

# 3DGS-Avatar: Animatable Avatars via Deformable 3D Gaussian Splatting

Zhiyin Qian<sup>1</sup> Shaofei Wang<sup>1,2,3</sup> Marko Mihajlovic<sup>1</sup> Andreas Geiger<sup>2,3</sup> Siyu Tang<sup>1</sup>  
<sup>1</sup>ETH Zürich <sup>2</sup>University of Tübingen <sup>3</sup>Tübingen AI Center

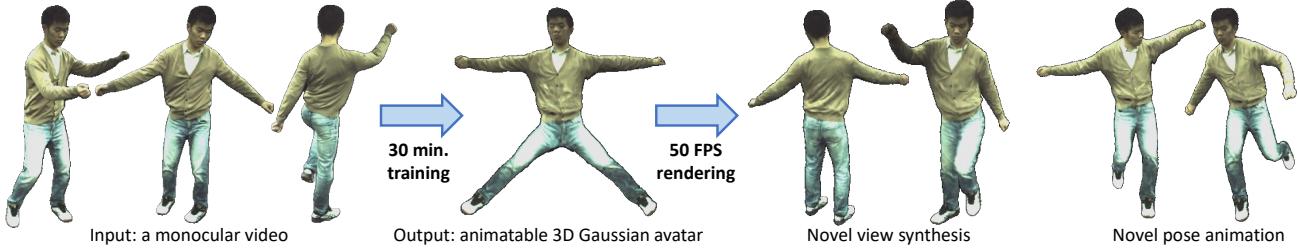


Figure 1. **3DGS-Avatar.** We develop an efficient method for creating animatable avatars from monocular videos, leveraging 3D Gaussian Splatting [14]. Given a short sequence of dynamic human with a tracked skeleton and foreground masks, our method creates an avatar within **30 minutes** on a single GPU, supports animation and novel view synthesis at over **50 FPS**, and achieves comparable or better rendering quality to the state-of-the-art [61, 62] that requires over 8 GPU days to train, takes several seconds to render a single image, and relies on pre-training on clothed human scans [61].

## Abstract

We introduce an approach that creates animatable human avatars from monocular videos using 3D Gaussian Splatting (3DGS). Existing methods based on neural radiance fields (NeRFs) achieve high-quality novel-view/novel-pose image synthesis but often require days of training, and are extremely slow at inference time. Recently, the community has explored fast grid structures for efficient training of clothed avatars. Albeit being extremely fast at training, these methods can barely achieve an interactive rendering frame rate with around 15 FPS. In this paper, we use 3D Gaussian Splatting and learn a non-rigid deformation network to reconstruct animatable clothed human avatars that can be trained within 30 minutes and rendered at real-time frame rates (50+ FPS). Given the explicit nature of our representation, we further introduce as-isometric-as-possible regularizations on both the Gaussian mean vectors and the covariance matrices, enhancing the generalization of our model on highly articulated unseen poses. Experimental results show that our method achieves comparable and even better performance compared to state-of-the-art approaches on animatable avatar creation from a monocular input, while being 400x and 250x faster in training and inference, respectively. Please see our project page at <https://neuralbodies.github.io/3DGS-Avatar>.

## 1. Introduction

Reconstructing clothed human avatars from image inputs presents a significant challenge in computer vision, yet holds immense importance due to its applications in virtual reality, gaming, and e-commerce. Traditional methods often rely on dense, synchronized multi-view inputs, which may not be readily available in more practical scenarios. Recent advances in implicit neural fields [29, 33, 35, 39, 51, 54, 55, 57, 59, 69, 70] have enabled high-quality reconstruction of geometry [8, 41, 61, 65] and appearance [13, 22, 24, 34, 38, 40, 45, 62, 74] of clothed human bodies from sparse multi-view or monocular videos. Animation of such reconstructed clothed human bodies is also possible by learning the geometry and appearance representations in a predefined canonical pose [13, 22, 38, 61, 62, 74].

To achieve state-of-the-art rendering quality, existing methods rely on training a neural radiance field (NeRF) [29] combined with either explicit body articulation [8, 12, 13, 22, 38, 41, 61, 62, 74] or conditioning the NeRF on human body related encodings [34, 40, 52, 65]. They often employ large multi-layer perceptrons (MLPs) to model the neural radiance field, which are computationally demanding, leading to prolonged training (days) and inference (seconds) time. This computational expense poses a significant challenge for practical applications of these state-of-the-art methods in real-time applications.

With recent advances in efficient learning of implicit

neural fields, training time of NeRFs has been reduced to minutes [3, 21, 32, 50, 56]. There are also works targeting fast inference of pretrained NeRFs [47, 71, 73]. Inspired by these developments, several avatar reconstruction methods have been tailored to fast training [7, 12] or fast inference [6, 18, 42]. However, to the best of our knowledge, there currently exists no published method that simultaneously achieves both fast training and *real-time* inference for *animatable* avatar reconstruction from just monocular videos.

Point-based rendering [43, 48, 53, 66, 75, 77, 78] has emerged as an efficient alternative to NeRFs for fast inference. With the recently proposed 3D Gaussian Splatting (3DGS) [14] model, it is possible to achieve state-of-the-art rendering quality using only a fraction of NeRFs’ inference time and comparatively fast training for static scene reconstruction.

Leveraging the capabilities of 3DGS, we demonstrate its application in modeling animatable clothed avatars using monocular videos. Our approach effectively integrates rigid human articulation with a non-rigid deformation field within the 3DGS framework. We use a small multi-layer perceptron (MLP) to decode color. This MLP is designed to be responsive to local non-rigid deformations and dynamic lighting conditions, ensuring a more realistic and responsive rendering of the avatar’s appearance. Furthermore, we apply as-isometric-as-possible regularizations [15, 44] to *both* the Gaussian mean vectors and the covariance matrices, which helps maintain the geometric consistency and realistic deformation of the avatar, particularly in dynamic and varied poses.

Our experimental results show that our method is comparable to or better than current state-of-the-art [61, 62] in animatable avatar creation from monocular inputs, achieving training speed 400 times faster and inference speed 250 times quicker. Compared to methods that focus on fast training [7, 12], our method, despite being slower in training, can model pose-dependent non-rigid deformation and produce significantly better rendering quality, while being 3 times faster in terms of rendering. We provide an overview of the comparison to major prior works in Tab. 1. In summary, our work makes the following contributions:

- We introduce 3D Gaussian Splatting to animatable human avatars reconstruction from monocular videos.
- We develop a simple yet effective deformation network as well as regularization terms that effectively drive 3D Gaussian Splats to handle highly articulated and out-of-distribution poses.
- Our method is the first, to our knowledge, to simultaneously deliver high-quality rendering, model pose-dependent non-rigid deformation, generalize effectively to unseen poses, and achieve fast training (less than 30 minutes) and real-time rendering speed (50+ FPS).

	pose-dependent deformation	novel pose animation	fast training	real-time rendering	monocular input	
✗	✗	✗	✗	✓		NeuralBody [40]
✓	✗	✗	✗	✓		HumanNeRF [62]
✓	✓	✗	✗	✓		ARAH [61]
✓	✗	✓	✗	✓		Instant-NVR [7]
✗	✓	✓	✗	✓		InstantAvatar [12]
✓	✓	✗	✗	✓		MonoHuman [74]
✓	✓	✗	✗	✗		UV-Volumes [6]
✓	✓	✗	✓	✗		DELIFFAS [18]
✓	✓	✓	✓	✓		<b>3DGS-Avatar (Ours)</b>

Table 1. **Comparison to SoTA.** Instant-NVR [7] and InstantAvatar [12] achieve instant training within 5 minutes. For real-time rendering, we require a frame rate over 30 FPS. Note that while UV-Volumes [6] claims real-time freeview rendering, they only achieve 14 FPS on novel pose synthesis due to the slow generation of their UV Volume.

## 2. Related Works

### 2.1. Neural Rendering for Clothed Human Avatars

Since the seminal work of Neural Radiance Fields (NeRF) [29], there has been a surge of research on neural rendering for clothed human avatars. The majority of the works focus on either learning a NeRF conditioned on human body related encodings [34, 52, 65], or learning a canonical NeRF representation and warp camera rays from the observation space to the canonical space to query radiance and density values from the canonical NeRF [8, 12, 13, 22, 38, 41, 61, 62, 74]. Most of these works rely on large multi-layer perceptrons (MLPs) to model the underlying neural radiance field, which are computationally expensive, resulting in prolonged training (days) and inference (seconds) time.

With recent advances in accelerated data structures for neural fields, there has been several works targeting fast inference and fast training of NeRFs for clothed humans. [12] proposes to use iNGP [32] as the underlying representation for articulated NeRFs, which enables fast training (less than 5 minutes) and interactive rendering speed (15 FPS) but ignores pose-dependent non-rigid deformations. [7] also utilizes iNGP and represents non-rigid deformations in the UV space, which enables fast training and modeling of pose-dependent non-rigid deformations. However, as we will show in our experiments, [7]’s parametrization of non-rigid deformations result in blurry renderings. [6] proposes to generate a pose-dependent UV volume for efficient freeview synthesis. However, their UV-volume generation process is slow (20 FPS), making novel pose synthesis less effi-

cient (only 14 FPS). [18] also employs UV-based rendering to achieve real-time rendering of dynamic clothed humans, but only works on dense multi-view inputs. Extending [73], [58, 76] applied Fourier transform for compressing human performance capture data, albeit with limitations on dense multi-view data (60-80 views) and non-generalizability of the Fourier basis representation to unseen poses beyond the training dataset. In contrast to all these works, our method achieves state-of-the-art rendering quality and speed with less than 30 minutes of training time from a single monocular video input.

## 2.2. Accelerating Neural Radiance Fields

NeRF and its variants [35, 59, 70] can achieve state-of-the-art performance in tasks such as novel-view synthesis and geometry reconstruction. However, the high computational cost of the vanilla NeRF has been a major obstacle to their practical applications. To address this issue, there has been a surge of research on accelerating NeRF.

