Eurographics Symposium on Rendering (2025)
B. Wang and A. Wilkie (Editors)

Joint Gaussian Deformation in Triangle-Deformed Space for High-Fidelity Head Avatars

Jiawei Lu[1], Kunxin Guang[2], Conghui Hao[1], Kai Sun[3], Jian Yang[1], Jin Xie[2], Beibei Wang[†2]

[1]Nankai University, [2]Nanjing University, [3]China Mobile Zijin Innovation Institute

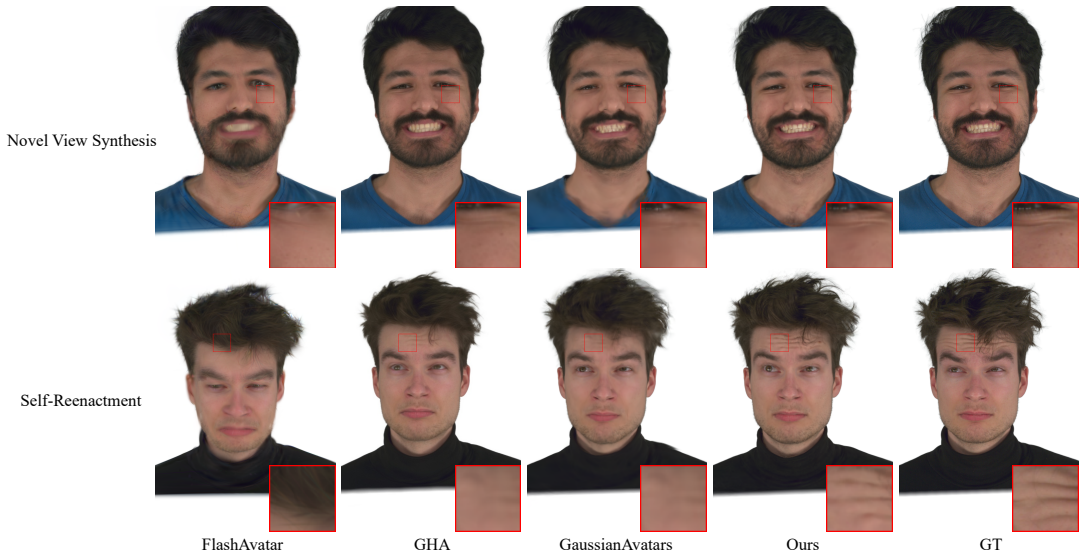

Figure 1: *We propose Joint Gaussian Deformation in Triangle-Deformed Space, decoupling the complex deformation of Gaussian into two simpler deformations, which are much simpler to represent or learn, consisting of a learnable displacement map-guided Gaussian-triangle binding and a neural-based deformation refinement, achieving high-fidelity animation and high-frequency details of head avatars.*

Abstract

Creating 3D human heads with mesoscale details and high-fidelity animation from monocular or sparse multi-view videos is challenging. While 3D Gaussian splatting (3DGS) has brought significant benefits into this task, due to its powerful representation ability and rendering speed, existing works still face several issues, including inaccurate and blurry deformation, and lack of detailed appearance, due to difficulties in complex deformation representation and unreasonable Gaussian placement. In this paper, we propose a joint Gaussian deformation method by decoupling the complex deformation into two simpler deformations, incorporating a learnable displacement map-guided Gaussian-triangle binding and a neural-based deformation refinement, improving the fidelity of animation and details of reconstructed head avatars. However, renderings of reconstructed head avatars at unseen views still show artifacts, due to overfitting on sparse input views. To address this issue, we leverage synthesized pseudo views rendered with fitted textured 3DMMs as priors to initialize Gaussians, which helps maintain a consistent and realistic appearance across various views. As a result, our method outperforms existing state-of-the-art approaches with about 4.3 dB PSNR in novel-view synthesis and about 0.9 dB PSNR in self-reenactment on multi-view video datasets. Our method also preserves high-frequency details, exhibits more accurate deformations, and significantly reduces artifacts in unseen views.

CCS Concepts

• ***Computing methodologies*** → *Rendering;*

1. Introduction

Creating animatable 3D human heads from monocular or sparse multi-view videos has been a longstanding problem in computer vi-

† Corresponding author

sion and graphics. It can enrich 3D face assets in many applications, including digital humans, film, and virtual reality. Unfortunately, it is still challenging to reconstruct detailed photorealistic appearance with high-fidelity animations driven by new poses or expressions, while maintaining real-time rendering. Particularly, the limited input views further raise difficulties in this task.

The combination of 3D Morphable Models (3DMMs) [BV99] and Neural Radiance Fields (NeRF) [MST*20] has brought the opportunity to achieve both detailed representation and animatable capability. However, these methods [GTZN21, ZAB*22, ZBT23] have difficulties in achieving real-time rendering, due to their volume rendering mechanism. Recently, 3D Gaussian splatting (3DGS) [KKLD23] has shown powerful representation capability and high-performance rendering speed. Several works [QKS*24, MWSZ24,XCL*24,XGGZ24] have introduced 3DGS into 3D head reconstruction and animation by binding the Gaussian representation and the 3DMMs explicitly or implicitly. While these methods have shown impressive capability, they still encounter several issues. First, these methods represent 3D deformation explicitly or implicitly, where the former has limited capability for the accessory regions (e.g., mouth, hair, etc.), and the latter has difficulties in representing the complex deformation mapping from the canonical space to the deformed space, leading to inaccurate and blurry deformation. Second, some mesoscale appearances, like the winkles, are missing from these approaches, due to unreasonable Gaussian placement. Last, renderings at unseen views show severe artifacts, due to the overfitting on the sparse input views.

In this paper, we aim to address the above three issues: **inaccurate and blurry deformation**, **lack of detailed appearance**, and **overfitting artifacts at unseen views**. Our key insight is that the deformation mapping from the canonical space to the deformed space is complex, which is non-trivial to be represented accurately with a lightweight neural network. Therefore, we propose a joint deformation method, including both explicit and implicit components, where the former binds the Gaussians with triangles directly and the latter defines a refinement mapping in the deformed space. This way, the deformation can be represented more accurately. Furthermore, to enhance detailed appearance, we use a displacement map to guide the placement of the Gaussians, so that the mesoscale details can be preserved. Although our joint deformation method can improve the fidelity and high-frequency details of reconstructed head avatars, the appearance at unseen views still suffers from artifacts. To address this, we leverage synthesized pseudo views rendered with fitted textured 3DMMs as priors to initialize Gaussians, which enhances a consistent and realistic appearance across various views. Consequently, our method outperforms existing state-of-the-art approaches numerically, achieving an improvement of about 4.3 dB PSNR in novel-view synthesis and about 0.9 dB PSNR in self-reenactment on multi-view video datasets [KQG*23]. In terms of visual quality, our method preserves high-frequency details such as wrinkles and hair, displays more accurately matched deformations, and significantly reduces artifacts at less common/unseen views. To summarize, our main contributions include:

- We propose a joint Gaussian deformation method, combining explicit Gaussian-triangle binding with neural-based Gaussian deformation refinement, resulting in high-fidelity 3D head reconstruction and animation.
- We introduce a displacement map guide for Gaussian placement to enhance the appearance of fine details.
- We utilize synthesized pseudo views as priors to initialize Gaussians, reducing overfitting artifacts at unseen views.

