2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3315-4364-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/CVPR52734.2025.01969

GeoAvatar: Geometrically-Consistent Multi-Person Avatar Reconstruction from Sparse Multi-View Videos

Soohyun Lee[†] Seoyeon Kim[†] HeeKyung Lee[‡] Won-Sik Jeong[‡] Joo Ho Lee[†]

[†]Sogang University [‡]Electronics and Telecommunications Research Institute

Abstract

Multi-person avatar reconstruction from sparse multi-view videos is challenging. The independent avatar reconstruction of each person often fails to reconstruct the geometric relationship among multiple instances, resulting in inter-penetrations among avatars. Some researchers resolve this issue via neural volumetric rendering techniques but they suffer from huge computational costs for rendering and training. In this paper, we propose a multi-person avatar reconstruction method that reconstructs a 3D avatar of each person while keeping the geometric relations among people. Our 2D Gaussian Splatting (2DGS)-based avatar representation allows us to represent geometrically-accurate surfaces of multiple instances that support sharp inside-outside tests. We utilize the monocular prior to alleviate the inter-penetration via surface ordering and to enhance the geometry in less-observed and textureless surfaces. We demonstrate the efficiency and performance of our method quantitatively and qualitatively on a multi-person dataset [49] containing close interactions.

1. Introduction

Human avatar reconstruction has been intensively studied, as the novel view synthesis achieves unprecedented advances in the past half a decade [17, 23, 24, 26, 32, 42]. Single-person avatar reconstruction combines pose tracking and novel view synthesis, where the human appearance is modeled in the canonical space articulated with the human skeleton [31]. The appearance in the target frame is then obtained by warping from the canonical to the target space based on the estimated target pose. The overall performance of single-person avatar reconstruction depends on the accuracy of pose estimation and the fidelity of appearance reconstruction.

In contrast to single-person reconstruction, multi-person scenarios present an additional challenge: inter-penetration between body parts and inter-person occlusions. Inter-penetrations occur when each avatar is processed independently without the consideration of adjacent avatars. Previous approaches [7, 18, 26, 51] maintain geometric relationships by modeling the occlusion-aware image formation model, detecting penetrating regions and pushing them outward. These methods require clear surface representations such as mesh [18] or implicit surface [37] for the inner-outer test which takes a huge computational cost. They may cause the shape deviation as the skinning model does not address the soft body deformation in contact regions.

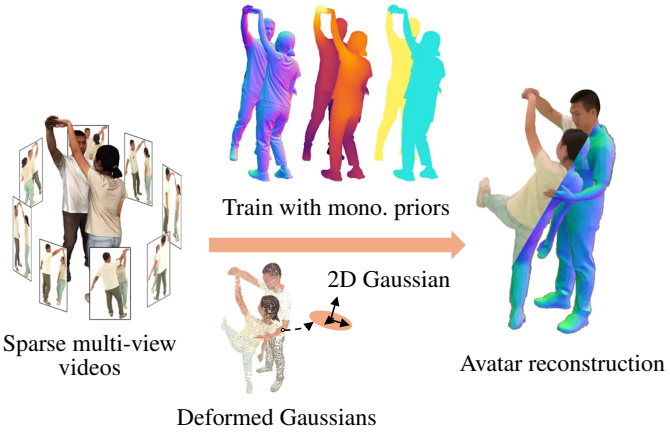

Figure 1. Teaser. We represent multi-person avatars using planar Gaussian splats. To overcome the view sparsity, we utilize the monocular priors during training. Our reconstruction achieves state-of-the-art performance with fast training.

The reconstruction becomes more challenging with sparse views, where silhouettes provide the coarse geometry bound of an avatar and triangulation-based depth priors become unreliable, particularly in less-observed and textureless areas. Recently, vision foundation models, especially a human vision foundation model Sapiens [21], achieve remarkable improvements in monocular geometric estimation. Inspired by MonoSDF [50], which shows that monocular inference improves static 3D reconstruction, we utilize the monocular prior to achieve geometrically consistent multiple avatar reconstruction.

In this paper, we propose a method that reconstructs

2575-7075/25/$31.00 ©2025 IEEE
DOI 10.1109/CVPR52734.2025.01969

multiple avatars from sparse multi-view videos while keeping their geometric relationships. To form the discrete surface boundaries without sacrificing the rendering speed, we adopt the 2D Gaussian splatting (2DGS) representation as an avatar representation of each person. We synthesize multi-person images by warping all planar Gaussian splats onto the target space, sorting them, and blending them in the volumetric manner. To overcome inter-penetrations, we efficiently reorder surfaces of foreground and background instances without explicitly addressing the physical interactions in contact regions. We utilize monocular geometric priors to improve the geometric details while mitigating the ambiguity of sparse views and textureless regions. Our method achieves the state-of-the-art performance among multi-person avatar reconstruction methods in terms of geometry and rendering quality.

2. Related Work

Dynamic Avatar Reconstruction Early approaches to avatar reconstruction focus on simultaneous surface scanning and non-rigid tracking [4, 13, 30, 39]. Avatar geometries are captured using a commercial depth camera such as KINECT [29]. The captured surfaces are aligned and integrated in the canonical space by estimating the non-rigid motion, controlled by the embedded deformation graph [38].

Recent studies tackle the problem of avatar reconstruction by estimating the motion offset field and reconstructing the human appearance, utilizing the human template body model and the deep pose prior [31] to establish coarse correspondence fields between the canonical and the target frames [32]. Building on the success of neural-based novel view synthesis [27, 28, 44, 48], various neural volumetric representations have been explored in avatar reconstruction, ranging from vanilla implicit neural networks [42] to sophisticated approaches using latent features stored in meshes [32] and lattice grids [26]. To focus on reconstruction near the template surfaces, some neural volumetric representations learn the radiance field in the surface-aligned volume, associated with texture space [15, 22, 45, 54]. Learning the background scene together helps automate the reconstruction pipeline [17]. These neural representations train the density-based radiance field defined in the canonical space by warping the camera rays. However, density-based radiance fields struggle to separate the geometry and the texture, resulting in physically inconsistent surface geometry. Although implicit signed distance field representations [10, 14] improve surface quality, neural volumetric representations still suffer from a long training time.