The study by [46] showcased real-time NeRF rendering by training of numerous smaller MLPs, an approach differing from the large MLP in the original NeRF. The research by [73] focused on distilling learned density values and spherical harmonic coefficients into an octree, resulting in considerable test-time acceleration. Further advancements in fast training were achieved by [50, 56], which optimized voxel grid-stored features w.r.t. to rendering losses. Mueller *et al.* [32] combined a multi-level hash grid with small mixed-precision MLPs [31], dramatically speeding up NeRF model training to around 5 minutes. [3] proposed to decompose high-dimensional tensors (3D for density and 4D for color) into vectors and planes, leading to state-of-the-art results with more compact model sizes.

On the other hand, point-based rendering [43, 48, 53, 66, 75, 77, 78] has also been shown to be an efficient alternative to NeRFs for fast inference and training. 3D Gaussian Splatting (3DGS) [14] models the rendering process as splatting a set of 3D Gaussians onto image plane via alpha blending, achieving state-of-the-art rendering quality with real-time inference speed and fast training given multi-view inputs. Our approach combines 3DGS with human body articulation and non-rigid deformation to create animatable avatar from monocular videos.

## 2.3. Dynamic 3D Gaussians

Given the great performance on both quality and speed of 3DGS, a rich set of works has further explored the 3D Gaussian representation for dynamic scene reconstruction. [14] proposed to optimize the position and shape of each 3D Gaussian on a frame-by-frame basis and simultaneously performed 6-DOF dense tracking for free. Their model size, however, increases with the temporal dimension. [63, 67] maintain a single set of 3D Gaussians in a

canonical space and deform them to each frame via learning a time-dependent deformation field, producing state-of-the-art results in terms of both rendering quality and speed. [68] augments 3D Gaussians with temporal dimension into 4D Gaussian primitives to approximate the underlying spatiotemporal 4D volume of the dynamic scene. While such methods show promising results, they are only applicable to either synthetic datasets with fast camera movement and slow object motion or forward-facing real scenes with limited object movements, thus unable to handle the immense displacement of the articulated human body. To address this problem, our approach utilizes a statistical human body model [26] for articulation and applies regularization to reduce the overfitting of the deformation field.

## 2.4. Concurrent Works

Concurrent with our method, many recent works also seek to combine 3DGS with human articulation prior for avatar reconstruction. We provide a comparison of our approach to concurrent works in Tab. 2. D3GA [79] proposed to embed 3D Gaussians in tetrahedral cages and utilize cage deformations for drivable avatar animation. However, they use dense calibrated multi-view videos as input and require an additional 3D scan to generate the tetrahedral mesh template. Li *et al.* [23] focused on generating avatars with a detailed appearance from multi-view videos by post-processing radiance field renderings with 2D CNNs, which limits their rendering speed. Along with [11, 30], these works fail to achieve fast training with relatively complex pipelines. Similar to our approach, Ye *et al.* [72] deforms 3D Gaussians in canonical space via pose-dependent deformation and rigid articulation, but they still require 2 hours for training and do not show results on monocular inputs. HUGS [17] learns a background model along with the animatable human avatar, but they fail to take pose-dependent cloth deformation into account. Several other works [10, 19, 25] also neglect pose-dependent cloth deformation to achieve even faster training (in 5 minutes) and rendering (150+ FPS). We argue that our method strikes a good balance between quality and speed compared to concurrent works, as being the only method simultaneously achieving the properties listed in Tab. 2.

## 3. Preliminary

In this section, we start by briefly reviewing the linear blend skinning (LBS) function for human articulation in Sec. 3.1. We then explain 3D Gaussians Splatting in Sec. 3.2.

### 3.1. Linear Blend Skinning

To model human articulations, a widely adopted paradigm is to represent geometry and appearance in a shared canonical space [8, 12, 13, 22, 38, 41, 61, 62] and use Linear Blend

pose-dependent deformation	novel pose animation	fast training	real-time rendering	monocular input	
✓	✓	✗	✓	✗	D3GA [79]
✓	✓	✗	✗	✗	Li <i>et al.</i> [23]
✓	✓	✗	✓	✓	SplatArmor [11]
✓	✓	✗	✗	✗	Moreau <i>et al.</i> [30]
✓	✓	✓	✓	✗	Ye <i>et al.</i> [72]
✗	✓	✓	✓	✓	HUGS [17]
✗	✓	✗	✗	✓	GART [19]
✗	✓	✗	✗	✓	Liu <i>et al.</i> [25]
✗	✗	✗	✗	✓	GauHuman [10]
✓	✓	✓	✓	✓	<b>3DGS-Avatar (Ours)</b>

Table 2. Comparison to Concurrent Works.

Skinning (LBS) [2, 9, 26, 36, 37, 64] to deform the parametric human body under arbitrary poses. Given a point  $\mathbf{x}_c$  in canonical space, the LBS function takes a set of rigid bone transformations  $\{\mathbf{B}_b\}_{b=1}^B$  and computes its correspondence  $\mathbf{x}_o$  in the observation space:

$$\mathbf{x}_o = LBS_{\sigma_w}(\mathbf{x}_c; \{\mathbf{B}_b\}) \quad (1)$$

Assuming an underlying SMPL model, we use a total of  $B = 24$  bone transformations, each represented by a  $4 \times 4$  rotation-translation matrix, which are then linearly blended via a set of skinning weights  $\mathbf{w} \in [0, 1]^B$ , s.t.  $\sum_{b=1}^B \mathbf{w}_b = 1$ , modeled by a coordinate-based neural skinning field  $f_{\sigma_w}(\mathbf{x}_c)$  [4, 5, 28, 49, 60]. The forward linear blend skinning function can thus be formulated as:

$$\mathbf{x}_o = LBS(\mathbf{x}_c; \{\mathbf{B}_b\}) = \sum_{b=1}^B f_{\sigma_w}(\mathbf{x}_c)_b \mathbf{B}_b \mathbf{x}_c \quad (2)$$

Compared to prior works that search canonical correspondences of points in observation space [12, 61, 62], our method requires no inverse skinning which is typically difficult to compute and often leads to multiple solutions [4, 5]. A similar technique has been employed in [77] for face avatar modeling.

### 3.2. 3D Gaussian Splatting

3DGS [14] utilizes a set of 3D Gaussian primitives  $\{\mathcal{G}\}$  as static scene representation which can be rendered in real-time via differentiable rasterization. Each 3D Gaussian  $\mathcal{G}$  is defined by its mean  $\mathbf{x}$ , covariance  $\Sigma$ , opacity  $\alpha$  and view-dependent color represented by spherical harmonics coefficients  $\mathbf{f}$ . To ensure positive semi-definiteness, the covariance matrix is represented by a scaling matrix  $\mathbf{S}$  and rotation matrix  $\mathbf{R}$ :

$$\Sigma = \mathbf{R} \mathbf{S} \mathbf{S}^T \mathbf{R}^T \quad (3)$$

In practice, we store the diagonal vector  $\mathbf{s} \in \mathbb{R}^3$  of the scaling matrix and a quaternion vector  $\mathbf{q} \in \mathbb{R}^4$  to represent the rotation matrix, which can be trivially converted to a valid covariance matrix.

The 3D Gaussians are projected to the 2D image plane during the rendering process and accumulated via alpha blending. Given a viewing transformation  $\mathbf{W}$  and the Jacobian of the affine approximation of the projective transformation  $\mathbf{J}$ , the 2D covariance matrix in camera coordinate [80] is given by:

$$\Sigma' = (\mathbf{J} \mathbf{W} \Sigma \mathbf{W}^T \mathbf{J}^T)_{1:2, 1:2} \quad (4)$$

The pixel color  $C$  is thus computed by blending 3D Gaussian splats that overlap at the given pixel, sorted according to their depth:

$$C = \sum_i \left( \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j) \right) c_i \quad (5)$$

where  $\alpha'_i$  denotes the learned opacity  $\alpha_i$  weighted by the probability density of  $i$ -th projected 2D Gaussian at the target pixel location.  $c$  denotes the view-dependent color computed from stored SH coefficients  $\mathbf{f}$ .

The 3D Gaussians  $\{\mathcal{G}\}$  are optimized via a photometric loss. During optimization, 3DGS adaptively controls the number of 3D Gaussians via periodic densification and pruning, achieving self-adaptive convergence to an optimal density distribution of 3D Gaussians that well represents the scene.

## 4. Methods

We illustrate the pipeline of our method in Fig. 2. The input to our method is a monocular video with a calibrated camera, fitted SMPL parameters, and foreground masks. Our method optimizes a set of 3D Gaussians in canonical space, which is then deformed to the observation space and rendered from the given camera. For a set of 3D Gaussians  $\{\mathcal{G}^{(i)}\}_{i=1}^N$ , we store the following properties at each point: position  $\mathbf{x}$ , scaling factor  $\mathbf{s}$ , rotation quaternion  $\mathbf{q}$ , opacity  $\alpha$  and a color feature vector  $\mathbf{f}$ . We start by randomly sampling  $N = 50k$  points on the SMPL [26] mesh surface under canonical pose as initialization of our canonical 3D Gaussians  $\{\mathcal{G}_c\}$ . Inspired by HumanNeRF [62], we decompose the complex human deformation into a non-rigid part that encodes pose-dependent cloth deformation, and a rigid transformation controlled by the human skeleton.

