SHAPEGAUSSIAN: DYNAMIC GAUSSIAN SPLATTING FOR MONOCULAR VIDEOS WITH NON-PARAMETRIC SHAPE REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we tackle the challenging underconstrained problem of reconstructing dynamic objects from monocular videos using a new method, which we term ShapeGaussian. This approach incorporates shape priors in an implicit way to enhance reconstruction accuracy and chance of success with few multi-view clues. Our methodology employs a two-phase process. In the first stage, we establish a temporally consistent deformation model across frames based on depth maps and keypoint estimations derived from a pre-trained model. The second stage obtains high-quality photorealistic reconstruction by optimizing 3D Gaussian jointly with non-parametric shape models. Through rendering this combined model into radiance fields, we achieve high-quality, photo-realistic reconstructions of dynamically deforming objects that maintain 3D consistency across novel views. Our results demonstrate a significant improvement over previous methods on human dynamics, particularly in scenarios with scarce multi-view cues, highlighting the persistent challenges and varied approaches in recent research aimed at this inherently complex task.

1 INTRODUCTION

Reconstruction of photo-realistic 3D scenes remains a fundamental challenge in computer vision, with a wide array of applications from video games to robotics. Significant advancements have been achieved in the reconstruction of static scenes in recent years. However, in real-world scenarios, the focus often shifts to capturing dynamic scenes, typically through monocular video recordings done in a casual manner. Consequently, addressing the dynamics of the scene and the limitations of monocular input during the reconstruction process is crucial.

A significant limitation of existing Gaussian-based reconstruction methods is their poor performance on monocular video captures featuring rapid human motion, particularly in novel view synthesis. This deficiency primarily stems from the inaccurately reconstructed geometry of the Gaussians, compounded by the unconstrained nature of the optimization problem due to the limited information available from monocular captures. Although recent efforts have attempted to address this issue by incorporating explicit mesh-based priors into monocular video-based Gaussian avatars, these methods still struggle with reconstructing loose clothing and hair, as well as the variable and smooth appearance of the human subject over time.

Most state-of-the-art Gaussian-based methods for human-centric dynamic scene reconstruction from monocular videos focus primarily on either avatar reconstruction with a strong template prior or general-purpose scene reconstruction using a deformation neural model. However, avatar-fitting techniques are limited by the expressiveness of the underlying template and do not capture time-dependent environmental changes. Consequently, the same pose vector invariably produces the same novel view rendering, regardless of changes in hair, clothing, and lighting. Another prevalent approach involves using a neural deformation model to calculate the offset from a canonical scene representation, but this often requires multi-view input data, which is impractical in most applications and leads to suboptimal results due to the absence of semantic priors.

In this paper, we address the challenge of dynamic human-centric novel-view synthesis using realistic monocular captures. We introduce a two-step method that initially learns a coarse, deformable

geometry using pretrained models that estimate geometry and pose information, and then utilizes this geometry as a basis for reconstruction. The first step involves creating a coarse model for each frame. In the second step, we utilize a neural deformation model to capture the dynamic deformation details, building upon the dynamic coarse point template.

In short, our contributions are:

- We propose a unified Gaussian splatting framework *ShapeGaussian*, leveraging implicit shape priors and neural deformation network for modeling the shape deformation, to support general-purpose reconstruction of dynamic environments and object reconstruction with enhanced accuracy and details. For this goal we implement an adaptive density control mechanism designed for dynamic base model.
- Instead of utilizing explicit human templates like SMPL, our method is more flexible to represent interactions between environment and human.
- Our approach integrates pose-independent details into our model, diverging from traditional template-based human reconstruction techniques.

2 Related Work

Given our primary objective of synthesizing novel views for dynamic 3D scenes involving significant human motion, there are two main areas of related work: general-purpose dynamic scene reconstruction and the incorporation of explicit mesh constraints for avatar reconstruction in canonical poses.

In this section, we provide a concise review of Gaussian-based dynamic 3D scene reconstruction in Section 2.2. We then examine Gaussian-based avatar methods in Section 2.3.

2.1 Implicit Neural Rendering of Dynamic Scenes

NeRF Mildenhall et al. (2020) has been adapted to dynamically model scenes, utilizing either a global MLP framework or a hybrid setup where neural features are linked to the nodes of a discrete data structure. This adaptation captures scene dynamics in two primary ways.

The first strategy, known as Space-Time Neural Fields, introduces an additional time dimension to the scene structure, enabling the reconstruction of dynamic scenes from either multi-view Attal et al. (2023); Li et al. (2022a;b); Lin et al. (2023a); Park et al. (2023); Song et al. (2023b); Wang et al. (2023b; 2022a; 2023c; 2022b) or monocular video sources Cao & Johnson (2023); Du et al. (2021); Fridovich-Keil et al. (2023); Gao et al. (2021); Li et al. (2021; 2023b); Shao et al. (2023); Song et al. (2023a); Xian et al. (2021). Hybrid techniques enhance this process by parameterizing the 4D scene representation through voxel grid Li et al. (2022a); Park et al. (2023); Song et al. (2023a); Wang et al. (2023b; 2022a; 2023c; 2022b) or planar factorizatin Attal et al. (2023); Cao & Johnson (2023); Fridovich-Keil et al. (2023); Lin et al. (2023a); Shao et al. (2023). While effective for forward-facing videos, these methods depend heavily on pre-calculated depth maps and optical flow for local motion tracking but are less capable of global information transfer necessary for rendering new viewpoints.

