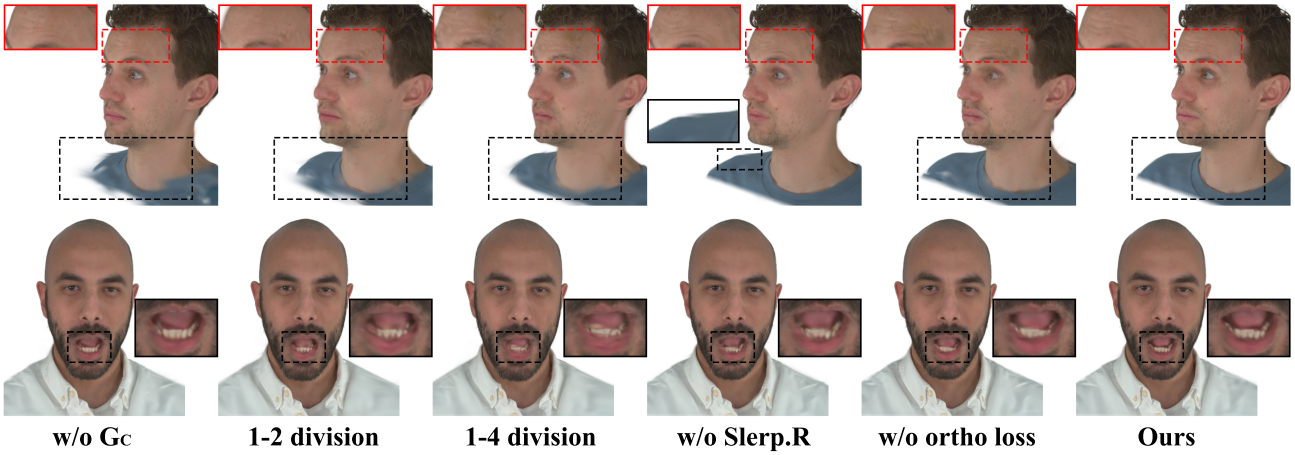

Figure 7. Our joint optimization strategy, division strategy and regularization strategy are all beneficial for head avatar to maintain a stable structure when rendering new expressions.

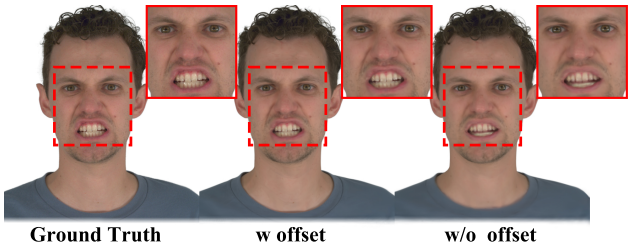

Figure 8. Failed to capture the dynamic details without the expression-dependent position offset.

ous analysis.

5. Discussion and Conclusion

5.1. Limitations

Due to the lack of authentic modeling for facial muscle movement, fully recovering expression-dependent facial details that align with a specific human head remains challenging. Moreover, our reliance on explicit mesh representation still prevents our method from capturing the dynamic changes in hair corresponding to pose variations. Beyond these limitations, our method requires approximately an extra 50 MB of storage space to store parameters related to expression offset bases and rotation control. Future research will focus on two aspects: (1) incorporating physical laws to model the motion of real human head components; (2) improving model compression to reduce the storage overhead caused by additional parameter storage.

5.2. Ethical Considerations

The proposed method enables the generation of photorealistic avatars and artificial portrait videos, which, while promising for legitimate uses like virtual interaction, also pose inherent risks of misuse—such as privacy violation, identity theft, misinformation dissemination, and eroded trust in media. These malicious applications could bring significant negative societal impacts. We condemn all harmful exploitation of this technology and emphasize the urgency of developing reliable detection methods to distinguish authentic from forged content, thus safeguarding media authenticity and mitigating adverse social consequences.

5.3. Conclusion

In this paper, we propose DoubleGaussianAvatar, a novel method for generating photorealistic head avatars from multi-view videos. Our approach leverages two sets of Gaussians to represent the human avatar, with each set serving a distinct purpose: one for maintaining expression controllability and the other for refining facial details. We introduce an expression-dependent position offset to directly adjust the Gaussians to capture subtle facial nuances. Through a carefully designed initialization and a joint optimization strategy, the two sets of Gaussians work synergistically to render the avatar in any given expression or pose. Our method outperforms state-of-the-art techniques in capturing facial details, demonstrating its potential for applications requiring dynamic, photorealistic facial representations.

