Supplemental Material: Neural 6-DoF Tracking and 3D Reconstruction of Unknown Objects

1. Implementation Details

During coarse pose initialization, if there is no immediate previous frame to compare with (*e.g.*, missing detection by the segmentation, or object reappearing after complete occlusion), the current frame will instead be compared with the memory frames. The memory frame which has more than 10 feature correspondences with the current frame is selected as the new reference frame for the coarse pose initialization. The following steps remain the same.

For online pose graph optimization, we constrain the maximum number of participating memory frames $K = 10$ for efficiency. When computing $\mathcal{L}_p$ we reject corresponding points whose distance is larger than 1 cm, or their normal angle is larger than 20°. The Gauss-Newton optimization iterates for 7 steps.

For Neural Object Field learning, we normalize the object into the neural volume bound of $[-1, 1]$, where the scale is computed as 1.5 times of the initial frame's point cloud dimension. The neural volume's coordinate system is based on the first frame's centered point cloud. The geometry network Ω consists of two-layer MLP with hidden dimension 64 and ReLU activation except for the last layer. The intermediate geometric feature $f_{\Omega(\cdot)}$ has dimension 16. The bias of the last layer is initialized to 0.1 for a small positive SDF prediction at the start of training. The appearance network Φ consists of three-layer MLP with hidden dimension 64 and ReLU activation except for the last layer, where we apply sigmoid activation to map the color prediction to $[0, 1]$. For Octree ray-tracing, the finest voxel size is set to 2 cm. We simplify the multi-resolution hash encoder [39] to 4 levels, with number of feature vectors from 16 to 128 for efficiency. Each level's feature dimension is set to 2. The hash table size is set to 2^{22}. In each iteration the ray batch size is 2048. For hierarchical point sampling, N and N' are set to 128 and 64, respectively. The truncation distance λ is set to 1 cm. For *uncertain free space*, ϵ is set to 0.001. In the training loss, $w_u = 100, w_e = 1, w_{surf} = 1000, w_c = 100, w_{eik} = 0.1$. We implement in PyTorch [46] with Adam optimizer. The initial learning rate is 0.01 with linear decay rate 0.1. The Neural Object Field training runs in a separate thread concurrently and interchanges data with the memory pool periodically after each training convergence (300 steps), which leads to sufficient pose refinement. The first training period starts when there are 10 memory frames in the pool. Upon training convergence, it returns the data to the memory pool and grabs memory frames newly added to the pool during its last training period, to repeat the training process. The next training reuses the latest updated frames' poses. But for the other trainable parameters, reusing their weights tend to get stuck in local minima if there is any sub-optimum in the previous training period, particularly due to noisy pose. Therefore, we re-initialize the network weights for the new training periods. This takes similar number of steps to refine the newly added memory frames' poses, compared to reusing the previous network weights.

2. Computation Time

All experiments were conducted on a standard desktop with Intel i9-10980XE CPU and a single NVIDIA RTX 3090 GPU. Our method consists of two threads running concurrently. The online tracking thread processes frames at around 10.2 Hz, where video segmentation takes 18 ms, coarse matching takes 24 ms, pose graph takes 56 ms on average. Concurrently, the neural object field thread runs in the background and takes 6.7 s averagely for each training round, at the end of which it exchanges data with the main thread. On the same hardware, competitive methods DROID-SLAM [61] and BundleTrack [69] run at 6.1 Hz and 11.2 Hz respectively.

3. Metrics

For evaluation, we decouple the pose estimation and shape reconstruction, so that they can be treated separately. For 6-DoF object pose evaluation, we compute the area under the curve (AUC) percentage of *ADD* and *ADD-S* metric:

$$\text{ADD} = \frac{1}{|\mathcal{M}|} \sum_{x \in \mathcal{M}} \left\| (Rx + t) - (\tilde{R}x + \tilde{t}) \right\|_2 \tag{1}$$

$$\text{ADD-S} = \frac{1}{|\mathcal{M}|} \sum_{x_1 \in \mathcal{M}} \min_{x_2 \in \mathcal{M}} \left\| (Rx_1 + t) - (\tilde{R}x_2 + \tilde{t}) \right\|_2, \tag{2}$$

where $\mathcal{M}$ is the object model. Since the novel unknown object's CAD model is inaccessible to the methods to define the coordinate system, we use the ground-truth pose in the first image to define the canonical coordinate frame of each video to evaluate the pose.

