HRAvatar: High-Quality and Relightable Gaussian Head Avatar

Dongbin Zhang[1,2*] Yunfei Liu[2] Lijian Lin[2] Ye Zhu[2] Kangjie Chen[2]
Minghan Qin[1] Yu Li[2†] Haoqian Wang[1†]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]International Digital Economy Academy (IDEA)

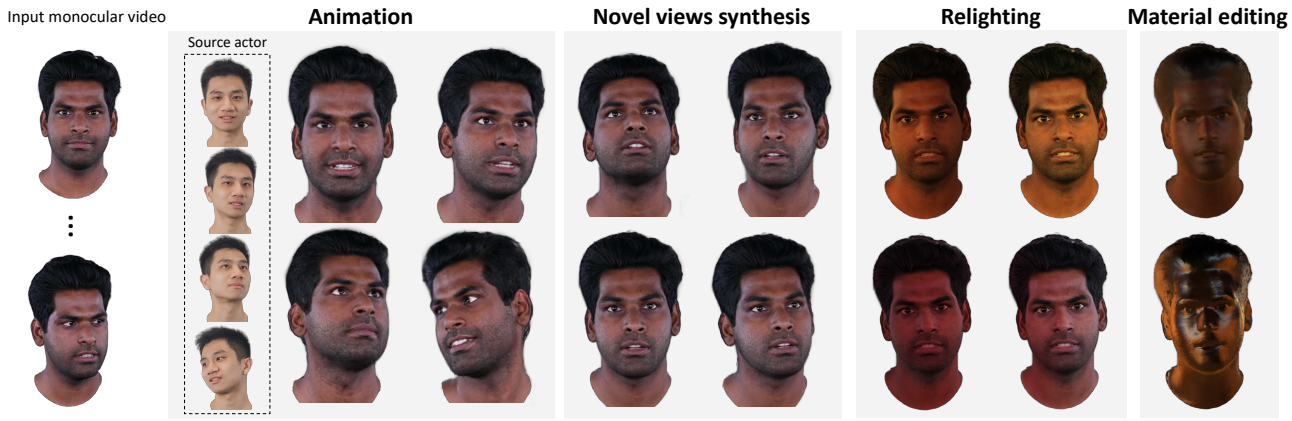

Figure 1. With monocular video input, HRAvatar reconstructs a high-quality, animatable 3D head avatar that enables realistic relighting effects and simple material editing.

Abstract

Reconstructing animatable and high-quality 3D head avatars from monocular videos, especially with realistic relighting, is a valuable task. However, the limited information from single-view input, combined with the complex head poses and facial movements, makes this challenging. Previous methods achieve real-time performance by combining 3D Gaussian Splatting with a parametric head model, but the resulting head quality suffers from inaccurate face tracking and limited expressiveness of the deformation model. These methods also fail to produce realistic effects under novel lighting conditions. To address these issues, we propose HRAvatar, a 3DGS-based method that reconstructs high-fidelity, relightable 3D head avatars. HRAvatar reduces tracking errors through end-to-end optimization and better captures individual facial deformations using learnable blendshapes and learnable linear blend skinning. Additionally, it decomposes head appearance into several physical properties and incorporates physically-based shading to account for environmental lighting. Extensive experiments demonstrate that HRAvatar not only reconstructs superior-quality heads but also achieves realistic visual effects under varying lighting conditions. Video results and code are available at the project page.

1. Introduction

Creating a 3D head avatar is essential for film, gaming, immersive meetings, AR/VR, etc. In these applications, the avatar must meet several requirements: animatable, real-time, high-quality, and visually realistic. However, achieving a highly realistic and animatable head avatar from widely-used monocular video remains challenging.

Research in this area spans many years. Early efforts [7, 35, 48] develop parametric head models based on 3D Morphable Models (3DMM) theory [3]. These methods allow registering 3D head scans to parametric models for 3D facial mesh reconstruction. With the rise of deep learning, methods [10, 17, 41, 79] use parametric model priors to simplify head mesh reconstruction from videos, either through estimation or frame-wise optimization, *i.e.*, 3D face tracking. While these methods generalize well for expressions and pose variations, their fixed topology limits complex hair modeling and fine-grained appearance reconstruction. To

* Intern at IDEA. † Corresponding authors.

address this issue, some researchers have turned to Neural Radiance Fields (NeRF) [45] for modeling head avatars [51, 60, 61, 76]. These approaches enable complete geometry and appearance reconstruction, including hair, glasses, earrings, *etc*. However, they are limited by slow rendering and long training time. Recently, 3D Gaussian Splatting (3DGS) [30] has gained significant attention for its fast rendering speed. Some methods [15, 54, 59, 69] have extended 3DGS to head avatar reconstruction, significantly improving rendering speed compared to NeRF-based methods.

Although previous 3DGS-based methods have made progress in animatability and real-time rendering, their reconstruction quality is constrained by two major factors: **limited deformation flexibility** and **inaccurate expression tracking**. Additionally, they are **unable to produce realistic relighting effects**. Specifically, our motivation primarily stems from the following three points. **1)** Head reconstruction requires a geometric model to deform from the compact canonical space to various states based on different expressions and poses. Recent methods [54, 59] model geometric deformations of Gaussian points by rigging them to universal parametric model mesh faces. However, parametric models may not accurately capture personalized deformations. **2)** Before training, these methods extract FALME parameters by fitting pseudo-2D facial keypoints, which are usually error-prone and lead to suboptimal results. Methods like PointAvatar [78] try to directly optimize these parameters during training. Such a design may introduce a mismatch from pre-tracked parameters and limit generalization to new expressions and poses. Consequently, such methods still require post-optimization during testing. **3)** Under monocular and unknown lighting settings, existing 3DGS-based methods directly fit the colors of the avatar, causing an inability to relight and mix the person's intrinsic appearance with ambient lighting.

