NeuralDome: A Neural Modeling Pipeline on Multi-View Human-Object Interactions

Juze Zhang [1,2,3,4,*], Haimin Luo [1,4,5,*], Hongdi Yang [1,4], Xinru Xu [1,4], Qianyang Wu [1,4], Ye Shi [1,4], Jingyi Yu [1,4], Lan Xu [1,4,†], Jingya Wang [1,4,†]
[1] ShanghaiTech University [2] Shanghai Advanced Research Institute, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences
[4] Shanghai Engineering Research Center of Intelligent Vision and Imaging [5] LumiAni Technology
{zhangjz,luohm,yanghd,xuxr2022,wuqy2022,shiye,yujingyi,xulan1,wangjingya}@shanghaitech.edu.cn

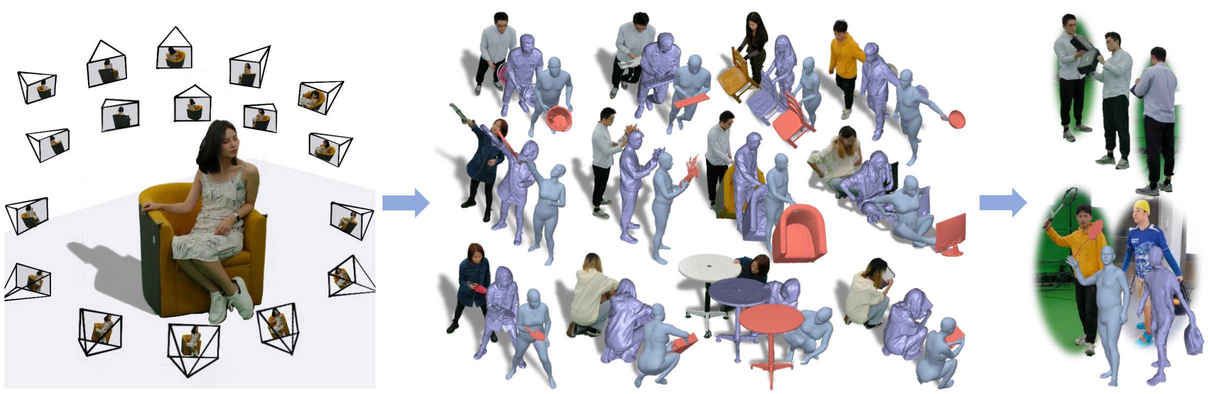

Figure 1. Our NeuralDome pipeline for processing multi-view video sequences on human object interactions. NeuralDome supports tracking, modeling, and rendering of individual human subjects and objects. To validate NeuralDome, we collect a large dataset HODome over a total of 71 million video frames across 76 viewpoints and process the datasets using NeuralDome for a variety of inference and neural modeling and rendering tasks.

Abstract

Humans constantly interact with objects in daily life tasks. Capturing such processes and subsequently conducting visual inferences from a fixed viewpoint suffers from occlusions, shape and texture ambiguities, motions, etc. To mitigate the problem, it is essential to build a training dataset that captures free-viewpoint interactions. We construct a dense multi-view dome to acquire a complex human object interaction dataset, named HODome, that consists of ~71M frames on 10 subjects interacting with 23 objects. To process the HODome dataset, we develop NeuralDome, a layer-wise neural processing pipeline tailored for multi-view video inputs to conduct accurate tracking, geometry reconstruction and free-view rendering, for both human subjects and objects. Extensive experiments on the HODome dataset demonstrate the effectiveness of NeuralDome on a variety of inference, modeling, and rendering tasks. Both the dataset and the NeuralDome tools will be disseminated to the community for further development, which can be found at https://juzezhang.github.io/NeuralDome

1. Introduction

A key task of computer vision is to understand how humans interact with the surrounding world, by faithfully capturing and subsequently reproducing the process via mod-

* These authors contributed equally.
†Corresponding author.

eling and rendering. Successful solutions benefit broad applications ranging from sports training to vocational education, digital entertainment to tele-medicine.

Early solutions [11, 12, 51] that reconstruct dynamic meshes with per-frame texture maps are time-consuming and vulnerable to occlusions or lack of textures. Recent advances in neural rendering [37, 62, 71] bring huge potential for human-centric modeling. Most notably, the variants of Neural Radiance Field (NeRF) [37] achieve compelling novel view synthesis, which can enable real-time rendering performance [38, 59, 67] even for dynamic scenes [45, 63, 77], and can be extended to the generative setting without per-scene training [23, 69, 80]. However, less attention is paid to the rich and diverse interactions between humans and objects, mainly due to the severe lack of dense-view human-object datasets. Actually, existing datasets of human-object interactions are mostly based on optical markers [61] or sparse RGB/RGBD sensors [6, 26], without sufficient appearance supervision for neural rendering tasks. As a result, the literature on neural human-object rendering [21,57] is surprisingly sparse, let alone further exploring the real-time or generative directions. Besides, existing neural techniques [52,77] suffer from tedious training procedures due to the human-object occlusion, and hence infeasible for building a large-scale dataset. In a nutshell, despite the recent tremendous thriving of neural rendering, the lack of both a rich dataset and an efficient reconstruction scheme constitute barriers in human-object modeling.

In this paper, we present *NeuralDome*, a neural pipeline that takes multi-view dome capture as inputs and conducts accurate 3D modeling and photo-realistic rendering of complex human-object interaction. As shown in Fig. 1, NeuralDome exploits layer-wise neural modeling to produce rich and multi-modality outputs including the geometry of dynamic human, object shapes and tracked poses, as well as a free-view rendering of the sequence.