For 3D shape reconstruction evaluation, we report the results of chamfer distance between the final reconstructed mesh and the ground-truth mesh, using the following symmetric formulation:

$$d_{CD} = \frac{1}{2|\mathcal{M}_1|} \sum_{x_1 \in \mathcal{M}_1} \min_{x_2 \in \mathcal{M}_2} \|x_1 - x_2\|_2 + \quad (3)$$

$$\frac{1}{2|\mathcal{M}_2|} \sum_{x_2 \in \mathcal{M}_2} \min_{x_1 \in \mathcal{M}_1} \|x_1 - x_2\|_2 \quad (4)$$

In our method, the mesh can be extracted by applying Marching Cubes over the zero level set in the Neural Object Field. For all methods, we use the same resolution (5 mm) to sample points for evaluation. Since most videos do not cover the complete surrounding view of the object, we cull the ground-truth mesh faces that are never visible in the video by a rendering test, given by the ground-truth mesh and pose.

4. Detailed Results

Recall curves for ADD-S and ADD for all three datasets are presented in Fig. 1 (HO3D), Fig. 2 (YCBInEOAT), and Fig. 3 (BEHAVE). Each plot shows the results for all videos of the respective dataset. As can be seen, the area-under-the-curve (AUC) for our method exceeds that of other methods for almost all datasets.

Per-video quantitative results for all three datasets are presented in Tab. 1 (HO3D), Tab. 2 (YCBInEOAT), and Tabs. 3-6 (BEHAVE). As can be seen, our method performs best on almost all videos of HO3D, more than half the videos of YCBInEOAT, and a large majority of videos of BEHAVE. Note that the last row of each table ("Mean") is included in the main paper.

Qualitative results are demonstrated in Figs. 4 and 5 (HO3D), Fig. 6 (YCBInEOAT), and Figs. 7 and 8 (BEHAVE). We encourage the reader to watch the supplemental video.

Details Regarding the Single-View Setup of BEHAVE. As mentioned in the paper, the BEHAVE Dataset was captured by a pre-calibrated multi-camera system with four cameras. Since our method only requires a monocular input, for fair evaluation, we run all methods on a single monocular input. That is, for each scene, we input only one of the cameras' captured video to the methods.

Although in theory we could run each method four times, once per video camera, this would be excessively time consuming for the little insight that it might bring. Moreover, since there are only four cameras placed at each corner around the scene, it is often the case that the object is severely occluded by the human in several cameras' views (including at the beginning of the video). Using such cameras would not lead to meaningful results for tracking evaluation, due to the very limited object visibility at initialization.

Instead, we decided to automatically select one of the four cameras from each scene for evaluation. More specifically, we select the video with the least amount of occlusion in each scene over the entire sequence. To do so, we compute the average visibility ratio of the object in each camera's video by comparing the ground-truth object mask against the rendered object mask using the ground-truth information. This is performed offline for all videos before evaluation. The selected single-view video is then used by all methods for evaluation, even though severe occlusions still occur frequently which exhibit challenges, as shown in Fig. 7, 8.

5. Robustness Analysis

In the following we discuss our approach's robustness under various challenges. We encourage the reader to watch our supplemental video for more complete appreciation of the system.

Dearth of Texture or Geometric Cues. In the case of dynamic object-centric setting, dearth of texture or geometric cues frequently occur given by the object itself. For instance, in Fig. 4, large areas on the blue pitcher lack texture, which challenge those methods heavily relying on optical flow (DROID-SLAM [61]), or keypoint matching (BundleTrack [69]), or photometric loss (NICE-SLAM [85]). Additionally, large areas of cylindrical surface also exhibit few geometric cues to leverage and can cause rotational ambiguity to those methods relying on point-to-surface matching (SDF2SDF [53], BundleTrack [69], KinectFusion [43]). In contrast, our method is robust to these challenges due to the synergy of pose graph optimization and Neural Object Field. More examples of such challenges can be found in Fig. 5, 7, 8.

