

Fresco: Frequency–Spatial Consistent Optimization for Fine-Grained Head Avatar Modeling

Shikun Zhang¹ Yong Li^{2*} Yiqun Wang^{2*} Qihong Ke¹ Cunjian Chen¹

¹Department of Data Science and AI, Monash University, Australia

²College of Computer Science, Chongqing University, China

{yongli,yiqun.wang}@cqu.edu.cn, {shikun.zhang,qihong.ke,cunjian.chen}@monash.edu

Abstract

We propose Fresco, a unified optimization pipeline designed to mitigate early over-sharpening, and cross-view drifting in head avatar reconstruction. Fresco combines a Laplacian-pyramid-based frequency curriculum with UV-space consistency regularization to progressively enhance reconstruction quality. The optimization begins by stabilizing low-frequency appearance in the image domain, which suppresses spurious details and promotes reliable convergence. As learning proceeds, consistency across different viewpoints is reinforced through pixel-level alignment on shared UV texture coordinates. Finally, high-frequency components are refined under explicit frequency-band constraints, and seam boundary regularization is applied to preserve local continuity. By optimizing in a frequency- and UV-aligned space, Fresco achieves robust convergence without pseudo high-frequency artifacts and yields consistent, high-fidelity results across views. Experiments on the NeRSemble dataset validate the effectiveness of our design, outperforming previous state-of-the-art methods. Our code is available at: <https://github.com/saralkun/Fresco>.

1. Introduction

Generating animatable human head avatars has long been a central research topic in computer vision and graphics. The ability to render dynamic, view-consistent, photorealistic facial images from arbitrary viewpoints is critical for applications such as gaming, film production, and AR/VR communication [8, 17, 20, 25]. For these scenarios, models must not only deliver high-fidelity visual results but also offer fine-grained control over facial expressions and head poses, while robustly generalizing to unseen pose–expression combinations.

In recent years, significant progress has been made in photorealistic animatable head avatar rendering based on

neural representations such as NeRF [19] and Gaussian fields [10]. However, achieving fine-grained control over facial expressions and head poses remains highly challenging. Specifically, the joint modeling of head geometry, facial expressions, and subtle dynamic details makes the reconstruction process highly under-constrained [15, 24, 30, 35]. As a result, early-stage reconstructions tend to exhibit pseudo high-frequency artifacts, and extreme expressions or complex details, such as wrinkles, intra-oral structures, and hair strands, often lead to cross-view drifting or inconsistency.

NeRF and its variants [1, 19, 29, 37] employ implicit volumetric representations to effectively encode head geometry and appearance. These methods leverage volumetric rendering with multi-view supervision to produce continuous and smooth surfaces, achieving strong cross-view consistency in novel view synthesis [9, 26, 33]. Their main advantage lies in the unified modeling of geometry and texture, which avoids the resolution limitations of explicit mesh or point-based representations. However, NeRF-based approaches typically require long training times and are less suitable for real-time rendering or large-scale data. Moreover, the implicit nature of these representations makes direct control over local expressions or fine dynamic details more challenging.

In contrast, Gaussian or deformable point-based approaches adopt explicit representations that enable more flexible control over facial dynamics and expressions, while supporting efficient or near real-time rendering [4, 23, 28, 31]. These methods directly model head structure and texture, allowing for intuitive manipulation and preservation of high-resolution details. However, explicit representations are highly sensitive to point positions and density, which can lead to seam artifacts and unstable training. Additional constraints or optimization strategies are therefore required to improve cross-view consistency.

To address these limitations, we introduce **Fresco** (Frequency–Spatial Consistent Optimization), a unified optimization framework that enhances cross-view consistency and detail robustness across head representations. At its

*Corresponding authors.

core, Fresco employs a Laplacian-pyramid frequency decomposition that progressively shifts learning from low to high spectral bands, following the curriculum learning principle [2]. This stabilizes early updates and improves structural fidelity. Concretely, we regulate frequency directly in the image domain using a Laplacian-pyramid decomposition implemented with differentiable Gaussian (G_σ) and Difference-of-Gaussian (DoG) filters. This spatially localized formulation is well suited to head avatars, where regions exhibit distinct spectral characteristics, with facial surfaces evolving smoothly under low-frequency regularization while hair and boundary areas benefit from targeted high-frequency refinement. In addition, we enforce cross-view texel alignment in a shared UV space so that frequency progression and geometric correspondence are optimized together, improving view consistency and boundary continuity. Our contributions are summarized below:

- We propose Fresco, which integrates Laplacian-pyramid-based frequency decoupling with UV texel alignment supervision to enhance structural fidelity and cross-view consistency in head avatar reconstruction.
- We develop a frequency-annealing curriculum and an efficient differentiable UV-baking operator that jointly stabilize early optimization, suppress pseudo high-frequency artifacts, and keep training overhead low.
- Experiments on the NeRSemble dataset show that our method produces higher-quality renderings for both novel-view reconstruction and self-reenactment tasks.

2. Related Works

2.1. Representations for Animatable Head Avatars

Reconstructing high-fidelity and animatable head avatar models from multi-view or monocular inputs has long been a central goal in computer vision and computer graphics. Early implicit neural representations, such as NeRF-based head avatars, jointly model geometry and appearance in a continuous volumetric space, achieving photorealistic rendering but lacking explicit controllability [5, 9, 41]. Explicit mesh-based models provide semantic topology and animation compatibility, yet they are constrained by fixed templates and limited capability in recovering fine-grained high-frequency details [6, 14, 18, 39]. More recently, explicit Gaussian-based representations, such as PointAvatar [42] and GaussianAvatar [23], as well as other 3D Gaussian Splatting methods [4, 7, 28, 40], enable efficient rendering and expression-driven animation through dense primitives. However, these approaches still suffer from detail inconsistency and cross-view drifting under free-viewpoint rendering or complex animation scenarios. This is mainly because their per-view photometric supervision lacks explicit geometric correspondence across views, causing each camera to overfit localized appearance [23, 42].