Similar to our approach, HexplaneCao & Johnson (2023) uses a low-rank temporal framework for managing spatially separated volumes, concurrently optimizing this basis with the detailed geometry and appearance. Conversely, our method first establishes a rough template and then refines the details in a subsequent phase, providing stronger regularization.

The second approach, Deformable NeRFs, employs a 4D deformation field that maps each timestep's observations back to a standard configuration, ensuring temporal alignment. This technique maintains consistency but restricts large variations and topological shifts from the standard model Choe et al. (2023); Park et al. (2021a;b); Pumarola et al. (2020); Tretschk et al. (2021); Wang et al. (2023a). Several models use a voxel-based hybrid system for quick multi-view reconstruction Liu et al. (2022); Tretschk et al. (2023). TiNeuVox Fang et al. (2022) enhances monocular reconstruction by utilizing a minimal MLP for the deformation field and compensating with improved temporal scene representations. However, backward warping can be problematic in sparse

capture settings due to its lack of smoothness, requiring strong regularization for effective new view synthesis.

2.2 Dynamic Reconstruction with 3D Gaussians

Several recent work have explored modeling dynamic secenes using 3D Gaussians. Dynamic3DGS Luiten et al. (2023) utilizes dense multi-view inputs to maintain the tracking of 3D Gaussians' positions and rotations from their initial pacement in the first frame throughout various timesteps. Similarly, Deformable3DGS Yang et al. (2023), GaussianFlow Lin et al. (2023b) and 4DGS Wu et al. (2023) implement a canonical space for initilizing 3D Gaussians, employing a neural deformation field to track changes in positions, rotations and scales over time. Notably, 4DGS also describes the deformation field with a Hexplane representation mixing temporal and spacial dimensions.

2.3 Human-centric Dynamic Reconstruction from Monocular Videos

Due to high costs of scanning and the labor-intensive nature, it is practically infeasible for early technologies that utilized RGB-D sensors to capture subject's shape and then manually attaching the captured surface to preset skeleton to generate the avatar Dou et al. (2016; 2017); Izadi et al. (2011); Newcombe et al. (2015); Yu et al. (2017; 2018), with the advent of parametric models like SMPL Loper et al. (2015) and SMPL-X Pavlakos et al. (2019), creating avatars economically became feasible. These models enable avatar creation using just RGB images, bypassing the need for expensive scanning equipment.

Recently, innovative approaches in avatar reconstruction have developed, leveraging a parameterized human body as a foundation while incorporating advanced rendering techniques such as vertex offsets Ma et al. (2020); Xiang et al. (2020), signed distance fields He et al. (2020); Saito et al. (2019; 2020); Varol et al. (2018); Xiu et al. (2022; 2023), neural radiance fields (NeRF) Jiang et al. (2022b;a); Kwon et al. (2021); Peng et al. (2021b;a); Weng et al. (2022), and the most recent 3D Gaussians Hu et al. (2023); Jung et al. (2023); Li et al. (2023a); Qian et al. (2023b;a); Saito et al. (2023); Yuan et al. (2024); Zielonka et al. (2023); Kocabas et al. (2023). These methods enhance the realism of avatars by capturing detailed individual shape features. Such improvements significantly boost the expressiveness of avatars, yet the reconstruction's fidelity largely hinges on precise pose estimation. Despite these advancements, the focus remains predominantly on the human form itself, often overlooking the dynamics of the background. While methods like HUGS Kocabas et al. (2023) can generate novel views of a scene, they tend to produce unnatural human movements and are limited to static backgrounds, highlighting a gap in current modeling capabilities that could benefit from further innovation.

3 Preliminaries

In this sections, we simply review the representation of 3D-GS Kerbl et al. (2023) in Sec. 3.1 and the formulation of priors used by our method in Sec. 3.2.

3.1 3D Gaussian Splatting

3D Gaussians Kerbl et al. (2023) serve as an explicit 3D scene representation through point clouds. Each Gaussian is defined by a 5-tuple $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{s}, o, \boldsymbol{c})$, where $\boldsymbol{\mu} \in \mathbb{R}^3, \boldsymbol{Sigma} \in \mathbb{SO}(3)$ are the 3D mean and orientation and $\boldsymbol{s} \in \mathbb{R}^3$ the scale, $o \in \mathbb{R}$ the opacity, and $\boldsymbol{c} \in \mathbb{R}^3$ the color. The rendering process would first project 3D Gaussians onto the 2D image plane. More specifically, given the world-to-camera extrinsics $\mathbf{E}$ and intrinsics $\mathbf{K}$, the projection of the 3D Gaussians can be obtained by formula

$$\boldsymbol{\mu}'(\mathbf{K}, \mathbf{E}) := \Pi(\mathbf{KE}\boldsymbol{\mu}) \in \mathbb{R}^2, \quad \Sigma'(\mathbf{K}, \mathbf{E}) := \mathbf{J}_{\mathbf{KE}} \Sigma_0 \mathbf{J}_{\mathbf{KE}}^T \in \mathbb{R}^{2\times 2}, \tag{1}$$

where Π is the perspective projection operator, and $\mathbf{J}_{\mathbf{KE}}$ is the Jacobian matrix from the affine approximation of the projective transformation determined by $\mathbf{E}$ and $\mathbf{K}$ at location $\boldsymbol{\mu}$. The projected 2D Gaussians can then be efficiently rasterized into RGB image along with the depth map via volume

rendering as

$$\hat{\mathbf{I}}(\boldsymbol{p}) := \sum_{i \in H(\boldsymbol{p})} T_i \alpha_i \boldsymbol{c}_i, \quad \hat{\mathbf{D}}(\boldsymbol{p}) := \sum_{i \in H(\boldsymbol{p})} T_i \alpha_i \boldsymbol{d}_i, \tag{2}$$

where $H(\boldsymbol{p})$ denotes the index set of Gaussians that intersect the ray shoot from pixel $\boldsymbol{p}$, and the equivalent opacity and transmittance is calculated by

$$\alpha_i := o_i \cdot \exp\big(-\tfrac{1}{2}(\boldsymbol{p}-\boldsymbol{\mu}')^T \boldsymbol{\Sigma}'(\boldsymbol{p}-\boldsymbol{\mu}')\big), \quad T_i := \prod_{j<i}(1-\alpha_j). \tag{3}$$