To tackle the aforementioned challenges, we propose HRAvatar, which utilizes 3D Gaussian points for high-quality head avatar reconstruction with realistic relighting from monocular videos, as Fig. 1. We propose a learnable blendshapes and learnable linear blend skinning strategy, allowing the Gaussian points for flexible deformation from canonical space to pose space. Additionally, we utilize an expression encoder to extract accurate facial expression parameters in an end-to-end training manner, which not only reduces the impact of tracking errors on reconstruction but also ensures the generalization of expression parameters estimation. To achieve realistic and real-time relighting, we model the head's appearance by using albedo, roughness, Fresnel reflectance, *etc*. with an approximate physically-based shading model. An albedo pseudo-prior is also employed to better decouple the albedo. For a detailed comparison and distinction from previous methods, please refer to the supporting materials. Benefiting from these techniques, HRAvatar can reconstruct fine-grained and expressive avatars while achieving realistic relighting effects.

In summary: **a)** We present HRAvatar, a method for monocular reconstruction of head avatars using 3D Gaussian points. HRAvatar leverages learnable blendshapes and learnable linear blend skinning for flexible and precise geometric deformations, with a precise expression encoder reducing tracking errors for high-quality reconstructions. **b)** We incorporate intrinsic priors to model head appearance under unknown lighting conditions. Combined with a physically-based shading model, we achieve realistic lighting effects across different environments. **c)** Experimental results demonstrate that HRAvatar outperforms existing methods in overall quality, enabling realistic relighting in real-time and simple material editing.

2. Related Work

2.1. 3D Radiance Fields

Image-based 3D reconstruction has become a vibrant research area due to its photorealistic visuals. NeRF [45] introduced a novel method using MLPs to represent a 3D scene as a continuous density and color field, enabling differentiable image rendering through volume rendering. This approach has inspired numerous follow-up studies [1, 16, 44, 56, 66]. However, NeRF faces heavy computational challenges due to extensive MLP queries. Instant-NGP [46] employs multi-resolution hash encoding to accelerate inference. Additionally, some methods, propose hybrid 3D representations [6, 9, 21] to improve efficiency. Recently, 3DGS introduces an explicit representation using Gaussian points, achieving real-time rendering with an efficient tile-based rasterizer. It rapidly gains attention, and researchers applying it to various fields [11, 12, 26, 32, 50, 57, 67, 68] to exploit its efficiency. Our work also builds upon 3DGS to achieve real-time rendering.

2.2. 3D Head Reconstruction

Geometric mesh reconstruction. Traditional 3DMM [3] uses Principal Component Analysis (PCA) to create a parameterized facial model that represents appearance and geometric variations in a linear space. BFM [48] improves on this by adding more scanned facial data, resulting in a richer model. FLAME [35] introduces extra joints for the eyes, jaw, and neck, enabling more realistic facial motion. DECA [20] builds on FLAME by estimating parameters like shape and pose from a single image and capturing finer wrinkles. SMIRK [52] enhances tracking accuracy by using an image-to-image module to provide more precise supervision signals. Besides geometry, some works [8, 18, 19, 33] also focus on learning intrinsic attributes for relightable mesh reconstruction from a single image.

Image-based head reconstruction. Recent advances in

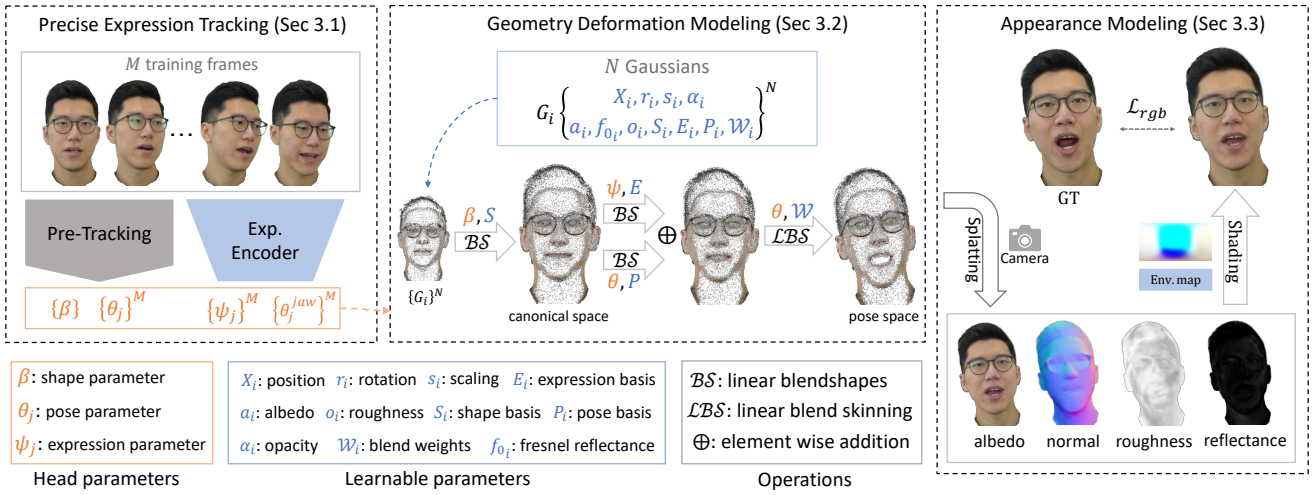

Figure 2. Given a monocular video with unknown lighting and M frames, we first track fixed shape parameter β and pose parameters $\{\theta_j\}^M$ through iterative optimization before training. Expression parameters $\{\psi_j\}^M$ and jaw poses θ^{jaw} are estimated via an expression encoder, which is optimized during training. With these parameters, we transform the Gaussian points into pose space using learnable linear blendshapes $\mathcal{BS}$ and linear blend skinning $\mathcal{LBS}$. We then render the Gaussian points to obtain albedo, roughness, reflectance, and normal maps. Finally, we compute pixel colors using physically-based shading with optimizable environment maps.

