Total-Recon: Deformable Scene Reconstruction for Embodied View Synthesis

SUPPLEMENTARY MATERIAL

The supplement is comprised of the following: additional details of Total-Recon's implementation (Section A), our dataset (Section B), and the baselines (Section C), additional metrics and results for the baseline comparisons (Tables 1 and 2, Figure 5), reconstructions and embodied view synthesis results on additional sequences and object removal results (Section D), additional ablation studies (Section E), and a societal impact statement (Section F).

A. Implementation Details

Data Preprocessing. Before training our composite scene representation, we follow BANMo [5] by resizing the raw RGB images and the ground truth depth maps from 960 × 720 and 256 × 192 resolution, respectively, to a resolution of 512 × 512, which is the resolution used during training. We also scale the ground-truth depth measurements and the translation component of the ARKit camera poses (used to initialize the background's root-body poses $\mathbf{G}_0^t$) by a scaling factor of 0.2, which we empirically found to improve the pre-training of the deformable objects. After training, we scale the ground-truth and rendered depth back to the original metric space and compute the evaluation metrics at 480 × 360 resolution.

Optimization. We optimize our composite scene representation by first pre-training each object field separately and then jointly finetuning them. We use the same batch size, sampled rays per batch, and sampled points per ray as BANMo [5] for both the pre-training and joint-finetuning stages. pre-training a deformable object takes 8.5 hours with 4 NVIDIA RTX A5000 GPUs, and pre-training the background takes 4.5 hours. Jointly finetuning one deformable object and the background takes an additional 1.5 hours with 4 NVIDIA RTX A5000 GPUs, and jointly finetuning two deformable objects and the background takes an additional 2.5 hours with 4 NVIDIA RTX A6000 GPUs.

Pre-training. For pre-training deformable objects, we follow the training procedure of and use the same hyperparameters as BANMo [5], which we augment with a depth reconstruction loss weighted by a default value of $\lambda_{\text{depth}} = 5$ (for the HUMAN 1 sequence, we use a loss weight of $\lambda_{\text{depth}} = 1.5$ for pre-training the deformable object). Following BANMo, we pre-train each deformable object in three training stages, each for 24k, 6k, and 24k iterations.

For pre-training the background model, we optimize color, flow, and depth reconstruction losses $\mathcal{L}_{\text{rgb}}$, $\mathcal{L}_{\text{flow}}$, $\mathcal{L}_{\text{depth}}$ on pixels outside the ground-truth object silhouettes, each with a default weight of $\lambda_{\text{rgb}} = 0.1$, $\lambda_{\text{flow}} = 1$, $\lambda_{\text{depth}} = 1$, respectively. We also optimize an eikonal loss $\mathcal{L}_{\text{SDF}}$ [6] with a weight of $\lambda_{\text{SDF}} = 0.001$ to encourage the reconstruction of a valid signed distance function (SDF):

$$\mathcal{L}_{\text{SDF}} = \sum_{\mathbf{x}^t} \sum_{\mathbf{X}_i^t} (||\nabla_{\mathbf{X}_i^*} \mathbf{MLP}_{\text{SDF}}(\mathbf{X}_i^*)||_2 - 1)^2, \tag{1}$$

where $\mathbf{x}^t \in \mathbb{R}^2$ denotes the pixel location at time t, $\mathbf{X}_i^* \in \mathbb{R}^3$ is the 3D point in the canonical world space corresponding to $\mathbf{X}_i^t \in \mathbb{R}^3$, the i^{th} sample in the camera space. To compute this eikonal loss, we sample 17 uniformly spaced points $\mathbf{X}_i^t$ along each camera ray $\mathbf{v}^t$ from a truncated region that is 0.2m long and centered at the surface point computed by backprojecting the ground-truth depth.

We pre-train the background model in two stages: in the first stage, we optimize the color, flow, depth, and eikonal losses with their respective default loss weights for 24k iterations. In the second stage, we optimize the same set of losses for another 24k iterations while fixing the background model's root-body poses $\mathbf{G}_0^t$, increasing the weight of the color loss from $\lambda_{\text{rgb}} = 0.1$ to $\lambda_{\text{rgb}} = 1$, and performing active sampling of pixels $\mathbf{x}^t$ to improve the background model's appearance, as was done in [5].

Joint Finetuning. For the joint-finetuning of all of the object models, we optimize the color, flow, depth, and per-object 3D-cycle consistency losses for another 6k iterations, each with a default weight of $\lambda_{\text{rgb}} = 1$, $\lambda_{\text{flow}} = 1$, $\lambda_{\text{depth}} = 5$, and $\lambda_{\text{cyc}, j} = 1$, respectively. Importantly, we freeze the background's appearance and shape models by default and only allow its root-body poses $\mathbf{G}_0^t$, the foregrounds' root-body poses $\mathbf{G}_j^t$, and the foregrounds' appearance and shape models to be optimized (for the HUMAN 1 sequence, we use a loss weight of $\lambda_{\text{depth}} = 1.5$, and for the CAT 1 and CAT 1 (V2) sequences, we allow the background's appearance and shape models to be optimized during joint-finetuning). We also perform active sampling of pixels $\mathbf{x}^t$ over all deformable foreground objects. Intuitively, the joint-finetuning stage improves the appearance of the foreground objects and helps the model learn correct object-to-object interactions.

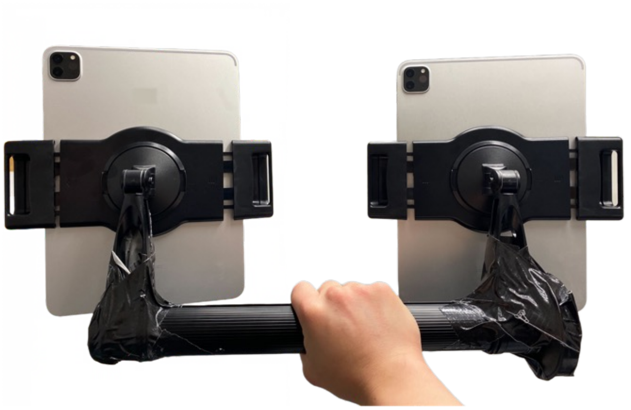

Figure 1: **Stereo Validation Rig Used Only for Evaluation**. To enable quantitative evaluation, we built a stereo rig comprised of two iPad Pros rigidly attached to a camera mount and captured 11 pairs of RGBD sequences. **We train each method *only* on the sequences captured from the left camera** and evaluate the images rendered from the viewpoint of the right camera.

B. Dataset Details

In this section, we describe the stereo validation rig we built for evaluation and elaborate on how the validation rig is used to evaluate novel-view synthesis.

As shown in Figure 1, our stereo validation rig is comprised of two iPad Pro's rigidly attached to a camera mount. Importantly, **we train each method *only* on the sequences captured from the left camera** and evaluate the images rendered from the viewpoint of the right camera *i.e.,* the "novel-view". To compute the pose of the "novel-view" camera, we compute the rigid transform between the left and right cameras and use this transform to map the optimized training-view cameras of our method to the novel-view cameras. For each sequence, we register the two cameras by solving a Perspective-n-Point (PnP) problem using manually annotated 2D-2D correspondences.