3.2 Data-driven Priors

We consider three types of data-driven priors produced by off-the-shelf pretrained models, which assist our model in inferring accurate geometry from videos that lack sufficient multi-view cues. For each training image $\mathbf{I}$, we have the following:

Semantic Mask. Represented by a map $\mathbf{M_I}$, where $\mathbf{M_I}(\boldsymbol{p}) = 1$ if and only if the pixel $\boldsymbol{p}$ is within the human silhouette.

Depth Map. Represented by a map $\mathbf{D_I}$, where $\mathbf{D_I}(\boldsymbol{p})$ indicates the distance of the point on the front-most surface from the viewpoint. In our method, only the human region is considered for the depth map.

Human Keypoints. Denoted by $\boldsymbol{z_I} \in \mathbb{R}^{2\times m}$, which corresponds to the m 2D keypoints estimated for image $\mathbf{I}$, where m is a constant in our model. It is important to note that, in general, the k-th keypoint may not be available for every training image $\mathbf{I}$. If a keypoint in $\boldsymbol{z_I}$ is unavailable, we set its value to $null$. These keypoints, $\boldsymbol{z_I}$, naturally provide alignment across frames. We will explain this in more detail in Sec. 4.1.

4 Method

In this section, we outline a 3-stage approach. Initially, we derive a coarse dynamic Gaussian model against semantic priors derived from the input monocular video. Subsequently, we fit a Gaussian deformation network that fits the coarse model established in the first stage. Finally we optimize the deformation model against the input video. This stage incorporates adaptive density control to ensure high-quality 4D reconstruction as well.

The input data required by our method includes the input views $\{\mathbf{I}_t\}$, SfM points $\boldsymbol{q}_i$, and prior information $\{(\mathbf{M_I}, \mathbf{D_I}, \boldsymbol{z_I})\}_{\mathbf{I}}$ associated to $\mathbf{I}$ as introduced in Sec.3.2. For real videos we estimate the camera poses and SfM points via COLMAP Schönberger & Frahm (2016); Schönberger et al. (2016), and the semantic depth maps and human keypoints are estimated by Sapiens model Khirodkar et al. (2024).

4.1 Stage I: Dynamic Coarse Model

A coarse model will be established in stage I to depict the dynamic geometry shape by utilizing the semantic priors. Simply speaking, the depth maps for human could help us determine a collection of 3D points associated to human bodies, while 2D keypoints priors enable picking important points and aligning them up.

Initialization. We define two distinct point sets: the static background points, denoted as $\mathcal{P}_b$, and the dynamic human points at each timestamp t, denoted as $\mathcal{P}_h^{(t)}$. For the static background, the setup is relatively straightforward, as we directly use the locations of the input structure-from-motion (SfM) points. On the other hand, the dynamic human points $\mathcal{P}_h^{(t)}$ are modeled by treating all available keypoints as Gaussian distributions, which we refer to as *key Gaussians*.

Starting from the first frame, we augment these key Gaussians by randomly sampling points uniformly within the human mask. In subsequent frames, we iteratively refine this sampling process to account for changes in the human motion and deformation.

It's important to note that we do not impose a strict constraint that the background must remain static throughout the process. The background is only fixed in the initial stage to provide a reference, but

