Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SRP-Avatar: Monocular 3D Head Avatar Reconstruction with Semi-Rigid Geometric Constraints and Pose Refinement

Xiaolin Ma
Zheren Gou
Hailan Kuang
kuanghailan@whut.edu.cn

Xinhua Liu

Research Article

Keywords: 3D Gaussian Splatting, Monocular Head Avatar, Pose Refinement, Semi-Rigid Constraint, Physical Consistency

Posted Date: March 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8794903/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SRP-Avatar: Monocular 3D Head Avatar Reconstruction with Semi-Rigid Geometric Constraints and Pose Refinement

Xiaolin Ma[1], Zheren Gou[1], Hailan Kuang[1*], Xinhua Liu[1]

[1]Hubei Key Laboratory of Broadband Wireless Communication and Sensor Networks , School of Information Engineering, Wuhan University of Technology , Wuhan , 430070 , Hubei , China.

*Corresponding author(s). E-mail(s): kuanghailan@whut.edu.cn;
Contributing authors: maxiaolin0615@whut.edu.cn;
288143@whut.edu.cn; liuxinhua@whut.edu.cn;

Abstract

Recent advancements in 3D Gaussian Splatting (3DGS) have established a strong benchmark for dynamic head avatar reconstruction. However, achieving improved physical consistency and local photorealism remains challenging, as inherent tracking jitter often blurs high-frequency details, and the complex oral topology frequently leads to non-physical geometric distortions. Furthermore, standard shading networks struggle to accurately disentangle material properties for regions with distinct optical characteristics, such as eyes and teeth. To address these limitations, we propose a 3D Gaussian head avatar framework that explicitly targets pose jitter, intra-oral geometric instability, and specular degradation. We counteract tracking instability by integrating a Perceptually-Guided Adaptive Pose Refinement module. This module incorporates LPIPS-driven gradients into the pose optimization loop, enhancing texture sharpness and reducing noise-induced blur. It does so by refining the pose parameters based on perceptual quality, to achieve more accurate and stable head geometry. Addressing anatomical plausibility, we redesign the deformation field with an Anatomically-Adaptive Semi-Rigid Constraint, coupling rigid mandibular transformations with soft skinning to mitigate geometric collapse in the oral cavity. Complementing these geometric improvements, a Semantically-Aware Specular Injection mechanism is introduced during the rendering stage, leveraging semantic priors to recalibrate material attributes for enhanced specularity in eyes and teeth. Experiments on INSTA and HDTF show improved perceptual quality and more stable intra-oral geometry compared with prior 3DGS head avatar baselines.

Keywords: 3D Gaussian Splatting, Monocular Head Avatar, Pose Refinement, Semi-Rigid Constraint, Physical Consistency

1 Introduction

The creation of photorealistic, animatable digital head avatars from monocular video sequences is a longstanding pursuit in computer vision and computer graphics, with extensive applications in telepresence, virtual reality, and movie production. Traditional approaches often relied on explicit textured meshes or parametric models like 3DMM [1], which offer robust control but frequently lack fine-scale geometric details. The advent of Neural Radiance Fields (NeRF) [2] revolutionized the field by introducing implicit representations that achieve photorealistic rendering. Early deformable NeRF methods, such as Nerfies **[3]** and HyperNeRF **[4]**, demonstrated the capability to handle non-rigid deformations, paving the way for head-specific methods [5–9].

However, the high computational cost of implicit sampling limits the real-time applicability of NeRF-based methods. Recently, 3D Gaussian Splatting (3DGS) [10] has emerged as a compelling alternative, combining the explicit nature of point clouds with differentiable rasterization to achieve real-time rendering speeds and high-quality visuals. Following this paradigm, recent works including GaussianAvatars [11], MonoGaussianAvatar [12], and Flash-avatar [13] have successfully integrated 3D Gaussians with parametric face models (e.g., FLAME [14]). By attaching Gaussians to the canonical space of the parametric model, these methods leverage the underlying geometry to drive facial deformations while learning residual details to compensate for the model's limited expressiveness. Additionally, approaches like Gaussian Blendshapes [15] have explored semantic control. State-of-the-art frameworks, represented by HRAvatar [16] and GaussianHead [17], have established a high benchmark in this domain by incorporating hybrid representations and physically-based shading to handle complex deformations and lighting.

Despite these significant advancements, achieving micro-scale fidelity and strict physical consistency from monocular videos remains a persistent challenge. While existing methods appear plausible at a macroscopic level, specific artifacts become apparent when scrutinized under high-definition conditions. A primary obstacle is the inevitable tracking jitter inherent to monocular settings. As discussed in BARF [18], inaccuracies in camera poses can severely blur high-frequency details. In the context of avatars, slight misalignments in the canonical space force the network to smooth out features like skin pores and hair strands to minimize reconstruction loss. Beyond tracking instability, the geometric modeling of rigid anatomical structures presents another difficulty. Widely used Linear Blend Skinning (LBS) [11, 19] treats the oral cavity as a continuous deformable surface. As noted in IMAvatar **[20]**, standard implicit surfaces often struggle to represent the topological changes inside the mouth, and similarly, Gaussian-based LBS often causes rigid structures like teeth to undergo unrealistic stretching or geometric collapse during large jaw movements. Furthermore,

current PBR pipelines often struggle with material disentanglement, resulting in visually flat rendering for complex regions like eyes and teeth, which fail to exhibit the vivid specularity and wetness inherent to biological tissues [21].