In another venue, the explicit scene representation, called Gaussian splatting [20], achieves the compatible performance in the novel-view synthesis with fast rendering and training thanks to the differentiable rasterization. In order to represent the dynamic appearance of avatars, the Gaussian splats in the canonical space are moved into the target space by using the skinning model [36]. Additional deformation fields improve the reconstruction quality by enhancing the correspondence between frames through pose-dependent deformation [33] and auxiliary deformation guided by latent bones [24]. Although these approaches achieve fast training and real-time rendering, their surfaces are not well-aligned to the actual surfaces [9]. We extend this approach to the planar Gaussian splats [12], achieving both geometric-consistent reconstruction and fast training.

Multi-Person Avatar Reconstruction Multi-person dynamic avatar reconstruction has been studied by constructing an independent canonical space for each avatar and aggregating their appearance through volumetric rendering [37, 52]. This simple extension to multi-person utilizes the multi-view consistency to find the corresponding instance for each pixel, which works in far-person scenarios.

In close-interaction scenarios, independent reconstruction of each avatar falls into the geometrically inconsistent results such as inter-penetrations between avatars. The inter-penetration hinders the appearance reconstruction of avatars, since a camera ray cannot reach the corresponding point. To resolve this inter-penetration problem, the major approach penalizes the simultaneous occupancy by multiple avatars [6, 7, 16, 18, 26, 49, 51]. Though this approach is physically sound, it requires the per-point inference for inner points and may cause unwanted shape deviation due to model inadequacy for elastic body deformation. The instance segmentation map helps to correct the ordering of visible surfaces along the ray [7, 16, 18, 52]. We use the surface ordering strategy to expose the penetrated parts without over-regulating unseen contact parts.

Exploring Monocular Priors Monocular prior models for geometry estimation [1, 5, 8, 11, 19, 35] and segmentation [34] have been proposed by training networks with large image datasets of general scenes. Some researchers focus on human-centric prior models. ECON [43] estimates the geometry of the clothed human supervised by the clothed-human scanning dataset. A large human-vision foundation model, Sapiens [21], trained with a curated human image dataset demonstrates the state-of-the-art performance in human-centric vision tasks such as 2D pose estimation, body-part segmentation, and monocular geometry estimation.

Recent works have demonstrated that monocular geometry priors enhance static scene reconstruction across various scene representations, e.g. the neural implicit functions [50] and the Gaussian-based models [3, 40]. In this paper, we exploit the monocular priors to enhance dynamic multi-person avatar reconstruction from sparse multi-view observations.

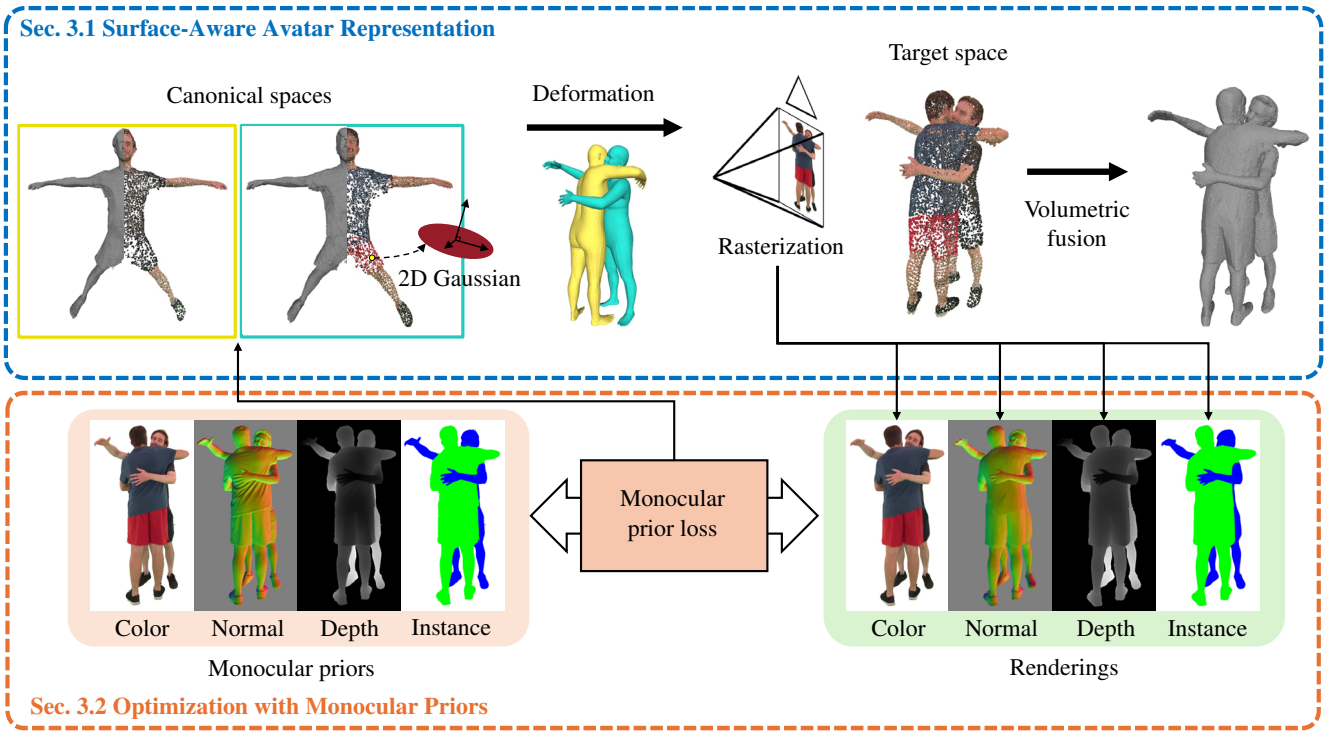

Figure 2. Overview. We define each avatar's representation as a set of planar Gaussian splats [12] in the different canonical spaces. The avatars are transformed to the integrated target space through the deformation field. We render the property maps of the scene such as RGB color, depth, normal, and instance, in the volumetric manner and encourage the similarity between renderings and monocular predictions during optimization.

3. Method

Our goal is to reconstruct the consistent geometric relationships among avatars by resolving the inter-penetration issues while estimating the geometrically accurate appearance of avatars. Sec. 3.1 proposes a geometrically consistent avatar representation that has well-defined surfaces that enables precise inside-outside testing. Sec. 3.2 explains objective functions to train the multiple avatar representation from multi-view videos.