2. Related Work

2.1. 3D human head modeling

Existing works on modeling 3D human heads from monocular or multi-view videos can be mainly categorized into two types: implicit model-based and explicit model-based approaches.

NeRF [MST*20] has been introduced into the 3D head reconstruction by Gafni et al. [GTZN21] by combining it with a low-dimensional morphable model to satisfy the dynamics of heads. Wang et al. [WBL*21] combine discrete and continuous volume representations to achieve high-resolution rendering of dynamic human heads and the upper body. While these methods have shown a remarkable capability for 3D head modeling, they have difficulties in generalizing to novel poses and expressions, and tend to suffer long rendering time. In addition, IMavatar [ZAB*22] reconstructs implicit head avatars using neural implicit functions and models head deformations with learnable blendshapes and skinning fields. Though it improves generalization beyond train-time expressions, their work is still limited by training time as well as the efficiency of rendering. Moreover, INSTA [ZBT23] constructs a surface-embedded implicit radiance field of 3D heads utilizing neural graphics primitives [MESK22], and achieves relatively high rendering efficiency.

Typical of explicit 3D head representations, 3DMMs [BV99] use Principal Component Analysis (PCA) to decompose shape priors of a 3D head into a low-dimensional space, making it convenient and stable to manipulate. Afterwards, a number of works introduce 3DMMs or their variants [VBPP06, CWZ*13, LBB*17, FFBB21] into the reconstruction of 3D heads and are able to drive head avatars based on novel poses and expressions. These works are typically based on optimization [TZS*16] or combined with deep neural networks [TBG*19, SSL*20], which are used to predict the corresponding offsets for novel poses or expressions. Another type of explicit models is based on point representations, overcoming the limitations of mesh-based methods, and enabling more efficient fitting of arbitrary topologies. Pointavatar [ZYW*23] uses this representation to model a deformable 3D head avatar and implements a self-supervised lighting disentanglement. However, their work requires a significant number of points, leading to a considerable computational burden.

2.2. Head reconstruction and animation with 3DGS

Recently, 3DGS [KKLD23] gains much attention for its fidelity in scene reconstruction as well as its rendering speed, which is migrated to 3D head reconstruction and animation within the last year [SSS*24, DNM*25]. Specifically, 3DGS-based head reconstruction and animation can also be broadly divided into two categories. One class of works [QKS*24] binds Gaussian points to the

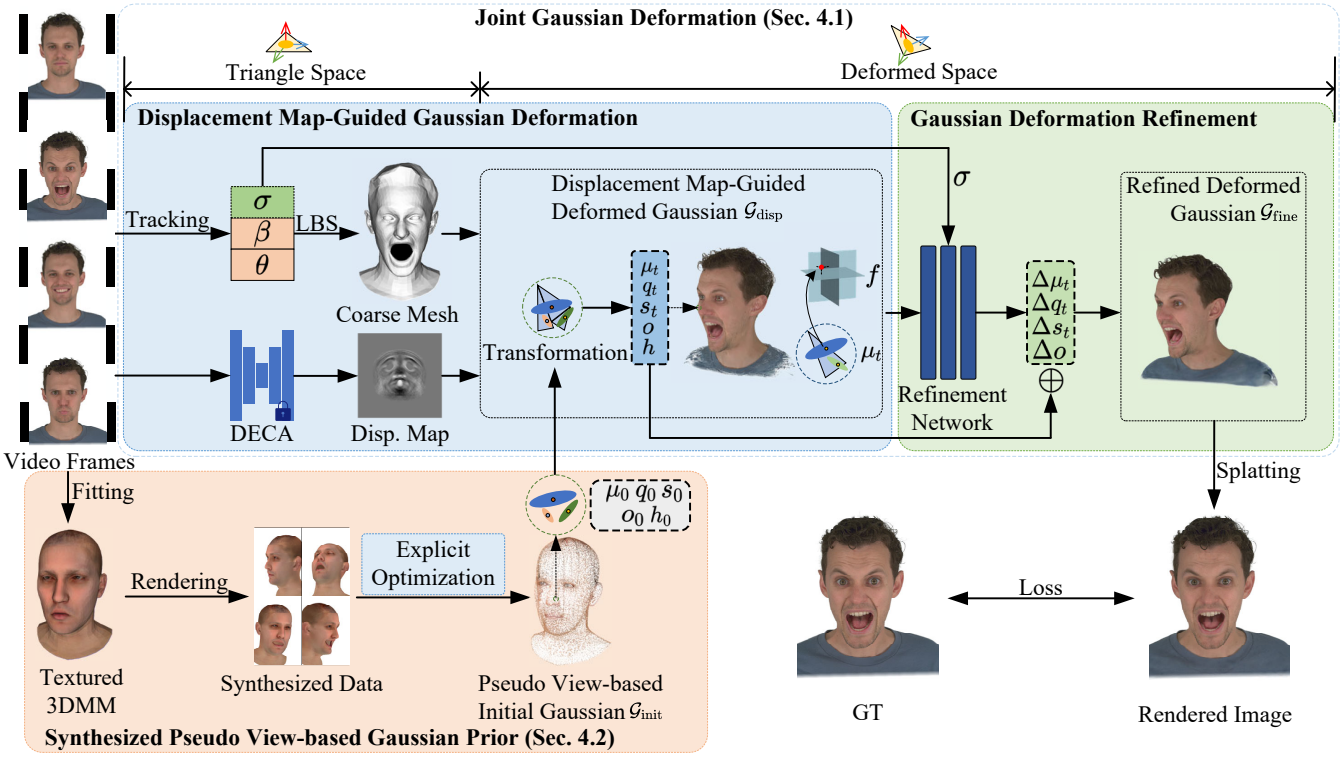

Figure 2: *The key of our method is a joint Gaussian deformation which represents the 3D head deformation of Gaussians with two components (an explicit component and an implicit component). Specifically, in the explicit deformation component, Gaussians parameterized in the triangle space are bound to triangles guided by displacement maps, with attributes initialized synthesized pseudo view-based Gaussian prior module. Then, Gaussians are mapped to the deformed world space via triangle-Gaussian transformation. In the implicit deformation component, Gaussian positions, together with a spatial semantic feature encoded by a learnable triplane and the expression, are fed into a refinement network to predict offsets, leading to final refined deformed Gaussians.*

triangular facets of a 3DMM, explicitly following the movement of vertices on the 3DMM driven by pose or expression. Similarly, Ma et al. [MWSZ24] model arbitrary facial expressions of a head as linear combinations of a base head model and a set of expression blendshapes using 3D Gaussian representations. While these methods offer low training and rendering costs, they are limited by the underlying 3DMMs, struggling with unmodeled regions and fine details like wrinkles.