To address these issues, we propose a refined 3D Gaussian head avatar framework that prioritizes physical correctness and perceptual sharpness. To counteract tracking noise, we introduce a Perceptually-Guided Adaptive Pose Refinement strategy. Inspired by the joint optimization of pose and geometry in NeRF– [22], we achieve sub-pixel alignment by embedding LPIPS-driven gradients [23], effectively recovering sharp high-frequency details. To ensure anatomical validity, drawing inspiration from the geometric constraints in GeoAvatar [24], we propose an Anatomy-Guided Rigid Motion Constraint. By strictly locking the motion of teeth Gaussians to the jaw bone transformation, we effectively mitigate non-physical deformations within the oral cavity. Finally, to further enhance visual realism, we incorporate a Semantically-Aware Specular Injection mechanism during the rendering phase, explicitly boosting the specular attributes of eyes and teeth to restore their natural vitality. Our main contributions are summarized as follows:

- We develop a joint optimization framework that combines Perceptually-Guided Adaptive Pose Refinement with perceptual loss to mitigate tracking jitter, effectively preserving high-frequency textural details.
- We implement an Anatomically-Adaptive Semi-Rigid Constraint that enforces rigid motion for the oral cavity, ensuring geometric stability and alleviating stretching artifacts during large expressions.
- We propose a Semantically-Aware Specular Injection strategy that enhances the specular rendering of eyes and teeth, improving the overall material plausibility of the avatar.

2 Related Work

2.1 Efficient Neural Head Avatars

The evolution of digital head reconstruction has transitioned from traditional explicit meshes to implicit neural representations, and recently, to hybrid primitives. Early approaches largely relied on 3D Morphable Models (3DMMs) [1, 14], which offer robust semantic control but are limited by the fixed topology and low resolution of the underlying mesh. The advent of Neural Radiance Fields (NeRF) [2]revolutionized the field by modeling the head as a continuous volumetric field, enabling photorealistic rendering of complex details like hair [20, 25]. However, the implicit nature of NeRFs imposes a severe computational bottleneck: rendering a single image requires volumetric ray-marching, which necessitates querying a coordinate-based Multi-Layer Perceptron (MLP) hundreds of times per pixel. This heavy computational load restricts most NeRF-based methods from achieving real-time performance.

To overcome these limitations, 3D Gaussian Splatting (3DGS) [10] has emerged as a paradigm shift, combining the efficiency of rasterization with the flexibility of point-based rendering. Unlike NeRFs, 3DGS represents the scene with explicit anisotropic

Gaussians, allowing for real-time inference without costly neural network evaluations during rendering. Recent state-of-the-art methods, such as GaussianAvatars [11], MonoGaussianAvatar [12], Flash-avatar [13]and HRAvatar [16], integrate this representation with FLAME [14] to enable animatable avatars. Despite their efficiency, these methods face a new challenge in monocular settings: sensitivity to tracking noise. While implicit fields naturally smooth out high-frequency jitter, the explicit nature of 3D Gaussians tends to overfit to noisy camera poses [12, 18], resulting in visual artifacts that require specific refinement strategies. In this paper, we address this issue by integrating a Perceptually-Guided Adaptive Pose Refinement strategy directly into the optimization loop, aiming to recover high-frequency details from noisy monocular tracking.

2.2 Geometry-Aware Facial Animation

To achieve controllable animation, prevalent 3DGS-based frameworks typically drive the scene using Linear Blend Skinning (LBS) derived from the underlying parametric model. Since the coarse FLAME mesh lacks fine geometric details, methods like HeadGaS [26] and GaussianAvatars [11] introduce learnable position offsets or neural deformation fields to capture non-rigid dynamics such as facial wrinkles and subtle expressions. However, a critical limitation in these hybrid representations is the uniform treatment of heterogeneous facial tissues. Most existing pipelines operate under the assumption that the entire head is a continuous deformable surface governed by smooth skinning weights. This assumption fails for anatomically rigid structures, particularly within the oral cavity. Because the skinning weights are typically smoothed to ensure seamless transitions on the skin, teeth Gaussians driven by these weights are often erroneously interpolated during large mandibular movements. This leads to non-physical stretching or "rubber-like" distortions, where the teeth appear to deform elastically rather than moving as a rigid body. Although recent attempts like Hugs [27] employ geometry-guided deformations to better align Gaussians with the mesh surface, they generally lack explicit constraints to enforce strict rigidity for the upper and lower dental structures, resulting in geometric inconsistencies and texture sliding artifacts during extreme expressions. To mitigate these geometric inconsistencies, we propose an anatomy-guided constraint that enforces strict rigid motion for the oral cavity, ensuring geometric stability even during extreme expressions.

2.3 Relightable Appearance Modeling

Decomposing appearance into intrinsic physical attributes (albedo, roughness, normal) from monocular video is essential for photorealistic relighting but remains a challenging inverse problem. For static scenes, implicit neural representations [28, 29] have successfully integrated BRDF models to enable relightable reconstruction. However, extending these methods to dynamic head avatars faces greater difficulties due to the lack of multi-view constraints. While recent point-based approaches like Relightable 3D Gaussian [19] and GaussianShader [30] have introduced physically based rendering (PBR) into the 3DGS pipeline, they often struggle with the ambiguity between lighting and material properties in monocular settings. This limitation is particularly

evident in complex facial regions such as eyes and teeth, where specular highlights are frequently baked into the albedo or overly smoothed, resulting in a flat appearance. In this work, rather than attempting a fully physically accurate decomposition, we adopt a semantic-guided inference strategy to improve perceptual realism. By utilizing semantic masks to explicitly modulate the reflectance parameters of highly specular regions during the rendering phase, we empirically enhance the visibility of specular highlights, mitigating the dullness often observed in purely learning-based decompositions.

3 Method

As mentioned, previous methods for head reconstruction are plagued by blurring artifacts caused by tracking jitter, geometric collapse of intra-oral structures due to non-physical deformations, and the struggle to render realistic specular highlights under novel lighting. To tackle these challenges, we first propose a Perceptually-Guided Adaptive Pose Refinement strategy to compensate for tracking noise via learnable residuals (Sec. 3.1). Subsequently, we introduce an Anatomically-Adaptive Semi-Rigid Constraint to strictly couple oral geometry with skeletal transformations (Sec. 3.2). Finally, a Semantic-Aware Specular Injection strategy is employed during inference to enforce physical plausibility for high-specular regions, enabling realistic relighting (Sec. 3.3). The overall pipeline is illustrated in Fig. 1.

