

Robust Dual Gaussian Splatting for Immersive Human-centric Volumetric Videos

YUHENG JIANG, ShanghaiTech University, China and NeuDim Digital Technology (Shanghai) Co.,Ltd., China

ZHEHAO SHEN, ShanghaiTech University, China

YU HONG, ShanghaiTech University, China

CHENGCHENG GUO, ShanghaiTech University, China

YIZE WU, ShanghaiTech University, China

YINGLIANG ZHANG, DGene Digital Technology Co., Ltd., China

JINGYI YU, ShanghaiTech University, China

LAN XU*, ShanghaiTech University, China



Fig. 1. We present a robust human performance tracking and rendering approach with a customized compression scheme. Our method serves as a “ticket” to a virtual world, offering immersive, high-fidelity viewing of multiple musicians performing.

*The corresponding author is Lan Xu (xulan1@shanghaitech.edu.cn).

Authors’ addresses: Yuheng Jiang, ShanghaiTech University, Shanghai, China and NeuDim Digital Technology (Shanghai) Co.,Ltd., China, zhaofq@shanghaitech.edu.cn; Zhehao Shen, ShanghaiTech University, Shanghai, China, shenzhh@shanghaitech.edu.cn; Yu Hong, ShanghaiTech University, Shanghai, China, hongyu@shanghaitech.edu.cn; Chengcheng Guo, ShanghaiTech University, Shanghai, China, guochch@shanghaitech.edu.cn; Yize Wu, ShanghaiTech University, Shanghai, China, wuyize25@163.com; Yingliang Zhang, DGene Digital Technology Co., Ltd., China, yingliang.zhang@dgene.com; Jingyi Yu, ShanghaiTech University, Shanghai, China, yujingyi@shanghaitech.edu.cn; Lan Xu, ShanghaiTech University, Shanghai, China, xulan1@shanghaitech.edu.cn.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2024 Copyright held by the owner/author(s).

0730-0301/2024/12-ART

<https://doi.org/10.1145/3687926>

Volumetric video represents a transformative advancement in visual media, enabling users to freely navigate immersive virtual experiences and narrowing the gap between digital and real worlds. However, the need for extensive manual intervention to stabilize mesh sequences and the generation of excessively large assets in existing workflows impedes broader adoption. In this paper, we present a novel Gaussian-based approach, dubbed *DualGS*, for real-time and high-fidelity playback of complex human performance with excellent compression ratios. Our key idea in *DualGS* is to separately represent motion and appearance using the corresponding skin and joint Gaussians. Such an explicit disentanglement can significantly reduce motion redundancy and enhance temporal coherence. We begin by initializing the *DualGS* and anchoring skin Gaussians to joint Gaussians at the first frame. Subsequently, we employ a coarse-to-fine training strategy for frame-by-frame human performance modeling. It includes a coarse alignment phase for overall motion prediction as well as a fine-grained optimization for robust tracking and high-fidelity rendering. To integrate volumetric video seamlessly into VR environments, we efficiently compress motion using entropy encoding and appearance using codec compression coupled with a persistent codebook. Our approach achieves a compression ratio of up to 120

times, only requiring approximately 350KB of storage per frame. We demonstrate the efficacy of our representation through photo-realistic, free-view experiences on VR headsets, enabling users to immersively watch musicians in performance and feel the rhythm of the notes at the performers' fingertips. Project page: <https://nowheretrix.github.io/DualGS/>.

CCS Concepts: • Computing methodologies → Image-based rendering

Additional Key Words and Phrases: human performance capture, neural rendering, Gaussian Splatting

ACM Reference Format:

Yuheng Jiang, Zhehao Shen, Yu Hong, Chengcheng Guo, Yize Wu, Yingliang Zhang, Jingyi Yu, and Lan Xu. 2024. Robust Dual Gaussian Splatting for Immersive Human-centric Volumetric Videos. *ACM Trans. Graph.* 43, 6 (December 2024), 15 pages. <https://doi.org/10.1145/3687926>

1 INTRODUCTION

As the distinction between digital and real worlds diminishes, 3D and 4D content is rapidly gaining prominence, reshaping societal expectations and applications across digital landscapes. Among these innovations, volumetric videos represent a significant advancement in visual media. They provide viewers with six degrees of freedom, enabling users to navigate virtual environments freely. Specifically, users can immerse themselves in a virtual musical odyssey, observing musicians perform up close and feeling the rhythm of the music as if standing beside them (see Fig. 1).

Over the past two decades, numerous studios [Collet et al. 2015; Işık et al. 2023; Vlasic et al. 2009; Zhang et al. 2022] have established multi-view domes worldwide, from west to east, for capturing volumetric videos. Yet, the predominant workflow for producing human-centric volumetric videos still relies on the explicit reconstruction and tracking of textured meshes. This method is prone to occlusions, often resulting in holes and noises that degrade texturing quality. Creating even minimally immersive segments requires substantial computational resources and meticulous cleanup by skilled artists. Moreover, the volumetric assets are often too large for storage and integration into immersive devices. As a result, volumetric video has not achieved widespread adoption.

Neural advances in photo-realistic rendering, notably through Neural Radiance Fields [Mildenhall et al. 2020], have facilitated bypassing explicit reconstructions and enhancing novel view synthesis. Recently, 3D Gaussian Splatting (3DGS) advances the explicit paradigm by using learnable Gaussians to achieve high-fidelity rendering at unprecedented frame rates. It emergently facilitates the development of various dynamic variants. For animatable avatar modeling, many works [Hu et al. 2024; Kocabas et al. 2024; Li et al. 2024b; Pang et al. 2024] transform 3D Gaussians to posed space using linear blend skinning. For volumetric video playback, some studies [Wu et al. 2024b; Yang et al. 2024] combine 3DGS with MLPs to model temporal coherence, sacrificing the explicit and GPU-friendly beauty of 3DGS. Yet, these methods are still fragile to challenging motions and require significant storage.

In this paper, we present a novel Gaussian-based representation for volumetric videos, achieving robust human performance tracking and high-fidelity rendering. Our core idea is to utilize Dual Gaussians, named *DualGS*, for disentangled and hierarchical motion and appearance representation. It significantly enhances temporal coherence and tracking accuracy and also enables a companion

compression strategy. Our approach achieves significant storage efficiency, requiring only approximately 350KB of storage per frame. DualGS also maintains highly competitive rendering quality and consistently delivers superior rendering and temporal consistency across various challenging cases.

In DualGS, inspired by the SMPL model [Loper et al. 2015], which represents skin motion by interpolating a few joints, we utilize a compact number of motion-aware *joint Gaussians* to capture global movements and a larger set of appearance-aware *skin Gaussians* for visual representation. For the initialization of our DualGS representation in the first frame, we randomly initialize joint Gaussians and carefully control their scale and size to effectively represent the overall movement of the performance. Once optimized, these joint Gaussians serve as the basis for initializing the skin Gaussians. To establish the relationship between dual Gaussians, each skin Gaussian is anchored to multiple joint Gaussians, facilitating the interpolation of position and rotation for sequential optimization. Then, for the subsequent frame-by-frame human performance tracking, we employ a novel coarse-to-fine optimization strategy that enhances both temporal coherence and rendering fidelity. During the coarse alignment phase, we perform optimization only on the joint Gaussians, using a locally as-rigid-as-possible regularizer while maintaining fixed appearance attributes. We also integrate a motion prediction module to aid this phase and ensure robust tracking. In the fine-grained optimization phase, we recompute the skin Gaussian motions from joint data as well as fine-tune the detailed positions and appearances using temporal regularizers in a differentiable manner. Such a coarse-to-fine optimization provides explicit disentanglement of the Gaussian attributes in our DualGS, and hence significantly improves the tracking accuracy.

Despite the advancements, integrating long-duration sequences into low-end devices like VR headsets remains challenging. Benefiting from our explicit DualGS representation, we effectively separate and compress the motion and appearance attributes. Specifically, for joint Gaussians, we employ Residual-Vector Quantization combined with entropy encoding to efficiently handle the motion attributes. For skin Gaussians, we first employ codec compression for spatial-temporal Look-up Tables, addressing both scaling and opacity attributes. Then, to manage the storage-intensive spherical harmonic (SH) attributes, we design a specialized persistent codebook. This codebook compresses SH attributes into persistent SH indices, coupled with length encoding. Our approach achieves a compression ratio of up to 120 times compared to the original 3DGS. It enables the seamless integration of multiple 4D assets (illustrated with 9 performers in Fig. 1) into VR environments for real-time rendering. This capability enables users to experience the notes pouring from the musician's dancing fingertips, embarking on a deeply immersive and enchanting musical odyssey.