neural radiance fields combine 3DMM for view-consistent, photorealistic 3D head reconstruction, which can be generally divided into two categories. *Multi-view-based methods*. Some studies explore multi-view video-based head [25, 42, 49, 55, 63] and full-body [36, 38, 39] reconstruction. However, these approaches require multiple synchronized cameras, making them more complex and less convenient than single-phone captures. Although multi-view-based methods can achieve impressive results, their setup limits the applicability of these approaches. *Monocular-based methods*. NeRFace [22] extends NeRF to dynamic forms by incorporating expression and pose parameters as conditional inputs, enabling animatable head reconstruction. IMavatar [77] models deformation fields for expression and pose motions, using iterative root-finding to locate the canonical surface intersection for each pixel. Point-avatar [78] introduces a novel point-based representation for more efficient animatable avatars. While Point-avatar learns person-specific deformation fields through a shared MLP, our method independently learns per-point blendshapes basis and blend weights, leading to a more flexible deformation modeling. INSTA [80] speeds up training by using multi-resolution hashing for 3D head representation. Recent works [54, 59] based on 3DGS achieve significant breakthroughs in rendering speed. 3D Gaussian Blendshapes (GBS) [43] learn Gaussian basis to better handle expression movements but struggle with pose variations. In contrast, our method utilizes learnable linear blend skinning for flexible point pose transformations, enabling better handling of person-specific head pose animation, while also providing realistic relighting effects.

2.3. Neural Relighting

Implementing relighting in reconstructed 3D scenes is difficult. For static scenes, some methods [23, 62, 72] use learning-based approaches to learn relightable appearances from images under varying lighting. In contrast, inverse rendering methods [4, 70, 73, 74] leverage reflection models like BRDF for more realistic relighting. Recent works [24, 27] integrate BRDF into 3DGS and methods Wu et al. [58], Ye et al. [65] introduce deferred shading for efficient relighting or specular rendering of static scenes. While simplified physical rendering models can be inaccurate, many methods [28, 37, 58] add fitting-based rendering branches to improve reconstruction results. Although some researchers combine physical reflection models with dynamic radiance fields to achieve relightable head avatars [34, 53, 64], they require data under controlled lighting conditions. Reconstructing relightable 3D head avatars under monocular unknown lighting is still underexplored. Point-avatar models lighting but relies on trained shading networks, unable to flexibly relight through environment maps. Unlike NeRF or 3DGS, FLARE [2] reconstructs avatars with meshes and uses a BRDF for relighting, but the reconstruction quality is limited. Our method not only reconstructs superior head avatars but also supports realistic and real-time relighting.

3. Method

As mentioned, previous methods for head reconstruction suffer from inaccurate 3D expression tracking and limited person-specific deformation. They also cannot achieve realistic relighting effects. To tackle these challenges, we en-

hance expression tracking through end-to-end optimization (Sec. 3.1). We also adopt learning strategy for both linear blendshapes and blend skinning for more flexible deformation of Gaussian points (Sec. 3.2). Physically-based shading is employed to realistically model head appearance, which makes our model achieve realistic relighting (Sec. 3.3). The overall pipeline is illustrated in Fig. 2.

3.1. Precise Expression Tracking

Although existing face tracking methods can accurately track head pose and shape parameters, they often struggle to precisely estimate expression parameters. Since these parameters control head expressions, inaccuracies can cause deformation errors, compromising reconstruction quality. To mitigate this issue while maintaining good generalization, we propose to use an expression encoder $\mathcal{E}$ to extract more accurate expression parameters, which is end-to-end trained with subsequent 3D avatar reconstruction:

$$\psi, \theta^{jaw} = \mathcal{E}(I), \tag{1}$$

where ψ and θ^{jaw} represent the expression and jaw pose parameters, respectively. Note that traditional fitting-based methods optimize face parameters using pseudo labels (e.g., pre-estimated 2D landmarks). In contrast, our encoder is trained end-to-end during reconstruction, utilizing photometric loss with ground-truth face images for supervision. Hence, the proposed encoder enables more precise expression tracking and maintains good generalization.

Since point transformations are sensitive to jaw pose parameters [35], we introduce a regularization loss that constrains the distance between the inferred and pre-tracked jaw poses $\hat{\theta}^{jaw}$:

$$\mathcal{L}_{jaw} = \left\| \hat{\theta}^{jaw} - \theta^{jaw} \right\|_2. \tag{2}$$

Other pose parameters in θ and shape parameters β are pre-tracked using [77], with β shared across all frames.

3.2. Geometry Deformation Modeling

Like most methods, we employ a deformation model to map points from canonical space to pose space based on expression and pose parameters. However, facial shapes, expressions, and pose deformations vary widely among individuals, making it difficult for parametric head models to accurately recover each person's unique shape and deformations. To address this, we independently learn per-point blendshapes basis and blend weights adaptively for more flexible geometric deformation.

Learnable linear blendshapes. Similar to FLAME [35], we use linear blendshapes to model geometric displacement. For each Gaussian point, we introduce three additional attributes: shape basis $S = \{S^1, ..., S^{|\beta|}\} \in \mathbb{R}^{N\times 3\times|\beta|}$, expression basis $E = \{E^1, ..., E^{|\psi|}\} \in \mathbb{R}^{N\times 3\times|\psi|}$ and pose basis $P = \{P^1, ..., P^{9K}\} \in \mathbb{R}^{N\times 3\times 9K}$. These are learnable parameters that fit the individual head shape and deformations. First, we compute the shape offset to displace the points to the canonical space X_c using shape blendshapes:

$$\mathcal{BS}(\beta, S) = \sum_{m=1}^{|\beta|} \beta^m S^m, \; X_c = X + \mathcal{BS}(\beta, S), \tag{3}$$

where $\mathcal{BS}$ denotes linear blendshapes and $\beta = \{\beta^1, ..., \beta^{|\beta|}\} \in \mathbb{R}^{|\beta|}$ is the shape parameter. Next, we compute expression and pose offsets in the same manner, using expression blendshapes and pose blendshapes to model facial expressions:

$$X_e = X_c + \mathcal{BS}(\psi, E) + \mathcal{BS}(\mathcal{R}(\theta^*) - \mathcal{R}(\theta^0), P), \tag{4}$$

where $\psi = \{\psi^1, ..., \psi^{|\psi|}\} \in \mathbb{R}^{|\psi|}$ is the expression parameter, and $\theta \in \mathbb{R}^{3(K+1)}$ is the pose parameter representing the axis-angle rotation of the points relative to the joints. θ^* excludes the global joint, with $K = 4$. $\mathcal{R}(\theta)$ is the flattened rotation matrix vector obtained by Rodrigues' formula, and θ^o represents zero pose.

