Instant-NVR: Instant Neural Volumetric Rendering for Human-object Interactions from Monocular RGBD Stream

Yuheng Jiang[1,2]* Kaixin Yao[1,2]* Zhuo Su[3] Zhehao Shen[1] Haimin Luo[1] Lan Xu[1]
[1]ShanghaiTech University [2]NeuDim [3]Pico IDL, ByteDance

Abstract

Convenient 4D modeling of human-object interactions is essential for numerous applications. However, monocular tracking and rendering of complex interaction scenarios remain challenging. In this paper, we propose Instant-NVR, a neural approach for instant volumetric human-object tracking and rendering using a single RGBD camera. It bridges traditional non-rigid tracking with recent instant radiance field techniques via a multi-thread tracking-rendering mechanism. In the tracking front-end, we adopt a robust human-object capture scheme to provide sufficient motion priors. We further introduce a separated instant neural representation with a novel hybrid deformation module for the interacting scene. We also provide an on-the-fly reconstruction scheme of the dynamic/static radiance fields via efficient motion-prior searching. Moreover, we introduce an online key frame selection scheme and a rendering-aware refinement strategy to significantly improve the appearance details for online novel-view synthesis. Extensive experiments demonstrate the effectiveness and efficiency of our approach for the instant generation of human-object radiance fields on the fly, notably achieving real-time photo-realistic novel view synthesis under complex human-object interactions. Project page: https://nowheretrix.github.io/Instant-NVR/.

1. Introduction

The accurate tracking and photo-realistic rendering for human-object interactions are critical for numerous human-centric applications like telepresence, tele-education or immersive experience in VR/AR. However, a convenient solution from monocular input, especially for on-the-fly setting, remains extremely challenging in the vision community.

Early high-end solutions [6, 9, 13, 18] require dense cameras for high-fidelity reconstruction. Recent approaches [11, 12, 17, 46, 47, 59, 63] need less RGB or RGBD video inputs (from 3 to 8 views) by using volumetric tracking techniques [19, 32]. Yet, the multi-view setting is still undesirable for consumer-level daily usage. Differently, the monocular method with a single handiest commercial RGBD camera is more practical and attractive. For monocular human-object modeling, most approaches [2, 15, 53, 57, 65, 66] track the rigid and skeletal motions of object and human using a pre-scanned template or parametric model. Besides, the monocular volumetric methods [32,41,43,58,64] obtain detailed geometry through depth fusion, while the recent advance [44] further extends it into the human-object setting. However, they fail to generate realistic appearance results, restricted by the limited geometry resolution.

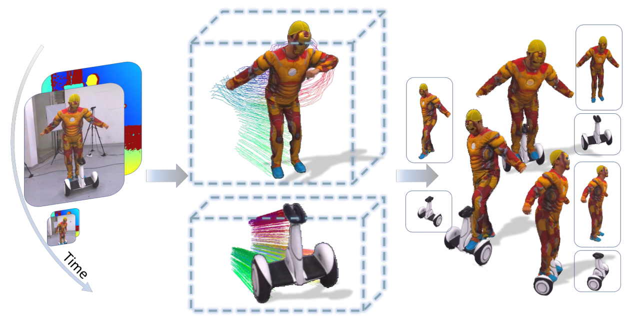

Figure 1. Our Instant-NVR adopts a separated instant neural representation to achieve photo-realistic rendering for human-object interacting scenarios.

Recent neural rendering advances, represented by Neural Radiance Fields (NeRF) [29], have recently enabled photo-realistic rendering with dense-view supervision. Notably, some recent dynamic variants of NeRF [21, 28, 50, 51, 55, 60, 67] obtain the compelling novel-view synthesis of human activities even under monocular capturing. However, they rely on tedious and time-consuming per-scene training to fuse the temporal observations into the canonical space, thus unsuitable for on-the-fly usage like telepresence. Only recently, Instant-NGP [30] enables fast radiance field generation in seconds, bringing the possibility for on-the-fly radiance field modeling. Yet, the original Instant-NGP can only handle static scenes. Few researchers explore the on-the-fly neural rendering strategies for human-object interactions, especially for monocular setting.

*Equal Contribution.

In this paper, we present *Instant-NVR* – an instant neural volumetric rendering system for human-object interacting scenes using a single RGBD camera. As shown in Fig. 1, Instant-NVR enables instant photo-realistic novel view synthesis via on-the-fly generation of the radiance fields for both the rigid object and dynamic human. Our key idea is to bridge the traditional volumetric non-rigid tracking with instant radiance field techniques. Analogous to the tracking-mapping design in SLAM, we adopt a multi-thread and tracking-rendering mechanism. The tracking front-end provides online motion estimations of both the performer and object, while the rendering back-end reconstructs the radiance fields of the interaction scene to provide instant novel view synthesis with photo-realism.

For the tracking front-end, we first utilize off-the-shelf instant segmentation to distinguish the human and object from the input RGBD stream. Then, we adopt an efficient non-rigid tracking scheme for both the performer and rigid object, where we adopt both embedded deformation [45] and SMPL [27] to model human motions. For the rendering back-end, inspired by Instant-NGP [30] we adopt a separate instant neural representation. Specifically, both the dynamic performer and static object are represented as implicit radiance fields with multi-scale feature hashing in the canonical space and share volumetric rendering for novel view synthesis. For the dynamic human, we further introduce a hybrid deformation module to efficiently utilize the non-rigid motion priors. Then, we modify the training process of radiance fields into a key-frame based setting, so as to enable graduate and on-the-fly optimization of the radiance fields within the rendering thread. For the dynamic one, we further propose to accelerate our hybrid deform module with a hierarchical and GPU-friendly strategy for motion-prior searching. Yet we observe that naively selecting key-frames with fixed time intervals will cause non-evenly distribution of the captured regions of the dynamic scene. It results in unbalanced radiance field optimization and severe appearance artifacts during free-view rendering. To that end, we propose an online key-frame selection scheme with a rendering-aware refinement strategy. It jointly considers the visibility and motion distribution across the selected key-frames, achieving real-time and photo-realistic novel-view synthesis for human-object interactions.