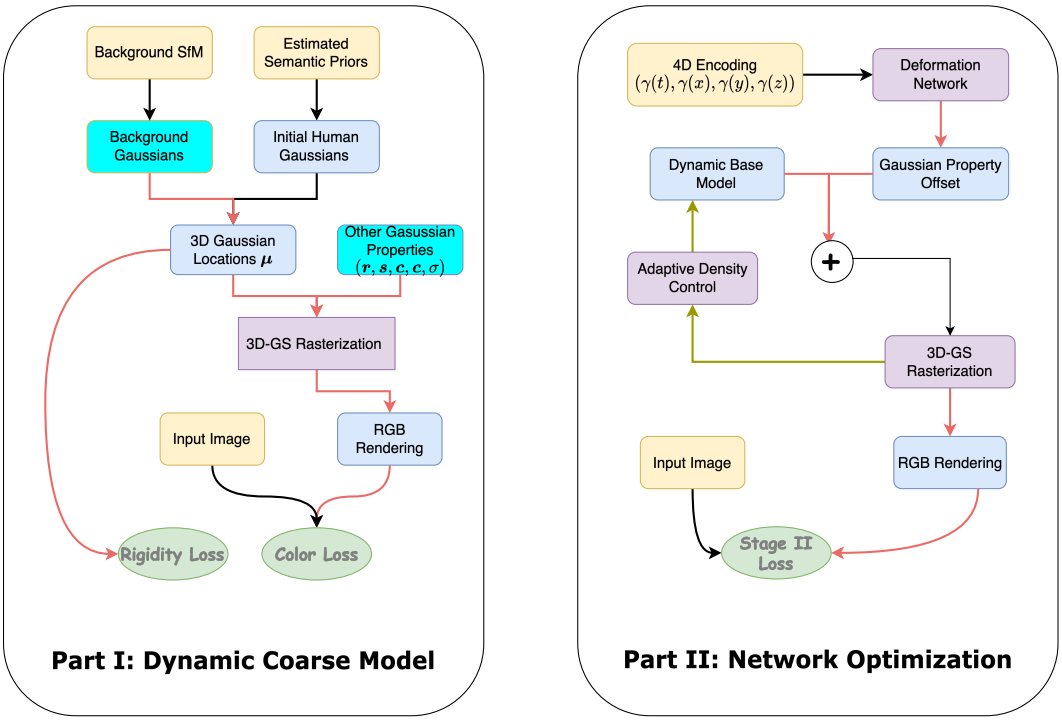

Figure 1: **Overview of our method.** Overview of the proposed method consisting of two parts. In part I, we learn a dynamic coarse 3D Gaussian model consisting of static background points and dynamic human points. In part II, we fit a deformable neural network that calculates the offsets of Gaussian location $\delta\boldsymbol{\mu}$, rotation $\delta\boldsymbol{r}$ and scales $\delta\boldsymbol{s}$ from the dynamic coarse model established in part I.

the deformation network is flexible enough to accommodate dynamic background elements. The separation between background and human Gaussians is designed to capture the highly deformable nature of human motion while maintaining a stable foundation for less dynamic scene components.

Let $\tilde{\mathbf{I}}_t$ represent the rendered image, which is rasterized from the Gaussians $\mathcal{P}_b \cup \mathcal{P}_h^{(t)}$. To refine these Gaussians, we adjust their properties frame-by-frame based on color loss. Specifically, we optimize the loss $\mathcal{L}^{(t)} = |\mathbf{I}_t - \tilde{\mathbf{I}}_t|_1$ by tuning Gaussian parameters such as intensity and variance, while keeping the spatial locations of the points fixed during this step.

Alignment. At this stage, the Gaussians across different frames are not yet fully aligned, as their locations have been frozen during the initial optimization. To address this, we perform a joint optimization of the Gaussians across all frames using the following objective:

$$\mathcal{L} = \sum_{t=1}^{N} \left(\mathcal{L}_{\text{color}}^{(t)} + \lambda_1 \mathcal{L}_{\text{rigid}}^{(t)} \right), \tag{4}$$

where two types of losses are involved. The color loss is simply defined by $\mathcal{L}_{\text{color}}^{(t)} = \|\mathbf{I}_t - \tilde{\mathbf{I}}_t\|_1$. The another loss, rigitidy loss, is introduced to ensure the human shape is consistent across frames.

$$\mathcal{L}_{\text{rigid}}^{(t)} := \sum_{\boldsymbol{\mu}_1 \in \mathcal{P}_h^{(t)}, \boldsymbol{\mu}_2 \in \mathcal{N}(\boldsymbol{\mu})} \left| \|\boldsymbol{\mu}_1^{(t)} - \boldsymbol{\mu}_2{}^{(t)}\|_2^2 - \|\boldsymbol{\mu}_1^{(t+1)} - \boldsymbol{\mu}_2^{(t+1)}\|_2^2 \right| \tag{5}$$

for $t \in \{1, \ldots, N-1\}$, where $\mathcal{N}(\mu)$ collects n_{neigh} nearest neighborhood points of point $\boldsymbol{\mu}$.

This strategy ensures a Gaussian initialization with high-fidelity geometric, setting a solid foundation before progressing to the learning phase of the neural deformation network.

4.2 STAGE II: DEFORMATION NETWORK INITIALIZATION

Denote by $\boldsymbol{g}_i = (\boldsymbol{\mu}_i, \boldsymbol{r}_i, \boldsymbol{s}_i)$. Utilizing the deformation modeling as outlined in Yang et al. (2023), our method processes the time t and the center location $\boldsymbol{\mu}_i$ of 3D Gaussians to generate offsets through the deformation MLP $\mathcal{F}_\Theta$:

$$\delta \boldsymbol{g}_i = (\delta \boldsymbol{\mu}_i, \delta \boldsymbol{r}_i, \delta \boldsymbol{s}_i) = \mathcal{F}_\Theta(\gamma(\boldsymbol{x}), \gamma(t)), \tag{6}$$

where $\gamma(x) := (\sin(2^k \pi x), \cos(2^k \pi x))_{k=0}^{L-1}$ provides the positional encoding. To be concise we directly write $\mathcal{F}_\Theta(\boldsymbol{x}, t)$

We fit $\mathcal{F}_\Theta$ in terms of $\mathcal{L} := \sum_t \mathcal{L}^{(t)}$ where

$$\mathcal{L}^{(t)} := \sum_{\boldsymbol{\mu}_i \in \mathcal{P}_h^{(t)} \cap \mathcal{P}_h^{(t)}} \left\| \boldsymbol{g}_i^{(0)} + \mathcal{F}_\Theta(\boldsymbol{\mu}_i^{(0)}, t) - \boldsymbol{g}_i^t \right\|. \tag{7}$$

In Eq. 7, we fit the deformation network by considering all available Gaussians at timestamp t. It is worth noting that not all Gaussians at timestamp t would appear at the initial frame. For those Gaussians $\boldsymbol{g}_i^{(0)}$ is just referring to the base properties, instead of an explicit Gaussian.

