GaussianBody: Clothed Human Reconstruction via 3d Gaussian Splatting

Mengtian Li[1,3], Shengxiang Yao[1], Zhifeng Xie[1,3,†] Keyu Chen[2†]
[1]Shanghai University, [2]Tavus Inc.
[3]Shanghai Engineering Research Center of Motion Picture Special Effects

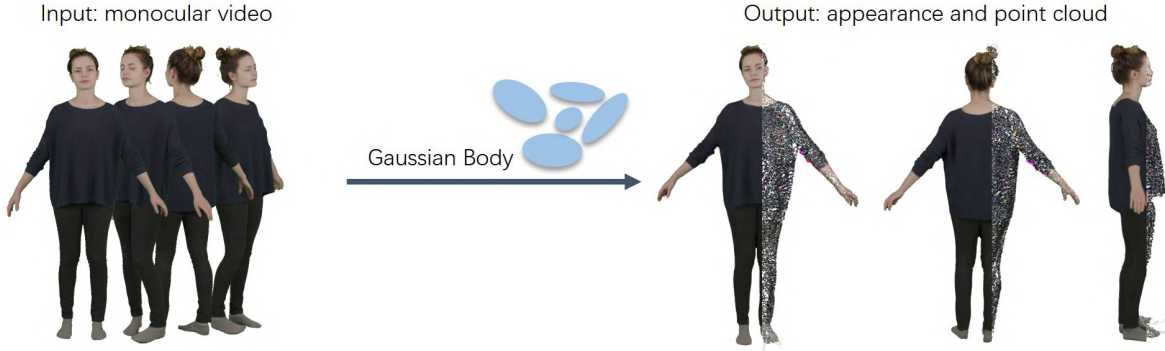{mtli,yaosx033,zhifeng_xie}@shu.edu.cn, keyu@tavus.dev

Figure 1. GaussianBody takes monocular RGB video as input, reconstructing a clothed human model from 1080×1080 images in around 1 hour on a single 4090 GPU. The resulting human model serves as a tool for simulating human performance in novel views. Furthermore, we offer the point cloud as a mechanism for deformation control.

Abstract

In this work, we propose a novel clothed human reconstruction method called GaussianBody, based on 3D Gaussian Splatting. Compared with the costly neural radiance-based models, 3D Gaussian Splatting has recently demonstrated great performance in terms of training time and rendering quality. However, applying the static 3D Gaussian Splatting model to the dynamic human reconstruction problem is non-trivial due to complicated non-rigid deformations and rich cloth details. To address these challenges, our method considers explicit pose-guided deformation to associate dynamic Gaussians across the canonical space and the observation space, introducing a physically-based prior with regularized transformations helps mitigate ambiguity between the two spaces. During the training process, we further propose a pose refinement strategy to update the pose regression for compensating the inaccurate initial estimation and a split-with-scale mechanism to enhance the density of regressed point clouds. The experiments validate that our method can achieve state-of-the-art photorealistic novel-view rendering results with high-quality details for dynamic clothed human bodies, along with explicit geom-etry reconstruction.

---

†Corresponding author.

1. Introduction

Creating high-fidelity clothed human models holds significant applications in virtual reality, telepresence, and movie production. Traditional methods involve either complex capture systems or tedious manual work from 3D artists, making them time-consuming and expensive, thus limiting scalability for novice users. Recently, there has been a growing focus on automatically reconstructing clothed human models from single RGB images or monocular videos.

Mesh-based methods [1–4] are initially introduced to recover human body shapes by regressing on parametric models such as *SCAPE* [5], *SMPL* [6], *SMPL-X* [7], and *STAR* [8]. While they can achieve fast and robust reconstruction, the regressed polygon meshes struggle to capture variant geometric details and rich clothing features. The addition of vertex offsets becomes an enhancement solution [9, 10] in this context. However, its representation ability is still strictly constrained by mesh resolutions and generally fails in loose-cloth cases.

To overcome the limitations of explicit mesh mod-

els, implicit methods based on occupancy fields [11, 12], signed distance fields (SDF) [13], and neural radiance fields (NeRFs) [14–21] have been developed to learn the clothed human body using volume rendering techniques. These methods are capable of enhancing the reconstruction fidelity and rendering quality of 3D clothed humans, advancing the realistic modeling of geometry and appearance. Despite performance improvements, implicit models still face challenges due to the complex volume rendering process, resulting in long training times and hindering interactive rendering for real-time applications. Most importantly, native implicit approaches lack an efficient deformation scheme to handle complicated body movements in dynamic sequences [18–21].

Therefore, combining explicit geometry primitives with implicit models has become a trending idea in recent works. For instance, point-based NeRFs [22, 23] propose controlling volume-based representations with point cloud proxy. Unfortunately, estimating an accurate point cloud from multi-view images is practically challenging as well due to the intrinsic difficulties of the multi-view stereo (MVS) problem.

In this work, we address the mentioned issues by incorporating 3D Gaussian Splatting (3D-GS) [24] into the dynamic clothed human reconstruction framework. 3D-GS establishes a differential rendering pipeline to facilitate scene modeling, notably reducing a significant amount of training time. It learns the explicit point-based model while rendering high-quality results with spherical harmonics (SH) representation. The application of 3D-GS to present 4D scenes has demonstrated superior results [25–27], motivating our endeavor to integrate 3D-GS into human body reconstruction. However, learning dynamic clothed body reconstruction is more challenging than other use cases, primarily due to non-rigid deformations of body parts and the need to capture accurate details of the human body and clothing, especially for loose outfits like skirts.

