RMAvatar: Photorealistic Human Avatar Reconstruction from Monocular Video Based on Rectified Mesh-embedded Gaussians

Sen Peng[1], Weixing Xie[2], Zilong Wang[3], Xiaohu Guo[3], Zhonggui Chen[4], Baorong Yang*[1], and Xiao Dong*[5]

[1]College of Computer Engineering, Jimei University, Xiamen, China
[2]National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China
[3]Department of Computer Science, The University of Texas at Dallas, Richardson, United States
[4]School of Informatics, Xiamen University, Xiamen, China
[5]Guangdong Provincial/Zhuhai Key Laboratory of IRADS, Beijing Normal-Hong Kong Baptist University, Zhuhai, China

Abstract

We introduce RMAvatar, a novel human avatar representation with Gaussian splatting embedded on mesh to learn clothed avatar from a monocular video. We utilize the explicit mesh geometry to represent motion and shape of a virtual human and implicit appearance rendering with Gaussian Splatting. Our method consists of two main modules: Gaussian initialization module and Gaussian rectification module. We embed Gaussians into triangular faces and control their motion through the mesh, which ensures low-frequency motion and surface deformation of the avatar. Due to the limitations of LBS formula, the human skeleton is hard to control complex non-rigid transformations. We then design a pose-related Gaussian rectification module to learn fine-detailed non-rigid deformations, further improving the realism and expressiveness of the avatar. We conduct extensive experiments on public datasets, RMAvatar shows state-of-the-art performance on both rendering quality and quantitative evaluations. Please see our project page at https://rm-avatar.github.io.

Keywords: 3D Reconstruction, Human Avatar, Monocular Video Reconstruction, Gaussian Splatting

1. Introduction

High-fidelity animatable human avatar modeling from videos has been a longstanding challenge in computer vision. The rendering of photorealistic avatars from arbitrary views is important due to its wide applications in telepresence [18], movie making and AR/VR [51]. Modeling human avatars from video requires fusing multiple 2D observations to synthesize a 3D consistent human model. Traditional methods usually rely on dense multi-view supervision to reconstruct the avatar, in most actual scenarios, such a complex multi-view camera system [17, 55] is not readily available. The under-constrained nature of monocular observation makes the task of reconstructing the unseen poses and viewpoints of human avatars more challenging. In addition, the distortion of clothes, hair and hand movements are also difficult to reconstruct, and the rendering quality of these parts will affect the realism of the avatar.

Recent methods [11, 16, 38, 53] based on implicit neural fields [32, 35, 36] usually learn a canonical avatar representation by mapping camera rays from observation space to canonical space. In NeRF framework [41, 53], the inverse mapping may project multiple points from the observation space onto same point in the canonical space. This ambiguous correspondence affects the rendering quality, especially for objects with significant motion or high-frequency details. In addition, the heavy MLPs used to model the underlying neural radiation fields are computationally expensive, resulting in long training and inference time. Other than implicit neural representations, point-based methods [54, 66, 23, 62] are efficient and can capture flexible topology, but may produce incomplete surface geometries.

The recent 3D Gaussian Splatting method [19] surpasses NeRF in both rendering quality and efficiency by optimizing discrete 3D Gaussian primitives to learn explicit representation of the scene. Follow-up works [6, 13] utilize the strong representation ability of 3D Gaussians and reconstruct human avatar via pose-dependent appearance mod-

*Corresponding authors. Emails: yangbaorong@jmu.edu.cn, xiaodong@uic.edu.cn

eling. Although the current work has made substantial progress, the authenticity of avatar needs to be further improved. Existing human model reconstruction works can be roughly divided into two categories. One is implicit neural avatar based on Gaussian Splatting [13, 43]. Given initial Gaussian primitives of avatar, these methods first learn 3D Gaussians in the canonical space, and transform Gaussians to the observation space based on the guidance of human pose and LBS [25] . These methods adopt the multi-layer perceptrons (MLP) for motion control [43, 58, 56], which is inferior to mesh-based representation to capture surface deformation. The other category is the hybrid avatar representation [45, 42, 52], which combines rendering quality of Gaussian splatting with geometry modeling of deformable meshes. Specifically, a Gaussian is attached to a mesh face and deformed with the face. The hybrid representation enables a compact and topology complete avatar, thus providing better regularization for Gaussians in novel poses. This representation is beneficial for learning avatars under monocular observation. However, the flexibility of the model needs to be improved. In addition to obtaining the avatar in the observation space based on the mesh guidance, the Gaussian primitives need to be further fine-tuned to enhance the ability to learn complex personal characters, such as twisted clothes and wispy hair.

To address these issues, we introduce a novel 3D avatar representation which is designed to model personalized human avatar with complex identity and motion. Our model, namely RMAvatar, is a hybrid 3D representation with Gaussians embedded on a mesh. The use of the mesh can more accurately represent body motion and surface deformation, providing accurate positioning of the Gaussians in observation space. Our method consists of two main modules: Gaussian initialization module and Gaussian rectification module. At the first, we load a template mesh and deform it to obtain posed mesh at a certain frame. Our method binds 3D Gaussian splats to the posed mesh locally and transform the splats to global space. The Gaussian splat now has a good initial position and other default properties. We optimize the Gaussians by minimizing color loss on the rendering. However, methods based on LBS formula deformation cannot capture the motion of fine non-surface regions, such as cloth distortions, hair and skin winkles. Thus the representation ability of Gaussians on posed mesh need to be further improved. We then propose the Gaussian rectification strategy, which is a pose-dependent non-rigid deformation module based on MLP. This module allows us to predict further positional adjustments and covariance shifts, significantly boosting the avatar's realism and expressiveness.

In summary, the contributions of our method are as follows:

- We propose RMAvatar to model personalized high-fidelity human avatar from monocular videos based on mesh-embedded Gaussian splats.
- We design Gaussian rectification module to accurately capture complex non-rigid deformation relate to pose to improve the realism of the avatar.
- We conduct extensive experiments on public datasets to demonstrate the superior reconstruction ability of our method on both rendering quality and quantitative evaluations.

2. Related Work

Reconstruction of animable human avatars from monocular videos is challenging to capture high-quality geometry deformation and appearance. Early works take parametric template models, e.g., FLAME [26] and SMPL [30, 37], which provide vertices with fixed connectivity as an explicit prior of the 3D human avatar. By unwrapping them to a unified UV space, texture atlas can be obtained through differentiable rendering [35]. These explicit mesh-based models can be easily fit into existing rendering pipeline and the vertices of mesh templates can be easily deformed to capture pose-dependent geometry deformation. However, since the mesh templates don't model clothes and have fixed topology, these methods often suffer from capturing fine-scale deformation of the human body, especially the distortion and color changes of clothes, hairs and faces. To address these problems, researchers have investigated ways, e.g., via learning 3D vertex offsets [1, 33, 49, 67] for clothes, or resorting to implicit representations, e.g., neural radiance fields (NeRF) [16, 53, 14] and Gaussian fields [13, 6, 43, 7, 29].

