Total-Recon: Deformable Scene Reconstruction for Embodied View Synthesis

Chonghyuk Song Gengshan Yang Kangle Deng Jun-Yan Zhu Deva Ramanan
Carnegie Mellon University

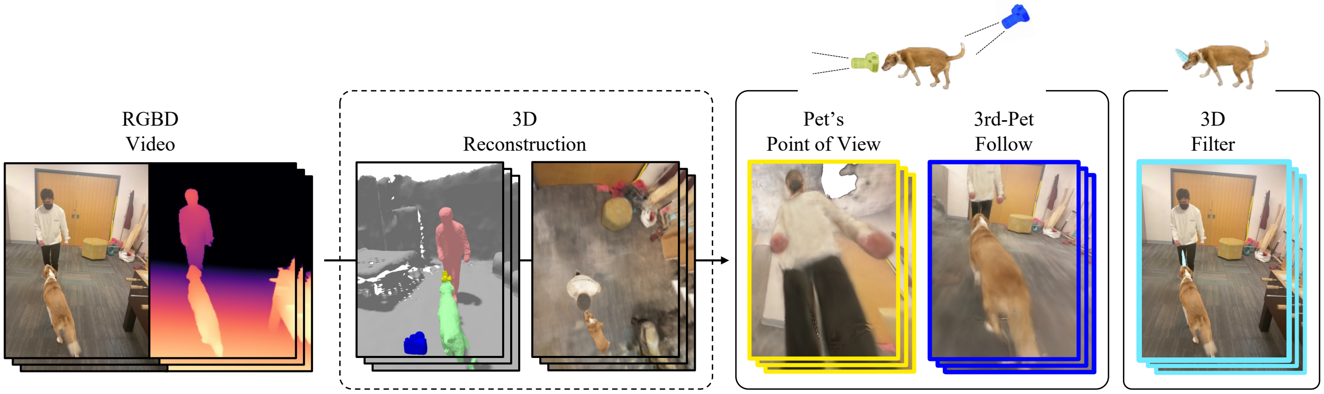

Figure 1: **Embodied View Synthesis.** Given a long video of deformable objects captured by a handheld RGBD sensor, Total-Recon renders the scene from novel camera trajectories derived from the *in*-scene motion of actors: (1) egocentric cameras that simulate the *point-of-view of a target actor* (such as the pet) and (2) 3rd-person (or pet) cameras that follow the actor from behind. Our method also enables (3) 3D video filters that attach virtual 3D assets to the actor. Total-Recon achieves this by reconstructing the geometry, appearance, root-body- and articulated motion of each deformable object in the scene and the background.

Abstract

We explore the task of embodied view synthesis from monocular videos of deformable scenes. Given a minute-long RGBD video of people interacting with their pets, we render the scene from novel camera trajectories derived from the in-scene motion of actors: (1) egocentric cameras that simulate the point of view of a target actor and (2) 3rd-person cameras that follow the actor. Building such a system requires reconstructing the root-body and articulated motion of every actor, as well as a scene representation that supports free-viewpoint synthesis. Longer videos are more likely to capture the scene from diverse viewpoints (which helps reconstruction) but are also more likely to contain larger motions (which complicates reconstruction). To address these challenges, we present Total-Recon, the first method to photorealistically reconstruct deformable scenes from long monocular RGBD videos. Crucially, to scale to long videos, our method hierarchically decomposes the scene into the background and objects, whose motion is decomposed into carefully initialized root-body motion and local articulations. To quantify such "in-the-wild" reconstruction and view synthesis, we collect ground-truth data from a specialized stereo RGBD capture rig for 11 challenging videos, significantly outperforming prior methods.

1. Introduction

We explore *embodied view synthesis*, a new class of novel-view synthesis tasks that renders deformable scenes from novel 6-DOF trajectories reconstructed from the *in*-scene motion of actors: egocentric cameras [45, 7] that simulate the point-of-view of moving actors and 3rd-person-follow cameras [54, 7] that track a moving actor from behind (Figure 1). We focus on everyday scenes of people interacting with their pets, producing renderings from the point-of-view of the person *and pet* (Figure 1). While such camera trajectories could be manually constructed (e.g., by artists via keyframing), building an *automated* system is an interesting problem of its own: spatial cognition theory [57] suggests that the ability to visualize behavior from another actor's perspective is necessary for action learning and imitation; in the context of gaming and virtual reality [7, 45], egocentric cameras offer high levels of user immersion, while 3rd-person-follow cameras provide a large field of view that is useful for exploring a user's environment.

Challenges. Building a system for embodied view synthesis is challenging for many reasons. First, to reconstruct everyday-but-interesting content, it needs to process long, monocular captures of multiple interacting actors. How-

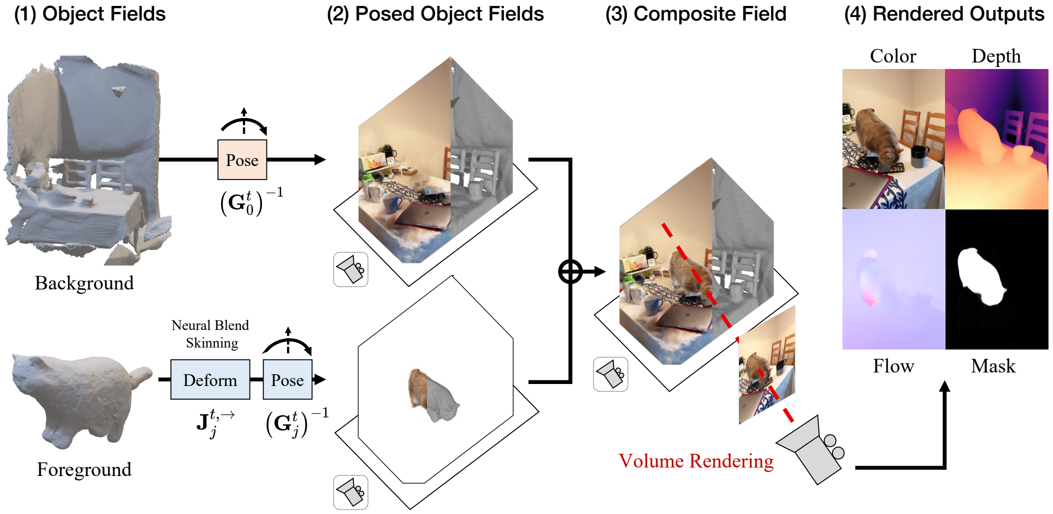

Figure 2: **Method Overview.** Total-Recon represents the entire scene as a composition of M object-centric neural fields, one for the rigid background and each of the $M-1$ deformable objects. To render a scene, (1) *each object field* j is transformed into the camera space with a rigid transformation $\left(\mathbf{G}_j^t\right)^{-1}$ that encodes root-body motion and, for each deformable object, an additional deformation field $\mathbf{J}_j^{t,\rightarrow}$ that encodes articulated motion. Next, all (2) *posed object fields* are combined into a (3) *composite field*, which is then volume-rendered into (4) *color, depth, optical flow, and object silhouettes*. Each rendered output defines a reconstruction loss that derives supervision from a monocular RGBD video captured by a moving iPad Pro.

ever, such videos are likely to contain large scene motions, which we demonstrate are difficult to reconstruct with current approaches. Second, it needs to produce a deformable 3D scene representation that supports free-viewpoint synthesis, which also would benefit from long videos likely to capture the scene from diverse viewpoints. Recent approaches have extended Neural Radiance Fields (NeRFs) [28] to deformable scenes, but such work is often limited to rigid-only object motion [18, 33], short videos with limited scene motion [41, 35, 21, 55, 36, 60, 10, 61, 58], or reconstructing single objects as opposed the entire scene [65, 66, 67, 4]. Third, it needs to compute global 6-DOF trajectories of root-bodies and articulated body parts (e.g., head) of multiple actors.