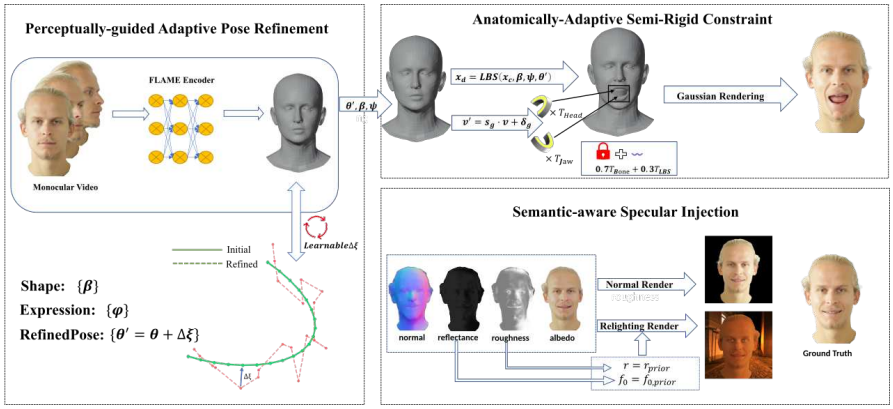


Fig. 1 Structure of SRP-Avatar. The framework comprises three core components tailored for high-fidelity head reconstruction. To mitigate monocular tracking jitter, the **Perceptually-Guided Adaptive Pose Refinement** module optimizes a learnable pose residual. This refines the camera trajectory from a noisy initial estimation into a stable path, while jointly optimizing shape and expression parameters. The **Anatomically-Adaptive Semi-Rigid Constraints** aim to prevent intra-oral geometric collapse. We employ a hybrid weighting strategy to explicitly couple the rigid bone transformation with soft skinning, enforcing physical plausibility for the teeth. Finally, the **Semantic-Aware Specular Injection** strategy disentangles PBR attributes (normal, roughness, reflectance, albedo). During inference, physical priors are injected to modulate the rendering equation, recovering realistic specular highlights for the eyes and teeth under novel lighting conditions.

3.1 Perceptually-Guided Adaptive Pose Refinement

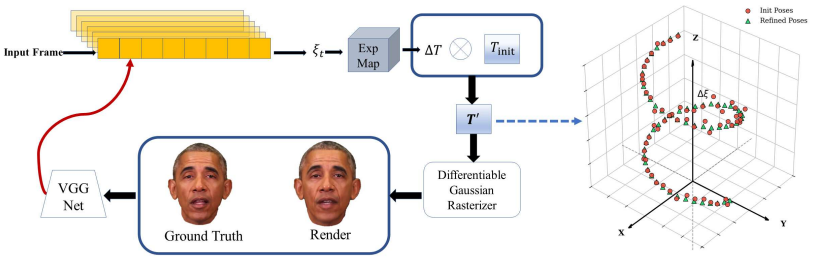


Fig. 2 Perceptually-Guided Adaptive Pose Refinement. To correct the jittery initial pose input, we introduce a learnable embedding ξ_t that generates a residual matrix ΔT via an exponential mapping. This correction is composited with the noisy input to produce the refined pose T', supervised by a VGG-based perceptual loss. As shown in the 3D visualization, the gradient-steered optimization effectively smooths the camera trajectory, ensuring temporal stability and structural consistency.

Structure-Aware Pose Optimization. In monocular settings, tracking errors inevitably propagate to the Gaussian distribution optimization, resulting in the blurring of high-frequency details [18]. While generic view-consistent deblurring strategies have been recently explored in Gaussian splatting [31], the specific micro-jitter in facial tracking remains a challenge that traditional photometric losses (e.g., L1, SSIM) struggle to rectify due to their insensitivity to structural misalignment. To address this, we introduce a Learnable Pose Correction Module structurally integrated into the 3DGS pipeline Figure 2. Instead of treating camera extrinsics as immutable constants, we reformulate the trajectory as a learnable manifold optimized alongside the canonical geometry.

We instantiate a learnable embedding layer $\mathcal{E}_{\text{pose}} = \{\xi_t\}_{t=1}^{M}$, where each frame t is assigned a 6-DoF residual vector $\xi_t = [\omega_t, t_t] \in \mathbb{R}^6$. Here, ω_t represents the axis-angle rotational deviation and t_t denotes the translational offset. This layer acts as a differentiable adapter that intercepts the noisy tracker output $[R_t|T_t]$ and injects precise corrections. The refined camera pose $[R'_t|T'_t]$ is synthesized via a Differentiable Lie Algebra Mapping:

$$\begin{aligned} R'_t &= \exp\left([\boldsymbol{\omega}_t]_\times\right) \cdot R_t \\ T'_t &= \exp\left([\boldsymbol{\omega}_t]_\times\right) \cdot T_t + \mathbf{t}_t \end{aligned} \tag{1}$$

where $[\cdot]_\times$ denotes the skew-symmetric operator converting a vector to a matrix, and $\exp(\cdot)$ applies the Rodrigues' rotation formula. By embedding this operation directly into the computational graph, gradients can flow from the rendering output back to the camera parameters, enabling end-to-end geometric calibration [22].

Perceptual Gradient Steering. A core challenge in pose refinement is that traditional pixel-wise metrics (e.g., L_1) often fall into local minima, favoring blurry

averages over sharp alignment. To drive our correction module towards sub-pixel accuracy, we implement a Perceptual Feedback Mechanism utilizing the LPIPS metric [23] .Unlike pixel-level supervision, this mechanism extracts deep features from a pre-trained VGG network ϕ [32] , forcing the optimization to align high-level structural patterns (such as teeth boundaries and hair strands). The loss function is formulated as:

$$\mathcal{L}_{\text{perc}} = \sum_{l} \frac{1}{H_l W_l} \left\| w_l \odot \left(\phi^l(I_{\text{rend}}) - \phi^l(I_{gt}) \right) \right\|_2^2 \tag{2}$$

where ϕ^l represents features from a pre-trained VGG network. Gradients from $\mathcal{L}_{\text{perc}}$ are back-propagated to the pose parameters ξ_t. This strategy enables the model to adaptively compensate for tracking errors, significantly improving both sharpness and temporal stability.Crucially, we utilize these gradients to explicitly steer the update of the pose embeddings. This effectively transforms the optimization landscape, allowing the network to sense and correct structural misalignments that are invisible to standard photometric losses.