2.1. Implicit function-based human avatar

Implicit models normally encode the 3D human avatar with implicit surface functions [39], e.g., SDF [65, 60], occupancy field [31, 5, 28] and NeRF [14, 53, 16, 38, 15, 61, 4]. NeRF-based methods learns the neural radiance fields of human from videos and render novel views with differentiable volumetric rendering. These methods present a deformable NeRF representation by unwrapping different poses to a shared canonical space with inverse kinematic transformations as well as residual deformations for modelling animatable human avatar from videos. Especially, HumanNeRF [53] models human motion by decomposing it into skeletal and non-rigid deformations and refines texture details by aggregating color and depth information from neighboring views. Anim-NeRF [4] learns pose conditioned inverse LBS field to capture the fine details of human. However, the implicit function-based methods usually adopt pure MLPs to model the human avatar, yielding smooth and blurry quality and low rendering speed [38]. To address the expensive computation of

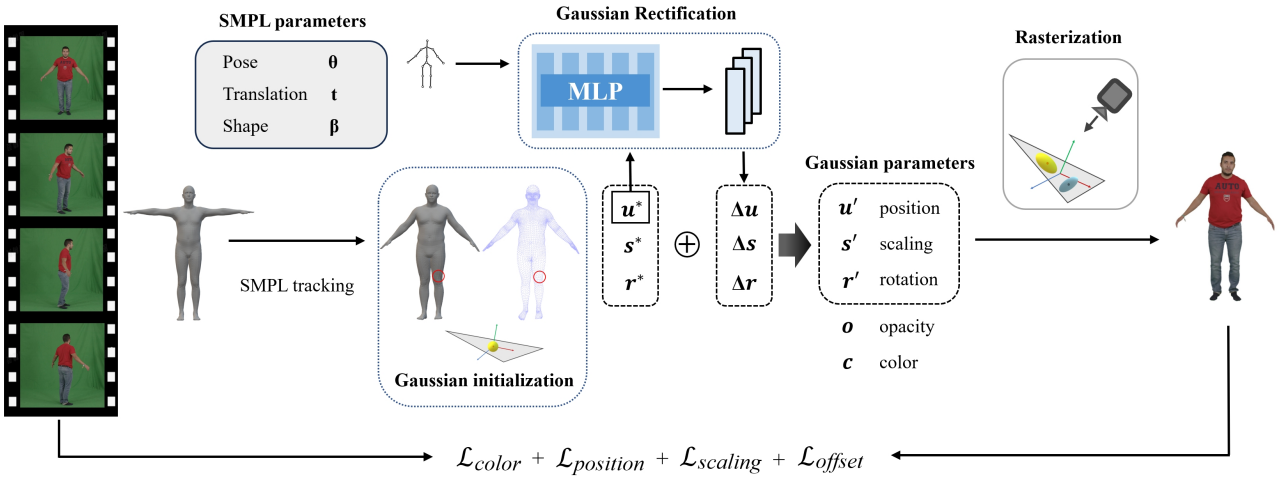

Figure 1: Overview of RMAvatar. Given a sequence of monocular images and a neutral SMPL template, we first obtain the deformed mesh under current pose of the person via SMPL tracking. Our initialization strategy is to embed Gaussian splats on mesh in the local coordinates of each triangle and then transform them to world space based on triangle's shape. To improve the representation ability of Gaussian splats for non-rigid cloth deformation, the rectification module is designed to further adjust Gaussians to learn pose-dependent appearance details of avatar. Finally, the Gaussians in observation frame and their respective color values are accumulated via differentiable Gaussian rasterization to render the image.

NeRF-based avatar, efforts has been made in accelerated data structures [48, 64, 15, 11, 22]. However, some works rely on dense multi-view inputs [48, 64, 22, 27, 3, 9, 59] to achieve good rendering quality of animatable human avatar. Instant-NVR [11] use iNGP [34] as the underlying representation for articulated NeRFs, and modelling non-rigid deformations in the UV space for fast training. However, it generates blurring renderings on the non-rigid deformations [43].

2.2. Point-based human avatar

Point cloud is also a commonly used 3D representation. DPF [40] and NPC [46] apply Point-NeRF [57] for producing explicit surface points and learning non-rigid deformation of human avatars. However, with the MLPs used in Point-NeRF, these methods still struggle with the blurry rendering and the computation efficiency. Point-based rendering [68, 66, 19] has been adopted as an effective alternative to NeRF in digital human reconstruction. PointAvatar [66] takes advantage of explicit point primitives in forward rasterization and learns a forward deformation from canonical to pose space, generating high-fidelity 3D avatars. As an explicit point-based 3D representation, 3D Gaussian Splatting (3DGS) [19] has shown its efficiency in real-time photo-realistic rendering of static scenes. Concurrent works [21, 24, 27, 42, 43, 44, 47] introduce 3DGS into animatable human avatars. GaussianAvatar [13] introduces 3D Gaussians to model digital humans, and jointly optimizes motion and appearance in the avatar modeling process. 3DGS-Avatar [43] leverages Gaussian splats and explicitly learns a non-rigid transformation network to model the pose-related Gaussian deformations, improving the reconstruction quality of clothed human avatars.

2.3. Hybrid human avatar

Hybrid 3D representations have also been used in modelling human avatars from videos [12, 8]. HDHumans [12] learns the deformation embedded as NeRF by the parameterized pose and embedded graph of template mesh prior. DELTA [10] propose to hybridly model human avatar with textured mesh for body and NeRF for hair and clothing. SplattingAvatar [45] proposes a hybrid avatar representation of 3D Gaussians and mesh to disentangle the human motion and appearance. Specifically, the pose-dependent deformation are explicitly defined by mesh, and the geometry and appearance details are modeled by the Gaussians. Similarly, GoMAvatar [52] introduces the Gaussian-on-Mesh representation that leverages 3D Gaussians for real-time rendering. GaussianAvatars [42] reconstructs head avatars by rigging 3D Gaussians to a parametric morphable face model with the binding inheritance strategy in which the Gaussian is parameterized with the index of its parent triangle. However, in SplattingAvatar [45] and GaussianAvatars [42], the Gaussian attributes are independent of the specific pose or expression of the avatar, so non-rigid deformations related to pose cannot be modeled. To solve this problem, based on mesh-embedded Gaussians, we design a Gaussian rectification module to help Gaussians represent areas that

SMPL [30] cannot model, such as clothes.

3. Method

Figure 1 shows the framework of our method RMAvatar. Given a sequence of monocular images and a SMPL template, we deform the mesh to a certain pose of a person via LBS. Each Gaussian function is embedded in a triangular face of the mesh and moves with the triangle. Except for position, each gaussian has its attributes of rotation, scaling, opacity and color. To reconstruct personalized high-fidelity avatars, we add pose-dependent transformations to each Gaussian to learn details caused by non-rigid transformations of the human avatar.