To summarize, our main contributions include:

- We present the first instant neural rendering system under human-object interactions from an RGBD sensor.
- We introduce an on-the-fly reconstruction scheme for dynamic/static radiance fields using the motion priors through a tracking-rendering mechanism.
- We introduce an online key frame selection scheme and a rendering-aware refinement strategy to significantly improve the online novel-view synthesis.

2. Related Work

Traditional Human Volumetric Capture. Human volumetric capture and reconstruction have been widely investigated to achieve detailed geometry reconstruction and accurate tracking. A series of works are proposed to make volumetric fusion more robust with SIFT features [16], multi-view systems [11, 12], scene flow [54], human articulated skeleton prior [62, 64], extra IMU sensors [70], data-driven prior [43, 44], learned correspondences [5], neural deformation graph [4, 23] or implicit function [17, 63]. Starting from the pioneering work DynamicFusion [32] which benefits from the GPU solvers, the high-end solutions [11, 12] rely on the multi-view camera system and complex calibration. VolumeDeform [16] combines depth-based correspondences with sparse SIFT features to reduce drift. KillingFusion [41] and SobolevFusion [42] support topology changes via more constraints on the motion fields. Thanks to the human parametric model [27], DoubleFusion [64] proposes the two-layer representation to capture scene more robustly. UnstructuredFusion [59] extends it to an unstructured multi-view setup. RobustFusion [44] further handles the challenging human-object interaction scenarios. Besides, Function4d [63] and NeuralHOFusion [17] marry the non-rigid-tracking with implicit modeling. However, these methods are dedicated to getting detailed geometry without focusing on high-quality texture and most methods can not handle human-object interactions. Comparably, our approach bridges the traditional volumetric capture and neural rendering advances, achieving photo-realistic rendering results under human-object interactions.

Static Neural Scene Representations. Coordinates-based neural scene representations in static scenes produce impressive novel view synthesis results and show huge potential. Various data representations are adopted to obtain better performance and characteristics, such as point-clouds [1, 48, 56], voxels [26], textured meshes [25, 49], occupancy [33, 40] or SDF [34, 52]. Meanwhile, Since the vanilla NeRF which requires hours of training is time-consuming, some NeRF extensions [30, 39, 61] are proposed to accelerate both training and rendering. Plenoctrees [61] utilizes the octree to skip the empty regions. Plenoxels [39] parameterizes the encoding using spherical harmonics on the explicit 3D volume. Instant-NGP [30] utilizes the multi-scale feature hashing and TCNN to speed up. Though its rendering speed seems possible to train on-the-fly, they do not have a specific design for streaming input and only can recover static scenes. Comparably, our Instant-NVR achieves on-the-fly efficiency based on the Instant-NGP [30].

Dynamic Neural Scene Representations. Novel view synthesis in dynamic scenes is an important research problem. D-NeRF [38] and Non-rigid NeRF [50] leverage the displacement field to represent the motion while Ner-

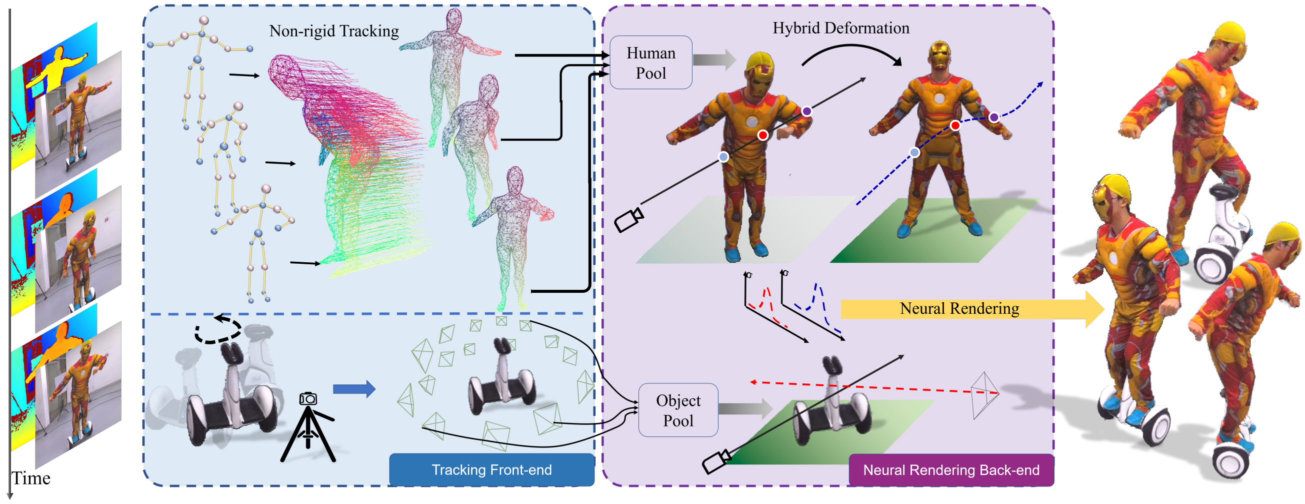

Figure 2. Our approach consists of two stages. The tracking front-end (Sec. 4.1) captures human and object motions, while the rendering back-end (Sec. 4.2) separately reconstructs the human-object radiance fields on-the-fly, for instant novel view synthesis with photo-realism.

fies [35] and HyperNeRF [36] use the SE(3) field. Moreover, some researchers focus on human reconstruction and utilize the human prior. Neuralbody anchors latent code on the SMPL [27] vertices. Humannerf [68] combines the SMPL warping and deformation net to construct the motion field. TAVA [20] learns the skinning weight for joints via root-finding and can generalize to novel pose. DeVRF [24] incorporates 4D-motion volume into the NeRF pipeline. NDR [7] defines a bijective function which naturally compatible with the cycle consistency. However, most methods rely on multi-view camera input and the training is costly. Comparably, our Instant-NVR bridges the non-rigid volumetric capture with the instant radiance field training, achieving photo-realistic rendering results from monocular RGBD stream.