2 RELATED WORK

Human Performance Capture. Recent research on human performance capture aims to achieve detailed registration for various applications [Habermann et al. 2019; Li et al. 2021; Shao et al. 2022; Slavcheva et al. 2017; Wang et al. 2021; Xiang et al. 2020; Zhang et al. 2023; Zhao et al. 2022a]. Starting with the pioneering work

DynamicFusion [Newcombe et al. 2015], which benefits from the GPU solver to achieve real-time capture, VolumeDeform [Innmann et al. 2016] combines depth-based correspondences with sparse SIFT features to reduce drift. Fusion4d [Dou et al. 2016] and Motion2fusion [Dou et al. 2017] utilize a key-frame strategy to handle topological changes. KillingFusion [Slavcheva et al. 2017] and SobolevFusion [Slavcheva et al. 2018] address these variations by introducing additional constraints on the motion fields. For more robust tracking, DoubleFusion [Yu et al. 2019] proposes a two-layer representation aided by a human parametric model, extended by UnstructureFusion [Xu et al. 2019b] for unstructured setups. RobustFusion [Su et al. 2020, 2022] further addresses the challenging human-object interaction scenarios. DDC [Habermann et al. 2021a] learns the deformations with skeletons and embedded graph [Sumner et al. 2007] and DELIFFAS [Kwon et al. 2024] parameterized the light field based on DDC. Other efforts [Jiang et al. 2022, 2023b; Yu et al. 2021b] marry the non-rigid deformation with implicit neural advances for better performance. Nevertheless, these methods rely on parametric template priors, focusing more on overall tracking accuracy, which limits their ability to capture fine details like wrinkles and high-frequency texture.

Neural Human Modeling. In the domain of digital human neural representation, various approaches [Lin et al. 2023, 2022; Liu et al. 2020; Shetty et al. 2024; Sun et al. 2021; Suo et al. 2021; Xiang et al. 2022] have been proposed to address this challenge. A collection of studies [Pumarola et al. 2021; Tretschk et al. 2021; Xian et al. 2021] model time as an additional latent variable into the NeRF’s MLP. For dynamic human modeling, some methods [Habermann et al. 2023, 2021b; Liu et al. 2021; Luvizon et al. 2023; Zhu et al. 2023] leverage skeleton-based and graph embedding representations, while another line of studies [Jiang et al. 2023a; Li et al. 2022; Shen et al. 2023; Wang et al. 2022a] build upon the SNARF [Chen et al. 2021] framework, which learns skinning weights through root-finding, resulting in enhanced reconstruction accuracy and improved animation quality. Humannerfs [Weng et al. 2022; Zhao et al. 2022b] utilize the human prior SMPL [Loper et al. 2015] model as an anchor to warp the radiance field. NeuVV [Zhang et al. 2022] and Fourier PlenOctrees [Wang et al. 2022b] leverage advanced PlenOctree [Yu et al. 2021a] and volumetric fusion to achieve real-time rendering of dynamic scenes with significant acceleration. Recent methods [İşik et al. 2023; Song et al. 2023; Wang et al. 2023a] draw inspiration from advanced framework [Chen et al. 2022; Müller et al. 2022] and incorporate explicit optimizable embeddings into the implicit representation to accelerate training times and rendering speeds. Building on the pioneering work of 3DGS, several dynamic variants [Jena et al. 2023; Moreau et al. 2024; Qian et al. 2024b; Wu et al. 2024b; Yang et al. 2024] utilize MLPs and human parametric models to establish temporal correspondences. GPS-Gaussian [Zheng et al. 2024] develops an NVS system to regress Gaussian maps. Spacetime Gaussians [Li et al. 2024a] extend this approach by incorporating polynomials. ASH [Pang et al. 2024] and Animatable Gaussians [Li et al. 2024c] parameterize mesh positions in 2D space and infer Gaussian maps using a UNet architecture. GaussianAvatars [Chen et al. 2024a; Qian et al. 2024a] bind Gaussians to the FLAME mesh

for animation, while D3GA [Zielonka et al. 2023] relies on tetrahedral cages. HiFi4G [Jiang et al. 2024] leverages embedded deformation [Sumner et al. 2007] to accelerate training. However, most existing methods suffer from blurred results or struggle with fast motions. In contrast, our approach employs dual Gaussians coupled with a coarse-to-fine training strategy, enabling robust tracking and high-fidelity rendering.

Data Compression. Compact representation plays a pivotal role in 3D/4D reconstruction, attracting significant research interest. For traditional animated meshes, numerous studies use PCA [Alexa and Müller 2000; Luo et al. 2013; Vasa and Skala 2007] or mesh pre-segmentation [Gupta et al. 2002; Mamou et al. 2009] to identify geometric parts of the human body to ensure connectivity consistency while others [Ibarria and Rossignac 2003; Luo et al. 2013; Mamou et al. 2009] predict vertex trajectories to maintain temporal coherence in vertex groups. For neural fields, several studies propose compact neural representations through CP-decomposition [Chen et al. 2022], rank reduction [Tang et al. 2022], codec [Wang et al. 2023b] and tri-planes [Hu et al. 2023; Reiser et al. 2023]. Recent works focus on the compression of 3D Gaussian representations. Compact3D [Navaneet et al. 2023], C3DGGS [Niedermayr et al. 2024a] and Compact-3DGGS [Lee et al. 2024] use vector quantization and entropy encoding while LightGaussian [Fan et al. 2023] prunes Gaussians and adopts octree-based compression for positions. RDO-Gaussian [Wang et al. 2024b] and Reduced3DGGS [Papantonakis et al. 2024] combine redundant Gaussian culling with vector quantization, whereas SOG [Morgenstern et al. 2023] maps Gaussian attributes onto 2D grids and utilizes image codec compression techniques. Scaffold-GS [Lu et al. 2024] leverages anchor points to significantly reduce the number of redundant Gaussians, while HAC [Chen et al. 2024b] further enhances compression with a combination of hash tables and learnable features. EAGLES [Girish et al. 2023] compresses attributes using quantized latent codes and a trainable decoder. In the dynamic domain, 4K4D [Xu et al. 2024] employs a 4D feature grid and IBR module, VideoRF [Wang et al. 2024a] encodes 4D radiance fields into 2D feature streams, TeTriRF [Wu et al. 2024a] bakes density grid sequences into the tri-plane representation, while HiFi4G [Jiang et al. 2024] utilizes residual computation and entropy encoding. However, these methods either fail to achieve high compression ratios or compromise quality. In contrast, our method requires only 350KB per frame while maintaining high-fidelity rendering.

3 DUAL-GAUSSIAN REPRESENTATION

Given multi-view videos capturing a dynamic 3D scene, our objective is to robustly track human performance and achieve high-quality novel view rendering in real-time. The methodology is visually summarized in Fig. 2. We first introduce a Dual Gaussian(DualGS) representation, which comprises a small number of motion-aware joint Gaussians to capture global movements and a large set of appearance-aware skin Gaussians to express visual appearance. Additionally, we propose a novel coarse-to-fine optimization strategy with a motion prediction module to ensure temporal consistency and produce high-fidelity Gaussian assets. Our method

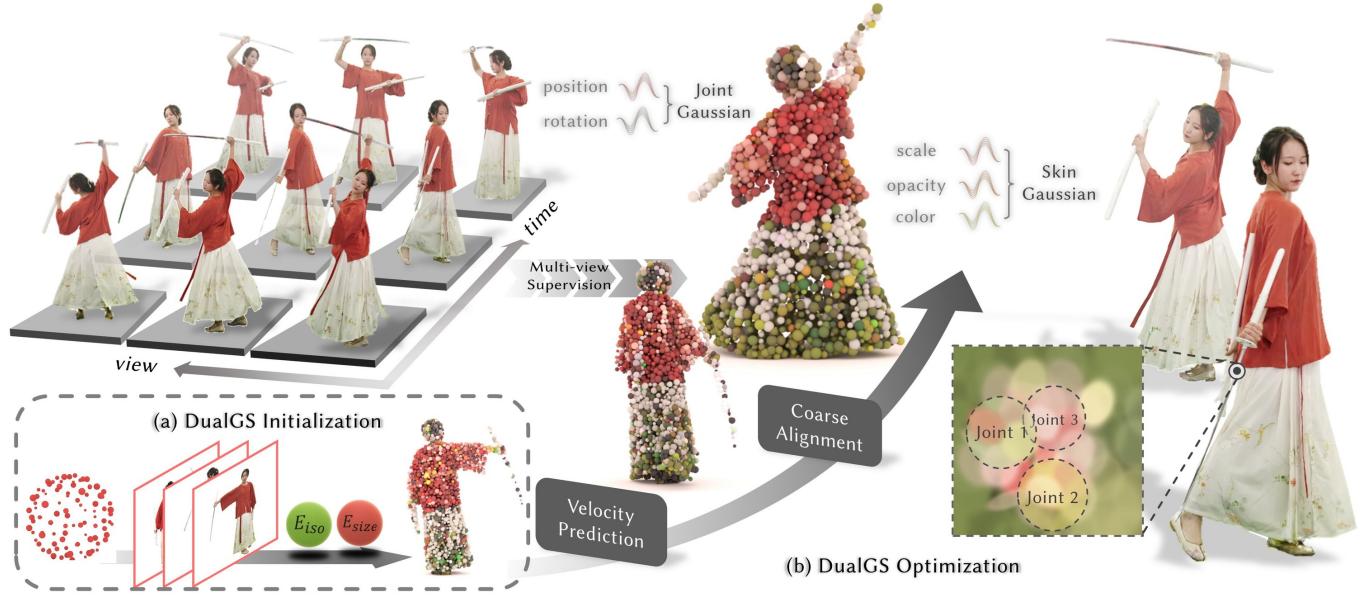


Fig. 2. We propose a novel Dual Gaussian representation to capture challenging human performance from multi-view inputs. We first optimize joint Gaussians from a random point cloud, then use them to initialize skin Gaussians, expressing their motion through interpolation. In the following optimization, we employ a coarse-to-fine strategy, with a coarse alignment for overall motion prediction and fine-grained optimization for robust tracking and high-fidelity rendering.

enables accurate tracking and realistic rendering at 4K resolution, outperforming existing approaches in performance and quality.