3.1. Preliminary

3D Gaussian Splatting [19] employs a set of anisotropic Gaussian primitives to explicitly represent a static 3D scenes. Each Gaussian splat is characterized by a covariance matrix Σ at position μ, which is referred as the mean of Gaussian:

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

To ensure positive semi-definite nature of the covariance matrix, it can be decomposed into a scaling matrix S and a rotation matrix R:

$$\Sigma = RSS^TR^T. \tag{2}$$

In practice, we store diagonal vector $s \in \mathbb{R}^3$ of scaling matrix S and a quaternion vector $r \in \mathbb{R}^4$ of rotation matrix R for a Gaussian.

During the rendering, the 3D Gaussians are projected to 2D image plane and accumulated via alpha blending. As introduced by [69], using a viewing transform matrix W and the Jacobian matrix J of the affine approximation of the projective transformation, the covariance matrix Σ' in 2D camera space can be computed as:

$$\Sigma' = JW\Sigma W^TJ^T. \tag{3}$$

The color of a pixel is then calculated by blending 3D Gaussian splats that overlap at that pixel, and these Gaussians are sorted according to their depth:

$$C = \sum_{i\in N} c_i\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j). \tag{4}$$

Here, α_i is blending weight calculated by opacity σ_i multiplied with the probability density of projected 2D Gaussian at the target pixel location. c_i is the view-dependent color of Gaussian G_i represented by spherical harmonic coefficients.

We denote Gaussian properties as $G = \{\mu, r, s, \sigma, c\}$. After rasterization, the Gaussian properties are optimized through appearance and other losses to obtain a 3D representation of the scene. In addition, the adaptive control of the Gaussian can improve its representation ability, mainly including three operations: splitting, cloning and pruning. Splitting and cloning are performed on Gaussians with large position gradients, which augments the number of Gaussians. Pruning operation periodically eliminates Gaussians with excessively small opacity to suppress floating artifacts.

3.2. Gaussian initialization on mesh

Our method is a hybrid representation of avatar with Gaussians embedded on mesh [42, 45, 52]. Given a set of monocular images, the registered mesh corresponding to each frame can be obtained by deforming the neutral SMPL template via shape and pose parameters. The mesh consists of a set of vertices $V = \{v_i\}_1^n$ and faces $F = \{f_j\}_1^m$.

Given a triangle with vertices (v_a, v_b, v_c) and normal $\boldsymbol{n}$, we bind a Gaussian to the triangle. Specifically, we construct a local coordinate system with the mean position $\boldsymbol{M}$ of three vertices as the origin. The local coordinate system is determined by an edge vector $\boldsymbol{e} = v_b - v_a$, the normal of the triangle $\boldsymbol{n}$, and their cross product $\boldsymbol{e} \times \boldsymbol{n}$. These three vectors constitute a rotation matrix $\boldsymbol{R}$ of the triangle in the global space, describing the orientation of the triangle. The scaling $\boldsymbol{w}$ of triangle is measured by the mean length of one edge $\boldsymbol{e}$ and its perpendicular $\boldsymbol{e_p}$.

We initialize the position of the Gaussian on triangle to the local origin, the rotation $\boldsymbol{r}$ to identity matrix, and the scaling to unit vector. We then convert these properties from local coordinate system of the triangle to the global coordinate system. The global position, rotation and scaling of the Gaussian μ^*, $\boldsymbol{r}^*$ and s^* are computed by:

$$\mu^* = \boldsymbol{w}\boldsymbol{R}\mu + \boldsymbol{M}, \tag{5}$$

$$\boldsymbol{r}^* = \boldsymbol{R}\boldsymbol{r}, \tag{6}$$

$$s^* = \boldsymbol{w}s. \tag{7}$$

Assigning a single Gaussian to each triangle is not enough to capture complex details. For example, there may be only one triangle under the distorted clothing or curly hair, and a single Gaussian splat on it is not enough to represent such complex appearance. We adaptively adjust the density of Gaussian splats on mesh using strategies such as splitting, cloning and pruning. Specifically, the splitting operation is performed on Gaussians with magnitude of scaling matrix larger than a threshold, and cloning operation is performed with magnitude of scaling matrix smaller than a threshold. The pruning operation is periodically applied to reset the opacity of all Gaussians close to zero and removes Gaussians with opacity below a threshold. A Gaussian stores the index of its parent triangle, and we make sure that every triangle has at least one Gaussian attached to it.

3.3. Gaussian rectification for non-rigid deformation

Gaussian motion guided by triangle mesh can capture rough human movements, but constrained by the linear representation of LBS, Gaussians that move with the mesh are insufficient for capturing non-rigid distortions and intricate dynamic textures. We decompose complex human motion to rigid transformation guided by human skeleton and non-rigid transformation caused by pose-dependent cloth distortions [53]. It is necessary to design a separate module for non-rigid deformation to further adjust the Gaussian properties dependent on poses.

Given a human pose $\boldsymbol{P}$ in current observation and Gaussian positions μ^* on the posed mesh, we formulate the non-rigid deformation module $\mathcal{F}_\theta$ based on MLP as:

$$(\delta\mu, \delta r, \delta s) = \mathcal{F}_\theta(\gamma(\mu^*), \boldsymbol{P}). \tag{8}$$

We use an embedding function $\gamma(\cdot)$ to encode Gaussian positions with a specific frequency. The change of Gaussian attributes is related to the pose of each frame, which enables Gaussian to learn the non-rigid changes caused by pose, making up for the shortcomings of LBS direct linear representation and further improving the realism of avatar. Based on the predicted offset, the rectified Gaussian attributes are calculated as follows:

$$\mu' = \mu^* + \delta\mu, \tag{9}$$
$$r' = r^* \cdot \delta r, \tag{10}$$
$$s' = s^* + \delta s. \tag{11}$$

Here, the updated μ', r' and s' are in global space, and r^* is the quaternion vector corresponding to the rotation in Equation 6. Applying $\cdot$ operation to the quaternion vectors is equivalent to multiplying the corresponding rotation matrices. Combined with opacity and color, we get the Gaussian property set $G = \{\mu', r', s', \sigma, c\}$ for subsequent rasterization.

Based on above analysis, the motion guidance of the mesh ensures that Gaussians learn the geometry of the avatar, while the pose-based recitification enables the Gaussians to model the appearance changes dependent on pose. By modeling both rigid and non-rigid deformations, our approach is able to reconstruct more realistic avatars.

3.4. Optimization

We utilize RGB loss $\mathcal{L}_{rgb}$, SSIM loss [50] $\mathcal{L}_{ssim}$ and LPIPS loss [63] $\mathcal{L}_{lpips}$ with the corresponding weights λ_{rgb}, λ_{ssim} and λ_{lpips} to optimize the rendered images:

$$\mathcal{L}_{color} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{ssim}\mathcal{L}_{ssim} + \lambda_{lpips}\mathcal{L}_{lpips}. \tag{12}$$

Note that $\mathcal{L}_{rgb}$, $\mathcal{L}_{ssim}$ and $\mathcal{L}_{lpips}$ are L_1-norm losses. Besides, we apply some regularization terms to constrain the position and covariance learning for Gaussians.