Optimization Objective. The overall training objective $\mathcal{L}_{\text{total}}$ is a weighted combination of photometric reconstruction, perceptual guidance, and geometric regularization. It is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{photo}} + \lambda_{pr}\mathcal{L}_{pr} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} \tag{3}$$

The photometric term $\mathcal{L}_{\text{photo}}$ combines L1 distance and Structural Dissimilarity (D-SSIM):

$$\mathcal{L}_{\text{photo}} = (1 - \lambda_{\text{dssim}})\mathcal{L}_1 + \lambda_{\text{dssim}}\mathcal{L}_{D\text{-SSIM}} \tag{4}$$

To recover high-frequency details from noisy tracking, we incorporate a **masked** perceptual loss $\mathcal{L}_{pr}$. We compute the LPIPS metric using a VGG network solely within the eroded foreground mask M_{in} to avoid boundary artifacts:

$$\mathcal{L}_{pr} = \text{LPIPS}(I_{\text{rend}} \odot M_{in}, I_{gt} \odot M_{in}) \tag{5}$$

Finally, to prevent the learnable pose residuals from overfitting to large deviations, we apply an L2 regularization $\mathcal{L}_{\text{pose}} = \|\Delta\xi\|_2^2$. Empirically, we set the weights as $\lambda_{pr} = 0.05$ and $\lambda_{\text{pose}} = 0.01$.

3.2 Anatomically-Adaptive Semi-Rigid Constraint

Standard Formulation and Limitations. After resolving the pose jitter issue, the geometric distortion of 3D Gaussian Splatting (3DGS) is mainly concentrated in rigid structures such as the oral cavity. Thus, on the basis of pose optimization, we further design a semi-rigid constraint module . Formally, the deformation from canonical space to pose space is modeled via linear blendshapes ($\mathcal{BS}$) and linear blend skinning ($\mathcal{LBS}$). A canonical vertex $\boldsymbol{v}_c$ is typically synthesized by superimposing shape and expression offsets onto the template $\overline{\boldsymbol{T}}$:

$$\boldsymbol{v}_c = \overline{\boldsymbol{T}} + \mathcal{BS}(\boldsymbol{\beta}, \mathcal{S}) + \mathcal{BS}(\boldsymbol{\psi}, \mathcal{E}) \tag{6}$$

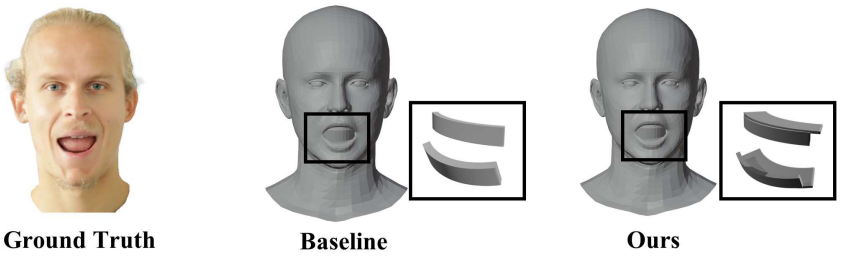


Fig. 3 Visualization of mouth structure. The baseline approach employs LBS to deform teeth as elastic skin, resulting in volume collapse and non-physical artifacts during extreme jaw articulation. To address this, we introduce the Anatomically-Adaptive Semi-Rigid Constraint to enforce structural integrity. Our method incorporates comprehensive supporting tissues, specifically the palate and jaw base, to construct a rigid volumetric foundation that effectively eliminates distortion problems.

where $\mathcal{BS}(\boldsymbol{c}, \mathcal{B}) = \sum c_m \mathcal{B}_m$ represents the linear combination of basis vectors. Subsequently, the final vertex $\boldsymbol{v}_p$ is transformed into pose space via LBS:

$$\boldsymbol{v}_p = \mathcal{LBS}(\boldsymbol{v}_c, \mathcal{J}, W) = R_{\text{LBS}} \boldsymbol{v}_c + T_{\text{LBS}} \tag{7}$$

While this formulation effectively models elastic skin deformation, it is physically ill-posed for rigid intra-oral structures. The expression term $\mathcal{BS}(\boldsymbol{\psi}, \mathcal{E})$ introduces non-physical stretching to the teeth, while the smooth weights in $\mathcal{LBS}$ cause rubber-like distortion during jaw articulation, as illustrated in Figure 3. To resolve these issues, drawing inspiration from GeoAvatar [24] ,we propose the Anatomically-Adaptive Semi-Rigid Constraint.

Semantic Decoupling and Spatial Calibration. We first segment the primitives into upper and lower dental groups. Recognizing that dental geometry is anatomically invariant, we explicitly nullify the expression blendshapes ($\mathcal{BS}(\boldsymbol{\psi}, \mathcal{E}) \to 0$) for these regions. Addressing the potential misalignment between the generic template and the subject's anatomy, we introduce learnable spatial parameters for each group $g \in \{\text{up}, \text{low}\}$: a scaling factor s_g and a translation vector $\boldsymbol{\delta}_g$. The calibrated vertex $\boldsymbol{v}'$ replaces the standard canonical form:

$$\boldsymbol{v}' = s_g \cdot \left(\overline{\boldsymbol{T}} + \mathcal{BS}(\boldsymbol{\beta}, \mathcal{S}) - \text{center}_g\right) + \text{center}_g + \boldsymbol{\delta}_g \tag{8}$$

This formulation removes elastic artifacts while allowing the rigid geometry to adaptively resize and shift for precise static alignment.