### 4.1. Pose-dependent Non-rigid Deformation

We formulate the non-rigid deformation module as:

$$\{\mathcal{G}_d\} = \mathcal{F}_{\theta_{nr}}(\{\mathcal{G}_c\}; \mathcal{Z}_p) \quad (6)$$

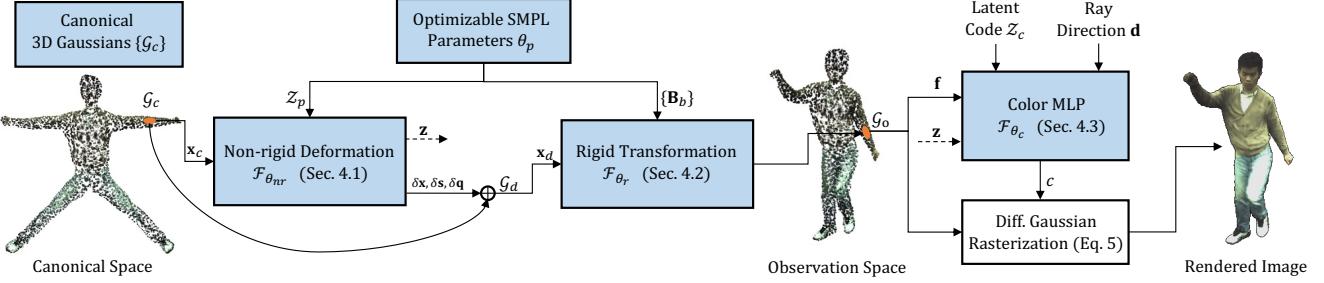


Figure 2. **Our framework for creating animatable avatars from monocular videos.** We first initialize a set of 3D Gaussians in the canonical space via sampling points from a SMPL mesh. Each canonical Gaussian  $\mathcal{G}_c$  goes through a non-rigid deformation module  $\mathcal{F}_{\theta_{nr}}$  (Sec. 4.1) to account for pose-dependent non-rigid cloth deformation. This module outputs a non-rigidly deformed 3D Gaussian  $\mathcal{G}_d$  and a pose-dependent latent feature  $\mathbf{z}$ . The non-rigidly deformed 3D Gaussian  $\mathcal{G}_d$  is transformed to the observation space  $\mathcal{G}_o$  (Sec. 4.2) via LBS with learned neural skinning  $\mathcal{F}_{\theta_r}$ . The Gaussian feature  $\mathbf{f}$ , the pose-dependent feature  $\mathbf{z}$ , a per-frame latent code  $\mathcal{Z}_c$ , and the ray direction  $\mathbf{d}$  are propagated through a small MLP  $\mathcal{F}_{\theta_c}$  to decode the view-dependent color  $c$  for each 3D Gaussian. Finally, the observation space 3D Gaussians  $\{\mathcal{G}_o\}$  and their respective color values are accumulated via differentiable Gaussian rasterization (Eq. (5)) to render the image.

where  $\{\mathcal{G}_d\}$  represents the non-rigidly deformed 3D Gaussians.  $\theta_{nr}$  represents the learnable parameters of the non-rigid deformation module.  $\mathcal{Z}_p$  is a latent code which encodes SMPL pose and shape  $(\theta, \beta)$  using a lightweight hierarchical pose encoder [28]. Specifically, the deformation network  $f_{\theta_{nr}}$  takes the canonical position  $\mathbf{x}_c$ , the pose latent code  $\mathcal{Z}_p$  as inputs and outputs the offsets of the Gaussian’s position, scale, rotation, along with a feature vector  $\mathbf{z}$ :

$$(\delta\mathbf{x}, \delta\mathbf{s}, \delta\mathbf{q}, \mathbf{z}) = f_{\theta_{nr}}(\mathbf{x}_c; \mathcal{Z}_p) \quad (7)$$

We use a multi-level hash grid [32] to encode 3D positions as spatial features, which are then concatenated with the pose latent code  $\mathcal{Z}_p$  and fed into a shallow MLP with 2 hidden layers and a width of 128. The canonical Gaussian is deformed by:

$$\mathbf{x}_d = \mathbf{x}_c + \delta\mathbf{x} \quad (8)$$

$$\mathbf{s}_d = \mathbf{s}_c \cdot \exp(\delta\mathbf{s}) \quad (9)$$

$$\mathbf{q}_d = \mathbf{q}_c \cdot [1, \delta q_1, \delta q_2, \delta q_3] \quad (10)$$

note that the  $\cdot$  operator on quaternions is equivalent to multiplying the two rotation matrices derived by the two quaternions. Since the quaternion  $[1, 0, 0, 0]$  corresponds to the identity rotation matrix, we have  $\mathbf{q}_d = \mathbf{q}_c$  when  $\delta\mathbf{q} = \mathbf{0}$ .

## 4.2. Rigid Transformation

We further transform the non-rigidly deformed 3D Gaussians  $\{\mathcal{G}_d\}$  to the observation space via a rigid transformation module:

$$\{\mathcal{G}_o\} = \mathcal{F}_{\theta_r}(\{\mathcal{G}_d\}; \{\mathbf{B}_b\}_{b=1}^B) \quad (11)$$

where a skinning MLP  $f_{\theta_r}$  is learned to predict skinning weights at the position  $\mathbf{x}_d$ . We transform the position and

the rotation matrix of 3D Gaussians via forward LBS described in Sec. 3.1:

$$\mathbf{T} = \sum_{b=1}^B f_{\theta_r}(\mathbf{x}_d)_b \mathbf{B}_b \quad (12)$$

$$\mathbf{x}_o = \mathbf{T}\mathbf{x}_d \quad (13)$$

$$\mathbf{R}_o = \mathbf{T}_{1:3,1:3} \mathbf{R}_d \quad (14)$$

where  $\mathbf{R}_d$  is the rotation matrix derived from the quaternion  $\mathbf{q}_d$ .

## 4.3. Color MLP

Prior works [63, 67, 68] follow the convention of 3DGs [14], which stores spherical harmonics coefficients per 3D Gaussian to encode the view-dependent color. Treating the stored color feature  $\mathbf{f}$  as spherical harmonics coefficients, the color of a 3D Gaussian can be computed by the dot product of the spherical harmonics basis and the learned coefficients:

$$c = \langle \gamma(\mathbf{d}), \mathbf{f} \rangle \quad (15)$$

where  $\mathbf{d}$  represents the viewing direction, derived from the relative position of the 3D Gaussian wrt. the camera center.  $\gamma$  denotes the spherical harmonics basis function. While conceptually simple, we argue that this approach does not suit our monocular setting. Since only one camera view is provided during training, the viewing direction in the world space is fixed, leading to poor generalization to unseen test views. Similar to [41], we use the inverse rigid transformation from Sec. 4.2 to canonicalize the viewing direction:

$$\hat{\mathbf{d}} = \mathbf{T}_{1:3,1:3}^{-1} \mathbf{d} \quad (16)$$

where  $\mathbf{T}$  is the forward transformation matrix defined in Eq. (12). Theoretically, canonicalizing viewing direction

also promotes consistency of the specular component of canonical 3D Gaussians under rigid transformations.

On the other hand, we observe that the pixel color of the rendered clothed human avatar also largely depends on local deformation. Local fine wrinkles on clothes, for instance, would cause self-occlusion that heavily affects shading. Following [40], we also learn a per-frame latent code  $\mathcal{Z}_c$  to compensate for different environment light effects across frames caused by the global movement of the subject. Hence, instead of learning spherical harmonic coefficients, we enhance color modeling by learning a neural network that takes per-Gaussian color feature vector  $\mathbf{f} \in \mathbb{R}^{32}$ , local pose-dependent feature vector  $\mathbf{z} \in \mathbb{R}^{16}$  from the non-rigid deformation network, per-frame latent code  $\mathcal{Z}_c \in \mathbb{R}^{16}$ , and spherical harmonics basis of canonicalized viewing direction  $\gamma(\hat{\mathbf{d}})$  with a degree of 3 as input and predicts the color of the 3D Gaussian:

$$c = \mathcal{F}_{\theta_c}(\mathbf{f}, \mathbf{z}, \mathcal{Z}_c, \gamma(\hat{\mathbf{d}})) \quad (17)$$

In practice, we find a tiny MLP with 1 hidden layer and a width of 64 sufficient to model the appearance. Increasing the size of the MLP leads to overfitting and performance drop.

#### 4.4. Optimization

We jointly optimize canonical 3D Gaussians  $\{\mathcal{G}_c\}$  and the parameters  $\theta_{nr}, \theta_r, \theta_c$  of the non-rigid deformation network, the skinning network and the color network, respectively.

**Pose correction.** SMPL [26] parameter fittings from images can be inaccurate. To address this, we additionally optimize the per-sequence shape parameter as well as per-frame translation, global rotation, and local joint rotations. We initialize these parameters  $\theta_p$  with the given SMPL parameters and differentiably derive the bone transformations  $\{\mathbf{B}_b\}$  as input to the network, enabling direct optimization via backpropagation.

**As-isometric-as-possible regularization.** With monocular video as input, only one view of the human is visible in each frame, making it extremely hard to generalize to novel views and novel poses. Considering the sparsity of input, the non-rigid deformation network is highly under-constrained, resulting in noisy deformation from the canonical space to the observation space. Inspired by [44], we leverage the as-isometric-as-possible constraint [15] to restrict neighboring 3D Gaussian centers to preserve a similar distance after deformation. We further augment the con-

straint to Gaussian covariance matrices:

$$\mathcal{L}_{isopos} = \sum_{i=1}^N \sum_{j \in \mathcal{N}_k(i)} \left| d(\mathbf{x}_c^{(i)}, \mathbf{x}_c^{(j)}) - d(\mathbf{x}_o^{(i)}, \mathbf{x}_o^{(j)}) \right| \quad (18)$$

$$\mathcal{L}_{isocov} = \sum_{i=1}^N \sum_{j \in \mathcal{N}_k(i)} \left| d(\Sigma_c^{(i)}, \Sigma_c^{(j)}) - d(\Sigma_o^{(i)}, \Sigma_o^{(j)}) \right| \quad (19)$$

where  $N$  denotes the number of 3D Gaussians.  $\mathcal{N}_k$  denotes the  $k$ -nearest neighbourhood, and we set  $k$  to 5. We use L2-norm as our distance function  $d(\cdot, \cdot)$ .

**Loss function.** Our full loss function consists of a RGB loss  $\mathcal{L}_{rgb}$ , a mask loss  $\mathcal{L}_{mask}$ , a skinning weight regularization loss  $\mathcal{L}_{skin}$  and the as-isometric-as-possible regularization loss for both position and covariance  $\mathcal{L}_{isopos}, \mathcal{L}_{isocov}$ . For further details of the loss definition and respective weights, please refer to Appendix A.