Another class of works [CWL*24, GKR*24, WXL*23, TKG*24] relies on neural networks to predict changes in Gaussian point attributes based on input expression codes, using neutral meshes obtained from methods such as 3DMMs or implicit signed distance field (SDF). FlashAvatar [XGGZ24] embeds a uniform 3D Gaussian field on the surface of a parametric face model and learns additional spatial offsets to capture details. However, their work forgoes adaptive density control for Gaussian points in order to achieve faster rendering speed, limiting its ability to capture high-frequency details of heads. Xu et al. [XCL*24] utilize a fully learned multilayer perceptron (MLP)-based deformation field to animate neutral 3D Gaussians of heads based on a geometry-guided initialization. Due to the heavy reliance on neural components, their method requires a considerable amount of training time.

Different from the above methods, we present a joint Gaussian deformation through decoupling the complicated deformation into two simpler components including displacement map-guided explicit deformation and implicit deformation refinement.

3. Preliminary

3DGS. 3DGS [KKLD23] reconstructs a static scene using a series of anisotropic 3D Gaussian based on input images and camera parameters. Each 3D Gaussian is defined as:

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

where μ is the position of the Gaussian, and Σ is the anisotropic covariance matrix of the Gaussian. Σ is further decomposed into a scaling matrix S and a rotation matrix R, represented by a scaling vector s and a unit quaternion q respectively as $\Sigma = RSS^TR^T$. In addition, each Gaussian contains a series of spherical harmonic coefficients h to represent color c as well as an opacity parameter o. When an image is to be rendered, the color of each pixel is calculated by blending all the Gaussians overlapping the pixel as:

$$C = \sum_{i \in \mathbb{N}} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{2}$$

where α represents the blending weight obtained by the 2D projection of the 3D Gaussian multiplied by the opacity o.

GaussianAvatars. GaussianAvatars [QKS*24] explicitly binds Gaussian points to the tracked meshes and establishes a mapping of Gaussian attributes between the local triangle space and the deformed global space. Specifically, the attributes of each Gaussian point in the local space are defined as $\mathcal{G}_{loc} = \{\mu, q, s, o, h\}$, with the corresponding triangle transformation as follows:

$$\mu_t = SQ\mu + P, q_t = Qq, s_t = Ss, \tag{3}$$

where S denotes the scaling of the triangle, Q represents the orientation of the triangle in the global space, and P indicates the mean position of the triangle's three vertices. For simplicity, we define the above three metrics of the triangles as M. For each time step, the local Gaussian attributes $\mathcal{G}_{loc}$ are converted into the corresponding global Gaussian attributes $\mathcal{G}_{glob} = \{\mu_t, q_t, s_t, o, h\}$ used for image rendering.

4. Our method

In this paper, our aim is to reconstruct photorealistic appearance of 3D heads with mesoscale details, maintain high-fidelity animations driven by new poses or expressions given monocular or sparse multi-view videos.

Since 3DGS has shown impressive representation capability and high-performance rendering speed, we opt for 3D Gaussians as our representation. The key to achieving our target is the representation capability of 3D Gaussians for complex deformation of 3D heads under various expressions and poses and detailed appearance given sparse views. To achieve this, we propose a joint method with explicit-implicit Gaussian deformation (Sec. 4.1) for accurate deformation. We also introduce a displacement map-guided Gaussian placement to preserve detailed appearance. Then, we propose a synthesized pseudo view-based Gaussian prior (Sec. 4.2) to prevent overfitting artifacts at unseen views. Our pipeline is illustrated in Figure 2.

4.1. Joint Gaussian deformation

Since we target an animatable 3D Gaussian representation under various poses and expressions, the deformation operation needs to be represented accurately. On one hand, existing explicit approaches represent the deformation directly by binding Gaussian and triangles, leading to blurriness for accessory regions, which lack 3DMM definition. On the other hand, other methods learn the deformation from the canonical space to the deformed space with a small MLP, leading to inaccurate deformation, due to its limited representation ability. To this end, we propose a joint Gaussian deformation method representing the deformation step by step, which includes both explicit and implicit components. The explicit component binds the Gaussians with triangles in a learnable manner, and the implicit component further refines the deformation in the deformed space with a small MLP.

Displacement map-guided explicit deformation. Given multi-view or monocular videos, we first extract the corresponding shape parameters β, expression parameters σ, and pose parameters θ using FLAME tracking [QKS*24]. We also build a coarse mesh with linear blend skinning (LBS) [FFBB21]. Then, we bind each Gaussian point to the corresponding triangular facet of the coarse mesh, similar to GaussianAvatar, as shown in Sec. 3. The Gaussian points are defined in the local triangle space, and are mapped to the global/world space considering the triangle transformation (Eqn. 3).

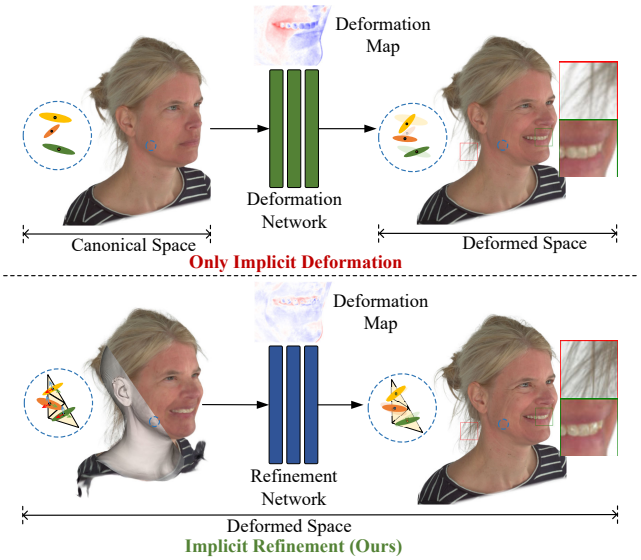

Figure 3: *Instead of directly predicting Gaussian attribute offsets from the canonical space, we adopt the implicit refinement in the deformed space, which predicts smaller deformation that is easier to learn, significantly improving the fidelity of head avatars.*

To improve the representation of detailed appearances, we introduce a displacement map that serves as a guide for the placement of the Gaussians, which allows for the preservation of the mesoscale details. Specifically, we use a predefined dense FLAME template to map the coarse mesh onto a finer topology by subdividing each triangle uniformly into ten facets, following DECA [FFBB21]. Then, we adjust the detailed mesh vertices according to the displacement maps extracted from the multi-view/monocular videos, resulting in about 110k triangles. Thus, each Gaussian point is bound to a facet of the detailed mesh, with the corresponding displacement map-guided global Gaussian attributes $\mathcal{G}_{disp}$ after the triangle transformation, based on the metrics M of the subdivided triangles, represented as follows:

$$\mathcal{G}_{disp} = \{\mu_t, q_t, s_t, o, h\}. \tag{4}$$

Please note that, the expression parameters σ, and pose parameters θ are updated during training. Therefore, the metrics M of the triangular facets are all learnable. This is a key difference from FlashAvatar [XGGZ24].

Implicit deformation refinement with per-Gaussian feature embeddings. On top of the Gaussians deformed by the triangle transformation, we further refine the Gaussian attributes (position, rotation, scaling, and opacity) with a small MLP, as illustrated in Figure 3. We encode each Gaussian into a feature with a learnable triplane representation T, consisting of three orthogonal feature planes aligned with axes, in order to store spatial information of the head avatar. We then decode the Gaussian basis, feature embedding,

and expression into Gaussian attribute offsets, which are added to the Gaussian basis, forming the refined Gaussian attributes.