Key Ideas. To address these challenges, we introduce Total-Recon, the first monocular NeRF that enables embodied view synthesis for deformable scenes with large motions. Given a monocular RGBD video, Total-Recon reconstructs the scene as a composition of object-centric representations, which encode the 3D appearance, geometry, and motion of each deformable object and the background. Crucially, Total-Recon hierarchically decomposes scene motion into the motion of individual objects, which itself is decomposed into global root-body movement and the local deformation of articulated body parts. We demonstrate that such decomposition of object motion, along with appropriate initialization of root-body pose, allows reconstruction to scale to longer videos, enabling free-viewpoint synthesis. By reconstructing such motions in a globally-consistent coordinate frame, Total-Recon can generate renderings from egocentric and 3rd-person-follow cameras, as well as static but extreme viewpoints like bird's-eye-views.

Evaluation. Due to the difficulty of collecting ground-truth data for embodied view synthesis on in-the-wild videos, we evaluate our method on the proxy task of stereo-view synthesis [35], which compares rendered views to those captured from a stereo pair. To this end, we build a stereo RGBD sensor capture rig for ground-truthing and collect a dataset of 11 long video sequences in various indoor environments, including people interacting with their pets. Total-Recon outperforms the state-of-the-art monocular deformable NeRF methods [36, 60], even when modified to use depth sensor measurements.

Contributions. In summary, our contributions are: (1) Total-Recon, a hierarchical 3D representation that models deformable scenes as a composition of object-centric representations, each of which decomposes object motion into its global root-body motion and its local articulations; (2) a system based on Total-Recon for automated embodied view synthesis from casual, minute-long RGBD videos of highly dynamic scenes; (3) a dataset of stereo RGBD videos containing various deformable objects, such as humans and pets, in a host of different background environments. Our code, models, and data can be found at https://andrewsonga.github.io/totalrecon.

Method	Entire Scenes	Deform. Objects	Beyond Humans	Global 6-DOF Traj.	Long Videos	Extreme Views
BANMo [67]	✗	✓	✓	✗	✓	✓
PNF [18]	✓	✗	✓	✓	✗	✗
NeuMan [16]	✓	✓	✗	✓	✗	✗
SLAHMR [69]	✗	✓	✗	✓	✗	✗
HyperNeRF [36]	✓	✓	✓	✗	✗	✗
D^{2}NeRF [60]	✓	✓	✓	✗	✗	✗
DynIBaR [22]	✓	✓	✓	✗	✓	✗
SUDS [56]	✓	✓	✓	✗	✓	✗
Ours	✓	✓	✓	✓	✓	✓

Table 1: **Comparison to Related Work.** Unlike prior work, Total-Recon targets embodied view synthesis of scenes containing humans and pets, requiring the ability to (1) reconstruct *entire scenes*, (2) model *deformable objects*, (3) extend *beyond humans*, (4) recover *global 6-DOF trajectories* of objects' root-bodies and their articulated parts, (5) process *minute-long videos* of dynamic scenes, and (6) render *extreme views*.

2. Related Work

Neural Radiance Fields. Prior works on Neural Radiance Fields (NeRF) optimize a continuous scene function for novel view synthesis given a set of multi-view images, usually under the assumption of a rigid scene and densely sampled views [28, 26, 27, 23, 14, 59]. DS-NeRF [6] and Dense Depth Priors [44] extend NeRFs to the sparse-view setting by introducing depth as additional supervision. Total-Recon also operates in the sparse-view regime and uses depth supervision to reduce the ambiguities inherent to monocular, multibody, non-rigid reconstruction [62, 34]. Another line of work [18, 33] represents rigidly moving scenes as a composition of multiple object-level NeRFs. Total-Recon also leverages such an object-centric scene representation, but models scenes containing *non-rigidly* moving objects, such as humans and pets.

Deformable NeRFs. Recent approaches extend NeRF to monocular deformable scene reconstruction either by learning an additional function that deforms observed points in the camera space to a time-independent canonical space [41, 35, 55, 36, 60] or explicitly modeling density changes over time [10, 61, 58, 21]. Such methods are typically limited to short videos containing little scene and camera motion. They also perform novel-view synthesis only over small baselines. Total-Recon belongs to the former category of prior monocular deformable NeRFs, but unlike them, our method hierarchically decomposes scene motion into the motion of each object, which is further decomposed into global root-body motion and local articulations. The proposed motion decomposition is what enables embodied view synthesis: it allows Total-Recon to scale to *minute-long* videos and reconstruct a deformable 3D scene representation that supports free-viewpoint synthesis; it also makes it easy to extract an object's root-body motion, the key motion primitive required for 3rd-person-follow view synthesis. Several works have taken different approaches to making non-rigid reconstruction more tractable. One group of work leverages human-specific priors [38, 53, 32, 39, 24, 16, 19, 37] such as human body models (e.g., SMPL), 3D skeletons, or 2D poses to achieve high reconstruction quality. We achieve similar levels of fidelity *without* relying on such shape priors, allowing Total-Recon to generalize to pets and, by extension, reconstruct human-pet interaction videos. Another body of work [15, 48, 20] achieves high-fidelity scene reconstructions by relying on synchronized multi-view video captured from a specialized camera rig ranging from 8 to 18 static cameras. In contrast, Total-Recon only requires a single video captured from a moving RGBD camera equipped with inertial sensors, which has now become widely accessible in consumer products with the advent of Apple's iPhone and iPad Pro.