Occlusions. In the dynamic object setting, occlusions include self-occlusions and external occlusions introduced by the interaction agent (*e.g.*, human hand, human body, robotic arm). For instance, in Fig. 5, there are moments when the "meat can" only exhibits a single flat face (2nd column) after extreme rotations, causing severe self-occlusion. In other observations, external occlusion introduced by the human hand (4th column) also challenges the comparison methods. More examples of such challenges can be found in Fig. 4, 7, 8, 6. As can be observed, our method is robust to either case and keeps tracking accurately throughout the video thanks to the memory mechanism, whereas the comparison methods struggle.

Specularity. Due to the object's surface smoothness, material and complex environmental lighting, specularity could happen, introducing challenges for those methods heavily replying on optical flow (DROID-SLAM [61]), keypoint matching (BundleTrack [69]) or photometric loss (NICE-SLAM [85]). As shown in Fig. 4, 5, 7, 6, despite the specularity on metalic or highly smooth surfaces, our method keeps tracking accurately throughout the video, whereas the comparison methods become brittle.

Abrupt Motion and Motion Blur. Fig. 9 illustrates an example of abrupt object motion due to the human freely swinging the box. Aside from challenges for 6-DoF pose tracking under large displacement, it causes motion blur in RGB, leading to additional challenge for keypoint matching and Neural Object Field learning. However, our method has shown robustness under these adverse conditions and even yields more accurate pose than ground-truth.

Noisy Segmentation. Figs. 10 and 11 demonstrate examples of noisy masks (purple) from the video segmentation network, including both false positive and false negative predictions. The false negative segmentation leads to ignorance of the texture-rich areas, intensifying the issue of dearth of texture. The false positive segmentation introduces deformable part from the interaction agent or undesired scene background, causing inconsistency in multi-view. However, our downstream modules are robust to the segmentation noise and maintain accurate tracking.

Noisy Depth. As shown in Fig. 12, in our setting, the noisy depth comes from two sources. First, the consumer-level RGBD camera has observable sensing noise. This is especially the case for BEHAVE [4] and YCBInEOAT [72] Dataset, where the images are captured at a distance from the camera, which challenges depth sensing. Second, due to the noisy segmentation, false positive predictions include undesired background areas in the depth point cloud. In Fig. 12 (left), when naively fusing the per-frame depth point cloud using ground-truth pose, the result is highly cluttered, which implies the noisy depth sensing and segmentation. However, despite such noise, our simultaneous pose tracking and reconstruction produce high quality mesh, as shown on the right.

6. Limitation and Failure Modes

While our method is robust to a variety of challenging conditions, it fails when multiple types of challenges appear together. For instance, in Fig. 13, the occurrence of severe occlusion, segmentation error, dearth of texture and geometric cues together lead to tracking failure. When the object re-appears, the recovered pose is affected by symmetric geometry. Besides, our method requires depth modality which limits its application to certain types of objects where depth sensing fails, such as transparent objects. Finally, our method assumes the object to be rigid. In future work, generalizing to both rigid and non-rigid objects at the same time would be of interest.