## 5. Experiments

In this section, we first compare the proposed approach with recent state-of-the-art methods [7, 12, 40, 61, 62], demonstrating that our proposed approach achieves superior rendering quality in terms of LPIPS, which is more informative under monocular setting, while achieving fast training and real-time rendering speed, respectively **400x and 250x faster** than the most competitive baseline [62]. We then systematically ablate each component of the proposed model, showing their effectiveness in better rendering quality.

### 5.1. Evaluation Dataset

**ZJU-MoCap** [40]. This is the major testbed for quantitative evaluation. We pick six sequences (377, 386, 387, 392, 393, 394) from the ZJU-MoCap dataset and follow the training/test split of HumanNeRF [62]. The motion of these sequences is repetitive and does not contain a sufficient number of poses for meaningful novel pose synthesis benchmarks. Thus we focus on evaluating novel view synthesis (PSNR/SSIM/LPIPS) and show qualitative results for animation on out-of-distribution poses. Note that LPIPS in all the tables are scaled up by 1000.

**PeopleSnapshot** [1]. We also conduct experiments on 4 sequences of the PeopleSnapshot dataset, which includes monocular videos of people rotating in front of a camera. We follow the data split of InstantAvatar [12] and compare to [12] on novel pose synthesis. For a fair comparison, we use the provided poses optimized by Anim-NeRF [38] and do not further optimize it during our training.

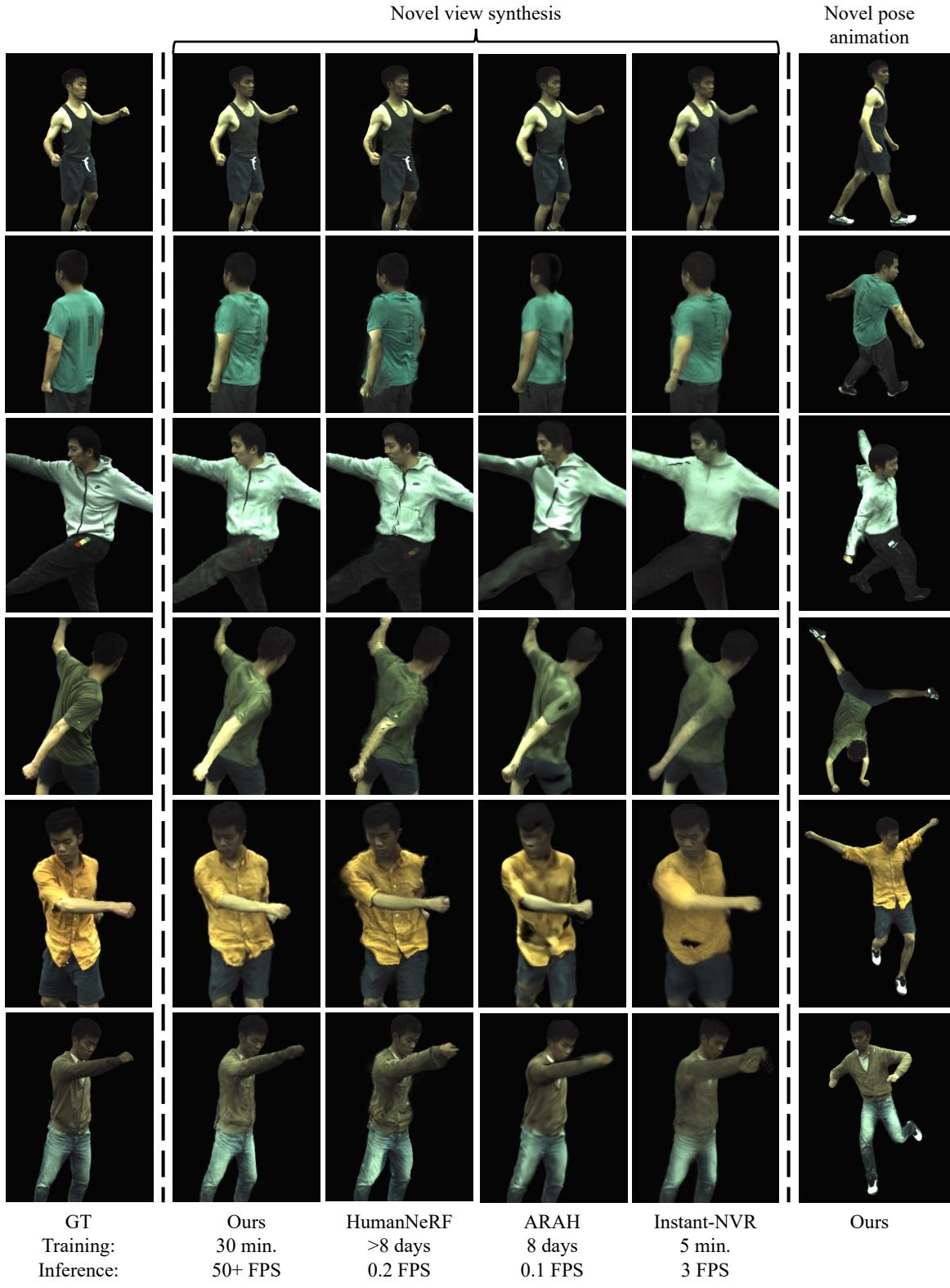


Figure 3. **Qualitative Comparison on ZJU-MoCap [40].** We show the results for both novel view synthesis and novel pose animation of all sequences on ZJU-MoCap. Our method produces high-quality results that preserve cloth details even on out-of-distribution poses.

Table 3. **Quantitative Results on ZJU-MoCap [40].** We outperform both competitive baselines [61, 62] in terms of LPIPS while being two orders of magnitude faster in training and rendering. Cell color indicates **best** and **second best**. Instant-NVR [7] is trained and tested on a refined version of ZJU-MoCap, thus is not directly comparable to other baselines quantitatively. We train our model on the refined dataset for fair quantitative comparison to Instant-NVR and the metrics are reported in the last two rows of the table.

Subject:	377				386				387				392				393				394			
Metric:	GPU↓	FPS↑	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓																
NeuralBody [40]	12h	2	29.11	0.9674	40.95	30.54	0.9678	46.43	27.00	0.9518	59.47	30.10	0.9642	53.27	28.61	0.9590	59.05	29.10	0.9593	54.55				
HumanNeRF [62]	>8d	0.2	30.41	0.9743	24.06	33.20	0.9752	28.99	28.18	0.9632	35.58	31.04	0.9705	32.12	28.31	0.9603	36.72	30.31	0.9642	32.89				
MonoHuman [74]	4d	0.1	29.12	0.9727	26.58	32.94	0.9695	36.04	27.93	0.9601	41.76	29.50	0.9635	39.45	27.64	0.9566	43.17	29.15	0.9595	38.08				
ARAH [61]	8d	0.1	30.85	0.9800	26.60	33.50	0.9781	31.40	28.49	0.9656	40.43	32.02	0.9742	35.28	28.77	0.9645	42.30	29.46	0.9632	40.76				
Ours	0.5h	50	30.64	0.9774	20.88	33.63	0.9773	25.77	28.33	0.9642	34.24	31.66	0.9730	30.14	28.88	0.9635	35.26	30.54	0.9661	31.21				
Instant-NVR* [7]	0.1h	3	31.28	0.9789	25.37	33.71	0.9770	32.81	28.39	0.9640	45.97	31.85	0.9730	39.47	29.56	0.9641	46.16	31.32	0.9680	40.63				
Ours*	0.5h	50	30.96	0.9778	19.85	33.94	0.9784	24.70	28.40	0.9656	32.96	32.10	0.9739	29.20	29.30	0.9645	34.03	30.74	0.9662	31.00				

Table 4. **Quantitative Results on PeopleSnapshot [1].**

Subject:	female-3-casual				female-4-casual				male-3-casual				male-4-casual				
Metric:	GPU↓	FPS↑	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
InstantAvatar [12]	5 min.	15	27.66	0.9709	21.00	29.11	0.9683	16.70	29.53	0.9716	15.50	27.67	0.9626	30.7			
Ours	45 min.	50	30.57	0.9581	20.86	33.16	0.9678	15.74	34.28	0.9724	14.92	30.22	0.9653	23.05			

## 5.2. Comparison with Baselines

We compare our approach with NeuralBody [40], Human-NeRF [62], MonoHuman [74], ARAH [61] and Instant-NVR [7] under a monocular setup on ZJU-MoCap. Since ARAH and MonoHuman are not trained on HumanNeRF’s data split, we re-train new models using their public code and report metrics on these re-trained models. The quantitative results are reported in Tab. 3.

NeuralBody is underperforming compared to other approaches. Overall, our proposed approach produces comparable performance to ARAH on PSNR and SSIM, while significantly outperforming all the baselines on LPIPS. We argue that LPIPS is more informative compared to the other two metrics, as it is very difficult to reproduce exactly the ground-truth appearance for novel views due to the monocular setting and the stochastic nature of cloth deformations. Meanwhile, our method is also capable of fast training and renders at a real-time rendering frame rate, being 400 times faster for training (30 GPU minutes *vs.* 8 GPU days) and 250 – 500 times faster for inference (50 FPS *vs.* 0.1 FPS for ARAH and 0.2 FPS for HumanNeRF). We also note that Instant-NVR trains on a refined version of ZJU-MoCap, which provides refined camera parameters, SMPL fittings, and more accurate instance masks with part-level annotation that is essential for running their method. Hence their metrics are not directly comparable to other methods in Tab. 3. We train our model on the refined dataset for a fair quantitative comparison, which clearly shows that our method outperforms Instant-NVR in most scenarios.

Qualitative comparisons on novel view synthesis can be found in Fig. 3. We observe that our method preserves sharper details compared to ARAH and does not produce fluctuating artifacts as in HumanNeRF caused by noisy de-

Table 5. **Ablation Study on ZJU-MoCap [40].** The proposed model achieves the lowest LPIPS, demonstrating the effectiveness of all components.