The PnP problem aims to estimate the pose of a *calibrated camera* given n 3D-2D correspondences *i.e.,* a set of n 3D points defined in some *world frame* and their corresponding 2D image projections. We formulate the problem of estimating the left-to-right camera transform as a PnP problem where the left camera of our validation rig corresponds to the *world frame*, and the right camera of our validation rig corresponds to the *calibrated camera*.

To obtain 3D-2D correspondences, we first manually annotate at least 20 2D-2D correspondences for each sequence. Next, we obtain the 3D points defined in the frame of the left camera by backprojecting its ground-truth depth using the provided intrinsics. Finally, we feed 1) the 3D points in the left-camera frame, 2) the 2D annotations, and 3) the intrinsics of the right camera to a generic PnP solver to compute the desired left-to-right camera transform.

Using the stereo validation rig, we captured a dataset containing 11 pairs of RGBD sequences featuring 3 different cats, 1 dog, and 2 human subjects in 4 different indoor environments. For each sequence, we provide 1) the RGBD frames (and object masks) captured from both cameras of our validation rig, 2) their camera pose trajectories, 3) their camera intrinsics, and 4) the left-to-right camera transform.

C. Details of Baseline Comparisons

Baseline Experiment Details. We provide additional details of the experiment settings of the baselines. For the sake of fair comparison, we set up augmented versions of the baselines D^{2}NeRF [4] and HyperNeRF [2], whereby we replace their COLMAP [3] camera poses with the iPad Pro's camera poses provided by ARKit - the same camera poses used to initialize the root-body transforms $\mathbf{G}_0^t$ of our method's background model. We also compare our method to depth-supervised variants of HyperNeRF and D^{2}NeRF, which uses the same losses and hyperparameters as the raw baselines, with the exception of an additional depth loss with weight $\lambda_{\text{depth}} = 0.1$. We empirically observe that using a higher depth-loss weight significantly deteriorates the baseline methods' rendered appearance. As was done for our method, before training, we scale the ground-truth depth measurements and the translation component of the ARKit camera poses by a scaling factor of 0.2; after training, we scale both the ground-truth and rendered depth back to the original metric space and compute the evaluation metrics at 480 × 360 resolution.

Additional Qualitative Results. In Figure 5, we show qualitative comparisons on the remaining 10 sequences of our RGBD dataset that were not shown in the main paper, namely sequences HUMAN 2 & CAT1, HUMAN 2, DOG 1 (v1), DOG 1 (v2), CAT 1 (v1), CAT 1 (v2), CAT 2 (v1), CAT 2 (v2), CAT 3, HUMAN 1. Due to space limits, we only display the visualizations for the depth-supervised variants of the baselines, but we showcase the complete set of baselines at https://andrewsonga.github.io/totalrecon/nvs.html. Total-Recon outperforms all of the baselines, which can only reconstruct the rigid background. On the other hand, Total-Recon reconstructs the *entire* scene, including all dynamic objects.

Additional Quantitative Metrics. In Tables 1 and 2, we display the full set of quantitative metrics for our method and all of the baselines. In addition to the LPIPS and the average depth accuracy at 0.1m reported in the main paper, we also report the PSNR, SSIM, and RMS depth error. Our method significantly outperforms all of the baselines in terms of LPIPS, PSNR, SSIM, the average depth accuracy at 0.1m, and the RMS depth error for all sequences.

	Dog 1 (v1) (626 images)			Dog 1 (v2) (531 images)			Cat 1 (v1) (641 images)			Cat 1 (v2) (632 images)			Cat 2 (v1) (834 images)			Cat 2 (v2) (901 images)		
	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑
HyperNeRF [2]	.634	12.84	.673	.432	14.27	.721	.521	14.86	.632	.438	14.87	.597	.641	12.32	.632	.397	15.68	.657
D^{2}NeRF [4]	.540	13.37	.694	.546	11.74	.685	.687	10.92	.545	.588	11.88	.548	.556	12.55	.664	.595	12.71	.604
HyperNeRF (w/ depth)	.373	16.86	.730	.425	16.95	.740	.532	14.37	.621	.371	15.65	.617	.330	18.47	.728	.376	16.56	.670
D^{2}NeRF (w/ depth)	.507	13.44	.698	.532	11.88	.690	.685	10.81	.534	.580	12.00	.563	.561	12.59	.656	.553	12.76	.629
Ours (w/ depth)	**.271**	**17.60**	**.745**	**.313**	**17.78**	**.768**	**.382**	**15.77**	**.657**	**.333**	**16.44**	**.652**	**.237**	**21.22**	**.793**	**.281**	**18.52**	**.713**

	Cat 3 (767 images)			Human 1 (550 images)			Human 2 (483 images)			Human - Dog (392 images)			Human - Cat (431 images)			Mean		
	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑
HyperNeRF [2]	.592	13.74	.624	.632	11.94	.603	.585	14.97	.620	.487	15.04	.699	.462	13.52	.512	.531	14.00	.635
D^{2}NeRF [4]	.759	11.03	.578	.588	11.88	.638	.630	12.13	.599	.576	12.41	.652	.628	10.41	.453	.611	11.97	.608
HyperNeRF (w/ depth)	.514	14.86	.635	.501	13.25	.664	.445	15.58	.665	.450	15.01	.704	.456	14.40	.535	.428	15.80	.667
D^{2}NeRF (w/ depth)	.730	11.08	.582	.585	12.14	.638	.609	12.11	.612	.608	12.30	.633	.645	10.51	.451	.599	12.02	.611
Ours (w/ depth)	**.261**	**19.89**	**.734**	**.213**	**18.39**	**.778**	**.264**	**16.73**	**.712**	**.256**	**16.69**	**.756**	**.233**	**17.67**	**.630**	**.278**	**18.11**	**.724**

Table 1: **Quantitative Comparisons on Novel View Synthesis (Visual Metrics)**. We compare our method to HyperNeRF [2], D^{2}NeRF [4], and their depth-supervised variants on the 11 sequences of our stereo RGBD dataset, in terms of LPIPS, PSNR, and SSIM. Our method significantly outperforms all baselines for all sequences.