3. Overview

From monocular RGBD input, Instant-NVR bridges the real-time non-rigid capture with instant neural rendering, allowing for high-quality novel-view synthesis under human-object interactions. As illustrated in Fig. 2, our system consists of two cooperating threads: a tracking front-end (Sec. 4.1) and a neural rendering back-end (Sec. 4.2).

Tracking Front-end. We extend the traditional volumetric tracking [32, 59, 64] into a human-object setting. For non-rigid human capture, we adopt the embedded deformation(ED) [45] and SMPL [27] as motion representations. For object, we directly track its rigid motions via the Iterative Closest Point(ICP) algorithm. This thread provides accurate per-frame human-object motion priors, enabling the integration of all the radiance information into the global canonical space. Note that the reconstructed volumetric geometry suffers from discrete and low-resolution artifacts. Thus, we only transmit the motion priors with the RGBD images to the rendering thread and discard the explicit volumetric geometry prior.

Neural Rendering Back-end. We extend instant radiance fields [30] to the monocular and dynamic human-centric scenes, where we maintain the canonical instant radiance fields for both dynamic human and rigid objects separately. We introduce a lightweight pose-conditioned deformation module to learn the residual motion to refine the initial warping provided by motion priors. To enable on-the-fly radiance field generation and rendering, we adapt the training process into a key frame setting with the aid of efficient motion-prior caching. We introduce a key frame selection method to jointly consider the diversity of capturing view and human pose, visibility maps, and the input image quality. We further adaptively refine the appearance output in the rendering view with more analogous spatial-temporal capturing views. Note that our rendering thread reconstructs the radiance fields online to provide instant novel view synthesis with photo realism.

4. Method

4.1. Tracking Front-end

Human Non-rigid Tracking. We follow the traditional volumetric capture methods [44, 64] to track human non-rigid motions. Specifically, we parameterize human non-rigid motions as an embedded deformation graph $W = \{dq_i, x_i\}$, where x_i is the coordinates of the sampled ED node in canonical space and dq_i is the dual quaternions representing the corresponding rigid transformation in $SE(3)$ space. Each 3D point v_c in the canonical space can be wrapped into the live space using an efficient and accurate motion interpolation method Dual-Quaternion Blend-

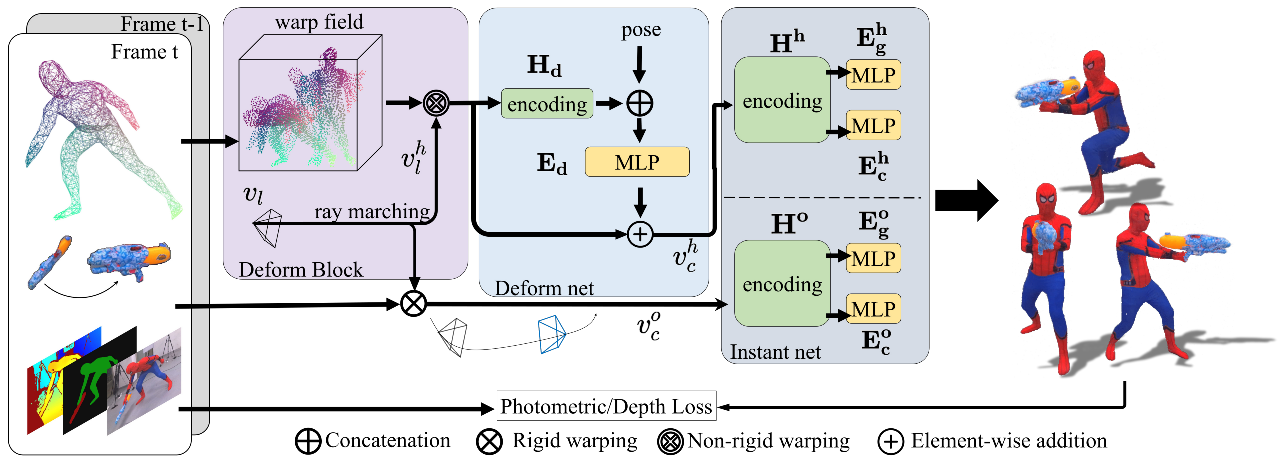

Figure 3. Our Neural Rendering back-end adopts a separated neural representation. Left are the input RGBD images with motions. Middle is a separate rendering engine that includes a hybrid deformation module and volumetric rendering. Right are rendering results.

ing (DQB):

$$DQB(v_c) = \sum_{i \in \mathcal{N}(v_c)} w(x_i, v_c) dq_i, \qquad (1)$$
$$\tilde{v}_c = SE3(DQB(v_c))v_c.$$

where $\mathcal{N}(v_c)$ is a set of neighboring ED nodes of v_c, $w(x_i, v_c)$ is the influence weight of the $i-th$ node x_i to v_c and formulated as $w(x_i, v_c) = exp(-\left\|\mathbf{v}_c - \mathbf{x}_i\right\|_2^2 / r^2)$. r is the influence radius (0.1 in our setting). Note that the ED-only-based human tracking is fragile since the non-rigid ICP often fails at fast articulated human motions due to losing correspondence. Therefore, we also introduce the SMPL inner body with shape parameters β and pose parameters θ as the skeleton prior and utilize θ with the skinning weight to wrap 3D point v_c, which further constrain the ED motion tracking within a reasonable motion scale. Please refer to [64] for details about the ED-sampling and double layer motion representation.