Metric:	PSNR↑	SSIM↑	LPIPS↓
Full model	<b>30.61</b>	<b>0.9703</b>	<b>29.58</b>
w/o color MLP	30.55	0.9700	31.24
w/o $\mathcal{L}_{isocov}$	<b>30.61</b>	<b>0.9703</b>	29.84
w/o $\mathcal{L}_{isopos}, \mathcal{L}_{isocov}$	30.59	0.9699	30.25
w/o pose correction	30.60	<b>0.9703</b>	29.87

formation fields. Instant-NVR produces an oversmooth appearance and tends to generate noisy limbs.

Additionally, we animate our learned avatars with pose sequences from AMASS [27] and AIST++ [20], shown in the rightmost column of Fig. 3. This shows that our model could generalize to extreme out-of-distribution poses.

The quantitative comparison of our method with InstantAvatar [12] on PeopleSnapshot is reported in Tab. 4. Our approach significantly outperforms InstantAvatar on PSNR and LPIPS, while being more than 3x faster during inference.

## 5.3. Ablation Study

We study the effect of various components of our method on the ZJU-MoCap dataset, including the color MLP, the as-isometric-as-possible regularization and the pose correction module. The average metrics over 6 sequences are reported in Tab. 5. We show that all proposed techniques are required to reach the optimal performance, best reflected by LPIPS which is the most informative metric for novel view synthesis evaluation under a monocular setup.

We further show qualitative comparison on out-of-

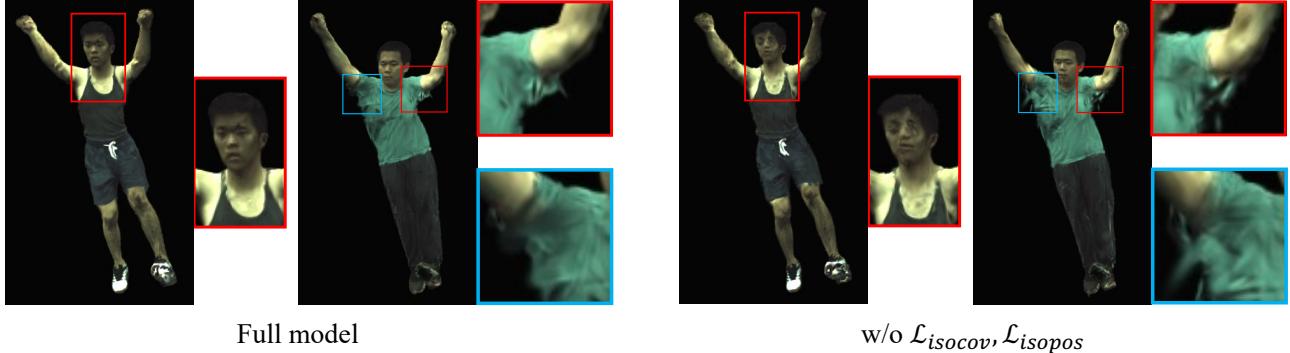


Figure 4. **Ablation Study** on as-isometric-as-possible regularization, which removes the artifacts on highly articulated poses.

distribution poses in Fig. 4, which demonstrates that the as-isometric-as-possible loss helps to constrain the 3D Gaussians to comply with consistent movement during deformation, hence improving generalization on novel poses. Albeit marginally, each individual component contributes to a better novel-view rendering quality and particularly generates more plausible results with respect to novel pose animation.

## 6. Conclusion

In this paper, we presented 3DGS-Avatar, one of the first methods that utilize the explicit representation of 3DGS for efficient reconstruction of clothed human avatars from monocular videos. Our method achieves photorealistic rendering, awareness of pose-dependent cloth deformation, generalization to unseen poses, fast training, and real-time rendering all at once.

Experiments show that our method is comparable to or even better than the state-of-the-art methods in terms of rendering quality while being two orders of magnitude faster in both training and inference. Furthermore, we propose to replace spherical harmonics with a shallow MLP to decode 3D Gaussian color and regularize deformation with geometric constraints, both proved to be effective in enhancing rendering quality. We hope that our new representation could foster further research in fast, high-quality animatable clothed human avatar synthesis from a monocular view.

**Acknowledgement.** SW and AG were supported by the ERC Starting Grant LEGO-3D (850533) and the DFG EXC number 2064/1 - project number 390727645. SW and ST acknowledge the SNF grant 200021 204840.

## References

- [1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proc. of CVPR*, 2018. [6](#), [8](#), [7](#)
- [2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. *ACM Transactions on Graphics*, 24, 2005. [4](#)
- [3] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Proc. of ECCV*, 2022. [2](#), [3](#)
- [4] Xu Chen, Yufeng Zheng, Michael Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proc. of ICCV*, 2021. [4](#)
- [5] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J. Black, and Otmar Hilliges. Fast-snarf: A fast deformer for articulated neural fields. *Pattern Analysis and Machine Intelligence (PAMI)*, 2023. [4](#)
- [6] Yue Chen, Xuan Wang, Xingyu Chen, Qi Zhang, Xiaoyu Li, Yu Guo, Jue Wang, and Fei Wang. Uv volumes for real-time rendering of editable free-view human performance. In *Proc. of CVPR*, 2023. [2](#)
- [7] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *CVPR*, 2023. [2](#), [6](#), [8](#), [5](#)
- [8] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proc. of CVPR*, 2023. [1](#), [2](#), [3](#)
- [9] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A Statistical Model of Human Pose and Body Shape. *Computer Graphics Forum*, 28:337–346, 2009. [4](#)
- [10] Shoukang Hu and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. *arXiv preprint arXiv:*, 2023. [3](#), [4](#)
- [11] Rohit Jena, Ganesh Subramanian Iyer, Siddharth Choudhary, Brandon Smith, Pratik Chaudhari, and James Gee. Splatarmor: Articulated gaussian splatting for animatable humans from monocular rgb videos. *arXiv preprint arXiv:2311.10812*, 2023. [3](#), [4](#)
- [12] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavator: Learning avatars from monocular video in 60 seconds. In *Proc. of CVPR*, 2023. [1](#), [2](#), [3](#), [4](#), [6](#), [8](#), [5](#), [7](#)

- [13] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proc. of ECCV*, 2022. 1, 2, 3
- [14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 1, 2, 3, 4, 5
- [15] Martin Kilian, Niloy J. Mitra, and Helmut Pottmann. Geometric modeling in shape space. *ACM Transactions on Graphics (SIGGRAPH)*, 26(3), 2007. 2, 6
- [16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of ICLR*, 2015. 2
- [17] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats, 2023. 3, 4
- [18] Youngjoong Kwon, Lingjie Liu, Henry Fuchs, Marc Habermann, and Christian Theobalt. Deliffs: Deformable light fields for fast avatar synthesis. *Proc. of NeurIPS*, 2023. 2, 3
- [19] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models, 2023. 3, 4
- [20] Rui long Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proc. of ICCV*, 2021. 8
- [21] Rui long Li, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: A general nerf acceleration toolbox. *arXiv preprint arXiv:2210.04847*, 2022. 2
- [22] Rui long Li, Julian Tanke, Minh Vo, Michael Zollhoefer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *Proc. of ECCV*, 2022. 1, 2, 3
- [23] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling, 2023. 3, 4
- [24] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 1
- [25] Yang Liu, Xiang Huang, Minghan Qin, Qinwei Lin, and Haoqian Wang. Animatable 3d gaussian: Fast and high-quality reconstruction of multiple human avatars, 2023. 3, 4
- [26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transasctions Graphics*, 34(6), 2015. 3, 4, 6
- [27] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proc. of ICCV*, 2019. 8
- [28] Marko Mihajlovic, Yan Zhang, Michael J. Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proc. of CVPR*, 2021. 4, 5
- [29] 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 *Proc. of ECCV*, 2020. 1, 2
- [30] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars, 2023. 3, 4
- [31] Thomas Müller, Fabrice Rousselle, Jan Novák, and Alexander Keller. Real-time neural radiance caching for path tracing. *ACM Trans. Graph.*, 40(4), 2021. 3
- [32] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transasctions Graphics*, 41(4), 2022. 2, 3, 5
- [33] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. of CVPR*, 2020. 1
- [34] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proc. of ICCV*, 2021. 1, 2
- [35] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proc. of ICCV*, 2021. 1, 3
- [36] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. Star: Sparse trained articulated human body regressor. In *Proc. of ECCV*, 2020. 4
- [37] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proc. of CVPR*, 2018. 4
- [38] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proc. of ICCV*, 2021. 1, 2, 3, 6
- [39] Songyou Peng, Chiyu "Max" Jiang, Yiyi Liao, Michael Niemeyer, Marc Pollefeys, and Andreas Geiger. Shape as points: A differentiable poisson solver. In *Proc. of NeurIPS*, 2021. 1
- [40] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proc. of CVPR*, 2021. 1, 2, 6, 7, 8, 3
- [41] Sida Peng, Shangzhan Zhang, Zhen Xu, Chen Geng, Boyi Jiang, Hujun Bao, and Xiaowei Zhou. Animatable neural implicit surfaces for creating avatars from videos. *ArXiv*, abs/2203.08133, 2022. 1, 2, 3, 5
- [42] Sida Peng, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Representing volumetric videos as dynamic mlp maps. In *Proc. of CVPR*, 2023. 2
- [43] Sergey Prokudin, Michael J Black, and Javier Romero. Smplpix: Neural avatars from 3d human models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1810–1819, 2021. 2, 3
- [44] Sergey Prokudin, Qianli Ma, Maxime Raafat, Julien Valentin, and Siyu Tang. Dynamic point fields. In *Proc. of ICCV*, 2023. 2, 6
- [45] Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. Anr-articulated neural rendering for virtual avatars. In *Proc. of CVPR*, 2021. 1