Firstly, we extended the 3D-GS representation to clothed human reconstruction by utilizing an articulated human model for guidance. Specifically, we use forward linear blend skinning (LBS) to deform the Gaussians from the canonical space to each observation space per frame. Secondly, we optimize a physically-based prior for the Gaussians in the observation space to mitigate the risk of overfitting Gaussian parameters. We transform the local rigidity loss [28] to regularize over-rotation across the canonical and observation space. Finally, we propose a split-with-scale strategy to enhance point cloud density and a pose refinement approach to address the texture blurring issue.

We evaluate our proposed framework on monocular videos of dynamic clothed humans. By comparing it with baseline approaches and other works, our method achieves superior reconstruction quality in rendering details and ge-

ometry recovery, while requiring much less training time (approximately one hour) and almost real-time rendering speed. We also conduct ablation studies to validate the effectiveness of each component in our method.

2. Related Work

In this section, we briefly review the related literature with 3D clothed human reconstruction.

2.1. 3D Human Reconstruction

Reconstructing 3D humans from images or videos is a challenging task. Recent works [9, 10, 29] use template mesh models like SMPL [6] to reconstruct 3D humans from monocular videos or single images. However, template mesh models have limitations in capturing intricate clothing details. To address these limitations, neural representations have been introduced [11, 12, 30] for 3D human reconstruction. Implicit representations, like those used in PIFU [11] and its variants, achieve impressive results in handling complex details such as hairstyle and clothing. Some methods, like ICON [13] and ECON [31], leverage SMPL as a prior to handle extreme poses. However, most of these methods are designed for static scenes and struggle with dynamic scenarios. Other methods [32–34] use parametric models to handle dynamic scenes and obtain animatable 3D human models.

Recent advancements involve using neural networks for representing dynamic human models. Extensions of NeRF [14] into dynamic scenes [35–37] and methods for animatable 3D human models in multi-view scenarios [18–21, 38, 39] or monocular videos [15–17, 40] have shown promising results. Signal Distance Function (SDF) is also employed [41–43] to establish a differentiable rendering framework or use NeRF-based volume rendering to estimate the surface. Our method enhances both speed and robustness by incorporating 3D-GS [24].

2.2. Accelerating Neural Rendering

Several methods [44–47] focus on accelerating rendering speed, primarily using explicit representations or baking methods. However, these approaches are tailored for static scenes. Some works [48, 49] aim to accelerate rendering in dynamic scenes, but they often require dense input images or additional geometry priors. InstantAvatar [17], based on instant-NGP [50], combines grid skip rendering and a quick deformation method [51] but relies on accurate pose guidance for articulate weighting training. In contrast, 3D-GS [24] offers fast convergence and easy integration into graphics rendering pipelines, providing a point cloud for explicit deformation. Our method extends 3D-GS for human reconstruction, achieving high-quality results and fast rendering.
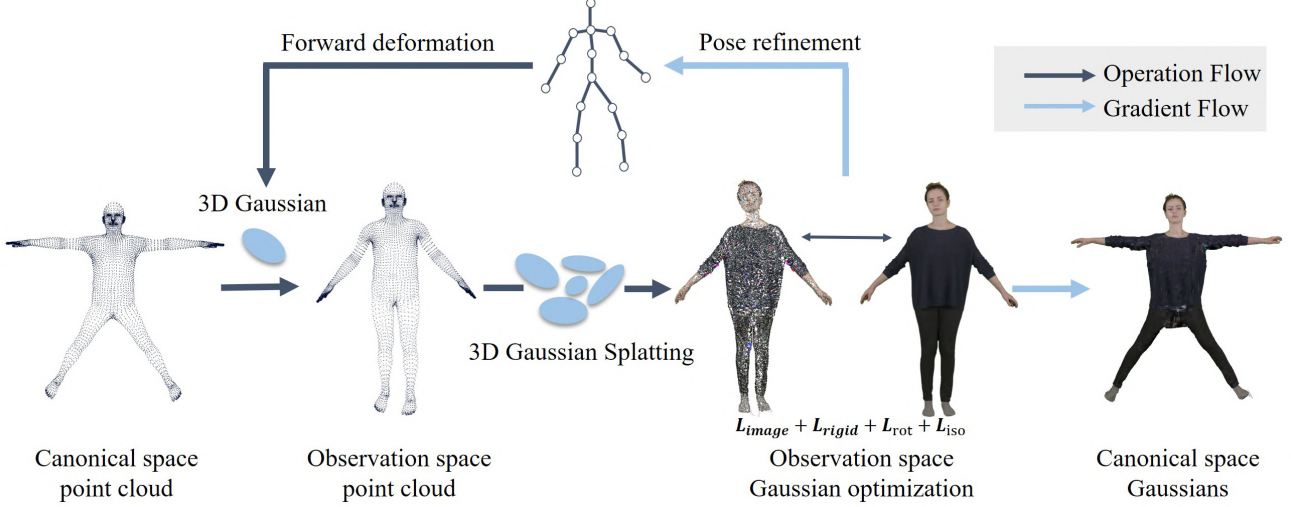
Figure 2. **Overview of our pipeline.** We initialize the point cloud using SMPL vertices, deforming the position and rotation parameters of Gaussians through SMPL forward linear blend skinning (LBS) to transform them into the observation space. The canonical model is then optimized, taking into account the physically-based prior $\mathcal{L}_{rigid}, \mathcal{L}_{rot}, \mathcal{L}_{iso}$. To address image blurriness, we optimize the pose parameters. The output includes both the point cloud and the appearance of the reconstructed human.