Moreover, Gaussian primitives are optimized independently without enforcing global regularization, which leads to misaligned details and unstable texture transitions under viewpoint changes [7, 10, 38].

To improve geometric controllability and photorealistic appearance, hybrid representations such as mesh-Gaussian structures have attracted increasing attention as a complementary alternative to purely implicit or purely point-based models. Early attempts, including Neural Head Avatars [6] and IMAvatar [41], combine neural textures or implicit fields with mesh templates to enable semantically controllable reconstruction, but fine details often degrade under free-viewpoint rendering. More recent mesh-Gaussian frameworks, such as MeGA [27] and HERA [3], attach Gaussian primitives to mesh vertices to establish explicit correspondences, yielding more faithful expression-driven animation. However, training in these methods typically relies on per-view photometric supervision, similar to early radiance-field pipelines [5], which can lead to unstable gradient propagation in multi-view settings, early over-sharpening, and the accumulation of pseudo high-frequency artifacts. To alleviate these issues, several works introduce spatial or normal regularization, such as GaussianAvatar [23] and PointAvatar [42], which suppress noise and improve consistency. This often comes at the cost of excessive smoothing and slower convergence. Although representation designs have become increasingly mature, training stability and cross-view consistency remain difficult to guarantee under existing learning paradigms.

2.2. Optimization for Stable and Consistent Reconstruction

Beyond structural design, recent studies have also begun to address the aforementioned issues from an optimization perspective. In the frequency domain, Recent works [13, 16, 22, 32, 36] have sought to stabilize optimization through spectral regulation. BARF [16] employs frequency annealing on positional encodings to mitigate coarse-to-fine misalignment in implicit fields, while FreNeRF [36] regularizes spectral amplitude distributions to suppress high-frequency noise. FreGS [38] extends this idea to Gaussian Splatting for general scene reconstruction, by explicitly decomposing rendered images into amplitude and phase spectra, and expanding the effective frequency bands during training to prevent premature over-sharpening. However, these global spectral constraints treat all spatial regions uniformly. The Fourier representation captures only global frequency statistics and ignores spatial locality and geometric correspondence, which limits their ability to ensure cross-view consistency and boundary continuity. In contrast, we regulate frequency in the spatial domain with localized Laplacian filters, preserving position-sensitive details and better respecting geometry. In addition, we align texels

across views in a shared UV space so that frequency progression and spatial correspondence are optimized jointly, improving view consistency and boundary continuity.

3. Method

We propose Fresco, a unified optimization framework for stable and cross-view consistent head modeling. Fig. 1 provides an overview of the overall framework. Rather than modifying the underlying representation, Fresco focuses on training dynamics by jointly scheduling a Laplacian-pyramid-based frequency curriculum with UV-space supervision. Early iterations emphasize low-frequency learning to suppress pixel-driven pseudo high frequencies. A differentiable UV bake enforces per-textel cross-view alignment and seam continuity. High-frequency components are then progressively introduced to recover edge-sensitive details.

3.1. Preliminary

We adopt a hybrid head representation that integrates an explicit facial mesh with a volumetric Gaussian field for hair and peripheral regions. This formulation combines the semantic controllability of mesh-based models with the expressive rendering capacity of 3D Gaussian primitives. Below we summarize the key components that constitute the geometric and appearance foundation of our optimization framework.

The facial region is modeled by a parameterized mesh controlled by shape β , expression ψ , and pose ϕ . To capture subject-specific geometric nuances, a learnable UV displacement map $\hat{G}_d$ is applied on the base mesh to refine local surface details. A disentangled neural texture set $\hat{T} = \{\hat{T}_{di}, \hat{T}_v, \hat{T}_{dy}\}$ is decoded into per-pixel RGB attributes, corresponding to diffuse, view-dependent, and dynamic components, respectively.

Hair and surrounding regions are represented by an anisotropic 3D Gaussian set $\mathcal{G} = \{(\mu_i, \Sigma_i, c_i, \alpha_i)\}_{i=1}^N$, where $\mu_i \in \mathbb{R}^3$ is the center, $\Sigma_i = R_i S_i^2 R_i^\top$ is the covariance with rotation R_i and scale is S_i , c_i the color coefficient modeled by spherical harmonics, and α_i the opacity. Each Gaussian is animated through rigid and non-rigid transformations driven by the mesh pose parameters, enabling coherent motion and view-dependent appearance.

During rendering, the mesh-rasterized facial image and the Gaussian-splat hair field are composited. Given a camera projection $\pi(\cdot)$, the final color at pixel $\mathbf{x}$ is obtained by depth-sorted alpha blending:

$$C(\mathbf{x}) = \sum_i c_i \alpha'_i \prod_{j < i} (1 - \alpha'_j), \quad (1)$$

where α'_i denotes the projected opacity of the i -th Gaussian on the image plane. Photometric supervision is applied against multi-view ground-truth images. This representation provides an explicit and differentiable structural

foundation, enabling the subsequent frequency regularization and UV-space consistency optimization to achieve fine-grained, cross-view consistent reconstruction within a unified framework.

3.2. Frequency Curriculum with Laplacian Pyramids

Pixel-domain supervision in neural head modeling often leads to unstable convergence and pseudo high-frequency artifacts during the early training stages. To address this issue, we introduce a frequency curriculum grounded in Laplacian pyramid decomposition, which progressively schedules the optimization from low- to high-frequency components. This mechanism consists of a low-frequency regularization branch that stabilizes early optimization and a high-frequency enhancement branch that refines fine structures and boundary details, forming a unified frequency-driven training framework.