Hybrid Transformation Integration. To handle dynamic motion, we address the trade-off between structural rigidity and tracking tolerance. Pure LBS creates distortions, while a strict rigid bone transformation $\mathcal{T}_{\text{bone}}$ is susceptible to shearing artifacts caused by tracking jitter. We propose a Hybrid Transformation Integration mechanism. We synthesize a final transformation matrix $\mathcal{T}_{\text{final}}$ by linearly interpolating between the strict rigid bone transformation and the original smoothed LBS transformation $\mathcal{T}_{\text{LBS}}$ (derived from $R_{\text{LBS}}, T_{\text{LBS}}$):

$$\mathcal{T}_{\text{final}} = \lambda \cdot \mathcal{T}_{\text{bone}} + (1 - \lambda) \cdot \mathcal{T}_{\text{LBS}} \tag{9}$$

The rigidity anchoring coefficient λ is empirically set to 0.7. The dominant weight from $\mathcal{T}_{\text{bone}}$ serves as a rigid anchor, enforcing the kinematic constraints of the mandible to prevent volume collapse. Simultaneously, the residual 30% influence from $\mathcal{T}_{\text{LBS}}$ provides necessary spatial tolerance to compensate for high-frequency tracking noise, ensuring robust motion fluidity.

3.3 Semantic-Aware Specular Injection

The Ambiguity of Monocular Decomposition. While our framework successfully reconstructs geometry, disentangling intrinsic material properties—specifically albedo, roughness, and specular reflectance—from monocular video remains a fundamentally ill-posed inverse problem. Without multi-view photometric cues, self-supervised networks inevitably suffer from parameter ambiguity [21], often "baking" specular highlights directly into the albedo texture. This phenomenon leads to severe artifacts in highly specular biological regions, such as the eyes and teeth, rendering them dull, foggy, or plastic-like under novel illumination conditions.

Physically-Grounded Specular Enhancement. To restore the dynamic vitality of biological tissues, we propose a Semantic-Aware Specular Enhancement strategy that incorporates domain-specific physical priors into the inference pipeline. To bypass the uncertainty of network predictions for high-frequency specular areas, we explicitly model the dielectric properties of human biological tissues. We utilize semantic parsing masks [33] to identify specific semantic regions $\mathcal{M}_{\text{sem}}$ (i.e., eyes and teeth). For these regions, we override the network-predicted parameters by injecting physically-calibrated priors:

$$r_i = r_{\text{prior}}, \quad f_{0,i} = f_{0,\text{prior}}, \quad \forall i \in \mathcal{M}_{\text{sem}} \tag{10}$$

$$f_{0,\text{prior}} = \left(\frac{n-1}{n+1}\right)^2 \tag{11}$$

where r_{prior} and $f_{0,\text{prior}}$ denote the pre-calibrated physical constants specific to each tissue type.

Specifically, for the eye region, we adopt the refractive index of the cornea ($n_{\text{eye}} \approx 1.376$). Based on the Fresnel equations, this yields a base reflectance $f_{0,\text{eye}} \approx 0.025$. Furthermore, we enforce a near-zero roughness ($r_{\text{eye}} \approx 0.03$) to account for the fluid surface tension that maintains an optically smooth liquid interface. This effectively creates a "wet" surface response, eliminating milky or foggy artifacts. Similarly, for the teeth region, we reference the optical properties of dental enamel ($n_{\text{enamel}} \approx 1.63$), resulting in a theoretical reflectance $f_{0,\text{tooth}} \approx 0.058$. We assign a low roughness ($r_{\text{tooth}} \approx 0.08$) to model the dense, polished crystalline structure of the enamel surface, preventing the rubber-like appearance common in baseline methods.

Crucially, this strategy acts as strong physical guidance during the physically-based rendering phase. While the underlying albedo may retain residual static lighting, our explicit specular term serves to significantly boost the dynamic range and responsiveness of the highlights, preventing the dead eye effect observed in baselines.

4 Experiments

4.1 Experimental Setup

Implementation Details. We implement our framework in PyTorch [34] and optimize the model using the Adam optimizer [35] on a single NVIDIA 3090 GPU. For data preprocessing and initialization, we strictly adhere to the protocols established in HRAvatar [16]. Specifically, all video sequences are cropped and resized to a resolution of 512×512. To isolate the avatar, we apply matting to extract the foreground and set the background to black. Regarding initialization, we utilize pre-tracked FLAME parameters and leverage pre-trained weights from SMIRK [36] for the encoder. The model is trained for 15 epochs per subject.

Datasets and Preprocessing. We conducted a comprehensive evaluation on two publicly available benchmarks: the INSTA dataset [37] and the HDTF dataset [38]. Specifically, we utilized 10 subjects from the INSTA dataset, which provides pre-processed sequences. Following standard protocols, we held out the last 350 frames of each video for self-reenactment testing. To further assess robustness, we included 8 subjects collected from the internet-based HDTF dataset, with the final 500 frames reserved for testing. Consistent with baselines, all sequences underwent identical cropping and background segmentation. Additionally, to facilitate our Semantic-Aware specular injection, we employed a pre-trained face parsing network (BiSeNet [33]) to generate pixel-wise semantic masks for all frames, specifically extracting the eye and oral cavity (including teeth) regions to guide the material attribute optimization.

Baselines and Metrics. We conduct a comprehensive comparison with several state-of-the-art methods, including INSTA [37], Point-Avatar [39], Flash-Avatar [40], 3D Gaussian Blendshapes (GBS) [41], and our direct baseline HRAvatar [16]. Following the official evaluation protocol of HRAvatar, all baseline methods are compared using the quantitative results re-implemented and reported in HRAvatar under identical data splits and rendering resolution. We evaluate SRP-Avatar using the same evaluation pipeline to ensure a fair and consistent comparison. Standard image quality metrics, including PSNR, SSIM, and LPIPS, are used for quantitative evaluation. In addition, we report MAE^* (scaled by 10^2) to measure pixel-level reconstruction accuracy.

4.2 Evaluation

Quantitative Results. We evaluate self-reenactment performance on both the INSTA and HDTF datasets, with results summarized in Table 1. As shown, our method outperforms the baselines, achieving improvements in both LPIPS and PSNR scores. The reduction in LPIPS is particularly significant, indicating that our Perceptually-Guided Adaptive Pose Refinement module effectively mitigates rendering blur caused by high-frequency tracking jitter, resulting in sharper visual details. Regarding structural and pixel-level metrics (SSIM and MAE), our method maintains performance on par with the HRAvatar baseline. This demonstrates that our Anatomically-Adaptive Semi-Rigid Constraint successfully corrects local artifacts (e.g., texture misalignment

in the oral cavity) and enhances perceptual realism without compromising the global structural fidelity of the reconstruction.

