HiFi4G: High-Fidelity Human Performance Rendering via Compact Gaussian Splatting

Yuheng Jiang[1,2] Zhehao Shen[1] Penghao Wang[1] Zhuo Su[3] Yu Hong[1]
Yingliang Zhang[4] Jingyi Yu[1] Lan Xu[1]

[1]ShanghaiTech University [2]NeuDim [3]ByteDance [4]DGene

Figure 1. **High-fidelity rendering with our compact Gaussian Splatting**. From multi-view human performance video, HiFi4G marries the traditional non-rigid fusion with differentiable rasterization advance to efficiently produce compact 4D assets.

Abstract

We have recently seen tremendous progress in photo-real human modeling and rendering. Yet, efficiently rendering realistic human performance and integrating it into the rasterization pipeline remains challenging. In this paper, we present HiFi4G, an explicit and compact Gaussian-based approach for high-fidelity human performance rendering from dense footage. Our core intuition is to marry the 3D Gaussian representation with non-rigid tracking, achieving a compact and compression-friendly representation. We first propose a dual-graph mechanism to obtain motion priors, with a coarse deformation graph for effective initialization and a fine-grained Gaussian graph to enforce subsequent constraints. Then, we utilize a 4D Gaussian optimization scheme with adaptive spatial-temporal regularizers to effectively balance the non-rigid prior and Gaussian updating. We also present a companion compression scheme with residual compensation for immersive experiences on various platforms. It achieves a substantial compression rate of approximately 25 times, with less than 2MB of storage per frame. Extensive experiments demonstrate the effectiveness of our approach, which significantly outperforms existing approaches in terms of optimization speed, rendering quality, and storage overhead. Project page: https://nowheretrix.github.io/HiFi4G/.

1. Introduction

Volumetric recording and realistic rendering of 4D (space-time) human performance diminish the boundaries between viewers and performers. It brings numerous immersive experiences like telepresence or tele-education in VR/AR.

Early solutions [8–10, 25] reconstruct textured meshes from captured videos by explicitly leveraging non-rigid registration [44, 62]. Yet, they remain vulnerable to occlusions and lack of textures which cause holes and noise in the reconstruction results. Recent neural advances, represented by NeRF [41], bypass explicit reconstruction and instead optimize a coordinate-based multi-layer perceptron (MLP) to conduct volume rendering at photo-realism. Some dynamic variants [13, 46–48, 66, 68] of NeRF attempt to maintain a canonical feature space to reproduce features in each live frame with an extra implicit deformation field. However, such a canonical design is fragile to large mo-

tions or topology changes. Recent approaches [22, 59, 69] remove the deformation fields and compactly represent the 4D feature grid through planar factorization [7, 12] or Hash-encoding [42]. They notably accelerate both the training and rendering speed for interactive applications but the challenges of runtime memory and storage still exist. The recent 3D Gaussian Splatting (3DGS) [26] marks a significant return to an explicit paradigm for static scene representation. Based on GPU-friendly rasterization of 3D Gaussian primitives, it allows real-time and high-quality radiance field rendering unseen before. Various concurrent works [40, 71, 76, 77] adapt 3DGS for dynamic scenes. Some [40] focus on extracting the non-rigid motions from dynamic Gaussians yet sacrificing the rendering quality. Others [71, 76] adopt extra implicit deformation fields to compensate for the motion information, and hence fall short of handling long-duration motions and lose the explicit and GPU-friendly beauty of the original 3DGS.

In this paper, we present *HiFi4G* – a totally explicit and compact Gaussian-based approach for high-fidelity, real-time playback of human performance from dense footage (see Fig. 1). Our key idea is to marry the 3D Gaussian representation [26] with non-rigid tracking [44, 62], so as to explicitly disentangle motion and appearance information for a compact and compression-friendly representation. HiFi4G significantly outperforms existing implicit rendering approaches, in terms of optimization speed, rendering quality, and storage overhead. Our explicit representation also enables seamlessly integrating our results into the GPU-based rasterization pipeline., i.e., immersively watching high-fidelity human performances with VR headsets.

To organically bridge the Gaussian representation with non-rigid tracking, we first introduce a dual-graph mechanism, which consists of a coarse deformation graph and a fine-grained Gaussian graph. For the former, we obtain per-frame geometry proxy via the NeuS2 [69] and then employ embedded deformation (ED) [62] in a key-frame manner. Such an explicit tracking process splits the sequence into segments and provides rich motion prior within each segment. Analogous to the key-volume update [9], we follow 3DGS to prune the incorrect Gaussians from the previous segment and update new ones to restrict the number of Gaussians in the current segment. Then, we build a fine-grained Gaussian graph and interpolate the motion of each Gaussian from the coarse ED graph for subsequent initialization. Naïvely warping the Gaussian graph with the ED graph and splatting it onto screen space will cause severe unnatural artifacts, while continuous optimization without any constraints leads to jittery artifacts. Thus, we propose a 4D Gaussian optimization scheme to carefully balance the non-rigid motion prior and the updating of Gaussian attributes. We adopt a temporal regularizer to enforce the appearance attributes of each Gaussian, i.e., spherical harmonic (SH), opacity, and scaling coefficients, to be consistent. We also propose a smooth term for the motion attributes (position and rotation) to produce locally as-rigid-as-possible motions between the adjacent Gaussians. These regularizers are further enhanced with an adaptive weighting mechanism to penalize the flicking artifacts on the regions with slight non-rigid motions. Once optimized, we obtain spatial-temporally compact 4D Gaussians. To make our HiFi4G practical for users, we demonstrate a companion compression scheme that follows standard residual compensation, quantization, and entropy encoding for the Gaussian parameters. It achieves a substantial compression rate of approximately 25 times and requires less than 2 MB storage per frame, enabling immersively viewing human performances on various platforms like VR headsets.