Low-Frequency Regularization. In the early training phase, the model tends to amplify unstable gradients in local textures before achieving consistent global geometry and color. To stabilize convergence, we constrain the learning process to the smoothed, low-frequency domain. Given the rendered image $\hat{I} \in \mathbb{R}^{H \times W \times 3}$ and the ground-truth image I , both are convolved with a Gaussian kernel $\mathcal{G}_\sigma(\cdot)$ to extract low-frequency components:

$$\hat{I}_{LF} = \mathcal{G}_\sigma(\hat{I}), \quad I_{LF} = \mathcal{G}_\sigma(I), \quad (2)$$

where $\mathcal{G}_\sigma(\cdot)$ denotes Gaussian filtering with standard deviation σ . The low-frequency reconstruction loss is formulated as:

$$L_{LF} = \|\hat{I}_{LF} - I_{LF}\|_1, \quad (3)$$

where $\|\cdot\|_1$ represents the mean absolute error. A temporal weight $\lambda_{LF}(t)$ gradually decreases with iteration t , guiding the model to first capture global structural and chromatic consistency while suppressing early pseudo high-frequency responses.

High-Frequency Enhancement. As training progresses, the model is encouraged to recover high-frequency details such as facial boundaries, wrinkles, and hair texture. We adopt a Difference-of-Gaussian (DoG) filtering operation to extract band-pass responses that emphasize fine structural details:

$$H(\cdot) = \mathcal{G}_{\sigma_1}(\cdot) - \mathcal{G}_{\sigma_2}(\cdot), \quad \sigma_1 < \sigma_2, \quad (4)$$

where σ_1 and σ_2 control the target frequency range. An edge-aware high-frequency loss focuses supervision on regions of significant structural variation:

$$L_{HF}^{edge} = M_{edge} \cdot \|H(\hat{I}) - H(I)\|_1, \quad (5)$$

where $M_{edge} = \text{normalize}(\|\nabla I\|)$ is a normalized edge mask derived from the gradient magnitude ∇I . To prevent

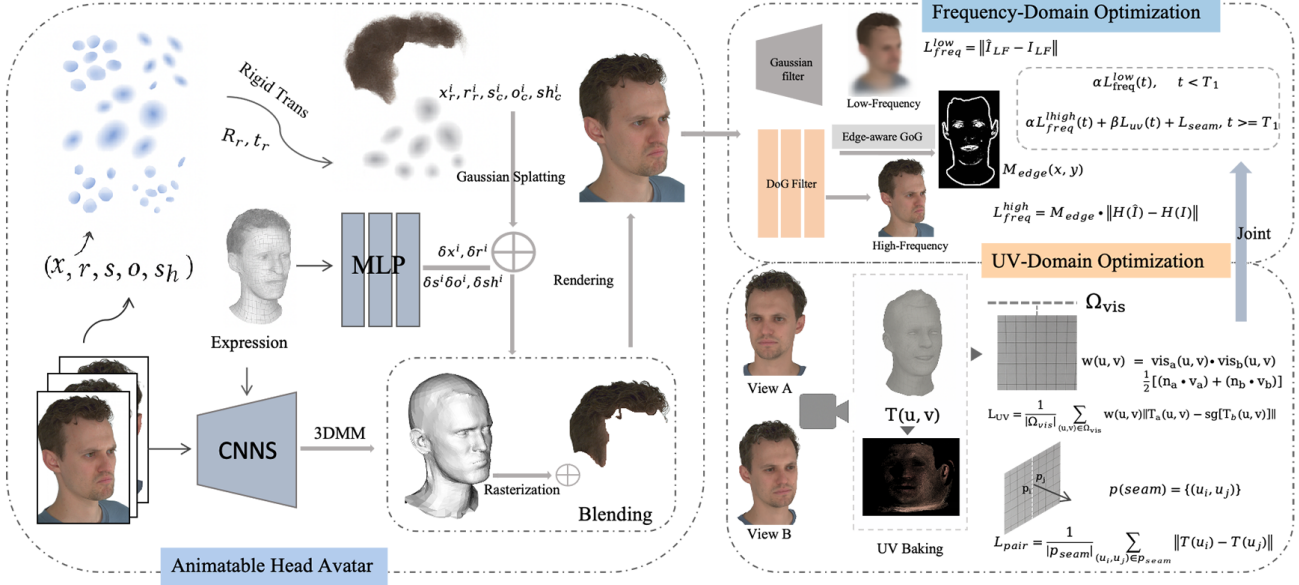


Figure 1. Overview of the proposed Fresco. Multi-view images drive a CNN to estimate 3DMM and an MLP to refine Gaussian hair parameters $\{x, r, s, o, sh\}$. The facial mesh and Gaussian field are rendered and blended to obtain $\hat{I}$. In the *frequency branch*, a Gaussian low-pass stabilizes global structure (low-freq) and a DoG band-pass with edge-aware weighting M_{edge} restores details (high-freq) under progressive scheduling. In the *UV branch*, differentiable UV baking maps multi-view renderings to a shared atlas $T(u, v)$ for cross-view texel alignment within Ω_{vis} , with confidence weights $w(u, v)$ and seam regularization for boundary continuity.

over-sharpening and ensure local smoothness, we introduce a Gradient Difference Loss (GDL):

$$L_{GDL} = \sum_{i,j} \left\| |\nabla \hat{I}_{i,j}| - |\nabla I_{i,j}| \right\|_1, \quad (6)$$

where (i, j) indexes pixel coordinates. $|\nabla I|$ denotes the gradient magnitude computed by finite differences along the x and y directions. The combined high-frequency loss is thus given by:

$$L_{HF} = \lambda_{h1} L_{HF}^{edge} + \lambda_{h2} L_{GDL}, \quad (7)$$

where λ_{h1} and λ_{h2} control the trade-off between edge sharpness and local consistency. The temporal weight $\lambda_{HF}(t)$ gradually increases during training, activating high-frequency supervision once the global structure has stabilized.

