

EVA-Gaussian: 3D Gaussian-based Real-time Human Novel View Synthesis under Diverse Multi-view Camera Settings

Yingdong Hu¹ Zhening Liu¹ Jiawei Shao² Zehong Lin^{1*} Jun Zhang¹

¹ The Hong Kong University of Science and Technology

² Institute of Artificial Intelligence (TeleAI), China Telecom

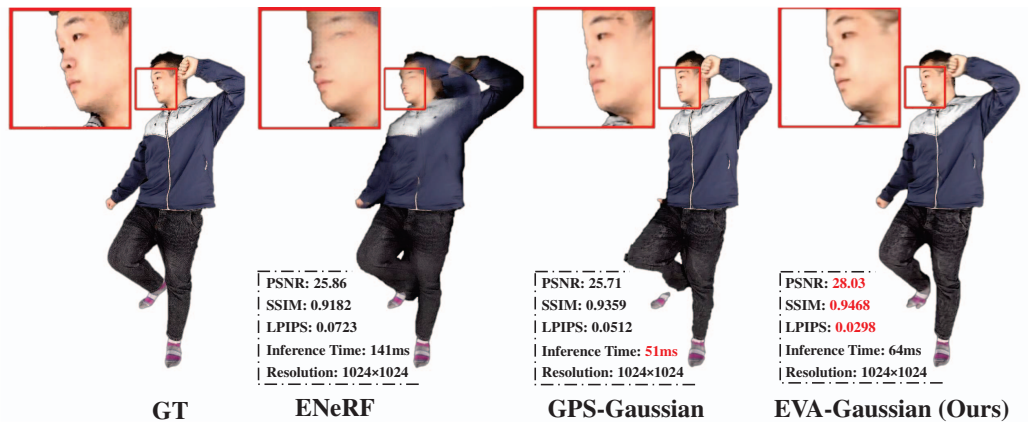


Figure 1. Qualitative comparison of novel view synthesis on the THuman2.0 dataset, with the angle between the stereo views being 72 degree and GT representing the ground truth. We compare our proposed EVA-Gaussian against the state-of-the-art approaches GPS-Gaussian [37] and ENeRF [11]. The quantitative metrics of PSNR \uparrow , SSIM \uparrow , LPIPS \downarrow , and inference time \downarrow demonstrate that EVA-Gaussian achieves superior reconstruction quality, while enabling real-time reconstruction under sparse-view conditions and high-resolution settings.

Abstract

Feed-forward based 3D Gaussian Splatting methods have demonstrated exceptional capability in real-time novel view synthesis for human models. However, current approaches are confined to either dense viewpoint configurations or restricted image resolutions. These limitations hinder their flexibility in free-viewpoint rendering across a wide range of camera view angle discrepancies, and also restrict their ability to recover fine-grained human details in real time using commonly available GPUs. To address these challenges, we propose a novel pipeline named EVA-Gaussian for 3D human novel view synthesis across diverse multi-view camera settings. Specifically, we first design an Efficient Cross-View Attention (EVA) module to effectively fuse cross-view information under high resolution inputs and sparse view settings, while minimizing temporal and computational overhead. Additionally, we introduce a feature refinement mechanism to predict the attributes of the 3D Gaussians and assign a feature value to each Gaussian, enabling the correction of artifacts caused by geometric in-

accuracies in position estimation and enhancing overall visual fidelity. Experiments on the THuman2.0 and THuman-sit datasets showcase the superiority of EVA-Gaussian in rendering quality across diverse camera settings.

1. Introduction

3D reconstruction and novel view synthesis have long been fundamental yet complex tasks in visual data representation and computer vision. Recent advancements in fast 3D reconstruction and novel view synthesis for humans have shown immense potential in applications such as holographic communication, real-time teaching, and augmented/virtual reality (AR/VR), where time efficiency and the clarity of human body representations are critical for delivering satisfactory user experiences. Nonetheless, existing methods either rely on dense input views and precise templates as prior knowledge [5, 7–9, 14, 22, 30] or are restricted to specific camera poses [27, 37]. None of these approaches has fully developed a pipeline for real-time human reconstruction under diverse, especially sparse, camera viewpoints while maintaining high-resolution input quality. In recent years, Neural Radiance Fields (NeRFs) [18]

*Corresponding author, email: eezhlin@ust.hk.

have emerged as a promising technique for 3D reconstruction. These models employ neural networks to predict the color and density of sampled 3D points along camera rays and aggregate these predictions to synthesize novel images with high fidelity. Despite their effectiveness, NeRFs suffer from substantial time consumption during both the training and rendering phases. Although various advancements, such as multi-resolution hash encoding [19] and feed-forward neural scene prediction [32, 33], have been achieved to mitigate these inefficiencies, the achievable speeds remain insufficient for real-time applications.

More recently, 3D Gaussian Splatting (3DGS) [6] has been introduced as a solution to the rendering bottleneck. 3DGS utilizes a set of discrete Gaussian representations to model complex 3D scenes and leverages the α -blending technique to enable real-time novel view synthesis. In the field of 3D human avatar reconstruction, previous works [5, 7–9, 22, 30] have employed 3DGS as a representation for humans to achieve animatable full-body human avatar reconstruction. These methods, however, rely on precise human templates as priors, and their training and reconstruction processes can span from minutes to hours, which impedes their use in real-time applications such as holographic communication. While a feed-forward human reconstruction method [37] has achieved fast reconstruction and real-time rendering with stereo inputs at a resolution of 1024, its reliance on stereo settings and limited camera angle variations restricts its reconstruction quality under sparse camera settings and leads to sub-optimal performance when more than two input views are utilized.