3.1 Dual-Gaussian Initialization

We first initialize the DualGS to establish the mapping between the skin Gaussians and joint Gaussians. To simplify the description, we first categorize the attributes of Gaussians into two groups: 1) motion-aware parameters, which include position p and rotation q . 2) appearance-aware parameters, comprising spherical harmonic C , opacity σ , and scaling s . Inspired by the human parametric model, where skin vertices are represented through the interpolation of a minimal number of joints, our approach utilizes a dual Gaussian scheme to separately encode motion and appearance. We utilize a compact set of Gaussians ($\sim 15,000$) to encapsulate the overall motion of the performer, while a more extensive set of Gaussians ($\sim 180,000$) captures the nuanced appearance details. Once established, the number of Gaussians remains constant over time, with only the attributes subject to continuous updates. This formulation yields two main benefits: 1) The relative positions of the skin Gaussians in local space maintain stability, driven consistently by the same joint Gaussians motion, thereby enhancing spatial-temporal consistency. 2) The reduced motion parameters are highly conducive to subsequent compression processes.

Analogous to the original 3DGs [Kerbl et al. 2023], we commence by initializing a small number of joint Gaussians to represent global motion dynamics. The Gaussian model is trained on the first frame using a uniform random initialization. During the training process, we regulate the number of Gaussians to strike an optimal balance between efficient motion representation and compact storage. Specifically, we perform densification and pruning before 15,000 iterations. The joint Gaussians are then downsampled to approximately 15,000,



Fig. 3. Sampled results from our DualGS optimization pipeline. With the aid of our coarse-to-fine training strategy, we can produce high-fidelity 4D assets.

fixing this number and subsequently optimizing only their values. Skinny kernels are generated to effectively fit local appearance details but lack geometric information, leading to unexpected plush artifacts in the following optimizations. To address this, we follow PhysGaussian [Xie et al. 2024] that employs an isotropic loss that constrains overly skinny scaling:

$$E_{\text{iso}} = \frac{1}{N} \sum_{i=1}^N \text{ReLU}(e^{\max(s_i) - \min(s_i)} - r), \quad (1)$$

where s_i represents the i -th joint Gaussian scaling parameters, and e is the activation function. We enforce a constraint that the ratio between the length of the major and minor axis does not exceed r . Additionally, we propose another term to constrain oversized

Gaussians, preventing local over-reconstruction:

$$E_{\text{size}} = \sum_{i=1}^N \text{ReLU} \left(s_i - \alpha \frac{1}{N} \sum_{i=1}^N sg[s_i] \right), \quad (2)$$

where sg stands for the stop-gradient operator. E_{size} penalizes those Gaussians whose scale exceeds the average size by a factor of α .

For skin Gaussians initialization, we use the initialized joint Gaussians kernels as inputs and perform training to achieve high-fidelity quality. During the training process, the position of skin Gaussians is updated through differentiable rasterization. According to human parametric models [Li et al. 2017; Loper et al. 2015] where skin vertices are driven by predefined joint motions and skinning weights, we then bind each skin Gaussian to the k-nearest joint Gaussians. Specifically, for each skin Gaussian position p_i^s , we identify the k-nearest neighbors(KNN, $k = 8$) joint Gaussians p_k^j within the encompassing ellipsoid to serve as its anchor joints. The blending weight is defined as:

$$w(p_i^s, p_k^j) = \exp \left(-\|p_i^s - p_k^j\|_2^2 / l^2 \right), \quad (3)$$

where l is the influence radius. Here, the superscripts j and s denote the joint Gaussians and skin Gaussians, respectively. This KNN graph and the corresponding blending weights are integral to the subsequent optimization process and remain fixed throughout.

Notably, our approach offers greater flexibility compared to the parametric model, which relies on predefined joints, skinning weights, or a fixed topology. Experimental results demonstrate that our method can handle a wide range of dynamic sequences.

Implementation. We perform 30,000 training iterations for DualGS initialization separately. For joint Gaussians, the complete loss function is as follows:

$$E_{\text{init}} = \lambda_{\text{iso}} E_{\text{iso}} + \lambda_{\text{size}} E_{\text{size}} + E_{\text{color}}, \quad (4)$$

where E_{color} is the photometric loss. We use the following empirically determined parameters: $r = 4, \alpha = 3, l = 0.001, \lambda_{\text{iso}} = 0.005, \lambda_{\text{size}} = 1$.

3.2 Dual-Gaussian Optimization

For sequential training, we fix the number of DualGS and optimize the motion of joint Gaussians as well as the appearance of skin Gaussians. We observe that Gaussians tend to alter appearance rather than update positions to the desired location to fit the photometric loss. To address this, we adopt a coarse-to-fine training strategy that starts with isolated coarse alignment and advances to integrated fine-grained optimization to achieve robust human performance tracking and high-fidelity rendering.

Coarse Alignment. Upon initializing the joint Gaussians on the first frame, we fix the color, opacity, and scaling attributes to concentrate on capturing the human dynamic motions. In this phase, we solely fine-tune the motions of the joint Gaussians. Inspired by dynamic 3d Gaussian [Luiten et al. 2024], we employ a smooth

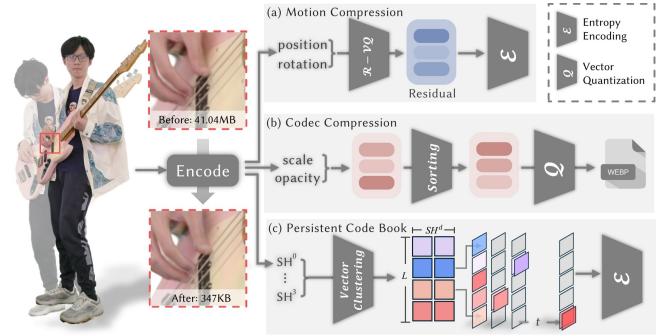


Fig. 4. Illustration of our hybrid compression strategy. We compress joint Gaussian motions using residual vector quantization, encode opacity and scaling via codec compression, and represent spherical harmonics with a persistent codebook. Our approach achieves a compression ratio of up to 120-fold.

regularizer to constrain the joint Gaussians motion locally as-rigid-as-possible(ARAP):

$$E_{\text{smooth}} = \sum_i \sum_{k \in N(i)} w_{i,k} \|R(q_{i,t}^j * q_{i,t-1}^{j-1})^{-1} (p_{k,t-1}^j - p_{i,t-1}^j) - (p_{k,t}^j - p_{i,t}^j)\|_2^2 \quad (5)$$

where $N(i)$ represents the set of neighboring joint Gaussian kernels of i , $R(\cdot)$ converts quaternion back into a rotation matrix and $w_{i,k}$ corresponds to the blending weights defined in Eq. 3. These weights remain fixed throughout the optimization process to avoid additional storage overhead. Additionally, following the original 3DGS [Kerbl et al. 2023], we incorporate the \mathcal{L}_1 photometric loss combined with a D-SSIM term during the coarse alignment process. The color energy is defined as:

$$E_{\text{color}} = (1 - \lambda_{\text{color}}) \mathcal{L}_1 + \lambda_{\text{color}} \mathcal{L}_{\text{D-SSIM}}, \quad (6)$$

the complete energy for coarse alignment is as follows:

$$E_{\text{coarse}} = \lambda_{\text{smooth}}^j E_{\text{smooth}} + E_{\text{color}}^j, \quad (7)$$

where E_{color}^j is computed by comparing the blended color after joint Gaussians rasterization with the ground truth input images.

Motion Prediction. To handle challenging motions, we further maintain a velocity attribute for each Gaussian and use the position changes between the latest two frames for weighted updates. Before the coarse alignment, we first estimate the new frame Gaussian positions based on the last one and velocities, then apply non-rigid constraint(ARAP) to restrict the unreasonable motions.