To summarize, our main contributions include:

- We present a compact 4D Gaussian representation for human performance rendering, which bridges Gaussian Splatting and non-rigid tracking.
- We propose a dual-graph mechanism with various regularization designs to effectively recover spatial-temporally consistent 4D Gaussians.
- We showcase a companion compression scheme, supporting immersive experience of human performance with low storage, even under various platforms.

2. Related Work

Human Performance Capture. Recently, human performance capture [1, 2, 16, 17, 20, 29, 30, 45, 50, 58, 67, 72] has been widely investigated to achieve detailed registration for various applications. Zollhöfer *et al*. [83] capture the rigid template first but DynamicFusion [44] removes this explicit template prior and enables real-time performance which benefits from the GPU solvers. Guo *et al*. [14] model the geometry, surface albedo, and appearance on the reference volume. Fusion4d [9] and Motion2fusion [10] rely on a key-frame-based strategy to handle topological changes. Based on the human parametric model [39], DoubleFusion [78] proposes a two-layer representation for more robust scene capture, while Xu *et al*. [75] extend it to sparse view setup. Su *et al*. [60, 61] further address the challenging motions and human-object interaction scenarios. Additionally, several studies [23, 32, 33, 79] combine explicit volumetric fusion and implicit modeling to capture more dynamic details. Nevertheless, these methods primarily focus on detailed geometry rather than high-quality texture.

Neural Human Modeling. In the domain of digital human neural representation, various approaches [18, 19, 27, 35, 38, 63, 64, 73, 80] have been proposed to address this challenge. Non-rigid NeRF [66] utilizes a displacement field to represent the motion, while Neuralbody [49] uses latent codes anchored to SMPL [39] vertices. Humannerfs [70, 82] combine the SMPL with a deformation net. TAVA [28]

and X-avatar [57] learn the skinning weight through root-finding. NDR [3] defines a bijective function that satisfies the cycle consistency. With recent advancements in Instant-NGP [42], some works [22, 24, 35, 59, 69] demonstrate the efficient training and rendering speed. However, most methods produce blurriness, particularly in high-frequency regions. The recent 3DGS [26] marks a significant return to an explicit paradigm for high-performance static scene representation. However, per-frame 3DGS disregards temporal consistency, resulting in visual jitteriness. Some concurrent studies [40, 71, 76, 77] adapt 3DGS for dynamic scenes. Yet, these methods typically offer real-time performance only at low resolution and are not equipped to handle large motions. In contrast, HiFi4G leverages dual-graph to generate compact 4D Gaussians, enabling high-fidelity real-time rendering with challenging motions.

Compact Representation. Compact representation plays a pivotal role in dynamic rendering, engaging the interest of numerous researchers. A series of works are proposed for early point cloud compression with Octree [54, 65], Wavelet [43]. These are formalized into MPEG-PCC [55] standards by the Moving Picture Experts Group (MPEG), which are categorized into video-based (VPCC) and geometry-based (GPCC). Following, learning-based methods [34, 51, 52] emerge, focusing on enhancing efficiency. For neural fields, several studies introduce compact neural representations through tensor [7] and scene [59] decomposition, tri-planes [21, 53] and multi-planes [4, 12, 56]. Instant-NSR [81] leverages the tracked mesh and texture video while HumanRF [22] employs temporal matrix-vector decomposition. Despite their advancements, these methods often compromise rendering quality and speed to minimize storage requirements. Comparably, HiFi4G only requires less than 2 MB storage per frame to enable high-quality rendering results.

3. Method

Given human performance videos captured by multi-view RGB cameras, HiFi4G integrates recent advancements in differentiable rasterization with traditional non-rigid tracking, significantly outperforming existing rendering approaches [22, 36, 69, 81] in terms of optimization speed, rendering quality, and storage overhead. The methodology is visually summarized in Fig. 2. Our approach starts with a dual graph mechanism, which consists of a coarse deformation graph and a fine-grained Gaussian graph, detailed in Sec 3.1. Subsequently, this representation is employed along with corresponding temporal and smooth regularization, leading to the generation of spatial-temporally compact 4D Gaussians in Sec 3.2. In addition, we introduce a companion compression scheme in Sec 3.3. This allows for immersive viewing of high-fidelity human performances with a storage requirement of less than 2 MB per frame.

3.1. Dual Graph Mechanism

We employ a dual graph structure to explicitly disentangle motion and appearance, resulting in a compact and compression-friendly representation. This design facilitates expedited convergence and enhances visual quality.

Coarse Deformation Graph. Instead of using an additional implicit deformation network [71, 76] to handle non-rigid motion, which could potentially affect the high performance and GPU-friendliness of the original 3DGS, we opt for the Embedded Deformation [62] to establish model-to-model correspondences by leveraging conventional non-rigid deformation techniques [15, 44, 74]. To achieve this, we first generate per-frame geometry proxies using NeuS2 [69]. We then apply non-rigid tracking to the resulting mesh sequences following a key-frame manner. Specifically, we parameterize the dynamic motions as an ED graph $W = \{dq_i, x_i\}$, where x_i represents the coordinates of sampled ED nodes in keyframe space, and dq_i denotes the dual quaternions representing the corresponding rigid transformation in $SE(3)$ space. Subsequently, we acquire each point v_c using Dual-Quaternion Blending:

$$DQB(v_c) = \sum_{i \in \mathcal{N}(v_c)} w(x_i, v_c) dq_i, \tag{1}$$

where $\mathcal{N}(v_c)$ is a set of neighboring ED nodes of v_c and $w(x_i, v_c)$ denotes the influence weight of the ith node x_i on v_c. At frame t, we identify correspondence points between the warped key mesh and the current mesh. Subsequently, we optimize the motion by constructing the terms:

$$E = \lambda_{\text{data}} E_{\text{data}} + \lambda_{\text{reg}} E_{\text{reg}}, \tag{2}$$

where E_{data} and E_{reg} represent the energies associated with the data term and the regularization term, respectively. Please refer to Dynamicfusion [44] for more details. To explicitly handle coarse topological changes and reduce severe misalignment issues, we implement the key-volume strategy as described in Fusion4d [9]. This strategy involves segmenting the sequence into multiple key volumes.