3.1. Surface-Aware Avatar Representation

A surface-aware appearance model enables the reconstruction of geometrically consistent 3D avatar surfaces by producing clear boundaries between the avatar's interior and exterior, allowing detection of physically-incoherent events such as inter-penetrations. To achieve this, we propose a planar Gaussian mixture appearance model articulated with the human body template model [31]. For each avatar, we splat planar Gaussian disks onto the independent canonical space aligned with the reference human pose called DA-pose [17]. Each Gaussian splat has a spatial density distribution G parameterized by the rotation $\mathbf{R}$, the center position $\mathbf{t}$, and the scale matrix $\mathbf{S}$:

$$G(\mathbf{u}) = e^{-\frac{1}{2}\mathbf{u}^T\mathbf{S}^{-T}\mathbf{S}^{-1}\mathbf{u}},$$

where $\mathbf{u}$ is a local 2D coordinate (u, v) of the i-th Gaussian splat corresponding to the 3D coordinate $\mathbf{x} = \mathbf{R}\mathbf{u} + \mathbf{t}$. It contains surface properties such as view-dependent color $\mathbf{c}$, opacity α, and surface normal $\mathbf{n}$ to represent the geometric and radiometric appearance.

To represent the appearance at the target frame, each Gaussian splat in the canonical space is moved to the corresponding position in the target space. The spatial deformation between the canonical and target spaces is defined as linear blending of transformations of adjacent human body joints $\{\mathbf{T}_k\}$ called linear-blending skinning (LBS):

$$\hat{\mathbf{T}}(\mathbf{x}) = \sum_{k=1}^{n_b} w_k(\mathbf{x})\mathbf{T}_k, \hat{\mathbf{x}} = \hat{\mathbf{T}}(\mathbf{x})\mathbf{x}, \quad (1)$$

where w_k is the blending weight for the k-th body joint, $\mathbf{x}$ is a canonical point, $\hat{\mathbf{x}}$ is the corresponding target point, n_b is the number of adjacent human joints, and $\hat{\mathbf{T}}(\mathbf{x})$ is a deformation field from the canonical space to the target space. In addition, we add the auxiliary deformation field [24] to cope with general motions of clothes.

The translated Gaussian splats of avatars are gathered in the target space then are sorted by depth and projected onto the image space via volumetric alpha blending in order to

render the animated avatar appearance:

$$\mathbf{p}(\mathbf{r}) = \sum_{i=1} \mathbf{p}_i \alpha_i G_i(\mathbf{u}_i(\mathbf{r})) \prod_{j=1}^{i-1} (1 - \alpha_j G_j(\mathbf{u}_j(\mathbf{r}))) \quad (2)$$

where i is the index of a Gaussian splat, $\mathbf{p}_i \in \{\mathbf{c}_i(\mathbf{r}), \mathbf{n}_i, d_i\}$ and α_i are a property of Gaussian splats and an opacity of the Gaussian splat, $\mathbf{r}$ is a pixel ray, and $\mathbf{u}_i(\mathbf{x})$ is the ray-splat intersection point. Thanks to the nature of planar Gaussian splats, the mixture of them estimates the well-defined surfaces and the surface-aligned texture appearance by regularizing their relative distances under multi-view supervision [12].

3.2. Optimization with Monocular Priors

In the previous sections, we build the differentiable rendering pipeline for our multiple avatar representations. We train each person's avatar by rendering all visible avatars into video frames and penalizing the photometric differences. To address the ambiguity of sparse-view photometric reconstruction, we adopt vision foundation models [21, 34] to utilize monocular priors such as instance segmentation and monocular geometry estimation. Besides, we encourage the planar Gaussian splats to construct surface-aligned appearance and estimate the accurate human body pose by regularizing the Gaussian splats and the auxiliary motion field introduced in [24, 50].

Photometric Reconstruction Loss We optimize the multi-person avatar representations from multi-view RGB videos via the photometric reconstruction loss. We compute the $L1$ difference and the structure similarity image metric:

$$\mathcal{L}_c = \sum_{\mathbf{r} \in \mathcal{R}} ||\mathbf{c}(\mathbf{r}) - \hat{\mathbf{c}}(\mathbf{r})||_1 + \lambda_{SSIM} \cdot \text{SSIM}(\mathbf{C}_k, \hat{\mathbf{C}}_k), \quad (3)$$

where $\mathbf{r}$ is a pixel or its ray, $\mathcal{R}$ is a set of pixels at the iteration, $\hat{\mathbf{c}}(\cdot)$ is the pixel observation value, n_i is the number of all video frames, λ_{ssim} is a linear coefficient set to 0.1, and $\mathbf{C}_i$ and $\hat{\mathbf{C}}_k$ is the k-th rendering image and the observation image, respectively.

Surface Ordering Loss In order to encourage clear boundaries between avatars and prevent incorrect depth orders of their surfaces, we use the instance segmentation map as the surface visibility map. We compute the cross entropy loss to correct the surface ordering:

$$\mathcal{L}_{so} = -\sum_{\mathbf{r} \in \mathcal{R}} \sum_{o=1}^{n_o} \log \frac{\exp(\mathbf{s}_o(\mathbf{r}))}{\sum_{i=1}^{n_o} \exp(\mathbf{s}_i(\mathbf{r}))} y_o(\mathbf{r}), \quad (4)$$

where $y_o(\cdot)$ is a mask map for instance o and $s_i(\cdot)$ is a probability map of the i-th avatar computed as the opacity map. The segmentation map is inferred by SAM2 [34]. We manually prompt the segmentation map to fix the incorrectly segmented regions and refine it via guide filtering as the initial inferred map is not accurate near the instance boundaries.

Monocular Geometry Loss To improve the reconstruction quality in under-observed and textureless areas, we aggregate the monocular geometry priors from all video frames. We use the state-of-the-art human vision foundation model Sapiens [21] as a teacher model that infers plausible and natural monocular depth and normal maps for clothed humans. Due to the scale and shift ambiguity in monocular depth $\hat{d}$, we compute the scale-shift invariant loss for depth supervision [50]:

$$\mathcal{L}_d = \min_{w,q} \sum_{\mathbf{r} \in \mathcal{R}} \|(\omega \cdot d(\mathbf{r}) + q) - \hat{d}(\mathbf{r})\|_2^2, \quad (5)$$

where scale w and translation q are estimated for each RGB frame. For normal supervision, we compute the cosine distance and the L1 distance with monocular normal $\hat{\mathbf{n}}$:

$$\mathcal{L}_n = \sum_{\mathbf{r} \in \mathcal{R}} \|\mathbf{n}(\mathbf{r}) - \hat{\mathbf{n}}(\mathbf{r})\|_1 + \|1 - \mathbf{n}(\mathbf{r}) \cdot \hat{\mathbf{n}}(\mathbf{r})\|_1. \quad (6)$$