To address these limitations and enable high-resolution real-time 3D reconstruction of humans across diverse camera positions and varying numbers of cameras, we propose *EVA-Gaussian*, a novel 3D Gaussian-based pipeline for real-time human novel view synthesis. Our method attaches 3D Gaussians to the surface of the human body through multi-view depth estimation and aligns their positions closely with point cloud locations. A key innovation is the introduction of an Efficient cross-View Attention (EVA) module for multi-view 3D Gaussian position estimation (see Sec. 3.2). Specifically, we employ a U-Net [23] as the backbone and incorporate a dedicated window-embedded cross-view attention mechanism to infer multi-view position correspondences. This design leverages the observation that, in most human capture systems, human subjects are typically centered within each frame. As a result, the epipolar lines tend to align nearly parallel to the x-axis of the image, which enables us to significantly reduce the memory for multi-view attention while maintaining accuracy under large view discrepancies. Besides, we incorporate a Gaussian attribute estimation module that takes the EVA output and the original RGB images as input to estimate the remaining 3D Gaussian attributes (see Sec. 3.3).

Furthermore, we embed an additional attribute, referred to as feature, into each Gaussian for further feature splatting and image quality refinement, thereby mitigating the position estimation errors introduced by the EVA module (see Sec. 3.4). In addition, we introduce an anchor loss to regularize the opacity and scale attributes of the 3D Gaussians (see Sec. 3.5). This ensures consistency between the point cloud depth and the 3D Gaussian position map, enhancing overall stability and accuracy. We conduct extensive experiments on the THuman2.0 [34] and THumanSit [35] datasets. The results, as exemplified in Fig. 1, demonstrate that *EVA-Gaussian* outperforms existing feed-forward synthesis approaches in rendering quality, while enabling real-time reconstruction and rendering. Moreover, our approach generalizes well to settings with varying numbers of cameras and significant changes in camera viewpoint angles. Our main contributions are summarized as follows:

- We propose *EVA-Gaussian*, a novel pipeline for fast feed-forward 3D human reconstruction that comprises three main stages: 1) 3D Gaussian position estimation, 2) 3D Gaussian attributes estimation, and 3) feature refinement.
- We introduce an EVA module to enhance multi-view correspondence retrieval, leading to improved 3D Gaussian position estimation and enhanced novel view synthesis under diverse view numbers and sparse camera settings.
- We embed a feature value to each 3D Gaussian through a feed-forward neural network and employ a recurrent feature refiner that fuses splatted images and feature maps to mitigate artifacts caused by position estimation errors.
- Extensive experiments on the THuman2.0 and THumanSit datasets demonstrate the effectiveness and superiority of our proposed pipeline over existing methods in terms of rendered novel view quality and inference speed, especially under sparse camera settings.

2. Related Works

Instant Human Reconstruction. Recovering a 3D human model from sparse view inputs and generating novel views in real time has been a long-standing challenge in computer vision and graphics. Despite continuous advancements, achieving both accuracy and efficiency remains an unresolved problem. PIFu [24] is one of the first methods to successfully reconstruct the human surface and its color map from one or several RGB images, leveraging the strong spatial encoding capabilities of implicit function representations. However, it remains constrained by low speed and resolution. Subsequent works attempt to address these limitations by employing specific rendering methods [10, 25] or incorporating additional depth information [34]. Nonetheless, their reliance on traditional 3D representations results in suboptimal rendering quality and efficiency.

3DGS-based Human Avatar Reconstruction. 3D Gaussian Splatting has recently emerged as an effective tech-

nique for 3D human reconstruction. However, most previous works [5, 7, 9, 20, 22, 30] bind 3D Gaussians to a predefined human mesh model, such as SMPL [15] or SMPL-X [21]. This approach generates 3D Gaussians and human models in a canonical space and then transforms them to match the target human pose using predefined weights. This iterative binding process, however, is extremely time-consuming. Moreover, these methods require human templates as inputs at each frame, which incurs extra computational costs and potentially misleads the reconstruction due to errors in pose estimation. These limitations significantly hinder their applicability in real-world scenarios.

Fast Generalizable 3D Reconstruction. In the field of NeRF rendering, pixelNeRF [33] pioneers the approach of predicting per-pixel features from a single image in a feed-forward manner for 3D reconstruction. While subsequent works [3, 11, 28] have followed this feed-forward NeRF pipeline, they still suffer from the extensive time consumption of the NeRF rendering process. Besides, their reconstruction results are often unsatisfactory under sparse camera settings. The introduction of 3DGS has helped mitigate the rendering speed issue of high-quality novel view synthesis. Notably, pixelSplat [2] and Splatter Image [26] are the first to combine feed-forward inference with 3DGS, which predict 3D Gaussian attributes for each pixel and project them back into the 3D space for real-time novel view synthesis. Nevertheless, they still struggle with inaccurate estimation of Gaussian positions. MVSplat [4] and MVGaussian [12] address this issue by leveraging cross-attention mechanisms and cost-volume modules, thereby achieving superior novel view quality. Moreover, latentSplat [31] attaches a latent vector to each 3D Gaussian and refines novel views through a diffusion decoder and generative loss, significantly improving image quality in extrapolation views. Despite these advancements, existing methods fail to fully exploit prior knowledge about human images and camera settings, which limits their performance on real-time human reconstruction and novel view synthesis.

The work most closely related to ours is GPS-Gaussian [37], which proposes a stereo matching network for 3D Gaussian position estimation and employs two 3-layer U-Nets to predict 3D Gaussian scales, rotations, and opacities. Although GPS-Gaussian demonstrates the potential for real-time human reconstruction and novel view synthesis, it suffers from severe distortions under sparse camera settings and mismatch across multiple viewpoints. Subsequent works have attempted to alleviate these issues. For instance, Tele-Aloha [27] introduces an image blending and cascaded disparity estimation method for human reconstruction with four input views. However, this approach is tailored to a specific system and struggles to generalize to sparser camera settings. Although GHG [8] achieves real-time 3D Gaussian-based human novel view synthe-

sis in a feed-forward manner, it requires additional human template priors, thus inheriting the limitations of template-based methods. In contrast, our method eliminates the need for human templates and is specifically designed to generalize effectively across diverse sparse camera settings.