Fine-grained Gaussian Graph. To bypass the tedious process of creating 3D Gaussians from Structure-from-Motion (SfM) points for each frame, we utilize a more efficient initialization method. For the first frame, we construct the 3D Gaussians from the NeuS2 mesh using an importance sampling strategy. We increase the sampling density in the hand and face regions to significantly improve visual quality. For subsequent keyframes, analogous to the key-volume update strategy, we follow 3DGS to prune incorrect Gaussians from the previous keyframe and densify new ones at the current keyframe. We then restrict the number of Gaussians within the current segment. Afterward, we establish a fine-grained Gaussian graph, consisting of refined Gaussian kernels for subsequent constraints, determined by the

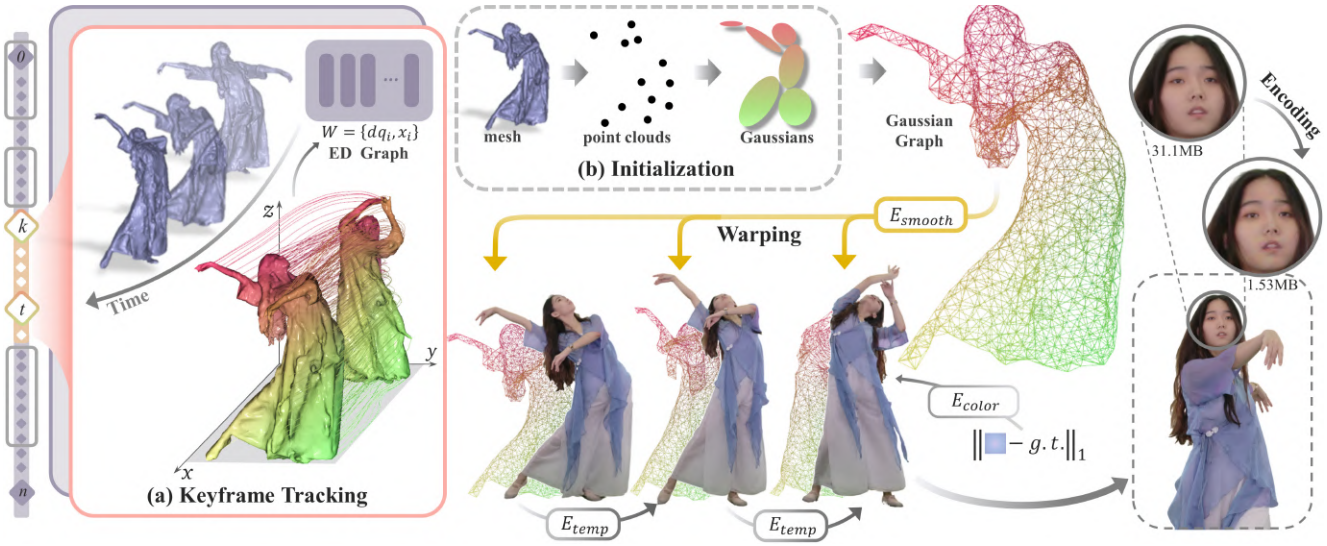

Figure 2. **Overview of HiFi4G.** (a) The non-rigid tracking establishes a **coarse deformation graph** and tracks the motions for Gaussian optimization. (b) HiFi4G initializes the first frame Gaussians from NeuS2 and constructs a **fine-grained Gaussian graph** to enhance temporal coherence. We then employ the ED graph to warp 4D Gaussians, applying both E_{smooth} and E_{temp} constraints to the Gaussian graph, which yields spatial-temporally compact and compression-friendly 4D Gaussians, thus facilitating efficient compression.

k-nearest neighbors (KNN, k = 16). In addition, for each Gaussian kernel in the fine-grained graph, we also find the KNN (k = 8) from the ED nodes, which assists in calculating the influence weight for motion interpolation. The initialization is still crucial for non-key frames to prevent falling into local optima during the back-propagation of differentiable rasterization. To this end, we warp the Gaussian graph from the keyframe to other frames within the segment according to the ED nodes' motion interpolation:

$$\begin{aligned} p'_{i,t} &= SE3(DQB(p_{i,k}))p_{i,k}, \\ q'_{i,t} &= ROT(DQB(p_{i,k}))q_{i,k}, \end{aligned} \tag{3}$$

$SE3(\cdot)$ converts dual quaternion back into a transformation matrix, while $ROT(\cdot)$ extracts the rotation component from dual quaternion. $p_{i,k}, q_{i,k}$ denote the position and rotation of the i-th Gaussian kernel at keyframe k, respectively. $p'_{i,t}$ and $q'_{i,t}$ represent the initial position and rotation at frame t. They will be further optimized in the subsequent stage.

3.2. 4D Gaussians Optimization

Directly warping the fine-grained Gaussian graph with tracking prior and splatting it onto screen space can lead to noticeable and unnatural artifacts. To mitigate this, we do not use the Gaussians' densification and pruning within the segment. Instead, we impose a constraint on their number and execute sequential optimization.

For the frame t, we categorize attributes for each 4D Gaussian kernel i into two groups: 1). Appearance-aware parameters, which include spherical harmonic $\mathcal{C}_{i,t}$, opacity $\sigma_{i,t}$, and scaling $s_{i,t}$. 2). Motion-aware parameters, which include position $p_{i,t}$ and rotation $q_{i,t}$. Leveraging the initialization from the warped Gaussian graph reduces the training time to one-third while still yielding vivid results. However, despite incorporating non-rigid tracking priors, we observe notable temporal jitters in the rendered results. Concurrent studies [40, 71, 76] address this issue by decoupling the deformation field from canonical 3D Gaussians. They employ a consistent set of Gaussians across dynamic sequences, which substantially diminishes view-dependent effects and sacrifices rendering quality. To mitigate temporal jitters while maintaining rendering quality, we introduce temporal and smooth regularization to delicately balance the dual graph prior and the updating of Gaussian attributes, thereby enforcing 4D consistency. First, we introduce the temporal regularization term E_{temp}. This term promotes coherent appearances by constraining the 4D Gaussian appearance attributes($\mathcal{C}_{i,t}$, $\sigma_{i,t}$, $s_{i,t}$) to be consistent with the previous frame:

$$E_{\text{temp}} = \sum_{a \in \{\mathcal{C}, \sigma, s\}} w_{i,t} \lambda_{\text{a}} \| a_{i,t} - a_{i,t-1} \|_2^2, \tag{4}$$