Reconstruction with RGBD Sensors. Depth sensors represent the third class of attempts to make non-rigid reconstruction more tractable, reducing the need for a predefined shape template. Kinect-fusion [30] creates a real-time system for indoor scene localization and mapping. Dynamic Fusion [29] builds a template-free dense SLAM system for dynamic objects. Later works improve RGBD reconstruction to be able to deal with topology changes [50, 51] and use correspondence matching for registration over large motions [2, 3]. Recent works have incorporated neural implicit representations to reconstruct the surface geometry and 3D motion fields for deformable objects [43, 4] or large-scale rigid scenes [1, 42] in isolation. Other works have reconstructed humans alongside small-scale objects and furniture [8, 2], but not the entire background. We aim to go even further by reconstructing the entire scene, which includes the background and multiple deformable targets such as humans and pets; not only do we reconstruct the geometry, but we also recover a radiance field that allows for photorealistic scene rendering from embodied viewpoints and other novel 6-DOF trajectories.

Concurrent Work. Concurrent work exhibits a subset of the design choices necessary for embodied view synthesis. SLAHMR [69] reconstructs the geometry and in-scene motion of human actors but not the scene appearance. Nerflets [71] models the appearance, geometry, and motion of each scene element but is limited to rigidly moving objects. RoDynRF [25], NeRF-DS [63], HexPlane [5], and K-planes [9] reconstruct other types of dynamic scene elements, such as deformable or specular objects, but these methods have been demonstrated on only short videos and/or videos containing limited object root-body motion [40, 47, 63, 11]. DynIBaR [22] scales dynamic view synthesis to longer videos with complex camera and scene motion, and SUDS [56] scales reconstruction to urban-scale dynamic scenes captured from 1.2 million frames. However, neither demonstrates extreme-view synthesis, a prerequisite for rendering embodied views. We summarize and compare prior work to Total-Recon in Table 1.

3. Method

3.1. Limitations of Prior Art

The state-of-the-art monocular deformable NeRFs [36, 60] decompose a deformable scene into a rigid, canonical template model and a deformation field $\mathbf{J}^{t,\leftarrow}$ that maps the world space $\mathbf{G}_0^t\mathbf{X}^t$ to the canonical space $\mathbf{X}^*$, where $\mathbf{G}_0^t$ is the *known* camera pose at time t, and $\mathbf{X}^t$ is a camera space point at time t:

$$\mathbf{X}^* = \mathcal{W}^{t,\leftarrow}\left(\mathbf{X}^t\right) = \mathbf{J}^{t,\leftarrow}(\mathbf{G}_0^t\mathbf{X}^t). \tag{1}$$

In theory, this formulation is sufficient to represent all continuous motion; it performs well on short videos containing near-rigid scenes, as the deformation field only has to learn minute deviations from the template model. However, this motion model is difficult to scale to minute-long videos, which are more likely to contain deformable objects undergoing large translations (e.g., a person walking into another room) and pose changes (e.g., a person sitting down). Here, the deformation field must learn large deviations from the canonical model, significantly complicating optimization.

Another critical limitation of HyperNeRF and D^{2}NeRF is that they cannot track separate deformable objects and therefore cannot perform 3rd-person-follow view synthesis for scenes with *multiple* actors.

3.2. Component Radiance Fields

To address the limitations of existing monocular deformable NeRFs, we propose Total-Recon, a novel 3D representation that models a deformable scene as a composition of M object-centric neural fields, one for the rigid background and each of the $M-1$ deformable objects (Figure 2). Crucially, Total-Recon hierarchically decomposes scene motion into the motion of each object, which itself is decomposed into global root-body motion and local articulations. This key design choice scales our method to minute-long videos containing highly dynamic and deformable objects.

Background Radiance Field. We begin by modeling the background environment as a Neural Radiance Field (NeRF) [28]. For a 3D point $\mathbf{X}^* \in \mathbb{R}^3$ and a viewing direction $\mathbf{v}^*$ in the canonical world space, NeRF defines a color $\mathbf{c}$ and density σ represented by an MLP. We follow contemporary variants [26] that include a time-specific embedding code ω_e^t to model illumination changes over time and model density with as a function of a neural signed distance function (SDF) $\mathbf{MLP}_\sigma(\cdot) = \alpha\Gamma_\beta(\mathbf{MLP}_{\text{SDF}}(\cdot))$ [68] to encourage the reconstruction of a valid surface:

$$\sigma = \mathbf{MLP}_\sigma(\mathbf{X}^*), \qquad \mathbf{c}^t = \mathbf{MLP}_\mathbf{c}(\mathbf{X}^*, \mathbf{v}^*, \omega_e^t). \tag{2}$$

The pixel color can then be computed with differentiable volume rendering equations (Section 3.3).

Most NeRF methods, including HyperNeRF [36] and D^{2}NeRF [60], assume images with known cameras. While our capture devices are equipped with inertial sensors, we find their self-reported camera poses have room for improvement. As such, we also model camera pose as an *optimizable* rigid-body transformation $\mathbf{G}_0^t \in SE(3)$ that maps points in a time-specific camera space $\mathbf{X}^t \in \mathbb{R}^3$ to the world space (where we assume homogenous notation):

$$\mathbf{X}^* = \mathbf{G}_0^t\mathbf{X}^t. \tag{3}$$

Deformable Field (for Object j). We model the deformable radiance field of object $j \in \{1, \cdots, M-1\}$ with BANMo [67], which consists of a canonical rest shape and time-*dependent* deformation field. The canonical rest shape is represented by the same formulation described by Equation 2, but now defined in a local *object-centric canonical space* rather than the world space. BANMo represents object motion with a warping function $\mathcal{W}_j^{t,\leftarrow} : \mathbf{X}^t \rightarrow \mathbf{X}_j^*$ that maps the camera space points $\mathbf{X}^t$ to canonical space points $\mathbf{X}_j^*$ with a rigid-body transformation $\mathbf{G}_j^t \in SE(3)$ and a deformation field $\mathbf{J}_j^{t,\leftarrow}$ modeled by linear blend skinning [13]:

$$\mathbf{X}_j^* = \mathcal{W}_j^{t,\leftarrow}\left(\mathbf{X}^t\right) = \mathbf{J}_j^{t,\leftarrow}\left(\mathbf{G}_j^t\mathbf{X}^t\right). \tag{4}$$

Note that our choice of deformation field differs from the $SE(3)$-field used in HyperNeRF and D^{2}NeRF, which has been shown to produce irregular deformation in the presence of complex scene motion [67]. Intuitively, rigid-body transformation $\mathbf{G}_j^t$ captures the global root-body pose of object j relative to the camera at time t, while deformation field $\mathbf{J}_j^{t,\leftarrow}$ aligns more fine-grained articulations relative to its local canonical space (Figure 2). Explicitly disentangling these two sources of object motion (as opposed to conflating them) enables easier optimization of the deformation field, because local articulations are significantly easier to learn than those modeled relative to the world space (Equation 1). Furthermore, this motion decomposition makes the deformation field invariant to rigid-body transformations of the object. A motion model similar to ours was proposed by ST-NeRF [15], but their model encodes an object's global root-body motion with a 3D axis-aligned bounding box that does not explicitly represent object orientation, a prerequisite for embodied view synthesis from 3rd-person-follow cameras.