Learnable linear blend skinning. After applying linear displacement, we transform Gaussian points into pose space using Linear Blend Skinning (LBS). Each Gaussian point is assigned with a learnable blend weight attribute $\mathcal{W} \in \mathbb{R}^{N\times K}$ to accommodate individual pose deformations. $\mathcal{LBS}$ rotates the points X_e around each joints $\mathcal{J}(\beta)$ and linearly weighted by $\mathcal{W}$, defined as:

$$X_p = \mathcal{LBS}(X_e, \mathcal{J}(\beta), \mathcal{W}) = R_{lbs}X_e + T_{lbs}, \tag{5}$$

where $\mathcal{J}(\beta) \in \mathbb{R}^{K\times 3}$ represents the positions of the neck, jaw, and eyeball joints. To maintain geometric consistency, the rotation attributes of the Gaussians are also transformed by the weighted rotation matrix R_{lbs}: $R_p = R_{lbs}R$.

Geometry initialization. To facilitate easier learning, we leverage FLAME's geometric and deformation priors. We initialize the positions of the Gaussian points through linear interpolation on the FLAME mesh faces. The same method is applied to initialize the blendshapes basis and blend weights. Other geometric attributes, like rotation and scale, are initialized similarly to 3DGS [30].

3.3. Appearance Modeling

3DGS uses spherical harmonics to model the view-dependent appearance of each point, but it cannot simulate visual effects under new lighting conditions. To overcome this, we introduce a novel appearance modeling approach that decomposes the appearance into three properties: albedo a, roughness o, and Fresnel base reflectance f_0. We then utilize a BRDF model [5] for physically-based shading of the image. To enhance efficiency, we apply the

SplitSum approximation technique [29] to precompute the environment map.

Shading. First, we render the albedo map $\mathbf{A}$, roughness map $\mathbf{O}$, reflectance map $\mathbf{F_0}$, and normal map $\mathbf{N}$ using rasterizer. The specular and diffuse maps are then calculated as follows:

$$I_{specular} = I_{env}(\mathbf{R}, \mathbf{O}) \cdot (ks \cdot I_{BRDF}(\mathbf{O}, \mathbf{N} \cdot \mathbf{V})[0] + I_{BRDF}(\mathbf{O}, \mathbf{N} \cdot \mathbf{V})[1]), \tag{6}$$

$$I_{diffuse} = \mathbf{A} \cdot I_{irr}(\mathbf{N}), \tag{7}$$

where $\mathbf{V}$ is the view direction map derived from the camera parameters and $\mathbf{R}$ is the reflection direction map, computed as $\mathbf{R} = 2(\mathbf{N} \cdot \mathbf{V})\mathbf{N} - \mathbf{V}$. I_{BRDF} is a precomputed map of the simplified BRDF integral. We use an approximate Fresnel equation $\tilde{\mathcal{F}}$ to compute the specular reflectance ks:

$$ks = \tilde{\mathcal{F}}(\mathbf{N} \cdot \mathbf{V}, \mathbf{O}, \mathbf{F_0}) = \mathbf{F_0} + (\max(1 - \mathbf{O}, \mathbf{F_0}) - \mathbf{F_0}) \cdot 2^{(-5.55473(\mathbf{N}\cdot\mathbf{V}) - 6.698316)\cdot(\mathbf{N}\cdot\mathbf{V})}. \tag{8}$$

The final shaded image is computed as: $I_{shading} = I_{diffuse} + I_{specular}$. During training, we optimize two cube maps: the environment irradiance map I_{irr} and the prefiltered environment map I_{env}. $I_{env}(\mathbf{R}, \mathbf{O})$ provides radiance values based on the reflection directions and roughness, while $I_{irr}(\mathbf{N})$ provides irradiance values based on the normal directions.

Normal estimation. Smooth and accurate normals are essential for physical rendering, as rough normals can cause artifacts during relighting. Following Jiang et al. [27], we use the shortest axis of each Gaussian point as its normal n. To ensure the correct direction and geometric consistency, we supervise the rendered normal map $\mathbf{N}$ with the normal map $\hat{\mathbf{N}}$ obtained from depth derivatives:

$$\mathcal{L}_{normal} = \left\| \mathbf{1} - \mathbf{N} \cdot \hat{\mathbf{N}} \right\|_1. \tag{9}$$

Intrinsic prior. Disentangling material properties under constant unknown lighting is challenging due to inherent uncertainties. When reconstructing heads under non-uniform lighting, local lighting effects can be erroneously coupled into the albedo, resulting in unrealistic relighting. To address this, we use an existing model [14] to extract pseudo-ground-truth albedos $\mathbf{A}^{gt}$, supervising the rendered albedos for a more realistic appearance, as Eq. (10). We also constrain the roughness and base reflectance within predefined ranges: $o \in [\tau^o_{min}, \tau^o_{max}]$, $f_0 \in [\tau^{f_0}_{min}, \tau^{f_0}_{max}]$.

$$\mathcal{L}_{albedo} = \left\| \mathbf{A} - \mathbf{A}^{gt} \right\|_1. \tag{10}$$