E_{temp} helps to reduce jitteriness. However, it may be not sufficient, especially when motion parameters change significantly, particularly in feature-less areas. Moreover, applying this regularization directly to motion attributes can also result in unnatural artifacts. Inspired by works [40, 44, 62] on non-rigid registration, we introduce a smooth term targeted at the motion attributes ($p_{i,t}$, $q_{i,t}$) within the fine-grained Gaussian graph. We define this term as follows:

$$\begin{aligned} E_{\text{smooth}} = \sum_i \sum_{j \in \mathcal{N}(i)} w_{i,t} \| SO3(q_{i,t} * q_{i,t-1}^{-1}) \\ (p_{j,t-1} - p_{i,t-1}) - (p_{j,t} - p_{i,t}) \|_2^2, \end{aligned} \tag{5}$$

$SO3(\cdot)$ converts a quaternion into a rotation matrix. Kernel i and j are neighbors on the Gaussian graph. The smooth term produces locally as-rigid-as-possible deformations to constrain the consistent 4D Gaussian motion on the spatial-temporal domain. Furthermore, it's observed that the Human Visual System is more sensitive to detail changes in static regions as opposed to dynamic ones [6]. Thus, we incorporate an adaptive weight that takes into account the displacement of positions between adjacent frames:

$$w_{i,t} = \exp(-\alpha \|p'_{i,t} - p'_{i,t-1}\|^2), \tag{6}$$

This adaptive weight indicates the degree of motion change in a corresponding local region. It penalizes the flicking artifacts in regions with slight non-rigid motions and reduces penalties in areas with large movements. This significantly improves the visual quality. Additionally, we employ the photometric loss during the training process:

$$E_{\text{color}} = \|\hat{\mathbf{C}} - \mathbf{C}\|_1, \tag{7}$$

$\hat{\mathbf{C}}$ is the blended color after rasterization and $\mathbf{C}$ is the ground truth. The complete energy is as follows:

$$E = \lambda_{\text{temp}} E_{\text{temp}} + \lambda_{\text{smooth}} E_{\text{smooth}} + \lambda_{\text{color}} E_{\text{color}}. \tag{8}$$

3.3. Compact 4D Gaussians

After optimization, we obtain spatial-temporally compact 4D Gaussians, resulting in high-fidelity rendering results. However, each frame requires the same amount of storage as the keyframe. This leads to significant memory consumption and presents challenges when handling lengthy sequences. To address this problem, we introduce a companion compression scheme on top of our compact 4D Gaussians. This scheme adheres to the traditional method of residual compensation, quantization, and entropy encoding, as depicted in Fig. 3.

Residual Compensation. In contrast to the broad distribution range of the original attributes, we opt to retain the keyframe attributes and calculate residuals for the following frames within the segment, effectively narrowing the range. In terms of appearance attributes($\mathcal{C}_{i,t}$, $\sigma_{i,t}$, $s_{i,t}$), the impact of E_{temp} results in minimal variations. Therefore, we can directly derive the residual appearance through subtraction. However, for position p and rotation q, simple subtraction is not sufficient as large motions still exist within a segment. To address this, we employ motion compensation as outlined in Eq. 1 and Eq. 3. We subtract the warped key Gaussians $p'_{i,k}, q'_{i,k}$ from $p_{i,t}, q_{i,t}$, ensuring a narrower range.

Quantization. We scale and round attribute values based on their range and quantization bits Q_{bit}, making the data ready for entropy encoding.

Entropy Encoding. Residual computation combined with motion compensation yields a residual distribution for attributes that cluster around zero. To leverage this distribution for real-time encoding and decoding, we apply the Ranged Arithmetic Numerical System (RANS) [11]. RANS enhances compression by taking advantage of the distribution 's skewness, a key factor for meeting the high-performance demands of HiFi4G. We compress our data by calculating the frequency of each quantized attribute and constructing a frequency distribution. This distribution helps to encode each attribute efficiently using the RANS algorithm, where each attribute and the current state of the encoder are processed to update the state, representing the encoded data sequence. The final state is stored as an integer stream for subsequent decoding. This compression scheme achieves a substantial compression rate of approximately 25 times, reducing the storage requirement to less than 2 MB per frame. This capability facilitates the immersive viewing of high-fidelity human performances on various platforms, including VR/AR HMDs.

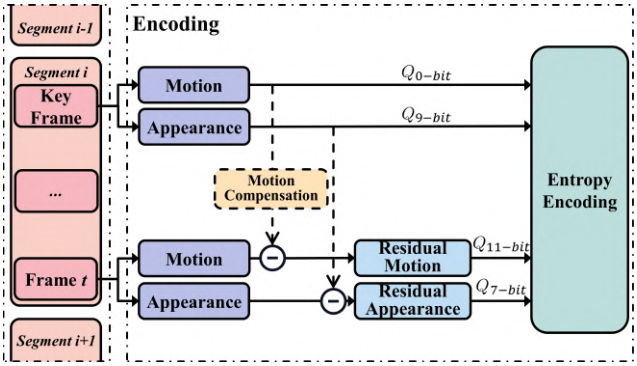

Figure 3. Illustration of compression strategy for 4D Gaussians.