- [46] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proc. of ICCV*, 2021. 3
- [47] Christian Reiser, Richard Szeliski, Dor Verbin, Pratul P. Srinivasan, Ben Mildenhall, Andreas Geiger, Jonathan T. Barron, and Peter Hedman. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *ACM TOG*, 42(4), 2023. 2
- [48] Darius Rückert, Linus Franke, and Marc Stamminger. Adop: Approximate differentiable one-pixel point rendering. *ACM Transactions on Graphics*, 41(4), 2022. 2, 3
- [49] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proc. of CVPR*, 2021. 4
- [50] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proc. of CVPR*, 2022. 2, 3
- [51] Vincent Sitzmann, Semon Rezhikov, William T. Freeman, Joshua B. Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Proc. of NeurIPS*, 2021. 1
- [52] Shih-Yang Su, Frank Yu, Michael Zollhoefer, and Helge Rhodin. A-neRF: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Proc. of NeurIPS*, 2021. 1, 2
- [53] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. Npc: Neural point characters from video. In *Proc. of ICCV*, 2023. 2, 3
- [54] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *Proc. of ECCV*, 2022. 1
- [55] Mohammed Suhail1, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Light field neural rendering. In *Proc. of CVPR*, 2022. 1
- [56] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proc. of CVPR*, 2022. 2, 3
- [57] Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In *Proc. of ECCV*, 2022. 1
- [58] Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yan-shun Zhang, Yingliang Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Fourier plenoctrees for dynamic radiance field rendering in real-time. In *Proc. of CVPR*, 2022. 3
- [59] 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. In *Proc. of NeurIPS*, 2021. 1, 3
- [60] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. In *Proc. of NeurIPS*, 2021. 4
- [61] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *Proc. of ECCV*, 2022. 1, 2, 3, 4, 6, 8, 5, 7
- [62] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humanerf: Free-viewpoint rendering of moving people from monocular video. In *Proc. of CVPR*, 2022. 1, 2, 3, 4, 6, 8, 7
- [63] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Wang Xinggang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023. 3, 5
- [64] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proc. of CVPR*, 2020. 4
- [65] Hongyi Xu, Thimo Alldieck, and Cristian Sminchisescu. H-neRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *Proc. of NeurIPS*, 2021. 1, 2
- [66] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proc. of CVPR*, 2022. 2, 3
- [67] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023. 3, 5, 2
- [68] Zeyu Yang, Hongye Yang, Zijie Pan, Xiatian Zhu, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv 2310.10642*, 2023. 3, 5
- [69] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Proc. of NeurIPS*, 2020. 1
- [70] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Proc. of NeurIPS*, 2021. 1, 3
- [71] Lior Yariv, Peter Hedman, Christian Reiser, Dor Verbin, Pratul P. Srinivasan, Richard Szeliski, Jonathan T. Barron, and Ben Mildenhall. Bakedsdf: Meshing neural sdfs for real-time view synthesis. In *Proc. of SIGGRAPH*, 2023. 2
- [72] Keyang Ye, Tianjia Shao, and Kun Zhou. Animatable 3d gaussians for high-fidelity synthesis of human motions, 2023. 3, 4
- [73] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for real-time rendering of neural radiance fields. In *Proc. of ICCV*, 2021. 2, 3
- [74] Zhengming Yu, Wei Cheng, xian Liu, Wayne Wu, and Kwan-Yee Lin. MonoHuman: Animatable human neural field from monocular video. In *Proc. of CVPR*, 2023. 1, 2, 8, 6, 7
- [75] Qiang Zhang, Seung-Hwan Baek, Szymon Rusinkiewicz, and Felix Heide. Differentiable point-based radiance fields for efficient view synthesis. In *SIGGRAPH Asia Conference Proceedings*, 2022. 2, 3
- [76] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, and Jingyi Yu. Human performance modeling and rendering via neural animated mesh. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 41(6), 2022. 3

- [77] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proc. of ECCV*, 2023. [2](#), [3](#), [4](#)
- [78] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yan-dong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proc. of CVPR*, 2022. [2](#), [3](#)
- [79] Wojciech Zelenka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. *arXiv preprint arXiv:2311.08581*, 2023. [3](#), [4](#)
- [80] M. Zwicker, H. Pfister, J. van Baar, and M. Gross. Ewa volume splatting. In *Proceedings Visualization, 2001. VIS '01.*, pages 29–538, 2001. [4](#)

# 3DGS-Avatar: Animatable Avatars via Deformable 3D Gaussian Splatting

## Supplementary Material

### A. Loss Definition

In Sec. 4.4 of the main paper we describe our loss term which can be formulated as follows:

$$\mathcal{L} = \lambda_{l1}\mathcal{L}_{l1} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{skin}\mathcal{L}_{skin} + \lambda_{isopos}\mathcal{L}_{isopos} + \lambda_{isocov}\mathcal{L}_{isocov} \quad (20)$$

We describe how each loss term is defined below:

**RGB Loss:** We use an  $l1$  loss to compute pixel-wise error and a perceptual loss to provide robustness to local misalignments, which is critical for the monocular setup. Following [62], we optimize LPIPS as the perceptual loss with VGG as the backbone. However, unlike NeRF-based methods which train on random ray samples, we render the whole image via rasterization and thus do not require patch sampling. For computational efficiency, we crop the tight enclosing bounding box with the ground truth mask and compute the VGG-based LPIPS as our perceptual loss.

**Mask Loss:** To boost the convergence of 3D Gaussian positions, we use an explicit mask loss. For each pixel  $p$ , we compute the opacity value  $O_p$  by summing up the sample weights in the rendering equation Eq. (5) in the main paper , namely:

$$O_p = \sum_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j) \quad (21)$$

We thus supervise it with the ground truth foreground mask via an  $l1$  loss. Experiments show that the  $l1$  loss provides faster convergence than the Binary Cross Entropy (BCE) loss.

**Skinning Loss:** We leverage SMPL prior by sampling 1024 points  $\mathbf{X}_{skin}$  on the surface of the canonical SMPL mesh and regularizing the forward skinning network with corresponding skinning weights  $\mathbf{w}$  interpolated with barycentric coordinates.

$$\mathcal{L}_{skin} = \frac{1}{|\mathbf{X}_{skin}|} \sum_{\mathbf{x}_{skin} \in \mathbf{X}_{skin}} \|f_{\theta_r}(\mathbf{x}_{skin}) - \mathbf{w}\|^2 \quad (22)$$

**As-isometric-as-possible Loss:** Please refer to the second paragraph of Sec. 4.4 in the main paper for details.

We set  $\lambda_{l1} = 1, \lambda_{perc} = 0.01, \lambda_{mask} = 0.1, \lambda_{isopos} = 1, \lambda_{isocov} = 100$  in all experiments. For  $\lambda_{skin}$ , we set it to 10 for the first 1k iterations for fast convergence to a reasonable skinning field, then decreased to 0.1 for soft regularization.

### B. Implementation Details

We initialize the canonical 3D Gaussians with  $N = 50k$  random samples on the SMPL mesh surface in canonical pose. During optimization, we follow the same strategy from [14] to densify and prune the 3D Gaussians, using the view-space position gradients derived from the transformed Gaussians  $\mathcal{G}_o$  in the observation space as the criterion for densification.

We then describe the network architectures of our learned neural components. For the forward skinning network  $f_{\theta_r}$ , we use an MLP with 4 hidden layers of 128 dimensions which takes  $\mathbf{x}_c \in \mathbb{R}^3$  with no positional encoding and outputs a 25-dimension vector. This vector is further propagated through a hierarchical softmax layer that is aware of the tree structure of the human skeleton to obtain the skinning weights  $\mathbf{w}$  that sum up to 1. To normalize the coordinates in the canonical space, we proportionally pad the bounding box enclosing the canonical SMPL mesh instead of using the same length in all axes as in [61]. This allows us to use a lower resolution in the flat  $z$ -dimension of the human body.

For the non-rigid deformation network  $f_{\theta_{nr}}$ , the 3D position  $\mathbf{x}_d$  is normalized with the aforementioned bounding box and first encoded into representative features with a multi-level hash grid, whose parameters are defined in Tab. 6. The concatenation of the hash grid features and the pose latent code  $\mathcal{Z}_p$  then go through a shallow MLP with 3 hidden layers of 128 dimensions to decode pose-dependent local deformation.

The details of our color network structure  $\mathcal{F}_{\theta_c}$  are well elaborated in Sec. 4.3 of the main paper. For frames outside the training set, we follow [61] and use the latent code of the last frame in the training sequence.

Parameter	Value
Number of levels	16
Feature dimension per level	2
Hash table size	$2^{16}$
Coarsest resolution	16
Finest resolution	2048

Table 6. Hash table parameters.

To reduce overfitting, we add noise to the pose and viewing direction input. Specifically, we add a noise drawn from the normal distribution  $\mathcal{N}(0, 0.1)$  to the SMPL pose parameters  $\theta$  with a probability of  $p = 0.5$  during training. The viewing direction  $d$  is first canonicalized to the canonical space and then augmented with a random rotation derived from uniformly sampled roll, pitch, and yaw degrees  $\in [0, 45]$ . Adding noise to training signals helps the model to better generalize to novel poses and views.

Our model is trained for a total of  $15k$  iterations on the ZJU-MoCap dataset in 30 minutes and  $30k$  iterations on PeopleSnapshot in 45 minutes on a single NVIDIA RTX 3090 GPU. We use Adam [16] to optimize our model and the per-frame latent codes with hyperparameters  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . The learning rate of 3D Gaussians is exactly the same as the original implementation from [14]. We set the learning rate for forward skinning network  $\theta_r$  to  $1 \times 10^{-4}$  and  $1 \times 10^{-3}$  for all the others. An exponential learning rate scheduler is employed to gradually decrease the learning rate by a factor of 0.1 on neural networks. We also apply a weight decay with a weight of 0.05 to the per-frame latent codes.

Following prior works [62, 67], we split the training stage and learn the whole model in a coarse-to-fine manner. In the first  $1k$  iterations, we freeze everything except the forward skinning network  $f_{\theta_r}$  to learn a coarse skinning field with  $\mathcal{L}_{skin}$  and prevent the noisy gradients from moving the 3D Gaussians away from the initialization. We then enable optimization on the 3D Gaussians after  $1k$  steps. To decouple rigid and non-rigid motion, we start to optimize the non-rigid deformation network  $f_{\theta_{nr}}$  after  $3k$  iterations. Lastly, we turn on pose correction after  $5k$  iterations.