Rendered features $\hat{\mathbf{f}}$ at pixel $\mathbf{x}^t$	Corresponding 3D features $\mathbf{f}_{ij}(\mathbf{X}_i^t)$
color $\hat{\mathbf{c}}(\mathbf{x}^t)$	$\mathbf{c}_i^t\left(\mathcal{W}_j^{t,\leftarrow}(\mathbf{X}_i^t)\right)$
flow $\hat{\mathcal{F}}(\mathbf{x}^t, t \rightarrow t')$	$\Pi^{t'}\left(\mathcal{W}_j^{t',\rightarrow}\left(\mathcal{W}_j^{t,\leftarrow}(\mathbf{X}_i^t)\right)\right) - \mathbf{x}^t$
depth $\hat{\mathbf{d}}(\mathbf{x}^t)$	$[0,\ 0,\ 1] \cdot \mathbf{X}_i^t$

Table 2: Rendered 2D features $\hat{\mathbf{f}}$ and their corresponding 3D features $\mathbf{f}_{ij}$. $\Pi^{t'}$ denotes the camera intrinsics at time t'.

As did BANMo, Total-Recon also models a forward warp $\mathbf{X}_j^t = \mathcal{W}_j^{t,\rightarrow}(\mathbf{X}^*) = \left(\mathbf{G}_j^t\right)^{-1} \mathbf{J}_j^{t,\rightarrow}(\mathbf{X}^*)$ that maps the canonical space to the camera space, which is used to establish the surface correspondences required for egocentric view synthesis and 3D video filters.

3.3. Composite Rendering of Multiple Objects

Given a set of M object representations (the background is treated as an object as well), we use the composite rendering scheme from prior work [31, 52] to combine the outputs of all object representations and volume-render the entire scene. To volumetrically render the image at frame t, we sample multiple points along each camera ray $\mathbf{v}^t$. Denoting the i^{th} sample as $\mathbf{X}_i^t$, we write the density and color observed at sample i due to object j as:

$$\sigma_{ij} = \mathbf{MLP}_{\sigma,j}\left(\mathbf{X}_{ij}^*\right), \qquad \mathbf{c}_{ij} = \mathbf{MLP}_{\mathbf{c},j}\left(\mathbf{X}_{ij}^*, \mathbf{v}_j^*, \omega_e^t\right),$$

where $\mathbf{X}_{ij}^* = \mathcal{W}_j^{t,\leftarrow}(\mathbf{X}_i^t)$ and $\mathbf{v}_j^* = \mathcal{W}_j^{t,\leftarrow}(\mathbf{v}^t)$ are sample i and camera ray $\mathbf{v}^t$ backward-warped into object j's canonical space, respectively. The composite density σ_i at sample i along the ray is then computed as the sum of each object's density σ_{ij}; the composite color $\mathbf{c}_i$ is computed as the weighted sum of each object's color $\mathbf{c}_{ij}$, where the weights are the normalized object densities σ_{ij}/σ_i:

$$\sigma_i = \sum_{j=0}^{M-1} \sigma_{ij}, \quad \mathbf{c}_i = \frac{1}{\sigma_i} \sum_{j=0}^{M-1} \sigma_{ij}\mathbf{c}_{ij}. \tag{5}$$

We can then use the standard volume rendering equations to generate an RGB image of the scene, where N is the number of sampled points along camera ray $\mathbf{v}^t$, τ_i is the transmittance, α_i is the alpha value for sample point i and δ_i is the distance between sample point i and the $(i+1)$:

$$\hat{\mathbf{c}} = \sum_{i=1}^{N} \tau_i \alpha_i \mathbf{c}_i, \quad \tau_i = \prod_{k=1}^{i-1}(1-\alpha_k), \quad \alpha_i = 1 - e^{-\sigma_i \delta_i}.$$

Rendering Flow, Depth, and Silhouettes. Our composite rendering scheme can be used to render different quantities by replacing the object color $\mathbf{c}_{ij}$ in Equation 5 with the appropriately defined 3D *feature* $\mathbf{f}_{ij}$ (Table 2) and rendering the resulting composite feature $\mathbf{f}_i$. To render occlusion-aware object silhouettes, we follow ObSURF [52] to produce a categorical distribution over the M objects:

$$\hat{\mathbf{o}}_j = \sum_{i=1}^{N} \tau_i \alpha_{ij}, \quad \text{where} \quad \tau_i = \prod_{k=1}^{i-1}(1-\alpha_k), \tag{6}$$

$$\alpha_i = 1 - e^{-\sigma_i \delta_i}, \qquad \alpha_{ij} = 1 - e^{-\sigma_{ij}\delta_i}. \tag{7}$$

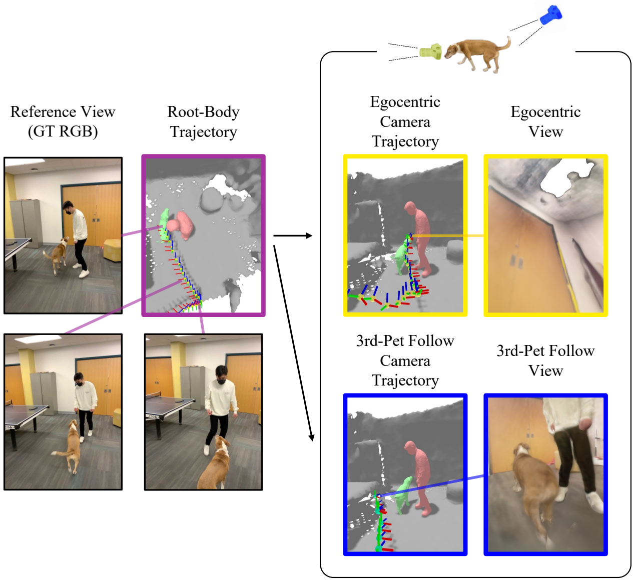

Figure 3: **6-DOF Trajectories for Embodied View Synthesis.** To synthesize embodied views from egocentric or actor-following cameras, Total-Recon reconstructs the entire background, every individual actor in the scene, as well as global 6-DOF trajectories of its root-body and its articulated body parts (e.g. head).