Method	INSTA Dataset				HDTF Dataset			
	PSNR↑	MAE*↓	SSIM↑	LPIPS↓	PSNR↑	MAE*↓	SSIM↑	LPIPS↓
INSTA	27.85	1.309	0.9110	0.1047	25.03	2.333	0.8475	0.1614
Point-avatar	26.84	1.549	0.8970	0.0926	25.14	2.236	0.8385	0.1278
Flash-avatar	29.13	1.133	0.9255	0.0719	27.58	1.751	0.8664	0.1095
GBS	29.64	1.020	0.9294	0.0823	27.81	1.601	0.8915	0.1297
HRAvatar	<u>30.36</u>	<u>0.845</u>	<u>0.9482</u>	<u>0.0569</u>	<u>28.55</u>	**1.373**	<u>0.9089</u>	<u>0.0825</u>
SRP-Avatar (Ours)	**30.60**	**0.833**	**0.9494**	**0.0464**	**28.81**	<u>1.340</u>	**0.9107**	**0.0717**

Table 1 Quantitative Performance Comparison on the INSTA and HDTF Datasets. Baseline results for INSTA, Point-Avatar, Flash-Avatar, GBS, and HRAvatar are taken from the HRAvatar evaluation, where all methods are re-implemented and evaluated under identical data splits and rendering settings [16]. ↑ indicates that higher values are better, while ↓ indicates that lower values are better. **Bold** denotes the best result and <u>underline</u> denotes the second-best result.

Visual Comparisons The visual comparison of our method with the HRAvatar baseline on self-reenactment is shown in Figure 4. While HRAvatar generates generally plausible results, it often struggles with high-frequency tracking jitter, resulting in blurring artifacts around detailed areas such as hair strands and skin textures. This is largely attributed to the tracking lag inherent in monocular pose estimation, which the baseline fails to fully compensate for. In contrast, our method, aided by Pose Refinement, effectively aligns these temporal inconsistencies, restoring sharper textures and cleaner details. Regarding geometric accuracy, HRAvatar tends to exhibit instability in the oral cavity during challenging poses or extreme expressions (e.g., wide mouth openings), occasionally leading to floating teeth artifacts. Our method handles these scenarios more effectively; the Semi-Rigid Constraint ensures the teeth remain rigidly anchored to the jaw, preserving precise geometric structure even under large deformations. We also qualitatively compare the visual differences in relighting. As shown in Figure 5 , due to the entangled modeling of texture and lighting, HRAvatar often reconstructs eyes with a dull, diffuse appearance under novel lighting conditions. In contrast, by incorporating Physics-based Material Injection during the relighting process, our method successfully synthesizes realistic eye specularity and vivid reflections, leading to more consistent and lifelike lighting effects.

4.3 Ablation Studies

To rigorously validate the individual contribution of each proposed module to the overall reconstruction fidelity, we conducted a comprehensive ablation study on the INSTA dataset. We benchmarked our full SRP-Avatar framework against three degraded variants: w/o Pose Refinement, w/o Semi-Rigid Constraint (reverting to standard learnable LBS), and w/o Material Injection. The quantitative performance gaps and visual comparisons are detailed in Table 2 and Figure 6.

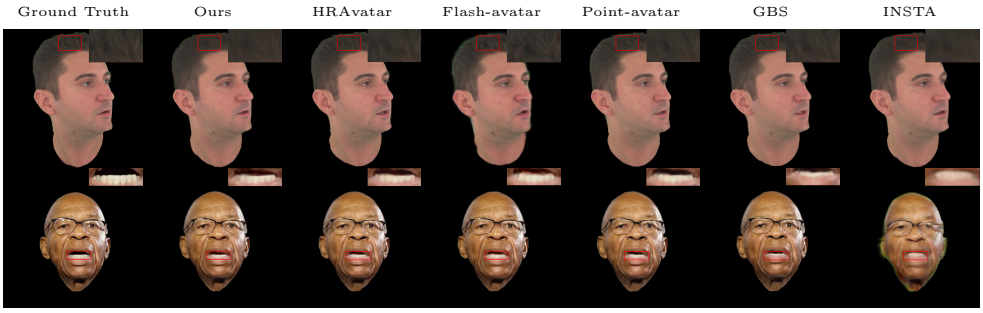


Fig. 4 Qualitative comparison of key facial details (e.g., hair and teeth) among Ground Truth, baselines, and our method. All baseline results are obtained from our own re-implementations under the same input conditions. Red boxes highlight regions where our method exhibits improved fine-grained texture fidelity compared to baseline methods. Notably, our approach recovers sharper hair strands and skin details by mitigating tracking blur, while preserving precise dental geometry via the semi-rigid constraint.

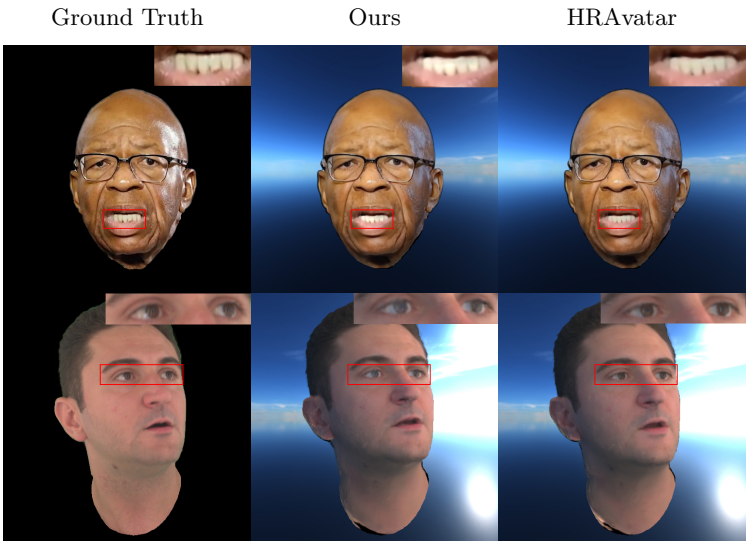


Fig. 5 Qualitative comparison of relighting results under frontal illumination. In the eye region, baseline methods exhibit dull reflections, whereas our Semantic-Aware Specular Injection synthesizes vivid and physically plausible specular highlights. Similarly, regarding the teeth, our approach effectively disentangles material properties to recover realistic dental enamel reflections, contrasting sharply with the visually flat shading produced by the baselines.