## C. Implementation Details for Baselines

In this section, we elaborate on the implementation details of baselines used for comparison to our proposed method, *i.e.* NeuralBody [40], HumanNeRF [62], ARAH [61], Instant-NVR [7], MonoHuman [74] and InstantAvatar [12].

### C.1. NeuralBody

For the quantitative evaluation, we use the results of NeuralBody [40] reported in HumanNeRF [62] which follows the same data split.

### C.2. HumanNeRF

We use pre-trained models provided by the official code repository<sup>1</sup> for both quantitative and qualitative evaluation.

### C.3. ARAH

For the quantitative evaluation, we use the same setup as HumanNeRF (*i.e.* same data split with a reduced image size of  $512 \times 512$ ) and train the models using the code from official code repository<sup>2</sup> for 500 epochs. All other hyperparameters remain unchanged. The trained models are then used for qualitative evaluation and out-of-distribution pose animation.

### C.4. Instant-NVR

For quantitative and qualitative evaluation, we retrain the models using the code from official code repository<sup>3</sup> on the refined ZJU-MoCap dataset provided by the author. We change the data split to match other baselines while keeping all other hyperparameters the same.

<sup>1</sup><https://github.com/chungiyweng/humannerf>

<sup>2</sup><https://github.com/taconite/arah-release>

<sup>3</sup><https://github.com/zju3dv/instant-nvr>

Table 7. **Additional Ablation Study on ZJU-MoCap [40].** We present the average metrics over 6 sequences.

Metric:	PSNR↑	SSIM↑	LPIPS↓
Full model	<b>30.61</b>	<b>0.9703</b>	<b>29.58</b>
w/o $\mathcal{L}_{mask}$	30.58	<b>0.9703</b>	29.90
Random initialization	<b>30.61</b>	0.9701	30.90
7k iterations	30.56	0.9698	31.73

### C.5. MonoHuman

We note that MonoHuman uses a different data split from HumanNeRF with the last fifth of the training frames being used for novel pose synthesis evaluation instead. For fair comparison we retrain the model from official code repository<sup>4</sup> on the same data split of HumanNeRF with the provided configs for 400k iterations and recompute the metrics on novel view synthesis. The trained models are then used for qualitative evaluation and out-of-distribution pose animation.

### C.6. InstantAvatar

We follow the original setup and use the provided poses optimized by Anim-NeRF [38] without further pose correction. For quantitative results we copy the metrics from their table, while for qualitative results we train the model from official code repository<sup>5</sup> as they do not release pretrained checkpoints.

## D. Ablation Study

We conduct additional ablation study and report the average metrics on ZJU-MoCap in Tab. 7.

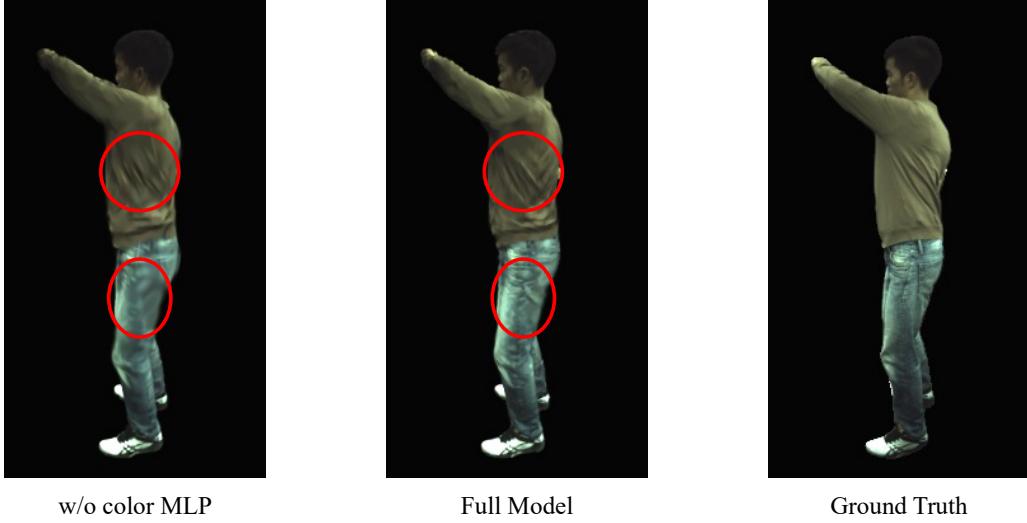


Figure 5. **Qualitative Ablation of Color MLP.**

### D.1. Ablation on Color MLP

We show in Tab. 5 of the main paper that our proposed color MLP produces rendering with higher quality compared to learning spherical harmonics coefficients. We hereby show qualitative comparison to corroborate this enhancement in Fig. 5. Our proposed color MLP helps generate more realistic cloth wrinkles and sharper textures with pose-dependent feature  $z$  and per-frame latent code  $\mathcal{Z}_c$  as additional inputs.

<sup>4</sup><https://github.com/Yzmblog/MonoHuman>

<sup>5</sup><https://github.com/tijiang13/InstantAvatar>



Figure 6. **Qualitative Ablation of Mask Loss.**

## D.2. Ablation on Mask Supervision

Explicit supervision from ground-truth foreground masks only seems to gain slight improvement, as shown in Tab. 7. However, we observe that the mask loss is useful for removing floating blobs in the empty space. Fig. 6 shows an example for this, without mask loss, the floating Gaussian with the background color could occlude the subject in novel views.

## D.3. Ablation on Gaussian Initialization

Instead of initializing the canonical 3D Gaussians from a SMPL mesh surface, we tried to perform random initialization. Specifically, we randomly sample  $N = 50k$  points in the enclosing bounding box around the canonical SMPL mesh. Experimental results from Tab. 7 demonstrate that our method could as well converge starting from random initialization, with little performance drop compared to the SMPL initialization scheme. Despite this interesting observation, we decide to use SMPL initialization as it is more intuitive and does not incur any overhead.

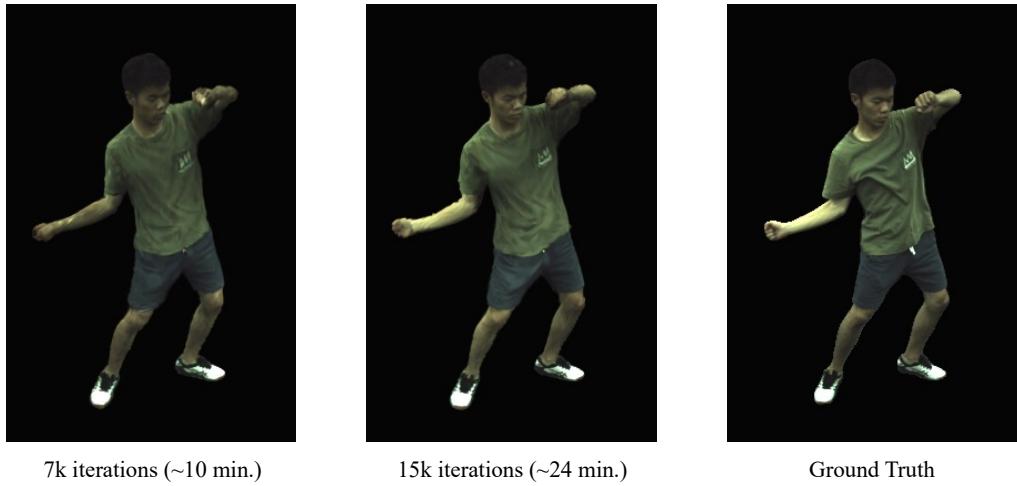


Figure 7. **Qualitative Ablation of Training Iterations.**

#### D.4. Ablation on Training Iterations

Training for  $15k$  iterations on ZJU-MoCap takes precisely around 24 minutes. We further show that our method can already achieve high-quality results at  $7k$  iterations in Tab. 7, which takes around 10 minutes, not far away from [12] and [7] that claim instant training within 5 minutes. Qualitative comparison is shown in Fig. 7.

### E. Additional Qualitative Results

We show more qualitative results in this section. **For better visualization, we strongly recommend to check our supplementary video.**

#### E.1. Qualitative Results of Novel View Synthesis on ZJU-MoCap

Additional qualitative comparison of novel view synthesis on ZJU-MoCap is shown in Fig. 8. HumanNeRF and MonoHuman preserves sharp details, but often produces undesired distortions and cloud-like effect around the contour. ARAH gives more rigid body thanks to their explicit modeling of geometry, while they show misalignment and lack fine details. Instant-NVR synthesizes blurry appearance and obvious artifacts on the limbs. Overall, our method can generate high-quality images with realistic cloth deformations.

#### E.2. Qualitative Results of Out-of-distribution Poses on ZJU-MoCap

We present qualitative comparison of extreme out-of-distribution pose animation in Fig. 9. Our method does not produce obvious artifacts compared to baselines, demonstrating good generalization to unseen poses.

#### E.3. Qualitative Results on PeopleSnapshot

We show qualitative results on the test set of PeopleSnapshot in Fig. 10. Compared to InstantAvatar, our method produces sharper results, especially in the face region.

### F. Limitations

While our proposed approach achieves state-of-the-art rendering quality of clothed human avatars with an interactive frame rate of rendering, the training time of our model still does not match those fast grid-based methods [7, 12]. On the other hand, our method may produce blurry results in areas with high-frequency texture or repetitive patterns, such as striped shirts. Lastly, our method does not provide accurate geometry reconstruction of the avatar, unlike ARAH [61]. Despite reasonable rendering quality, our method generates noisy surface normal resulting from the inconsistency of Gaussian splat depth. It would be particularly interesting to study how to extract a smooth, detailed geometry from the 3DGS avatar model, possibly by applying regularization to the normal map or attaching 3D Gaussians to an underlying mesh.