To calculate the final ED-based motion, we jointly optimize the skeleton pose θ and ED non-rigid motion field W as follows:

$$\begin{aligned} E(W, \theta) = & \lambda_{\text{data}} E_{\text{data}} + \lambda_{\text{bind}} E_{\text{bind}} + \lambda_{\text{reg}} E_{\text{reg}} + \\ & \lambda_{\text{prior}} E_{\text{prior}} + \lambda_{\text{pose}} E_{\text{pose}} + \lambda_{\text{inter}} E_{\text{inter}}. \end{aligned} \qquad (2)$$

The data term E_{data} measures the point-to-plane distances between the deformed model and the current input depth map:

$$\boldsymbol{E}_{\text{data}} = \sum_{(\mathbf{v}_c, \mathbf{u}) \in \mathcal{P}} \psi(\mathbf{n}_{\mathbf{u}}^T (\tilde{\mathbf{v}}_c - \mathbf{u})), \qquad (3)$$

where $\mathbf{u}$ is a sampled point in the depth map, $\mathbf{n_u}$ is its normal, and $\mathbf{v}_c$ denotes its closest point on the fused surface. $\mathcal{P}_i$ is the set of correspondences found via a projective local search [32]. Besides, the binding term E_{bind} constrains both skeleton and final ED motions to be consistent

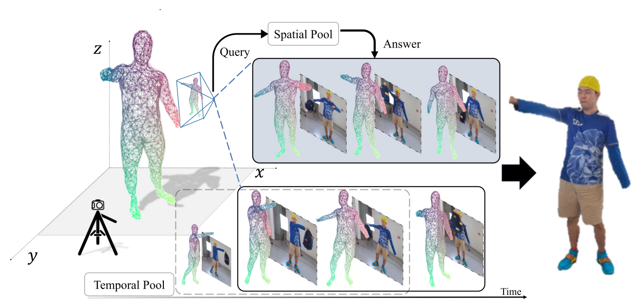

Figure 4. Illustration of our online refinement strategy.

while the geometry regularity term E_{reg} produces locally as-rigid-as-possible (ARAP) motions to prevent overfitting to depth inputs. These two terms are detailed in [14, 64]. The pose prior term E_{prior} from [3] penalizes the unnatural poses. Both the pose term E_{pose} and interaction term E_{inter} are form [44] to encourage natural motion capture during human-object interactions. Note that the optimization nonlinear least squares problem in Eqn. 2 is solved using LM method with the PCG solver on GPU [12, 14].

Object Rigid Tracking. For rigid tracking of objects, we follow [44] to optimize the rigid motions and transform them to camera pose T^t under the ICP framework, in which we fuse the depth map to a canonical TSDF volume to maintain the stable correspondence for robust object tracking.

4.2. Neural Rendering Back-end

To enable efficient photo-realistic neural rendering of the interaction scenes, our neural rendering back-end adopts a separated instant neural representation based on the on-the-fly key frame selection strategy.

Separated instant neural representation. We design the instant neural representation to reconstruct the human and object separately. For object branch, given the RGBD image I_t and D_t with the camera pose T^t as the training set, we leverage the original Instant-NGP [30] to extract the 3D point v_c^o features on the hash table $\mathbf{H^o}$ and then feed them into the geometry MLP $\mathbf{E^o_g}$ and color MLP $\mathbf{E^o_c}$ to acquire the density and color.

For dynamic human, in contrast to recent approaches [35, 36, 38, 50] which can't handle long sequences via pure MLP and human NeRFs [37, 55, 69] that heavily rely on SMPL [27] which do not align well with the surface and easily cause artifacts, we introduce a hybrid deformation module to efficiently leverage the motion priors. The explicit non-rigid warping and an implicit pose-conditioned deformation net jointly aggregate the corresponding point information in the canonical space.

Specifically, given this human non-rigid motion $\{dq_i^t\}$, SMPL pose $\vec{\theta^t}$ and a sampling point v_l^h at frame t, we construct the warping function to map v_l^h back to the canonical space v_t^h. We calculate the deformed ED nodes $x_i^t = dq_i^t x_i$, and then the point v_l^h in the influence radius r of these nodes can be warped into canonical surface via neighboring ED nodes weight blending:

$$v_t^h = SE3(DQB^{-1}(v_l^h))v_l^h. \tag{4}$$

To reduce the warping error and improve the rendering quality, we further integrate pose-conditioned deformation net here to correct the misalignment, where we concatenate the encoded v_t^h via hash-encoding with the human pose $\vec{\theta_t}$ and predict the residual displacement δv^h through an MLP. Finally, we feed $v_c^h = v_t^h + \delta v^h$ into the canonical hash-encoding $\mathbf{H^h}$, geometry as well as color MLPs $\mathbf{E^h_g}, \mathbf{E^h_c}$.

On-the-fly Radiance Fields. To ensure accurate tracking, hundreds of ED-nodes are maintained to query live points neighbors which is time-consuming and lead to bottlenecks. The time consumption is $O(n)$ even if we query a small number of neighbors for each sampled point, in which n is the number of ED nodes. To enhance on-the-fly efficiency, we introduce a look-up-table-based fast search strategy here to speed up. Specifically, we only initialize the canonical KNN(k-nearest-neighbors) field in the beginning, whose resolution is 512^3, and each voxel saves s neighboring ED nodes index(4 in our setting). We then concatenate non-rigid motions in each frame to form a look-up table. At frame t, for a voxel with index k and coordinates v_k in the canonical, we warp it via Eqn. 1 to the live space and obtain its corresponding voxel index f. We save the canonical index k in live voxel f. Afterward, for each sampling point, we can acquire the live space index f and obtain the canonical index k. k links the 4 neighbor ED nodes index. Offsetting the index to frame t on the look-up-table, we can acquire the corresponding motions and calculate the blending weight as well as warped point via DQB in $O(1)$ manner. In addition, we are able to construct the live KNN field for each voxel in $O(1)$ time by utilizing custom CUDA kernels.