Progressive Frequency Curriculum. Rather than two isolated components, the low- and high-frequency branches together define a spectrally continuous optimization paradigm, where supervision bandwidth expands adaptively as training progresses. Inspired by Laplacian pyramid decomposition, the overall frequency-aware loss is expressed as:

$$L_{freq}(t) = \sum_{i=1}^N w_i(t) \|\hat{I}_i - I_i\|_1. \quad (8)$$

where I_i and $\hat{I}_i$ denote the i -th frequency band of the ground-truth and rendered images. In practice, the low-frequency component is assigned a relatively larger weight in the early stage and gradually decayed during training following an empirical schedule. For higher-frequency bands, the weighting coefficient $w_i(t) \in [0, 1]$ controls their progressive activation at iteration t . The frequency weights follow a smooth cosine annealing schedule:

$$w_i(t) = \frac{1}{2} \left[1 - \cos \left(\pi \cdot \frac{t - T_i}{T - T_i} \right) \right], \quad t > T_i, \quad (9)$$

where T_i is the threshold iteration for activating the i -th frequency band, and T is the total number of training iterations. As training advances, higher-frequency components are gradually activated, refining textures and edges. This coarse-to-fine spectral scheduling enables smooth optimization without modifying the underlying representation, effectively balancing convergence stability and detail fidelity.

3.3. UV-Space Consistency and Seam Regularization

Although the frequency curriculum stabilizes optimization in the image domain, subtle inconsistencies across view-points may still emerge due to view-dependent shading and limited multi-view constraints. To address this, we introduce a UV-space optimization paradigm that enforces view-consistent alignment on a shared parameterization and fur-

ther regularizes seam continuity to ensure appearance coherence across the entire surface.

Multi-view UV Consistency. We employ a differentiable UV-baking operator $\mathcal{B}(\cdot)$ that maps rendered images onto the mesh atlas, producing UV textures $\hat{T}^v = \mathcal{B}(\hat{I}^v; \Pi_v, \mathcal{M})$, where Π_v denotes camera parameters and $\mathcal{M}$ the deformable mesh with per-vertex UVs. Unlike per-view pixel supervision, this UV-space projection allows explicit pixel-to-textel correspondence and enables geometry-aware multi-view consistency learning. To construct robust correspondences, we perform visibility-constrained pairing, where only texels visible in both views participate in supervision:

$$L_{UV} = \frac{1}{|\Omega_{vis}|} \sum_{(u,v) \in \Omega_{vis}} w(u,v) \left\| \hat{T}^a(u,v) - \text{sg}[\hat{T}^b(u,v)] \right\|_1, \quad (10)$$

where $\hat{T}^a(u,v)$ and $\hat{T}^b(u,v)$ denote the RGB values of texel (u,v) baked from the rendered images of views a and b , respectively, via the differentiable UV-baking operator $\mathcal{B}(\cdot)$. Here, Ω_{vis} is the mutual-visibility region, $\text{sg}[\cdot]$ is the stop-gradient operator applied to the target branch to prevent bidirectional chasing. The confidence weight $w(u,v) \in [0, 1]$ is computed as:

$$w(u,v) = \text{vis}_a(u,v) \text{vis}_b(u,v) \cdot \frac{1}{2} \left[(\mathbf{n}_a(u,v) \cdot \mathbf{v}_a) + (\mathbf{n}_b(u,v) \cdot \mathbf{v}_b) \right]. \quad (11)$$

Here, $\text{vis}_a(u,v)$ and $\text{vis}_b(u,v)$ are binary visibility masks obtained from differentiable UV baking; $\mathbf{n}_a$ and $\mathbf{n}_b$ denote the surface normals under views a and b , and $\mathbf{v}_a, \mathbf{v}_b$ are the corresponding viewing directions. This weighting emphasizes texels jointly visible from both views and facing the cameras, effectively reducing the impact of self-occluded or grazing-angle regions.

To suppress early-stage noise, we initially down-weight the UV consistency term and gradually increase its contribution as the image-domain optimization stabilizes. Together, these strategies establish a progressive UV-space consistency mechanism that complements the image-domain frequency curriculum: the former aligns cross-view geometry in the parameter domain, while the latter regularizes spectral evolution in the image domain.

Seam Regularization. When a 3D surface is unwrapped into UV space, the atlas are split into multiple charts, inevitably producing discontinuities along their borders. These discontinuities may lead to shading gaps or color inconsistencies between adjacent texels, especially around high-curvature regions such as hair and ear boundaries. To alleviate this issue, we introduce a seam regularization term that explicitly constrains both cross-chart color continuity and local smoothness near the seam area.

$$L_{seam} = \lambda_{pair} L_{seam}^{pair} + \lambda_{tv} L_{seam}^{tv}, \quad (12)$$

where $L_{seam}^{pair} = \frac{1}{N_p} \sum_{i=1}^{N_p} \|T(u_i) - T(v_i)\|_1$ measures the color difference between paired texels (u_i, v_i) lying on two sides of the same 3D surface boundary. The texel pairs $\{(u_i, v_i)\}$ are precomputed from the UV mapping of the mesh by identifying duplicated vertices along chart borders in the topology of the FLAME model. This pairing establishes one-to-one correspondences across seams:

$$L_{seam}^{tv} = \frac{1}{|\Omega_s|} \sum_{(u,v) \in \Omega_s} (|\nabla_u T(u,v)| + |\nabla_v T(u,v)|). \quad (13)$$

The second term is a total-variation regularizer that smooths local fluctuations within a narrow seam band Ω_s , which is automatically dilated from the seam mask by a fixed pixel width. The weighting coefficients λ_{pair} and λ_{tv} are empirically set to balance texel consistency and smoothness.