3. Methodology

3.1. Overview

In this paper, we focus on fast human 3D reconstruction and novel view synthesis under diverse camera settings. Our objective is to reconstruct a 3D scene from a set of n sparse-view RGB images $\{\mathbf{I}_i\}_{i=1}^n$, $\mathbf{I}_i \in \mathbb{R}^{H \times W \times 3}$, captured from different viewpoints surrounding a human subject, where the angle between any two adjacent camera views is denoted by Δ , and synthesize arbitrary novel view images at any camera position in real time. To achieve this, we propose *EVA-Gaussian*, a method that utilizes deep neural networks and 3D Gaussian Splatting to enhance novel image quality while achieving real-time reconstruction.

Specifically, we employ 3DGS to represent each source image \mathbf{I}_i as a set of 3D Gaussians. Each pixel in the foreground corresponds to a unique 3D Gaussian. We use U_i to denote the number of Gaussians for source image i . The proposed EVA-Gaussian predicts the positions and attributes of 3D Gaussians in the form of attribute maps $\{\mathbf{M}_i\}_{i=1}^n = \{\mathbf{P}_i, \mathbf{O}_i, \mathbf{S}_i, \mathbf{Q}_i, \mathbf{F}_i\}_{i=1}^n$ from the image set $\{\mathbf{I}_i\}_{i=1}^n$, where \mathbf{P}_i , \mathbf{O}_i , \mathbf{S}_i , \mathbf{Q}_i , and \mathbf{F}_i denote the attribute maps for Gaussian positions, opacities, scales, quaternions, and features of source image i , respectively. Notably, in the feature map $\mathbf{F}_i = \{\mathbf{f}_i^u\}_{u=1}^{U_i}$, each element $\mathbf{f}_i^u \in \mathbb{R}^{32}$ serves as a new attribute associated with each 3D Gaussian, which will be used later in Sec. 3.4 to remove artifacts caused by geometric errors in $\{\mathbf{P}_i\}_{i=1}^n$. Mathematically, the procedure of EVA-Gaussian is expressed as:

$$\{\mathbf{M}_i\}_{i=1}^n = \mathcal{D}_\theta(\{\mathbf{I}_i\}_{i=1}^n), \quad (1)$$

where θ denotes the learnable parameters of the network.

The framework of EVA-Gaussian is depicted in Fig. 2. EVA-Gaussian splits the process of predicting Gaussian maps into three stages. In the first stage, it employs a U-Net architecture with an Efficient cross-View Attention module (EVA) to obtain enhanced multi-view predictions of the 3D Gaussian position maps $\{\mathbf{P}_i\}_{i=1}^n$, as elaborated in Sec. 3.2. In the second stage, a Gaussian attribute prediction network, detailed in Sec. 3.3, takes the predicted position maps $\{\mathbf{P}_i\}_{i=1}^n$ and the original RGB images $\{\mathbf{I}_i\}_{i=1}^n$ as input to estimate the remaining attributes of 3D Gaussians. The predicted 3D Gaussians from all source images are then aggregated to render target views using differential rasterization [6]. In the final stage, the rendered image \mathbf{I}^0 and its corresponding feature map $\mathbf{F}_{\text{novel}}$ are fused for further refinement using the network described in Sec. 3.4. In addition, an an-

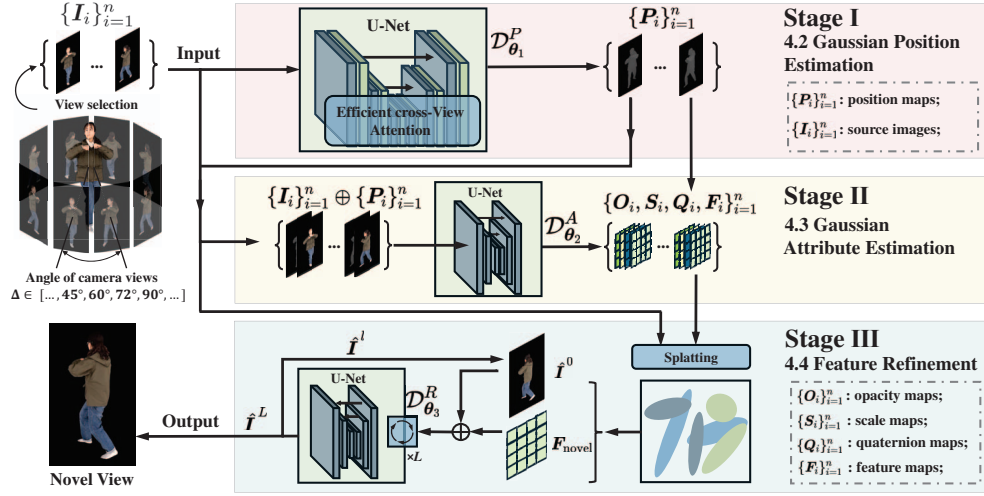


Figure 2. **Framework of EVA-Gaussian.** EVA-Gaussian takes sparse-view images captured around a human subject as input and performs three key stages: (1) estimating the positions of 3D Gaussians, (2) inferring the remaining attributes (i.e., opacities, scales, quaternions, and features) of these Gaussians, and (3) refining the output image in a recurrent manner.

chor loss is introduced during the training stage to enhance the overall reconstruction quality, as depicted in Sec. 3.5.

3.2. Gaussian Position Estimation

The variations in depth across the surface of human body may appear minimal. However, these nuances are critically important, particularly in regions such as the face and hands that contain a wealth of semantic information. Even slight inaccuracies in depth estimation within these areas can lead to significant degradation in visual quality and fidelity. This underscores the necessity for precise estimation of 3D Gaussian positions to enable effective and high-fidelity human reconstruction.