	Dog 1 (v1) (626 images)		Dog 1 (v2) (531 images)		Cat 1 (v1) (641 images)		Cat 1 (v2) (632 images)		Cat 2 (v1) (834 images)		Cat 2 (v2) (901 images)	
	Acc@0.1m↑	ϵ_{depth}↓	Acc@0.1m↑	ϵ_{depth}↓	Acc@0.1m↑	ϵ_{depth}↓	Acc@0.1m↑	ϵ_{depth}↓	Acc@0.1m↑	ϵ_{depth}↓	Acc@0.1m↑	ϵ_{depth}↓
HyperNeRF [2]	.107	.687	.176	.870	.316	.476	.314	.564	.277	.765	.252	.811
D^{2}NeRF [4]	.219	.463	.220	.456	.346	.334	.403	.314	.333	.371	.339	.361
HyperNeRF (w/ depth)	.352	.331	.357	.338	.552	.206	.596	.209	.605	.154	.612	.170
D^{2}NeRF (w/ depth)	.338	.423	.270	.445	.510	.325	.362	.313	.438	.298	.376	.318
Ours (w/ depth)	**.841**	**.165**	**.790**	**.167**	**.889**	**.184**	**.894**	**.124**	**.967**	**.050**	**.925**	**.081**

	Cat 3 (767 images)		Human 1 (550 images)		Human 2 (483 images)		Human - Dog (392 images)		Human - Cat (431 images)		Mean	
	Acc@0.1m↑	ϵ_{depth}↓	Acc@0.1m↑	ϵ_{depth}↓	Acc@0.1m↑	ϵ_{depth}↓	Acc@0.1m↑	ϵ_{depth}↓	Acc@0.1m↑	ϵ_{depth}↓	Acc@0.1m↑	ϵ_{depth}↓
HyperNeRF [2]	.213	.800	.053	.821	.067	1.665	.072	.894	.162	.862	.198	.855
D^{2}NeRF [4]	.231	.523	.066	1.063	.128	.890	.078	.847	.126	.880	.247	.739
HyperNeRF (w/ depth)	.451	.285	.211	.591	.249	.611	.283	.565	.214	.613	.439	.374
D^{2}NeRF (w/ depth)	.243	.496	.086	.984	.131	.813	.154	.789	.176	.757	.302	.549
Ours (w/ depth)	**.949**	**.066**	**.909**	**.142**	**.849**	**.142**	**.827**	**.204**	**.914**	**.104**	**.895**	**.131**

Table 2: **Quantitative Comparisons on Novel View Synthesis (Depth Metrics)**. We compare our method to HyperNeRF [2], D^{2}NeRF [4], and their depth-supervised variants on the 11 sequences of our stereo RGBD dataset, in terms of the average accuracy at 0.1m (Acc@0.1m) and the RMS depth error ϵ_{depth} (units: meters). Our method significantly outperforms all baselines for all sequences.

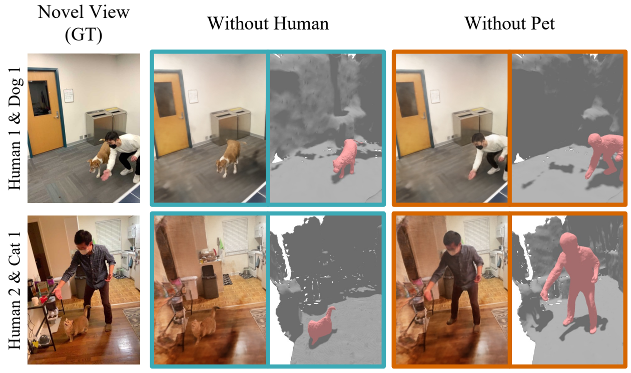

Figure 2: **Object Removal.** Our compositional scene representation enables object removal. We remove the HUMAN and then the PET object from our composite rendering process and display the resulting renderings.

D. Reconstruction and Applications

Geometry and Embodied View Synthesis. In Figure 4, we display the novel-view reconstructions and the corresponding embodied view synthesis results for the remaining 5 sequences of our RGBD dataset that were not shown in the main paper: sequences HUMAN 1, DOG 1 (v2), CAT 1 (v2), CAT 2 (v2), CAT 3.

Object Removal. Our compositional scene representation allows for easy object removal. To remove object k from our trained scene representation, one skips $j = k$ in the summation that appears in the compositing process described by Equation 5 of the main paper. We showcase object removal in Figure 2.

	Dog 1		Cat 1		Cat 2		Human 1		Human - Dog		Human - Cat		Mean	
	(626 images)		(641 images)		(834 images)		(550 images)		(392 images)		(431 images)			
	LPIPS↓	Acc@0.1m↑	LPIPS↓	Acc@0.1m↑	LPIPS↓	Acc@0.1m↑	LPIPS↓	Acc@0.1m↑	LPIPS↓	Acc@0.1m↑	LPIPS↓	Acc@0.1m↑	LPIPS↓	Acc@0.1m↑
w/o depth	.307	.296	.496	.051	.287	.193	.314	.125	.376	.206	.519	.017	.372	.154
Full (w/ depth)	**.271**	**.841**	**.382**	**.889**	**.237**	**.967**	**.213**	**.909**	**.256**	**.827**	**.233**	**.914**	**.268**	**.898**

Table 3: **Ablation Study on Depth Supervision.** Depth supervision improves our model both in terms of the visual (LPIPS) and depth (Acc@0.1m) metrics, an observation that is consistent with the qualitative results displayed in Figure 7.

Methods	Optimizes Camera	Deformation Field	Deformable Objects	Root-Body Initialization	Root-Body Motion	Dog 1		Cat 1		Cat 2		Human 1		Human - Dog		Human - Cat		Mean	
						(626 images)		(641 images)		(834 images)		(550 images)		(392 images)		(431 images)			
						LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑	LPIPS↓	Acc↑
(1) **Full**	✓	NBS	✓	✓	✓	**.271**	**.841**	**.382**	.889	**.237**	**.967**	.213	.909	**.256**	.827	**.233**	**.914**	**.268**	**.898**
(2) w/o cam. opt.	✗	NBS	✓	✓	✓	.315	.801	.407	**.898**	.270	.959	**.202**	**.920**	.268	**.833**	.283	.851	.294	.885
(3) w/ SE(3)-field	✓	SE(3)-field	✓	✓	✓	.274	.833	.443	.786	.257	.930	.217	.893	.395	.619	.245	.898	.302	.841
(4) w/o deform. field	✓	None	✗	✓	✓	.297	.833	.408	.872	.250	.940	.243	.862	.298	.798	.285	.833	.296	.867
(5) w/o root-body init.	✓	NBS	✓	✗	✓	.311	.821	.410	.848	.251	.951	.214	.892	.322	.747	.250	.899	.293	.870
(6) w/o root-body	✗[†]	NBS	✓	✗	✗	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
(7) w/o root-body (SE3)	✗	SE(3)-field	✓	✗	✗	.373	.703	.437	.749	.311	.892	.376	.640	.326	.758	.328	.805	.360	.766

Table 4: **Ablation Study on Motion Modeling.** Ablating camera-pose optimization (row 2), changing the deformation field (row 3), removing deformation modeling (row 4), or removing PoseNet initialization of object root-body poses (row 5) moderately hurts the visual and depth metrics. Importantly, removing root-body modeling entirely (row 6) prevents our method from converging (N/A), as the deformation field alone has to explain global object motion (see Figure 2 in the main paper). We perform another ablation (row 7) that replaces Total-Recon's neural blend skinning (NBS) function with the more flexible SE(3)-field [1], which does converge but still performs worse than other converging ablations. These experiments justify Total-Recon's hierarchical motion representation, which decomposes object motion into global root-body motion and local articulations. [†]When ablating root-body poses, we freeze the camera poses to prevent the object fields, which are now all defined in the world space, from learning different camera poses during their separate pre-training processes.