Specifically, we first capture a novel human-object dome (HODome) dataset that consists of 274 human-object interacting sequences, covering 23 diverse 3D objects and 10 human subjects (5 males and 5 females) in various apparels. We record multi-view video sequences of natural interactions between the human subjects and the objects where each sequence is about 60s in length using a dome with 76 RGB cameras, resulting in 71 million video frames. We also provide an accurate pre-scanned 3D template for each object and utilize sparse optical markers to track individual objects throughout the sequences.

To process the HODome dataset, we adopt an extended Neural Radiance Field (NeRF) pipeline. The brute-force adoption of off-the-shelf neural techniques such as Instant-NSR [79] and Instant-NGP [38], although effective, do not separate objects from human subjects and therefore lack sufficient fidelity to model their interactions. We instead introduce a layer-wise neural processing pipeline. Specifically, we first perform a joint optimization based on the dense inputs for accurately tracking human motions using the parametric SMPL-X model [43] as well as localizing objects using template meshes. We then propose an efficient layer-wise neural rendering scheme where the humans and objects are formulated as a pose-embedded dynamic NeRF and a static NeRF with tracked 6-DoF rigid poses, respectively. Such a layer-wise representation effectively exploits temporal information and robustly tackles the occluded regions under interactions. We further introduce an object-aware ray sampling strategy to mitigate artifacts during layer-wise training, as well as template-aware geometry regularizers to enforce contact-aware deformations. Through weak segmentation supervision, we obtain the decoupled and occlusion-free appearances for both the humans and the objects at a high fidelity amenable for training the input multi-view inputs for a variety of tasks from monocular motion capture to free-view rendering from sparse multi-view inputs.

To summarize, our main contributions include:

- We introduce NeuralDome, a neural pipeline, to accurately track humans and objects, conduct layer-wise geometry reconstruction, and enable novel-view synthesis, from multi-view HOI video inputs.
- We collect a comprehensive dataset that we call HODome that will be disseminated to the community, with both raw data and the output modalities including separated geometry and rendering of individual objects and human subjects, their tracking results, free-view rendering results, etc.
- We demonstrate using the dataset to train networks for a variety of visual inference tasks with complex human object interactions.

2. Related Works

2.1. Neural Human Rendering

Various 3D data representations have been explored for neural human rendering, such as point-clouds [4, 41, 72], textured meshes [14, 28, 29, 54], and volumes [31, 32]. Implicit occupancy function-based methods [17, 19, 48, 49] can recover detailed 3D human geometry from sparse 2D images without faithful appearance synthesis. The recent NeRF [37] technique brings huge potential for 3D photo-realistic view synthesis [9, 10, 30, 34, 36, 38, 39, 64, 69, 75] and geometry modeling [7, 33, 68]. Further explorations extend it to dynamic scenes [7, 35, 42, 46, 67, 77], especially for humans. Existing works equip NeRF with pose-embeddings [23, 27, 40, 45, 80], learnable skinning weights [25, 44, 70] and even generalization across individuals [23, 66, 80]. However, these works only focus on

Datasets	#Cam-view	# Frame(M)	Resolution	Fps	Marker	Obj. Num.	Human Annot.	Novel views	Geometry	Human Textured	Neural Representation	Object Appearance
NTU [26]	3	$\sim$ 34	1920×1080	30	$\times$	NA	✓	NA	$\times$	$\times$	$\times$	$\times$
PiGr [50]	1	0.1	960×540	5	$\times$	NA	✓	NA	$\times$	$\times$	$\times$	✓
GRAB [61]	$\times$	NA	NA	NA	✓	51	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$
PROX [16]	1	0.1	1920×1080	30	✓	NA	✓	$\times$	$\times$	$\times$	$\times$	✓
BEHAVE [6]	4	0.15	2048×1536	30	$\times$	20	$\times$	$\times$	$\times$	$\times$	$\times$	✓
InterCap [18]	6	0.07	1920×1080	30	$\times$	10	✓	$\times$	$\times$	$\times$	$\times$	✓
Our	76	71	3840×2160	60	✓	23	✓	✓	✓	✓	✓	✓

Table 1. **Dataset Comparisons**. We compare our proposed HODome dataset with the existing publicly available human-object dataset. HODome has the largest scale of human-object interactions in terms of the number of frames (#Frame), camera view (#Cam-view), and modality. "Obj. Num." represents the object number. "Human Annot." represents annotation from professional annotators.

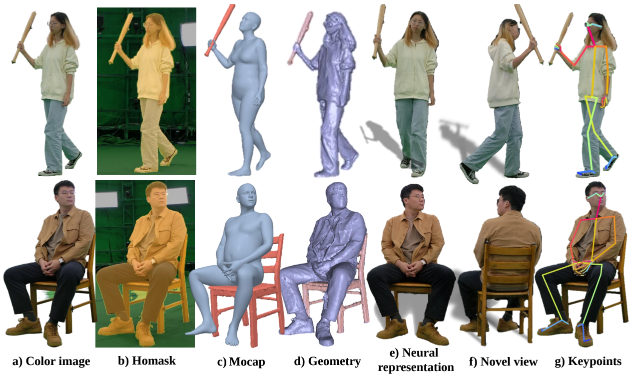

Figure 2. **HODome Modality.** HODome features multiple modalities of data format and annotations, including a) Color image, b) Human-object mask (Homask), c) MoCap (SMPL-X parameters and located object template), d) Geometry, e) digital assets in neural representations, f) Novel view and g) Keypoints (25 for bodies and 42 for hands with human annotation).

a single person. The most relevant works [53, 77] aim to model multi-person interactions, but they cannot handle more complex interactions due to the lack of large-scale human-object interactions dataset.