3. GaussianBody

In this section, we first introduce the preliminary method 3D-GS [24] in Section 3.1. Next, we describe our framework pipeline for 3D-GS-based clothed body reconstruction (Section 3.2). We then discuss the application of a physically-based prior to regularize the 3D Gaussians across the canonical and observation spaces (Section 3.3). Finally, we introduce two strategies, split-with-scale and pose refinement, to enhance point cloud density and optimize the SMPL parameters, respectively (Section 3.4).

3.1. Preliminary

3D-GS [24] is an explicit 3D scene reconstruction method designed for multi-view images. The static model comprises a list of Gaussians with a point cloud at its center. Gaussians are defined by a covariance matrix Σ and a center point X, representing the mean value of the Gaussian:

$$G(X) = e^{-\frac{1}{2}X^T \Sigma^{-1} X}. \quad (1)$$

For differentiable optimization, the covariance matrix Σ can be decomposed into a scaling matrix S and a rotation matrix R:

$$\Sigma = RSS^T R^T, \quad (2)$$

The gradient flow computation during training is detailed in [24]. To render the scene, the regressed Gaussians can be projected into camera space with the covariance matrix Σ':

$$\Sigma' = JW\Sigma W^T J^T, \quad (3)$$

Here, J is the Jacobian of the affine approximation of the projective transformation, and W is the world-to-camera matrix. To simplify the expression, the matrices R and S are preserved as rotation parameter r and scaling parameter s. After projecting the 3D Gaussians to 2D, the alpha-blending rendering based on point clouds bears a resemblance to the volumetric rendering equation of NeRF [14] in terms of its formulation. During volume rendering, each Gaussian is defined by an opacity α and spherical harmonics coefficients c to represent the color. The volumetric rendering equation for each pixel contributed by Gaussians is given by:

$$C = \sum_{i \in \mathbf{N}} c_i \alpha_i, \quad (4)$$

Collectively, the 3D Gaussians are denoted as $G(x, r, s, \alpha, c)$.

3.2. Framework

In our framework, we decompose the dynamic clothed human modeling problem into the canonical space and the motion space. First, We define the template 3D Gaussians in the canonical space as $G(\bar{x}, \bar{r}, \bar{s}, \bar{\alpha}, \bar{c})$. To learn the template 3D Gaussians, we employ pose-guidance deformation fields to transform them into the observation space and render the scene using differentiable rendering. The gradients of the pose transformation are recorded for each time and used for backward optimization in the canonical space.

Specifically, we utilize the parametric body model SMPL [6] as pose guidance. The articulated SMPL model
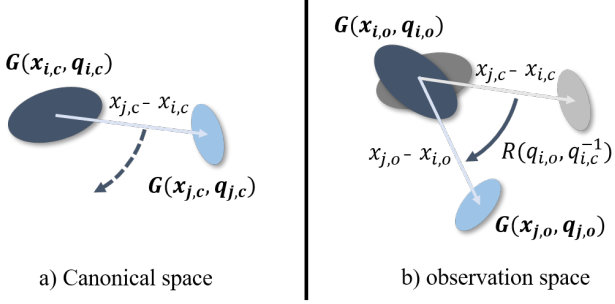
Figure 3. **Local-rigidity loss.** With the Gaussians i rotating between the two spaces, the neighbour Gaussians j should move to follow the rigid-transform in the coordinate system of Gaussians i

$M(\beta, \theta)$ is defined with pose parameters $\theta \in R^{3n_k+3}$ and shape parameters $\beta \in R^{10}$. The transformation of each point is calculated with the skinning weight field $w(\beta)$ and the target bone transformation $B(\theta) = \{B_1(\theta), ..., B_n(\theta)\}$. To mitigate computational costs, we adopt the approach from InstantAvatar [17], which diffuses the skinning weight of the SMPL [6] model vertex into a voxel grid. The weight of each point is then obtained through trilinear interpolation from the grid weighting, denoted as $w(\bar{x}), \beta$. The transformation of the canonical points to deform space via forward linear blend skinning is expressed as:

$$\mathcal{D}(\bar{x}, \theta, \beta) = \sum_{i=1}^{n} w_i(\bar{x}, \beta) B_i(\theta). \quad (5)$$

With the requirements of the 3D-GS[24] initial setting, we initialize the point cloud with the template SMPL[6] model vertex in the canonical pose(as shown in Figure.2). For each frame, we deform the position $\bar{x}$ and rotation $\bar{r}$ of the canonical Gaussians $G(\bar{x}, \bar{r}, \bar{s}, \bar{\alpha}, \bar{c})$ with the pose parameter θ_t of current frame and the global shape parameter β to the observation space :

$$\begin{aligned} x &= \mathcal{D}(\bar{x}, \theta_t, \beta)\bar{x}, \\ r &= \mathcal{D}(\bar{x}, \theta_t, \beta)\bar{r} \end{aligned} \quad (6)$$

where $\mathcal{D}$ is the deformation function defined in Eq.5.