4. Implementation Details

First, we use the background matting [37] to extract the foreground masks from all captured frames. For global initialization, we use openpose [5] to estimate the hand and face regions for importance sampling. The sampling ratio across the body, hands, and face regions is approximately 8:1:1. We perform 30000 training iterations with densification and pruning on the keyframes, followed by resetting the tracking and reconstructing the dual-graph. For non-key frames, training iterations are reduced to 9000. In the optimization stage, we use the following empirically determined parameters: $\alpha = 50, \lambda_{\mathcal{C}} = 1, \lambda_{\sigma} = 0.05, \lambda_{s} = 0.05, \lambda_{\text{smooth}} = 0.002, \lambda_{\text{temp}} = 0.0005, \lambda_{\text{color}} = 1.0$. During compression, we first quantize the appearance attributes, then fix these parameters and fine-tune motion p and q of 4D Gaussians over an additional 1000 iterations. Afterward, we quantize the motion. We apply different precision levels for various attributes to balance storage and quality. For the keyframes, we keep the motion uncompressed(0-bit) and apply 9-bit quantization for appearance. For non-key frames, we use 11-bit quantization for motion and 7-bit quantization for appearance due to their more compact range.

Figure 4. Gallery of our results. HiFi4G delivers real-time high-fidelity rendering of human performance across challenging motions, such as "playing instruments", "dancing" and "changing clothes".

5. Experimental Results

To demonstrate the capabilities of HiFi4G, we deploy 81 pre-calibrated Z-CAM cinema cameras to capture complex human performances with a resolution of 3840 × 2160 at 30 fps, and then evaluate our method. The dataset covers a variety of costumes, from traditional Chinese attire to casual clothes and cosplay. It also includes a wide range of activities such as dance, fitness, and interaction with various objects. As shown in Fig. 4, HiFi4G enables real-time, high-fidelity rendering of human performance in high resolution. It effectively handles complex motions like playing instruments, dancing, and changing clothes.

5.1. Comparison

We compare HiFi4G with the SOTA methods including Instant-NSR [81], NeuS2 [69], HumanRF [22] and concurrent work Dynamic 3D Gaussians [40] on our captured dataset and ActorsHQ [22]. As depicted in Fig. 5, Instant-NSR [81] suffers from severe artifacts due to the heavy reliance on geometry. Volume rendering methods such as NeuS2 [69] and HumanRF [22] produce blurry results, over-smoothing on high-frequency details. Meanwhile, Dynamic 3D Gaussians [40] loses the advantages of 3DGS [26] due to fixed appearance attributes, failing to recover detailed appearance and view-dependency. In contrast, HiFi4G surpasses these existing methods by merging 3D Gaussian representation with keyframe-update-based non-rigid tracking, providing detailed and high-quality human performance rendering. For quantitative comparison, we evaluate each method on three 200-frame sequences from our dataset and ActorsHQ separately. We use various metrics, including PSNR, SSIM, LPIPS, the temporal metric VMAF [31], and per-frame storage. As seen

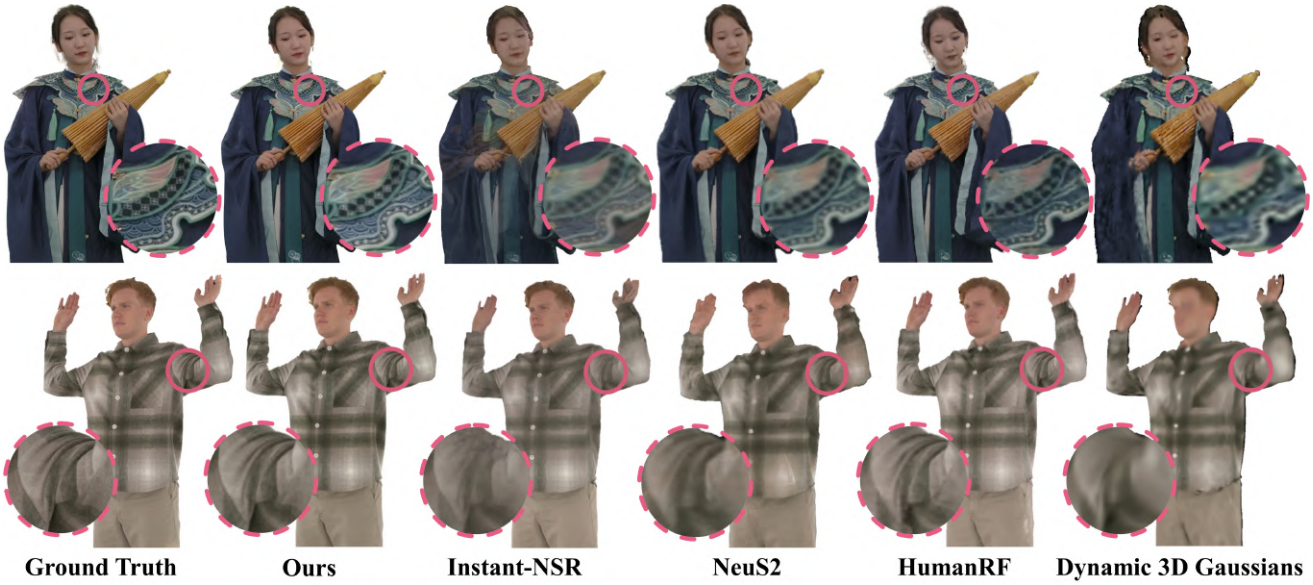

Figure 5. Qualitative comparison of our method against Instant-NSR [81], NeuS2 [69], HumanRF [22] and Dynamic 3D Gaussians [40] on both our dataset and ActorsHQ [22]. Our method achieves the highest rendering quality.