Losses. Given a monocular RGBD video, we optimize all parameters in our composite scene representation, which for each of the M objects includes the appearance and shape MLPs ($\mathbf{MLP}_{c,j}$, $\mathbf{MLP}_{\sigma,j}$), rigid-body transformations $\mathbf{G}_j^t$, and forward, backward deformation fields $\mathbf{J}_j^{\leftarrow}$, $\mathbf{J}_j^{\rightarrow}$. We optimize three reconstruction losses: a color loss $\mathcal{L}_{\text{rgb}}$, a flow loss $\mathcal{L}_{\text{flow}}$, and a depth loss $\mathcal{L}_{\text{depth}}$, where the ground truth color $\mathbf{c}$ and depth $\mathbf{d}$ are provided by the RGBD video, and the "ground truth" flow $\mathcal{F}$ is computed by an off-the-shelf network [64]. The model also optimizes a 3D-cycle consistency loss $\mathcal{L}_{\text{cyc},j}$ [67] for each deformable object to encourage their forward and backward warps to be consistent, where $\mathbf{x}^t \in \mathbb{R}^2$ denotes pixel location at time t:

$$\mathcal{L}_{\text{rgb}} = \sum_{\mathbf{x}^t} ||\mathbf{c}(\mathbf{x}^t) - \hat{\mathbf{c}}(\mathbf{x}^t)||^2, \tag{8}$$

$$\mathcal{L}_{\text{flow}} = \sum_{\mathbf{x}^t} ||\mathcal{F}(\mathbf{x}^t) - \hat{\mathcal{F}}(\mathbf{x}^t)||^2, \tag{9}$$

$$\mathcal{L}_{\text{depth}} = \sum_{\mathbf{x}^t} ||\mathbf{d}(\mathbf{x}^t) - \hat{\mathbf{d}}(\mathbf{x}^t)||^2, \tag{10}$$

$$\mathcal{L}_{\text{cyc},j} = \sum_{i} \tau_i \alpha_{ij} \left\| \mathcal{W}_j^{t', \rightarrow} \left(\mathcal{W}_j^{t, \leftarrow}(\mathbf{X}_i^t) \right) - \mathbf{X}_i^t \right\|^2. \tag{11}$$

Initialization. We initialize the rigid-body transformations of each deformable object $\mathbf{G}_j^t$ using a pre-trained category-specific PoseNet [67]; we initialize the rigid-body transformation of the background $\mathbf{G}_0^t$ with the camera poses provided by the iPad Pro.

Embodied View Synthesis and 3D Filters. To enable embodied view synthesis and 3D video filters (Figure 3), we design a simple interface that allows a user to select a point on a target object's surface in its reconstructed canonical mesh, and use its forward warping function $\mathcal{W}_j^{t,\rightarrow} : \mathbf{X}^* \rightarrow \mathbf{X}^t$ followed by the rigid-body transformation $\mathbf{G}_0^t$ to place the egocentric camera (or virtual 3D asset) in the world space. The surface normal to the object's mesh at the user-defined point provides a reference frame to align the egocentric camera's viewing direction and place the 3D asset. To implement a 3rd-person-follow camera, we add a user-defined offset to the object's local reference frame, which is defined by its root-body pose.

4. Experiments

Implementation Details. In practice, we train our composite scene representation by first pre-training each object field separately. For deformable objects, we pre-train using a depth loss (Equation 10) combined with the losses optimized by BANMo [67]. This includes a silhouette loss $\mathcal{L}_{\text{mask}} = \sum_{\mathbf{x}^t} ||\mathbf{o}_j(\mathbf{x}^t) - \hat{\mathbf{o}}_j(\mathbf{x}^t)||^2$, where the "ground truth" object silhouette $\mathbf{o}_j$ is computed by an off-the-shelf instance segmentation engine [17]. For pre-training the background, we optimize color, flow, and depth losses (Equations 8, 9, 10) on pixels outside the ground truth object silhouettes. Importantly, we don't supervise the object fields on frames that are not provided an object silhouette since it cannot be determined whether the absence of detection is a true or false negative.

After pre-training, we composite-render the pre-trained object fields and jointly finetune them using only the color, depth, flow, and object-specific 3D-cycle consistency losses. Since the silhouette loss is no longer used, the scene representation is supervised on *all* frames of the training sequence during joint-finetuning. We provide a complete description of the implementation details in the supplement.

Dataset. We evaluate Total-Recon on novel-view synthesis for deformable scenes. To enable quantitative evaluation, we built a stereo rig comprised of two iPad-Pros rigidly attached to a camera mount, a setup similar to that of Nerfies [35]. Using the stereo rig, we captured 11 RGBD sequences containing 3 different cats, 1 dog, and 2 human subjects in 4 different indoor environments. The RGBD videos were captured using the Record3D iOS App [49], which also automatically registers the frames captured by each camera. These video sequences, which were subsampled at 10 fps, range from 392 to 901 frames, amount to, on average minute-long videos that are significantly longer and contain more dynamic motion than the datasets introduced by [35, 36, 60, 11]. The left and right cameras were registered by solving a Perspective-n-Point (PnP) problem using manually annotated correspondences, and their videos were synchronized based on audio. We provide a complete description of our dataset in the supplement.

Reconstruction and Applications. By hierarchically decomposing scene motion into the motion of each object, which itself is decomposed into root-body motion and local articulations, Total-Recon *automatically* computes novel 6-DoF trajectories such as those traversed by egocentric cameras and 3rd-person follow cameras (Figure 3). In turn, these trajectories enable automated embodied view synthesis and 3D occlusion-aware video filters (Figure 4). These tasks are also enabled by Total-Recon's ability to recover an accurate deformable 3D scene representation, which is currently out of reach for the best of related methods (Figure 5). As shown in the bird's eye view, each reconstructed object is properly situated with respect to the background and other objects, a direct consequence of our use of depth supervision. Furthermore, even though the iPad Pro can only measure depth up to 4m, Total-Recon can *render* depth *beyond* this sensor limit by pooling the measurements from other frames into a single metric scene reconstruction. We provide results on additional sequences in the supplement.

Baselines and Evaluation. In Figure 5 and Table 3, we compare Total-Recon to D^{2}NeRF [60] and HyperNeRF [36], and their depth-supervised equivalents on the proxy task of stereo-view synthesis, a prerequisite for *embodied* view synthesis: we train each method on the RGBD frames captured from the left camera of our dataset and evaluate the images rendered from the viewpoint of the right camera. The depth-supervised versions of the baselines contain the same depth loss used in Total-Recon. We report LPIPS [70] and the average (depth) accuracy at 0.1m [42] in all subsequent experiments, and we include a more complete set