Online Key Frame Selection. To achieve online performance and high quality rendering, we choose key frames to organize our neural rendering training dataset. Before choosing, we discard blurry RGB frames caused by fast motion based on the blurriness measure [10]. Besides, we observe that naively selecting key-frames with fixed time intervals brings the time-related details but causes the non-evenly distribution of the captured regions. Inspired by [22, 31], we introduce a key frame selection scheme here to keep the diversity of motion distribution and complement visibility. Specifically, we formulate the visibility map for each ED node $x_i^t = (x', y', z')$ in frame t as follows:

$$s_i^t = \begin{cases} 1, & if \quad |z' - D^t(\pi(x_i^t))| < \epsilon \\ 0, & othersize \end{cases}, \tag{5}$$

where $\pi(\cdot)$ denotes the projection matrix, $D^t(\cdot)$ represents the depth value of the corresponding pixel at frame t, ϵ is the visibility degree (0.01 in our setting). we continue to define the similarity for two frames:

$$\begin{aligned} E_h(t_1, t_2) = &\beta_{\text{pose}}|\vec{\theta_{t_1}} - \vec{\theta_{t_2}}|^2 + \beta_{\text{vis}} \sum_i s_i^{t_1} \oplus s_i^{t_2} + \\ &\beta_{\text{h}}|t_1 - t_2|^2, \end{aligned} \tag{6}$$

where $\oplus$ is the xor operation and t_1, t_2 are frame indexes.

For an object with the pose T^t which includes rotation $\mathbf{R}^t$ and translation $\vec{d^t}$, we define the similarity as follows:

$$E_o(t_1, t_2) = \beta_{\text{d}}\|\vec{d^{t_1}} - \vec{d^{t_2}}\|_2^2 + \beta_{\text{o}}|t_1 - t_2|^2, \tag{7}$$

Furthermore, we define γ to determine the diversity of the spatial pool. At the start time, the pool is empty and imported the first frame. Once the similarity of each two among the latest frame received from tracking and frame(s) in the pool is greater than γ, we push this frame to the pool. When the pool capacity reaches its peak (100 in our setting), we will continually update the pool using the new frame by removing the frame with the biggest similarity. In this manner, our spatial pool constantly updates in all frames.

To achieve photo-realistic novel-view synthesis, our Instant-NVR further refines the rendering view via training short iterations on the carefully selected frames in a spatial-temporal pool. The spatial-temporal pool includes m frames from the spatial pool, which have the most similarity with the rendering view and m latest frames received from the front-end. Our selection strategy ensures high-quality rendering without losing temporal detail.

4.3. Implementation Details

To train the dynamic NeRF under human-object interactions, we first apply the semantic segmentation MIVOS [8] to decouple the scene and obtain the human and object masks separately. To assemble human and object in a novel

Figure 5. The rendering results of Instant-NVR on various interaction sequences, including “driving a balance car”,“shaking a bag”, and “playing a water gun”.

view, we additionally render the depth maps and then combine the RGB images according to the depth occlusion. We implement our entire pipeline on GPU based on the Instant-NGP [30] using two Nvidia GeForce RTX3090 GPU. One GPU for the tracking front-end(14 GB memory consumption) and another for the neural rendering(15 GB for human branch and 4 GB for object). For deformation net, the input is the 32-dim hash feature and the 72-dim pose. The hidden layer is 4 and hidden dimension is 128. For key frame selection, we use the bigger weight for the joint rotation of the torso to domain the human pose diversity, specifically, for $\vec{\beta}_{\text{pose}}$, we set each torso weight $\beta_{pose(torso)} = 0.1$ and each limbs weight as 0.02. Besides, we use the following empirically determined parameters: $\beta_{\text{vis}} = 0.01, \beta_{\text{h}} = 0.02, \beta_{\text{d}} = 1.0, \beta_{\text{o}} = 0.02, \gamma = 2.5, \lambda_{\text{data}} = 1.0, \lambda_{\text{bind}} = 1.0, \lambda_{\text{reg}} = 4.0, \lambda_{\text{prior}} = 0.01, \lambda_{\text{pose}} = 0.02, \lambda_{\text{inter}} = 1.0$. For efficiency, we choose $m = 10$ in the spatial-temporal pool. We use the photometric loss and depth loss to supervise human NeRF and object NeRF separately as follows:

$$
\begin{aligned}
\mathcal{L}_{\text{color}} &= \sum_{\mathbf{r}\in\mathcal{R}} \|M(\mathbf{r})(\hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r}))\|_2, \\
\mathcal{L}_{\text{depth}} &= \sum_{\mathbf{r}\in\mathcal{R}} \|M(\mathbf{r})(\hat{\mathbf{D}}(\mathbf{r}) - \mathbf{D}(\mathbf{r}))\|_1
\end{aligned}
\tag{8}
$$

where $M(r)$ is human or object mask.

5. Experimental Results

In this section, we compare the state-of-the-art methods and evaluate Instant-NVR on various challenging human-object interaction scenarios. Besides, various rendering results of Instant-NVR are shown in Fig. 5, such as driving a balance car, shaking a bag and playing with a water gun. Please also kindly refer to our video.