To tackle this challenge, we employ a U-Net based architecture, denoted by $\mathcal{D}_{\theta_1}^P$, to estimate the 3D Gaussian position maps $\{P_i\}_{i=1}^n$ from multi-view images $\{I_i\}_{i=1}^n$. This process is expressed as:

$$\{P_i\}_{i=1}^n = \mathcal{D}_{\theta_1}^P(\{I_i\}_{i=1}^n). \quad (2)$$

To ensure accurate depth estimation across diverse camera angles and arbitrary input views at high resolutions, while maintaining computational and temporal efficiency, we propose an EVA module, as illustrated in Fig. 3. This module is integrated into the three lowest resolution layers of the U-Net backbone $\mathcal{D}_{\theta_1}^P$ to facilitate the multi-view correspondence retrieval and information exchange. We use j to index these three layers, with $j = -1, j = -2$, and $j = -3$ representing the lowest, second-lowest, third-lowest resolution layers, respectively. The EVA module takes intermediate image features $E_i^j \in \mathbb{R}^{R^j \times C^j}, \forall i \in \{1, \dots, n\}, \forall j \in \{-1, -2, -3\}$, as input, where R^j and C^j denote the total number of pixels and the channel dimension of each pixel at layer j , respectively. The module outputs enhanced im-

age features \tilde{E}_i^j , which incorporate cross-view information. Before the execution of attention mechanisms, a learnable positional embedding γ is added to encode spatial coordinates, improving the understanding of image geometry.

While cross-view attention mechanisms have been explored in various vision tasks that require multiple image inputs, existing approaches in feed-forward 3D Gaussian reconstruction, such as epipolar attention in pixelSplat [2] and vanilla cross-attention in MVSpLat [4], are limited to low-resolution inputs (256×256) and fail to recover fine-grained details. In contrast, our approach targets high-resolution (1024×1024) human-centric reconstruction, a setting that has been demonstrated to be critical for recovering intricate human details [25]. The key novelty of our EVA module lies in its localized 1D window-based attention mechanism, which exploits the observation that corresponding pixels from reference views are typically located in adjacent positions along the x-axis in human-centric camera settings. Unlike global attention mechanisms used in prior works [2, 4, 12], EVA computes cross-attention only within a local 1D window aligned with the x-axis, and the window is shifted by half its size at each iteration to expand the receptive field. This design significantly reduces computational complexity and improves GPU memory utilization, while maintaining high performance, as demonstrated in Tab. 1.

Notably, when the scale of each Gaussian is sufficiently small, the 3D Gaussian position of a pixel aligns precisely with its corresponding value on the depth map. A detailed proof of this property is provided in Appendix D. Based on this observation, we train the position estimation network \mathcal{D}_{θ}^P to obtain the position maps $\{P_i\}_{i=1}^n$ with the mean squared error (MSE) loss function:

$$\mathcal{L}_{\text{depth}} = \|\mathbf{P}_i - \mathbf{P}_i^{\text{gt}}\|_2, \quad (3)$$

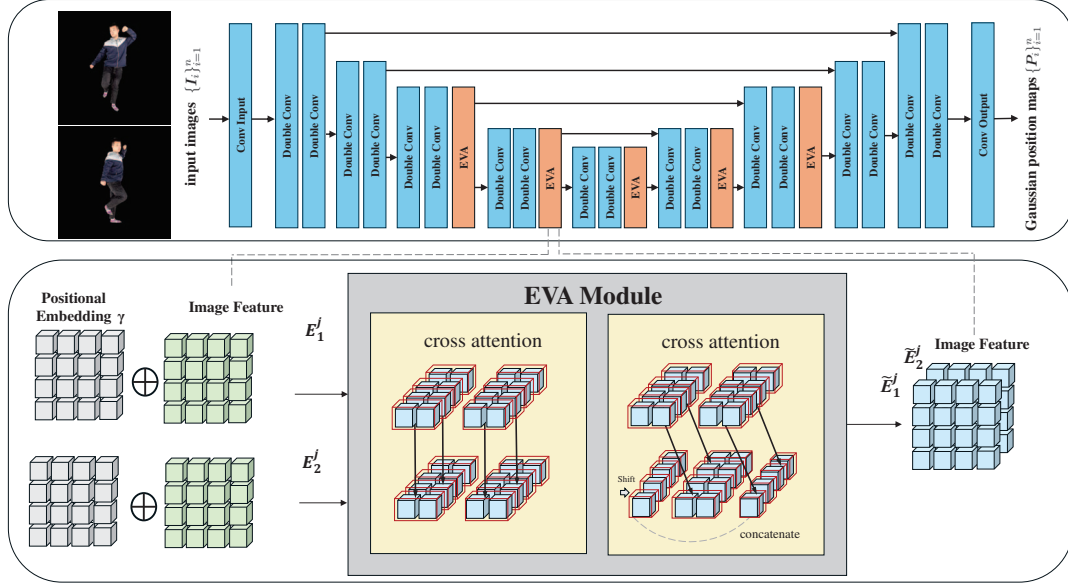


Figure 3. **Efficient cross-View Attention (EVA) module** for 3D Gaussian position estimation. EVA takes multi-view image features as input, embeds them into window patches using a shifted algorithm, and performs cross attention between the features from different views.

where P_i^{gt} denotes the ground truth depth map.

3.3. Gaussian Attribute Estimation

To complete the estimation of 3D Gaussian maps $\{M_i\}_{i=1}^n$, we employ a shallow U-Net $\mathcal{D}_{\theta_2}^A$ to estimate the remaining attributes O_i, S_i, Q_i, F_i . This network takes the estimated position maps $\{P_i\}_{i=1}^n$ from the first stage in Sec. 3.2 and the original RGB images $\{I_i\}_{i=1}^n$ as input, and outputs 3D Gaussian attributes O_i, S_i, Q_i, F_i , which is expressed as:

$$\{O_i, S_i, Q_i, F_i\}_{i=1}^n = \mathcal{D}_{\theta_2}^A(\{I_i\}_{i=1}^n \oplus \{P_i\}_{i=1}^n). \quad (4)$$

The resulting estimated 3D Gaussian maps $\{M_i\}_{i=1}^n = \{P_i, O_i, S_i, Q_i, F_i\}_{i=1}^n$ are then utilized to render novel views through the α -blending mechanism, as in the vanilla 3DGS [6]. The network $\mathcal{D}_{\theta_2}^A$ is trained by using a combination of MSE loss and structural similarity index measure (SSIM) [29] loss between the rendered novel view image \hat{I}^0 and the ground truth I^{gt} as follows:

$$\mathcal{L}_{\text{render}} = \|\hat{I}^0 - I^{\text{gt}}\|_2 + \lambda_{\text{render}}(1 - \text{SSIM}(\hat{I}^0, I^{\text{gt}})), \quad (5)$$

where λ_{render} denotes the weighting factor for the SSIM loss.