E. Additional Ablation Studies

E.1. Ablation Study on Depth Supervision

In this section, we perform an ablation study on depth supervision for additional sequences in our dataset. Figure 7 shows that while removing depth supervision from Total-Recon does not significantly deteriorate the training-view RGB renderings, it induces critical failure modes as shown in the *novel-view* 3D reconstructions: (a) *Floating objects*: for the Human 1 & Dog 1, Dog 1, Human 1, and Cat 2 sequences, the foreground objects float above the ground, as evidenced by their shadows. (b) *Objects that sink into the background*: for the Human 2 & Cat 1 sequence, the reconstructed cat is halfway sunk into the ground. (c) *Incorrect occlusions*: for the Human 1 & Dog 1 sequence, the human is incorrectly occluding the dog. (d) *Lower reconstruction quality*: for the Human 2 & Cat 1 sequence, the cat displays lower reconstruction quality, and for all sequences except Human 1 & Dog 1 and Cat 2, the background exhibits lower reconstruction quality.

These observations are corroborated by Table 3, which shows that depth supervision significantly improves our method's visual and depth metrics over all sequences. Another reason for the large difference in metrics is that the novel-view cameras computed for the non-depth-supervised version may not be entirely accurate. This is because our method optimizes the camera poses during training, meaning that in the absence of the depth loss, the training-view camera poses may converge to a different scale to the ground-truth left-to-right camera transform from Section B, resulting in slightly misaligned novel-view cameras.

E.2. Ablation Study on Motion Modeling

In this section, we perform ablation studies on Total-Recon's motion model for a more comprehensive set of design choices than those presented in the main paper. Table 4 and Figure 6 show that ablating camera-pose optimization (row 2) worsens the metrics but does not result in qualitative deterioration of the scene reconstruction. This suggests that the ARKit camera poses used to initialize $\mathbf{G}_0^t$ (Equation 3 of main paper) are already reasonably accurate. Changing the deformation field from Total-Recon's neural blend skinning (NBS) function to the SE(3)-field used in HyperNeRF [2] (row 3) further worsens the metrics, which are reflected in the minor artifacts that appear in the foreground reconstructions. Removing the deformation field entirely (row 4) also worsens the results, as our method now has to explain each object's (non-rigid) motion solely via 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. Ablating PoseNet-initialization of root-body poses (row 5) is just as detrimental, resulting in noisy appearance and geometry and sometimes even failed object reconstructions (see Dog 1 sequence in Figure 6).

Most notably, Table 4 shows that removing object root-body poses entirely (row 6) prevents our method from con-

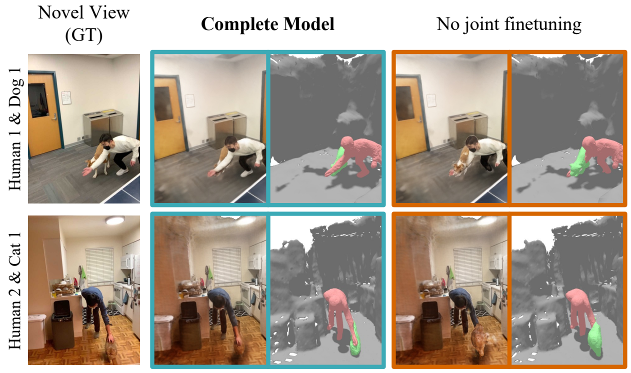

Figure 3: **Ablation Study on Joint-Finetuning.** Joint-finetuning enables Total-Recon to learn the correct human-pet interactions, particularly for frames without any detected object masks.

verging, even though the deformation field should be sufficient (in theory) to represent all continuous motion. However, when root-body poses are removed from our method, each object's canonical model is defined in the *static, world* coordinate frame (Equation 1 from main paper) as opposed to the moving, object-centric coordinate frame (Equation 4 from main paper). Therefore, the deformation field alone has to explain *global* object motion by learning potentially large deviations from the canonical model, significantly complicating optimization. We posit that Total-Recon's neural blend skinning (NBS) function is too constraining of a deformation field to model global object motion, so we perform another ablation that replaces NBS with the more flexible SE(3)-field (row 7). This ablation does converge but still performs worse than other converging ablations.

These diagnostics justify Total-Recon's hierarchical motion representation, which explicitly models objects' root-body motion; even root-bodies without a deformation field (row 4) or poorly initialized root-bodies (row 5) outperform no root-bodies (row 6). Our ablations also suggest that the poor performance of the baseline methods may be attributed to the lack of object-centric motion modeling, especially since the baseline method D^{2}NeRF (w/ depth) and the ablation w/o root-body (SE3) both exhibit the ghosting artifacts that indicate failed foreground reconstruction (see Figures 5 and 6, respectively). Note that these two methods are not strictly equivalent.

E.3. Ablation Study on Joint Finetuning

In Figure 3, we show that joint-finetuning is indispensable by visualizing its effects on frames without any detected object segmentation masks, which often exist in human-pet interaction videos due to partial occlusions. Since our method does not supervise on frames without segmentation masks during pre-training, the appearance, deformation, and root-body pose of the deformable foreground objects remain uncertain for such frames.

For the human1 & dog1 sequence, the dog ends up penetrating the human arm; for the human2 & cat1 sequence, the cat lies in front of the human hand rather than behind it. Joint-finetuning resolves these issues as it does not optimize a silhouette loss, enabling our scene representation to be supervised on all frames of the training sequence. By jointly optimizing all objects in the scene, our scene representation learns the correct human-pet interactions by reasoning about occlusions, resulting in a general improvement of the visual metrics, as shown in Table 5. Note that joint-finetuning doesn't always improve the depth metrics. We posit that the depth supervision during pre-training was sufficient in learning a metric model that the qualitative improvements brought by joint-finetuning are not always reflected in the metrics.

F. Societal Impact

We have demonstrated that our method can holistically reconstruct a dynamic scene containing multiple deformable objects, such as humans and pets - all from a single RGBD video captured from a commodity consumer device. We believe that a truly holistic reconstruction of the background geometry, each moving object, its own deformation, and camera pose would enable a number of new applications ranging from augmented reality to asset generation for virtual worlds, especially given the ubiquity of consumer-grade RGBD sensors.

However, the reconstruction capabilities of our method could be a double-edged sword; the very ease with which one could reconstruct a realistic 3D human model from nothing but a casually captured RGBD video poses potential privacy concerns. For instance, one could extract sensitive personal information such as height and other body measurements from a metric human model reconstructed with our method. In terms of appearance synthesis, our method poses similar types of risks as Deepfakes pose to society, especially given that the deformable object model used in our method is animatable (*i.e.*, user-drivable) [5]. An important future direction of research that needs to accompany 3D reconstruction research would therefore be methods of distinguishing photorealistic rendered videos from genuine content.

	Dog 1 (626 images)		Cat 1 (641 images)		Cat 2 (834 images)		Human 1 (550 images)		Human - Dog (392 images)		Human - Cat (431 images)		Mean	
	LPIPS↓	Acc@0.1m↑	LPIPS↓	Acc@0.1m↑	LPIPS↓	Acc@0.1m↑	LPIPS↓	Acc@0.1m↑	LPIPS↓	Acc@0.1m↑	LPIPS↓	Acc@0.1m↑	LPIPS↓	Acc@0.1m↑
w/o joint-ft	.273	**.859**	**.379**	**.889**	.239	.963	**.207**	**.916**	.259	**.840**	.241	.901	.268	**.902**
Full (w/ joint-ft)	**.271**	.841	.382	.889	**.237**	**.967**	.213	.909	**.256**	.827	**.233**	**.914**	**.268**	.898

Table 5: **Ablation Study on Joint-Finetuning.** Joint-finetuning improves LPIPS across most sequences but does not always improve the average depth accuracy at 0.1m . We posit that using a depth signal during pre-training of the individual objects is sufficient for learning a metric model, such that the qualitative improvements induced by joint-finetuning (Figure 3) are not always reflected in the metrics.