4.3 STAGE III: DEFORMATION NETWORK OPTIMIZATION

In Stage III, the primary objective is to optimize the deformation network by minimizing the discrepancy between the rendered images and the ground-truth images from the training dataset. This involves refining the alignment between the predicted scene geometry and the actual observations, ensuring more accurate and visually consistent reconstructions.

Specifically, we adjust the 3D Gaussians of the base model at each timestep by applying offsets predicted by the deformation network. These offsets capture the dynamic changes in the scene, particularly for non-rigid objects such as humans. Both the deformation network and the 3D Gaussians are optimized simultaneously to reduce error, leveraging a combination of L1 loss and D-SSIM loss together with the depth prior regularization:

$$\mathcal{L} = \mathcal{L}_{\text{color}} + \lambda_2 \mathcal{L}_{\text{D-SSIM}} + \lambda_3 \mathcal{L}_{\text{depth}}, \tag{8}$$

with $\lambda_2 = 0.2$ consistently applied in all experiments.

5 EXPERIMENTS

	Seattle			**Citron**			**Parking**		
	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
4DGS	24.03	0.89	161.1	22.84	0.88	143.2	23.14	0.84	156.4
Deformable-GS	24.39	0.88	169.4	23.99	0.87	148.2	24.53	0.87	163.9
HUGS	25.94	0.85	130	25.54	0.86	150	26.86	0.85	220
GART + Background	29.14	0.95	27.3	28.72	0.94	32.5	28.14	0.93	29.4
Ours	30.21	0.96	23.3	29.95	0.96	23.1	30.79	0.95	29.7
	Bike			**Jogging**			**Lab**		
4DGS	24.55	0.88	182.3	24.24	0.87	153.2	25.13	0.90	134.5
Deformable-GS	24.97	0.88	172.9	23.97	0.86	191.2	25.52	0.91	129.9
HUGS	25.46	0.84	130	23.75	0.78	220	26.00	0.92	90
GART + Background	28.75	0.94	19.5	28.49	0.92	22.1	29.01	0.95	20.8
Ours	30.47	0.97	18.4	30.03	0.94	21.3	30.82	0.97	0.12

Table 1: Comparison of ours method with previous work on test images of the NeuMan dataset Jiang et al. (2022a) using PSNR, SSIM and 1000x LPIPS metrics.

In this section, we assess our approach using real monocular datasets and conduct ablation studies to showcase its capability in reconstructing photo-realistic dynamic scenes featuring dramatic character action.

Datasets. First, we assess our method on the widely utilized ZJU dataset Peng et al. (2021b) under monocular setting and demonstrate that it achieves state-of-the-art performance on objects suited

	377			386			387		
	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
4DGS	27.18	0.84	97.2	26.85	0.55	152.1	28.18	0.84	86.4
Deformable-GS	28.49	0.86	97.7	27.84	0.61	131.3	27.81	0.65	105.1
HUGS*	30.80	0.98	20	34.11	0.98	20	29.97	0.97	30
GART	31.90	0.97	18.8	33.50	0.97	29.9	27.74	0.95	40.3
Ours	32.18	0.97	18.9	33.51	0.97	29.8	28.04	0.95	39.8
	392			**393**			**394**		
4DGS	28.14	0.86	131.4	27.18	0.87	121.5	26.23	0.84	98.1
Deformable-GS	27.28	0.85	91.9	27.91	0.84	131.7	27.59	0.89	105.8
HUGS*	31.36	0.97	30	29.80	0.97	30	30.54	0.97	30
GART	31.92	0.96	32.6	29.34	0.95	37.9	31.08	0.96	31.5
Ours	32.04	0.97	29.7	29.97	0.96	35.8	31.31	0.96	31.0

Table 2: Comparison of ours method with previous work on test images of the ZJU Mocap dataset Peng et al. (2021b) using PSNR, SSIM and 1000x LPIPS metrics. The evaluation results of HUGS are quoted directly from Kocabas et al. (2023), and suffer from low precision for LPIPS.

	Seattle			Citron			Parking		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Stage I	16.43	0.43	0.40	16.81	0.41	0.56	16.04	0.38	0.62
w/o $\mathcal{L}_{rigid}$	30.05	0.96	0.15	29.84	0.95	0.15	30.41	0.95	0.16
w/o DBM	24.91	0.82	0.19	24.31	0.88	0.21	24.58	0.84	0.16
Complete	30.21	0.96	0.14	29.95	0.96	0.14	30.79	0.95	0.15
	Bike			**Jogging**			**Lab**		
w/o Stage I	17.41	0.56	0.60	14.32	0.62	0.65	21.56	0.69	0.50
w/o $\mathcal{L}_{rigid}$	23.99	0.78	0.26	24.63	0.81	0.26	25.43	0.80	0.31
w/o DBM	24.19	0.86	0.19	23.95	0.87	0.18	25.18	0.21	0.17
Complete	30.47	0.97	0.14	30.03	0.94	0.14	30.82	0.97	0.12

Table 3: **Ablation study.** The performance is evaluated over full images using PSNR, SSIM and LPIPS metrics.

to our model. Subsequently, we conduct quantitative, qualitative, and ablation studies on the more realistically captured Neuman dataset Jiang et al. (2022a) to show that our method effectively provides adequate regularization for this underconstrained scenario, characterized by a camera moving around a human in motion.