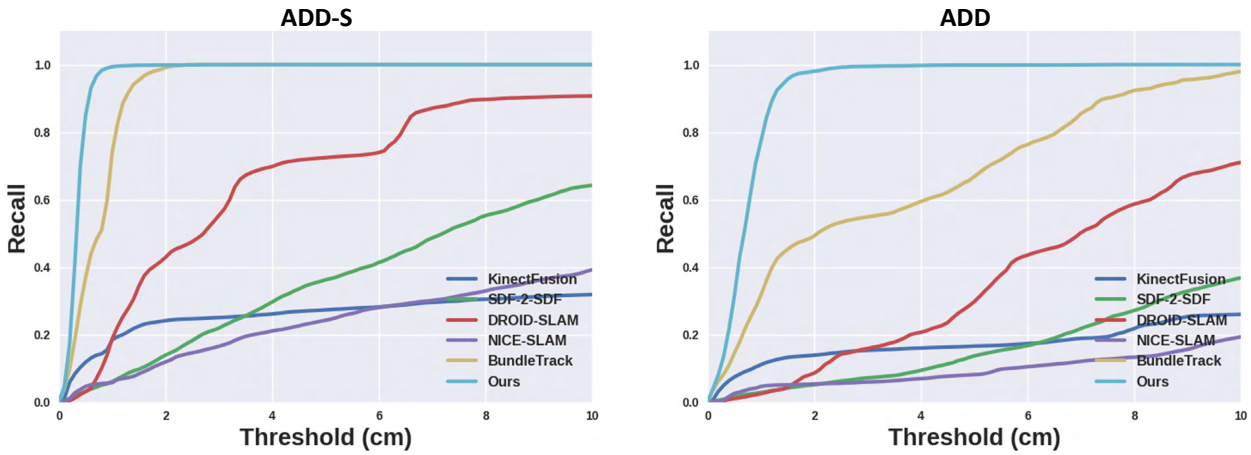

Figure 1. Recall curve of ADD-S (left) and ADD (right) metric including all videos on HO3D Dataset.

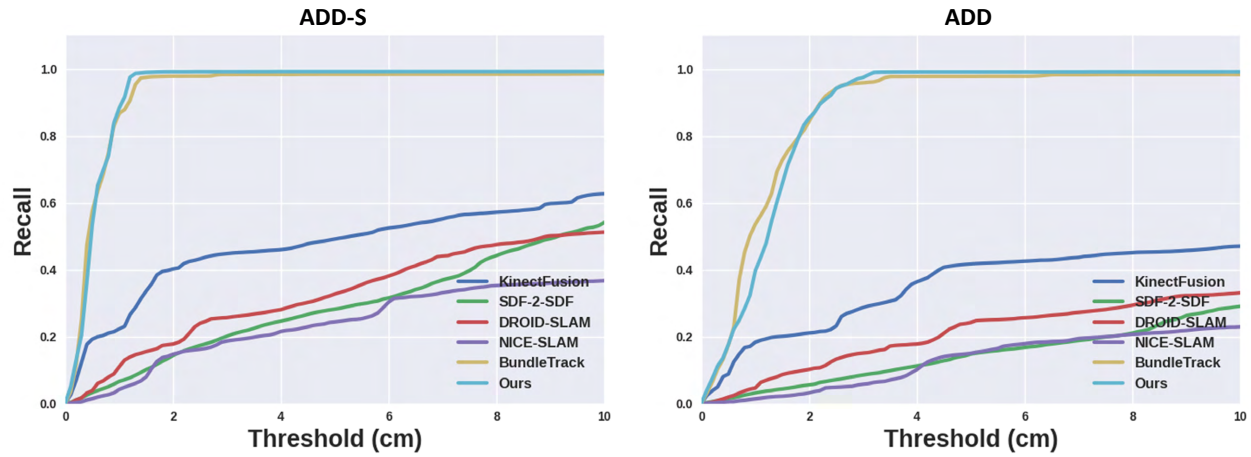

Figure 2. Recall curve of ADD-S (left) and ADD (right) metric including all videos on YCBInEOAT Dataset.