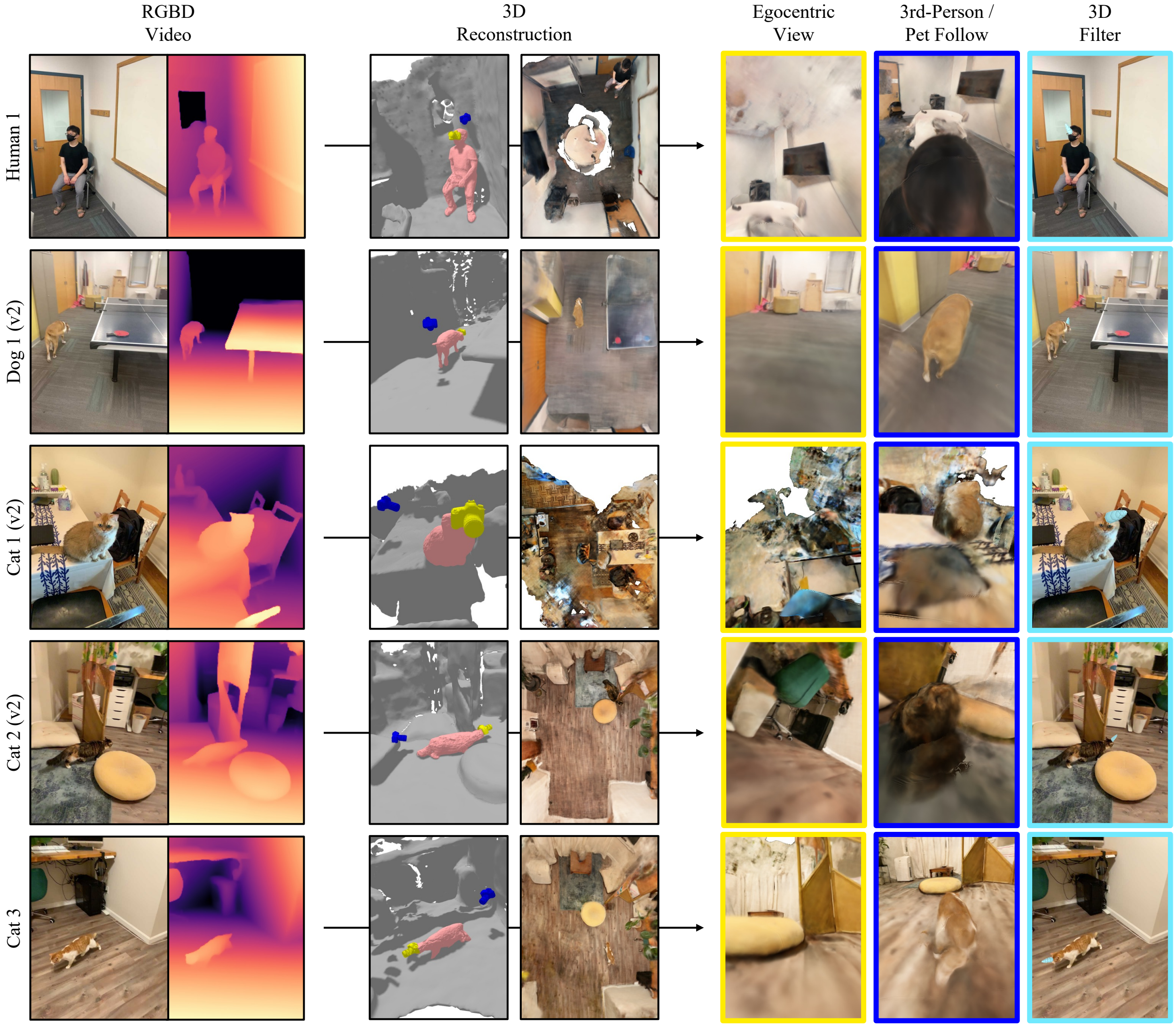

Figure 4: **Embodied View Synthesis and 3D Filters (Additional Sequences).** We visualize the 3D reconstructions (rendered from the novel view) and the applications enabled by Total-Recon for the remaining 5 sequences of our RGBD dataset that were not shown in the main paper. The yellow and blue camera meshes in the mesh renderings represent the egocentric and 3rd-person-follow cameras, respectively. To showcase the 3D filter, we attach a sky-blue unicorn horn to the forehead of the foreground object, which is automatically propagated across all frames. Full-length videos can be found at `https://andrewsonga.github.io/totalrecon/applications.html`.