Comparison Methods. We evaluate our methods against two classes of approaches: neural deformation methods and avatar fitting methods. The former category includes 4D Gaussian Wu et al. (2023) and Deformable Gaussian Yang et al. (2023). The latter category comprises GART Lei et al. (2023) and HUGS Kocabas et al. (2023).

Implementation Details. For the optimization in stage I, we set $\lambda_1 = 0.9$. In stage II, we set the depth of the deformation network $n_d = 8$ and the dimension of the hidden layer $n_h = 256$. The dimension L of positional encoding are set to 10 for both location $\boldsymbol{x}$ and time t.

5.1 Results

Comparison on NeuMan dataset. Jiang et al. (2022a) Table 1 provides a comparative analysis of our method against existing methods on the NeuMan dataset. Notably, GART is not capable of modeling the environment other than the human. Hence we synthesize the complete novel view by combining the human rendering generated by GART and the background scene rendering provided by 4DGS in light of the human segmentation.

Our method consistently shows superior performance across all metrics and scenarios, highlighting its effectiveness in generating high-quality reconstructions from monocular recordings. In addition,

the significant performance advantage of GART over 4DGS and Deformable-GS clearly indicates the critical importance of the SMPL prior.

Figure 2: **Qualitative comparison on Seattle and Citron in NeuMan dataset.**
Left column: ground truth. Middle column: our method. Right column:4DGS Wu et al. (2023).

Figure 2 offers a qualitative comparison between our method and 4DGS, which lacks a human shape prior, on the Neuman dataset. Typically, 4DGS produces lower-quality renderings from novel viewpoints, and in some cases, it even fails to capture the human figure in the scene due to missing geometric information.

Comparison on ZJU Mocap dataset. Peng et al. (2021b) Table 2 compares the performance of our method with other state-of-the-art techniques, namely 4DGS, Deformable-GS, HUGS*, and GART, on test images from the ZJU Mocap dataset. The evaluation focuses on various scenarios identified by their dataset numbers: 377, 386, 387, 392, 393, and 394. We adopt the same camera view settings as Peng et al. (2021a) for training and testing. Our method consistently demonstrates superior performance in terms of PSNR and SSIM across all scenarios. In the context of the LPIPS metric, our method generally achieves competitive results, occasionally outperforming others.

5.2 Ablation Study

As outlined in Table 3, our ablation studies evaluated the impact of removing several key components on the reconstruction performance, using the real-world NeuMan dataset for assessment. The first component examined is the necessity of learning the coarse model. We experimented with initializing Stage II using only the mesh vertices derived from SMPL pose estimation. The second component analyzed is the importance of the rigidity loss. The third component involves the dynamic base model(DBM), where we explored initializing with a static point cloud directly from COLMAP in stage I. Table 3 demonstrates that each of these three components positively contributes to the model's performance.

5.3 Limitations

While our method can produce high-quality dynamic reconstructions from monocular videos featuring fast human motion, it faces several challenges and limitations. Firstly, our approach relies on the semantic priors provided by an external model. Moreover a set of accurate camera parameters have to be given and our reconstruction is highly sensitive to the errors of camera poses. Secondly, unlike most human reconstruction work employing an explicit human template, our method cannot accom-

modate renderings with novel human poses. Lastly, designing a more complex model is necessary to effectively handle scenarios involving multiple people, particularly in cases with occlusions.

6 CONCLUSION

We introduce *ShapeGaussian*, a high-quality view synthesis approach for human with motion captured using a monocular camera. Our method utilizes a two-stage optimization process that first establishes a dynamic coarse 3D Gaussian model from estimated pose parameters. This provides robust shape regularization, enabling consistent synthesis of new views from sparse observations. We have demonstrated that, unlike previous efforts in monocular reconstruction, our approach can produce consistent reconstruction even in demanding scenarios.

REFERENCES

Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim. HyperReel: High-Fidelity 6-DoF Video with Ray-Conditioned Sampling, May 2023. URL http://arxiv.org/abs/2301.02238. arXiv:2301.02238 [cs].

Ang Cao and Justin Johnson. HexPlane: A Fast Representation for Dynamic Scenes, March 2023. URL http://arxiv.org/abs/2301.09632. arXiv:2301.09632 [cs].

Jaesung Choe, Christopher Choy, Jaesik Park, In So Kweon, and Anima Anandkumar. Spacetime Surface Regularization for Neural Dynamic Scene Reconstruction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17825–17835, Paris, France, October 2023. IEEE. ISBN 9798350307184. doi: 10.1109/ICCV51070.2023.01638. URL https://ieeexplore.ieee.org/document/10376928/.

Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi. Fusion4D: real-time performance capture of challenging scenes. *ACM Transactions on Graphics*, 35(4):114:1–114:13, July 2016. ISSN 0730-0301. doi: 10.1145/2897824.2925969. URL https://dl.acm.org/doi/10.1145/2897824.2925969.

Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. Motion2fusion: real-time volumetric performance capture. *ACM Transactions on Graphics*, 36(6):246:1–246:16, November 2017. ISSN 0730-0301. doi: 10.1145/3130800.3130801. URL https://dl.acm.org/doi/10.1145/3130800.3130801.

Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural Radiance Flow for 4D View Synthesis and Video Processing, September 2021. URL http://arxiv.org/abs/2012.09790. arXiv:2012.09790 [cs].

Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast Dynamic Radiance Fields with Time-Aware Neural Voxels. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, November 2022. doi: 10.1145/3550469.3555383. URL http://arxiv.org/abs/2205.15285. arXiv:2205.15285 [cs].