Table 1. **Quantitative comparison on our dataset**. Green and yellow cell colors indicate the best and the second-best results.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	VMAF↑	Per-frame Storage(MB) ↓
Instant-NSR [81]	29.385	0.958	0.0370	68.309	11.63
NeuS2 [69]	32.952	0.961	0.0682	79.102	24.16
HumanRF [22]	31.174	0.977	0.0298	80.942	11.38
Dynamic 3D Gaussians [40]	30.244	0.965	0.0847	52.224	4.523
Ours(Before Compression)	36.205	0.989	0.0184	85.127	43.42
Ours(After Compression)	35.788	0.986	0.0214	84.312	1.648

Table 2. **Quantitative comparison on ActorsHQ dataset**.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	VMAF↑	Per-frame Storage(MB) ↓
Instant-NSR [81]	27.145	0.932	0.0949	55.635	13.72
NeuS2 [69]	32.124	0.939	0.1579	81.374	29.09
HumanRF [22]	34.106	0.963	0.0549	85.198	10.95
Dynamic 3D Gaussians [40]	22.413	0.911	0.2191	41.374	2.325
Ours(Before Compression)	35.029	0.969	0.0909	88.427	50.91
Ours(After Compression)	34.704	0.967	0.1013	87.543	2.143

in Tab. 1 and Tab. 2, HiFi4G surpasses other methods in both quality and storage. Note that our compression strategy significantly reduces per-frame storage requirements without compromising quality. Remarkably, even on the VMAF metric [31], which evaluates the perceptual quality and temporal consistency, our explicit method outperforms HumanRF which benefits from the inherent smoothness of the MLP.

5.2. Ablation Study

Compact 4D Gaussians. We conduct a qualitative ablation on the dual-graph and the regularization term to assess their impact on post-compression rendering results. As shown in Fig. 6, the removal of the coarse ED-graph prior typically causes severe artifacts. Excluding the Gaussian graph often results in significant precision loss and unnatural rendering. Regarding regularizers, the omission of E_{temp} usually triggers unrealistic artifacts post-compression. Meanwhile, the absence of E_{smooth} produces blurry results, with both leading to flickering in the video. Additionally, to evaluate the impact of the adaptive weight $w_{i,t}$, we replace it with a fixed weight of 0.1. This adjustment generally leads to noticeable blurriness, especially in areas with significant movement. In contrast, our full pipeline generates spatially and temporally compact 4D Gaussians, maintaining high-fidelity rendering even after compression. The quantitative results are as demonstrated in Tab. 3, in which our full approach achieves the highest accuracy.

Residual Compensation. As illustrated in Fig. 7 (b), we allocate 48.24MB of storage for the 4D Gaussians of each frame before compression. Applying high-bit quantization (0-bit for motion and 9-bit for appearance) without residual compensation results in a storage requirement of 7.41MB, as shown in Fig. 7 (c). Using low-bit quantization (11-bit for motion and 7-bit for appearance), again without residual compensation, reduces storage to 3.67MB but compromises rendering quality, as illustrated in Fig. 7 (d). In contrast, applying the same low-bit quantization but with residual compensation significantly reduces storage needs to under 2MB per frame while maintaining the same level of rendering quality, as shown in Fig. 7 (e).

The Number of 4D Gaussians. We assess the impact of changing the number of 4D Gaussians on the quality of results across three sequences. As depicted in Fig. 8, using 200,000 4D Gaussians is adequate for generating high-quality results. This amount enables effective compression to less than 2MB, supporting immersive viewing on diverse platforms, including VR and AR.

Table 3. Quantitative evaluation of compact 4D Gaussians.

	PSNR ↑	SSIM ↑	LPIPS ↓	VMAF ↑
w/o ED graph	29.142	0.9534	0.0662	69.724
w/o Gaussian graph	31.185	0.9541	0.0586	76.873
w/o E_{temp}	33.555	0.9661	0.0496	76.308
w/o E_{smooth}	33.889	0.9657	0.0518	81.944
w/o $w_{i,t}$	33.577	0.9678	0.0425	81.236
Ours	35.085	0.9828	0.0219	83.133

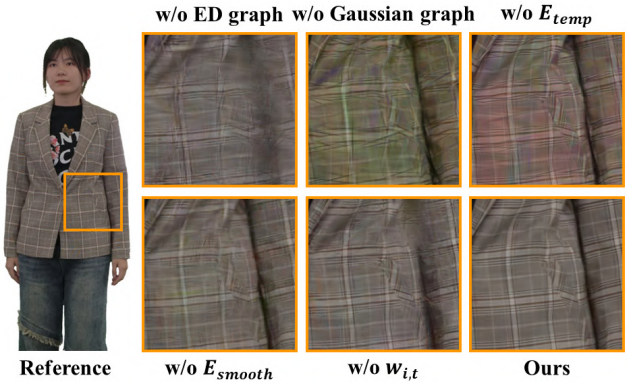

Figure 6. Qualitative evaluation of compact 4D Gaussians.

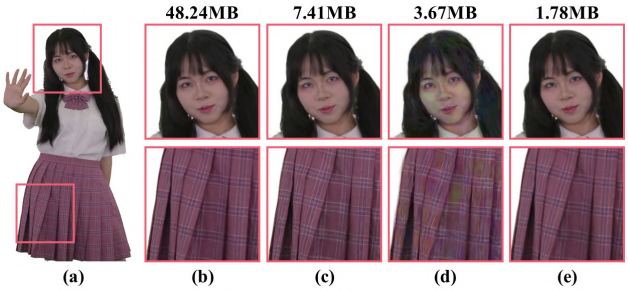

Figure 7. Qualitative evaluation of our residual strategy. (a) Reference image; (b) 4D Gaussians results before compression; (c) Per-frame encoding using high-bit quantization without residual; (d) Per-frame encoding using low-bit quantization without residual; (e) Ours results using low-bit quantization with residual.

Table 4. Run-time evaluation of each step.

Procedure	Time
Background matting	∼ 1 min
Meshing	∼ 1 min
non-rigid tracking	∼ 100 ms
4D Gaussians Optimization	∼ 4 mins
4D Gaussians Compression	∼ 100 ms

Run-time Evaluation of Each Step. As shown in Tab. 4, we also provide the runtime for each step on a PC with an Nvidia GeForce RTX3090 GPU, which includes both the preprocessing and training stages. Our method can generate 4D assets efficiently, taking less than 7 minutes per frame.