3.4. Optimization

During optimization, we retain the point densification and pruning strategy from 3DGS, with additional attributes inherited similarly. In addition to the previously mentioned losses, we use the Mean Absolute Error (MAE) and D-SSIM to calculate the error between the rendered image and ground truth, as Eq. (12). We also apply Total Variation (TV) loss $\mathcal{L}_{tv}$ to the rendered roughness map $\mathbf{O}$ to ensure smoothness. The total loss function is given in Eq. (11). The weights for each loss component are set as follows: $\lambda_{jaw} = 0.1$, $\lambda_1 = 0.8$, $\lambda_{\mathcal{W}} = 0.1$, $\lambda_{normal} = 10^{-5}$, $\lambda_{albedo} = 0.25$, $\lambda_{tv} = 0.02$.

$$\mathcal{L}_{total} = \mathcal{L}_{rgb} + \lambda_{jaw}\mathcal{L}_{jaw} + \lambda_{normal}\mathcal{L}_{normal} + \lambda_{albedo}\mathcal{L}_{albedo} + \lambda_{tv}\mathcal{L}_{tv}(\mathbf{O}), \tag{11}$$

$$\text{where} \quad \mathcal{L}_{rgb} = \lambda_1 \left\| I_{shading} - I_{gt} \right\|_1 + (1 - \lambda_1)\mathcal{L}_{\text{D-SSIM}}(I_{shading}, I_{gt}). \tag{12}$$

4. Experiment

4.1. Experimental Setup

Implementation details. We build our model using PyTorch [47] and train it with the Adam optimizer [31] on a single NVIDIA 3090 GPU. Each monocular head video is trained for 15 epochs. All videos are cropped and resized to a resolution of 512×512. We run matting (*e.g.* [13, 40]) to extract the foreground, setting the background to black. Moreover, we follow Zheng et al. [77] to pre-track FLAME parameters for the videos. For our encoder $\mathcal{E}$, we utilize the pre-trained weight from SMIRK [52].

Dataset. We evaluate different methods on 10 subjects from the INSTA dataset [80], which provides pre-cropped and segmented images. Following INSTA, we use the last 350 frames of each video as the test set for self-reenactment evaluation. For a more robust assessment, we include 8 subjects from the HDTF dataset [75], which is collected from the internet. We also include 5 self-captured subjects using a mobile phone. For these two datasets, the last 500 frames are used as the test set. All methods adopt the same cropped and segmented process.

Baseline and metrics. We compare our method against several SOTA methods: Point-avatar [78], INSTA [80], Splatting-avatar [54], Flash-avatar [59], and 3D Gaussian Blendshapes (GBS) [43], as well as FLARE [2] for relighting. For each method, we use the official code to generate the results. Note that we disable the post-training optimization of test images' parameters in Point-avatar to ensure fairness. We use PSNR, MAE* (MAE $\times 10^2$), SSIM, and LPIPS [71] to evaluate the image quality.

4.2. Evaluation

Quantitative results. We evaluate all methods for self-reenactment, as shown in Tab. 1. Our method outperforms others across all four metrics, especially in LPIPS. This highlights that our method reconstructs more detailed

Method	INSTA dataset				HDTF dataset				self-captured dataset			
	PSNR↑	MAE*↓	SSIM↑	LPIPS↓	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	25.91	1.910	0.8333	0.1833
Point-avatar	26.84	1.549	0.8970	0.0926	25.14	2.236	0.8385	0.1278	25.83	1.692	0.8556	0.1241
Splatting-avatar	28.71	1.200	0.9271	0.0862	26.66	2.01	0.8611	0.1351	26.47	1.711	0.8588	0.1550
Flash-avatar	29.13	1.133	0.9255	0.0719	27.58	1.751	0.8664	0.1095	27.46	1.632	0.8348	0.1456
GBS	29.64	1.020	0.9394	0.0823	27.81	1.601	0.8915	0.1297	28.59	1.331	0.8891	0.1560
HRAvatar (Ours)	**30.36**	**0.845**	**0.9482**	**0.0569**	**28.55**	**1.373**	**0.9089**	**0.0825**	**28.97**	**1.123**	**0.9054**	**0.1059**

Table 1. Average quantitative results on the INSTA, HDTF, and self-captured datasets. Our method outperforms others in PSNR, MAE* (MAE $\times 10^2$), SSIM, and LPIPS metrics.

Figure 3. Qualitative comparison results on self-reenactment. Compared to others, ours captures finer texture details and renders high-fidelity images. Ours also achieves more accurate expression deformations and reconstructs better geometric details.

and high-quality animatable avatars, with the improved LPIPS score suggesting sharper images. Moreover, we test HRAvatar's rendering speed for animation and relighting, achieving about **155 FPS**. Further details are in the supplementary material.

Qualitative results. The visual comparison of our method with baseline methods on self-reenactment is shown in Fig. 3. INSTA and Splatting-avatar often struggle with chal-