In this way, we obtain the deformed Gaussians in the observation space. After differentiable rendering and image loss calculation, the gradients will be passed through the inverse of the deformation field $\mathcal{D}$ and optimized for the canonical Gaussians.

3.3. Physically-Based Prior

Although we define the canonical Gaussians and explicitly deform them to the observation space for differentiable rendering, the optimization is still an ill-posed problem because there could be multiple canonical positions mapped to the same observation position, leading to overfitting in

the observation space and visual artifacts in the canonical space. In the experiment, we also observed that this optimization approach might easily result in the novel view synthesis showcasing numerous Gaussians in incorrect rotations, consequently generating unexpected glitches. Thus we follow [28] to regularize the movement of 3D Gaussians by their local information. Particularly we employ three regularization losses to maintain the local geometry property of the deformed 3D Gaussians, including local-rigidity loss $\mathcal{L}_{rigid}$, local-rotation loss $\mathcal{L}_{rot}$ losses and a local-isometry loss $\mathcal{L}_{iso}$. Different from [28] that attempts to track the Gaussians frame by frame, we regularize the Gaussian transformation from the canonical space to the observation space.

Given the set of Gaussians j with the k-nearest-neighbors of i in canonical space (k=20), the isotropic weighting factor between the nearby Gaussians is calculated as:

$$w_{i,j} = exp(-\lambda_w||x_{j,c} - x_{i,c}||_2^2), \quad (7)$$

where $||x_{j,c} - x_{i,c}||$ is the distance between the Gasussians i and Gasussians j in canonical space, set $\lambda_w = 2000$ that gives a standard deviation. Such weight ensures that rigidity loss is enforced locally and still allows global non-rigid reconstruction. The local rigidity loss is defined as:

$$\mathcal{L}_{i,j}^{rigid} = w_{i,j}||(x_{j,o} - x_{i,o} - R_{i,o}R_{i,c}^{-1}(x_{j,c} - x_{i,c}))||_2, \quad (8)$$

$$\mathcal{L}_{rigid} = \frac{1}{k|\mathcal{S}|} \sum_{i \in \mathcal{S}} \sum_{j \in knn_{i;k}} \mathcal{L}_{i,j}^{rigid}, \quad (9)$$

when the Gaussians i transform from canonical space to observation space, the nearby Gaussians j should move in a similar way that follows the rigid-body transform of the co-ordinate system of the Gaussians i between two spaces. The visual explanation is shown in Figure.6.

While the rigid loss ensures that Gaussians i and Gaussians j share the same rotation, the rotation loss could enhance convergence to explicitly enforce identical rotations among neighboring Gaussians in both spaces:

$$\mathcal{L}_{rot} = \frac{1}{k|\mathcal{S}|} \sum_{i \in \mathcal{S}} \sum_{j \in knn_{i;k}} w_{i,j}||q_{j,o}q_{j,c}^{-1} - q_{i,o}q_{i,c}^{-1}||_2, \quad (10)$$

where q is the normalized Quaternion representation of each Gaussian's rotation, the $q_o q_c^{-1}$ demonstrates the rotation of the Gaussians with the deformation. We use the same Gaussian pair sets and weighting function as before.

Finally, we use a weaker constraint than $\mathcal{L}^{rigid}$ to make two Gaussians in different space to be the same one, which only enforces the distances between their neighbors:

$$\mathcal{L}_{iso} = \frac{1}{k|\mathcal{S}|} \sum_{i \in \mathcal{S}} \sum_{j \in knn_{i;k}} w_{i,j}\{||x_{j,o} - x_{i,o}||_2 - ||x_{j,c} - x_{i,c}||_2\},$$

$$(11)$$

after adding the above objectives, our objective is :

$$\mathcal{L} = \mathcal{L}_{image} + \lambda_{rigid}\mathcal{L}_{rigid} + \lambda_{rot}\mathcal{L}_{rot} + \lambda_{iso}\mathcal{L}_{iso}. \quad (12)$$

3.4. Refinement Strategy

Split-with-scale. Because the monocular video input for 3D-GS [24] lacks multi-view supervision, a portion of the reconstructed point cloud (3D Gaussians) may become excessively sparse, leading to oversized Gaussians or blurring artifacts. To address this, we propose a strategy to split large Gaussians using a scale threshold ϵ_{scale}. If a Gaussian has a size s larger than ϵ_{scale}, we decompose it into two identical Gaussians, each with half the size.

Pose refinement. Despite the robust performance of 3D-GS [24] in the presence of inaccurate SMPL parameters, there is a risk of generating a high-fidelity point cloud with inaccuracies. The inaccurate SMPL parameters may impact the model's alignment with the images, leading to blurred textures. To address this issue, we propose an optimization approach for the SMPL parameters. Specifically, we designate the SMPL pose parameters θ as the optimized parameter and refine them through the optimization process, guided by the defined losses.

4. Experiment

In this section, we evaluate our method on monocular training videos and compare it with the other baselines and state-of-the-art works. We also conduct ablation studies to verify the effectiveness of each component in our method.