5.3. Limitation

Although HiFi4G achieves high-fidelity 4D human performance rendering via compact Gaussian Splatting, it still has some limitations. First, HiFi4G heavily relies on segmentation, poor segmentation can lead to significant artifacts, especially in scenes with human-object interactions. Moreover, our method necessitates per-frame reconstruction and mesh tracking, which presents an interesting direction in exploring a more synergistic relationship between tracking and rendering. Even though HiFi4G is efficient in generating 4D assets, the Gaussian optimization process still requires several minutes, forming a major bottleneck. Accelerating this training process is vital for future research. Additionally, the current dependence of 4D Gaussian on fast GPU sorting limits the deployment of HiFi4G on web viewers and mobile devices.

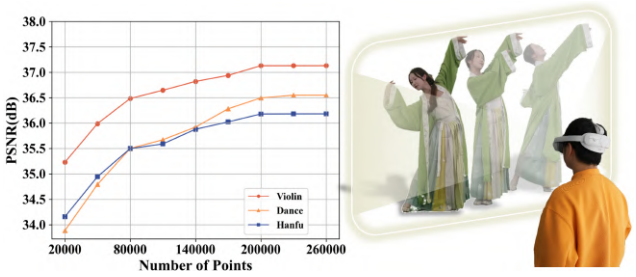

Figure 8. Evaluation of 4D Gaussians number. With ∼200,000 4D Gaussians, HiFi4G achieves high-fidelity human performance rendering, suitable for integration in VR applications.

6. Conclusion

We have presented an explicit and compact Gaussian-based approach for 4D human performance rendering from RGB inputs. By bridging 3D Gaussian Splatting with non-rigid tracking, our approach achieves high-fidelity rendering results, outperforming previous methods in terms of quality, efficiency, and storage. Our dual-graph mechanism provides sufficient non-rigid motion priors in a keyframe-based manner, while our Gaussian optimization scheme with novel regularization designs effectively ensures spatial-temporal consistency of the 4D Gaussian Splatting. We also demonstrate the compactness of our representation with a companion compression scheme which substantially reduces storage requirements. Our experimental results further demonstrate the effectiveness of our approach for delivering lifelike human performances. With its explicit and compact characteristics, we believe our approach makes a solid step forward to faithfully recording and providing immersive experiences of human performances on various platforms like VR headsets.

Acknowledgements. We thank reviewers for their feedback. This work was supported by National Key R&D Program of China (2022YFF0902301), Shanghai Local college capacity building program (22010502800). We also acknowledge support from Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI).

References

[1] Aljaz Bozic, Pablo Palafox, Michael Zollhofer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1450–1459, 2021. 2

[2] Andrei Burov, Matthias Nießner, and Justus Thies. Dynamic surface function networks for clothed human bodies, 2021. 2

[3] Hongrui Cai, Wanquan Feng, Xuetao Feng, Yan Wang, and Juyong Zhang. Neural surface reconstruction of dynamic scenes with monocular rgb-d camera. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3

[4] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 3

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[6] Matteo Carandini, Jonathan B Demb, Valerio Mante, David J Tolhurst, Yang Dan, Bruno A Olshausen, Jack L Gallant, and Nicole C Rust. Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–10597, 2005. 5

[7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022. 2, 3

[8] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34(4):69, 2015. 1

[9] 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 Trans. Graph.*, 35(4), 2016. 2, 3

[10] 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 Trans. Graph.*, 36(6): 246:1–246:16, 2017. 1, 2

[11] Jarek Duda. Asymmetric numeral systems: entropy coding combining speed of huffman coding with compression rate of arithmetic coding. *arXiv preprint arXiv:1311.2540*, 2013. 5

[12] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2, 3

[13] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2021. 1

[14] Kaiwen Guo, Feng Xu, Yangang Wang, Yebin Liu, and Qionghai Dai. Robust Non-Rigid Motion Tracking and Surface Reconstruction Using L0 Regularization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3083–3091, 2015. 2

[15] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo and motion reconstruction using a single rgbd camera. *ACM Transactions on Graphics (TOG)*, 2017. 3

[16] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 38(2):14:1–14:17, 2019. 2

[17] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 2

[18] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics (ToG)*, 40(4):1–16, 2021. 2

[19] Marc Habermann, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhoefer, and Christian Theobalt. Hdhumans: A hybrid approach for high-fidelity digital humans. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–23, 2023. 2

[20] Yannan He, Anqi Pang, Xin Chen, Han Liang, Minye Wu, Yuexin Ma, and Lan Xu. Challencap: Monocular 3d capture of challenging human performances using multi-modal references. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[21] Wenbo Hu, Yuling Wang, Lin Ma, Bangbang Yang, Lin Gao, Xiao Liu, and Yuewen Ma. Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19774–19783, 2023. 3

[22] 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, 3, 6, 7

[23] Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, and Lan Xu. Neuralhofusion: Neural volumetric rendering under human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6155–6165, 2022. 2

[24] Yuheng Jiang, Kaixin Yao, Zhuo Su, Zhehao Shen, Haimin Luo, and Lan Xu. Instant-nvr: Instant neural volumetric rendering for human-object interactions from monocular rgbd

stream. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 595–605, 2023. 3

[25] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[26] 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, 6

[27] Youngjoong Kwon, Lingjie Liu, Henry Fuchs, Marc Habermann, and Christian Theobalt. Deliffas: Deformable light fields for fast avatar synthesis. *Advances in Neural Information Processing Systems*, 2023. 2

[28] 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. 2

[29] Yue Li, Marc Habermann, Bernhard Thomaszewski, Stelian Coros, Thabo Beeler, and Christian Theobalt. Deep physics-aware inference of cloth deformation for monocular human performance capture. In *2021 International Conference on 3D Vision (3DV)*, pages 373–384. IEEE, 2021. 2

[30] Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, and Matthias Nießner. 4dcomplete: Non-rigid motion estimation beyond the observable surface. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[31] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, Megha Manohara, et al. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 6(2):2, 2016. 6, 7