5.1. Comparison

We compare Instant-NVR against the fusion-based methods RobustFusion [44], NeuralHOFusion [17] and NeRF-based methods NeuralBody [37], HumanNerf [68], both in efficiency and rendering quality. For comparison with fusion-based methods, as illustrated in Fig. 6 (b), RobustFusion [44] generates blurry appearance results, which are restricted by the limited geometry resolution. For a fair comparison, NeuralHOFusion [17] is modified to a single view setting and suffers from artifacts as shown in Fig. 6 (c). For NeRF-based methods, we employ the RGBD input to estimate the SMPL [27] as their prior and adopt RGBD loss terms. Both NeuralBody [37] and HumanNerf [68] give erroneous and blurry rendering results in the monocular setting (Fig. 6 (d-e)), which rely heavily on SMPL [27] and can not handle human-object interactions. In addition, training in these methods is time-consuming and novel

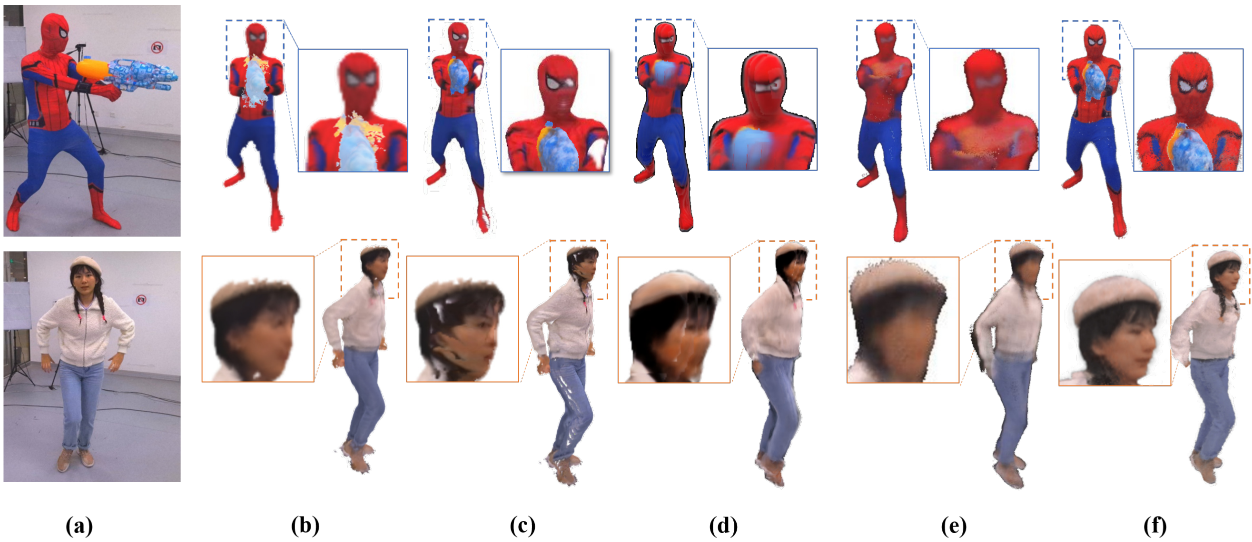

Figure 6. Qualitative comparison with fusion-based methods and NeRF-based methods. (a) Reference view. (b) RobustFusion [44] (c) NeuralHOFusion [17] (d) NeuralBody [37] (e) HumanNerf [68] (f) Ours.

view synthesis remains slow. In contrast, our Instant-NVR achieves more detailed and photo-realistic rendering results under human-object interactions, as shown in Fig. 6 (f). The quantitative results in Tab. 1 also demonstrate that our approach can achieve consistently better rendering quality and achieve efficient training as well as rendering speed to support on-the-fly performance. Note that both NeuralBody [37] and HumanNerf [68] take several hours to train, while training for our method is online.

Table 1. Comparison against fusion and NeRF-based methods.

Method	PSNR↑	SSIM ↑	Rendering Time↓
RobustFusion [44]	20.59	0.935	0.123s
NeuralHOFusion [17]	21.09	0.942	0.151s
NeuralBody [37]	19.71	0.928	2.420s
HumanNerf [68]	18.68	0.892	5.103s
Ours	**27.81**	**0.976**	**0.023s**

5.2. Evaluation

Online Human rendering. As shown in Fig. 7 (b), per-vertex texture extracted from the fused albedo volume [14] is blurry. Naively selecting key-frames with fixed time intervals generates noising rendering results in Fig. 7 (c) due to the non-evenly distribution of the captured regions. In contrast, our online key frame selection strategy based on the diversity of motion distribution and complement visibility can achieve much clearer rendering results, as shown in Fig. 7 (d). To boost the rendering quality, the further refinement scheme can help us to achieve more photo-realistic rendering results, as shown in Fig. 7 (e). As for quantitative analysis, we evaluate the rendering quality in Tab. 2, which highlights the contributions of each component.

Online Object Rendering. As for the evaluation of online object rendering in Fig. 8, we can observe that per-vertex texture failed to generate high-quality appearance which is restricted by the limited geometry resolution. Moreover, naively selecting key-frames with fixed time interval brings the noises. Fig. 8 (d) shows that applying our key frame selection strategy without refinement is still unclear. In contrast, we can achieve the best rendering results with our full training pipeline. Moreover, the quantitative evaluation is as demonstrated in Tab. 3, in which our full pipeline with the online key frame selection and rendering refinement achieves the highest accuracy.

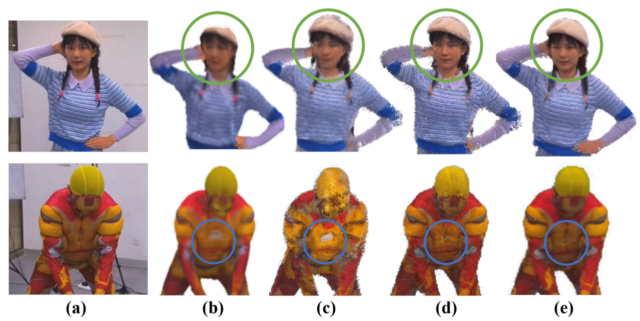

Figure 7. Quantitative evaluation of Online Human Rendering. (a) Input image; (b) Per-vertex texture; (c) Key frame selection using fixed interval; (c) Key frame selection w/o refinement; (d) Key frame selection w refinement.