At initialization, we bind a Gaussian to the local origin of the triangle. During training, the position of Gaussian splat is optimized and may deviate away from its parent triangle. Large drifts do not affect rendering quality from visible perspectives, but may cause artifacts when animating the Gaussians to a new pose via SMPL. Following [42], we regularize the local position of each Gaussian by:

$$\mathcal{L}_{pos} = \| \max(\mu - \epsilon_{pos}, 0) \|_2, \tag{13}$$

where ϵ_{pos} is set to 1 to constrain the Gaussians around their parent triangles and give them certain freedom to adjust position. In addition to the position, the scaling of the Gaussian also affects the rendering quality. A Gaussian that is too large compared to the parent triangle is sensitive to the motion of triangle, thus can easily introduce jitter and artifacts when rotating with the mesh. We regularize the scaling of each Gaussian and utilize the threshhold $\epsilon_{scaling}$ to prevent it from shrinking excessively:

$$\mathcal{L}_{scaling} = \| \max(s - \epsilon_{scaling}, 0) \|_2, \tag{14}$$

where $\epsilon_{scaling}$ is set to 0.6 for the maximum allowable scaling. When μ and s are below these thresholds, the corresponding loss terms are disabled.

The Gaussians on the mesh are able to capture the rough motion and appearance of the avatar, and the rectification module is proposed only to slightly adjust the Gaussian properties to support its learning of complex details caused by non-rigid deformations. Thus we design a regularization term $\mathcal{L}_{offset}$ to constrain the values of $\delta\mu$, δr and δs:

$$\mathcal{L}_{offset} = \|(\delta\mu, \delta r, \delta s)\|_2. \tag{15}$$

Taking the color losses and regularization losses together, we define the final loss function as follows:

$$\begin{aligned}\mathcal{L} = {} & \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{ssim}\mathcal{L}_{ssim} + \lambda_{lpips}\mathcal{L}_{lpips} \\ & + \lambda_{pos}\mathcal{L}_{pos} + \lambda_{scaling}\mathcal{L}_{scaling} + \lambda_{offset}\mathcal{L}_{offset}.\end{aligned} \tag{16}$$

Here, the $\mathcal{L}_{pos}$, $\mathcal{L}_{scaling}$ and $\mathcal{L}_{offset}$ are L_2-norm losses with corresponding weights λ_{pos}, $\lambda_{scaling}$ and λ_{offset}.

4. Experiments

In this section, we evaluate RMAvatar on PeopleSnapshot dataset [2], ZJU-MoCap dataset [39], and DynVideo [13] dataset by comparing with the state-of-the-art human avatar modeling methods in monocular videos, and systematically ablate each component of RMAvatar.

4.1. Datasets and metrics

PeopleSnapshot [2] dataset. We select 4 sequences of the PeopleSnapshot dataset as in InstantAvatar [15] and follow the same data split. We compare our approach with

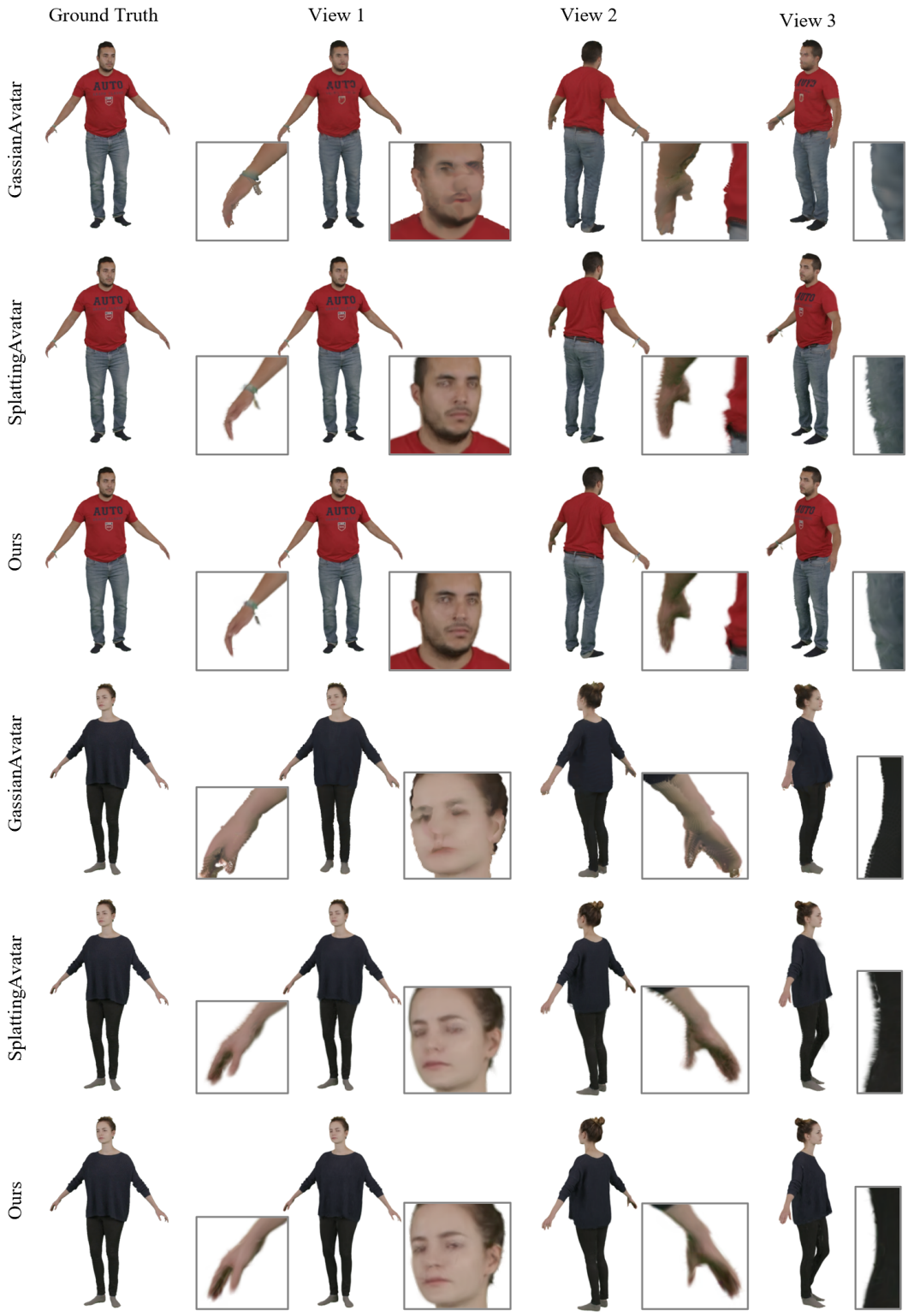

Figure 2: Comparison of novel view synthesis on PeopleSnapshot [2]. Our method is able to reconstruct intricate texture details.

Table 1: Train/test split of the ZJU-MoCap [39] dataset.

	train	test		train	test
377	1-456	457-617	386	1-456	457-646
387	1-456	457-654	392	1-456	457-556
393	1-456	457-658	394	1-656	657-859

Anim-NeRF [4], InstantAvatar [15], GaussianAvatar [13] and SplattingAvatar [45] on this dataset. Like SplattingAvatar, we use the poses optimized by Anim-NeRF [4].