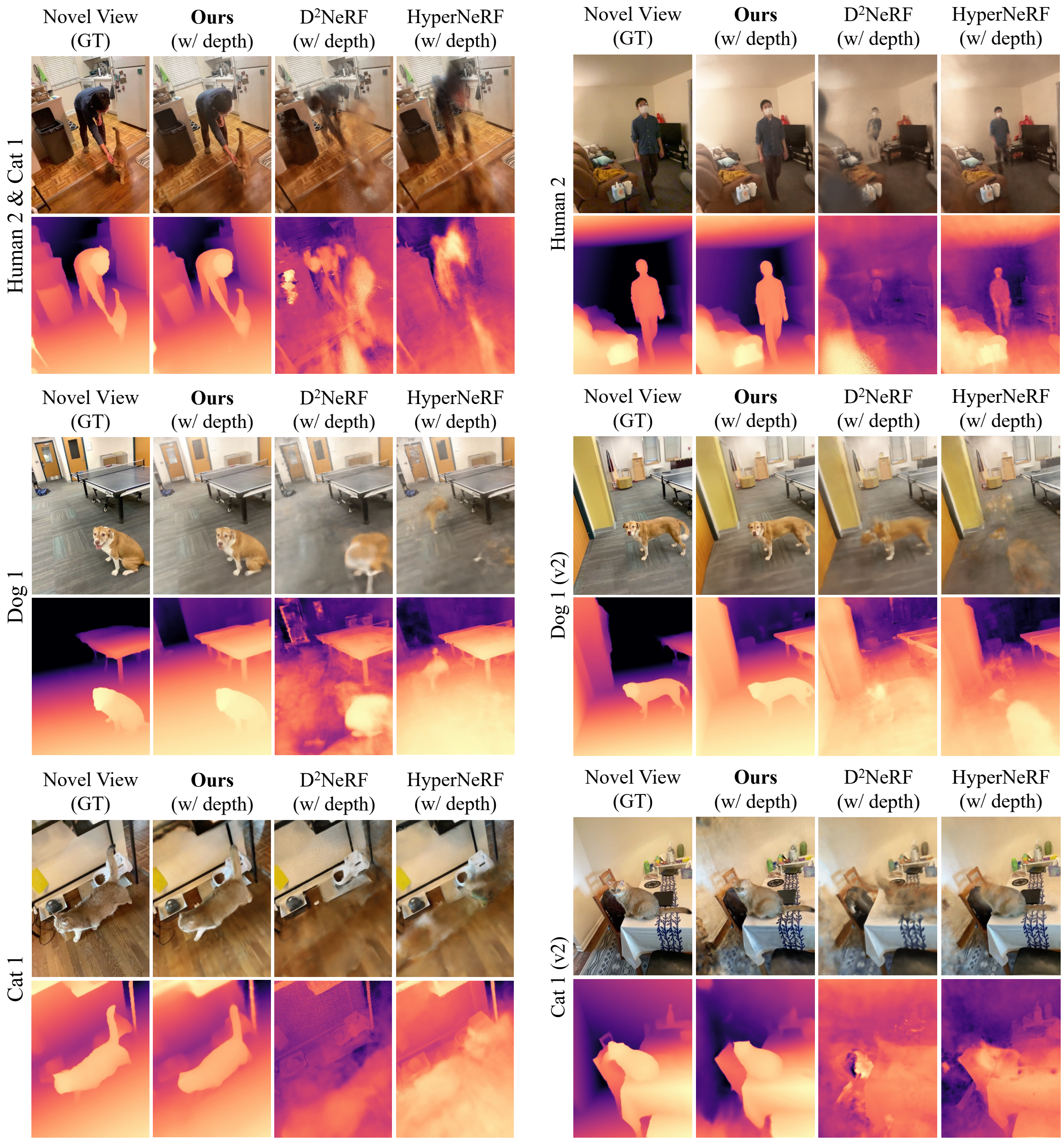

Figure 5: **Qualitative Comparisons on Novel View Synthesis (Additional Sequences).** We compare Total-Recon to depth-supervised variants of HyperNeRF [2] and D^{2}NeRF [4] on the task of stereo-view synthesis (the left camera is used for training and the images are rendered to the right camera). While the baselines are only able to reconstruct the background at best, Total-Recon is able to reconstruct *both* the background and the moving deformable object(s), demonstrating holistic scene reconstruction. Full-length videos can be found at https://andrewsonga.github.io/totalrecon/nvs.html.

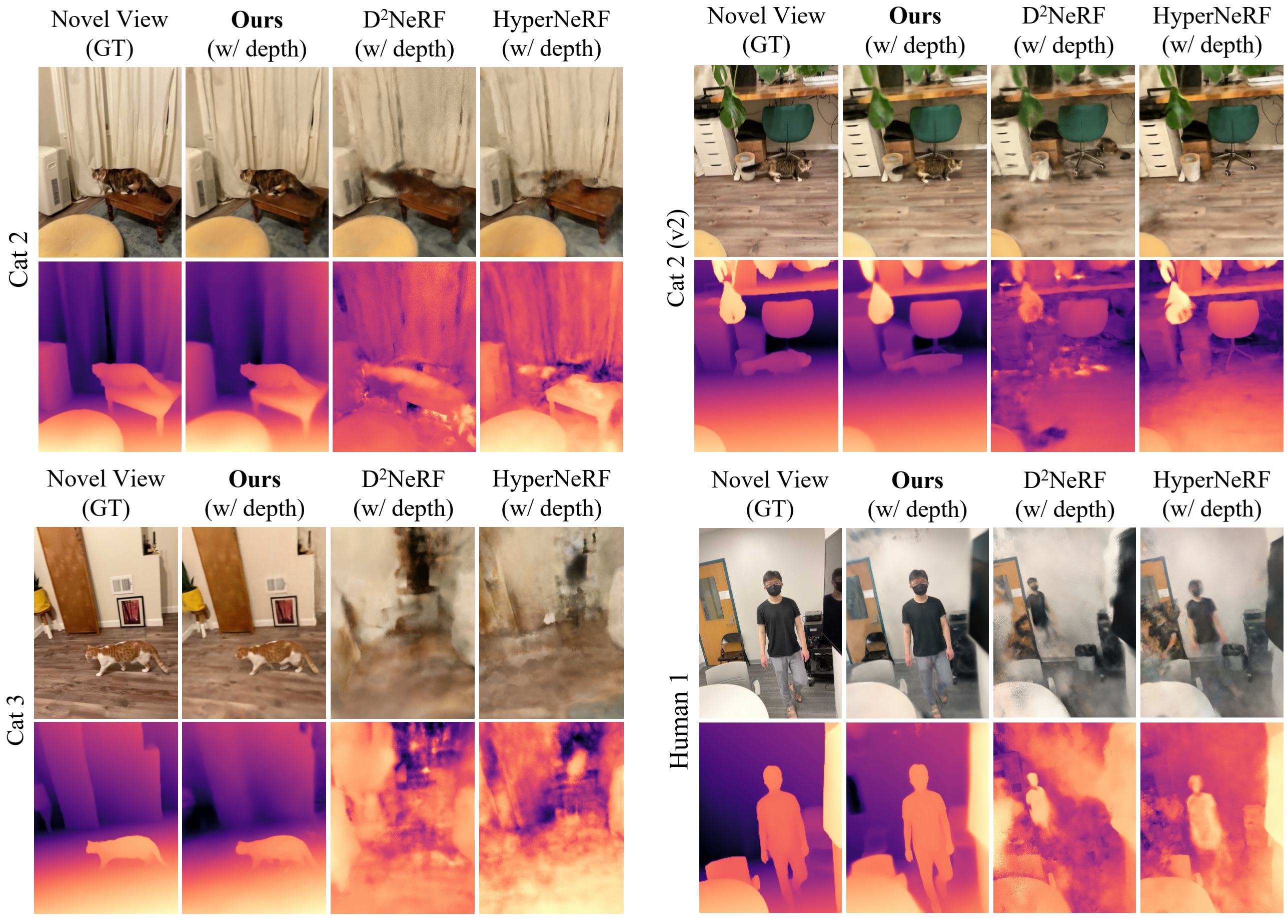

Figure 5: **[Continued] Qualitative Comparisons on Novel View Synthesis (Additional Sequences).** We compare Total-Recon to depth-supervised variants of HyperNeRF [2] and D^{2}NeRF [4] on the task of stereo-view synthesis (the left camera is used for training and the images are rendered to the right camera). While the baselines are only able to reconstruct the background at best, Total-Recon is able to reconstruct *both* the background and the moving deformable object(s), demonstrating holistic scene reconstruction. Full-length videos can be found at `https://andrewsonga.github.io/totalrecon/nvs.html`.