Joint Objective. The overall UV-space objective is expressed as:

$$L_{UV-total} = \lambda_{uv}(t) L_{UV} + L_{seam}, \quad (14)$$

where $\lambda_{uv}(t)$ follows a staged schedule consistent with the frequency curriculum. In combination, the frequency-based image supervision and geometry-aware UV-space alignment form a dual-domain optimization framework, achieving stable convergence, cross-view consistency, and seamless surface appearance without altering the underlying representation.

3.4. Integration Strategy and Overall Optimization

Building upon the progressive frequency curriculum introduced in Sec. 3.2, the overall optimization in Fresco follows a coarse-to-fine trajectory. In the early stage, the model focuses on stabilizing low-frequency structures to suppress pseudo high-frequency artifacts and ensure geometric consistency. As training proceeds, UV-space consistency is activated to enhance cross-view coherence and guide appearance alignment. In the later stage, high-frequency components are emphasized to recover fine-grained details and improve boundary continuity. Let $t \in [0, T]$ denote the current training iteration and T the total number of iterations. The three components are activated in a staged manner rather than simultaneously. In the first stage ($t < T_1$), the model emphasizes low-frequency reconstruction for stable geometry; in the intermediate stage ($T_1 \leq t < T_2$), UV-space supervision is introduced to enforce cross-view texture alignment; in the final stage ($t \geq T_2$), the high-frequency enhancement branch and seam regularization are enabled to refine structural details and eliminate boundary artifacts. This staged progression naturally transitions the optimization from structural stability, to cross-view coherence, and finally to detail refinement, forming a self-balanced optimization trajectory rather than a static combination of inde-

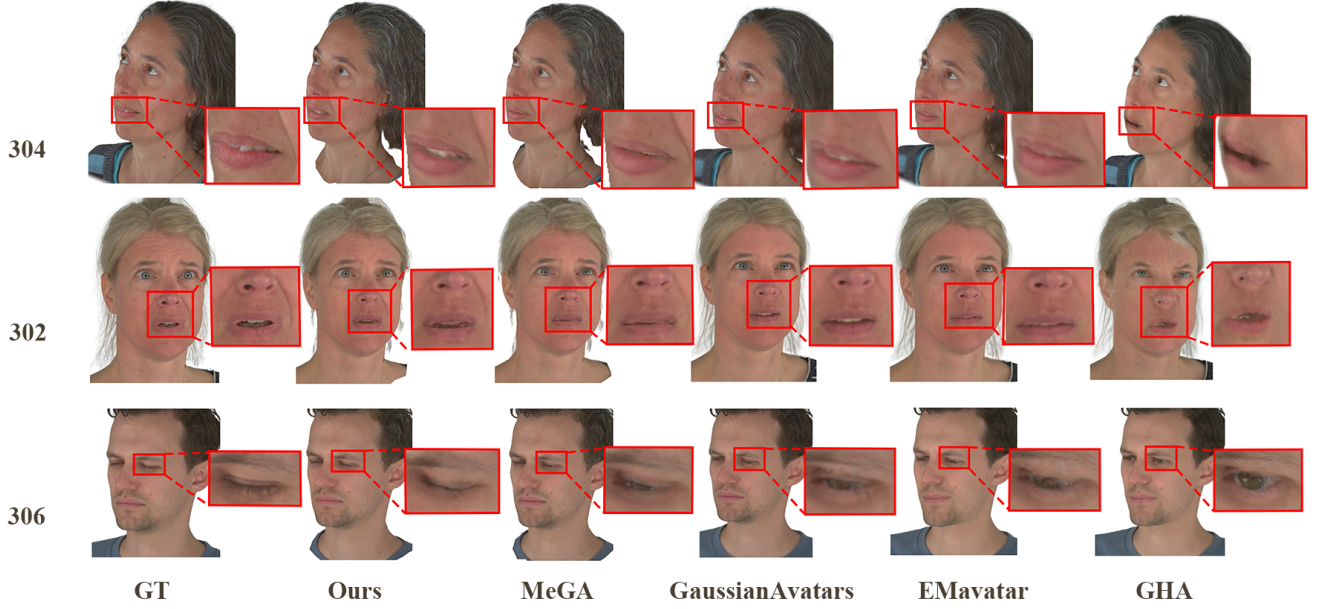


Figure 2. Qualitative comparison on self-reenactment of head avatars. Our method reproduces expressions more faithfully, with accurate mouth opening and eye blinking and clearer tooth details.

pendent losses. Formally, the overall objective can be written as:

$$L_{\text{total}}^{(t)} = \begin{cases} \alpha L_{\text{freq}}^{\text{low}}(t), & t < T_1, \\ \alpha L_{\text{freq}}^{\text{low}}(t) + \beta L_{\text{UV}}(t), & T_1 \leq t < T_2, \\ \alpha L_{\text{freq}}^{\text{high}}(t) + \beta L_{\text{UV}}(t) + L_{\text{seam}}, & t \geq T_2. \end{cases} \quad (15)$$

Here, T_1 and T_2 denote the stage boundaries, and α, β are weight factors. This three-stage schedule first stabilizes low-frequency appearance and geometry, then enforces cross-view consistency, and finally introduces high-frequency refinement, achieving stable convergence across both frequency and view domains without increasing structural complexity. All hyperparameters and hardware configurations remain consistent across experiments, and full details are provided in our supplementary material.

4. Experiments

4.1. Setup

Dataset. We evaluate our approach on the NeRSemble dataset [11], which provides multi-view video recordings of multiple subjects captured under controlled studio environments using sixteen calibrated cameras. The dataset already covers a wide range of identities and expression types, allowing cross-identity generalization to be effectively evaluated within this benchmark. Following the settings of GaussianAvatars [23] and MeGA [27], we select eight subjects for our experiments. Each subject contains

eleven sequences: ten with predefined facial expressions or emotions, and one unscripted sequence. Following the pre-processing procedures in [23, 27], all frames are downsampled to a resolution of 802×550, and per-frame foreground masks are generated. We further obtain facial parsing results using an open-source algorithm [12] and generate per-frame depth maps via Metashape [21]. During the training of Fresco, we keep the same train/test split as [23, 27]: for each subject, nine of the ten predefined sequences and fifteen of the sixteen cameras are used for training, while the remaining sequence and one camera are reserved for evaluation.