ZJU-MoCap [39] dataset. We use six subjects (377, 386, 387, 392, 393, and 394) [53] of the ZJU-MoCap dataset. We select a video from a camera viewpoint and use a segment of the video during training. Specifically, we select a clip that contains a complete turning action of the avatar for training. We show the training/test split in Table 1 and use the same split in comparison with other methods. Note that due to the small motion of the characters, we select 1 frame out of every 4 frames during training. We use Anim-NeRF [38] to obtain the refined poses of ZJU-MoCap.

DynVideo [13] dataset. DynVideo dataset proposed by GaussianAvatar records humans via a mobile phone with loose clothing and large movements, which serves as a valuable resource for evaluating reconstruction quality.

Evaluation metrics. We consider three metrics: PSNR, SSIM and LPIPS to evaluate the reconstruction quality, which measure the pixel intensity similarity, structural similarity, and perceptual image patch similarity between the rendered and ground truth images.

4.2. Implementation details

Our model is trained for 50,000 iterations on a single NVIDIA RTX 3090 GPU. During training, we use Adam [20] to optimize our model. We set the learning rate to 0.008 for the position and exponentially decay it with a factor of 0.01 to 10^{-5}. The learning rate for the scaling, rotation and opacity of 3D Gaussians are 0.017, 0.001 and 0.05 respectively. The learning rate for the Gaussian rectification module is 10^{-4}. For densification and pruning for the 3D Gaussians, we apply density control operations every 500 iterations, and reset the opacity of Gaussians every 5,000 iterations from iteration 10,000, and turn off the density controller as well as the opacity-rest operation at iteration 35,000. For the target loss, we set the parameters of different loss terms to $\lambda_{rgb} = 1.5$, $\lambda_{ssim} = 0.2$, $\lambda_{lpips} = 0.05$, $\lambda_{pos} = 0.01$, $\lambda_{scaling} = 1$ and $\lambda_{offset} = 1$, respectively.

For the Gaussian rectification module, we designed a 5-layer MLP to predict the offsets of Gaussian attributes such as position, rotation, and scaling. The MLP network takes the encoded Gaussian position and the pose parameters as input, which contains 105 channel features. The input channels for subsequent hidden layer are (128, 164, 128), and contains a skip connection in the fourth layer. The output of the last layer of the MLP is a 10-dimensional feature, which contains the position offset δu, scale offset δs, and quaternion offset δr. In order to limit the offset value to be small, we design offset loss for these three attributes.

4.3. Comparisons on human avatar reconstruction

We evaluate the reconstruction quality of human avatars by novel view synthesis and avatar animation experiments on two datasets. In Table 2 and Table 3 we report quantitative results of different methods on novel view synthesis on PeopleSnapshot and ZJU-MoCap, respectively. Our method far exceeds the SoTA methods in terms of PSNR, SSIM, and LPIPS, which reflect rendering quality of avatars, indicating that our method can learn the complex textures more clearly.

Novel view synthesis. We show the novel view synthesis results of GaussianAvatar, SplattingAvatar and our method in Figure 2. From the zoomed-in details of the face and hands, it is obvious that our method produces clear details and smooth boundaries. The contour of the reconstructed human avatar of GaussianAvatar [13] is obviously larger than the ground truth, which indicates that the position of the Gaussians it learned are not accurate, and it is difficult to capture big motion and intricate textures, such as human face and cloth wrinkles. The rendering images of SplattingAvatar [45] are better than GaussianAvatar, which shows the advantages of hybrid avatar representation with Gaussians on mesh. SplattingAvatar embeds Gaussians on canonical mesh and guides the motion of Gaussians by mesh warping from canonical space to posed space. As pointed out by SplattingAvatar, the quality of its reconstruction is highly dependent on the motion accuracy of the underlying mesh. Due to the lack of cloth and hair meshes, they failed to learn clear texture of head and cloth. In addition, this method produces obvious artifacts on the avatar surface due to inaccurate Gaussian positions, which is probably because Walking on mesh strategy does not constrain the amplitude of Gaussian motion effectively. In contrast, our approach designs a separate module to learn non-rigid cloth and hair deformations, which makes up for the defect that Gaussians on SMPL cannot represent clothes, thereby improving the reconstruction accuracy of rendered avatars.

The comparison of novel view synthesis on the ZJU-MoCap dataset can better reflect the advantages of our method. The avatars have large movements and there are certain errors in estimated poses, which leads to poor visual effects of hybrid methods combining mesh and Gaussian, such as SplattingAvatar. Our method further adjusts the Gaussian properties on the basis of mesh guidance, reduces the reconstruction error caused by inaccurate pose, and ensures the high quality of Avatar.