Total Loss Our total loss includes all losses introduced above and the regularization terms to encourage planar Gaussian splats to construct surface-aware appearance [12] and to mitigate the over-deviation in the representation and the motion field [24]:

$$\begin{aligned}\mathcal{L} =& \lambda_c \mathcal{L}_c + \lambda_{so} \mathcal{L}_{so} + \lambda_d \mathcal{L}_d + \lambda_n \mathcal{L}_n \\ &+ \lambda_{rn} \mathcal{L}_{rn} + \lambda_{rd} \mathcal{L}_{rd} + \lambda_{reg} \mathcal{L}_{reg},\end{aligned}$$

where $\lambda_{(\cdot)}$ is a loss coefficient of each term, $\mathcal{L}_{rn}$ is the normal consistency term between Gaussian's normals and the gradients of depth map, $\mathcal{L}_{rd}$ is the depth concentration term for Gaussian splats intersected with a ray, $\mathcal{L}_{reg}$ is the regularization term introduced in [24]. Please refer to [12, 24] for details.

4. Implementation Details

Pose Initialization and Refinement The accurate initial body pose and hand pose are critical for reconstructing the high-quality avatar appearance. To this end, we first train an SMPL-X [31], a human body template model with hand pose and facial expression, via the keypoint projection consistency [31, 53] for each person in multi-view video frames. We penalize the self-collision and force the naturalness and the temporal smoothness of estimated human models during training to alleviate the noticeable artifacts. Note that human keypoints of each avatar are estimated via the human vision model [21].

Method	Sequence	Appearance			Geometry			
		PSNR↑	SSIM↑	LPIPS↓	P2S↓	CD↓	COS↓	L2↓
EasyMocap	Backhug02	29.3978	0.9616	0.0499	0.5132	0.9070	0.0355	0.0701
	Hug01	28.3863	0.9508	0.0661	0.5910	0.9936	0.0477	0.0929
	Talk22	27.1694	0.9403	0.0764	0.7020	1.3125	0.0521	0.0997
	Dance15	24.9144	0.9373	0.0873	0.6329	1.1100	0.0524	0.1018
	Football21	24.2895	0.9337	0.0790	0.9163	1.5744	0.0543	0.1013
Multi-GART	Backhug02	29.9026	0.9633	0.0622	1.2458	1.7710	0.0290	0.0611
	Hug01	29.3392	0.9558	0.0746	1.5944	2.0274	0.0363	0.0767
	Talk22	28.8042	0.9493	0.0814	1.2339	1.7077	0.0354	0.0765
	Dance15	25.9205	0.9464	0.0821	1.1377	1.6761	0.0355	0.0751
	Football21	25.7526	0.9415	0.0894	1.3394	1.9340	0.0368	0.0769
Ours	Backhug02	**31.8115**	**0.9705**	**0.0417**	**0.4806**	**0.5894**	**0.0113**	**0.0311**
	Hug01	**31.2331**	**0.9643**	**0.0517**	**0.5034**	**0.6136**	**0.0138**	**0.0387**
	Talk22	**30.1691**	**0.9558**	**0.0586**	**0.5640**	**0.6359**	**0.0140**	**0.0401**
	Dance15	**27.8146**	**0.9546**	**0.0563**	**0.4769**	**0.4971**	**0.0157**	**0.0419**
	Football21	**27.3634**	**0.9490**	**0.0640**	**0.5550**	**0.5915**	**0.0158**	**0.0420**

Table 1. Quantitative comparison on the Hi4D Dataset [49] that captures multi-person close-interaction scenarios. We compare our method against EasyMocap [37] and multi-GART [24], with the latter extended by us for multi-person reconstruction. Our method achieves state-of-the-art performance in both rendering quality and geometry accuracy.

However, the human poses estimated above suffer from the inter-penetration among avatars. We refine the initial pose parameters via the surface ordering loss between the projected human meshes and the segmentation map defined in Eq. 4. Our refined human poses agree with the actual surfaces without modeling surface deformation and the inner-outer test.

Training Details We initialize Gaussian splats at vertices of the human template model in the canonical space. We optimize poses and Gaussian splats for every avatar simultaneously during training. We adaptively control the number of Gaussian splats in a heuristic way [12, 20]. We observed that training poses in the earlier stage suffer from quality loss due to incorrectly estimated appearance. We first train the avatar appearance with fixed initial avatar poses. After the avatar appearance is roughly estimated, we start joint optimization of poses and avatar appearances at the half-time of the training iterations. Our method completes total 3000 training iterations which takes 10 minutes on a single NVIDIA RTX A6000 GPU. Empirically, we set the scaling factors $(\lambda_c, \lambda_{so}, \lambda_d, \lambda_n, \lambda_{rn}, \lambda_{rd}, \lambda_{reg})$ to $(1.0, 1.0, 0.05, 0.025, 0.1, 0.05, 0.01, 0.5)$, respectively.

Geometry Extraction To extract the mesh of each frame space, we utilize the volumetric fusion method [2] introduced in [12]. We transform Gaussian splats to the target space, render them for uniformly sampled camera views, and obtain multi-view depth maps. The truncated signed distance function (TSDF) of each depth map is then cumulated and stored in the lattice grid. The final mesh is extracted from the integrated TSDF field using the marching cube algorithm [25]. We sample 100 camera views on the bounding hemisphere positioned at distances similar to the observation views.

5. Experiments

Dataset In order to analyze the overall performance of our multi-person reconstruction including geometric accuracy, we use the Hi4D dataset [49] that captures multiple multi-person interaction scenarios through 8 synchronized static cameras and provides raw 3D scans and highly accurate estimated individual poses per frame as ground-truth. Each scenario consists of close human interaction between two individuals across 200 video frames. For our evaluation, we select video sequences that clearly exhibit close contacts: *backhug02*, *hug01*, *talk22*, *dance15*, and *football21*. To evaluate the general performance for novel poses, we split the frames into training and testing sets. We sample training frames every 4 frames starting from frame 0, and test frames every 8 frames starting from frame 2.

Evaluation Metrics We evaluate the rendering quality using PSNR, SSIM, and LPIPS metrics, and the geometric quality using point-to-surface (P2S) and Chamfer distance (CD) metrics. For surface orientation assessment, we measure normal consistency by cosine distance (COS) and L_2 distance (L2).