2.2. Human-object Modeling

Only a few works [6,16,18,61,73,78] consider to jointly model whole-body interactions. Early solutions [11, 12, 51] that reconstruct meshes with per-frame texture maps, which are vulnerable to occlusions and lack of textures. Recently, several works [6, 16, 18, 61] explore the relationship via several interaction constraints, such as contact map [6, 18, 61], spatial arrangement [78] and physically plausible constraint [74]. However, these methods only produce a relative arrangement. The most relevant works [21, 56, 58] aim to render human-object interactions with volumetric fusion [56], neural texturing blending [58] and volumetric rendering [21]. However, without sufficient appearance supervision, their quality is limited. Comparably, our layer-wise neural representations can be converted to high-quality geometry and support separate free-view rendering for human or object.

2.3. Human-centric Dataset

A variety of human-centric datasets have been developed for human-only capturing or rendering tasks. Early datasets combine multi-camera RGB video capture with synchronized ground-truth 3D skeletons [20, 22, 55], parametric model [24, 65], scanned mesh [15, 76], without human-object interaction. Recent datasets capture human-object interactions using optical markers [61] or sparse RGB sensors [6, 26], without sufficient appearance supervision. As a result, the literature on neural human-object rendering [21, 57] is surprisingly rare. High-end works [11, 13] use dense cameras for reconstruction and rendering of humans and objects through mesh reconstruction and motion tracking, but without texture and neural representations. To fill this gap, we propose NeuralDome for capturing and rendering human-object interactions, facilitating various generative human-object tasks.

3. HODome Dataset

We present the HODome dataset for capturing and rendering photo-realistic human-object interactions. It consists of 274 human-object interacting sequences, covering 23 diverse 3D objects and 10 subjects (5 males and 5 females) under various apparels. For each sequence, we record the naturally interacting scene for about 60s using 76 RGB cameras for dense-view at 3840 $\times$ 2160 resolution and 60 frame-per-seconds (Fps), resulting in roughly 71 million frames. HODome consists of rich labels covering different aspects of HOI capturing and rendering labels (see Fig.2). See Tab. 1 for comparison with other datasets.

3.1. Data Capturing System

To construct the HODome dataset, we use 76 Z-CAM cinema cameras with sufficient appearance supervision for neural rendering tasks and 16 Optitrack MoCap cameras [1] for accurate human-object tracking tasks. The Z-CAM system and Optitrack system are synchronized to record the RGB and MoCap data together. We use a publicly available tool [2] to estimate the intrinsic camera parameters and extrinsic camera parameters.

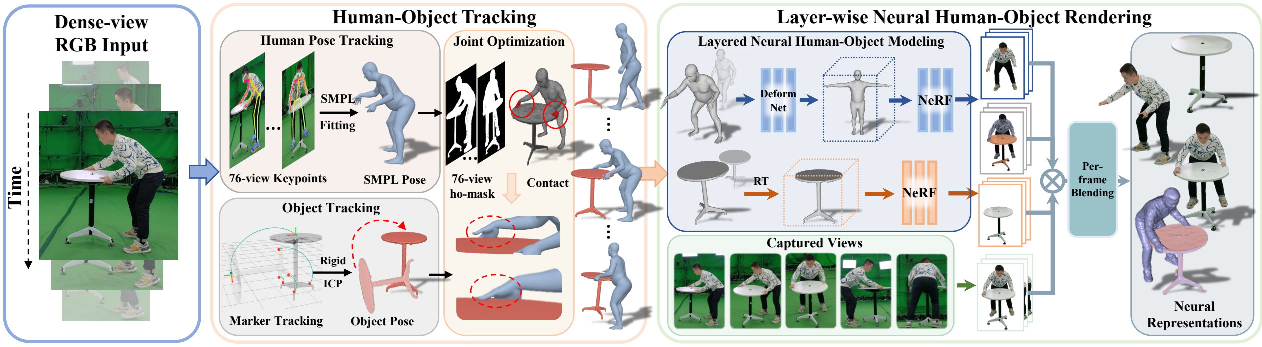

Figure 3. **Overview of NeuralDome.** Given the 76-view RGB stream as input, we first jointly track human skeletal motion and object rigid motion. Then with the tracked motion priors, we decouple the human-object interaction scenes via a layer-wise neural rendering scheme, to generate human/object renderings separately and corresponding acceptable segmentation maps. Blending with the high-fidelity captured views, we obtain the layer-wise neural representations in our HODome dataset.

3.2. Dataset Modality

Human MoCap and Object 6D Pose. For appearance realism, we do not place markers on human actors. Thus we detect the human joints from 76 RGB images by running the whole-body Openpose [8] to perform markerless MoCap. And we treat each object as rigid and solve the rigid object pose estimation from markers using the Iterative Closest Point(ICP) algorithm [5, 81]. See Fig. 3 for the pipeline of our method.