[32] Zhe Li, Tao Yu, Zerong Zheng, Kaiwen Guo, and Yebin Liu. Posefusion: Pose-guided selective fusion for single-view human volumetric capture. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[33] Zhe Li, Zerong Zheng, Hongwen Zhang, Chaonan Ji, and Yebin Liu. Avatarcap: Animatable avatar conditioned monocular human volumetric capture. In *ECCV*, 2022. 2

[34] Zujie Liang and Fan Liang. Transpcc: Towards deep point cloud compression via transformers. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 1–5, 2022. 3

[35] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia Conference Proceedings*, 2022. 2, 3

[36] Haotong Lin, Sida Peng, Zhen Xu, Tao Xie, Xingyi He, Hujun Bao, and Xiaowei Zhou. High-fidelity and real-time novel view synthesis for dynamic scenes. In *SIGGRAPH Asia 2023 Conference Papers*, New York, NY, USA, 2023. Association for Computing Machinery. 3

[37] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 5

[38] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 2020. 2

[39] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 2

[40] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 2, 3, 4, 6, 7

[41] 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 *Computer Vision – ECCV 2020*, pages 405–421, Cham, 2020. Springer International Publishing. 1

[42] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 2, 3

[43] Marcus J Nadenau, Julien Reichel, and Murat Kunt. Wavelet-based color image compression: exploiting the contrast sensitivity function. *IEEE Transactions on image processing*, 12(1):58–70, 2003. 3

[44] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 1, 2, 3, 4

[45] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12695–12705, 2021. 2

[46] 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. 1

[47] 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), 2021.

[48] 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. 1

[49] 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

[50] Sergey Prokudin, Qianli Ma, Maxime Raafat, Julien Valentin, and Siyu Tang. Dynamic point fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7964–7976, 2023. 2

[51] Maurice Quach, Giuseppe Valenzise, and Frederic Dufaux. Learning convolutional transforms for lossy point cloud geometry compression. In *2019 IEEE international conference on image processing (ICIP)*, pages 4320–4324. IEEE, 2019. 3

[52] Maurice Quach, Giuseppe Valenzise, and Frederic Dufaux. Improved deep point cloud geometry compression. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2020. 3

[53] Christian Reiser, Rick Szeliski, Dor Verbin, Pratul Srinivasan, Ben Mildenhall, Andreas Geiger, Jon Barron, and Peter Hedman. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 3

[54] Ruwen Schnabel and Reinhard Klein. Octree-based point-cloud compression. *PBG@ SIGGRAPH*, 3, 2006. 3

[55] Sebastian Schwarz, Marius Preda, Vittorio Baroncini, Madhukar Budagavi, Pablo Cesar, Philip A Chou, Robert A Cohen, Maja Krivokuća, Sébastien Lasserre, Zhu Li, et al. Emerging mpeg standards for point cloud compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(1):133–148, 2018. 3

[56] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16632–16642, 2023. 3

[57] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16911–16921, 2023. 3

[58] Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1395, 2017. 2

[59] 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. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 2, 3

[60] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *Computer Vision – ECCV 2020*, pages 246–264, Cham, 2020. Springer International Publishing. 2

[61] Zhuo Su, Lan Xu, Dawei Zhong, Zhong Li, Fan Deng, Shuxue Quan, and Lu Fang. Robustfusion: Robust volumetric performance reconstruction under human-object interactions from monocular rgbd stream. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[62] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. *ACM Transactions on Graphics (TOG)*, 26(3):80, 2007. 1, 2, 3, 4

[63] Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingya Wang, and Jingyi Yu. Neural free-viewpoint performance rendering under complex human-object interactions. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 2

[64] Xin Suo, Yuheng Jiang, Pei Lin, Yingliang Zhang, Minye Wu, Kaiwen Guo, and Lan Xu. Neuralhumanfvv: Real-time neural volumetric human performance rendering using rgb cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6226–6237, 2021. 2

[65] Dorina Thanou, Philip A Chou, and Pascal Frossard. Graph-based compression of dynamic 3d point cloud sequences. *IEEE Transactions on Image Processing*, 25(4):1765–1778, 2016. 3

[66] 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. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2021. 1, 2

[67] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally aware piecewise transformation fields for 3d human mesh registration. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[68] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *European Conference on Computer Vision*, 2022. 1

[69] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3295–3306, 2023. 2, 3, 6, 7

[70] 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 (CVPR)*, pages 16210–16220, 2022. 2

[71] 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. *arXiv preprint arXiv:2310.08528*, 2023. 2, 3, 4

[72] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *2020 International Conference on 3D Vision (3DV)*, pages 322–332. IEEE, 2020. 2

[73] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, Yaser Sheikh, Jessica Hodgins, and Chenglei Wu. Dressing avatars: Deep

photorealistic appearance for physically simulated clothing. *ACM Trans. Graph.*, 41(6), 2022. 2

[74] Lan Xu, Wei Cheng, Kaiwen Guo, Lei Han, Yebin Liu, and Lu Fang. Flyfusion: Realtime dynamic scene reconstruction using a flying depth camera. *IEEE transactions on visualization and computer graphics*, 27(1):68–82, 2019. 3

[75] Lan Xu, Zhuo Su, Lei Han, Tao Yu, Yebin Liu, and Lu Fang. Unstructuredfusion: realtime 4d geometry and texture reconstruction using commercial rgbd cameras. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2508–2522, 2019. 2

[76] 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, 3, 4

[77] 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. 2, 3

[78] 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. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019. 2

[79] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5746–5756, 2021. 2

[80] Hongwen Zhang, Siyou Lin, Ruizhi Shao, Yuxiang Zhang, Zerong Zheng, Han Huang, Yandong Guo, and Yebin Liu. Closet: Modeling clothed humans on continuous surface with explicit template decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[81] 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 Trans. Graph.*, 41(6), 2022. 3, 6, 7

[82] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7753, 2022. 2

[83] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rhemann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew W. Fitzgibbon, Charles T. Loop, Christian Theobalt, and Marc Stamminger. Real-time nonrigid reconstruction using an rgb-d camera. *ACM Trans. Graph.*, 33(4):156:1–156:12, 2014. 2