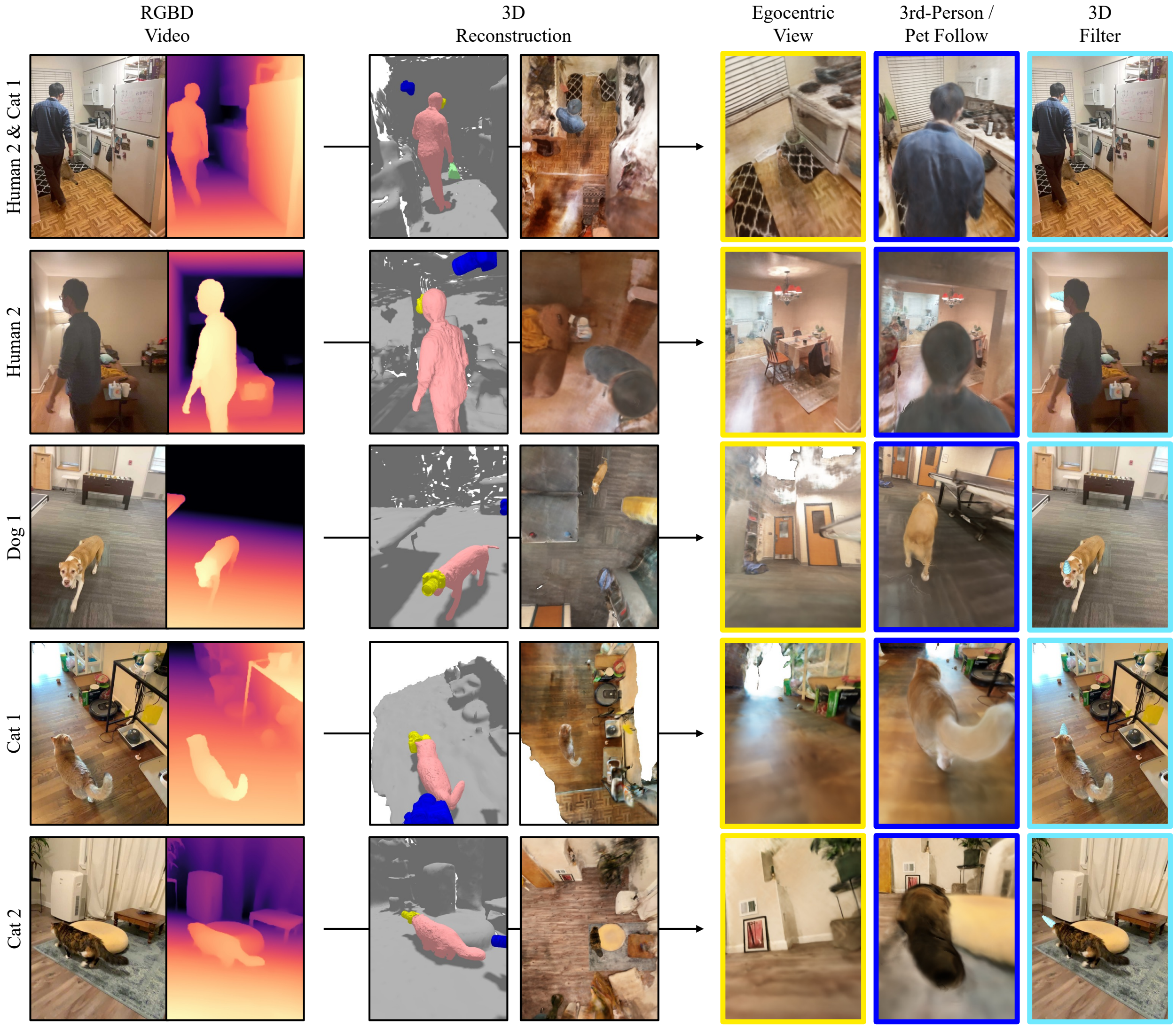

Figure 4: **Embodied View Synthesis and 3D Filters.** For select sequences of our RGBD dataset, we visualize the scene geometry and appearance reconstructed by our method (3D reconstruction) and the resulting downstream applications. The yellow and blue camera meshes in the mesh renderings represent the egocentric and 3rd-person-follow cameras, respectively. To showcase the 3D video filter we attach a sky-blue unicorn horn to the forehead of the target object, which is then automatically propagated across all frames. Full-length videos can be found at https://andrewsonga.github.io/totalrecon/applications.html.

of metrics (PSNR, SSIM, RMS depth error) in the supplement. Because D^{2}NeRF and HyperNeRF were not designed to recover a *metric* scene representation, we replaced their COLMAP [46] camera poses with those provided by the iPad Pro (which are metric measurements) for the sake of fair comparison.

Comparisons. Total-Recon qualitatively and quantitatively outperforms all of the baselines. As shown in Figure 5, Total-Recon successfully reconstructs the entire scene, whereas the baselines are only able to reconstruct the rigid background at best. As shown in Table 3, Total-Recon significantly outperforms all baselines in terms of LPIPS and the average accuracy at 0.1m (Acc@0.1m). We attribute this huge gap to the baselines' inability to reconstruct highly dynamic objects. We provide more details regarding the baselines and additional visualizations in the supplement.

4.1. Ablation Studies

Table 4 (and Figures 6 and 7) analyzes the importance of Total-Recon's design choices (see Section 3) by ablating its key components: the depth loss $\mathcal{L}_{\text{depth}}$ (row 2), the deformation field $\mathbf{J}_j^t$ (row 3), PoseNet-initialization of the root-body pose (row 4), and the root-body pose $\mathbf{G}_j^t$ itself (row 5), where j denotes a deformable actor. For all ablations, we use the same set of training losses used in Total-

	Dog 1 (626 images)		Dog 1 (v2) (531 images)		Cat 1 (641 images)		Cat 1 (v2) (632 images)		Cat 2 (834 images)		Cat 2 (v2) (901 images)		Cat 3 (767 images)		Human 1 (550 images)		Human 2 (483 images)		Human - Dog (392 images)		Human - Cat (431 images)		Mean	
	LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑
HyperNeRF	.634	.107	.432	.176	.521	.316	.438	.314	.641	.277	.397	.252	.592	.213	.632	.053	.585	.067	.487	.072	.462	.162	.531	.198
D^{2}NeRF	.540	.219	.546	.220	.687	.346	.588	.403	.556	.333	.595	.339	.759	.231	.588	.066	.630	.128	.576	.078	.628	.126	.611	.247
HyperNeRF (+depth)	.373	.352	.425	.357	.532	.552	.371	.596	.330	.605	.376	.612	.514	.451	.501	.211	.445	.249	.450	.283	.456	.214	.428	.439
D^{2}NeRF (+depth)	.507	.338	.532	.270	.685	.510	.580	.362	.561	.438	.553	.376	.730	.243	.585	.086	.609	.131	.608	.154	.645	.176	.599	.302
Total-Recon	**.271**	**.841**	**.313**	**.790**	**.382**	**.889**	**.333**	**.894**	**.237**	**.967**	**.281**	**.925**	**.261**	**.949**	**.213**	**.909**	**.264**	**.849**	**.256**	**.827**	**.233**	**.914**	**.278**	**.895**

Table 3: **Baseline Comparisons**. We train Total-Recon, HyperNeRF [36], D^{2}NeRF [60], and their depth-supervised variants on the *left* video captured with our stereo rig, and evaluate the novel view synthesis results on the *held-out* right video. Total-Recon significantly outperforms all of the baselines for all 11 sequences. These sequences are sampled at 10 fps, amounting to minute-long videos, on average.