Fine-grained Optimization. We then optimize the motion of joint Gaussians and the appearance of skin Gaussians via the differentiable tracking and rendering process. Using the joint Gaussians motion from the coarse alignment phase, we interpolate the position and rotation of the skin Gaussians, balancing the rendering quality



Fig. 5. Examples of data captured by our multi-view system. Our DualGS dataset includes a diverse range of musical instruments from both Western and Eastern traditions.

and temporal consistency:

$$\begin{aligned} q_{i,t}^s &= \sum_{k \in N(p_{i,1}^s)} w(p_{i,1}^s, p_{k,1}^j) q_{k,t}^j, \\ p_{i,t}^s &= \sum_{k \in N(p_{i,1}^s)} w(p_{i,1}^s, p_{k,1}^j) (R(q_{k,t}^j) p_{i,1}^s + p_{k,t}^j), \end{aligned} \quad (8)$$

where $N(p_{i,1}^s)$ and $w(p_{i,1}^s, p_{k,1}^j)$ represent the precomputed KNN graph and blending weights from the initialization stage respectively. During backpropagation, the gradients on the skin Gaussians $q_{i,t}^s, p_{i,t}^s$ are further propagated along the computation graph to the joint Gaussians $q_{k,t}^j$ and $p_{k,t}^j$. Furthermore, to enhance temporal consistency, we incorporate a temporal regularization term inspired by HiFi4G [Jiang et al. 2024]. This term constrains the 4D Gaussian appearance attributes $(C_{i,t}, \sigma_{i,t}, s_{i,t})$ from undergoing significant updates between consecutive frames:

$$E_{\text{temp}} = \sum_{a \in \{C, \sigma, s\}} \lambda_a \|a_{i,t} - a_{i,t-1}\|_2^2, \quad (9)$$

E_{temp} efficiently improves the visual quality while enabling higher compression ratios in the subsequent later stage. We define the overall energy as follows:

$$E_{\text{fine}} = \lambda_{\text{smooth}}^s E_{\text{smooth}} + \lambda_{\text{temp}} E_{\text{temp}} + E_{\text{color}}^s, \quad (10)$$

We visualize the joint Gaussian kernels and the corresponding skin Gaussians rendering in Fig. 3. With the aid of our coarse-to-fine training strategy, DualGS efficiently achieves robust human performance tracking and high-fidelity rendering. Regarding our implementation, we first employ velocity prediction to initialize the motions, then conduct 10,000 iterations of training in each phase. The hyperparameters are set as follows: $\lambda_{\text{color}} = 0.2$, $\lambda_{\text{smooth}}^j = 0.05$, $\lambda_{\text{smooth}}^s = 0.001$, $\lambda_C = 1$, $\lambda_\sigma = 0.003$, $\lambda_s = 0.003$, $\lambda_{\text{temp}} = 0.00003$.

4 COMPRESSION

Our goal is to seamlessly integrate the high-quality 4D assets generated by DualGS into low-end devices with limited memory, such as head-mounted displays. For example, users can immersively navigate a musical odyssey in a VR environment. However, integrating such multiple volumetric videos (9 people in Fig. 1) totaling 2700 frames is non-trivial, requiring over 130GB of storage and even more runtime memory. Thanks to the effective disentanglement provided by DualGS, we organically compress the motion and appearance separately from joint Gaussians and skin Gaussians. Our strategy achieves a compression ratio of up to 120 times, while still enabling the decoding of high-fidelity rendering results in real-time. We first divide the sequences into multiple segments, with each segment consisting of f frames(50 in our setting).

Residual Compression. As mentioned in C3DGS [Niedermayr et al. 2024b] and CompGS [Liu et al. 2024], the precision of Gaussian position plays a crucial role in the quality of the scene, where even minor errors can severely impact rendering quality. Therefore, they opt for high-bit quantization. To address this, we first employ Residual-Vector Quantization(R-VQ) on the joint Gaussians motion. We retain the position of the first frame $R_{i,1} = p_{i,1}^j$ in the current segment, then perform temporal quantization(11-bit in our setting) as follows:

$$R_{i,t} = Q \left(p_{i,t}^j - (R_{i,1} + \sum_{k=2}^{t-1} Q^{-1}(R_{i,k})) \right), t > 1 \quad (11)$$

Q and Q^{-1} represent the quantization and dequantization respectively. We also apply R-VQ to the rotation q . Compared to solely quantizing adjacent frame residuals, our scheme effectively prevents error accumulation. We further employ Ranged Arithmetic Numerical System(RANS) encoding for lossless compression.

Codec Compression. Although we can apply residual compression to opacity and scaling parameters, the significantly larger number of skin Gaussians results in a notably higher storage requirement for compressed opacity and scaling. To achieve an optimal balance between data accuracy and storage overhead, we leverage the spatial-temporal relationships of these two Gaussian attributes. By benefiting from the temporal regularizer, we embed the opacity and scaling into separate Look-up Tables (LUT) and then apply image codec compression for encoding. Specifically, the opacity and scaling attributes are arranged into a 2D LUT, with the height corresponding to the number of skin Gaussians and the width corresponding to the segment frame length. To enhance 2D consistency, we further sort the LUT by the average value of each row. We then quantize and compress the 2D LUT using an image codec(WebP/JPEG), encoding it as an 8-bit image with a quality level of 100.

Persistent Code Book. The color attributes take up the majority of storage, occupying 48 out of 59 parameters. Effectively compressing them can yield significant storage savings. However, applying residual or codec compression to these coefficients still requires considerable storage overhead. To this end, we design a novel compression strategy – a persistent codebook that leverages the temporal consistency of skin Gaussian SH parameters, achieving up to a 360-fold compression. In particular, we apply K-Means clustering

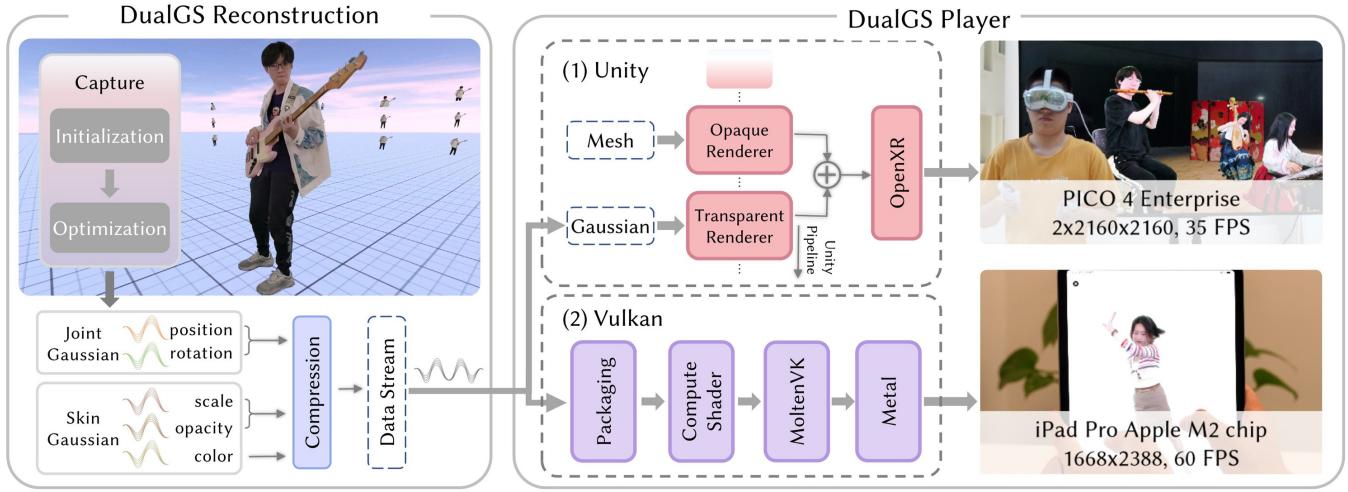


Fig. 6. Illustration of our DualGS player implementation for the seamless integration of 4D sequences into Unity and mobile platforms, enhancing real-time immersive rendering across multiple devices.

to the d-order ($d = 0, 1, 2, 3$) SH coefficients across all frames within this segment. The codebook \mathcal{Z}_d is initialized with a uniform distribution and iteratively updated by randomly selecting a batch of d -order coefficients. After optimization, we obtain four codebooks of length L (8192 in our setting). The skin Gaussian SH attributes are compactly encoded to SH indices via these codebooks:

$$\tau_{i,t}^d = \underset{k \in \{1, \dots, L\}}{\operatorname{argmin}} \left\| \mathcal{Z}_d[k] - C_{i,t}^d \right\|_2^2, \quad (12)$$

where $\tau_{i,t}^d$ is the d -order SH index for skin Gaussian i at frame t . We also can recover the compressed SH parameters $\hat{C}_{i,t}^d$ by indexing into the codebooks $\hat{C}_{i,t}^d = \mathcal{Z}_d[\tau_{i,t}^d]$. Using this representation, the SH attributes, originally consisting of $n \times f \times 48$ float parameters, are encoded as $n \times f \times 4$ integer indices along with four distinct codebooks, where n is the number of skin Gaussian. Furthermore, we observe that temporally coherent SH coefficients still maintain high consistency after being converted into indices. According to our calculation, on average, only one percent of the skin Gaussians SH indices change between adjacent frames. Therefore, instead of saving the spatial-temporal SH indices for each frame, we only save the first frame indices and the positions where the indices change in adjacent frames. Specifically, if $\tau_{i,t}^d \neq \tau_{i,t-1}^d$, we update it to the new index $\tau_{i,t}^d = k$ and save this integer quadruples (t, d, i, k) . In real-time decoding, we can instantly decode the spatial-temporal SH indices for each frame based on these quadruples. Additionally, the order of the quadruples does not affect the decoding process. We sort them in ascending order based on the first two variables and then apply length encoding.