Sara Fridovich-Keil, Giacomo Meanti, Frederik Warburg, Benjamin Recht, and Angjoo Kanazawa. K-Planes: Explicit Radiance Fields in Space, Time, and Appearance, March 2023. URL http://arxiv.org/abs/2301.10241. arXiv:2301.10241 [cs].

Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic View Synthesis from Dynamic Monocular Video, May 2021. URL http://arxiv.org/abs/2105.06468. arXiv:2105.06468 [cs].

Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-PIFu: Geometry and Pixel Aligned Implicit Functions for Single-view Human Reconstruction, December 2020. URL http://arxiv.org/abs/2006.08072. arXiv:2006.08072 [cs].

Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. GaussianAvatar: Towards Realistic Human Avatar Modeling from a Single Video via Animatable 3D Gaussians, December 2023. URL http://arxiv.org/abs/2312.02134. arXiv:2312.02134 [cs].

Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and Andrew Fitzgibbon. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, pp. 559–568, New York, NY, USA, October 2011. Association for Computing Machinery. ISBN 978-1-4503-0716-1. doi: 10.1145/2047196.2047270. URL https://dl.acm.org/doi/10.1145/2047196.2047270.

Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. NeuMan: Neural Human Radiance Field from a Single Video, September 2022a. URL http://arxiv.org/abs/2203.12575. arXiv:2203.12575 [cs].

Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. SDFDiff: Differentiable Rendering of Signed Distance Fields for 3D Shape Optimization, February 2022b. URL http://arxiv.org/abs/1912.07109. arXiv:1912.07109 [cs].

HyunJun Jung, Nikolas Brasch, Jifei Song, Eduardo Perez-Pellitero, Yiren Zhou, Zhihao Li, Nassir Navab, and Benjamin Busam. Deformable 3D Gaussian Splatting for Animatable Human Avatars, December 2023. URL http://arxiv.org/abs/2312.15059. arXiv:2312.15059 [cs].

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering, August 2023. URL http://arxiv.org/abs/2308.04079. arXiv:2308.04079 [cs].

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for Human Vision Models, 2024. URL https://arxiv.org/abs/2408.12569. _eprint: 2408.12569.

Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. HUGS: Human Gaussian Splats, November 2023. URL http://arxiv.org/abs/2311.17910. arXiv:2311.17910 [cs].

Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural Human Performer: Learning Generalizable Radiance Fields for Human Performance Rendering, September 2021. URL http://arxiv.org/abs/2109.07448. arXiv:2109.07448 [cs].

Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. GART: Gaussian Articulated Template Models, November 2023. URL http://arxiv.org/abs/2311.16099. arXiv:2311.16099 [cs].

Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, and Ping Tan. Streaming Radiance Fields for 3D Video Synthesis, October 2022a. URL http://arxiv.org/abs/2210.14831. arXiv:2210.14831 [cs].

Mingwei Li, Jiachen Tao, Zongxin Yang, and Yi Yang. Human101: Training 100+FPS Human Gaussians in 100s from 1 View, December 2023a. URL http://arxiv.org/abs/2312.15258. arXiv:2312.15258 [cs].

Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, and Zhaoyang Lv. Neural 3D Video Synthesis from Multi-view Video, May 2022b. URL http://arxiv.org/abs/2103.02597. arXiv:2103.02597 [cs].

Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes, April 2021. URL http://arxiv.org/abs/2011.13084. arXiv:2011.13084 [cs].

Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. DynIBaR: Neural Dynamic Image-Based Rendering, April 2023b. URL http://arxiv.org/abs/2211.11082. arXiv:2211.11082 [cs].

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, October 2023a. URL http://arxiv.org/abs/2310.08585. arXiv:2310.08585 [cs].

Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-Flow: 4D Reconstruction with Dynamic 3D Gaussian Particle, December 2023b. URL http://arxiv.org/abs/2312.03431. arXiv:2312.03431 [cs].

Jia-Wei Liu, Yan-Pei Cao, Weijia Mao, Wenqiao Zhang, David Junhao Zhang, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. DeVRF: Fast Deformable Voxel Radiance Fields for Dynamic Scenes, June 2022. URL http://arxiv.org/abs/2205.15723. arXiv:2205.15723 [cs].

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):1–16, November 2015. ISSN 0730-0301, 1557-7368. doi: 10.1145/2816795.2818013. URL https://dl.acm.org/doi/10.1145/2816795.2818013.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis, August 2023. URL http://arxiv.org/abs/2308.09713. arXiv:2308.09713 [cs].

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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6468–6477, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.00650. URL https://ieeexplore.ieee.org/document/9157608/.

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, August 2020. URL http://arxiv.org/abs/2003.08934. arXiv:2003.08934 [cs].

Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 343–352, Boston, MA, USA, June 2015. IEEE. ISBN 978-1-4673-6964-0. doi: 10.1109/CVPR.2015.7298631. URL http://ieeexplore.ieee.org/document/7298631/.

Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable Neural Radiance Fields, September 2021a. URL http://arxiv.org/abs/2011.12948. arXiv:2011.12948 [cs].

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, September 2021b. URL http://arxiv.org/abs/2106.13228. arXiv:2106.13228 [cs].

Sungheon Park, Minjung Son, Seokhwan Jang, Young Chun Ahn, Ji-Yeon Kim, and Nahyup Kang. Temporal Interpolation Is All You Need for Dynamic Neural Radiance Fields, March 2023. URL http://arxiv.org/abs/2302.09311. arXiv:2302.09311 [cs].