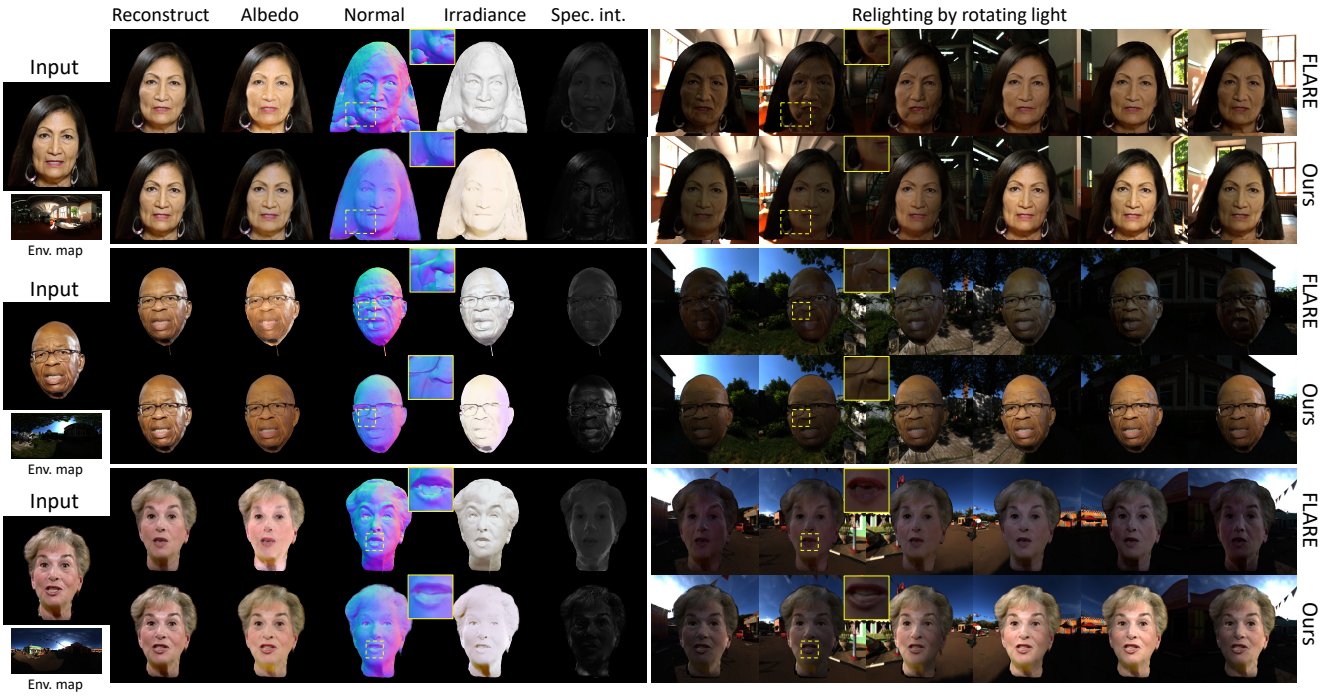

Figure 4. Visual comparison with FLARE on relighting. "Spec. int." denotes the specular intensity coefficient. FLARE exhibits some artifacts due to partially corrupted normals, while our method learns smoother normals, enabling more reasonable and consistent relighting. Notably, due to differences in pre-filtering environment maps, our method and FLARE exhibit variations in lighting brightness.

	PSNR↑	MAE*↓	SSIM↑	LPIPS↓
full (ours)	**30.36**	**0.845**	**0.9482**	0.0569
rigged to FLAME	29.79	0.937	0.9431	0.0695
MLP deform	29.67	0.966	0.941	0.0706
w/o exp. encoder	29.70	0.933	0.9438	0.0667
w/o learnable deform	29.83	0.923	0.9440	0.0684
w/o PBS	30.34	0.850	0.9480	**0.0563**

Table 2. Ablation quantitative results on the INSTA dataset. **Bold** marks the best results, and underline marks the second best results.

lenging poses, resulting in significant artifacts. Point-avatar maintains decent rendering in such poses but suffers from point artifacts and lacks detail in the mouth. Flash-avatar shows improvements but still loses some fine textures and has expression inaccuracies. GBS achieves relatively accurate facial expressions in normal poses but introduces blurring around edges, like the ears, hair, and neck. In contrast, our method accurately restores fine textures, such as hair and eye luster, while preserving precise geometric details like ears and teeth. Ours handles wrinkles and blinking more effectively due to the flexible deformation model and accurate tracking.

We qualitatively compare the visual differences in relighting between FLARE and our method. As shown in Fig. 4, FLARE incorrectly reconstructs some of the subject's geometric normals, causing blocky artifacts during relighting. In contrast, our method learns smoother normals, leading to more consistent and realistic lighting effects. Additional comparisons with FLARE are provided in the supplementary material.

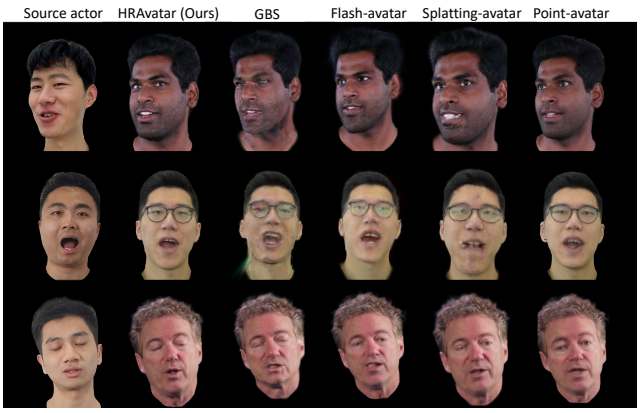

Figure 5. Visual comparison on cross-reenactment. HRAvatar accurately simulates actors' poses and expressions, preserving textures and geometric details, while others exhibit artifacts.

We also present cross-reenactment visual comparisons. As shown in Fig. 5, our method better retains the source actor's expressions and preserves original head details, even

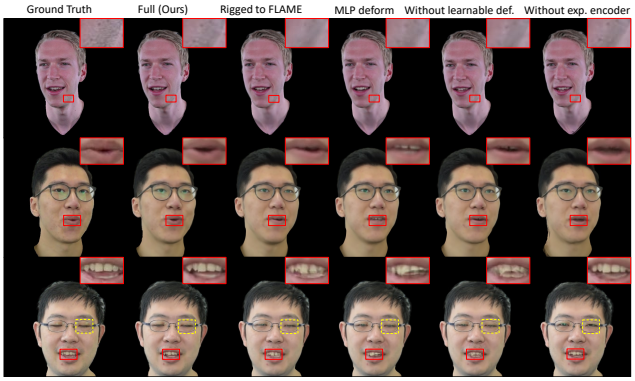

Figure 6. Qualitative results of the ablation study. Our full method renders better texture and geometry details and captures more accurate facial expressions, including mouth shapes and blinking.