5 IMPLEMENTATION

5.1 Dataset and Training Details

We utilize 81 Z-CAM cinema cameras to capture challenging human performances with a resolution of 3840×2160 at 30 fps under global illumination. To minimize motion blur during fast actions, all cameras are configured with a shutter speed of $640 \mu\text{s}$, ensuring

crisp and clear video quality. We showcase data examples in Fig. 5. Our DualGS dataset features 8 actors performing a wide range of musical instruments from both Western and Eastern traditions, such as violin, guitar, piano, flute, lute, and guzheng. Each sequence in the dataset starts with a standard pose to mitigate the close-to-open issue. These performances span various styles – from graceful classical melodies to contemporary pop music and vibrant subcultural pieces, providing a detailed portrayal of the performers' nimble finger movements and expressive demeanors. Additionally, for each instrument, the performers play pieces with slow, medium, and fast tempos, allowing us to demonstrate the robustness of our method across different motions. As illustrated in Fig. 7, DualGS enables robust tracking and high-fidelity rendering of human-centric volumetric video in real-time at high resolutions. For data pre-processing, we apply the background matting [Lin et al. 2021] to extract the foreground masks from all captured frames. Regarding DualGS optimization, due to limited GPU memory, we employ the per-frame training strategy. In the coarse alignment phase, we use the same learning rate as 3DGS [Kerbl et al. 2023]. For the Fine-grained phase, we reduce the learning rate and schedules by a factor of 10, resetting them at the beginning of each frame. We train the multi-view sequences on a single NVIDIA GeForce RTX3090, achieving a processing time of 12 minutes per frame. For rendering, DualGS adds an extra 10 ms for attribute decoding, memory copying, and skin Gaussians motion interpolation, achieving 77 fps for 4K rendering. Notably, the decoding process leverages CPU resources, running in parallel with Gaussian rasterization for acceleration.

5.2 DualGS player

As illustrated in Fig. 6, for the compressed data stream, we develop a companion Unity plugin that seamlessly integrates long-duration 4d sequences into standard CG engines and VR headsets, allowing conventional 3D rendering pipelines to efficiently deliver immersive environments. Additionally, we implement a DualGS player that enables real-time rendering on low-end mobile devices, offering a more user-friendly and interactive experience.

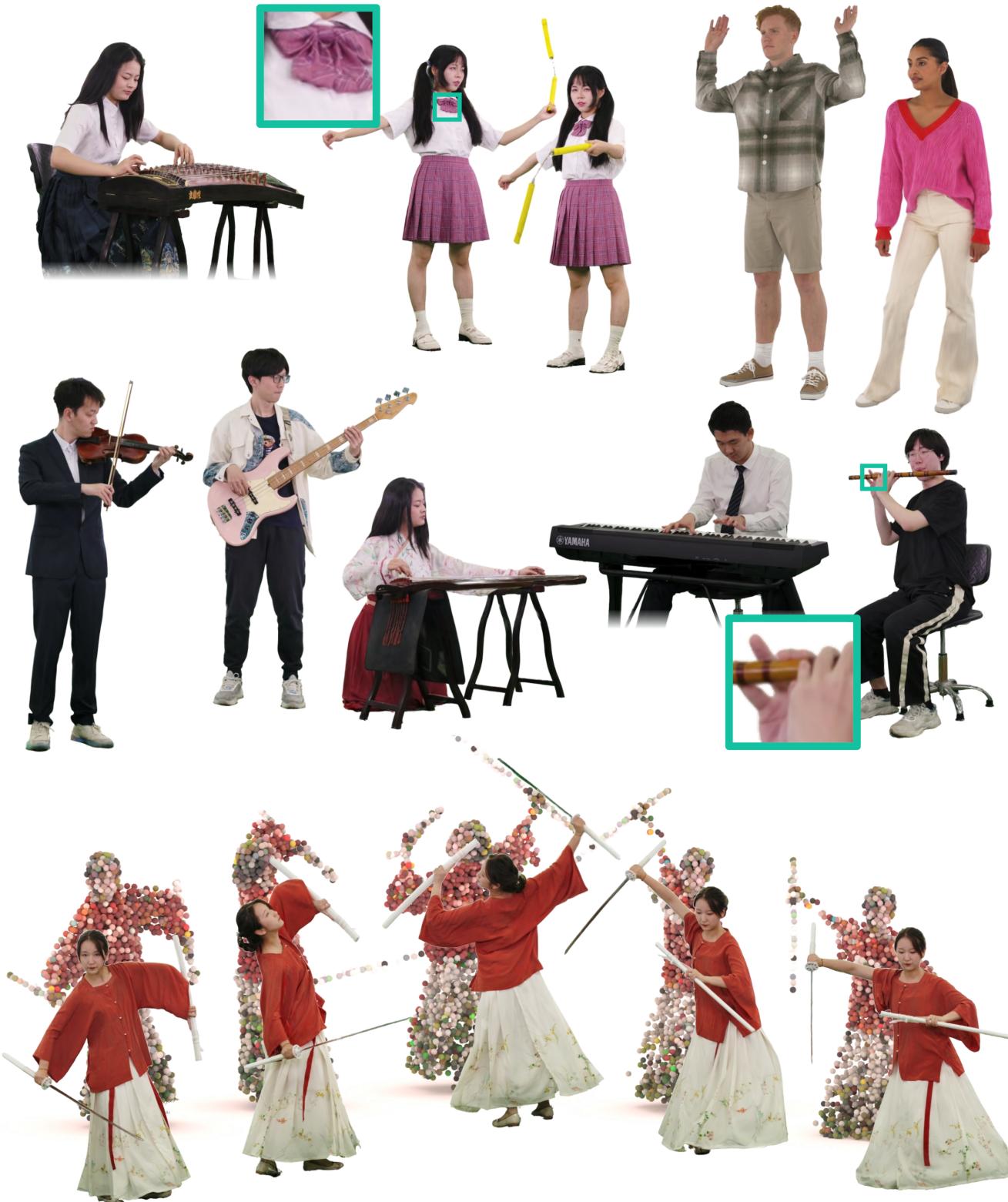


Fig. 7. We present a comprehensive results gallery showcasing our robust Dual Gaussian Splatting pipeline, featuring complex scenarios such as nunchuck swinging, musical instrument playing, and dancing. Additionally, we visualize dynamic sequences along with the corresponding joint Gaussians tracking. Even in the presence of challenging motions, DualGS achieves a 120-fold compression while maintaining real-time, high-fidelity rendering of human performances.



Fig. 8. Qualitative comparison of our method against HumanRF [Işık et al. 2023], NeuS2 [Wang et al. 2023a], Spacetime Gaussian [Li et al. 2024a] and HiFi4G [Jiang et al. 2024] on our challenging dataset. Our method achieves the highest rendering quality.

Unity Plugin. Based on the open-source Unity Renderer [aras p 2024], we implement a rendering plugin in Unity based on OpenXR that not only supports importing 4D assets generated by DualGS into Unity but also addresses shading and occlusions, seamlessly fusing the environment with Gaussian rasterization results in delivering an immersive, high-fidelity experience. Firstly, we decode the Gaussian point cloud of the current frame from the compressed data stream. Following the differentiable rasterization, the rendered images and corresponding alpha channels are stored in an additional texture buffer. This texture buffer is then combined with those produced by the standard mesh rendering pipeline, performing alpha blending from back to front to correctly handle occlusions. Leveraging Unity’s cross-platform capabilities, we can further stream the content to VR headsets, allowing users to experience immersive viewing in the virtual environment.

Gaussian Renderer. To enable high-fidelity dynamic rendering on low-end devices like iPhones and iPads, we developed a companion rendering application based on Vulkan [shg8 2024], removing the reliance on high-end GPU hardware. Our compression strategy explicitly divides sequences into multiple segments (50 frames per segment), allowing seamless playback of any length. Specifically, we employ a multi-threaded approach to parallelize data loading, decoding, and rendering. Once a frame is decoded, the Gaussian kernels are packaged into storage buffers compatible with compute shaders, which are then rendered directly to the swapchain image using alpha blending. The application is compatible with multiple platforms, including Windows, Linux, and Android. To extend support to iOS devices, such as Vision Pro, iPhones, and iPads, we leverage the MoltenVK library to map Vulkan API calls to Metal API. Within our DualGS player, users can drag, rotate, pause, and play the volumetric video, enhancing both accessibility and versatility across a wide range of devices.

Table 1. Quantitative comparison with SOTA dynamic rendering methods on our DualGS dataset. Green and yellow cell colors indicate the best and the second-best results.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	VMAF↑	Storage(MB / frame) ↓
HumanRF [Işık et al. 2023]	29.701	0.969	0.0461	79.171	7.566
NeuS2 [Wang et al. 2023a]	29.417	0.970	0.0593	77.912	24.163
Spacetime Gaussian [Li et al. 2024a]	29.532	0.964	0.0362	70.923	0.846
HiFi4G [Jiang et al. 2024]	33.503	0.988	0.0239	84.737	1.581
Ours(Before Compression)	35.577	0.990	0.0196	86.504	42.020
Ours(After Compression)	35.243	0.989	0.0221	86.171	0.323

6 EXPERIMENTS

6.1 Comparison

Rendering Comparison. We compare DualGS against SOTA implicit Instant NGP-based methods, HumanRF [Işık et al. 2023] and NeuS2 [Wang et al. 2023a] as well as explicit Gaussian-based methods Spacetime Gaussian [Li et al. 2024a] and HiFi4G [Jiang et al. 2024] using our captured dataset. As illustrated in Fig. 8, HumanRF [Işık et al. 2023] produces blurry results, whereas NeuS2 [Wang et al. 2023a] struggles with high-frequency details. Spacetime Gaussian [Li et al. 2024a] is prone to oversmoothing, losing fine details such as clothing wrinkles, while HiFi4G [Jiang et al. 2024] heavily relying on explicit mesh reconstruction and non-rigid tracking, generates severely unnatural results where deformation fails. Additionally, these methods exhibit artifacts or fail in rapid-motion areas. In contrast, our template-free DualGS leverages a dual Gaussian representation coupled with a tailored compression scheme for precise tracking and high-fidelity rendering. Our approach not only produces GPU-friendly and memory-efficient 4D assets but also demonstrably outperforms the compared methods. For quantitative comparison, we evaluate each method across three sequences, each consisting of 200 frames. In addition to traditional metrics such as PSNR, SSIM, and LPIPS, we introduce per-frame storage and VMAF [Li et al. 2016] to evaluate temporal consistency. As shown in Tab. 1, our method achieves the highest rendering quality and surpasses existing methods in compression efficiency.