4.1. Datasets and Baseline

PeopleSnapshot. PeopleSnapshot [10] dataset contains eight sequences of dynamic humans wearing different outfits. The actors rotate in front of a fixed camera, maintaining an A-pose during the recording. We train the model with the frames of the human rotating in the first two circles and test with the resting frames. The dataset also provides inaccurate shape and pose parameters. So we first process both the train dataset and test dataset to get the accurate pose parameters. Note that the coefficients of the Gaussians remain fixed. We evaluate the novel view synthesis quality with frame size in 540×540 with the quantitative metrics including peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS), and train our model and other baselines in 1080×1080 for visual comparison. The PeopleSnapshot dataset doesn't have the corresponding ground truth point cloud, we only provide qualitative results.

iPER. iPER [52] dataset consists of two sets of monocular RGB videos depicting humans rotating in an A-pose or engaging in random motions before a static camera. In the experiment, we leverage the A-pose series, adopting a similar setting to PeopleSnapshot. Subsequently, we visualize

novel views and compare them with baselines to demonstrate the robustness of our method.

Baselines. We compare our method with original 3D-GS [24], Neural body[15], Anim-NeRF [16] and InstantAvatar [17]. Neural body[15] adopts the SMPL vertexes as the set of latent code to record the local feature and reconstruct humans in NeRF. Anim-nerf [16] uses the explicit pose-guidance deformation that deforms the query point in observation space to canonical space with inverse linear blend skinning. InstantAvatar [17] builds the hash grid to restore the feature of NeRF and control the query points with Fast-SNARF [51], which uses the root-finding way to find the corresponding points and transform it to the observation space to optimize the articulate weighting.

4.2. Implementation Details

GaussianBody is implemented in PyTorch and optimized with the Adam [53]. We optimize the full model in 30k steps following the learning rate setting of official implementation, while the learning rate of position is initial in $6e^{-6}$ and the learning rate of pose parameters is $1e^{-3}$. We set the hyper-parameters as $\lambda_{rigid} = 4e^{-2}$, $\lambda_{rot} = 4e^{-2}$, $\lambda_{iso} = 4e^{-2}$. For training the model, it takes about 1 hour on a single RTX 4090.

4.3. Results

Novel view synthesis. In Table 1, our method consistently outperforms other approaches in various metrics, highlighting its superior performance in capturing detailed reconstructions. This indicates that our method excels in reconstructing intricate cloth textures and human body details.

Figure 5 visually compares the results of our method with others. 3D-GS [24] struggles with dynamic scenes due to violations of multi-view consistency, resulting in partial and blurred reconstructions. Our method surpasses InstantAvatar in cloth texture details, such as sweater knit patterns and facial features. Even with inaccurate pose parameters on iPER [52], our method demonstrates robust results. InstantAvatar's results, on the other hand, are less satisfactory, with inaccuracies in pose parameters leading to deformation artifacts.

Figure 4 showcases realistic rendering results from different views, featuring individuals with diverse clothing and hairstyles. These results underscore the applicability and robustness of our method in real-world scenarios.

3D reconstruction. In Figure 4, the qualitative results of our method are visually compelling. The generated point clouds exhibit sufficient details to accurately represent the human body and clothing. Examples include the organic wrinkles in the shirt, intricate facial details, and well-defined palms with distinctly separated fingers. This level of detail in the point cloud provides a solid foundation for

	male-3-casual			male-4-casual			female-3-casual			female-4-casual		
	PSNR↑	**SSIM↑**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
3D-GS[24]	26.60	0.9393	0.082	24.54	0.9469	0.088	24.73	0.9297	0.093	25.74	0.9364	0.075
NeuralBody[15]	24.94	0.9428	0.033	24.71	0.9469	0.042	23.87	0.9504	0.035	24.37	0.9451	0.038
Anim-NeRF[16]	29.37	0.9703	0.017	28.37	0.9605	**0.027**	28.91	**0.9743**	**0.022**	28.90	0.9678	**0.017**
InstantAvatar[17]	29.64	0.9719	**0.019**	28.03	0.9647	0.038	28.27	0.9723	0.025	29.58	**0.9713**	0.020
Ours	**35.66**	**0.9753**	0.021	**32.65**	**0.9769**	0.049	**33.22**	0.9701	0.037	**31.43**	0.9630	0.040