Comparison with Other MP Reconstruction We compare our method with two multi-person reconstruction methods as baselines, EasyMocap [37] and multi-GART,

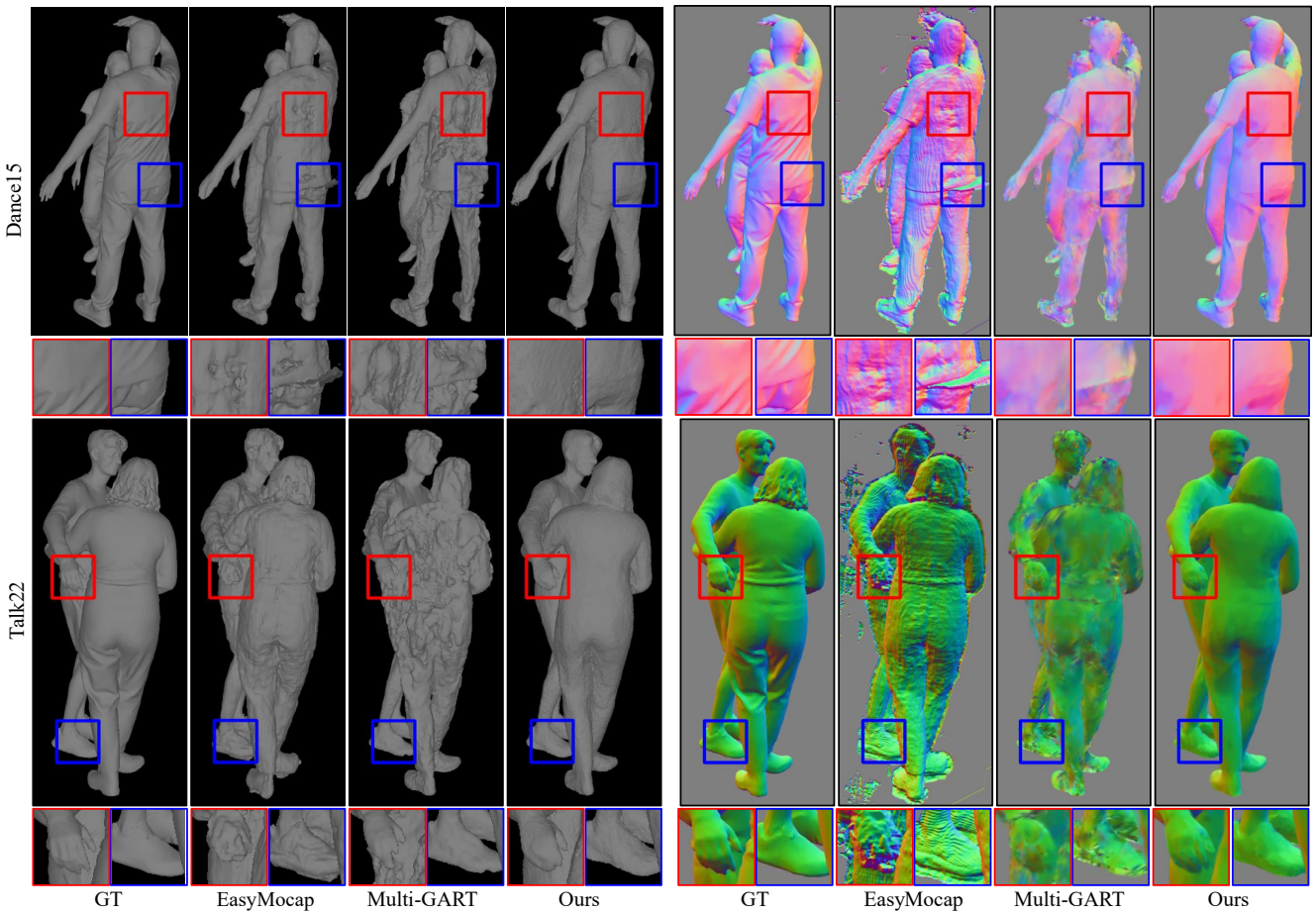

Figure 3. Qualitative comparison against baselines [24, 37]. EasyMocap [37] suffers from rugged surfaces, incorrect surfacing near boundaries, and cloud artifacts. Multi-GART, the multi-person extension of GART [24], struggles with surface alignment and fails to reconstruct thin structures such as fingers. Our method achieves the geometrically consistent surface reconstruction.

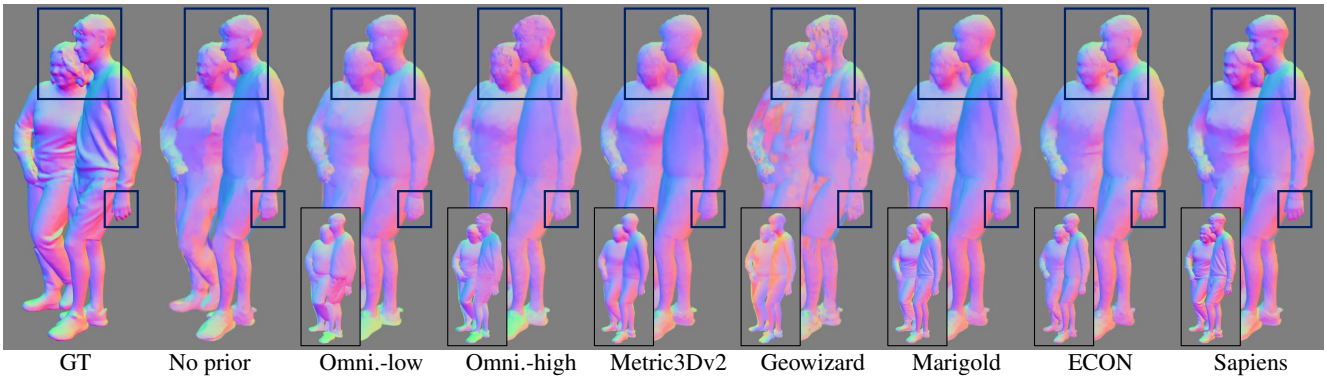

Figure 4. Qualitative comparison of monocular normal priors [5, 8, 11, 19, 21, 40, 43]. We leverage monocular normal priors (thumbnail) to reconstruct multiple avatars (large). Reconstruction with Sapiens [21], the human-centric vision model, achieves the top performance in geometric accuracy.