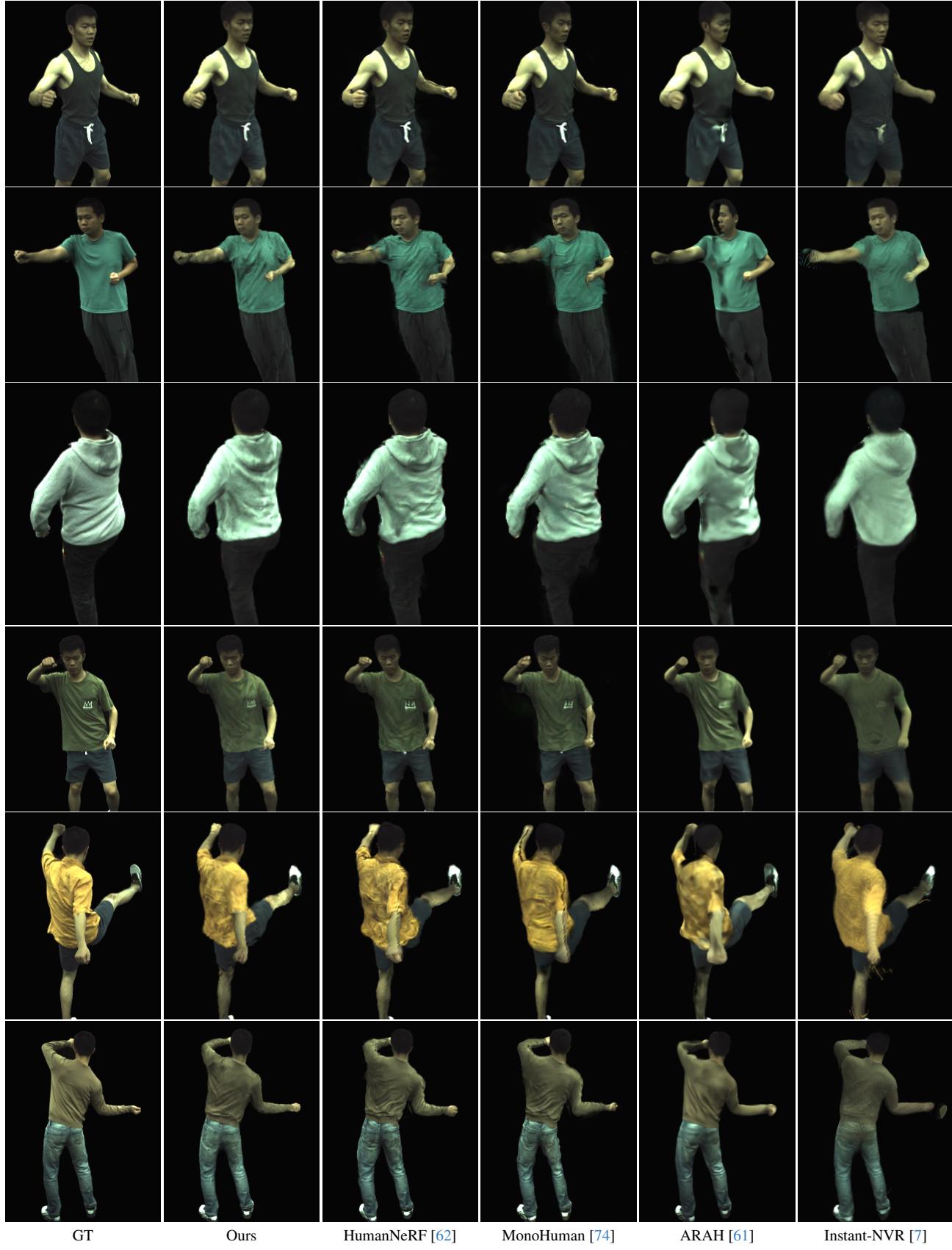


Figure 8. Qualitative Comparison of Novel View Synthesis on ZJU-MoCap.

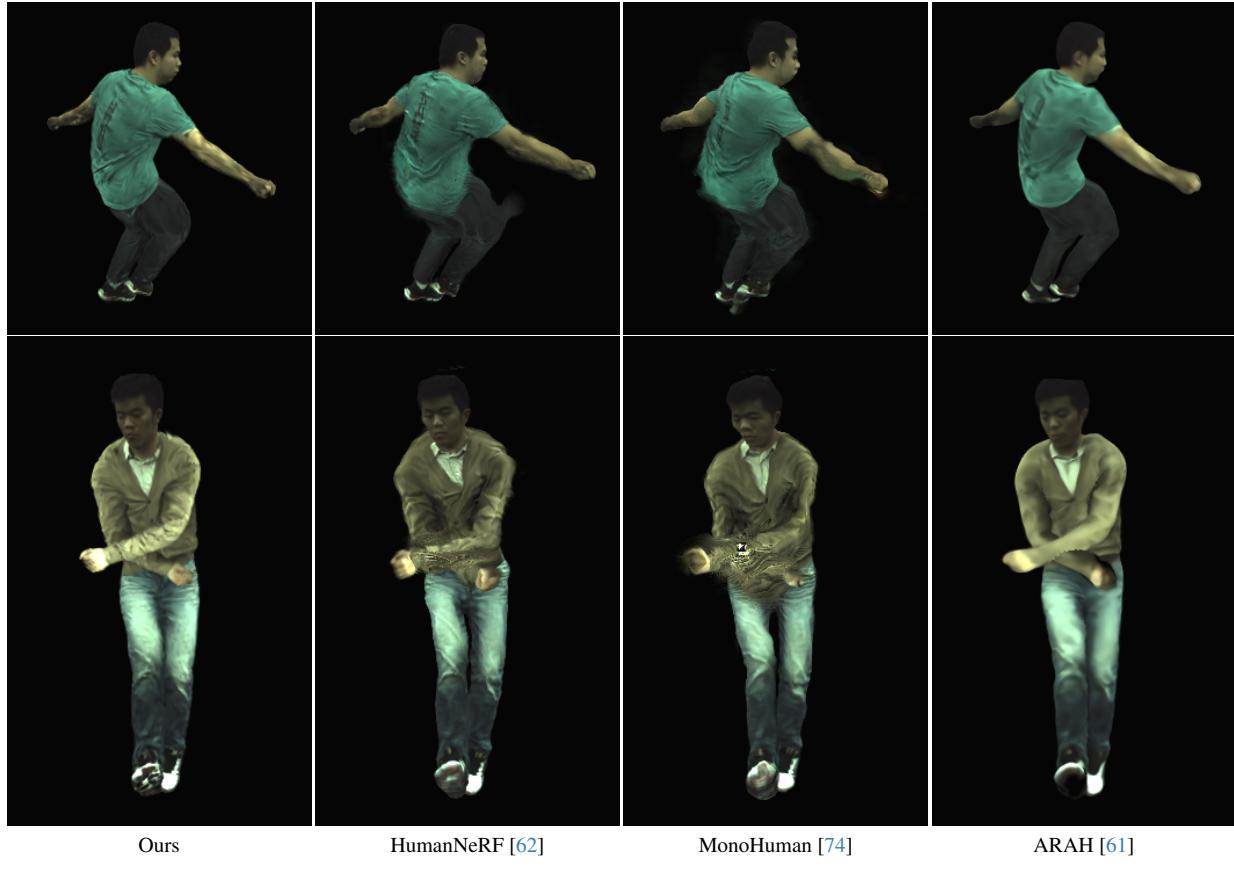


Figure 9. Qualitative Comparison of Out-of-distribution Pose Animation on ZJU-MoCap.

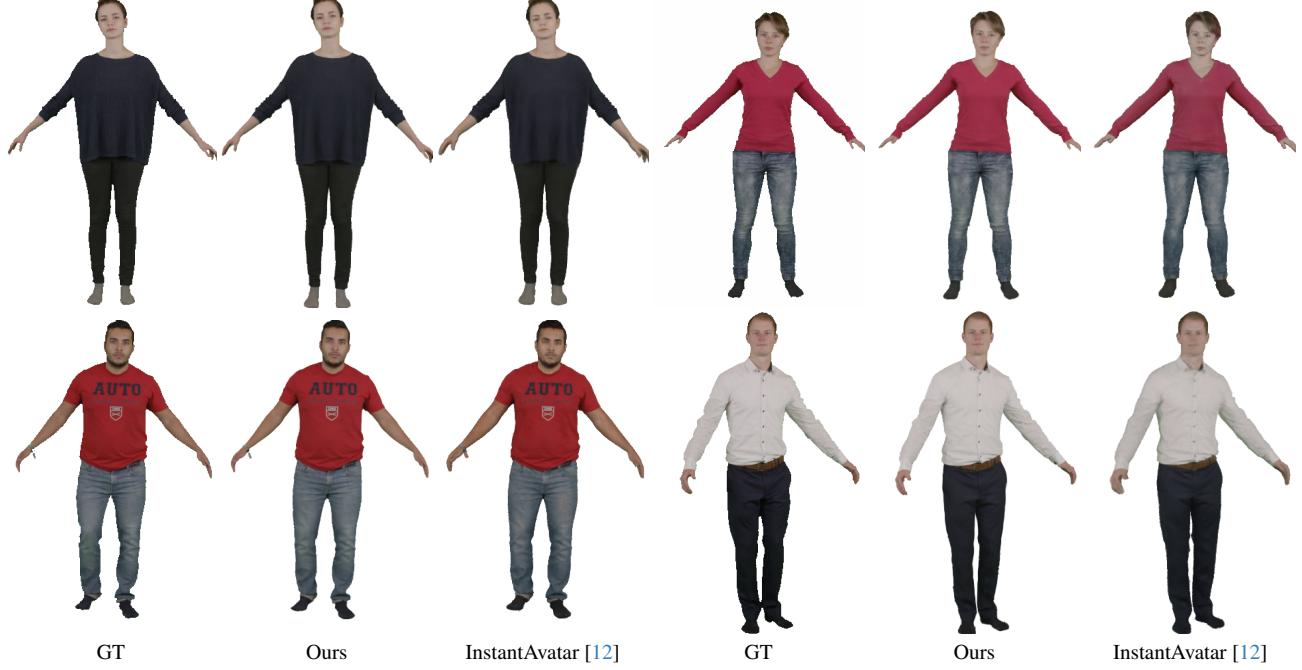


Figure 10. Qualitative Comparison on PeopleSnapshot [1]. Best viewed zoomed-in.