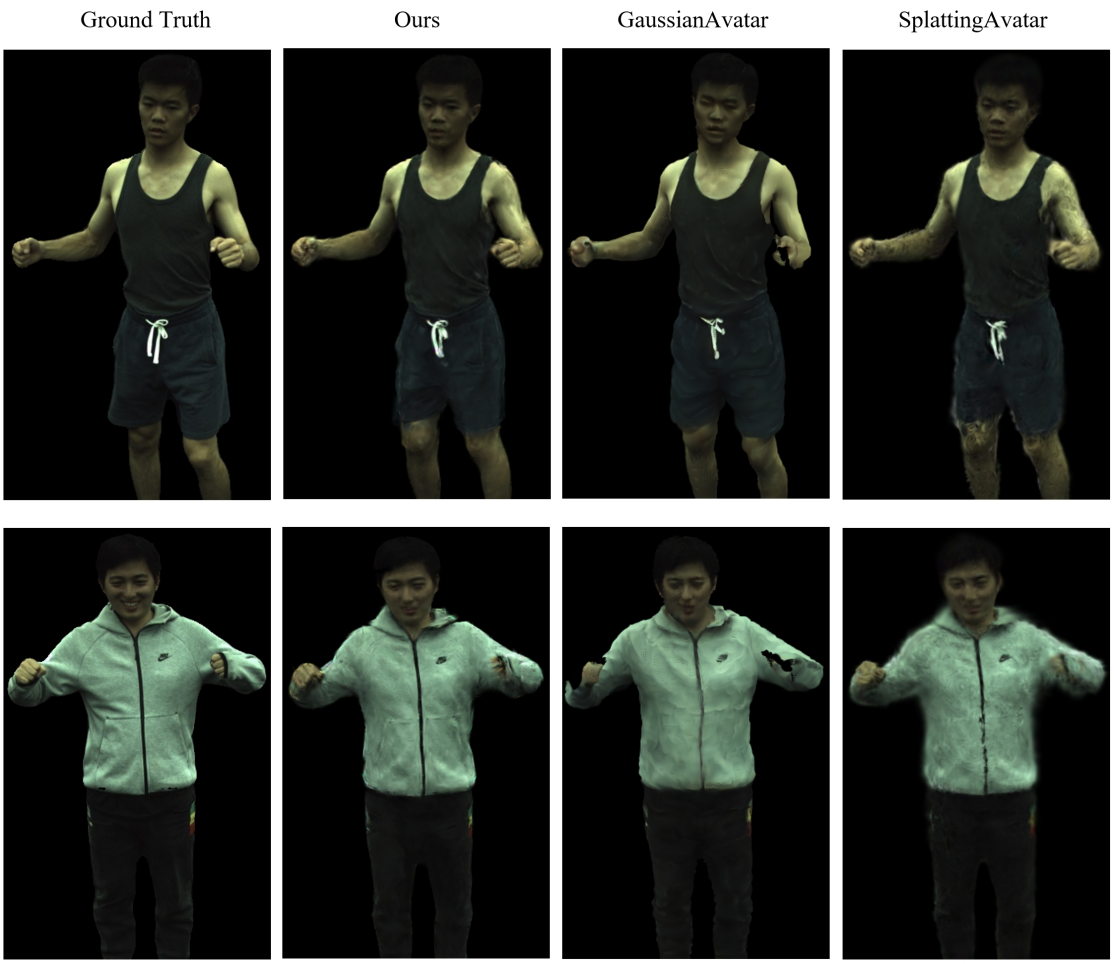

Figure 3: Comparison of novel view synthesis on ZJU-MoCap [39]. Our method reconstructs complicated cloth textures.

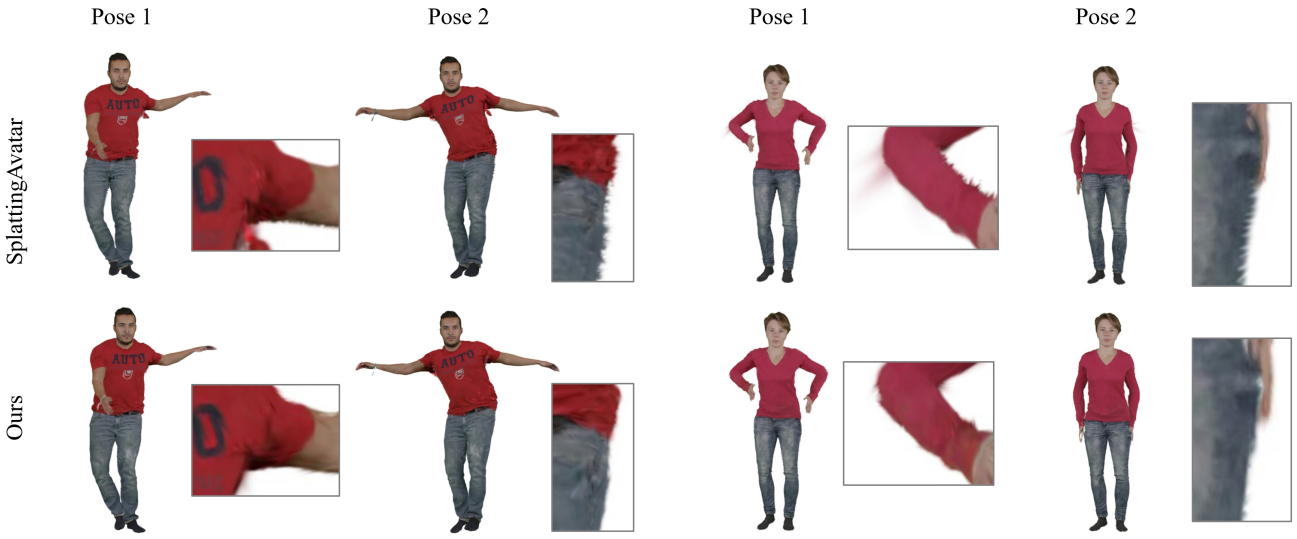

Figure 4: Comparison of animation on out-of-distribution poses on PeopleSnapshot [2]. Our method generates consistent representations for avatars on novel poses.

Table 2: Quantitative comparison on PeopleSnapshot [2]. Compared with state-of-the-art methods, our method achieves significant improvement in rendering quality of novel view synthesis on all evaluation metrics. The best results are in bold.

	male-3-casual			male-4-casual			female-3-casual			female-4-casual		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Anim-NeRF [4]	29.37	0.970	0.017	28.37	0.960	0.027	28.91	0.974	0.022	28.90	0.968	0.017
InstantAvatar [15]	30.91	0.977	0.022	29.77	0.980	0.025	29.73	0.975	0.025	30.92	0.977	0.021
GaussianAvatar [13]	30.98	0.979	0.015	28.78	0.975	0.023	29.55	0.976	0.023	30.84	0.977	0.014
SplattingAvatar [45]	32.31	0.978	0.031	30.51	0.978	0.041	30.42	0.976	0.044	31.12	0.976	0.032
Ours	**34.12**	**0.985**	**0.013**	**31.23**	**0.983**	**0.022**	**31.42**	**0.980**	**0.021**	**33.06**	**0.982**	**0.013**

Table 3: Quantitative comparison on ZJU-MoCap [39]. Our method achieves best reconstruction quality on novel view synthesis of avatars with large motions. The best results are in bold.

	377			386			387			392			393			394		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GaussianAvatar [13]	24.86	0.944	0.063	27.10	0.923	0.074	25.63	0.947	0.043	26.18	0.929	0.088	23.90	0.919	0.099	26.11	0.925	0.084
SplattingAvatar [45]	32.24	0.977	0.028	30.31	0.953	0.073	30.69	0.967	0.045	33.41	0.975	0.040	30.02	0.962	0.047	32.36	0.965	0.044
Ours	**32.68**	**0.982**	**0.015**	**30.61**	**0.955**	**0.055**	**31.01**	**0.973**	**0.021**	**34.30**	**0.979**	**0.024**	**31.09**	**0.969**	**0.025**	**32.70**	**0.969**	**0.025**

Table 4: Quantitative comparison on DynVideo [13]. Our method achieves better reconstruction for avatars with loose clothes and large motions. The best results are in bold.

	Male			Female		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GaussianAvatar [13]	24.42	0.9505	**0.0247**	22.69	0.9335	**0.0445**
SplattingAvatar [45]	23.58	0.9329	0.0778	23.06	0.9311	0.0943
Ours	**24.92**	**0.9511**	0.0356	**23.11**	**0.9336**	0.0707

Table 5: Training and running time comparisons.

Methods	Training Time	Running Time
Anim-NeRF [4]	$\sim$ 25 hs	$\sim$ 0.03 FPS
InstantAvatar [15]	$\sim$ 2 mins	$\sim$ 3.5 FPS
GaussianAvatar [13]	$\sim$ 38 mins	$\sim$ 15 FPS
SplattingAvatar [45]	$\sim$ 40 mins	$\sim$ 270 FPS
Ours	$\sim$ 50 mins	$\sim$ 210 FPS

Table 4 presents the comparative results for novel view synthesis on the DynVideo dataset, with pose parameters sourced from GaussianAvatar. The results indicate that our method achieves superior performance in terms of PSNR and SSIM. However, the LPIPS metric of our method and SplattingAvatar are inferior to GaussianAvatar. This is mainly due to the inaccurate pose estimation for humans with large motions. Hybrid Gaussian representations [42, 45] utilize SMPL parameters as the geometric prior of the Gaussians, the perceptual consistency of the rendered image is affected in parts with inaccurate postures such as hands, feet, and head.