the multi-person version of GART [24]. We implemented multi-GART that includes volumetric rendering with multiple avatars transformed in the target space. Multi-GART does not apply any monocular priors and does not use the surface ordering loss. While EasyMocap supports human pose and shape estimation, multi-GART requires pose ini-

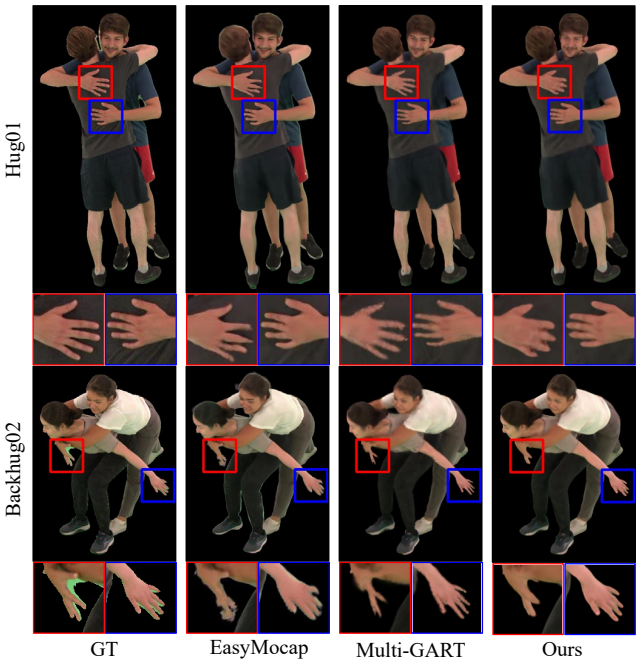

Figure 5. Qualitative comparison against EasyMocap [37] and Multi-GART [24].

tialization. We input the ground-truth poses included in the dataset into Multi-GART.

Tab. 1 presents a quantitative comparison with other state-of-the-art methods. EasyMocap, which uses an implicit neural representation [27], exhibits the longest training time (2–3 days) yet shows the lowest performance.In contrast, multi-GART, a state-of-the-art method in single-person avatar reconstruction, achieves high rendering quality even in multi-person scenarios. However, due to its surface ambiguity [9, 12], multi-GART shows inaccurate geometry reconstruction. On the other hand, our method achieves superior results in both rendering quality and geometric accuracy.

Fig. 3 and Fig. 5 qualitatively demonstrate the geometric and radiometric performance of our method against the baselines. Although EasyMocap builds a background color radiance field model, it fails to separate background colors from foreground instances. EasyMocap reconstructs noisy surfaces in the textureless regions and generates layered surfaces near boundaries and small clouds in the empty region. Multi-GART produces inconsistent surface geometry due to unwanted shape-texture entanglement in the volumetric radiance field [9, 12]. Multi-GART suffers from blob artifacts near the instance boundaries or thin structures like fingers during image rendering. Thanks to our surface-aware radiance field representation with the surface regularization terms, our method constructs smooth and complete surfaces that are well-aligned with the ground-truth surfaces and produces smooth detailed texture appearance.

Models	PSNR↑	SSIM↑	LPIPS↓	P2S↓	CD↓	COS↓	L2↓
W/o mono. prior	30.1219	0.9554	0.0603	0.6696	0.7720	0.0180	0.0485
Omnidata-low [5]	30.1674	0.9553	0.0617	0.6573	0.8038	0.0186	0.0516
Omnidata-high [40, 50]	30.1799	0.9553	0.0617	0.6948	0.8841	0.0190	0.0529
Metric3Dv2 [11]	30.1618	0.9555	0.0606	0.5927	0.6648	0.0164	0.0463
Geowizard [8]	**30.2167**	0.9547	0.0633	0.9213	1.4880	0.0284	0.0685
Marigold [19]	30.1811	0.9556	0.0599	0.5868	0.6791	0.0156	0.0439
ECON [43]	30.1871	0.9557	0.0598	0.6001	0.7038	0.0165	0.0470
Sapiens [21]	30.1678	**0.9558**	**0.0586**	**0.5656**	**0.6480**	**0.0142**	**0.0403**

Table 2. Quantitative comparison of monocular normal priors on sequence talk22.

Models	PSNR↑	SSIM↑	LPIPS↓	P2S↓	CD↓	COS↓	L2↓
W/o mono. prior	30.1325	0.9555	0.0602	0.6720	0.7691	0.0181	**0.0485**
ZoeDepth [1]	30.1540	0.9555	0.0602	0.6710	0.7654	0.0181	0.0486
Metric3D [11]	30.1478	0.9555	0.0602	**0.6682**	0.7721	0.0181	0.0486
Geowizard [8]	30.1504	**0.9556**	**0.0601**	0.6727	0.7730	0.0181	0.0486
Marigold [19]	**30.1617**	**0.9556**	0.0602	0.6716	**0.7627**	**0.0180**	**0.0485**
DA-v2 [47]	30.1320	0.9555	**0.0601**	0.6685	0.7666	0.0181	**0.0485**
Sapiens [21]	30.1500	0.9555	**0.0601**	0.6690	0.7820	0.0181	0.0487

Table 3. Quantitative comparison of monocular depth priors on sequence talk22.

Methods	PSNR↑	SSIM↑	LPIPS↓	P2S↓	CD↓	COS↓	L2↓
W/o L_{rn}, L_{rd}	29.8348	0.9591	0.0553	0.5856	0.6972	0.0152	0.0406
W/o L_r, L_n	29.6619	0.9586	0.0561	0.5836	0.6846	0.0172	0.0452
Ours	29.6783	0.9588	0.0545	0.5160	0.5855	0.0141	0.0388

Table 4. Quantitative analysis for ablation study of geometric losses. We examine both rendering and geometric performance while excluding each objective term. Our full model achieves the best geometric reconstruction while maintaining the rendering quality.

	mIoU Low 1%↑	mIoU Low 5%↑	mIoU Low 10%↑	mIoU↑
Multi-GART	0.6966	0.7263	0.7561	0.9145
W/o L_{so}	0.6128	0.6787	0.7385	0.9349
Ours	**0.7003**	**0.7475**	**0.7882**	**0.9381**

Table 5. Quantitative comparison of instance segmentation maps. We compute mean Intersection over Union (mIoU) over the foreground area of every test frame. We divide the frame into 200 patches and select the lowest 1%, 5%, and 10% IoU patches from W/o $\mathcal{L}_{so}$ and evaluate other methods at the same patch locations.