References

[1] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1999)*, pages 187–194. ACM Press, 1999. 1, 2

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

[3] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5543–5552, 2016. 2

[4] Z. Chai, T. Zhang, T. He, X. Tan, T. Baltrusaitis, H. Wu, R. Li, S. Zhao, C. Yuan, and J. Bian. Hiface: High-fidelity 3d face reconstruction by learning static and dynamic details. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9087–9098, 2023. 2

[5] Y. Chen, L. Wang, Q. Li, H. Xiao, S. Zhang, H. Yao, and Y. Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 3

[6] Y. Duan, F. Wei, Q. Dai, Y. He, W. Chen, and B. Chen. 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[7] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)*, 39(5):1–38, 2020. 2

[8] Y. Feng, H. Feng, M. J. Black, and T. Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 2

[9] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 2

[10] L. Gao, J. Yang, B.-T. Zhang, J.-M. Sun, Y.-J. Yuan, H. Fu, and Y. Lai. Real-time large-scale deformation of gaussian splatting. *ACM Transactions on Graphics*, 2024. 3

[11] X. Gao, C. Zhong, J. Xiang, Y. Hong, Y. Guo, and J. Zhang. Reconstructing personalized semantic facial nerf models from monocular video. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 41(6), 2022. 2

[12] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter. Morphable face models-an open framework. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 75–82. IEEE, 2018. 2

[13] S. Giebenhain, T. Kirschstein, M. Rünz, L. Agapito, and M. Nießner. Npga: Neural parametric gaussian avatars. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3

[14] Y. Hong, B. Peng, H. Xiao, L. Liu, and J. Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 2

[15] Y.-H. Huang, Y.-T. Sun, Z. Yang, X. Lyu, Y.-P. Cao, and X. Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4220–4230, 2024. 3

[16] Y. Jiang, C. Yu, T. Xie, X. Li, Y. Feng, H. Wang, M. Li, H. Lau, F. Gao, Y. Yang, et al. Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–1, 2024. 3

[17] K. Kania, S. J. Garbin, A. Tagliasacchi, V. Estellers, K. M. Yi, J. Valentin, T. Trzciński, and M. Kowalski. Blendfields: Few-shot example-driven facial modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 404–415, 2023. 2

[18] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 2, 3, 4, 5

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] T. Kirschstein, S. Qian, S. Giebenhain, T. Walter, and M. Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), jul 2023. 6

[21] B. Lei, J. Ren, M. Feng, M. Cui, and X. Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 394–403, 2023. 2

[22] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2, 5

[23] Z. Li, Z. Chen, Z. Li, and Y. Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8508–8520, 2024. 3

[24] Z. Li, Z. Zheng, L. Wang, and Y. Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. *arXiv*, 2023. 3

[25] Y. Lin, Z. Dai, S. Zhu, and Y. Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21136–21145, 2024. 3

[26] J. Luiten, G. Kopanas, B. Leibe, and D. Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 3

[27] S. Ma, Y. Weng, T. Shao, and K. Zhou. 3d gaussian blendshapes for head avatar animation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024. 3

[28] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[29] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2

[30] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 1, 2

[31] S. Qian, T. Kirschstein, L. Schoneveld, D. Davoli, S. Giebenhain, and M. Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 2, 3, 4, 5, 6, 8, 9

[32] Z. Shao, Z. Wang, Z. Li, D. Wang, X. Lin, Y. Zhang, M. Fan, and Z. Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 6, 8

[33] J. Sun, H. Jiao, G. Li, Z. Zhang, L. Zhao, and W. Xing. 3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20675–20685, 2024. 3

[34] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 3

[35] D. Wang, P. Chandran, G. Zoss, D. Bradley, and P. Gotardo. Morf: Morphable radiance fields for multiview neural head modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2

[36] J. Wang, J.-C. Xie, X. Li, F. Xu, C.-M. Pun, and H. Gao. Gaussianhead: High-fidelity head avatars with learnable gaussian derivation, 2024. 3

[37] L. Wang, Z. Chen, T. Yu, C. Ma, L. Li, and Y. Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20333–20342, 2022. 2

[38] Y. Wang, X. Wang, R. Yi, Y. Fan, J. Hu, J. Zhu, and L. Ma. 3d gaussian head avatars with expressive dynamic appearances by compact tensorial representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21117–21126, 2025. 2, 3, 6, 8

[39] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024. 3

[40] B. Xu, J. Zhang, K.-Y. Lin, C. Qian, and Y. He. Deformable model-driven neural rendering for high-fidelity 3d reconstruction of human heads under low-view settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17924–17934, 2023. 2

[41] Y. Xu, B. Chen, Z. Li, H. Zhang, L. Wang, Z. Zheng, and Y. Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[42] Y. Xu, L. Wang, X. Zhao, H. Zhang, and Y. Liu. Avatarmav: Fast 3d head avatar reconstruction using motion-aware neural voxels. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 2

[43] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 601–610, 2020. 2

[44] Z. Yang, X. Gao, W. Zhou, S. Jiao, Y. Zhang, and X. Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20331–20341, 2024. 3

[45] Z. Yang, H. Yang, Z. Pan, and L. Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*, 2023. 3

[46] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[47] Y. Zheng, W. Yifan, G. Wetzstein, M. J. Black, and O. Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21057–21067, 2023. 2

[48] Y. Zhuang, H. Zhu, X. Sun, and X. Cao. Mofanerf: Morphable facial neural radiance field. In *European conference on computer vision*, pages 268–285. Springer, 2022. 2

[49] W. Zielonka, T. Bolkart, T. Beeler, and J. Thies. Gaussian eigen models for human heads. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15930–15940, 2025. 2, 3

[50] W. Zielonka, T. Bolkart, and J. Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022. 2

[51] W. Zielonka, T. Bolkart, and J. Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4574–4584, 2023. 2