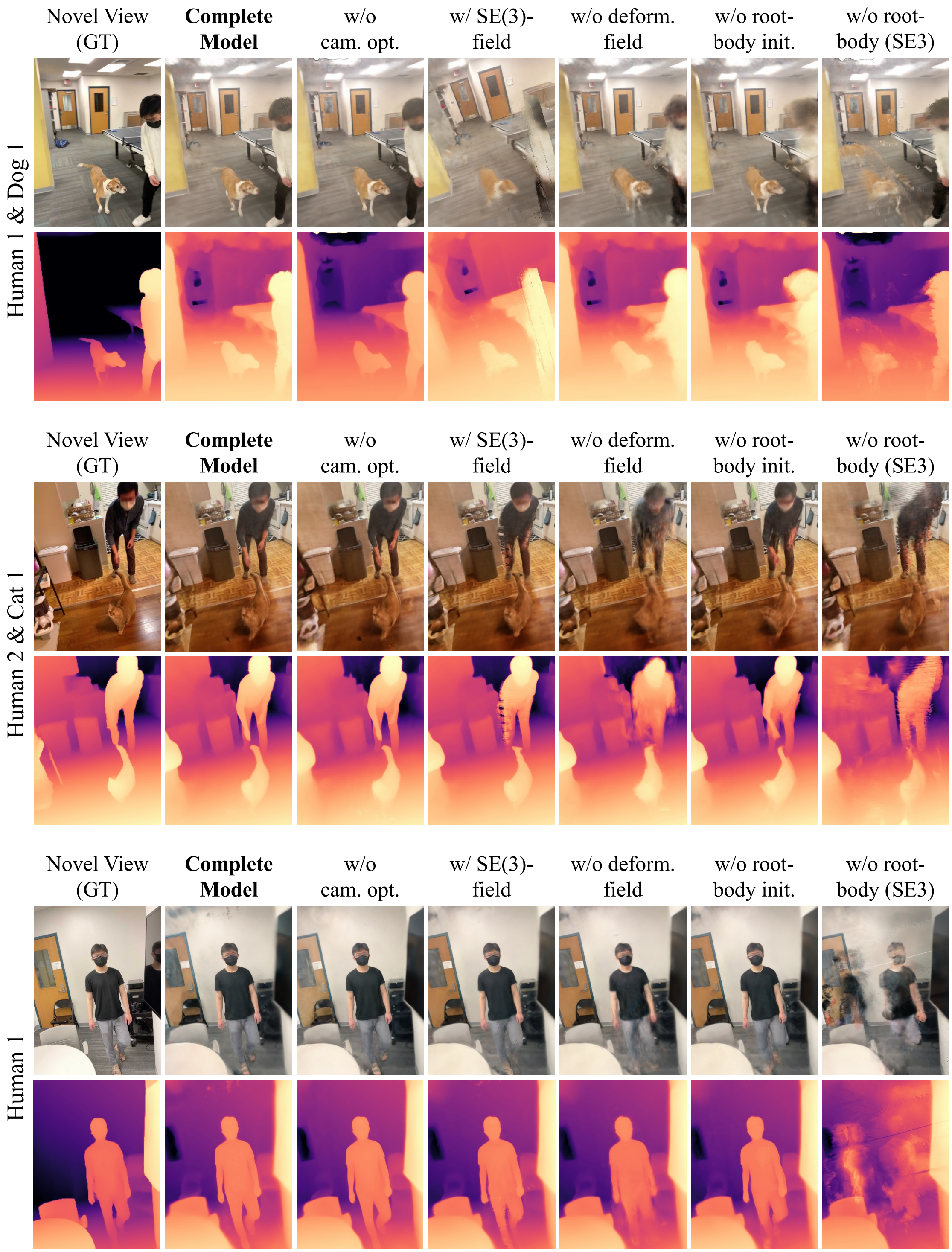

Figure 6: **Ablation Study on Motion Modeling.** We render novel views of the ablations in Table 4. Ablating camera-pose optimization (*w/o cam. opt.*) does not qualitatively change the scene reconstruction. Changing the deformation field from Total-Recon's neural blend skinning function to an SE(3)-field (*w/ SE(3)-field*) results in minor artifacts in the foreground reconstruction. Removing the deformation field entirely (*w/o deform. field*) produces coarse object reconstructions that fail to model moving body parts such as limbs. Removing PoseNet-initialization of object root-body poses (*w/o root-body init.*) results in noisy and sometimes even failed object reconstructions. We omit the ablation without root-body poses (*w/o root-body*) as it does not converge, and instead present a version that does converge (*w/o root-body (SE3)*). However, this ablation also performs significantly worse than previous ablations, as evidenced by the ghosting artifacts indicative of failed foreground reconstruction. These experiments justify Total-Recon's hierarchical motion representation, which explicitly models objects' root-body motion. Full-length videos can be found at https://andrewsonga.github.io/totalrecon/ablation_objmotion.html.