Human-Object Neural Representations. Under the dense-view setting, one can apply off-the-shell neural techniques, i.e., Instant-NSR [79] and Instant-NGP [38] to efficiently obtain per-frame geometry and novel-view synthesis of the whole human-object sequence, respectively. However, neither schemes separate object from human subject. We therefore propose a layer-wise neural scheme to separately recover neural representations of human and object, enabling both high-fidelity geometry reconstruction and free-view appearance rendering (Sec. 4.2).

Data Annotation. Our dataset includes 23 objects with varying scales and interaction types. Each object was pre-scanned using off-the-shell multi-view software packages [2]. Further, following the previous methods [6, 61], we annotated its pseudo contact label that computes from a threshold distance. To provide more accurate results for the quantitative benchmark, we annotate a separate *quantitative subset* as our test set with human-annotated segmentation and hand joints by the professional annotator. Our datasets will be available for research purposes.

4. Neural Modeling on HODome

We introduce a neural pipeline to produce the rich digital assets from the dense-view input of each human-object interaction (HOI) sequence in HODome, including accurate tracking, high-quality geometry reconstruction and novel-view rendering. As illustrated in Fig. 3, given 76-view videos, we first perform a joint optimization scheme to accurately capture both the human skeletal motions and object rigid motion (Sec. 4.1). Then, based on the tracked human-object motions and shape priors, we decouple the humans and objects in HOI scenarios via a layer-wise neural human-object rendering scheme and a corresponding HOI-aware optimization strategy (Sec. 4.2). Such digital assets of humans, objects and the entire HOI scenes can naturally enable further geometry and appearance analysis of HOI scenarios.

4.1. Human-object Tracking

Here we introduce our human-object tracking scheme with the aid of object markers.

Tracking initialization. We initialize the human tracking process by fitting the SMPL-X model to the keypoints of each view. Note that we utilize the off-the-shell toolbox Easymocap [3] to get the initial per-frame parameters, i.e., pose θ_t, shape β_t, facial expression ψ_t, and translation γ_t on each frame t. As for the object, following GRAB [61], we regard each object as rigid and only estimate the rotation $R_t \in \mathbb{SO}(3)$ and rigid translation $T_t \in \mathbb{R}^3$ with respect to its pre-scanned template. Specifically, we conduct the rigid-ICP technique [5, 81] to compute the rigid transformation between the per-frame markers, which serves as the initial object pose.

Joint optimization for human-object tracking. With human and object pose initialization, we further perform a joint optimization scheme to ensure correct human-object contact. We impose constraints to ensure plausible interactions as the following form:

$$E(\beta_t, \theta_t, \psi_t, \gamma_t, R_t, T_t) = E_{\text{smpl}} + \lambda_{\text{contact}} E_{\text{contact}} + \lambda_{\text{homask}} E_{\text{homask}} + \lambda_{\text{maker}} E_{\text{maker}}, \quad (1)$$

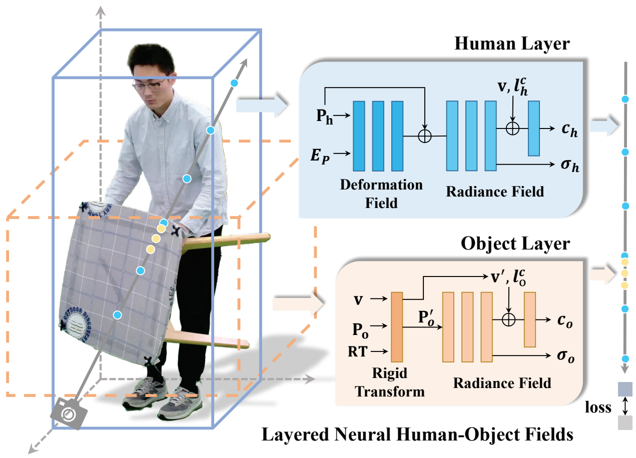

Figure 4. **Layer-wise Neural Human-Object Rendering Details.** We sample a ray uniformly in the bounding box of the performer and near the template for the object. We condition both the human and object layer with appearance codes (l_o^c, l_h^c). The dynamic human layer contains an additional pose embedding (E_p) conditioned deformation field. The sample colors and densities are merged, sorted and accumulated into pixel colors.

where E_{smpl} is the multi-view data fitting term describing the ℓ_2 distance between estimated joints and detected joints, following SMPLify-X [43] and PROX [16]. Besides, we impose contact term E_{contact}, human-object silhouette loss term E_{homask}, and object marker align term E_{maker} to ensure the plausibility of human-object interaction tracking. Due to page limitation, we have to defer more details of these terms in the Appendix.

4.2. Layer-wise Neural Human-Object Rendering

Here we introduce our layer-wise image generation pipeline via neural rendering, without tedious manual efforts for geometry separation.

Layered neural human-object modeling. For ease of instances separation, we represent the human-object interaction (HOI) scenes as a continuous layered neural radiance field, following ST-NeRF [77], as shown in Fig. 4. To leverage the shape prior imposed by the captured parametric SMPL-X model, we adopt a pose embedded dynamic NeRF similar to HumanNeRF [80] as the human layer. We use the skeletal pose to bridge the live frames to canonical space and an additional deformation MLP to learn subtle non-rigid deformation. Rather than using the pixel-aligned image feature, we leverage latent codes for time-varying human appearance capture to get rid of input images. Opposite to the human layer, we model objects as rigid static radiance fields in canonical space. The live frames are transformed to canonical space via object poses to maintain a globally consistent density field. Similarly, we adopt appearance latent codes for the time-varying shadow during interactions.