Comparison among Monocular Geometric Priors We first examine the effect of monocular normal priors including Omnidata-low [5], Omnidata-high [40, 50], Metric3Dv2 [11], GeoWizard [8], Marigold [19], ECON [43], and Sapiens [21]. To this end, we only activate monocular normal loss during training. in Tab. 2 and Fig. 4. Omnidata-low and Omnidata-high degenerate geometry reconstruction especially in normal orientations. Geowizard fails to estimate geometrically consistent normals, resulting in inconsistent avatar reconstruction. Metric3Dv2 and ECON enhance the reconstruction quality but the reconstructions are overly smooth in faces and clothes. Marigold helps with consistent reconstruction on clothes but shows low quality on face region. Sapiens, a human centric vision foundation model, successfully detects details of clothes and faces, achieving the top performance. We use Sapiens as our teacher network.

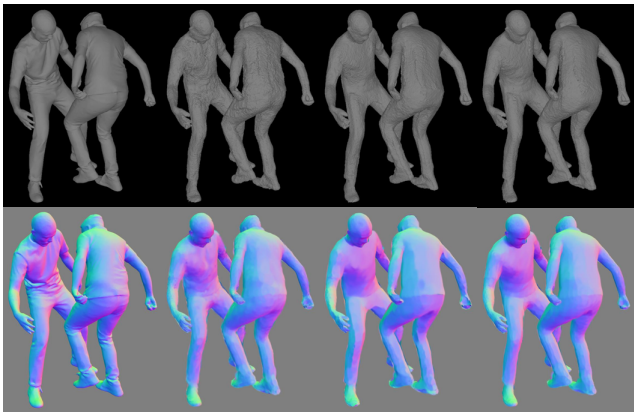

Figure 6. Qualitative analysis of ablation study. The Gaussian regularization terms $\mathcal{L}_{rn}$ and $\mathcal{L}_{rd}$ reconstruct the smooth surfaces. Training monocular geometric priors $\mathcal{L}_n$ and $\mathcal{L}_d$ reduce the angled surfaces. Our full model reconstructs high quality surfaces.

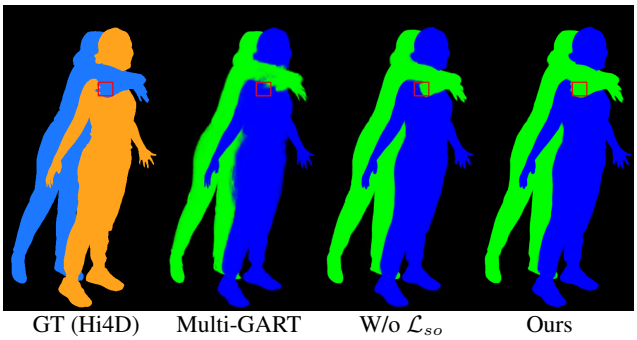

Figure 7. Qualitative comparison for surface ordering. The surface-aware avatar model allows the reconstruction of clear surface boundaries and the surface ordering loss encourages the extraction of penetrating parts.

Tab. 3 shows the influence of state-of-the-art monocular depth priors. We observe that the monocular depth loss makes subtle improvement with any depth priors that are trained to be consistent with normal priors. This implies that aligning orientations of Gaussian splats effectively construct the surface-consistent appearance while translating the Gaussian splat doesn't make a significant improvement.

Ablation Study Tab. 4 and Fig. 6 demonstrate the effectiveness of each geometric objective term in both quantitative and qualitative results. The regularization term for planar Gaussian splats [12] enhances the surface smoothness. Training monocular geometric priors enables the reconstruction of curved surfaces through their ability to predict geometric gradients from images, reducing the angled surfaces. By combining all of them, our full model achieves state-of-the-art performance with competitive rendering quality.

To demonstrate the effectiveness of our surface ordering loss, we introduce a novel metric that measures the mean intersection-over-union (mIoU) in tightly interacting regions. We first find potentially-contacted patch regions that could suffer from poor surface ordering. We select patches with the lowest mIoUs trained without the surface ordering loss. Then, we compare them with the corresponding patches (red box in Fig. 7) computed with surface ordering. Tab. 5 and Fig. 7 illustrate that our surface ordering loss enhances depth ordering in potentially-contacted regions and the quality of avatar boundaries. Compared to Multi-GART, our full model produces more distinct object boundaries and alleviates the inter-penetration issues.

6. Discussion

Despite achieving state-of-the-art performance in multi-person reconstruction, our method has some limitations. Since we fit avatar models in the canonical space using only the deformation field, a more flexible motion field is needed to capture fine dynamic details like wrinkles. Additionally, our model reconstructs static surface colors and normals, leading to smooth surfaces when temporal variations occur. Integrating Vid2Avatar [10], which enables pose-dependent surface colors, could help preserve dynamic appearance changes. Lastly, our method relies on monocular geometric priors, inheriting any biases they introduce. Physically-based inverse rendering could mitigate this [41, 46] by leveraging shape-from-shading constraints to refine surface details.

7. Conclusion

We propose the surface-aware avatar representation that uses planar Gaussian splats with surface regularization terms, facilitating the surface-aligned appearance reconstruction. To robustly reconstruct multi-person avatars against inter-penetrations and occlusions between avatars in close-interaction scenarios, we apply the monocular geometric and instance priors by modeling the objective functions between the monocular predictions and the estimations. We compare the state-of-the-art monocular geometric priors in multi-person reconstruction to find the best associated prior. Our full model achieves state-of-the-art performance on dynamic multi-person reconstruction from multi-view videos while taking around 5 minutes to train.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant (No. RS-2023-00212828) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No. 2018-0-00207, Immersive Media Research Laboratory), both funded by the Korea government (MSIT).