Settings. The evaluation protocol follows prior head-avatar works. For novel-view synthesis, we assess generalization to unseen viewpoints by rendering from a camera not used during training and comparing the outputs with the corresponding ground-truth images from that viewpoint. For self-reenactment, we evaluate the ability to generalize to unseen motions of the same subject by driving the avatar with poses and expressions from a held-out sequence and rendering across all sixteen camera views. For cross-identity reenactment, we examine cross-subject retargeting by driving a target avatar with poses and expressions extracted from another subject, rendering under the target subject’s camera configuration. The cross-identity results are shown in our supplementary materials. Following [27], both qualitative and quantitative results are reported, and all quantitative metrics are computed within the same rasterization mask for fair comparison.

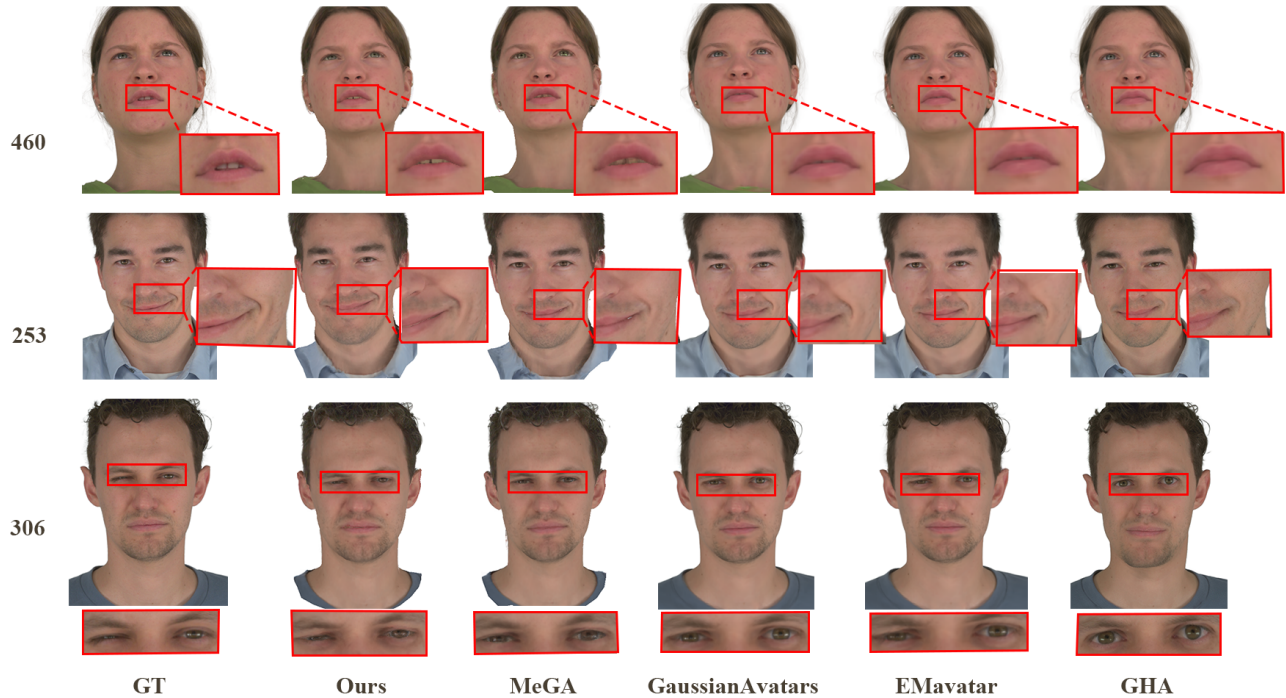


Figure 3. Qualitative comparison on novel-view synthesis of head avatars. Our results show sharper facial wrinkles and clearer teeth across viewpoints, with improved cross-view consistency.

4.2. Comparisons with State-of-the-Art Methods

We compare our Fresco framework with several state-of-the-art methods, including Gaussian Head Avatar [34], EMavatar [40], GaussianAvatars [23], and MeGA [27], to validate its effectiveness in both novel-view synthesis and self-reenactment tasks. All baselines are reimplemented using their official codes and trained under identical settings for fair comparison, and the training details are provided in our supplementary materials. As summarized in Tab. 1, our method consistently achieves the best overall performance, attaining the highest PSNR, and the lowest LPIPS across all evaluation metrics. And as shown in Fig. 2, our method achieves more faithful expression reproduction in the self-reenactment task. For example, the mouth opening in the second row and the eye blinking in the third row are consistent with the ground truth, and the first row shows more accurate preservation of tooth details under expression changes. In Fig. 3, our approach also produces sharper facial wrinkles and clearer teeth for novel-view synthesis compared with prior methods.

4.3. Ablation Study

Following previous head avatar works [23, 27], we chose to perform ablation on one subject (ID 253) to facilitate consistent visual comparison. Tab.2 reports quantitative indicators, while Figs. 4–5 show qualitative results.

Table 1. Quantitative comparison with previous methods on novel-view synthesis and self-reenactment task. Best results are **bold**, and second-best are underlined.