w/o Pose Refinement. To evaluate the necessity of our trajectory optimization strategy, we disabled the Perceptually-Guided Adaptive Pose Refinement module. In this configuration, the 3DGS optimization is strictly conditioned on the fixed, noisy camera extrinsics provided by the off-the-shelf tracker. As evidenced inTable 2 , removing this module precipitates a marked decline in perceptual quality, with the LPIPS score increasing significantly. Mechanistically, when the camera poses are inaccurate, the Gaussian splats are forced to blur themselves to minimize the photometric

error across misaligned frames, leading to an over-smoothed representation. Visual inspection confirms that without refinement, high-frequency features—particularly hair strands and skin pores—suffer from severe ghosting and blurring artifacts caused by the accumulated temporal tracking jitter. This underscores the critical role of our learnable pose manifold in resolving the structural ambiguity inherent in monocular sequences.

w/o Semi-Rigid Constraint. We assess the role of anatomical constraints by reverting the dental deformation model to standard Linear Blend Skinning (LBS). In this baseline setup, the movement of the teeth is driven by unconstrained soft skinning weights rather than a rigid mandibular anchor. Consequently, vertices in the oral cavity are frequently assigned erroneous non-zero weights from the surrounding lips or cheeks. Quantitative results show a distinct drop in PSNR, indicating compromised geometric fidelity. Qualitatively, this formulation fails to preserve the volumetric rigidity of the dentition. During extreme mouth expressions (e.g., wide opening), the teeth undergo non-physical elastic deformation, appearing to stretch like rubber or suffering from texture sliding artifacts. Our semi-rigid constraint effectively eliminates these distortions by enforcing a kinematic coupling between the dental geometry and the jaw bone.

w/o Material Injection. We verified the efficacy of our material injection strategy for relighting. Without this explicit inference-time prior, the network fails to disentangle specular properties. This causes the eye region to exhibit a dull, diffuse-only appearance under novel illumination. Our strategy explicitly modifies the reflectance parameters of the eyes and teeth during inference. This approach effectively synthesizes physically plausible specular highlights (e.g., eye glints) and enhances realism in relighting scenarios

	PSNR↑	MAE*↓	SSIM↑	LPIPS↓
Baseline	30.36	0.845	0.9482	0.0569
w/o Pose Refinement	<u>30.55</u>	0.854	0.9426	0.0574
w/o Semi-Rigid Constraint	30.41	<u>0.841</u>	<u>0.9486</u>	<u>0.0482</u>
Full(Ours)	**30.60**	**0.833**	**0.9494**	**0.0464**

Table 2 Quantitative ablation study of core components. We evaluate the contribution of each module on the test set. "w/o" denotes the removal of a specific component. ↑ indicates higher is better, while ↓ indicates lower is better. **Bold** marks the best results, and <u>underline</u> marks the second best results.

5 Discussion

Conclusion. In this paper, we have presented a framework for high-fidelity 3D head avatar reconstruction from monocular videos, aiming to enhance the preservation of fine-grained details and geometric stability. First, to address the instability of intra-oral structures often observed in existing methods, we introduced an Anatomically-Adaptive Semi-Rigid Constraint, which decouples dental geometry from soft skin

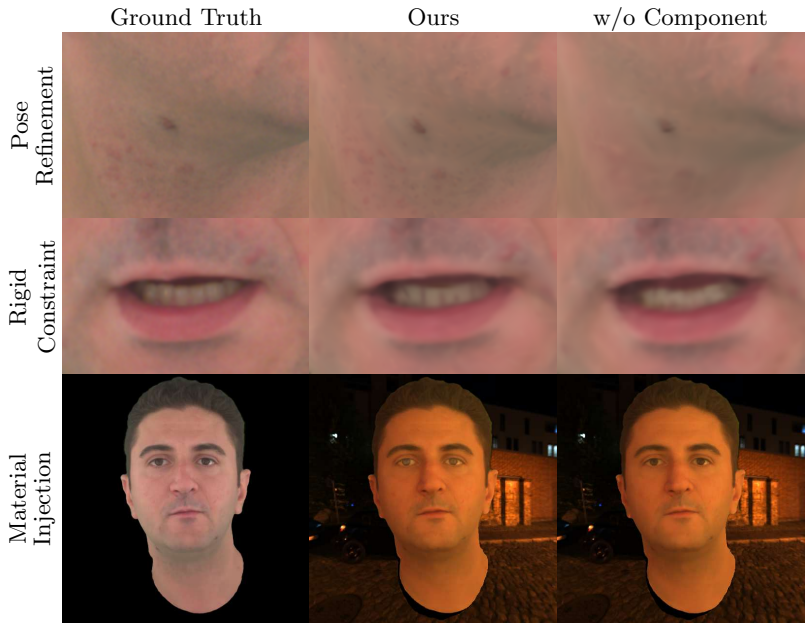


Fig. 6 Qualitative ablation study of core components. The columns represent (from left to right): the ground truth image, our full model, and the baseline result with the specific component removed. Regarding Pose Refinement, the baseline suffers from texture blurring due to misalignment, whereas our method recovers sharp high-frequency skin details (e.g., pores and moles). In terms of the Rigid Constraint, our method maintains the structural integrity and clear boundaries of the teeth, addressing the blurry and distorted geometry produced by the baseline. Finally, to demonstrate the necessity of Material Injection, we evaluate a relighting scenario with frontal illumination; our method synthesizes view-dependent specular highlights, improving visual realism compared to the flat, diffuse-only appearance of the baseline.

deformations. Second, we proposed a Perceptually-Guided Adaptive Pose Refinement strategy. By optimizing camera trajectories, we effectively mitigate micro-jitter, recovering sharper high-frequency textures in hair and skin. Finally, to enhance realism under novel lighting, we incorporated a Semantic-Aware Specular Injection module during inference. Experiments on the INSTA and HDTF datasets demonstrate that our method achieves competitive visual quality, particularly in maintaining structural integrity and texture sharpness.