Specifically, for any global Gaussian position attribute μ_t, we query the corresponding feature vector from the triplane T by projecting it onto the axis-aligned feature planes. We then concatenate three bilinear interpolated features to form the per-Gaussian feature as:

$$f = f_{\text{xy}} \oplus f_{\text{xz}} \oplus f_{\text{yz}}, \tag{5}$$

where f_{xy}, f_{xz}, and f_{yz} respectively denote the features of μ_t on the three feature planes.

Then, the triplane feature f and the global Gaussian position attribute μ_t, along with the expression parameters σ are decoded by a small MLP to predict the offsets for the global Gaussian attributes:

$$\{\Delta\mu_t, \Delta q_t, \Delta s_t, \Delta o\} = \phi(\mu_t, f; \sigma), \tag{6}$$

where ϕ represents the MLP used to predict offsets, while $\Delta\mu_t$, Δq_t, Δs_t, and Δo denote the predicted offsets for the global Gaussian attributes μ_t, q_t, s_t, and o, respectively. Then, the final global refined deformed Gaussian attributes are expressed as follows:

$$\mathcal{G}_{\text{fine}} = \{\mu_t \oplus \Delta\mu_t, q_t \oplus \Delta q_t, s_t \oplus \Delta s_t, o \oplus \Delta o, h\}. \tag{7}$$

4.2. Synthesized pseudo views as Gaussian prior

Thanks to the joint deformation, our method can effectively improve the fidelity and high-frequency details of reconstructed head avatars. However, when the viewing angles deviate significantly from those in the training set, the head avatars exhibit noticeable artifacts. This is essentially due to overfitting on the sparse training views. As adopted by Xu et al. [XCL*24], a straightforward approach is to apply screen-space super resolution. However, this inevitably compromises the 3D consistency of head avatars and leads to texture flickering. Our key insight is that although acquiring real dense-view data is challenging, we can leverage the accessibility of synthesized data and the flexibility of rendering views, incorporating them as geometry and color priors. One possible way is to leverage the reconstructed 3DMM models to synthesize pseudo-view images and supervise Gaussian optimization with these images. Unfortunately, these pseudo-view images exhibit low quality compared to the input videos, which leads to degraded reconstruction if using them for supervision directly. Alternatively, we introduce a simple yet effective way to inject these priors into the Gaussian optimization, which utilizes them for Gaussian initialization, and then updates the Gaussians with the real data.

Specifically, we reconstruct a textured FLAME mesh for each identity using Photometric FLAME Fitting [LBB*17], and generate different synthesized images under a densely distributed set of camera poses. Then, we use these synthesized pseudo-view images for initial Gaussian optimization. To make the optimization of the Gaussian attribute prior relatively lightweight and easy to converge, we only perform the explicit deformation. We first refine the mesh to create a finer version and initialize Gaussians by binding them to the mesh, then optimize thes e Gaussians using the synthesized pseudo views, adjusting their attributes, such as position and rotation. The optimized identity-specific Gaussian prior is injected into the Gaussian attributes in the triangle space of the identity as initialization for subsequent optimization:

$$\mathcal{G}_{init} = \{\mu_0, q_0, s_0, o_0, h_0\}. \tag{8}$$

5. Experiments

5.1. Implementation details

We implement our method based on PyTorch [PGC*17] and train it with the Adam optimizer [Kin14] on an NVIDIA GeForce RTX 3090 with the following loss:

$$\mathcal{L} = \mathcal{L}_{\text{c}} + \lambda_{Sobel}\mathcal{L}_{Sobel} + \lambda_{\mu}\mathcal{L}_{\mu} + \lambda_{\text{s}}\mathcal{L}_{\text{s}} + \lambda_{\phi}\mathcal{L}_{\phi}, \tag{9}$$

where $\mathcal{L}_{\text{c}}$ refers to the same loss function as in 3DGS [KKLD23], combining $\mathcal{L}_1$ with D-SSIM term, and $\mathcal{L}_{Sobel}$ is the $\mathcal{L}_2$ distance between the Sobel operator [KVB88] (with a radius of 1) results of rendered images and ground truth images. $\mathcal{L}_{\mu}$ and $\mathcal{L}_{\text{s}}$ are the position loss and scaling loss with threshold used in GaussianAvatars [QKS*24] to regularize the local position and scaling attributes of Gaussian points. $\mathcal{L}_{\phi}$ means $\mathcal{L}_2$ regularization on the position offset predicted by the refinement network, encouraging small offset predictions. All the λ in the above loss are the weights used to balance the loss terms. λ_{Sobel}, λ_{μ}, λ_{s}, and λ_{ϕ} are set as 1.0, 1e-2, 1.0, and 5e-3, respectively. For the joint Gaussian deformation optimization, the dimension of the triplane representation is $3\times64\times64\times32$ (with a resolution of 64×64 and 32 channels), trained with a learning rate of 1e-4, and the refinement MLP consists of 5 layers with a width of 256, trained with a learning rate of 4e-5. For Gaussian point densification, we adopt improved adaptive density control [GKR*24]. The training iterations for the synthesized pseudo view-based Gaussian prior are set to 300,000, taking about 4 hours. For the joint Gaussian deformation, training iterations are set to 600,000, taking about 30 hours.

5.2. Setup

Dataset. We conduct experiments on nine subjects from the multi-view dataset NeRSemble [KQG*23] and eight subjects from the monocular dataset INSTA [ZBT23]. For each subject in the NeRSemble dataset, each time step includes 16 images from different viewpoints, all downsampled to a resolution of 802×550. The image resolution for the INSTA dataset is uniformly 512×512. All other preprocessing steps are consistent with those used in GaussianAvatars [QKS*24].

Comparison methods. We compare our proposed method with three state-of-the-art approaches: FlashAvatar [XGGZ24], GHA [XCL*24], and GaussianAvatars [QKS*24]. We employ three widely-used metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [WBSS04], and Perceptual LPIPS [ZIE*18]. Furthermore, since GHA leverages super-resolution, we train it at a resolution of 2048×2048, as specified in their original paper. To ensure a fair comparison, we subsequently downsample the rendered images from GHA to match the resolution used by the other methods before conducting the metric evaluations.