References

[1] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2, 7

[2] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 5

[3] P. Dai, J. Xu, W. Xie, X. Liu, H. Wang, and W. Xu. High-quality surface reconstruction using gaussian surfels. In *ACM SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. 2

[4] M. Dou, H. Fuchs, and J. M. Frahm. Scanning and tracking dynamic objects with commodity depth cameras. In *2013 IEEE international symposium on mixed and augmented Reality (ISMAR)*, pages 99–106. Ieee, 2013. 2

[5] A. Eftekhar, A. Sax, Alexander, J. Malik, and A. Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, pages 10786–10796, 2021. 2, 6, 7

[6] M. Fieraru, M. Zanfir, E. Oneata, A. I. Popa, V. Olaru, and C. Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, pages 7214–7223, 2020. 2

[7] M. Fieraru, M. Zanfir, T. Szente, E. Bazavan, V. Olaru, Vlad, Sminchisescu, and Cristian. Remips: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. *NeurIPS*, 34:19385–19397, 2021. 1, 2

[8] X. Fu, W. Yin, M. Hu, K. Wang, Y. Ma, P. Tan, S. Shen, D. Lin, and X. Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *ECCV*, 2024. 2, 6, 7

[9] A. Guédon and V. Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *CVPR*, pages 5354–5363, 2024. 2, 7

[10] C. Guo, T. Jiang, X. Chen, J. Song, and O. Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *CVPR*, pages 12858–12868, 2023. 2, 8

[11] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024. 2, 6, 7

[12] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3, 4, 5, 7, 8

[13] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. VolumeDeform: Real-time Volumetric Non-rigid Reconstruction. 2016. 2

[14] B. Jiang, Y. Hong, H. Bao, and J. Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *CVPR*, 2022. 2

[15] T. Jiang, X. Chen, J. Song, and O. Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. *CVPR*, 2023. 2

[16] W. Jiang, N. Kolotouros, G. Pavlakos, X. Zhou, and K. Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020. 2

[17] W. Jiang, K. M. Yi, G. Samei, O. Tuzel, and A. Ranjan. Neuman: Neural human radiance field from a single video. In *ECCV*, 2022. 1, 2, 3

[18] Z. Jiang, C. Guo, M. Kaufmann, T. Jiang, J. Valentin, O. Hilliges, and J. Song. Multiply: Reconstruction of multiple people from monocular video in the wild. In *CVPR*, 2024. 1, 2

[19] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 2, 6, 7

[20] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4), 2023. 2, 5

[21] R. Khirodkar, T. Bagautdinov, J. Martinez, S. Zhaoen, A. James, P. Selednik, S. Anderson, and S. Saito. Sapiens: Foundation for human vision models. *ECCV*, 2024. 1, 2, 4, 6, 7

[22] S. Kim, J. Son, G. Ju, J. Lee, and S. Lee. High-quality geometry and texture editing of neural radiance field. 2024. 2

[23] M. Kocabas, J. H. R. Chang, J. Gabriel, O. Tuzel, and A. Ranjan. HUGS: Human gaussian splatting. In *CVPR*, 2024. 1

[24] J. Lei, Y. Wang, G. Pavlakos, L. Liu, and K. Daniilidis. Gart: Gaussian articulated template models. In *CVPR*, pages 19876–19887, 2024. 1, 2, 3, 4, 5, 6, 7

[25] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 5

[26] F. Lu, Z. Dong, J. Song, and O. Hilliges. Avatarpose: Avatar-guided 3d pose estimation of close human interaction from sparse multi-view videos. In *ECCV*, 2024. 1, 2

[27] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 7

[28] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 41(4):1–15, 2022. 2

[29] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011. 2

[30] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015. 2

[31] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands and face and body from a single image. In *CVPR*, 2019. 1, 2, 3, 4

[32] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. Neural body: Implicit neural representations with

structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 1, 2

[33] Z. Qian, S. Wang, M. Mihajlovic, A. Geiger, and S. Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *CVPR*, pages 5020–5030, 2024. 2

[34] N. Ravi, V. Gabeur, Y. T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C. Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 4

[35] R. René, B. Alexey, and K. Vladlen. Vision transformers for dense prediction. *ICCV*, 2021. 2

[36] Z. Shao, Z. Wang, Z. Li, D. Wang, X. Lin, Y. Zhang, M. Fan, and Z. Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *CVPR*, 2024. 2

[37] Q. Shuai, C. Geng, Q. Fang, Qi S. Peng, W. Shen, X. Zhou, and H. Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 1, 2, 5, 6, 7

[38] R. W. Sumner, J. Schmid, and M. Pauly. Embedded deformation for shape manipulation. In *ACM siggraph 2007 papers*, pages 80–es. 2007. 2

[39] A. Tevs, A. Berner, M. Wand I. Ihrke, M. Bokeloh, J. Kerber, and H. P. Seidel. Animation cartography—intrinsic reconstruction of shape and motion. *ACM TOG*, 31(2), 2012. 2

[40] M. Turkulainen, X. Ren, I. Melekhov, O. Seiskari, E. Rahtu, and J. Kannala. Dn-splatter: Depth and normal priors for gaussian splatting and meshing, 2024. 2, 6, 7

[41] S. Wang, B. Antić, A. Geiger, and S. Tang. Intrinsicavatar: Physically based inverse rendering of dynamic humans from monocular videos via explicit ray tracing. In *CVPR*, 2024. 8

[42] C. Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, pages 16210–16220, 2022. 1, 2

[43] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *CVPR*, 2023. 2, 6, 7

[44] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann. Point-nerf: Point-based neural radiance fields. In *CVPR*, pages 5438–5448, 2022. 2

[45] T. Xu, Y. Fujita, and E. Matsumoto. Surface-aligned neural radiance fields for controllable 3d human synthesis. In *CVPR*, 2022. 2

[46] Z. Xu, S. Peng, C. Geng, L. Mou, Z. Yan, J. Sun, H. Bao, and X. Zhou. Relightable and animatable neural avatar from sparse-view video. In *CVPR*, 2024. 8

[47] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 7

[48] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 2

[49] Y. Yin, C. Guo, M. Kaufmann, J. Zarate, J. Song, and O. Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *CVPR*, 2023. 1, 2, 5

[50] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *NeurIPS*, 2022. 1, 2, 4, 7

[51] A. Zanfir, E. Marinoiu, and C. Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, pages 2148–2157, 2018. 1, 2

[52] J. Zhang, X. Liu, X. Ye, F. Zhao, Y. Zhang, M. Wu, Y. Zhang, L. Xu, and J. Yu. Editable free-viewpoint video using a layered neural representation. *ACM TOG*, 40(4):1–18, 2021. 2

[53] Z. Zheng, T. Yu, Y. Liu, and Q. Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE TPAMI*, pages 1–1, 2021. 4

[54] H. Zhu, F. Zhan, C. Theobalt, and M. Habermann. Trihuman: A real-time and controllable tri-plane representation for detailed human geometry and appearance synthesis. *ACM TOG*, 44(1):1–17, 2024. 2