Dynamic human-object volume rendering. We utilize the layered volume rendering technique described in ST-NeRF [77] to render our neural human-object scenes. For a camera ray intersecting with the i_{th} entity at any timestamp, we compute the ray segment as the depth of intersection points d_f^j and d_f^i. We evenly partition each segment into N bins and sample one point uniformly from each bin:

$$p_j^i \sim \mathcal{U}\left[d_n^i + \frac{j-1}{N}(d_f^i - d_n^i), d_n^i + \frac{j}{N}(d_f^i - d_n^i)\right], j \in [1, 2, ..., N], \tag{2}$$

where $\mathcal{U}$ means uniform distribution. We impose the shape prior encoded in the tracked object template, on the ray sampling scheme for efficient training. We compute the ray-object intersection and uniformly sample a few points in the narrow segment where the first intersection point lies. For human, we simply sample in the ray-bounding box intersection segment. The samples from different segments are then merged and sorted by depth into M samples in total. We then compute a pixel's color by accumulating the radiance at sampled points:

$$C = \sum_{i=1}^{M} T(p_i)[(1 - e^{-\sigma_{p_i}\delta_{p_i}})c_{p_i}], \tag{3}$$

where δ_{p_i} is the distance between adjacent points, c_{p_i} and σ_{p_i} are color and density, and $T(p_i) = \prod_j^{i-1} e^{-\sigma_{p_j}\delta_{p_j}}$.

HOI-aware training scheme. Here we introduce an effective optimization scheme to train our neural layer-wise HOI model. We first leverage a photometric loss:

$$\mathcal{L}_c = \sum_{\mathbf{r} \in \mathcal{R}} \|C_{\mathbf{r}} - \hat{C}_{\mathbf{r}}\|_2^2, \tag{4}$$

where $\mathbf{r}$ is a ray in the training ray set $\mathcal{R}$, $C_{\mathbf{r}}$ and $\hat{C}_{\mathbf{r}}$ are the corresponding rendered and observed colors.

Note that in HOI scenarios, the human layer inevitably intersects with other object layers. This fact introduces ambiguities and cannot be addressed by a simple photometric loss. Thus we design three additional regularizers based on the tracked object template to alleviate this issue. To enforce the object layers to be solid surface radiance fields that do not contribute to human appearance, we adopt an occupancy and a sparsity regularizer inside and outside the object template, i.e.,

$$\mathcal{L}_o = \sum_{p \in \mathcal{O}} (\Omega_{-}(p, \mathcal{M})\|e^{(-\sigma_p)}\|_2^2 + \Omega_{+}(p, \mathcal{M})\|\sigma_p\|_2^2), \tag{5}$$

where $\mathcal{M}$ is the template mesh, $\mathcal{O}$ is a set of points randomly sampled in the object's bounding box, Ω_+ and Ω_- indicate whether a point locates outside or inside the mesh.

Recall that we have encoded human-object contact prior by jointly tracking human-object at the skeletal-pose level

Figure 5. **The geometry, MoCap and neural modeling results from our dataset**. HODome includes various interaction sequences, such as “holding a vase”, “opening the cupboard”, “sitting on the sofa” and “moving the table”.

(Eq. 1). However, due to the misalignment of SMPL shape and actual human geometry, the human body and clothes may still overlap the object space, which goes against real-world physics. Therefore, we implicitly constrain the deformation net to predict contact-aware non-rigid deformation:

$$\mathcal{L}_h = \sum_{p \in \mathcal{H}} \Omega\_(p, \mathcal{M}) \|\sigma_p\|_2^2, \tag{6}$$

where $\mathcal{H}$ is the set of random samples in the human's bounding box.

We further explore a weakly-supervising scheme that utilizes object texture cues to help decouple the human and object entities. The key observation is that, with accurate object tracking, the object radiance field fused across frames can be rendered with higher confidence. That is, we train the entire scene first and render the objects only to capture views, and the pixels similar to the captured ones are labeled as object pixels. In this way, we obtain coarse object segmentation maps $\mathcal{S}$ that serve as pseudo supervision: we adopt label-wise integration [53] to render the layer labels s of rays:

$$\mathbf{s} = \sum_{i=1}^{M} T(p_i)[(1 - e^{-\sigma_{p_i}\delta_{p_i}})\mathbf{l}_{p_i}], \tag{7}$$

where $\mathbf{l}_{p_i}$ is one-hot label indicates which layer the point p_i belongs to. We apply a semantic loss for object rays $\mathcal{S}_o$:

$$\mathcal{L}_s = \sum_{\mathbf{r} \in \mathcal{S}_o} \|\mathbf{s_r} - \hat{\mathbf{s}}_\mathbf{r}\|_2^2, \tag{8}$$

where $\hat{\mathbf{s}}_\mathbf{r}$ is the pseudo semantic label of ray $\mathbf{r}$. Note that additional loss can be applied if accurate labels are provided.

Layer-wise neural representation in HODome. To further use the data, we can render reliable segmentation maps and occlusion-free appearances for humans and objects respectively. However, the renderings tend to be blurred since the information of all the frames is fused. Hence we introduce an enhancement scheme to construct separate high-quality per-frame neural representation for human/object: we blend the visible regions of the high-fidelity observed input views to rendered human/object images using the segmentation maps. The Instant-NGP [38] and Instant-NSR [79] techniques are then applied to reconstruct separate human/object in tens of seconds. Finally, we obtain layer-wise representation which enables real-time photo-realistic rendering and benefits the thorough analysis of HOI scenarios.