Table 1: *Quantitative comparison between our method and comparison methods on NeRSemble dataset, with the best results in* ***bold****.*

	Novel View Synthesis			Self-Reenactment		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
FlashAvatar	23.92	0.859	0.148	20.43	0.833	0.174
GHA	30.34	0.926	0.045	24.91	0.903	0.057
GaussianAvatars	31.02	0.936	0.065	25.81	**0.910**	0.076
Ours	**35.30**	**0.958**	**0.033**	**26.67**	**0.910**	**0.053**

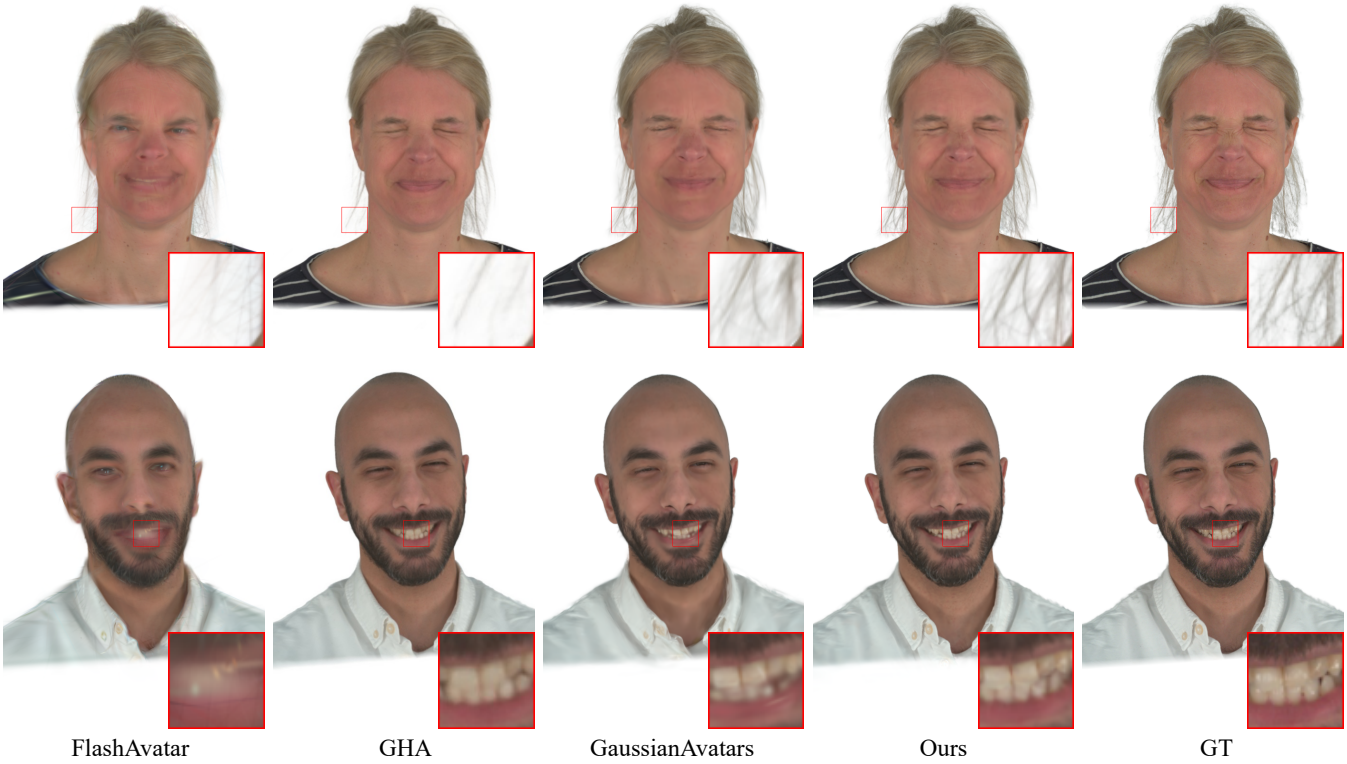

Figure 4: *Qualitative comparison between our method and comparison methods on novel-view synthesis of head avatars. Our method can reconstruct more high-frequency details.*

Table 2: *Quantitative comparison between our method and comparison methods on INSTA dataset, with the best results in* ***bold****.*

	Self-Reenactment		
	PSNR ↑	SSIM ↑	LPIPS ↓
FlashAvatar	27.69	0.937	0.065
GHA	27.76	0.931	0.040
GaussianAvatars	27.96	0.935	0.045
Ours	**28.97**	**0.942**	**0.037**

5.3. Comparison with previous methods

The training costs for previous work and our approach under a single RTX 3090 are as follows: GaussianAvatars (9 hours), GHA (48 hours), and ours (30 hours + 4 hours). Regarding inference time, previous work and our method using a single RTX 3090 are as follows: GaussianAvatars (187 fps), GHA (22 fps), and ours (35 fps).

Novel view synthesis. We trained on 15 view images from the NeRSemble dataset, reserving the remaining view for evaluating rendering quality of reconstructed head avatars under novel views driven by expressions from the training sequence. As illustrated in Table 1 and Figure 4, our method outperforms others, particularly in recovering high-frequency details.

Self-reenactment. We evaluate the rendering quality of reconstructed head avatars driven by unseen expressions in the 16-view NeRSemble dataset and the monocular INSTA dataset. As shown in Table 1, Table 2, and Figure 5, our method achieves the highest quality in both numerical and visual results, showing more accurate deformations and enhanced facial details with new expressions.

Cross-identity reenactment. We assess the rendering quality of the reconstructed head avatars driven by expressions from other subjects. As illustrated in Figure 6, our method accurately enables cross-identity expression transfer for reconstructed head avatars, recovering the details of the driving expressions while maintaining high-fidelity quality and reducing the occurrence of artifacts.