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies, October 2021a. URL http://arxiv.org/abs/2105.02872. arXiv:2105.02872 [cs].

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, March 2021b. URL http://arxiv.org/abs/2012.15838. arXiv:2012.15838 [cs].

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes, November 2020. URL http://arxiv.org/abs/2011.13961. arXiv:2011.13961 [cs].

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians, December 2023a. URL http://arxiv.org/abs/2312.02069. arXiv:2312.02069 [cs].

Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3DGS-Avatar: Animatable Avatars via Deformable 3D Gaussian Splatting, December 2023b. URL http://arxiv.org/abs/2312.09228. arXiv:2312.09228 [cs].

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization, December 2019. URL http://arxiv.org/abs/1905.05172. arXiv:1905.05172 [cs].

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization, April 2020. URL http://arxiv.org/abs/2004.00452. arXiv:2004.00452 [cs].

Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable Gaussian Codec Avatars, December 2023. URL http://arxiv.org/abs/2312.03704. arXiv:2312.03704 [cs].

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4D : Efficient Neural 4D Decomposition for High-fidelity Dynamic Reconstruction and Rendering, April 2023. URL http://arxiv.org/abs/2211.11610. arXiv:2211.11610 [cs].

Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. NeRFPlayer: A Streamable Dynamic Scene Representation with Decomposed Neural Radiance Fields, February 2023a. URL http://arxiv.org/abs/2210.15947. arXiv:2210.15947 [cs].

Liangchen Song, Xuan Gong, Benjamin Planche, Meng Zheng, David Doermann, Junsong Yuan, Terrence Chen, and Ziyan Wu. PREF: Predictability Regularized Neural Motion Fields, April 2023b. URL http://arxiv.org/abs/2209.10691. arXiv:2209.10691 [cs].

Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Dynamic Scene From Monocular Video, August 2021. URL http://arxiv.org/abs/2012.12247. arXiv:2012.12247 [cs].

Edith Tretschk, Vladislav Golyanik, Michael Zollhoefer, Aljaz Bozic, Christoph Lassner, and Christian Theobalt. SceNeRFlow: Time-Consistent Reconstruction of General Dynamic Scenes, August 2023. URL http://arxiv.org/abs/2308.08258. arXiv:2308.08258 [cs].

Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric Inference of 3D Human Body Shapes, August 2018. URL http://arxiv.org/abs/1804.04875. arXiv:1804.04875 [cs].

Chaoyang Wang, Lachlan Ewen MacDonald, Laszlo A. Jeni, and Simon Lucey. Flow supervision for Deformable NeRF, March 2023a. URL http://arxiv.org/abs/2303.16333. arXiv:2303.16333 [cs].

Feng Wang, Sinan Tan, Xinghang Li, Zeyue Tian, Yafei Song, and Huaping Liu. Mixed Neural Voxels for Fast Multi-view Video Synthesis, October 2023b. URL http://arxiv.org/abs/2212.00190. arXiv:2212.00190 [cs].

Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zhang, Minye Wu, Lan Xu, and Jingyi Yu. Fourier PlenOctrees for Dynamic Radiance Field Rendering in Real-time, February 2022a. URL http://arxiv.org/abs/2202.08614. arXiv:2202.08614 [cs].

Liao Wang, Qiang Hu, Qihan He, Ziyu Wang, Jingyi Yu, Tinne Tuytelaars, Lan Xu, and Minye Wu. Neural Residual Radiance Fields for Streamably Free-Viewpoint Videos, June 2023c. URL http://arxiv.org/abs/2304.04452. arXiv:2304.04452 [cs].

Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. NeuS2: Fast Learning of Neural Implicit Surfaces for Multi-view Reconstruction, December 2022b. URL http://arxiv.org/abs/2212.05231. arXiv:2212.05231 [cs].

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, June 2022. URL http://arxiv.org/abs/2201.04127. arXiv:2201.04127 [cs].

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering, December 2023. URL http://arxiv.org/abs/2310.08528. arXiv:2310.08528 [cs].

Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time Neural Irradiance Fields for Free-Viewpoint Video, June 2021. URL http://arxiv.org/abs/2011.12950. arXiv:2011.12950 [cs].

Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. MonoClothCap: Towards Temporally Coherent Clothing Capture from Monocular RGB Video, November 2020. URL http://arxiv.org/abs/2009.10711. arXiv:2009.10711 [cs].

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals, March 2022. URL http://arxiv.org/abs/2112.09127. arXiv:2112.09127 [cs].

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration, March 2023. URL http://arxiv.org/abs/2212.07422. arXiv:2212.07422 [cs].

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction, November 2023. URL http://arxiv.org/abs/2309.13101. arXiv:2309.13101 [cs].

Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. BodyFusion: Real-Time Capture of Human Motion and Surface Geometry Using a Single Depth Camera. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 910–919, Venice, October 2017. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.104. URL http://ieeexplore.ieee.org/document/8237366/.

Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. DoubleFusion: Real-time Capture of Human Performances with Inner Body Shapes from a Single Depth Sensor, April 2018. URL http://arxiv.org/abs/1804.06023. arXiv:1804.06023 [cs].

Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, and Umar Iqbal. GAvatar: Animatable 3D Gaussian Avatars with Implicit Mesh Learning, March 2024. URL http://arxiv.org/abs/2312.11461. arXiv:2312.11461 [cs].

Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3D Gaussian Avatars, November 2023. URL http://arxiv.org/abs/2311.08581. arXiv:2311.08581 [cs].

A APPENDIX

You may include other additional sections here.