Implementations. To build our neural representations, we train our layered models using Adam optimizer with a learning rate that starts from $5e^{-4}$ and decays exponentially. Using the 76-view 4K resolution videos, the training process takes 8-10 hours on a single NVIDIA 3090 GPU for acceptable appearance and segmentation rendering. Thanks to Instant-NGP [38] and Instant-NSR [79], the layer-wise neural modeling process is conducted in seconds per frame.

5. Experiments

In this section, we compare our neural pipeline with existing state-of-the-arts. We first demonstrate the effectiveness and the generalization ability of our neural modeling approach (5.1). Next, we implement NeuralDome on three specific tasks (5.2).

5.1. Analysis of Neural Rendering

Comparisons. Here we compare our layer-wise neural human-object rendering approach, denoted as 'Ours-layer', with recent state-of-the-art methods, i.e., ST-NeRF [77], NeuralBody(NB) [45], in human-object interaction scenario. For fair comparisons, we select 20 cameras surrounding the performer for training and others for evaluation. We

Figure 6. **Qualitative comparisons** of novel view synthesis results. We show ground truth and synthesized images of novel view for NeuralBody [45], and ST-NeRF [77] and our layered human-object representation. Our approach achieves the best performance. We further illustrate our high-fidelity neural representations in HODome in the last column.

Methods	NB [45]		ST-NeRF [77]		Ours	
Scenes	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Bigsofa	19.33	0.896	24.02	0.886	**32.28**	**0.958**
Sofa	26.73	0.965	28.49	0.958	**35.62**	**0.987**
Table	21.94	0.933	22.44	0.900	**27.88**	**0.949**
Average	22.67	0.931	24.99	0.915	**31.93**	**0.964**

Table 2. **Quantitative comparisons** of novel view synthesized appearance on different human-object interaction sequences.

remove the background layer and train ST-NeRF without the semantic label as ours. Fig. 6 shows several appearance synthesis results. NeuralBody has no ability to model objects which leads to artifacts in the density field. ST-NeRF fails as the bounding boxes of the performer and object almost completely overlap. Our method achieves both human and object modeling and rendering and thus further enables generating high-fidelity neural representation denoted as 'Ours-HODome'. Tab. 2 further illustrates our method significantly outperforms the baselines.

Evaluation. Here we further evaluate how our scheme contributes to the generated layer-wise neural representations. Fig. 7 shows the single human layer where the performer is "sitting" in the air. Note that we have no ground truth of the specific human/object layer, thus we conduct qualitative evaluations only. Let **w/o pseudo segmentation** and **w/o blending** denote our human assets generated without pseudo semantic loss and blending-based enhancement scheme. It demonstrates that the pseudo semantic loss effectively helps decouple the human and object layer and our blending scheme further boosts the appearance fidelity of the blurred temporal-fused model to a photo-realistic level.

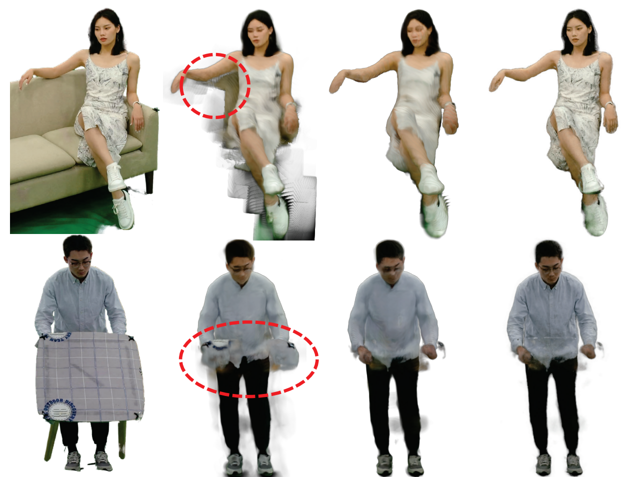

Figure 7. Qualitative evaluation of each strategy's contribution to layer-wise neural representation generation. The weakly-supervising scheme effectively helps decouple human and object and our blending scheme generates high-quality digital assets.

Method	Separated evaluation				Joint evaluation	
	MPJPE↓	PA-MPJPE↓	Chamfer$_h$ ↓	Chamfer$_o$ ↓	V2V↓	p.V2V↓
Fit to input	16.11	6.94	86.66	18.14	61.57	28.92
PHOSA [78]	14.88	6.94	77.62	16.64	54.59	26.09
CHORE [73]	10.21	6.14	86.58	7.69	44.93	14.93

Table 3. **Human object capture benchmark.** "Fit to input" represents the vanilla method that fits the object template to image and capture human with Frankmocap [47].

5.2. Task and benchmark

Human-object Capture Benchmark. HODome provides multi-view sequences with synchronized ground truth of capturing. To demonstrate the capability, we provide a benchmark for human-object capture, shown in Tab. 3. For detailed metric explanation please refer to [6, 73, 74]. PHOSA [78] used the contact map from hand-crafted annotations, suffering from coarse contact information. While CHORE [73] outperforms it with the distance fields predicted from neural networks. It shows the importance of a prepared human-object dataset that supports data-driven human-object capture model. Check out the full capturing benchmark of the dataset at https://arxiv.org/abs/2212.07626.