Table 2. Quantitative evaluation of Online Human Rendering.

Method	PSNR ↑	SSIM ↑	MAE ↓
Per-vertex texture	19.710	0.902	3.176
Key frame selection using fixed interval	23.071	0.922	1.571
Key frame selection w/o refinement	25.768	0.949	1.229
Key frame selection w refinement	**28.255**	**0.972**	**0.534**

Run-time Evaluation. In Tab. 4, we list the run-time of each step in our pipeline, including both the tracking front-end and the neural rendering back-end. For tracking front-end, the rigid tracking of the object takes 40ms while the hu-

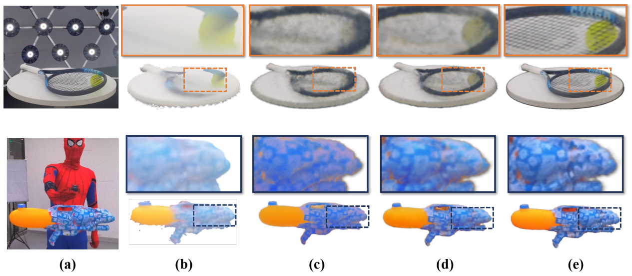

Figure 8. Quantitative evaluation of Online Object Rendering. (a) Input image; (b) Per-vertex texture; (c) Key frame selection using fixed interval; (c) Key frame selection w/o refinement ; (d) Key frame selection w refinement.

Table 3. Quantitative evaluation of Online Object Rendering.

Method	PSNR ↑	SSIM ↑	MAE ↓
Per-vertex texture	21.253	0.944	4.431
Key frame selection using fixed interval	25.248	0.954	1.043
Key frame selection w/o refinement	26.747	0.965	0.931
Key frame selection w refinement	**28.826**	**0.977**	**0.615**

Table 4. Quantitative evaluation of Run-time

Procedure	Time
rigid tracking	40ms
non-rigid-tracking	62ms
deformation net	5ms
training w/o fast search	205.53ms
training w fast search	17.95ms
rendering	23.38ms

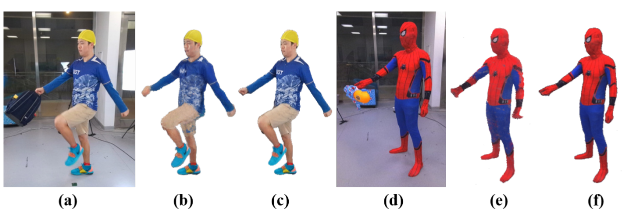

Figure 9. Evaluation of the hybrid deformation module. (a), (d) are the reference views. (b), (e) are the results with only deform block. (c),(f) use our deform block with the aid of deform net.

man non-rigid tracking takes 62ms. Besides, the deformation net costs 5ms. For rendering back-end, training without fast search strategy takes 205.53ms while using our fast search scheme, the training time reduces to 17.95ms. Besides, we use 15.38ms for the rendering process.

Hybrid Deformation Module Evaluation. We conduct further evaluation of our hybrid deformation module to demonstrate its advantages. As shown in Fig. 9 (b)(e), employing only the explicit deform block results in misalignment between the ground truth and warped space, leading to blurry images and erroneous silhouettes. Conversely, by utilizing the deform block with the aid of implicit deform net to learn the residual displacement in Fig. 9 (c)(f), the rendering results outcome exhibit superior alignment and significantly enhance texture.

5.3. Limitation

As the first instant neural rendering system from an RGBD sensor that performs real-time and photo-realistic novel-view synthesis under human-object interactions, the proposed Instant-NVR still has some limitations. First, although we adopt the hybrid deformation module to efficiently utilize the non-rigid motion priors since our method is in monocular RGB-D camera setting, non-rigid fusion fails when facing the fast movement and leads to inaccurate priors which affect the on-the-fly rendering. Due to limited resolution and inherent noise of the depth input, our method cannot reconstruct the extremely fine details of the performer, such as the fingers. Data-driven techniques on different human parts will be critical for such problem. Besides, Instant-NVR is committed to rendering photo-realistic results on-the-fly. Therefore, we choose the density field as geometry representation, analogous to Instant-NGP [30]. It is promising to integrate other SDF representations [7, 52], which can generate a more delicate geometry. Furthermore, to ensure efficient transmission between tracking front-end and rendering back-end, we discard the volumetric explicit geometry priors produced by the tracking step. It is an interesting direction to explore more complementary between tracking and rendering.

6. Conclusion

We have presented a practical neural tracking and rendering approach for human-object interaction scenes using a single RGBD camera. By bridging traditional non-rigid tracking with recent instant radiance field techniques, our system achieves a photo-realistic free-viewing experience for human-object scenes on the fly. Our non-rigid tracking robustly provides sufficient motion priors for both the performer and the object. Our separated instant neural representation with hybrid deformation and efficient motion-prior searching enables the on-the-fly reconstruction of both the dynamic and static radiance fields. Our online key frame selection with a rendering-aware refinement strategy further provides a more vivid and detailed novel-view synthesis for our online setting. Our experimental results demonstrate the effectiveness of Instant-NVR for the instant generation of dynamic radiance fields and photo-realistic novel view synthesis of human-object interactions in real time. We believe that our approach is a critical step to virtual but realistic teleport human-object interactions, with many potential applications like consumer-level telepresence in VR/AR.

Acknowledgements. This work was supported by Shanghai YangFan Program (21YF1429500), Shanghai Local college capacity building program (22010502800). We also acknowledge support from Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI).