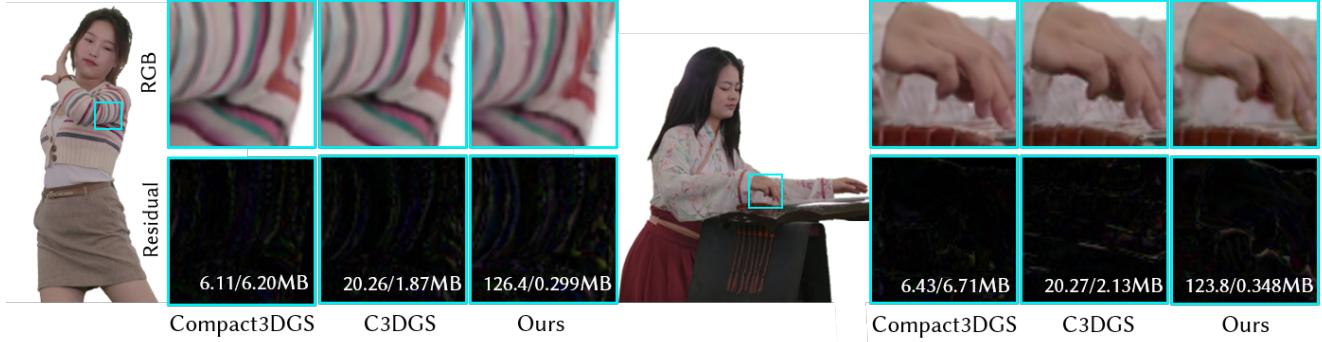


Fig. 9. Qualitative comparison of our method against Compact-3DGS [Lee et al. 2024], C3DGS [Niedermayr et al. 2024a] on our challenging dataset. we calculate the residual map between the predictions and ground truth. Our method achieves the highest compression ratio while maintaining comparable rendering quality.



Fig. 10. Qualitative evaluation of our Dual-Gaussian representation.

Table 2. Quantitative comparison with static compression methods.
Green and yellow cell colors indicate the best and the second-best results.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Running Time ↓	Storage(MB / frame) ↓
Compact-3DGS [Lee et al. 2024]	36.021	0.991	0.0193	14m 21s	6.576 (6.38x)
C3DGS [Niedermayr et al. 2024a]	35.823	0.991	0.0189	16m 09s	2.088 (20.11x)
Ours	35.243	0.989	0.0221	12m 12s	0.323 (122.76x)

Compression Comparison. We then compare our method with SOTA static compression methods, Compact-3DGS [Lee et al. 2024] and C3DGS [Niedermayr et al. 2024a]. Since we apply these methods to our human performance dataset, the resulting compression ratios differ from those reported in their respective papers for static scenes. In addition to rendering RGB images, we also calculate the residual map between the predictions and ground truth. As shown in Fig. 9, our DualGS compression achieves excellent storage efficiency while maintain the comparable rendering quality to other static Gaussian compression methods. A quantitative comparison is provided in Tab. 2, our method leverages spatial-temporal redundancy by using residual vector quantization to compress joint Gaussians motion, codec compression to encode the opacity and scaling of skin Gaussians, and a persistent codebook to represent spherical harmonics, achieving a compression ratio of up to 120 times and requiring only approximately 350KB of storage per frame.

Table 3. Compression Strategies on different attributes. Error is the mean absolute difference from the uncompressed data.

Method	Total size (KB)			Error		
	Motion	OP+Scale	SH	Motion	OP+Scale	SH
Raw(PLY)	462	2801	35573	0.0	0.0	0.0
Residual	96.3	362.3	1086.6	0.00124	0.20353	0.04265
Codec	27.44	126	226	0.01037	0.04661	0.00571
Codebook	38.187	658.36	98.76	0.03997	0.05519	0.01127

Table 4. Codebook sizes evaluation. Grey rows indicate our configurations.

Codebook Size	Metrics			
	PSNR ↑	SSIM ↑	LPIPS ↓	SIZE (KB) ↓
1024	35.447(-0.574)	0.99372(-0.000948)	0.02031(+0.00242)	54.8
2048	35.558(-0.463)	0.99386(-0.000807)	0.01999(+0.00210)	63.4
4096	35.653(-0.368)	0.99396(-0.000708)	0.01968(+0.00179)	77.3
8192	35.729(-0.292)	0.99404(-0.000628)	0.01944(+0.00155)	99.4
16384	35.771(-0.250)	0.99413(-0.000538)	0.01914(+0.00125)	152.09

6.2 Ablations

Dual Gaussian Representation. We conduct a qualitative ablation study on the dual Gaussian representation to evaluate its efficacy. As shown in Fig. 10, omitting the velocity prediction(w/o velocity) and relying solely on the smoothing term leads to inaccurate tracking during fast motions. Furthermore, the exclusion of the joint Gaussians(w/o joint) and optimizing all attributes of the skin Gaussians



Fig. 11. We develop a VR demo to showcase how users can immerse themselves in a virtual musical odyssey, standing beside musicians, observing their performances up close, and feeling the rhythm of the music.

produces severe artifacts and loses temporal consistency. Additionally, omitting the coarse alignment stage(w/o coarse) introduces noticeable artifacts, and excluding the fine-grained optimization(w/o fine) yields unnatural outputs due to the fixed appearances, despite relatively accurate motion capture. In contrast, our full pipeline significantly enhances tracking accuracy, ensuring temporal consistency and achieving high-fidelity rendering.

Hybrid Compression Strategy. As shown in Tab. 3, we evaluate three compression strategies for different attributes. For the motion attributes (position and rotation) of joint Gaussians, residual compression provides the highest precision with acceptable size. For the opacity and scaling of skin Gaussians, codec compression achieves optimal precision with minimal storage requirements. For

the storage-intensive SH attribute, we balance precision with storage efficiency by using a persistent codebook for compression.

Codebook Size. Spatial-temporal SH coefficients are compressed into persistent codebooks of predefined sizes. As shown in Tab. 4, enlarging the codebook size beyond 8,192 yields little effect in compression efficiency, while significantly increasing storage consumption. Consequently, we keep the storage overhead of SH coefficients at levels comparable to those required for storing XYZ coordinates.

6.3 Immersive Experiences

In Fig. 11, we showcase the application of watching a high-fidelity virtual concert using the PICO 4 VR headset. Viewers can immerse themselves in a virtual musical experience, observing musicians

perform up close as if standing beside them, even though the musicians are located in various places around the world. We can capture, process, and generate 4D assets at different times, integrating individual performers into a consistent environment within a standard CG engine. Additionally, we can edit their positions and align their actions temporally, ensuring that their performances are visually synchronized with the rhythm of the ensemble.

6.4 Limitations and Discussions

Our approach achieves template-free dynamic human modeling via the disentangled and hierarchical motion and appearance representation. With the specially designed compression strategy, we achieve a 120-fold compression ratio and still deliver accurate tracking and high-fidelity rendering for immersive experiences. Despite such compelling capabilities, our pipeline still yields some limitations. We provide detailed analysis and discuss potential future extensions.

Firstly, our method relies on image-based accurate segmentation to separate the foreground human performances, which may yield segmentation errors with slender objects such as lute strings and hair, compromising detailed tracking. It is an interesting direction to incorporate the view-consistent matting and 3D/4D understanding to enhance 4D modeling and rendering. Moreover, although our DualGS representation avoids explicit mesh reconstruction and non-rigid tracking, it still requires more training time compared to real-time tracking. The bottleneck lies in the coarse alignment and fine-grained optimization. To further accelerate the process, we observe that using 180,000 skin Gaussians to represent human appearance may introduce redundancy. A potential solution is to employ LightGaussian [Fan et al. 2023] to prune unimportant Gaussian kernels, thereby speeding up the rendering process as well. To ensure efficient motion tracking and temporal consistency, we fix the joint-skin KNN relationship after initialization. However, this sacrifices the ability to handle topological changes. Combing dynamic graph [Xu et al. 2019a] and keyframe strategy [Dou et al. 2016] may address this issue. Our method can produce vivid volumetric videos without relying on human parametric models or skeleton information. However, it does not support downstream tasks such as animatable avatar or motion transfer. Annotating 4D sequences and driving the 4D assets using multimodal inputs, such as text prompts, music, or human skeleton, is promising. Moreover, although we already integrate our 4D assets into standard CG engines, the lack of geometric or normal information prevents them from being re-lit under different lighting conditions, which presents an interesting avenue for future research.