3.4. Feature Splatting and Refinement

The 3D Gaussian position maps P_i estimated in Sec. 3.2 inevitably contain some degree of error, which may lead to distortions and artifacts in the rendered RGB images. To mitigate these issues, we propose a post-splatting refinement method to correct the position estimates. Recent studies [1] have demonstrated that feature vector representations can capture scene information more effectively than

spherical harmonics, resulting in significant improvements in novel view synthesis, particularly in scenarios with limited overlapping views. Inspired by this finding, we incorporate a feature vector, i.e., $f_i^u \in \mathbb{R}^{32}$ mentioned in Sec. 3.1, as an additional attribute for each Gaussian to more precisely capture its spatial characteristics.

During the splatting process, we first aggregate the 3D Gaussians from all source views. Then, the color values of these 3D Gaussians are rendered using α -blending mechanism from 3D Gaussian Splatting [6]. Concurrently, the feature values of the 3D Gaussians are splatted onto the image plane using a modified α -blending function:

$$f_{\text{pixel}} = \sum_{j=1}^N f_j \alpha_j \prod_{l=1}^{j-1} (1 - \alpha_l), \quad (6)$$

where f_{pixel} is the feature vector for the corresponding pixel on the feature map of the novel view image F_{novel} , f_j denotes the feature vector for the 3D Gaussian with the j -th greatest depth, and $N = \sum_{i=1}^n U_i$ is the total number of 3D Gaussians from all source views.

Moreover, we employ a carefully designed recurrent U-Net $\mathcal{D}_{\theta_3}^R$ that takes both the RGB and feature images as input and projects them onto the RGB space for the final output through L recurrent loops. This procedure is expressed as:

$$\hat{I}^l = \mathcal{D}_{\theta_3}^R(\hat{I}^{l-1} \oplus F_{\text{novel}}), l \in \{1 \cdots L\}, \quad (7)$$

where $\hat{I}^l \in \mathbb{R}^{H \times W \times 3}$ and $F_{\text{novel}} \in \mathbb{R}^{H \times W \times 32}$. Similar to the Gaussian attribute estimation, the loss function for

Table 1. Comparison of computational and temporal costs for different attention mechanisms. CVA refers to the cross-view attention in MVSpLat, EA denotes the epipolar attention in pixelSpLat, and w16, w32, w64 represent the window sizes of 16, 32, and 64, respectively, in EVA. G.M. indicates the GPU memory usage.

Input Size	2×64×128×128		2×64×256×256		2×32×256×256	
Module	Time	G.M.	Time	G.M.	Time	G.M.
CVA	35.3ms	3.72GB	30.4ms	35.44GB	26.3ms	31.77GB
EA	58.3ms	15.19GB	19.3ms	59.14GB	16.9ms	58.01GB
EVA (w16)	7.23ms	0.92GB	1.77ms	2.14GB	1.43ms	1.37GB
EVA (w32)	6.53ms	0.92GB	1.49ms	2.14GB	1.16ms	1.37GB
EVA (w64)	6.31ms	0.92GB	1.39ms	2.14GB	1.06ms	1.37GB

supervising the final output is a combination of the MSE loss and the SSIM loss between the refined image \hat{I}^L and the ground truth I^{gt} as follows:

$$\mathcal{L}_{\text{refine}} = \|\hat{I}^L - I^{\text{gt}}\|_2 + \lambda_{\text{refine}}(1 - \text{SSIM}(\hat{I}^L, I^{\text{gt}})), \quad (8)$$

where λ_{refine} denotes the weighting factor for the SSIM loss.

3.5. Attribute Regularization

Existing feed-forward 3D Gaussian Splatting methods, such as GPS-Gaussian [37] and MVSpLat [4], typically assume that the position of 3D Gaussians correspond directly to the estimated depth values in their respective views. However, this assumption does not always hold, as 3D Gaussians are not explicitly constrained to lie on the surface of objects. Moreover, the unprojection process in these methods is performed independently for each view using an orthogonal projection framework. This approach lacks explicit scene understanding and cross-view consistency, as it does not incorporate mechanisms for cross-view matching or global scene optimization during the standard training procedure.

To address these limitations and improve reconstruction quality, we introduce a regularization term to enhance the training process. Specifically, we propose an anchor loss to regularize the scales and opacities of Gaussians, ensuring consistency between the geometry of predicted depth maps and the 3D Gaussian positions, as proved in Appendix D. This loss also aligns the Gaussians from different views to force their locations to the same landmark. We leverage MediaPipe [17] to annotate human facial landmarks and compute the anchor loss to regularize the 3D landmark Gaussian scales, opacities, and positions as follows:

$$\mathcal{L}_{\text{anchor}} = \lambda_{\text{opacity}} \sum_{i=1}^N \|\mathbf{O}_i \log(\mathbf{O}_i)\|_1 + \lambda_{\text{scale}} \sum_{i=1}^N \|\mathbf{S}_i\|_2 + \sum_{i,j \in \mathbb{V}} \sum_{m_i \in \mathbb{M}_i, m_j \in \mathbb{M}_j} \max\{\text{Dist}, t\}, \quad (9)$$

where Dist represents the distance of human facial landmarks in 3D space with

$$\text{Dist} = \|\Pi^{-1}(\mathbf{m}_i, \mathbf{P}_i(\mathbf{m}_i)) - \Pi^{-1}(\mathbf{m}_j, \mathbf{P}_j(\mathbf{m}_j))\|_2,$$

$\{\mathbb{M}_i\}_{i=1}^n$ denotes the collection of all landmarks on the 2D image plane, \mathbb{V} denotes the collection of source views, and

Table 2. Comparison with feed-forward 3D reconstruction methods at a resolution of 256×256. Inference speeds are reported in the last column. Better results are marked in a deeper color.