Uncommon view rendering. We evaluate the rendering quality of

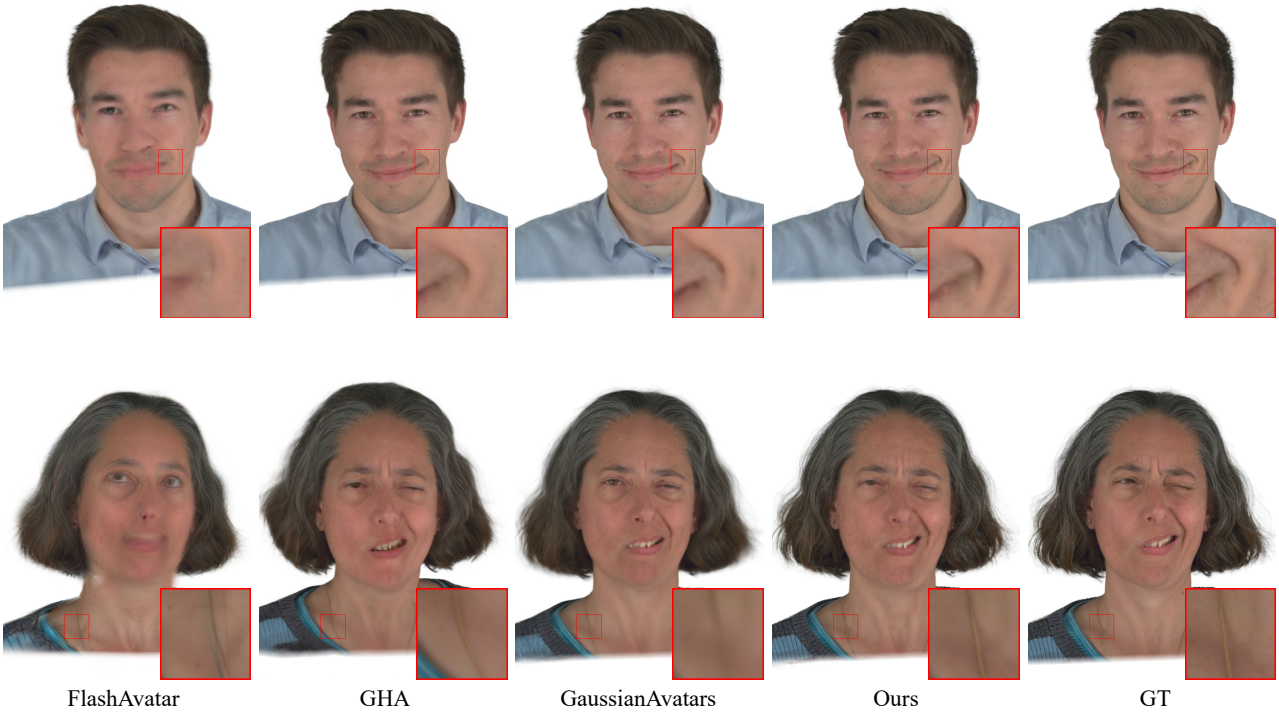

Figure 5: *Qualitative comparison between our method and comparison methods on self-reenactment of head avatars. Our method can exhibit more accurate deformations and finer facial details under new expressions.*

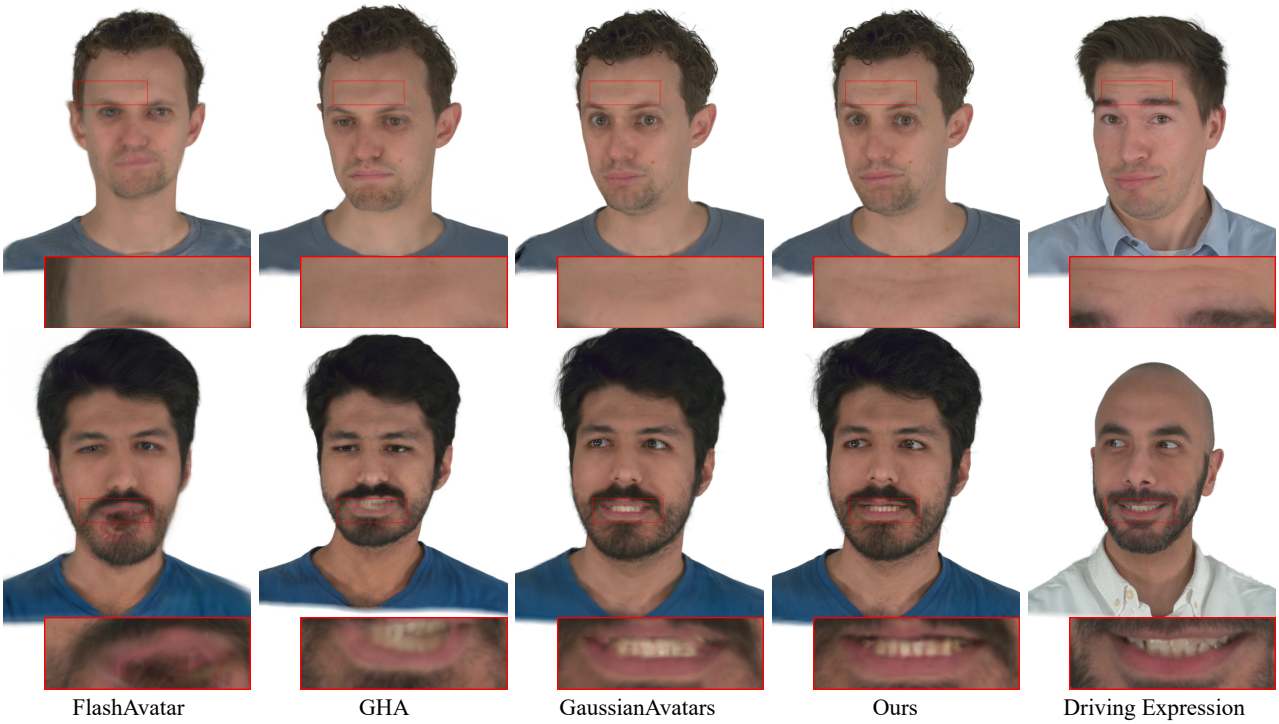

Figure 6: *Qualitative comparison between our method and comparison methods on cross-identity reenactment of head avatars. Our method can recover intricate details of driving expressions and mitigate appearance of artifacts.*

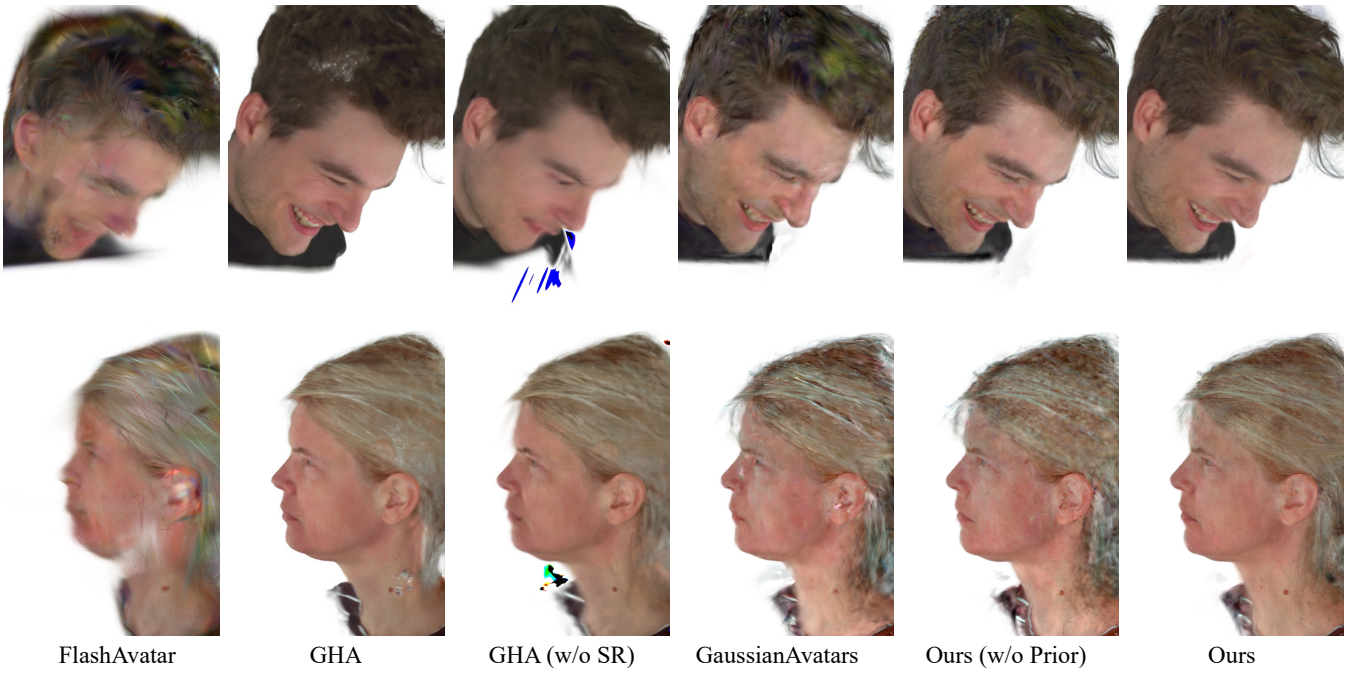

Figure 7: *Qualitative comparison between our method and comparison methods on uncommon view rendering of head avatars. Note that, as GHA incorporates super-resolution (SR), we also present results after training without the SR module, exhibiting noticeable overfitting artifacts. Additionally, we present results of our method without synthesized pseudo view-based Gaussian prior. Our method can significantly reduce overfitting artifacts.*