Table 1. **Quantitative comparison of novel view synthesis on PeopleSnapshot dataset.** Our approach exhibits a significant advantage in metric comparisons, showing substantial improvements in PSNR and SSIM metrics due to its superior restoration of image details.
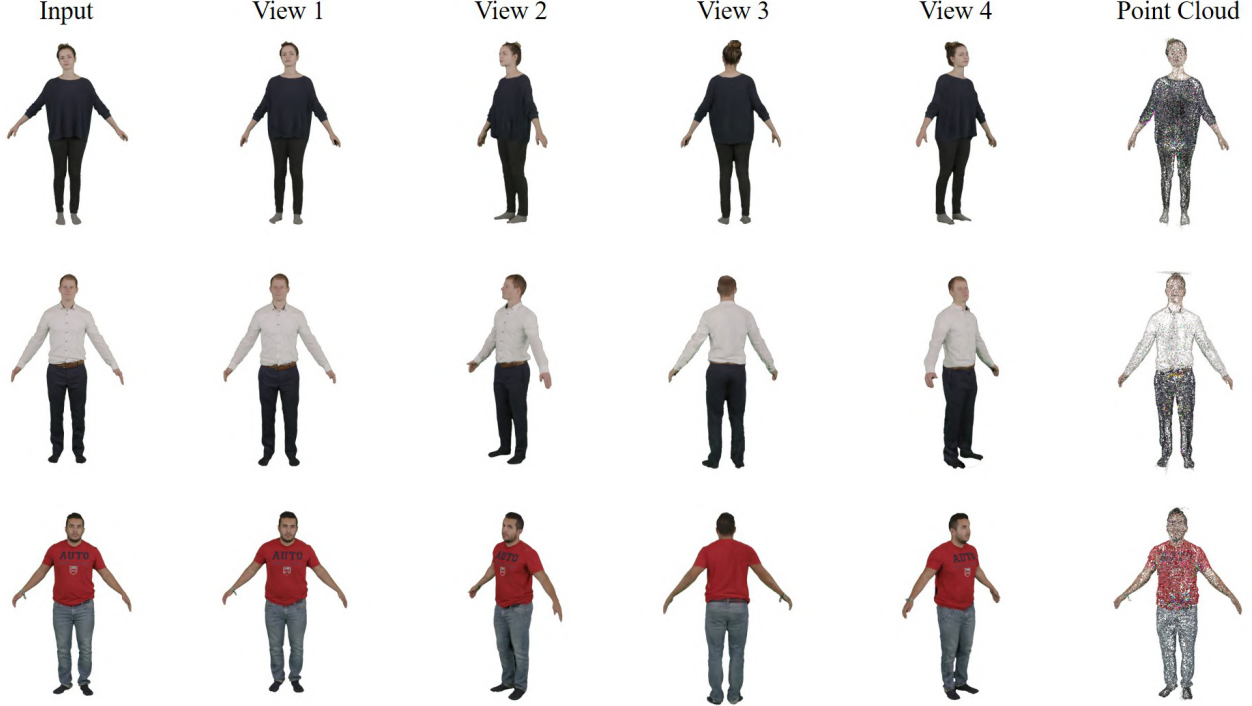

Figure 4. **Results of novel view synthesis and point cloud on PeopleSnapshot [10] dataset.** Our method effectively restores details on the human body, including intricate details in the hair and folds on the clothes. Moreover, the generated point cloud faithfully captures geometric details on the clothing, demonstrating a commendable separation between geometry and texture.

handling non-rigid deformations of the human body more accurately.

4.4. Ablation Study

Physically-based prior. To evaluate the impact of the physically-based prior, we conducted experiments by training models with and without the inclusion of both part-specific and holistic physically-based priors. Additionally, we visualized the model in the canonical space with the specified configurations.

Figure 6 illustrates the results. In the absence of the physically-based prior, the model tends to produce numerous glitches, especially leading to blurred facial features. Specifically, the exclusion of the rigid loss contributes to facial blurring. On the other hand, without the rotational loss,

the model generates fewer glitches, although artifacts may still be present. The absence of the isometric loss introduces artifacts stemming from unexpected transformations.

Only when incorporating all components of the physically-based prior, the appearance details are faithfully reconstructed without significant artifacts or blurring.

Pose refinement. We utilize SMPL parameters for explicit deformation, but inaccurate SMPL estimation can lead to incorrect Gaussian parameters, resulting in blurred textures and artifacts. Therefore, we introduce pose refinement, aiming to generate more accurate pose parameters, as depicted in Figure 7. This refinement helps mitigate issues related to blurred textures caused by misalignment in the observation space and avoids the need for the deformation
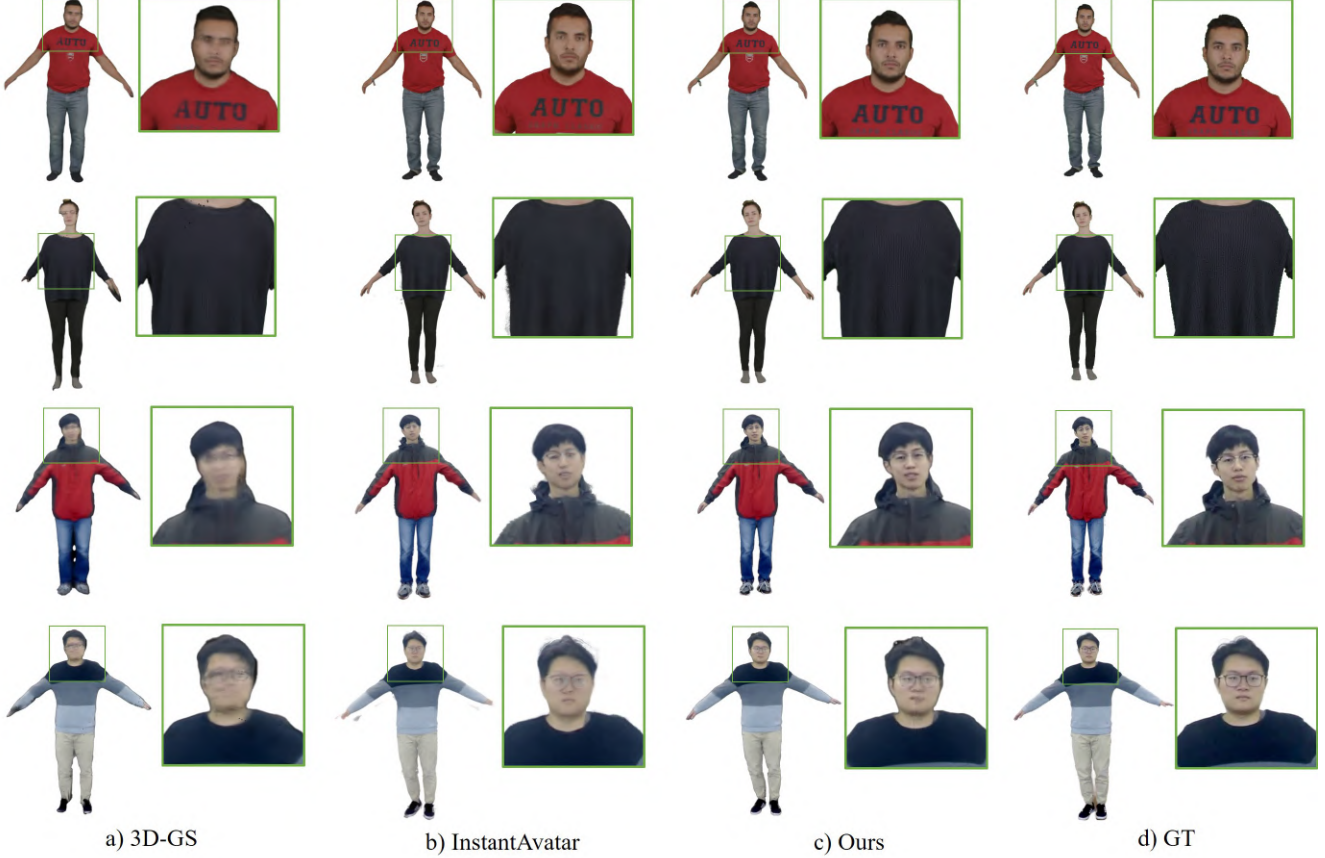
Figure 5. **Visual comparison** of different methods about novel view synthesis on PeopleSnapshot[10](column 1&2) and iPER[52](column 3&4). 3D-GS[24] rely on multi-view consistency to gain center subjection which failed to handle dynamic scenes. InstantAvatar[17] trade the quality and robust for time, might gain blur result on inaccurate parameters. Our method gains high-fidelity results, especially on the cloth texture and the robustness.