Human-object Geometry Reconstruction Benchmark. Benefiting from the dense-view setting, various human-only geometry reconstruction methods can step towards a human-object geometry reconstruction setting. We can benchmark the state-of-the-art algorithms in sparse view geometry reconstruction tasks. Hence, we evaluate PIFu [48] by training on a subset of HODome. As shown in Fig. 8, we compare the quality result between original PIFu and PIFu-trained on HODome and in-the-wild images. Compared to the original PIFu, our training set can improve the quality of

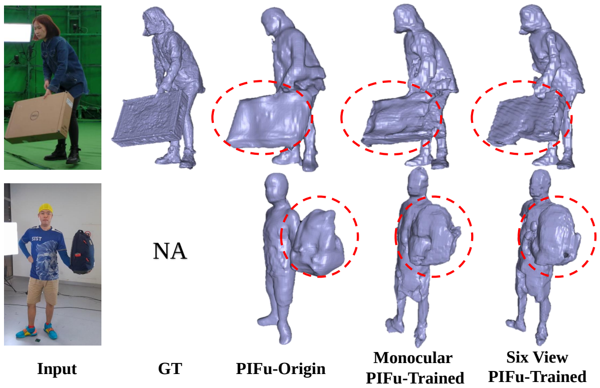

Figure 8. **Qualitative evaluation** on human-object reconstruction. We compare the original PIFu [48] with trained PIFu on both images from our dataset and in-the-wild images. The original PIFu fails to predict the object depth, and PIFu trained on our dataset can approximately predict the shape of the object.

Method	P2S $\times 10^{-4}\downarrow$	Chamfer $\times 10^{-4}\downarrow$
Origin PIFu [48]	38.726	40.947
Monocular PIFu-trained	14.653	14.483
6-View PIFu-trained	**3.376**	**4.901**

Table 4. **Geometry reconstruction benchmark**. Origin PIFu [48] uses the pre-trained model to infer. Monocular PIFu is trained on a single view. 6-View PIFu is trained on 6-view inputs.

Method	PSNR↑	SSIM↑
IBRnet [69]	21.43	0.892
Neuray [30]	23.34	0.909
NeuralHumanFVV [60]	21.69	0.914
NeuralHOIFVV	23.10	0.912

Table 5. **Neural rendering benchmark** in sparse-view setting.

the reconstructed shape of the objects. For further quantitative analysis, we evaluate the performance using **P2S** and **CD** in Tab. 4.

Sparse-View Human-object Rendering Benchmark. Our HODome supports various neural rendering tasks, even for human-object interactions. We provide a benchmark on sparse-view rendering tasks and evaluate on IBRNet [69], NeuRay [30] and NeuralHumanFVV [60]. Besides, we also provide a baseline, named NeuralHOIFVV, which uses the projection of trained PIFu's [48] results as depth input, then applies the Neural Blending method presented in [79] to obtain novel-view synthesis. More details about NeuralHOIFVV are referred to the Appendix. We analyze the above methods on HODome and in-the-wild images. The qualitative and quantitative results are shown in Fig. 9 and Tab. 5. Our proposed NeuralHOIFVV can obtain generative novel-view synthesis on HODome and in-the-wild inputs.

5.3. Limitations

Although NeuralDome can provide accurate capturing results and layer-wise neural representations, we also want

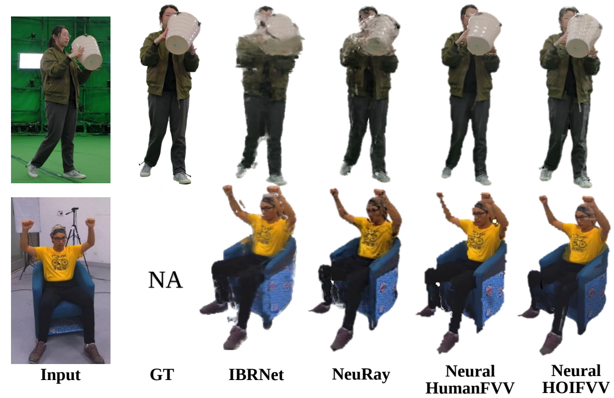

Figure 9. **Qualitative evaluation** on neural human-object rendering. We show the comparison of IBRNet [69], NeuRay [30], NeuralHumanFVV [60] and our proposed NeuralHOIFVV. We render one novel view between two input camera views.

to highlight some potential limitations of this pipeline. First, NeuralDome only considers single-person interaction with objects and the single-person task is already quite challenging for current research. It is non-trivial to extend the current algorithm to multi-person settings, especially in a crowd scene. Secondly, the reconstruction of 3D holistic scene is not covered in our pipeline. We leave the joint modeling of humans, objects and scene in the future work. Moreover, our dataset was collected under fixed illumination conditions with few backgrounds variance, limiting its generalization ability for other environments.

6. Conclusion

We have presented HODome, the first dataset to jointly capture and render human-object interaction using dense RGB cameras and the Optitrack system. To process the HODome dataset, we have further developed NeuralDome, a neural pipeline that can accurately track humans and objects, conduct layer-wise geometry reconstruction, and enable novel-view synthesis for both the subjects and objects. We believe the dataset could boost the development of capturing and rendering human-object interactions and is very valuable to the community for a wide range of human-object capturing and rendering tasks.

Acknowledgement This work was supported by the Shanghai Sailing Program (21YF1429400, 22YF1428800), Shanghai Local College Capacity Building Program (23010503100,22010502800), NSFC programs (61976138, 61977047), the National Key Research and Development Program (2018YFB2100500), STCSM (2015F0203-000-06), SHMEC (2019-01-07-00-01-E00003) and Shanghai Frontiers Science Center of Humancentered Artificial Intelligence (ShangHAI).