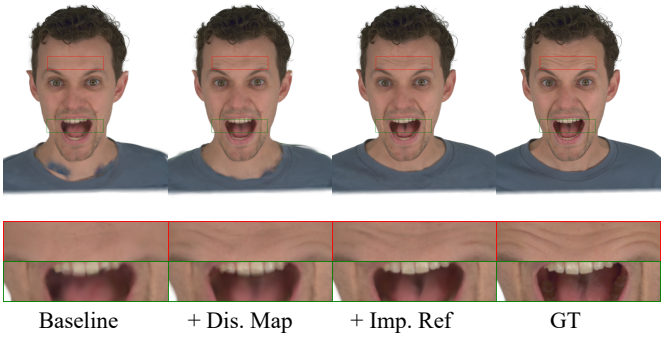

Figure 8: *Qualitative comparison of ablation study. "Disp. Map" means the displacement map, and "Imp. Ref." means the implicit Gaussian deformation refinement.*

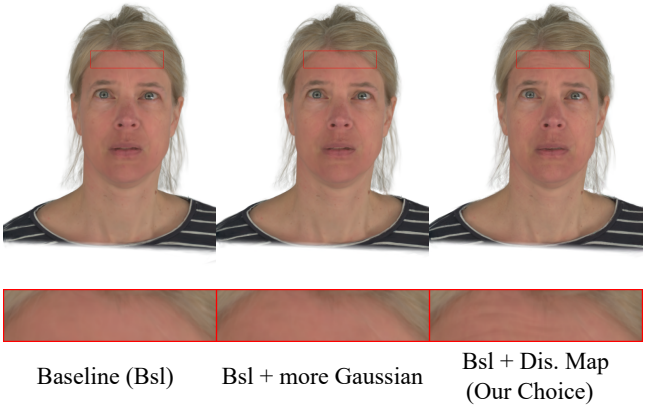

Figure 9: *Ablation study on displacement map guidance. Instead of directly increasing the Gaussian points, we leverage displacement maps for geometric guidance to aid in their placement.*

the reconstructed head avatars at uncommon views, which are distant from those in the training set, as shown in Figure 7. FlashAvatar and GaussianAvatars exhibit a noticeable degradation in visual quality, revealing a significant gap compared to our method. While GHA's fully implicit approach with super-resolution results in facial local smoothness due to the inductive bias of MLPs, it still shows unrealistic artifacts in the hair and neck regions, likely due to inaccuracies in the Gaussian points' opacity and color. In contrast, our method, with synthesized pseudo view information guidance, enhances consistent and realistic appearances at uncommon views.

5.4. Ablation study

We conduct ablation studies on three key components of our method. We progressively add our proposed components, investigating their impact from both numerical and visual perspectives, and the results are presented in Table 3 and Figure 8. The quantitative and qualitative quality gap highlights the effectiveness and importance of each component in our method. We begin with GaussianAvatars [QKS*24] as the baseline, which fails to capture high-

Table 3: *Ablation study on three key components, with the best/second-best results in **bold**/underscored. "Disp. Map" means the displacement map, "Imp. Ref." means the implicit Gaussian deformation refinement, and "Prior" means the synthesized pseudo view-based Gaussian prior. "Prior" enhances uncommon view rendering quality, though it slightly reduces self-reenactment numerical performance evaluated with common views in the dataset.*

	Novel View Synthesis			Self-Reenactment		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
	32.89	0.955	0.043	27.78	0.926	0.052
+Disp. Map	33.01	0.956	0.035	28.04	0.925	0.044
+Imp. Ref.	36.74	0.972	0.022	**28.85**	**0.931**	0.035
+Prior	**37.02**	**0.974**	**0.020**	28.67	0.929	**0.032**

frequency details and recover the appearance of parts not modeled by 3DMMs, as shown in the first column of Figure 8.

Displacement map guidance. We enhance the baseline with displacement maps, providing mesostructure-aware geometric guidance in the explicit Gaussian deformation, which aids in the recovery of high-frequency details such as wrinkles, as illustrated in the second column of Figure 8. To further validate the role of displacement map guidance, we increase the number of Gaussian points by solely adopting a lower densification gradient threshold and compare this approach with our proposed method. Figure 9 shows that displacement maps offer geometric guidance, enabling better Gaussian point placement, which helps recover details.

Implicit Gaussian deformation refinement. Next, we introduce implicit Gaussian deformation refinement, which further refines the Gaussian points after displacement map-guided explicit optimization. As shown in Table 3, this component provides significant improvements for novel view synthesis and self-reenactment. The third column of Figure 8 illustrates that high-frequency details, particularly inside the mouth, are further optimized, as the explicit Gaussian optimization relies on 3DMMs that lack accurate modeling for these regions.

Synthesized pseudo view-based Gaussian prior. Finally, we incorporate the synthesized pseudo view-based Gaussian prior. The last column of Figure 7 demonstrates that the introduction of the synthesized Gaussian prior effectively enhances the appearance realism and consistency of the reconstructed head avatars. Note that, the introduction of the synthesized pseudo view-based Gaussian prior results in a slight numerical decrease in performance for self-reenactment, as measured by PSNR and SSIM metrics. This reduction is attributed to the evaluation being performed using common views from the dataset.

Effect of synthesized pseudo view-based Gaussian prior with varying numbers of input views.

We perform additional ablation study of the synthesized pseudo view-based Gaussian prior with varying input view counts, using the four most extreme remaining views in the NeRSemble dataset for quantitative and qualitative comparisons. As illustrated in Table 4 and Figure 10, introducing the prior consistently enhances the fidelity of reconstructed head avatars under extreme views, effectively reducing overfitting artifacts. Notably, the fewer the input views, the more significant the prior's impact on the fidelity and realism of reconstructed head avatars.