$\Delta=45^\circ$	THuman2.0			THumansit			Speed
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
pixelSpLat	25.19	0.9156	0.0824	23.31	0.8880	0.0954	185ms
MVSpLat	28.05	0.9515	0.0346	24.97	0.9223	0.0532	70ms
MVSGaussian	26.44	0.9706	0.0283	25.20	0.9641	0.0297	71ms
ENeRF	29.62	0.9696	0.0238	27.06	0.9567	0.0334	136ms
GPS-Gaussian	30.30	0.9762	0.0224	28.02	0.9671	0.0251	40ms
EVA-Gaussian	31.11	0.9782	0.0198	29.16	0.9696	0.0249	55ms

Π^{-1} represents the process of reprojection from 2D image to 3D space. In addition, a tolerance factor t is introduced to mitigate the errors in the estimated landmarks.

By integrating the loss functions from the three stages, i.e., $\mathcal{L}_{\text{depth}}$, $\mathcal{L}_{\text{render}}$, $\mathcal{L}_{\text{refine}}$, and the proposed regularization term $\mathcal{L}_{\text{anchor}}$, the overall training loss for EVA-Gaussian is:

$$\mathcal{L}_{\text{EVA}} = \mathcal{L}_{\text{depth}} + \lambda_1 \mathcal{L}_{\text{render}} + \lambda_2 \mathcal{L}_{\text{refine}} + \lambda_3 \mathcal{L}_{\text{anchor}}, \quad (10)$$

where λ_1 , λ_2 , and λ_3 are weighting factors.

Since the 3D Gaussian position and attribute estimation stages can be executed within tens of milliseconds, and the feature refinement stage is lightweight, taking less than ten milliseconds, EVA-Gaussian is capable of rapidly reconstructing 3D human subjects from a collection of RGB images and rendering novel views in a real-time manner.

4. Experiments

4.1. Experiment Setup

Implementation details. Our EVA-Gaussian is trained on 1024×1024 pixel images across multiple training views using a single NVIDIA A800 GPU for 100K iterations with the AdamW [16] optimizer, unless otherwise specified. For the 3D Gaussian position estimation stage, it is first pre-trained under the supervision of ground truth depth maps. Baselines are trained using their publicly available code. More implementation details are provided in Appendix A.

Datasets. We conduct experiments on two open-source human body datasets: THuman2.0 [34] and THumanSit [35]. THuman2.0 contains 526 unique human models with their corresponding SMPL parameters, among which 100 individuals are randomly selected for our evaluation. The THumanSit dataset has a similar structure, containing 72 human models with around 60 poses for each, and we randomly choose 5 individuals with all poses for our evaluation.

Metrics. We report results on commonly used metrics: PSNR, SSIM [29], and LPIPS [36], computed over the entire image, as some methods may produce artifacts outside the human bounding box [11, 37]. We also include the inference time to demonstrate the real-time reconstruction capability of our method.

4.2. Stereo Reconstruction

Comparison with state-of-the-art feed-forward reconstruction methods. We first compare our approach against

Table 3. Comparison of feed-forward methods under different camera angle settings at a resolution of 1024×1024 . Better results are marked in a deeper color. GPS-Gaussian fails to work effectively when $\Delta = 90^\circ$, as it is unable to meet its rectification requirement.

THuman2.0		$\Delta = 45^\circ$			$\Delta = 60^\circ$			$\Delta = 72^\circ$			$\Delta = 90^\circ$		
1024 × 1024	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
ENeRF	27.94	0.9573	0.0367	26.16	0.9452	0.0516	24.61	0.9309	0.0705	22.85	0.8990	0.1147	
GPS-Gaussian	29.63	0.9703	0.0174	27.36	0.9630	0.0249	24.25	0.9519	0.0480	/	/	/	
EVA-Gaussian	30.46	0.9730	0.0178	28.29	0.9654	0.0248	27.54	0.9614	0.0297	26.31	0.9555	0.0391	

THumansit		$\Delta = 45^\circ$			$\Delta = 60^\circ$			$\Delta = 72^\circ$			$\Delta = 90^\circ$		
1024 × 1024	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
ENeRF	25.61	0.9397	0.0494	23.80	0.9168	0.0745	22.48	0.8956	0.0985	21.20	0.8571	0.1406	
GPS-Gaussian	27.05	0.9584	0.0227	25.19	0.9480	0.0351	21.48	0.9276	0.0713	/	/	/	
EVA-Gaussian	28.76	0.9621	0.0236	27.38	0.9543	0.0321	26.60	0.9498	0.0500	25.44	0.9416	0.0512	

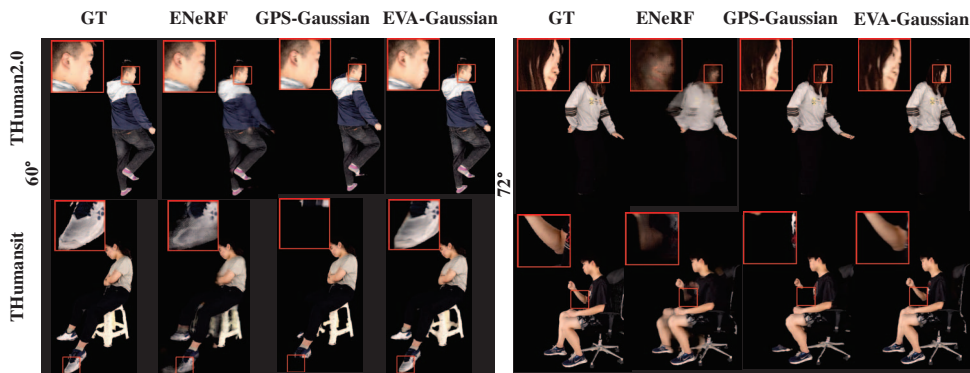


Figure 4. Qualitative comparison on THuman2.0 and THumansit. EVA-Gaussian achieves superior novel view rendering quality under diverse camera settings. Additional visualization results are provided in Appendix C.