7 CONCLUSION

We have presented a comprehensive solution for producing high-fidelity, human-centric volumetric video. Our core approach is based on a dual-Gaussian representation for challenging human performance, enabling accurate tracking and high-fidelity rendering. By organically combining a compact number of motion-aware joint Gaussians to capture global movements with a larger set of appearance-aware skin Gaussians for visual details, we adeptly manage challenging motions without sacrificing quality. For Dual-Gaussian initialization, we utilize a uniform random point cloud to initialize the

joint Gaussians and carefully control their number and scale. These joint Gaussians serve as the foundation for initializing the skin Gaussians and constructing the KNN field for subsequent optimization. Furthermore, we propose a coarse-to-fine training strategy to reduce optimization difficulty. To integrate long volumetric video sequences into VR platforms, we have developed a DualGS-based compression strategy to achieve a 120-fold compression ratio. We also implement a companion Unity plugin for hybrid rendering with a standard CG immersive environment as well as DualGS player that enables high-quality rendering on low-end mobile devices. Experimental results demonstrate that our method vividly produces high-quality renderings. We believe our method serves as a "ticket" to a virtual world, offering immersive and high-fidelity experiences. **Acknowledgements.** We thank reviewers for their feedback. This work was supported by NSFC programs (61976138), National Key R&D Program of China (2022YFF0902301), STCSM (2015F0203-000-06), Shanghai Local college capacity building program (22010502800). We also acknowledge support from Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI).

REFERENCES

- Marc Alexa and Wolfgang Müller. 2000. Representing animations by principal components. In *Computer Graphics Forum*, Vol. 19. Wiley Online Library, 411–418.
- hybridherbst pastasfuture JasonDeacutis aras p, b0nes164. 2024. UnityGaussianSplatting. <https://github.com/arasz-p/UnityGaussianSplatting>.
- Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*. Springer, 333–350.
- Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. 2021. SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11594–11604.
- Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. 2024a. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*. 1–9.
- Yihang Chen, Qianyi Wu, Jianfei Cai, Mehrtash Harandi, and Weiyo Lin. 2024b. HAC: Hash-grid Assisted Context for 3D Gaussian Splatting Compression. *arXiv preprint arXiv:2403.14530* (2024).
- Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 69.
- Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. 2017. Motion2Fusion: Real-time Volumetric Performance Capture. *ACM Trans. Graph.* 36, 6, Article 246 (Nov. 2017), 16 pages.
- 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. 2016. Fusion4D: real-time performance capture of challenging scenes. *ACM Trans. Graph.* 35, 4, Article 114 (jul 2016), 13 pages. <https://doi.org/10.1145/2897824.2925969>
- Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, and Zhangyang Wang. 2023. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. *arXiv preprint arXiv:2311.17245* (2023).
- Sharath Girish, Kamal Gupta, and Abhinav Shrivastava. 2023. Eagles: Efficient accelerated 3d gaussians with lightweight encodings. *arXiv preprint arXiv:2312.04564* (2023).
- Sumit Gupta, Kunal Sengupta, and Ashraf A Kassim. 2002. Compression of dynamic 3d geometry data using iterative closest point algorithm. *Computer Vision and Image Understanding* 87, 1–3 (2002), 116–130.
- Marc Habermann, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhoefer, and Christian Theobalt. 2023. Hdhumans: A hybrid approach for high-fidelity digital humans. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 6, 3 (2023), 1–23.
- Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. 2021a. Real-time Deep Dynamic Characters. *ACM Transactions on Graphics* 40, 4, Article 94 (aug 2021).
- Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. 2021b. Real-time deep dynamic characters. *ACM Transactions on Graphics (ToG)* 40, 4 (2021), 1–16.

- Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. 2019. LiveCap: Real-Time Human Performance Capture From Monocular Video. *ACM Transactions on Graphics (TOG)* 38, 2, Article 14 (2019), 17 pages.
- Shoukang Hu, Tao Hu, and Ziwei Liu. 2024. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20418–20431.
- Wenbo Hu, Yuling Wang, Lin Ma, Bangbang Yang, Lin Gao, Xiao Liu, and Yuewen Ma. 2023. Tri-MipRF: Tri-Mip Representation for Efficient Anti-Aliasing Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19774–19783.
- Lorenzo Lawrence Ibarria and Jaroslaw R Rossignac. 2003. *Dynapack: space-time compression of the 3D animations of triangle meshes with fixed connectivity*. Technical Report. Georgia Institute of Technology.
- Mustafa İşik, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. 2023. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–12. <https://doi.org/10.1145/3592415>
- Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. 2016. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European Conference on Computer Vision*. Springer, 362–379.
- Rohit Jena, Ganesh Subramanian Iyer, Siddharth Choudhary, Brandon Smith, Pratik Chaudhari, and James Gee. 2023. SplatArmor: Articulated Gaussian splatting for animatable humans from monocular RGB videos. *arXiv preprint arXiv:2311.10812* (2023).
- Tianjian Jiang, Xu Chen, Jie Song, and Ottmar Hilliges. 2023a. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16922–16932.
- Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, and Lan Xu. 2022. NeuralHOFusion: Neural Volumetric Rendering Under Human-Object Interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6155–6165.
- Yuheng Jiang, Zhehao Shen, Penghao Wang, Zhuo Su, Yu Hong, Yingliang Zhang, Jingyi Yu, and Lan Xu. 2024. Hifi4g: High-fidelity human performance rendering via compact gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19734–19745.
- Yuheng Jiang, Kaixin Yao, Zhuo Su, Zhehao Shen, Haimin Luo, and Lan Xu. 2023b. 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*. 595–605.
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)* 42, 4 (2023), 1–14.
- Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. 2024. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 505–515.
- Youngjoong Kwon, Lingjie Liu, Henry Fuchs, Marc Habermann, and Christian Theobalt. 2024. Deliffas: Deformable light fields for fast avatar synthesis. *Advances in Neural Information Processing Systems* 36 (2024).
- Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. 2024. Compact 3D Gaussian Representation for Radiance Field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21719–21728.
- Mengtian Li, Shengxiang Yao, Zhifeng Xie, Keyu Chen, and Yu-Gang Jiang. 2024b. Gaussianbody: Clothed human reconstruction via 3d gaussian splatting. *arXiv preprint arXiv:2401.09720* (2024).
- Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. 2022. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision*. Springer, 419–436.
- Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194–1.
- Yue Li, Marc Habermann, Bernhard Thomaszewski, Stelian Coros, Thabo Beeler, and Christian Theobalt. 2021. Deep physics-aware inference of cloth deformation for monocular human performance capture. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 373–384.
- Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, Megha Manohara, et al. 2016. Toward a practical perceptual video quality metric. *The Netflix Tech Blog* 6, 2 (2016), 2.
- Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. 2024a. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8508–8520.
- Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. 2024c. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19711–19722.
- Haotong Lin, Sida Peng, Zhen Xu, Tao Xie, Xingyi He, Hujun Bao, and Xiaowei Zhou. 2023. High-Fidelity and Real-Time Novel View Synthesis for Dynamic Scenes. In *SIGGRAPH Asia 2023 Conference Papers* (, Sydney, NSW, Australia,) (SA '23). Association for Computing Machinery, New York, NY, USA, Article 15, 9 pages. <https://doi.org/10.1145/3610548.3618142>
- Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2022. Efficient Neural Radiance Fields for Interactive Free-viewpoint Video. In *SIGGRAPH Asia Conference Proceedings*.
- Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2021. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8762–8771.
- Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. 2021. Neural Actor: Neural Free-view Synthesis of Human Actors with Pose Control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)* (2021).
- Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. 2020. Neural Human Video Rendering by Learning Dynamic Textures and Rendering-to-Video Translation. *IEEE Transactions on Visualization and Computer Graphics* PP (05 2020), 1–1. <https://doi.org/10.1109/TVCG.2020.2996594>
- Xiangrui Liu, Xinju Wu, Pingping Zhang, Shiqi Wang, Zhu Li, and Sam Kwong. 2024. CompGS: Efficient 3D Scene Representation via Compressed Gaussian Splatting. *arXiv preprint arXiv:2404.09458* (2024).
- Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-person Linear Model. *ACM Trans. Graph.* 34, 6, Article 248 (Oct. 2015), 16 pages.
- Tao Lu, Mulin Yu, Linnings Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. 2024. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20654–20664.
- Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. 2024. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. In *3DV*.
- Guoliang Luo, Frederic Cordier, and Hyewon Seo. 2013. Compression of 3D mesh sequences by temporal segmentation. *Computer Animation and Virtual Worlds* 24, 3-4 (2013), 365–375.
- Diogo Luvizton, Vladislav Golyanik, Adam Kortylewski, Marc Habermann, and Christian Theobalt. 2023. Relightable Neural Actor with Intrinsic Decomposition and Pose Control. *arXiv preprint arXiv:2312.11587* (2023).
- Khaled Mamou, Titus Zaharia, and Françoise Prêteux. 2009. TFAN: A low complexity 3D mesh compression algorithm. *Computer Animation and Virtual Worlds* 20, 2-3 (2009), 343–354.
- Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 405–421.
- Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. 2024. Human gaussian splatting: Real-time rendering of animatable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 788–798.
- Wieland Morgenstern, Florian Barthel, Anna Hilsmann, and Peter Eisert. 2023. Compact 3D Scene Representation via Self-Organizing Gaussian Grids. *arXiv preprint arXiv:2312.13299* (2023).
- Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* 41, 4, Article 102 (July 2022), 15 pages. <https://doi.org/10.1145/3528223.3530127>
- KL Navaneet, Kossar Pourahmadi Meibodi, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. 2023. Compact3D: Compressing Gaussian Splat Radiance Field Models with Vector Quantization. *arXiv preprint arXiv:2311.18159* (2023).
- Richard A Newcombe, Dieter Fox, and Steven M Seitz. 2015. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 343–352.
- Simon Niedermayr, Josef Stumpfegger, and Rüdiger Westermann. 2024a. Compressed 3D Gaussian Splatting for Accelerated Novel View Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10349–10358.
- Simon Niedermayr, Josef Stumpfegger, and Rüdiger Westermann. 2024b. Compressed 3d gaussian splatting for accelerated novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10349–10358.
- Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. 2024. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1165–1175.
- Panagiotis Papantoniakis, Georgios Kopanas, Bernhard Kerbl, Alexandre Lanvin, and George Drettakis. 2024. Reducing the Memory Footprint of 3D Gaussian Splatting. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 7, 1 (May 2024), 1–12.