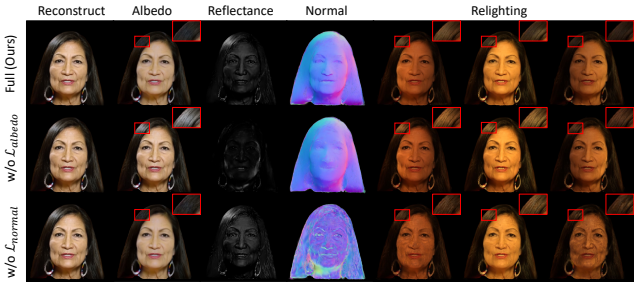

Figure 7. Ablation study for albedo and normal losses. Without $\mathcal{L}_{albedo}$, entangled attributes yield unrealistic relighting. Without $\mathcal{L}_{normal}$, chaotic normal maps cause artifacts when relighting.

in challenging poses and expressions, while other methods exhibit blurring and artifacts. It's worth noting that Flashavatar and GBS treat head poses as camera poses, which may cause minor scale discrepancies, resulting in variations in the size and positioning of rendered avatars.

Additionally, the supplementary material includes more relighting results under rotating environment maps, as well as material editing and novel view synthesis.

4.3. Ablation Studies

The quantitative results of the ablation study on self-reenactment are summarized in Tab. 2, with qualitative results in Fig. 6 and Fig. 7, validating the effectiveness of each component.

Rigged to FLAME. We replace HRAvatar's learnable blendshapes and LBS with the deformation method from Qian et al. [49], which rigs Gaussian points to the FLAME mesh. The results in Tab. 2 and Fig. 6 demonstrate that our model improves on metrics and achieves more accurate texture and tooth details.

MLP deform. To validate the superiority of independently learning per-point blendshapes basis and blend weights, we follow Point-avatar [78] and use a shared MLP to predict them for each point. The results highlight the advantages of our learning strategy.

Without learnable deform. We set the blendshapes basis and blend weights as non-learnable to assess the importance of adapting to individual deformations. This leads to reduced geometry and texture quality.

Without exp. encoder. To verify the expression encoder's effectiveness in extracting expression parameters, we use pre-tracked parameters instead. Results indicate our method better restores facial expressions, including mouth shapes and blinking, and improves performance metrics.

Without PBS. This means using the standard 3DGS appearance model instead of our shading model. While the fitting-based method of 3DGS performs well due to more learnable parameters and flexibility, our method achieves comparable results while enabling realistic relighting.

Without $\mathcal{L}_{normal}$. As shown in Fig. 7, removing normal consistency loss results in chaotic normal maps, causing blocky artifacts during relighting.

Without $\mathcal{L}_{albedo}$. Without the albedo prior loss, appearance attributes become entangled, causing incorrect coupling of local highlights with albedo. This results in unrealistic relighting effects, with highlights appearing in areas without actual lighting, as shown in Fig. 7.

5. Discussion

Conclusion. In this paper, we introduce HRAvatar, a novel method for high-fidelity, relightable 3D head avatar reconstruction from monocular video. To address errors incorporated from inaccurate facial expression tracking, we train an encoder in an end-to-end manner to extract more precise parameters. We model individual-specific deformations using learnable blendshapes and linear blend skinning for flexible Gaussian point deformation. By employing physically-based shading for appearance modeling, our method enables realistic relighting. Experimental results show that HRAvatar achieves state-of-the-art quality and real-time realistic relighting effects.

Limitation. While our method models effectively individual deformations well, it remains constrained by FLAME's priors when training data is insufficient, affecting control over elements like hair or accessories. Due to 3DGS's strong texture representation and the limitations of existing albedo estimation models, some shadows or wrinkles may still be mis-coupled into albedo or reflectance, leading to shortcomings in relighting, particularly for specular reflections or shadows. Besides, reconstructing the full head from a monocular video is infeasible for our method with unknown camera poses, even if the back of the head is visible. This is because monocular pose estimation relies on facial key points, which become unreliable when the yaw angle approaches 90 degrees.

Acknowledgement

This research was funded by the National Key Research and Development Program of China (Project No.2022YFB36066) and partially funded by the Shenzhen Science and Technology under Grant (KJZD20240903103210014).

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. 2

[2] Shrisha Bharadwaj, Yufeng Zheng, Otmar Hilliges, Michael J Black, and Victoria Fernandez-Abrevaya. Flare: Fast learning of animatable and relightable mesh avatars. *arXiv preprint arXiv:2310.17519*, 2023. 3, 5

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, page 187–194, 1999. 1, 2

[4] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. 3

[5] Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. In *Acm Siggraph*, pages 1–7. vol. 2012, 2012. 4

[6] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 2

[7] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 1

[8] Pol Caselles, Eduard Ramon, Jaime Garcia, Xavier Giro-i Nieto, Francesc Moreno-Noguer, and Gil Triginer. Sira: Relightable avatars from a single image. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 775–784, 2023. 2

[9] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 2

[10] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. Faceposenet: Making a case for landmark-free face alignment. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1599–1608, 2017. 1

[11] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 2

[12] Kangjie Chen, BingQuan Dai, Minghan Qin, Dongbin Zhang, Peihao Li, Yingshuang Zou, and Haoqian Wang. Sl-gaussian: Fast language gaussian splatting in sparse views. *arXiv preprint arXiv:2412.08331*, 2024. 2

[13] Xiangguang Chen, Ye Zhu, Yu Li, Bingtao Fu, Lei Sun, Ying Shan, and Shan Liu. Robust human matting via semantic guidance. In *Proceedings of the Asian Conference on Computer Vision*, pages 2984–2999, 2022. 5

[14] Xi Chen, Sida Peng, Dongchen Yang, Yuan Liu, Bowen Pan, Chengfei Lv, and Xiaowei Zhou. Intrinsicanything: Learning diffusion priors for inverse rendering under unknown illumination. *arXiv preprint arXiv:2404.11593*, 2024. 5

[15] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 2

[16] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. GPAvatar: Generalizable and precise head avatar from image(s). In *The Twelfth International Conference on Learning Representations*, 2024. 2

[17] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 1

[18] Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12819–12829, 2021. 2

[19] Abdallah Dib, Luiz Gustavo Hafemann, Emeline Got, Trevor Anderson, Amin Fadaeinejad, Rafael M. O. Cruz, and Marc-André Carbonneau. Mosar: Monocular semi-supervised model for avatar reconstruction using differentiable shading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1770–1780, 2024. 2