state-of-the-art (SOTA) feed-forward reconstruction methods, including ENeRF [11], pixelSplat [2], MVSpLat [4], MVSGaussian [13], and GPS-Gaussian [37]. All experiments are conducted in a stereo-view setting, where the angle between the two camera views $\Delta = 45^\circ$. The attention modules in the scene reconstruction methods [2, 4, 13] are inefficient in their utilization of GPU memory, limiting their ability to train effectively at a high resolution of 1024×1024 . Therefore, we also conduct a fair comparison of all methods at a resolution of 256×256 . The quantitative results presented in Table 2 demonstrate that EVA-Gaussian achieves the best novel view quality in terms of PSNR, SSIM, and LPIPS, while maintaining the second-fastest inference speed.

Comparison under diverse angle changes between camera views. We further evaluate the performance of our method across four different angles between the two camera views, i.e., $\Delta = 45^\circ, 60^\circ, 72^\circ$, and 90° , at a high resolution of 1024×1024 . As shown in Table 3, our EVA-Gaussian outperforms all baseline methods on all metrics, achieving a maximum PSNR advantage of 5.12 dB. Notably, thanks to our EVA module, EVA-Gaussian remains effective even under extremely sparse camera settings, e.g., $\Delta = 90^\circ$. In contrast, GPS-Gaussian fails to work effectively due to its reliance on stereo rectification. Fig. 4 presents the qualitative results of novel view rendering, where EVA-Gaussian outperforms previous SOTA methods in rendering quality,

especially in scenarios with large viewpoint discrepancies.

4.3. Multi-view Reconstruction

We conduct experiments under multi-view settings to evaluate EVA-Gaussian’s capability to handle more than two input images. Table 4 presents the quantitative results, where our method demonstrates a performance gain with more than 1.5 dB improvement over the baseline. While the performance of GPS-Gaussian drops significantly due to the mismatch among multiple inferences, our method maintains high performance, thanks to the cross-view consistency ensured by our proposed EVA module.

4.4. Generalizability Validation

We validate the generalizability of EVA-Gaussian by taking out-of-domain images as input. It is observed that EVA-Gaussian is inherently superior in generalizing to out-of-distribution human identities and postures, primarily due to the strong data processing ability of its attention modules in EVA-Gaussian. This allows EVA-Gaussian to maintain consistent performance when provided with sufficient data. As shown in Table 5, given that THumansit contains significantly more human models (over 4,000) than THuman2.0 (526 models), EVA-Gaussian shows a greater performance improvement (+2.11 dB in PSNR) compared to GPS-Gaussian (+1.97 dB in PSNR). This conclusion is further supported by the evaluation on THumansit, where models trained on THuman2.0 experience a performance

Table 4. Comparison with GPS-Gaussian under different camera numbers. Results in bold represent the best performance. EVA-Gaussian achieves SOTA performance across various metrics, primarily due to the multi-view consistency enabled by our proposed EVA module.

1024×1024	THuman2.0 ($\Delta = 45^\circ$)						THumansit ($\Delta = 45^\circ$)					
	3 views			4 views			3 views			4 views		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GPS-Gaussian	28.74	0.9655	0.0200	28.51	0.9636	0.0218	26.87	0.9523	0.0243	26.50	0.9498	0.0267
EVA-Gaussian	30.76	0.9722	0.0175	30.35	0.9707	0.0189	28.64	0.9596	0.0255	28.32	0.9582	0.0260

Table 5. Quantitative results of cross-domain validation. EVA-Gaussian consistently outperforms GPS-Gaussian.

Method	THumansit → THuman2.0			THuman2.0 → THumansit		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GPS-Gaussian	29.33	0.9733	0.0325	20.86	0.9243	0.0872
EVA-Gaussian	30.40	0.9751	0.0321	21.27	0.9275	0.0876

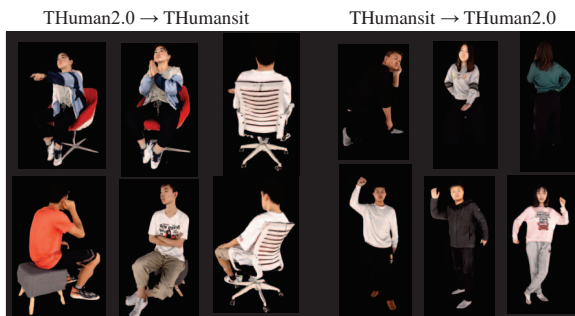


Figure 5. Visualization of cross-domain evaluation results for EVA-Gaussian. The left side displays the rendered results generated by EVA-Gaussian trained on the THuman2.0 dataset and evaluated on the THumansit dataset, while the right side shows the rendered results from EVA-Gaussian trained on the THumansit dataset and evaluated on the THuman2.0 dataset.

decline due to limited data availability. Despite this, EVA-Gaussian still outperforms GPS-Gaussian, achieving a performance gain of 0.41 dB in PSNR. In addition, Fig. 5 provides visual evidence of EVA-Gaussian’s robustness to out-of-domain data, since we have explicitly introduced inductive bias in the EVA module to mitigate cross-domain performance degradation. Additional validation results on a real-world dataset are provided in Appendix B.