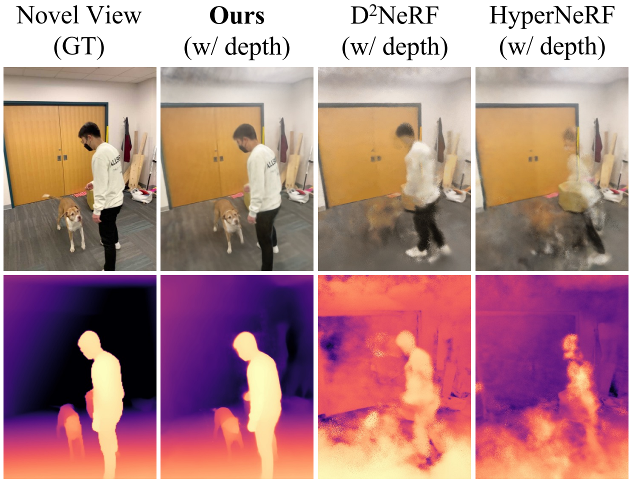

Figure 5: **Baseline Comparisons.** We compare Total-Recon to depth-supervised variants of HyperNeRF [36] and D^{2}NeRF [60] on the task of stereo-view synthesis (the left camera is used for training and the right is used for testing). While the baselines are able to reconstruct only the background, Total-Recon can reconstruct *both* the background and the moving deformable object(s), demonstrating holistic scene reconstruction.

Recon and initialize camera pose $\mathbf{G}_0^t$ with those reported by ARKit. For ablations that model root-body motions, we initialize each deformable actor's root-body pose $\mathbf{G}_j^t$ with predictions made by PoseNet [67] and optimize them during reconstruction; for row 4, we replace the PoseNet predictions with identity rotations. We report the novel-view metrics averaged over 6 select sequences of our dataset: DOG 1 (v1), CAT 1 (v1), CAT 2 (v1), HUMAN 1, HUMAN 1 & DOG 1, and HUMAN 2 & CAT 1.

Depth Supervision. Table 4 shows that removing depth supervision (row 2) significantly reduces the average accuracy at 0.1m (Acc). Figure 6 indicates that this reflects the incorrect arrangement of objects stemming from their scale inconsistency - while removing depth supervision does not significantly deteriorate the training-view RGB renderings, it induces critical failure modes as shown in the *novel-view* 3D reconstructions: (a) floating foreground objects, as evidenced by their shadows, and (b) the human incorrectly occluding the dog. In other words, without depth supervision, Total-Recon overfits the training view and learns a degenerate scene representation where the reconstructed objects fail to converge to the same scale. We show results on additional sequences in the supplement.

Methods	Depth Loss	Deform. Obj.	Root Init.	Root Motion	LPIPS↓	Acc@0.1m↑
(1) **Full model**	✓	✓	✓	✓	**.268**	**.898**
(2) w/o loss $\mathcal{L}_{\text{depth}}$	✗	✓	✓	✓	.372	.154
(3) w/o deform. $\mathbf{J}_j$	✓	✗	✓	✓	.296	.867
(4) w/o root-body init.	✓	✓	✗	✓	.293	.870
(5) w/o root-body $\mathbf{G}_j$	✓	✓	✗	✗	N/A	N/A

Table 4: **Ablation Study.** Removing depth supervision (2) significantly hurts performance, while removing the deformation field (3) and PoseNet-initialization of root-body poses (4) hurts moderately. Most importantly, removing root-body poses entirely (5) prevents convergence (N/A) as the deformation field alone has to explain *global* object motion (see Figure 2). These experiments justify our hierarchical modeling of motion, as even root-bodies without a deformation field (3) or poorly initialized root-bodies (4) are better than no root-bodies (5). We visualize these ablations in Figure 7 and explore other ablations in the supplement.

Motion Modeling. Table 4 shows that removing the deformation field (row 3) also hurts performance. This is because, without the deformation field, our method has to explain an object's non-rigid motion solely with its rigid, root-body poses. As a result, this ablation can only recover coarse object reconstructions that fail to model moving body parts such as limbs. Removing PoseNet-initialization of root-body poses (row 4) is just as detrimental, resulting in noisy appearance and geometry artifacts; see Figure 7 and additional visuals in the supplement. Most notably, Table 4 shows that removing object root-bodies entirely (row 5) causes the optimization to fail to converge (N/A), even though the deformation field can (in theory) represent all continuous motion. It appears difficult for deformation fields alone to explain *global* root-body motion because such motions can deviate significantly from a canonical model, complicating optimization.

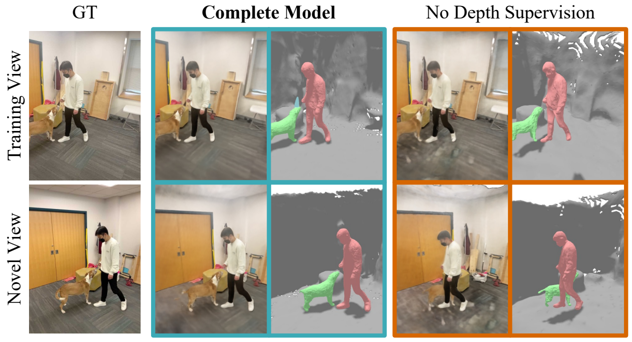

Figure 6: **Ablation Study on Depth Supervision.** While removing depth supervision does not significantly deteriorate the training-view RGB renderings, it significantly hurts the *novel-view* 3D reconstructions, as characterized by the following: (a) floating foreground objects (as evidenced by their shadows) and (b) the human incorrectly occluding the dog. These failure modes indicate that without depth supervision, Total-Recon overfits the training view, and the reconstructed objects fail to converge to the same scale.

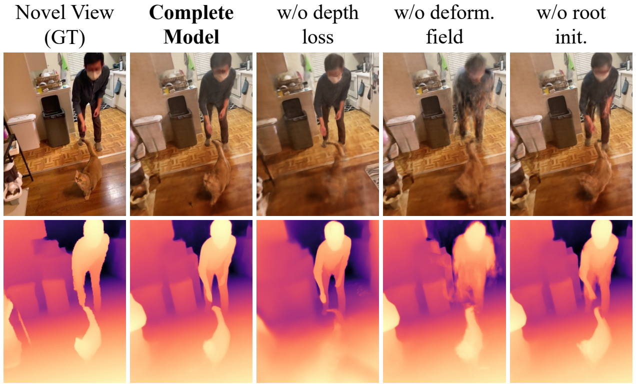

Figure 7: **Ablation Study.** We visualize the ablations from Table 4. Removing the depth loss causes the cat to sink into the ground due to inconsistent object scales. Removing the deformation field produces coarse object reconstructions that fail to capture articulated body parts such as moving limbs. Removing PoseNet-initialization of root-body poses results in noisier appearance and geometry, as shown by the human actor's left hand. We omit the ablation without root-body poses as it does not converge. We present additional visualizations in the supplement.

These diagnostics justify Total-Recon's hierarchical motion representation, which explicitly models objects' root-body poses; root-bodies without articulated deformations (row 3) or poorly initialized root-bodies (row 4) are better than no root-bodies at all (row 5). Our ablations also suggest that the poor performance of the baseline methods (on our challenging dataset) may be attributed to the lack of object-centric motion modeling. We provide a more detailed analysis with additional experiments and RGBD sequences in the supplement.