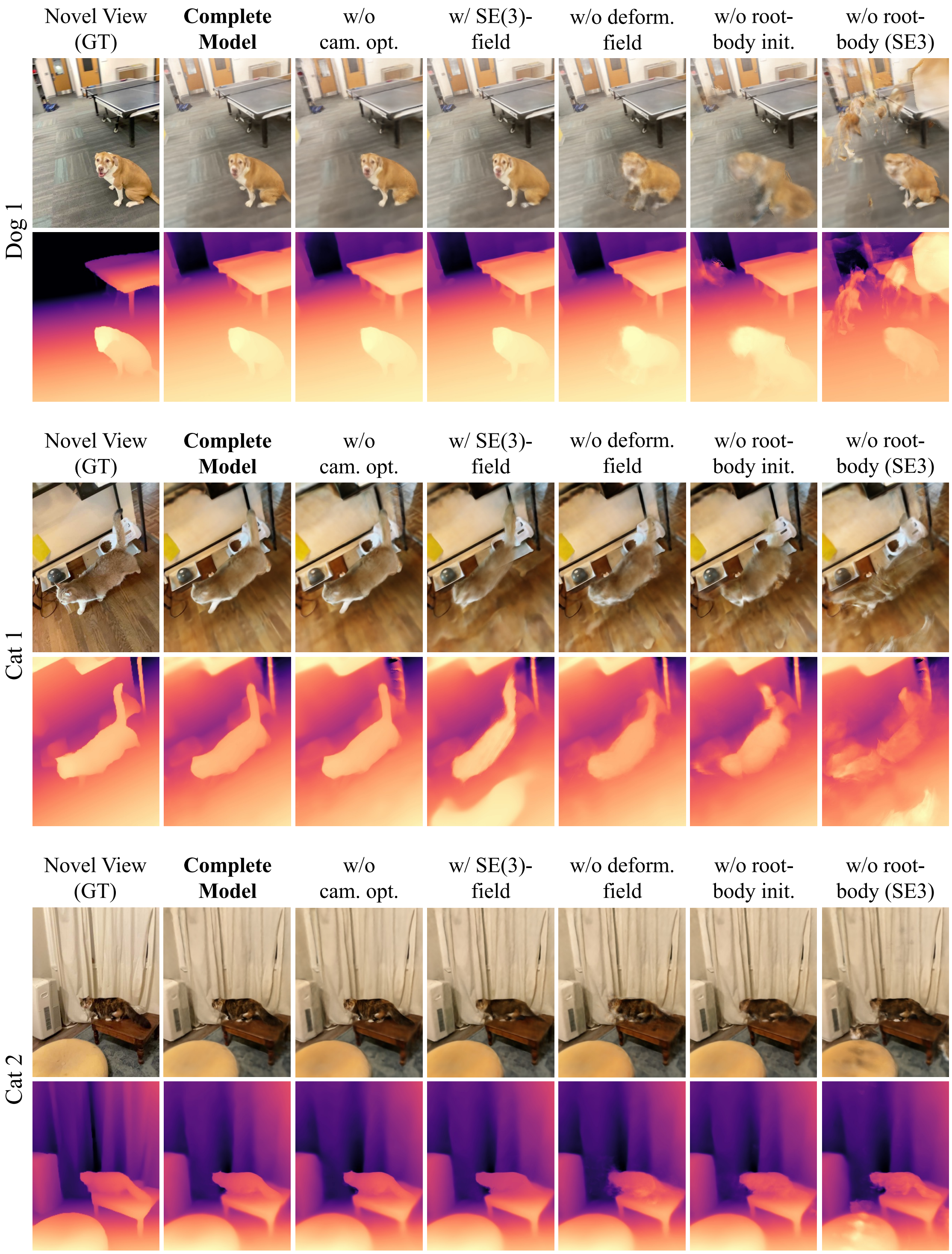

Figure 6: **[Continued] Ablation Study on Motion Modeling.** We render novel views of the ablations in Table 4. Ablating camera-pose optimization (*w/o cam. opt.*) does not qualitatively change the scene reconstruction. Changing the deformation field from Total-Recon's neural blend skinning function to an SE(3)-field (*w/ SE(3)-field*) results in minor artifacts in the foreground reconstruction. Removing the deformation field entirely (*w/o deform. field*) produces coarse object reconstructions that fail to model moving body parts such as limbs. Removing PoseNet-initialization of object root-body poses (*w/o root-body init.*) results in noisy and sometimes even failed object reconstructions. We omit the ablation without root-body poses (*w/o root-body*) as it does not converge, and instead present a version that does converge (*w/o root-body (SE3)*). However, this ablation also performs significantly worse than previous ablations, as evidenced by the ghosting artifacts indicative of failed foreground reconstruction. These experiments justify Total-Recon's hierarchical motion representation, which explicitly models objects' root-body motion. Full-length videos can be found at https://andrewsonga.github.io/totalrecon/ablation_objmotion.html.

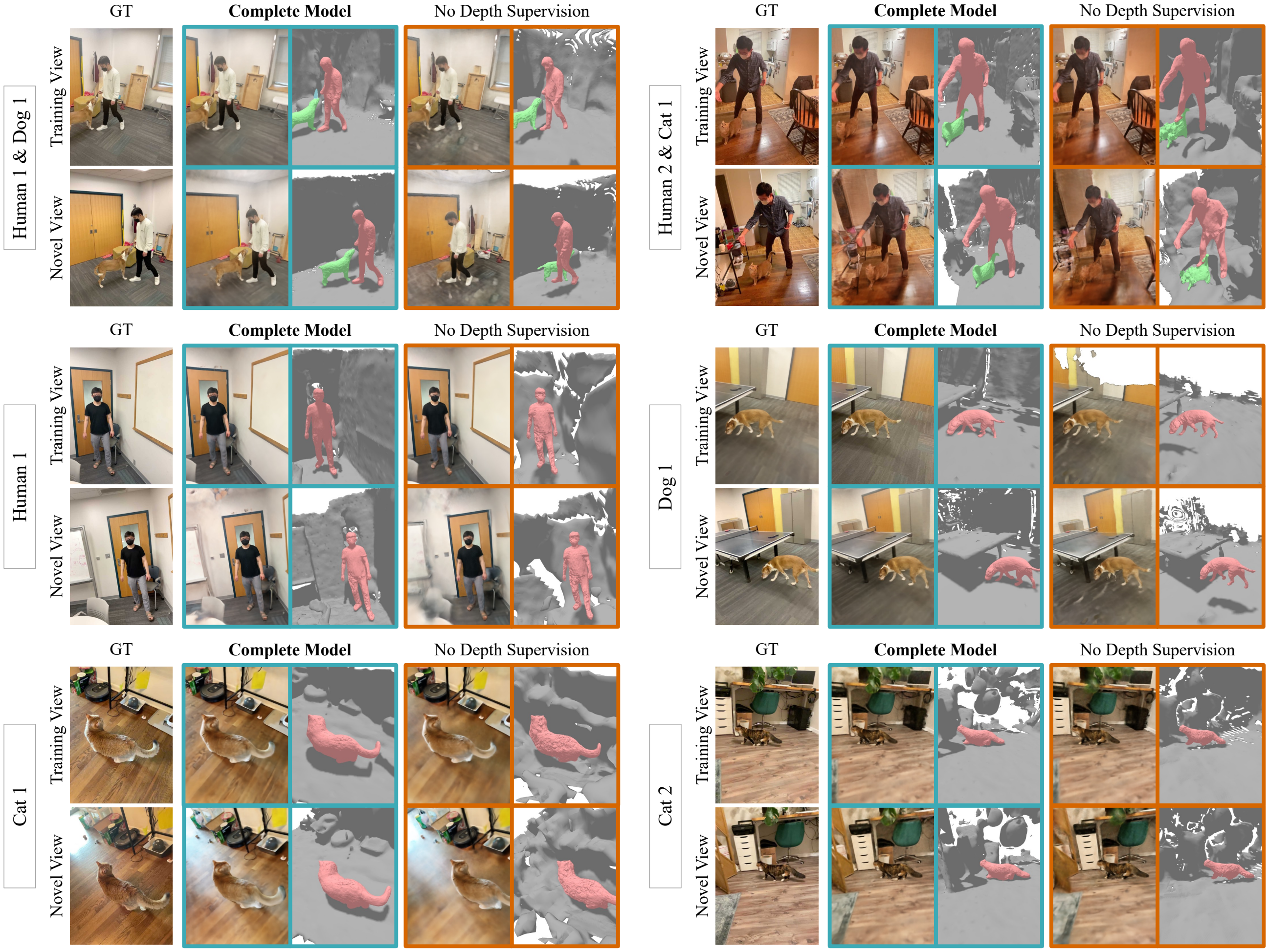

Figure 7: **Ablation Study on Depth Supervision.** While removing depth supervision from Total-Recon (COMPLETE MODEL) doesn't significantly hamper the training-view RGB renderings, it induces the following failure modes in the *novel-view* 3D reconstructions. (a) *Floating objects*: for the HUMAN 1 & DOG 1, DOG 1, HUMAN 1, and CAT 2 sequences, the foreground objects float above the ground, as evidenced by their shadows. (b) *Objects that sink into the background*: for the HUMAN 2 & CAT 1 sequence, the reconstructed cat is halfway sunk into the ground. (c) *Incorrect occlusions*: for the HUMAN 1 & DOG 1 sequence, the human is incorrectly occluding the dog. (d) *Lower reconstruction quality*: for the HUMAN 2 & CAT 1 sequences, we observe that the cat has lower reconstruction quality, and, for all sequences except HUMAN 1 & DOG 1 and CAT 2, we observe that the background object has lower reconstruction quality. Full-length videos can be found at https://andrewsonga.github.io/totalrecon/ablation_depth.html.

References

[1] 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. 4

[2] 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, 7, 8

[3] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[4] 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, 7, 8

[5] 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. 1, 5

[6] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1