- 2024). <https://repo-sam.inria.fr/fungraph/reduced-3dgs/>
- Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10318–10327.
- Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. 2024a. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20299–20309.
- Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 2024b. 3dgs-avat: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5020–5030.
- Christian Reiser, Rick Szeliski, Dor Verbin, Pratul Srinivasan, Ben Mildenhall, Andreas Geiger, Jon Barron, and Peter Hedman. 2023. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–12.
- Ruizhi Shao, Liliang Chen, Zerong Zheng, Hongwen Zhang, Yuxiang Zhang, Han Huang, Yandong Guo, and Yebin Liu. 2022. Floren: Real-time high-quality human performance rendering via appearance flow using sparse rgb cameras. In *SIGGRAPH Asia 2022 Conference Papers*. 1–10.
- Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Ottmar Hilliges. 2023. X-avat: Expressive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16911–16921.
- Ashwath Shetty, Marc Habermann, Guoxing Sun, Diogo Luvizou, Vladislav Golyanik, and Christian Theobalt. 2024. Holoported Characters: Real-time Free-viewpoint Rendering of Humans from Sparse RGB Cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1206–1215.
- cedric-chedaleux shg8. 2024. 3DG.cpp. <https://github.com/shg8/3DG.cpp>.
- Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. 2017. Killing-fusion: Non-rigid 3d reconstruction without correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1386–1395.
- Miroslava Slavcheva, Maximilian Baust, and Slobodan Ilic. 2018. Sobolevfusion: 3d reconstruction of scenes undergoing free non-rigid motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2646–2655.
- Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. 2023. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2732–2742.
- Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. 2020. RobustFusion: Human Volumetric Capture with Data-Driven Visual Cues Using a RGBD Camera. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 246–264.
- Zhuo Su, Lan Xu, Dawei Zhong, Zhong Li, Fan Deng, Shuxue Quan, and Lu Fang. 2022. Robustfusion: Robust volumetric performance reconstruction under human-object interactions from monocular rgbd stream. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- Robert W Sumner, Johannes Schmid, and Mark Pauly. 2007. Embedded deformation for shape manipulation. *ACM Transactions on Graphics (TOG)* 26, 3 (2007), 80.
- Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingya Wang, and Jingyi Yu. 2021. Neural Free-Viewpoint Performance Rendering under Complex Human-object Interactions. In *Proceedings of the 29th ACM International Conference on Multimedia*.
- Xin Suo, Yuheng Jiang, Pei Lin, Yingliang Zhang, Minye Wu, Kaiwen Guo, and Lan Xu. 2021. NeuralHumanFVV: Real-Time Neural Volumetric Human Performance Rendering using RGB Cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6226–6237.
- Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. 2022. Compressible-composable nerf via rank-residual decomposition. *Advances in Neural Information Processing Systems* 35 (2022), 14798–14809.
- Eduard Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. 2021. 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.
- Luboš Vasa and Václav Skala. 2007. Coddyac: Connectivity driven dynamic mesh compression. In *2007 3DTV Conference*. IEEE, 1–4.
- Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. 2009. Dynamic shape capture using multi-view photometric stereo. In *ACM SIGGRAPH Asia 2009 papers*. 1–11.
- Henan Wang, Hanxin Zhu, Tianyu He, Runsen Feng, Jiajun Deng, Jiang Bian, and Zhibo Chen. 2024b. End-to-End Rate-Distortion Optimized 3D Gaussian Representation. *arXiv preprint arXiv:2406.01597* (2024).
- Liao Wang, Qiang Hu, Qihan He, Ziyu Wang, Jingyi Yu, Tinne Tuytelaars, Lan Xu, and Minye Wu. 2023b. Neural Residual Radiance Fields for Streamably Free-Viewpoint Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 76–87.
- Liao Wang, Kaixin Yao, Chengcheng Guo, Zhirui Zhang, Qiang Hu, Jingyi Yu, Lan Xu, and Minye Wu. 2024a. VideoRF: Rendering Dynamic Radiance Fields as 2D Feature Video Streams. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 470–481.
- Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zhang, Minye Wu, Jingyi Yu, and Lan Xu. 2022b. Fourier plenoctrees for dynamic radiance field rendering in real-time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13524–13534.
- Shaofei Wang, Andreas Geiger, and Siyu Tang. 2021. Locally Aware Piecewise Transformation Fields for 3D Human Mesh Registration. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. 2022a. ARAH: Animatable Volume Rendering of Articulated Human SDFs. In *European Conference on Computer Vision*.
- Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. 2023a. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3295–3306.
- Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. 2022. HumanNeRF: Free-Viewpoint Rendering of Moving People From Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16210–16220.
- Guanjun Wu, Taoran Yu, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 2024b. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20310–20320.
- Minye Wu, Zehao Wang, Georgios Kouros, and Tinne Tuytelaars. 2024a. TeTriRF: Temporal Tri-Plane Radiance Fields for Efficient Free-Viewpoint Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6487–6496.
- Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. 2021. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9421–9431.
- 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. 2022. Dressing Avatars: Deep Photorealistic Appearance for Physically Simulated Clothing. *ACM Trans. Graph.* 41, 6, Article 222 (nov 2022), 15 pages. <https://doi.org/10.1145/3550454.3555456>
- Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. 2020. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *2020 International Conference on 3D Vision (3DV)*. IEEE, 322–332.
- Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. 2024. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4389–4398.
- Lan Xu, Wei Cheng, Kaiwen Guo, Lei Han, Yebin Liu, and Lu Fang. 2019a. Flyfusion: Realtime dynamic scene reconstruction using a flying depth camera. *IEEE transactions on visualization and computer graphics* 27, 1 (2019), 68–82.
- Lan Xu, Zhuo Su, Lei Han, Tao Yu, Yebin Liu, and Lu Fang. 2019b. UnstructuredFusion: realtime 4D geometry and texture reconstruction using commercial RGBD cameras. *IEEE transactions on pattern analysis and machine intelligence* 42, 10 (2019), 2508–2522.
- Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 2024. 4k4d: Real-time 4d view synthesis at 4k resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20029–20040.
- Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. 2024. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20331–20341.
- Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. 2021a. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5752–5761.
- Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. 2021b. 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*. 5746–5756.
- Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. 2019. 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).
- Hongwen Zhang, Siyu Lin, Ruizhi Shao, Yuxiang Zhang, Zerong Zheng, Han Huang, Yandong Guo, and Yebin Liu. 2023. CloSET: Modeling Clothed Humans on Continuous Surface with Explicit Template Decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Jiakai Zhang, Liao Wang, Xinhang Liu, Fuqiang Zhao, Minzhang Li, Haizhao Dai, Boyuan Zhang, Wei Yang, Lan Xu, and Jingyi Yu. 2022. NeuVV: Neural Volumetric Videos with Immersive Rendering and Editing. *arXiv preprint arXiv:2202.06088* (2022).

- Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, and Jingyi Yu. 2022a. Human Performance Modeling and Rendering via Neural Animated Mesh. *ACM Trans. Graph.* 41, 6, Article 235 (nov 2022), 17 pages. <https://doi.org/10.1145/3550454.3555451>
- Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. 2022b. HumanNeRF: Efficiently Generated Human Radiance Field from Sparse Inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7743–7753.
- Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. 2024. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19680–19690.
- Heming Zhu, Fangneng Zhan, Christian Theobalt, and Marc Habermann. 2023. TriHuman: A Real-time and Controllable Tri-plane Representation for Detailed Human Geometry and Appearance Synthesis. *arXiv preprint arXiv:2312.05161* (2023).
- Wojciech Zienonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. 2023. Drivable 3d gaussian avatars. *arXiv preprint arXiv:2311.08581* (2023).