Efficiency. Here we compare the inference speed of RMAvatar with other methods. As shown in Table 5, we measure the training and running time in the PeopleSnapshot dataset. Our method achieves an ultra-high rendering frame rate of 210 FPS with a training time 50 minutes. Our time cost is similar to SplattingAvatar, but produces higher reconstruction quality.

Avatar animation. In Figure 4, we show the avatar animation results for out-of-distribution poses generated by SplattingAvatar and our method. It is clear that our method demonstrates a more consistent 3D appearance and shape of avatars under challenging novel poses, while SplattingAvatar generates obvious wrong positioned Gaussians and non-smooth boundaries for large motion sequence.

4.4. Ablation study

We conduct ablation studies by comparing the full model with 1) removing the Gaussian rectification module, denoted as "w/o GauRec", 2) only removing the regularization term $\mathcal{L}_{offset}$, denoted as "w/o OffsetLoss" and 3) only removing the scaling loss term $\mathcal{L}_{scaling}$, denoted as "w/o ScalingLoss". The "w/o GauRec" model fixes the position of the Gaussian to its parent triangle, and properties such as rotation and scale of the Gaussian remain consistent in all poses. Specifically, the rectification MLP is removed in "w/o GauRec" model. The "w/o OffsetLoss" and "w/o ScalingLoss" models contain rectification MLP but discard the regularization loss $\mathcal{L}_{offset}$ and scaling loss $\mathcal{L}_{scaling}$, respectively.

The ablation results are shown in Table 6 and Figure 5. The results in Table 6 show that the GauRec module can significantly improve PSNR and SSIM by adjusting Gaussian properties to learn more accurate non-rigid distortions and complex dynamic textures (such as clothes and hair). The result in the second column of Figure 5 shows that removing the GauRec module results in burr textures on loose

Figure 5: Ablation on female-3-casual of PeopleSnapshot. Our full model mitigates rendering artifacts with Gaussian rectification module and regularization losses.

Table 6: **Ablation Study on PeopleSnapshot [2].** The Gaussian rectification module and regularization losses improve reconstruction quality of clothed avatars. The best results are in bold.

Metric:	PSNR↑	SSIM↑	LPIPS↓
w/o GauRec	31.92	0.974	0.017
w/o OffsetLoss	32.48	**0.983**	0.017
w/o ScalingLoss	32.45	**0.983**	0.017
Full model	**32.51**	**0.983**	0.017

clothing. However, large offsets predicted by the GauRec module may affect the consistency of the dynamic avatar representation. To address this issue, we introduce the offset regularization loss. The comparison between the "w/o OffsetLoss" model and our full model in Figure 5 shows that with the offset loss, our method can generate subtle Gaussian offsets to maintain the consistency of Gaussian splats and more accurately capture deformation details, resulting in improved rendering quality and color fidelity.

The scaling loss is to regularize the scaling of Gaussians and reduce Gaussians with elongated shapes. As shown in Table 6, without scaling loss, the PSNR on PeopleSnapshot decreases. The fourth column in Figure 5 shows that elongated Gaussian splats cause severe rendering artifacts.

5. Conclusion and Discussion

In this paper, we have proposed a hybrid representation for human avatar from monocular video based on mesh-embedded Gaussian splats. We utilize the explicit representation of the mesh to ensure correct shape and motion of avatar, and the implicit Gaussian splats located on mesh to render photorealistic appearance. To compensate for the limitation of linear transformations of LBS, we design a non-rigid deformation module that takes pose and Gaussian positions as input to further optimize the Gaussian properties, thereby learning the dynamic effects of flexible materials such as clothing and hair. Compared with the SoTA methods, our method achieves the best performance in both rendering quality and measurement indicators, and the reconstruction based on monocular video can further promote the application of avatars in multiple fields such as telepresentation and virtual reality.

Our method learns non-rigid deformation based on mesh-guided Gaussian motion, and reduces reconstruction errors caused by inaccurate poses. However, due to the lack of meshes for clothing and facial expressions, our method has limited ability to model these details. In the future, we consider combining layered meshes and Gaussians to further improve the reconstruction of complex details such as the head, clothing, and hands.

Acknowledgments

Sen Peng and Baorong Yang were partially supported by Scientific Research Start-up Fund of Jimei University (ZQ2021002) and Fujian Provincial Natural Science Foundation General Fund (2024J01118). Xiao Dong was partially supported by the Guangdong Provincial Key Laboratory of IRADS (2022B1212010006) and Guangdong Higher Education Upgrading Plan (2021-2025) (2023KQNCX091, 2024KTSCX223). Zhonggui Chen was partially supported by the National Natural Science Foundation of China (62372389), and the Natural Science Foundation of Fujian Province, China (2022J01001). Zilong Wang and Xiaohu Guo were partially supported by National Science Foundation (OAC-2007661).

References

[1] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 2

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 5, 6, 8, 9, 10

[3] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su. Mvsnerf: Fast generalizable radiance field recon-

struction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 3

[4] J. Chen, Y. Zhang, D. Kang, X. Zhe, L. Bao, X. Jia, and H. Lu. Animatable neural radiance fields from monocular rgb videos. *arXiv preprint arXiv:2106.13629*, 2021. 2, 7, 9

[5] X. Chen, T. Jiang, J. Song, M. Rietmann, A. Geiger, M. J. Black, and O. Hilliges. Fast-snarf: A fast deformer for articulated neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11796–11809, 2023. 2

[6] Y. Chen, L. Wang, Q. Li, H. Xiao, S. Zhang, H. Yao, and Y. Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 1, 2

[7] Y. Chen, L. Wang, Q. Li, H. Xiao, S. Zhang, H. Yao, and Y. Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 2

[8] Y. Chen, Z. Zheng, Z. Li, C. Xu, and Y. Liu. Meshavatar: Learning high-quality triangular human avatars from multi-view videos. *arXiv preprint arXiv:2407.08414*, 2024. 3

[9] J. Chibane, A. Bansal, V. Lazova, and G. Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7911–7920, 2021. 3

[10] Y. Feng, W. Liu, T. Bolkart, J. Yang, M. Pollefeys, and M. J. Black. Learning disentangled avatars with hybrid 3d representations. *arXiv preprint arXiv:2309.06441*, 2023. 3

[11] C. Geng, S. Peng, Z. Xu, H. Bao, and X. Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8759–8770, 2023. 1, 3

[12] M. Habermann, L. Liu, W. Xu, G. Pons-Moll, M. Zollhoefer, and C. Theobalt. Hdhumans: A hybrid approach for high-fidelity digital humans. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–23, 2023. 3

[13] L. Hu, H. Zhang, Y. Zhang, B. Zhou, B. Liu, S. Zhang, and L. Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 634–644, 2024. 1, 2, 3, 5, 7, 9

[14] M. Işık, M. Rünz, M. Georgopoulos, T. Khakhulin, J. Starck, L. Agapito, and M. Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics*, 42(4):1–12, 2023. 2