Method	Novel-View			Self-Reenactment		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Gaussian Head Avatar	29.84	0.895	0.083	22.34	0.851	0.145
EMavatar	30.16	0.898	0.082	22.56	0.852	0.145
GaussianAvatars	<u>31.50</u>	<u>0.935</u>	<u>0.060</u>	<u>29.75</u>	0.930	<u>0.066</u>
MeGA	30.79	0.932	0.066	29.62	0.926	0.071
Fresco (Ours)	31.80	0.938	0.058	30.30	<u>0.929</u>	0.064

High-Frequency Enhancement. The high-frequency branch aims to recover fine-grained texture and edge details that are often lost under purely low-frequency supervision. As shown in the second row of Tab.2, removing this branch (w/o $\mathcal{L}_{HF}$) leads to a clear drop across all metrics on both novel-view synthesis and self-reenactment tasks, indicating that high-frequency supervision is crucial for preserving detail fidelity. Qualitative comparisons in Fig. 4 further show that, without the high-frequency path, rendered results appear overly smooth. Facial wrinkles and hair strands become noticeably blurred or missing, confirming the necessity of progressive high-frequency refinement for photorealistic head reconstruction.

Seam Regularization. As reported in Tab.2, removing

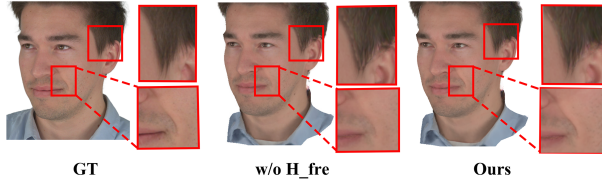


Figure 4. Qualitative comparison of high-frequency modules. Without high-frequency enhancement, the rendered result exhibits missing local details, with blurred facial wrinkles and hair textures. The complete model restores fine textures, achieving a more realistic reconstruction effect.

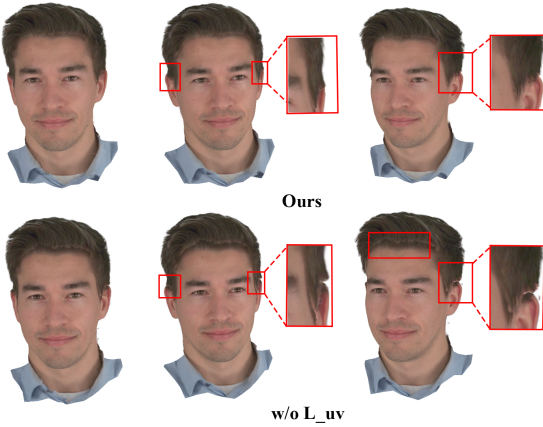


Figure 5. Cross-view comparison with and without UV supervision. w/o UV shows view-dependent artifacts near ear and hair-line, w/ UV remains coherent.

the seam term (w/o seam) produces only minor changes in global quantitative metrics on novel-view synthesis and self-reenactment. This outcome is expected since pixel-wise measures such as PSNR and SSIM are dominated by overall reconstruction accuracy and are relatively insensitive to localized artifacts around UV seams. In contrast, Fig. 6 presents clear qualitative gains in surface continuity. Without seam regularization, visible cracks and color discontinuities arise near view-sensitive regions such as the ear and hair boundaries, whereas the full model yields smooth and consistent transitions across these areas.

UV Consistency. The UV consistency term aims to align cross-view appearance within a unified texture coordinate space, thereby suppressing view-dependent drift and appearance discontinuities caused by occlusion changes during training. As shown in Tab.2, removing this term (w/o $\mathcal{L}_{UV}$) leads to a noticeable degradation in quantitative metrics for novel-view synthesis, confirming that UV consistency plays a crucial role in maintaining cross-view coherence. Similarly, consistent performance drops are observed in the self-reenactment task, further demonstrating

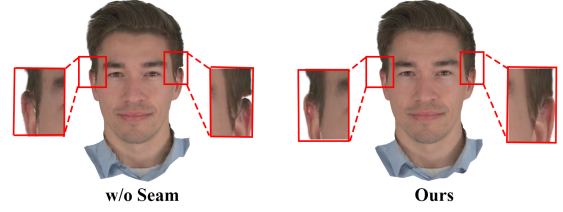


Figure 6. Qualitative comparison of seam regularization. Left w/o seam shows visible cracks and color discontinuities along ear and hair boundaries. Right w/ seam produces smoother and more consistent boundary transitions.

Table 2. Ablation study on subject #253. We validated the effectiveness of each module through two tasks: Novel-View and Self-Reenactment.

Method	Novel-View			Self-Reenactment		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Ours	34.02	0.957	0.036	32.46	0.949	0.043
w/o $\mathcal{L}_{HF}$	32.60	0.953	0.044	31.11	0.944	0.052
w/o Seam	<u>33.86</u>	<u>0.956</u>	<u>0.038</u>	<u>32.29</u>	0.950	<u>0.043</u>
w/o $\mathcal{L}_{UV}$	32.71	0.951	0.041	31.47	0.944	0.050
w/o $\mathcal{L}_{LF}$	33.48	0.952	0.039	32.07	0.947	0.045

that this mechanism not only improves generalization to unseen viewpoints but also enhances appearance stability under unseen motions and expressions.

Fig. 5 presents qualitative comparisons of cross-view rendering. Without UV consistency, even though the seam regularization alleviates visible cracks around texture boundaries (first column), the rendered appearances still exhibit inconsistencies and boundary artifacts under cross-view rendering (second and third columns). Typical cases include local discontinuities, and overlapping artifacts around the ears and hairline regions. With UV consistency enabled, the appearance remains coherent across views, yielding more stable texture alignment and smoother boundary transitions during viewpoint changes. More results are shown in our supplementary materials.

5. Conclusion

This work approaches head avatar reconstruction from an optimization standpoint, showing that a joint treatment in the spectral and UV domains can substantially reduce early instability and cross-view inconsistencies. Our studies and ablations indicate that deferring high-frequency learning mitigates pseudo-detail accumulation and training oscillation, while pixel-level alignment and boundary continuity in a shared UV coordinate space improve multi-view coherence and local transitions. Without altering the underlying representation, the framework achieves stable convergence, consistent quantitative and perceptual gains.