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020. 2

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022. 1

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 561–578, Cham, 2016. Springer International Publishing. 4

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

[5] Aljaz Bozic, Michael Zollhofer, Christian Theobalt, and Matthias Nießner. Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7002–7012, 2020. 2

[6] Derek Bradley, Tiberiu Popa, Alla Sheffer, Wolfgang Heidrich, and Tamy Boubekeur. Markerless garment capture. In *ACM SIGGRAPH 2008 papers*, pages 1–9. 2008. 1

[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. 3, 8

[8] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5559–5568, 2021. 5

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

[10] Frederique Crete, Thierry Dolmiere, Patricia Ladret, and Marina Nicolas. The blur effect: perception and estimation with a new no-reference perceptual blur metric. In *Human vision and electronic imaging XII*, volume 6492, pages 196–206. SPIE, 2007. 5

[11] 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. 1, 2

[12] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016. 1, 2, 4

[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. 1

[14] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo and motion reconstruction using a single rgbd camera. *ACM Transactions on Graphics (TOG)*, 2017. 4, 7

[15] Sanjay Haresh, Xiaohao Sun, Hanxiao Jiang, Angel X Chang, and Manolis Savva. Articulated 3d human-object interactions from rgb videos: An empirical analysis of approaches and challenges. In *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022. 1

[16] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European Conference on Computer Vision*, pages 362–379. Springer, 2016. 2

[17] 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. 1, 2, 6, 7

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

[19] Hao Li, Bart Adams, Leonidas J Guibas, and Mark Pauly. Robust single-view geometry and motion reconstruction. *ACM Transactions on Graphics (ToG)*, 28(5):1–10, 2009. 1

[20] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. 2022. 3

[21] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, and Zhaoyang Lv. Neural 3d video synthesis, 2021. 1

[22] Zhe Li, Tao Yu, Zerong Zheng, Kaiwen Guo, and Yebin Liu. Posefusion: Pose-guided selective fusion for single-view human volumetric capture. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2021. 5

[23] Wenbin Lin, Chengwei Zheng, Jun-Hai Yong, and Feng Xu. Occlusionfusion: Occlusion-aware motion estimation for real-time dynamic 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1736–1745, 2022. 2

[24] Jia-Wei Liu, Yan-Pei Cao, Weijia Mao, Wenqiao Zhang, David Junhao Zhang, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Devrf: Fast deformable voxel radiance fields for dynamic scenes. *arXiv preprint arXiv:2205.15723*, 2022. 3

[25] 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

[26] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, July 2019. 2

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, Oct. 2015. 2, 3, 5, 6

[28] 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. 1

[29] 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 Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 405–421, Cham, 2020. Springer International Publishing. 1

[30] 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. 1, 2, 3, 5, 6, 8

[31] Armin Mustafa, Hansung Kim, and Adrian Hilton. 4d match trees for non-rigid surface alignment. In *European Conference on Computer Vision*, pages 213–229. Springer, 2016. 5

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

[33] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 2

[34] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[35] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 3, 5

[36] 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. 3, 5

[37] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 5, 6, 7

[38] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2, 5

[39] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 2

[40] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yanpei Cao, Tao Yu, and Yebin Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *CVPR*, 2022. 2

[41] Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1395, 2017. 1, 2

[42] Miroslava Slavcheva, Maximilian Baust, and Slobodan Ilic. Sobolevfusion: 3d reconstruction of scenes undergoing free non-rigid motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2646–2655, 2018. 2

[43] 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. 1, 2

[44] Zhuo Su, Lan Xu, Dawei Zhong, Zhong Li, Fan Deng, Shuxue Quan, and Lu Fang. Robustfusion: Robust volumetric performance reconstruction under human-object interactions from monocular rgbd stream. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2, 3, 4, 6, 7

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

[46] 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. 1

[47] 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. 1

[48] Xin Suo, Minye Wu, Yanshun Zhang, Yingliang Zhang, Lan Xu, Qiang Hu, and Jingyi Yu. Neural3d: Light-weight neural portrait scanning via context-aware correspondence learning.

In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3651–3660, 2020. 2

[49] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2

[50] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2021. 1, 2, 5

[51] 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. 1

[52] 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, 8

[53] Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In *International Conference on 3D Vision (3DV)*, 2022. 1

[54] Zirui Wang, Shuda Li, Henry Howard-Jenkins, Victor Prisacariu, and Min Chen. Flownet3d++: Geometric losses for deep scene flow estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2

[55] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, June 2022. 1, 5

[56] 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

[57] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*. Springer, October 2022. 1

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

[59] Lan Xu, Zhuo Su, Lei Han, Tao Yu, Yebin Liu, and Lu Fang. Unstructuredfusion: realtime 4d geometry and texture reconstruction using commercial rgbd cameras. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2508–2522, 2019. 1, 2, 3

[60] Jae Shin Yoon, Duygu Ceylan, Tuanfeng Y Wang, Jingwan Lu, Jimei Yang, Zhixin Shu, and Hyun Soo Park. Learning motion-dependent appearance for high-fidelity rendering of dynamic humans from a single camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3407–3417, 2022. 1

[61] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 2

[62] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *The IEEE International Conference on Computer Vision (ICCV)*. ACM, October 2017. 2

[63] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5746–5756, 2021. 1, 2

[64] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019. 1, 2, 3, 4

[65] 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 (ECCV)*, 2020. 1

[66] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7376–7385, 2020. 1

[67] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, and Jingyi Yu. Human performance modeling and rendering via neural animated mesh. *ACM Trans. Graph.*, 41(6), nov 2022. 1

[68] 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. 3, 6, 7

[69] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15893–15903, 2022.

[70] Zerong Zheng, Tao Yu, Hao Li, Kaiwen Guo, Qionghai Dai, Lu Fang, and Yebin Liu. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In *European Conference on Computer Vision (ECCV)*, Sept 2018. 2