5. Discussion and Limitations

We have presented a new system for automated embodied view synthesis from monocular RGBD videos, focusing on videos of people interacting with their pets. Our main technical contribution is Total-Recon, a 3D representation for deformable scenes that hierarchically decomposes scene motion into the motion of each object, which is further decomposed into its root-body motion and local articulations; this key design choice enables appropriate initialization of root-body poses and hence easier optimization over long videos containing large motions. By explicitly reconstructing the geometry, appearance, root-body- and articulated motion of each object, Total-Recon enables seeing through the eyes of people and pets and generating game-like traversals of deformable scenes from behind a target actor.

Limitations. In Total-Recon, scene decomposition is primarily supervised by object silhouettes computed by an off-the-shelf segmentation model [17], which may be inaccurate, especially in partial occlusion scenarios. This may damage the resulting reconstructions and embodied-view renderings. We believe that incorporating the latest advances in video instance segmentation [12] will enable Total-Recon to be applied to more challenging scenarios. Second, Total-Recon initializes the root-body pose of each deformable object using a PoseNet [67] trained for humans and quadruped animals, which does not generalize to other object categories (e.g., birds, fish). We reserve the reconstruction of generic scenes for future work. Finally, our model needs to be optimized on a per-sequence basis for roughly 15 hours with 4 NVIDIA RTX A5000 GPUs and is therefore not suitable for real-time applications. Incorporating recent advances in fast neural field training methods is an interesting avenue for future research.

Acknowledgments. We thank Nathaniel Chodosh, Jeff Tan, George Cazenavette, and Jason Zhang for proofreading and Songwei Ge for reviewing the codebase. We also thank Sheng-Yu Wang, Daohan (Fred) Lu, Tamaki Kojima, Krishna Wadhwani, Takuya Narihira, and Tatsuo Fujiwara for providing valuable feedback. This work is supported in part by the Sony Corporation, Cisco Research, and the CMU Argo AI Center for Autonomous Vehicle Research.

References

[1] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[2] Aljaz Bozic, Pablo Palafox, Michael Zollhofer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. In

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 3

[3] Aljaz Bozic, Michael Zollhofer, Christian Theobalt, and Matthias Nießner. Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[4] Hongrui Cai, Wanquan Feng, Xuetao Feng, Yan Wang, and Juyong Zhang. Neural surface reconstruction of dynamic scenes with monocular rgb-d camera. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3

[5] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[6] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[7] Alena Denisova and Paul Cairns. First person vs. third person perspective in digital games: Do player preferences affect immersion? In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, page 145–148, 2015. 1

[8] Mingsong Dou, Jonathan Taylor, Henry Fuchs, Andrew Fitzgibbon, and Shahram Izadi. 3d scanning deformable objects with a single rgbd sensor. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[9] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[10] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[11] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 6

[12] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 9

[13] Alec Jacobson, Zhigang Deng, Ladislav Kavan, and J. P. Lewis. Skinning: Real-time shape deformation (full text not available). In *ACM SIGGRAPH 2014 Courses*, SIGGRAPH '14. Association for Computing Machinery, 2014. 4

[14] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[15] Zhang Jiakai, Liu Xinhang, Ye Xinyi, Zhao Fuqiang, Zhang Yanshun, Wu Minye, Zhang Yingliang, Xu Lan, and Yu Jingyi. Editable free-viewpoint video using a layered neural representation. In *ACM SIGGRAPH*, 2021. 3, 4

[16] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[17] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6, 9

[18] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J. Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[19] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[20] Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, and Zhaoyang Lv. Neural 3d video synthesis from multi-view video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[21] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[22] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[23] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[24] 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 SIGGRAPH Asia*, 2021. 3

[25] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[26] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 4

[27] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[28] 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 (ECCV)*, 2020. 2, 3, 4

[29] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[30] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136. IEEE, 2011. 3

[31] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[32] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[33] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[34] Kemal Egemen Ozden, Kurt Cornelis, Luc Van Eycken, and Luc Van Gool. Reconstructing 3d trajectories of independently moving objects using generic constraints. *Comput. Vis. Image Underst.*, 96(3):453–471, 2004. 3

[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 *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 6

[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 Transactions on Graphics (TOG)*, 40(6), 2021. 2, 3, 4, 6, 8

[37] Georgios Pavlakos*, Ethan Weber*, , Matthew Tancik, and Angjoo Kanazawa. The one where they reconstructed 3d humans and environments in tv shows. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[38] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[39] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[40] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[41] 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)*, 2020. 2, 3

[42] Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 6

[43] Zhongzheng Ren, Xiaoming Zhao, and Alex Schwing. Class-agnostic reconstruction of dynamic objects from videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[44] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[45] Richard Rouse. What's your perspective? *ACM SIGGRAPH Computer Graphics*, 33(3):9–12, August 1999. 1

[46] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[47] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[48] 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 *SIGGRAPH Conference Proceedings*, 2022. 3

[49] Marek Simonik. Record3d. `https://record3d.app`. 6

[50] Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[51] Miroslava Slavcheva, Maximilian Baust, and Slobodan Ilic. Sobolevfusion: 3d reconstruction of scenes undergoing free non-rigid motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[52] Karl Stelzner, Kristian Kersting, and Adam R Kosiorek. Decomposing 3d scenes into objects via unsupervised volume segmentation. *arXiv preprint arXiv:2104.01148*, 2021. 5

[53] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[54] Unity Technologies. 3rd person follow. `https://docs.unity3d.com/Packages/com.unity.cinemachine@2.8/manual/Cinemachine3rdPersonFollow.html`. 1

[55] 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)*, 2021. 2, 3

[56] Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 4

[57] Barbara Tversky. *Spatial Cognition Embodied and Situated*. Cambridge, 2008. 1

[58] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. Neural trajectory fields for dynamic novel view synthesis. *arXiv preprint arXiv:2105.05994*, 2021. 2, 3

[59] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 3

[60] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. D^{2}nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3, 4, 6, 8

[61] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[62] Jing Xiao, Jinxiang Chai, and Takeo Kanade. A closed-form solution to non-rigid shape and motion recovery. *International Journal of Computer Vision (IJCV)*, 67(2):233–246, April 2006. 3

[63] Zhiwen Yan, Chen Li, and Gim Hee Lee. Nerf-ds: Neural radiance fields for dynamic specular objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[64] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 5

[65] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. LASR: Learning articulated shape reconstruction from a monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[66] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[67] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4, 5, 6, 8, 9

[68] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 4

[69] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[70] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[71] Xiaoshuai Zhang, Abhijit Kundu, Thomas Funkhouser, Leonidas Guibas, Hao Su, and Kyle Genova. Nerflets: Local radiance fields for efficient structure-aware 3d scene representation from 2d supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3