Acknowledgment

This work is partly supported by the National Natural Science Foundation of China (52238003, 62202076) and the Project of Beijing Welinkirt DaoAI Technologies Co., Ltd. We also thank the NVIDIA Academic Hardware Grant Program for its support.

References

- [1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021. 1
- [2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 2
- [3] Hongrui Cai, Yuting Xiao, Xuan Wang, Jiafei Li, Yudong Guo, Yanbo Fan, Shenghua Gao, and Juyong Zhang. Hybrid explicit representation for ultra-realistic head avatars. *arXiv preprint arXiv:2403.11453*, 2024. 2
- [4] Helisa Dharmo, Yinyu Nie, Arthur Moreau, Jifei Song, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Headgas: Real-time animatable head avatars via 3d gaussian splatting. In *European Conference on Computer Vision*, pages 459–476. Springer, 2024. 1, 2
- [5] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias 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
- [6] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18653–18664, 2022. 2
- [7] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2024. 2
- [8] Zhenyi He, Ruofei Du, and Ken Perlin. Collabovr: A reconfigurable framework for creative collaboration in virtual reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 542–554. IEEE, 2020. 1
- [9] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong 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. 1, 2
- [10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2
- [11] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 6
- [12] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5549–5558, 2020. 6
- [13] Ru Li, Jia Liu, Guanghui Liu, Shengping Zhang, Bing Zeng, and Shuaicheng Liu. Spectralnerf: Physically based spectral rendering with neural radiance field. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3154–3162, 2024. 2
- [14] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2
- [15] Zhanfeng Liao, Yuelang Xu, Zhe Li, Qijing Li, Boyao Zhou, Ruifeng Bai, Di Xu, Hongwen Zhang, and Yebin Liu. Hhavatar: Gaussian head avatar with dynamic hairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1
- [16] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5741–5751, 2021. 2
- [17] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. 1
- [18] Xin Lu, Chuanqing Zhuang, Chenxi Jin, Zhengda Lu, Yiqun Wang, Wu Liu, and Jun Xiao. Lsf-animation: Label-free speech-driven facial animation via implicit feature representation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–12, 2025. 2
- [19] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1
- [20] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th annual symposium on user interface software and technology*, pages 741–754, 2016. 1
- [21] Jin-Si R Over, Andrew C Ritchie, Christine J Kranenburg, Jenna A Brown, Daniel D Buscombe, Tom Noble, Christopher R Sherwood, Jonathan A Warrick, and Phillipe A Wernet. Processing coastal imagery with agisoft metashape professional edition, version 1.6—structure from motion workflow documentation. Technical report, US Geological Survey, 2021. 6
- [22] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5865–5874, 2021. 2

- [23] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias 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. [1](#), [2](#), [6](#), [7](#)
- [24] Minghan Qin, Yifan Liu, Yuelang Xu, Xiaochen Zhao, Yebin Liu, and Haoqian Wang. High-fidelity 3d head avatars reconstruction through spatially-varying expression conditioned neural radiance field. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4569–4577, 2024. [1](#)
- [25] Zach Waggoner. *My avatar, my self: Identity in video role-playing games*. McFarland, 2009. [1](#)
- [26] Cong Wang, Di Kang, Yan-Pei Cao, Linchao Bao, Ying Shan, and Song-Hai Zhang. Neural point-based volumetric avatar: Surface-guided neural points for efficient and photorealistic volumetric head avatar. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. [1](#)
- [27] Cong Wang, Di Kang, Heyi Sun, Shenhan Qian, Zixuan Wang, Linchao Bao, and Song-Hai Zhang. Mega: Hybrid mesh-gaussian head avatar for high-fidelity rendering and head editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26274–26284, 2025. [2](#), [6](#), [7](#)
- [28] Jie Wang, Jiu-Cheng Xie, Xianyan Li, Feng Xu, Chi-Man Pun, and Hao Gao. Gaussianhead: High-fidelity head avatars with learnable gaussian derivation. *IEEE Transactions on Visualization and Computer Graphics*, 2025. [1](#), [2](#)
- [29] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. [1](#)
- [30] Yating Wang, Xuan Wang, Ran Yi, Yanbo Fan, Jichen Hu, Jingcheng Zhu, and Lizhuang 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. [1](#)
- [31] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1802–1812, 2024. [1](#)
- [32] Shuxiang Xie, Shuyi Zhou, Ken Sakurada, Ryoichi Ishikawa, Masaki Onishi, and Takeshi Oishi. G 2 f r: Frequency regularization in grid-based feature encoding neural radiance fields. In *European Conference on Computer Vision*, pages 186–203. Springer, 2024. [2](#)
- [33] Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. Avatarmav: Fast 3d head avatar reconstruction using motion-aware neural voxels. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–10, 2023. [1](#)
- [34] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2024. [7](#)
- [35] Yuelang Xu, Zhaoqi Su, Qingyao Wu, and Yebin Liu. Gphm: Gaussian parametric head model for monocular head avatar reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. [1](#)
- [36] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8254–8263, 2023. [2](#)
- [37] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in neural information processing systems*, 34:4805–4815, 2021. [1](#)
- [38] Jiahui Zhang, Fangneng Zhan, Muyu Xu, Shijian Lu, and Eric Xing. Fregs: 3d gaussian splatting with progressive frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21424–21433, 2024. [2](#)
- [39] Shikun Zhang, Fengyi Song, Ge Song, and Ming Yang. Sdr-net: Shape decoupled regression network for 3d face reconstruction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. [2](#)
- [40] Shikun Zhang, Cunjian Chen, Yiqun Wang, QiuHong Ke, and Yong Li. Eavatar: Expression-aware head avatar reconstruction with generative geometry priors. *arXiv preprint arXiv:2508.13537*, 2025. [2](#), [7](#)
- [41] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13545–13555, 2022. [2](#)
- [42] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar 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](#)