[20] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13, 2021. 2

[21] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2

[22] Guy Gafni, Justus Thies, Michael Zollhöfer, 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 (CVPR)*, pages 8649–8658, 2021. 3

[23] Duan Gao, Guojun Chen, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. Deferred neural lighting: free-viewpoint re-

lighting from unstructured photographs. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 3

[24] Jian Gao, Chun Gu, Youtian Lin, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussian: Real-time point cloud relighting with brdf decomposition and ray tracing. *arXiv preprint arXiv:2311.16043*, 2023. 3

[25] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Npga: Neural parametric gaussian avatars. *arXiv preprint arXiv:2405.19331*, 2024. 3

[26] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[27] Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5322–5332, 2024. 3, 5

[28] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensoir: Tensorial inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2023. 3

[29] Brian Karis and Epic Games. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 4(3):1, 2013. 5

[30] 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. 2, 4

[31] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[32] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads. *arXiv preprint arXiv:2406.09377*, 2024. 2

[33] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction "in-the-wild". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[34] Gengyan Li, Abhimitra Meka, Franziska Mueller, Marcel C Buehler, Otmar Hilliges, and Thabo Beeler. Eyenerf: a hybrid representation for photorealistic synthesis, animation and relighting of human eyes. *ACM Transactions on Graphics (ToG)*, 41(4):1–16, 2022. 3

[35] 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. 1, 2, 4

[36] Zhe Li, Zerong Zheng, Yuxiao Liu, Boyao Zhou, and Yebin Liu. Posevocab: Learning joint-structured pose embeddings for human avatar modeling. In *ACM SIGGRAPH Conference Proceedings*, 2023. 3

[37] Zhe Li, Yipengjing Sun, Zerong Zheng, Lizhen Wang, Shengping Zhang, and Yebin Liu. Animatable and relightable gaussians for high-fidelity human avatar modeling. *arXiv preprint arXiv:2311.16096v4*, 2024. 3

[38] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[39] Zhouyingcheng Liao, Vladislav Golyanik, Marc Habermann, and Christian Theobalt. Vinecs: video-based neural character skinning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1377–1387, 2024. 3

[40] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 238–247, 2022. 5

[41] Yunfei Liu, Lei Zhu, Lijian Lin, Ye Zhu, Ailing Zhang, and Yu Li. Teaser: Token enhanced spatial modeling for expressions reconstruction. *arXiv preprint arXiv:2502.10982*, 2025. 1

[42] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 3

[43] Shengjie Ma, Yanlin Weng, Tianjia Shao, and Kun Zhou. 3d gaussian blendshapes for head avatar animation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024. 3, 5

[44] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7210–7219, 2021. 2

[45] 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. In *ECCV*, 2020. 2

[46] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[48] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 1, 2

[49] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaus-

sianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 3, 8

[50] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 2

[51] 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. 2

[52] George Retsinas, Panagiotis P Filntisis, Radek Danecek, Victoria F Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2490–2501, 2024. 2, 5

[53] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2024. 3

[54] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 5

[55] Kartik Teotia, Hyeongwoo Kim, Pablo Garrido, Marc Habermann, Mohamed Elgharib, and Christian Theobalt. Gaussianheads: End-to-end learning of drivable gaussian head avatars from coarse-to-fine representations. *ACM Transactions on Graphics (TOG)*, 43(6):1–12, 2024. 3

[56] 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. 2

[57] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024. 2

[58] Tong Wu, Jia-Mu Sun, Yu-Kun Lai, Yuewen Ma, Leif Kobbelt, and Lin Gao. Deferredgs: Decoupled and editable gaussian splatting with deferred shading. *arXiv preprint arXiv:2404.09412*, 2024. 3

[59] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 5

[60] 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. 2

[61] Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Han Huang, Guojun Qi, and Yebin Liu. Latentavatar: Learning latent expression code for expressive neural head avatar. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–10, 2023. 2

[62] Yingyan Xu, Gaspard Zoss, Prashanth Chandran, Markus Gross, Derek Bradley, and Paulo Gotardo. Renerf: Relightable neural radiance fields with nearfield lighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22581–22591, 2023. 3

[63] 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. 3

[64] Haotian Yang, Mingwu Zheng, Chongyang Ma, Yu-Kun Lai, Pengfei Wan, and Haibin Huang. Vrmm: A volumetric relightable morphable head model. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[65] Keyang Ye, Qiming Hou, and Kun Zhou. 3d gaussian splatting with deferred reflection. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024. 3

[66] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 2

[67] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19447–19456, 2024. 2

[68] Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. *arXiv preprint arXiv:2403.15704*, 2024. 2

[69] Jiawei Zhang, Zijian Wu, Zhiyang Liang, Yicheng Gong, Dongfang Hu, Yao Yao, Xun Cao, and Hao Zhu. Fate: Full-head gaussian avatar with textural editing from monocular video. *arXiv preprint arXiv:2411.15604*, 2024. 2

[70] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. 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*, pages 5453–5462, 2021. 3

[71] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[72] Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, et al. Neural light transport for relighting and view synthesis. *ACM Transactions on Graphics (TOG)*, 40(1):1–17, 2021. 3

[73] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under

an unknown illumination. *ACM Transactions on Graphics (ToG)*, 40(6):1–18, 2021. 3

[74] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *CVPR*, 2022. 3

[75] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 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*, pages 3661–3670, 2021. 5

[76] Xiaochen Zhao, Lizhen Wang, Jingxiang Sun, Hongwen Zhang, Jinli Suo, and Yebin Liu. Havatar: High-fidelity head avatar via facial model conditioned neural radiance field. *ACM Transactions on Graphics*, 43(1):1–16, 2023. 2

[77] 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. 3, 4, 5

[78] 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, 3, 5, 8

[79] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022. 1

[80] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4574–4584, 2023. 3, 5