MLP to fine-tune in each frame, as illustrated in Figure 9.

Split-with-scale. Given the divergence in our input compared to the original Gaussian input, especially the absence of part perspectives, the optimization process tends to yield a relatively sparse point cloud. This sparsity affects the representation of certain details during pose changes. As shown in Figure 8, we address this issue by enhancing point cloud density through a scaling-based splitting approach.

4.5. Discussion on Gaussian Deformation

Deformation MLP. We observe that the parameters after the deformation MLP network might become random, potentially misleading the optimization in the canonical space with differentiable rendering. To capture non-rigid deformation after the pose-guidance deformation, we adopt an approach inspired by SCARF [54] and Deformable3dgs [25] by introducing a deformation MLP network into the 3D-GS [24] pipeline. We concatenate the Gaussians' point positions after the forward transform and their corresponding vertices into a fully connected neural network, aiming to obtain more accurate Gaussian parameters that capture the non-rigid deformation of the cloth in the observation space.

However, we encounter challenges where the canonical Gaussians lose generalization for novel views and pose synthesis, as illustrated in Figure 9. The deformation MLP tends to overfit the observation space and even influences the result of the rigid transformation. This issue needs further investigation and optimization to achieve a more balanced representation.

Novel pose synthesis. The explicit representation offers several advantages, including accelerated training, simplified interfacing, and effective deformation handling. However, challenges arise from the sparse Gaussians and imprecise deformation, affecting the representation of novel poses. The sparsity of Gaussians, combined with the absence of accurate deformation, makes it challenging to rep-

Full loss   wo L_{rigid*}

wo L_{rigid} wo L_{rot} wo L_{iso}

Figure 6. **Effect of physically-based prior.** This figure demonstrates what would be influenced without different losses. Each loss influences different parts of the canonical model, the rigid loss regularizes a part of the wired rotation, the rot loss mainly reduces the glitch, and the iso loss reduces the unexpected transformation.

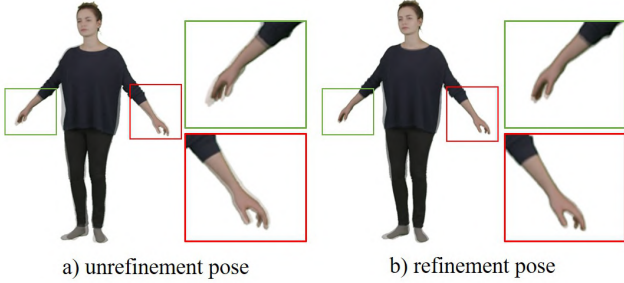a) unrefinement pose   b) refinement pose

Figure 7. **Effect of pose refinement.** In the figure, the semi-transparent is the human appearance of our method. The hand part can demonstrate our method could adjust the pose.
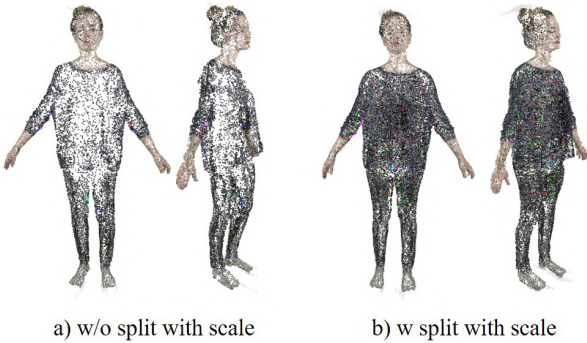

a) w/o split with scale   b) w split with scale

Figure 8. **Effect of splitting with the scale.** Without our optimization, the point cloud would be sparse. After the optimization, the point cloud could express more detail of the model