References

[1] Naturalpoint, inc. motion capture systems. https://optitrack.com/. 6. 3

[2] Reality capture. https://www.capturingreality.com/realitycapture. 3, 4

[3] Easymocap - make human motion capture easier. Github, 2021. 4

[4] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 696–712. Springer, 2020. 2

[5] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992. 4

[6] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 2, 3, 4, 7

[7] Hongrui Cai, Wanquan Feng, Xuetao Feng, Yan Wang, and Juyong Zhang. Neural surface reconstruction of dynamic scenes with monocular rgb-d camera. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 4

[9] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *arXiv preprint arXiv:2203.09517*, 2022. 2

[10] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2

[11] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34(4):69, 2015. 2, 3

[12] Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. Motion2fusion: Real-time volumetric performance capture. *ACM Trans. Graph.*, 36(6):246:1–246:16, Nov. 2017. 2, 3

[13] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 3

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

[15] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. A statistical model of human pose and body shape. In *Computer graphics forum*, volume 28, pages 337–346. Wiley Online Library, 2009. 3

[16] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 3, 5

[17] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11046–11056, 2021. 2

[18] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pages 281–299. Springer, 2022. 3

[19] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 2

[20] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 3

[21] Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, and Lan Xu. Neuralhofusion: Neural volumetric rendering under human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6155–6165, 2022. 2, 3

[22] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. 2017. 3

[23] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34:24741–24752, 2021. 2

[24] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[25] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. *arXiv preprint arXiv:2206.08929*, 2022. 2

[26] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-

scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019. 2, 3

[27] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 2

[28] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *arXiv preprint arXiv:2001.04947*, 2020. 2

[29] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)*, 38(5):1–14, 2019. 2

[30] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *CVPR*, 2022. 2, 8

[31] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 2

[32] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *arXiv preprint arXiv:2103.01954*, 2021. 2

[33] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. *ECCV*, 2022. 2

[34] H. Luo, A. Chen, Q. Zhang, B. Pang, M. Wu, L. Xu, and J. Yu. Convolutional neural opacity radiance fields. In *2021 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12, Los Alamitos, CA, USA, may 2021. IEEE Computer Society. 2

[35] Haimin Luo, Teng Xu, Yuheng Jiang, Chenglin Zhou, Qiwei Qiu, Yingliang Zhang, Wei Yang, Lan Xu, and Jingyi Yu. Artemis: Articulated neural pets with appearance and motion synthesis. *ACM Trans. Graph.*, 41(4), jul 2022. 2

[36] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6351–6361, 2021. 2

[37] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 2

[38] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 2, 4, 6

[39] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[40] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5762–5772, 2021. 2

[41] Anqi Pang, Xin Chen, Haimin Luo, Minye Wu, Jingyi Yu, and Lan Xu. Few-shot neural human performance rendering from sparse rgbd videos, 2021. 2

[42] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), dec 2021. 2

[43] 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. 2, 5

[44] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 2

[45] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2, 6, 7

[46] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021. 2

[47] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1749–1759, 2021. 7

[48] 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, 7, 8

[49] 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

[50] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 3

[51] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3

[52] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[53] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery. 3, 6

[54] Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, et al. Textured neural avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2387–2397, 2019. 2

[55] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1):4–27, 2010. 3

[56] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 246–264, Cham, 2020. Springer International Publishing. 3

[57] Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingya Wang, and Jingyi Yu. Neural free-viewpoint performance rendering under complex human-object interactions. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 2, 3

[58] Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingyi Yu, and Jingya Wang. Neural free-viewpoint performance rendering under complex human-object interactions. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4651–4660, 2021. 3

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

[60] Xin Suo, Yuheng Jiang, Pei Lin, Yingliang Zhang, Minye Wu, Kaiwen Guo, and Lan Xu. Neuralhumanfvv: Real-time neural volumetric human performance rendering using rgb cameras. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6222–6233, 2021. 8

[61] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020. 2, 3, 4

[62] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Niessner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhoefer, and Vladislav Golyanik. Advances in neural rendering, 2021. 2

[63] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhofer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 2

[64] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 2

[65] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 3

[66] Liao Wang, Ziyu Wang, Pei Lin, Yuheng Jiang, Xin Suo, Minye Wu, Lan Xu, and Jingyi Yu. ibutter: Neural interactive bullet time generator for human free-viewpoint rendering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4641–4650, 2021. 2

[67] Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Fourier plenoctrees for dynamic radiance field rendering in real-time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13524–13534, 2022. 2

[68] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 2

[69] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 2, 8

[70] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: animatable volume rendering of articulated human sdfs. *arXiv preprint arXiv:2210.10036*, 2022. 2

[71] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[72] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1682–1691, 2020. 2

[73] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. *arXiv preprint arXiv:2204.02445*, 2022. 3, 7

[74] Hongwei Yi, Chun-Hao P Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J Black. Human-aware object placement

for visual environment reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3959–3970, 2022. 3, 7

[75] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *arXiv*, 2021. 2

[76] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[77] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Trans. Graph.*, 40(4), July 2021. 2, 3, 5, 6, 7

[78] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European conference on computer vision*, pages 34–51. Springer, 2020. 3, 7

[79] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, et al. Human performance modeling and rendering via neural animated mesh. *arXiv preprint arXiv:2209.08468*, 2022. 2, 4, 6, 8

[80] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7753, 2022. 2, 5

[81] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 4