[15] T. Jiang, X. Chen, J. Song, and O. Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16922–16932, 2023. 2, 3, 5, 7, 9

[16] W. Jiang, K. M. Yi, G. Samei, O. Tuzel, and A. Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European Conference on Computer Vision*, pages 402–418. Springer, 2022. 1, 2

[17] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 1

[18] R. Kachach, P. Perez, A. Villegas, and E. Gonzalez-Sosa. Virtual tour: An immersive low cost telepresence system. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops*, pages 504–506. IEEE, 2020. 1

[19] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139–1, 2023. 1, 3, 4

[20] D. Kinga, J. B. Adam, et al. A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, volume 5, page 6. San Diego, California;, 2015. 7

[21] M. Kocabas, J.-H. R. Chang, J. Gabriel, O. Tuzel, and A. Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 505–515, 2024. 3

[22] Y. Kwon, L. Liu, H. Fuchs, M. Habermann, and C. Theobalt. Deliffas: Deformable light fields for fast avatar synthesis. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[23] C. Lassner and M. Zollhofer. Pulsar: Efficient sphere-based neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1440–1449, 2021. 1

[24] J. Lei, Y. Wang, G. Pavlakos, L. Liu, and K. Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19876–19887, 2024. 3

[25] J. P. Lewis, M. Cordner, and N. Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pages 165–172, 2000. 2

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

[27] Z. Li, Z. Zheng, L. Wang, and Y. 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*, pages 19711–19722, 2024. 3

[28] Z. Li, Z. Zheng, H. Zhang, C. Ji, and Y. Liu. Avatarcap: Animatable avatar conditioned monocular human volumetric capture. In *European Conference on Computer Vision*, pages 322–341. Springer, 2022. 2

[29] S. Lin, Z. Li, Z. Su, Z. Zheng, H. Zhang, and Y. Liu. Layga: Layered gaussian avatars for animatable clothing transfer. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[30] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2, 4

[31] M. Mihajlovic, Y. Zhang, M. J. Black, and S. Tang. Leap:

Learning articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10461–10471, 2021. 2

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

[33] A. Moreau, J. Song, H. Dhamo, R. Shaw, Y. Zhou, and E. Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 788–798, 2024. 2

[34] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):1–15, 2022. 3

[35] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 1, 2

[36] M. Oechsle, S. Peng, and A. Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 1

[37] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2

[38] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, and H. Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 1, 2, 7

[39] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 2, 5, 7, 8, 9

[40] S. Prokudin, Q. Ma, M. Raafat, J. Valentin, and S. Tang. Dynamic point fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7964–7976, 2023. 3

[41] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 1

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

[43] Z. Qian, S. Wang, M. Mihajlovic, A. Geiger, and S. Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5020–5030, 2024. 2, 3

[44] R. Shao, J. Sun, C. Peng, Z. Zheng, B. Zhou, H. Zhang, and Y. Liu. Control4d: Efficient 4d portrait editing with text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4556–4567, 2024. 3

[45] Z. Shao, Z. Wang, Z. Li, D. Wang, X. Lin, Y. Zhang, M. Fan, and Z. Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1606–1616, 2024. 2, 3, 4, 7, 9

[46] S.-Y. Su, T. Bagautdinov, and H. Rhodin. Npc: Neural point characters from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14795–14805, 2023. 3

[47] J. Wang, J. Fang, X. Zhang, L. Xie, and Q. Tian. Gaussianeditor: Editing 3d gaussians delicately with text instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20902–20911, 2024. 3

[48] L. Wang, J. Zhang, X. Liu, F. Zhao, Y. Zhang, Y. Zhang, M. Wu, J. Yu, and L. Xu. Fourier plenoctrees for dynamic radiance field rendering in real-time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13524–13534, 2022. 3

[49] S. Wang, K. Schwarz, A. Geiger, and S. Tang. Arah: Animatable volume rendering of articulated human sdfs. In *Proceedings of the European Conference on Computer Vision*, pages 1–19. Springer, 2022. 2

[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[51] F. Weidner, G. Boettcher, S. A. Arboleda, C. Diao, L. Sinani, C. Kunert, C. Gerhardt, W. Broll, and A. Raake. A systematic review on the visualization of avatars and agents in ar & vr displayed using head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2596–2606, 2023. 1

[52] J. Wen, X. Zhao, Z. Ren, A. G. Schwing, and S. Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2059–2069, 2024. 2, 3, 4

[53] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 1, 2, 5, 7

[54] O. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 1

[55] C.-h. Wuu, N. Zheng, S. Ardisson, R. Bali, D. Belko, E. Brockmeyer, L. Evans, T. Godisart, H. Ha, X. Huang, et al. Multiface: A dataset for neural face rendering. *arXiv preprint arXiv:2207.11243*, 2022. 1

[56] J. Xiang, X. Gao, Y. Guo, and J. Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1802–1812, 2024. 2

[57] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 3

[58] Y. Xu, B. Chen, Z. Li, H. Zhang, L. Wang, Z. Zheng, and Y. Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2024. 2

[59] Y. Xu, Z. Su, Q. Wu, and Y. Liu. Gphm: Gaussian parametric head model for monocular head avatar reconstruction. *arXiv preprint arXiv:2407.15070*, 2024. 3

[60] Z. Xu, S. Peng, C. Geng, L. Mou, Z. Yan, J. Sun, H. Bao, and X. Zhou. Relightable and animatable neural avatar from sparse-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 990–1000, 2024. 2

[61] Z. Yu, W. Cheng, X. Liu, W. Wu, and K.-Y. Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023. 2

[62] Q. Zhang, S.-H. Baek, S. Rusinkiewicz, and F. Heide. Differentiable point-based radiance fields for efficient view synthesis. In *Siggraph Asia 2022 Conference Papers*, pages 1–12, 2022. 1

[63] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5

[64] F. Zhao, Y. Jiang, K. Yao, J. Zhang, L. Wang, H. Dai, Y. Zhong, Y. Zhang, M. Wu, L. Xu, et al. Human performance modeling and rendering via neural animated mesh. *ACM Transactions on Graphics*, 41(6):1–17, 2022. 3

[65] Y. Zheng, V. F. Abrevaya, M. C. Bühler, X. Chen, M. J. Black, and O. Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13545–13555, 2022. 2

[66] Y. Zheng, W. Yifan, G. Wetzstein, M. J. Black, and O. Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21057–21067, 2023. 1, 3

[67] Y. Zheng, Q. Zhao, G. Yang, W. Yifan, D. Xiang, F. Dubost, D. Lagun, T. Beeler, F. Tombari, L. Guibas, and G. Wetzstein. Physavatar: Learning the physics of dressed 3d avatars from visual observations. In *Proceedings of the European Conference on Computer Vision*, pages 262–284, 2024. 2

[68] Z. Zheng, H. Huang, T. Yu, H. Zhang, Y. Guo, and Y. Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15893–15903, 2022. 3

[69] M. Zwicker, H. Pfister, J. Van Baar, and M. Gross. Ewa volume splatting. In *Proceedings of the IEEE Conference on Visualization 2001*, pages 29–538. IEEE, 2001. 4