a) Input   Animation w/o   Optimize with
Physical-based prior   Deform MLP
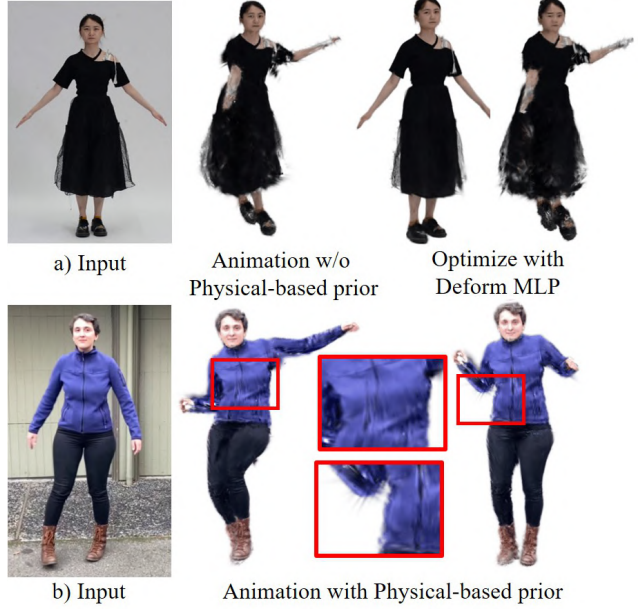
b) Input   Animation with Physical-based prior

Figure 9. **Failure cases in deformation.** This picture demonstrates some results of our tries in deformation. Due to conflicts between the deformation MLP and differentiable rendering, the model exhibits inaccuracies in both the canonical model and novel pose synthesis. The Gaussians represent an elliptical region, leading to artifacts in the absence of precise deformation.

resent a continuous surface. Despite attempts to mitigate these issues by reducing the size of Gaussians and applying regularization through the physically-based prior, unexpected glitches persist in novel poses, as illustrated in Figure 9. Overcoming these challenges to reconstruct a reasonable non-rigid transformation of the cloth surface remains an area for improvement.

5. Conclusion

In this paper, we present a novel method called Gaussian-Body for reconstructing dynamic clothed human models from monocular videos using the 3D Gaussians Splatting representation. By incorporating explicit pose deformation guidance, we extend the 3D-GS representation to clothed human reconstruction. To mitigate over-rotation issues between the observation space and the canonical space, we employ a physically-based prior to regularize the canonical space Gaussians. Additionally, we incorporate pose refinement and a split-with-scale strategy to enhance both the quality and robustness of the reconstruction. Our method achieves comparable image quality metrics with the baseline and other methods, demonstrating competitive performance, relatively fast training speeds, and the capability to train with higher resolution images.

References

[1] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 1

[2] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020.

[3] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11179–11188, 2021.

[4] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, pages 792–804. IEEE, 2021. 1

[5] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 1

[6] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015. 1, 2, 3, 4

[7] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 1

[8] Ahmed AA Osman, Timo Bolkart, and Michael J Black. Star: Sparse trained articulated human body regressor. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 598–613. Springer, 2020. 1

[9] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 1, 2

[10] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard 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. 1, 2, 5, 6, 7

[11] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 2

[12] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2

[13] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 2

[14] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[15] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 2, 5, 6

[16] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos. *arXiv preprint arXiv:2106.13629*, 2021. 5, 6

[17] Tianjian Jiang, Xu Chen, Jie Song, and Otmar 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, 4, 5, 6, 7

[18] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira 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. 2

[19] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision*, pages 419–436. Springer, 2022.

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

[21] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 2

[22] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich 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. 2

[23] Haitao Yu, Deheng Zhang, Peiyuan Xie, and Tianyi Zhang. Point-based radiance fields for controllable human motion synthesis, 2023. 2

[24] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time

radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 2, 3, 4, 5, 6, 7

[25] 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. 2, 7

[26] Zeyu Yang, Hongye Yang, Zijie Pan, Xiatian Zhu, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting, 2023.

[27] 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. 2

[28] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024. 2, 4

[29] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018. 2

[30] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12869–12879, 2023. 2

[31] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 2

[32] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3170–3184, 2021. 2

[33] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020.

[34] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11046–11056, 2021. 2

[35] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc 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. 2

[36] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.

[37] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-

Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), dec 2021. 2

[38] Haotong Lin, Sida Peng, Zhen Xu, Tao Xie, Xingyi He, Hujun Bao, and Xiaowei Zhou. Im4d: High-fidelity and real-time novel view synthesis for dynamic scenes. *arXiv preprint arXiv:2310.08585*, 2023. 2

[39] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 2

[40] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, et al. Human performance modeling and rendering via neural animated mesh. *ACM Transactions on Graphics (TOG)*, 41(6):1–17, 2022. 2

[41] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-fidelity clothed avatar reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8662–8672, 2023. 2

[42] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022.

[43] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023. 2

[44] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021. 2

[45] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021.

[46] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021.

[47] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16569–16578, 2023. 2

[48] Sida Peng, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Representing volumetric videos as dynamic mlp maps. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 4252–4262, 2023. 2

[49] Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zhang, Minye Wu, Jingyi Yu, and Lan 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. 2

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

[51] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 5

[52] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5904–5913, 2019. 5, 7

[53] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Ithaca, NYarXiv.org*, 2014. 5

[54] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 7