Limitations. While our approach leverages the FLAME topology to ensure robust facial kinetics, it remains constrained by this geometric prior when modeling highly dynamic non-facial elements; consequently, features like long loose hair or complex accessories that do not follow the underlying head motion may exhibit reconstruction artifacts. Similarly, although our deformation field handles most articulations, extreme expressions deviating significantly from the training distribution can still pose challenges, potentially leading to minor geometric inconsistencies. Regarding appearance modeling, due to the additive nature of our inference-time material injection, the superposition of physical highlights onto an albedo with residual baked-in lighting may occasionally lead to pixel saturation or compounding brightness artifacts in

highly specular regions. Furthermore, the disentanglement of complex global illumination effects, such as cast shadows from hair or self-occlusions, remains simplified, suggesting avenues for future physics-based simulations.

References

[1] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 296–301 (2009)

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

[3] Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5865–5874 (2021)

[4] Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. arXiv preprint arXiv:2106.13228 (2021)

[5] Qin, M., Liu, Y., Xu, Y., Zhao, X., Liu, Y., Wang, H.: High-fidelity 3d head avatars reconstruction through spatially-varying expression conditioned neural radiance field. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 4569–4577 (2024)

[6] Xu, Y., Wang, L., Zhao, X., Zhang, H., Liu, Y.: Avatarmav: Fast 3d head avatar reconstruction using motion-aware neural voxels. In: ACM SIGGRAPH 2023 Conference Proceedings, pp. 1–10 (2023)

[7] Xu, Y., Zhang, H., Wang, L., Zhao, X., Huang, H., Qi, G., Liu, Y.: Latentavatar: Learning latent expression code for expressive neural head avatar. In: ACM SIGGRAPH 2023 Conference Proceedings, pp. 1–10 (2023)

[8] Zhao, X., Wang, L., Sun, J., Zhang, H., Suo, J., Liu, Y.: Havatar: High-fidelity head avatar via facial model conditioned neural radiance field. ACM Transactions on Graphics **43**(1), 1–16 (2023)

[9] Yang, Q., Weng, D., Liu, Y.: Utilizing periodic feature-enhanced neural-field modeling for the photorealistic representation of human head avatars. The Visual Computer **40**(12), 9129–9140 (2024)

[10] Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for

real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)

[11] 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, pp. 20299–20309 (2024)

[12] 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, pp. 1–9 (2024)

[13] 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, pp. 1802–1812 (2024)

[14] 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), 194–1 (2017)

[15] Ma, S., Weng, Y., Shao, T., Zhou, K.: 3d gaussian blendshapes for head avatar animation. In: ACM SIGGRAPH 2024 Conference Papers, pp. 1–10 (2024)

[16] Zhang, D., Liu, Y., Lin, L., Zhu, Y., Chen, K., Qin, M., Li, Y., Wang, H.: Hravatar: High-quality and relightable gaussian head avatar. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 26285–26296 (2025)

[17] Wang, J., Xie, J.-C., Li, X., Xu, F., Pun, C.-M., Gao, H.: Gaussianhead: High-fidelity head avatars with learnable gaussian derivation. IEEE Transactions on Visualization and Computer Graphics (2025)

[18] Lin, C.-H., Ma, W.-C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5741–5751 (2021)

[19] Gao, J., Gu, C., Lin, Y., Li, Z., Zhu, H., Cao, X., Zhang, L., Yao, Y.: Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In: European Conference on Computer Vision, pp. 73–89 (2024). Springer

[20] 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, pp. 13545–13555 (2022)

[21] Guo, Y., Chen, K., Liang, S., Liu, Y.-J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5784–5794 (2021)

[22] Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V.A.: Nerf–: Neural radiance

fields without known camera parameters. arXiv preprint arXiv:2102.07064 (2021)

[23] 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, pp. 586–595 (2018)

[24] Moon, S., Lew, H.M., Lee, S., Kang, J.-S., Park, G.-M.: Geoavatar: Adaptive geometrical gaussian splatting for 3d head avatar. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12811–12821 (2025)

[25] 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, pp. 8649–8658 (2021)

[26] 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: European Conference on Computer Vision, pp. 459–476 (2024). Springer

[27] Kocabas, M., Chang, J.-H.R., Gabriel, J., Tuzel, O., Ranjan, A.: Hugs: Human gaussian splats. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 505–515 (2024)

[28] Boss, M., Braun, R., Jampani, V., Barron, J.T., Liu, C., Lensch, H.: Nerd: Neural reflectance decomposition from image collections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12684–12694 (2021)

[29] Zhang, K., Luan, F., Wang, Q., Bala, K., Snavely, N.: Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5453–5462 (2021)

[30] Jiang, Y., Tu, J., Liu, Y., Gao, X., Long, X., Wang, W., Ma, Y.: Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5322–5332 (2024)

[31] Zhang, Y., *et al.*: Vc-gs: view-consistent deblurring gaussian splatting via alternating branch optimization. The Visual Computer **41**(10), 7303–7317 (2025)

[32] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[33] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 325–341 (2018)

[34] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

[35] Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[36] Retsinas, G., Filntisis, P.P., Danecek, R., Abrevaya, V.F., Roussos, A., Bolkart, T., Maragos, P.: 3d facial expressions through analysis-by-neural-synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2490–2501 (2024)

[37] Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4574–4584 (2023)

[38] Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3661–3670 (2021)

[39] 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, pp. 21057–21067 (2023)

[40] 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, pp. 1802–1812 (2024)

[41] Ma, S., Weng, Y., Shao, T., Zhou, K.: 3d gaussian blendshapes for head avatar animation. In: ACM SIGGRAPH 2024 Conference Papers, pp. 1–10 (2024)