4.5. Ablation Study

We conduct a detailed ablation study on THuman2.0 in a stereo-view setting, where the angle between the two views $\Delta = 45^\circ$, as shown in Table 6 and Fig. 6. We gradually incorporate the EVA module, feature refinement module, and anchor loss to evaluate their individual contributions. Removing the EVA module results in significant degradation across all metrics, as the network struggles to perform multi-view 3D Gaussian geometry prediction. When feature refinement is excluded, artifacts appear in critical areas, such as the hands and feet. Moreover, the lack of anchor loss leads to unreliable geometry predictions, particularly in the facial region, which in turn degrades the performance across all metrics, with a notable impact on LPIPS.

Table 6. Quantitative results of the ablation study on THuman2.0 in a stereo-view setting, where the angle between the two views $\Delta = 45^\circ$, at a resolution of 1024×1024 .

1024×1024	THuman2.0 ($\Delta = 45^\circ$)			
	w/o EVA	w/o f.r.	w/o anchor loss	Full model
PSNR↑	23.41	29.31	30.34	30.46
SSIM↑	0.9380	0.9676	0.9724	0.9730
LPIPS↓	0.0659	0.0191	0.0186	0.0178

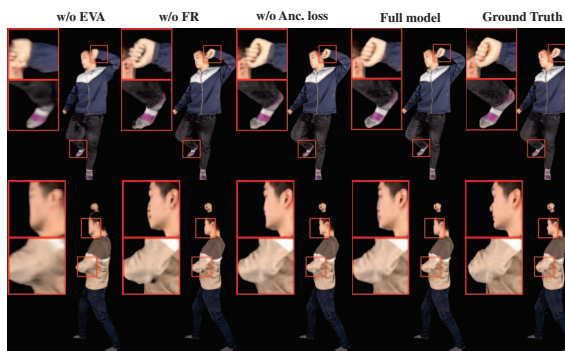


Figure 6. Qualitative visualization results of the ablation study on THuman2.0. Each module shows its effectiveness for a better visual output. The feature refinement (FR) module corrects geometric errors in the initial estimations, and the anchor loss further refines critical areas, such as the face, for generating novel view images with higher fidelity.

5. Conclusion

In this paper, we introduce EVA-Gaussian, a novel real-time 3D human reconstruction pipeline that employs multi-view attention-based 3D Gaussian position estimation and comprehensive feature refinement. To ensure robust performance, the method is trained using both photometric loss and anchor loss. Quantitative and qualitative evaluations on benchmark datasets demonstrate that EVA-Gaussian achieves state-of-the-art performance while maintaining a competitive inference speed, particularly under sparse camera settings.

While EVA-Gaussian synthesizes high-fidelity novel views, there remain several areas for improvement. For instance, the attention module can consume substantial GPU memory when processing a large number of input views or high-resolution images. In addition, the naive reprojection of pixels into 3D space may introduce conflicts in overlapping areas, leading to redundancy in the 3D representation. These limitations can be effectively addressed by incorporating RGB-D information or developing advanced techniques for detecting and resolving overlapping areas.

References

- [1] T Berriel Martins and Javier Civera. Feature splatting for better novel view synthesis with low overlap. *arXiv e-prints*, pages arXiv–2405, 2024. 5
- [2] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 3, 4, 7
- [3] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnrf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14124–14133, 2021. 3
- [4] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *arXiv preprint arXiv:2403.14627*, 2024. 3, 4, 6, 7
- [5] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 634–644, 2024. 1, 2, 3
- [6] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 3, 5
- [7] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 505–515, 2024. 1, 2, 3
- [8] Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, et al. Generalizable human gaussians for sparse view synthesis. *arXiv preprint arXiv:2407.12777*, 2024. 3
- [9] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19876–19887, 2024. 1, 2, 3
- [10] Ruilong Li, Yulian Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 49–67. Springer, 2020. 2
- [11] 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 2022 Conference Papers*, pages 1–9, 2022. 1, 3, 6, 7
- [12] Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Fast generalizable gaussian splatting reconstruction from multi-view stereo. *arXiv preprint arXiv:2405.12218*, 2024. 3, 4
- [13] Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. *arXiv preprint arXiv:2405.12218*, 2024. 7
- [14] Zhening Liu, Yingdong Hu, Xinjie Zhang, Rui Song, Jiawei Shao, Zehong Lin, and Jun Zhang. Dynamics-aware gaussian splatting streaming towards fast on-the-fly 4d reconstruction, 2025. 1
- [15] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3
- [16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [17] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuolin Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 6
- [18] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1
- [19] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2
- [20] Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors. *arXiv preprint arXiv:2406.12459*, 2024. 3
- [21] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 3
- [22] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5020–5030, 2024. 1, 2, 3
- [23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2
- [24] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 2

- [25] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 84–93, 2020. 2, 4
- [26] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10208–10217, 2024. 3
- [27] Hanzhang Tu, Ruizhi Shao, Xue Dong, Shunyuan Zheng, Hao Zhang, Lili Chen, Meili Wang, Wenyu Li, Siyan Ma, Shengping Zhang, et al. Tele-aloha: A low-budget and high-authenticity telepresence system using sparse rgb cameras. *arXiv preprint arXiv:2405.14866*, 2024. 1, 3
- [28] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2021. 3
- [29] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5, 6
- [30] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G Schwing, and Shenlong Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2059–2069, 2024. 1, 2, 3
- [31] Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. *arXiv preprint arXiv:2403.16292*, 2024. 3
- [32] Haofei Xu, Anpei Chen, Yuedong Chen, Christos Sakaridis, Yulun Zhang, Marc Pollefeys, Andreas Geiger, and Fisher Yu. Murf: Multi-baseline radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20041–20050, 2024. 2
- [33] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 2, 3
- [34] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, 2021. 2, 6
- [35] Jiajun Zhang, Yuxiang Zhang, Hongwen Zhang, Boyao Zhou, Ruizhi Shao, Zonghai Hu, and Yebin Liu. Ins-hoi: Instance aware human-object interactions recovery. *arXiv preprint arXiv:2312.09641*, 2023. 2, 6
- [36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6
- [37] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. 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*, pages 19680–19690, 2024. 1, 2, 3, 6, 7