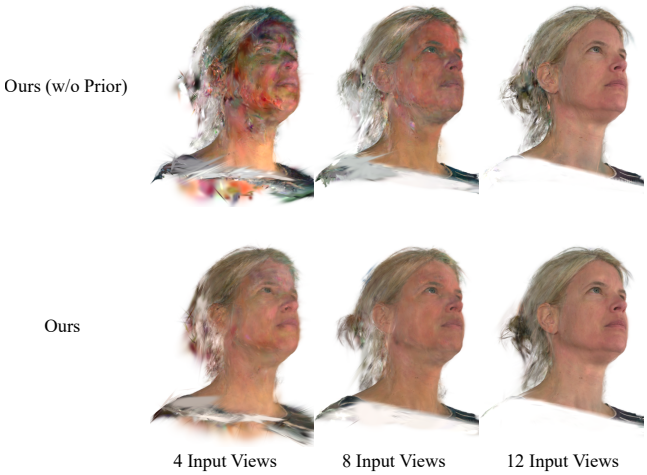

Figure 10: *Qualitative comparison of ablation study on synthesized pseudo views-based Gaussian prior. Introducing the prior reduces overfitting artifacts of head avatars under extreme views. Moreover, the fewer the input views, the more pronounced the prior's effect on head fidelity.*

5.5. Discussion and limitations

While our method demonstrates promising results, it has some limitations that need to be addressed. It is still constrained by the 3DMMs and the expression parameters derived from them. This dependency may introduce challenges in driving reconstructed head avatars for extreme expressions, potentially leading to noticeable artifacts that compromise the visual realism. Additionally, the synthesized pseudo view-based Gaussian prior proposed in our work is inherently limited by the quality of the synthesized data. We believe that refining the per-identity fitting model can help bridge the gap between synthesized data and real video data, ultimately further enhancing the fidelity of the reconstructed head avatars when viewed from uncommon views. Moreover, the introduction of the implicit refinement component and Gaussian prior inevitably incurs additional overhead during training and inference. During the self-reenactment with novel expressions and poses as input, flickering artifacts may occasionally appear in the hair region, mainly due to slight instability in inter-frame prediction. This issue stems from the implicit component, and it is also noticeable in the GHA results. Addressing this limitation requires considering the coherence among frames, and we leave it for future work.

6. Conclusion

In this paper, we present a novel method for creating 3D human heads with mesoscale details and high-fidelity animations from monocular or sparse multi-view videos. By integrating a joint deformation that combines a learnable displacement map-guided Gaussian-triangle binding with a neural-based deformation refinement, we significantly enhance the fidelity of the reconstructed head avatars. In addition, the incorporation of synthesized pseudo

Table 4: *Ablation study on synthesized pseudo views-based Gaussian prior, with the best results in **bold**. We gradually increase the number of training input views in the NeRSemble dataset, using four extreme views for quantitative comparison, both with and without the prior.*

	4 Input Views			8 Input Views			12 Input Views		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Ours (w/o Prior)	15.73	0.694	0.318	20.38	0.802	0.183	22.46	0.835	0.137
Ours	**16.72**	**0.761**	**0.250**	**21.68**	**0.845**	**0.133**	**23.28**	**0.865**	**0.110**

views rendered with fitted textured 3DMMs provides a robust prior, ensuring consistent and realistic appearances across various views. Many promising directions remain for future work. Creating 3D human heads in occluded views will be essential for broadening the applicability of our method in diverse real-world settings. Another interesting potential work is incorporating head-related priors from large language models for head reconstruction and animation.

Acknowledgments. We thank the reviewers for the valuable comments. This work has been partially funded by Nanjing University - China Mobile Communications Group Co., Ltd. Joint Institute and National Natural Science Foundation of China under grant No. 62172220.

References

[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)* (1999), pp. 187–194. 2

[CWL*24] CHEN Y., WANG L., LI Q., XIAO H., ZHANG S., YAO H., LIU Y.: Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers* (2024), pp. 1–9. 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. 2

[DNM*25] DHAMO H., NIE Y., MOREAU A., SONG J., SHAW R., ZHOU Y., PÉREZ-PELLITERO E.: Headgas: Real-time animatable head avatars via 3d gaussian splatting. In *Computer Vision – ECCV 2024* (Cham, 2025), Leonardis A., Ricci E., Roth S., Russakovsky O., Sattler T., Varol G., (Eds.), Springer Nature Switzerland, pp. 459–476. 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. 2, 4

[GKR*24] GIEBENHAIN S., KIRSCHSTEIN T., RÜNZ M., AGAPITO L., NIESSNER M.: Npga: Neural parametric gaussian avatars. *arXiv preprint arXiv:2405.19331* (2024). 3, 5

[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

[Kin14] KINGMA D. P.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 5

[KKLD23] KERBL B., KOPANAS G., LEIMKÜHLER T., DRETTAKIS G.: 3d gaussian splatting for real-time radiance field rendering. *ACM TOG* (2023). 2, 3, 5

[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. 2, 5

[KVB88] KANOPOULOS N., VASANTHAVADA N., BAKER R. L.: Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits 23*, 2 (1988), 358–367. 5

[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. 2, 5

[MESK22] MÜLLER T., EVANS A., SCHIED C., KELLER A.: Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG) 41*, 4 (2022), 1–15. 2

[MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV* (2020). 2

[MWSZ24] MA S., WENG Y., SHAO T., ZHOU K.: 3d gaussian blendshapes for head avatar animation. In *ACM SIGGRAPH 2024 Conference Papers* (2024), pp. 1–10. 2, 3

[PGC*17] PASZKE A., GROSS S., CHINTALA S., CHANAN G., YANG E., DEVITO Z., LIN Z., DESMAISON A., ANTIGA L., LERER A.: Automatic differentiation in pytorch. 5

[QKS*24] QIAN S., KIRSCHSTEIN T., SCHONEVELD L., DAVOLI D., GIEBENHAIN S., NIESSNER M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 20299–20309. 2, 4, 5, 8

[SSL*20] SHANG J., SHEN T., LI S., ZHOU L., ZHEN M., FANG T., QUAN L.: Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *European Conference on Computer Vision* (2020), Springer, pp. 53–70. 2

[SSS*24] SAITO S., SCHWARTZ G., SIMON T., LI J., NAM G.: Relightable gaussian codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 130–141. 2

[TBG*19] TEWARI A., BERNARD F., GARRIDO P., BHARAJ G., ELGHARIB M., SEIDEL H.-P., PÉREZ P., ZOLLHOFER M., THEOBALT C.: Fml: Face model learning from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 10812–10822. 2

[TKG*24] TEOTIA K., KIM H., GARRIDO P., HABERMANN M., ELGHARIB M., THEOBALT C.: Gaussianheads: End-to-end learning of drivable gaussian head avatars from coarse-to-fine representations. *ACM Transactions on Graphics (TOG) 43*, 6 (2024), 1–12. 3

[TZS*16] THIES J., ZOLLHOFER M., STAMMINGER M., THEOBALT C., NIESSNER M.: Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 2387–2395. 2

[VBPP06] VLASIC D., BRAND M., PFISTER H., POPOVIC J.: Face transfer with multilinear models. In *ACM SIGGRAPH 2006 Courses*. 2006, pp. 24–es. 2

[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

[WBSS04] WANG Z., BOVIK A. C., SHEIKH H. R., SIMONCELLI E. P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing 13*, 4 (2004), 600–612. 5

[WXL*23] WANG J., XIE J.-C., LI X., XU F., PUN C.-M., GAO H.: Gaussianhead: High-fidelity head avatars with learnable gaussian derivation. *arXiv preprint arXiv:2312.01632* (2023). 3

[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. 2, 3, 5

[XGGZ24] XIANG J., GAO X., GUO Y., ZHANG J.: Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 1802–1812. 2, 3, 4, 5

[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

[ZBT23] ZIELONKA W., BOLKART T., THIES J.: Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 4574–4584. 2, 5

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 586–595. 5

[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. 2