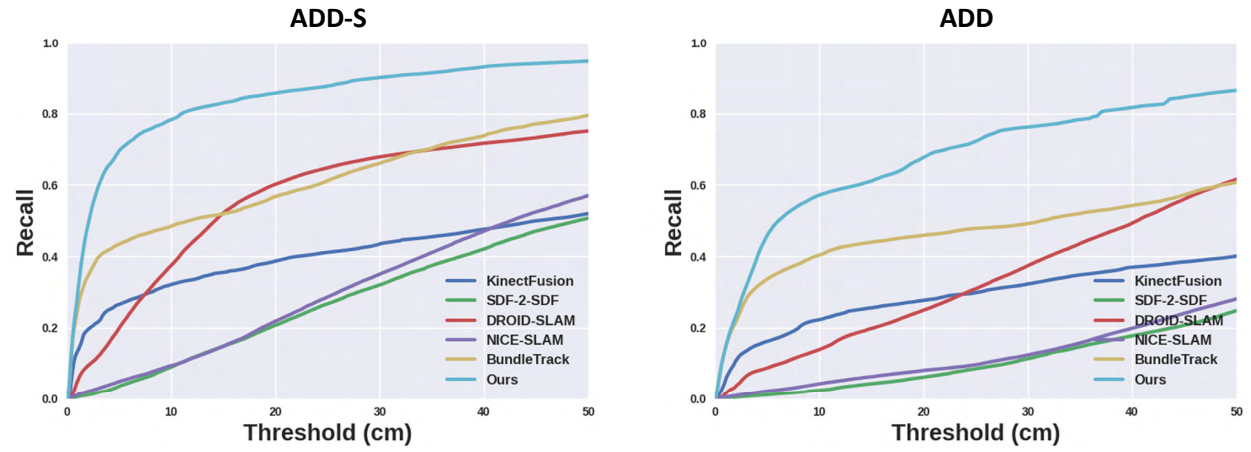

Figure 3. Recall curve of ADD-S (left) and ADD (right) metric including all videos on BEHAVE Dataset.

Video	Metric	DROID-SLAM [61]	BundleTrack [69]	KinectFusion [43]	NICE-SLAM [85]	SDF-2-SDF [53]	Ours
AP10	ADD-S (%) ↑	89.36	91.68	11.39	14.11	33.54	**96.10**
	ADD (%) ↑	50.06	36.60	9.99	2.62	16.35	**91.00**
	CD (cm) ↓	2.48	1.88	4.18	33.22	12.01	**0.47**
AP11	ADD-S (%) ↑	68.76	91.45	76.34	11.40	21.21	**96.18**
	ADD (%) ↑	26.24	41.28	30.99	3.62	7.65	**91.76**
	CD (cm) ↓	120.91	129.18	21.65	90.13	16.79	**0.56**
AP12	ADD-S (%) ↑	38.71	90.79	20.52	19.90	19.48	**97.06**
	ADD (%) ↑	7.15	50.82	9.13	4.11	2.78	**94.76**
	CD (cm) ↓	10.43	2.47	17.18	52.11	8.66	**0.59**
AP13	ADD-S (%) ↑	91.68	90.68	11.40	32.45	49.90	**96.16**
	ADD (%) ↑	73.67	49.03	9.46	6.11	18.16	**92.73**
	CD (cm) ↓	3.00	2.77	19.76	37.62	12.22	**0.63**
AP14	ADD-S (%) ↑	35.53	**96.02**	18.43	5.98	45.56	96.01
	ADD (%) ↑	0.06	90.30	15.81	0.34	32.54	**91.25**
	CD (cm) ↓	71.68	72.40	20.92	31.91	4.38	**1.28**
MPM10	ADD-S (%) ↑	0.33	94.94	12.82	29.20	41.85	**95.05**
	ADD (%) ↑	0.27	87.45	9.37	7.17	15.23	**88.92**
	CD (cm) ↓	1.38	0.97	16.81	54.71	5.86	**0.56**
MPM11	ADD-S (%) ↑	59.68	89.94	13.10	5.34	13.06	**96.20**
	ADD (%) ↑	20.32	53.20	9.74	3.55	6.15	**91.51**
	CD (cm) ↓	87.41	88.97	15.72	66.32	6.82	**0.49**
MPM12	ADD-S (%) ↑	84.43	95.66	12.59	3.99	26.08	**96.98**
	ADD (%) ↑	53.29	90.96	6.70	0.35	8.48	**93.13**
	CD (cm) ↓	1.70	121.33	15.92	51.38	10.24	**0.46**
MPM13	ADD-S (%) ↑	75.30	89.42	10.58	14.34	40.95	**95.80**
	ADD (%) ↑	22.61	38.78	7.27	6.67	9.49	**90.62**
	CD (cm) ↓	3.27	81.39	18.41	72.50	6.05	**0.57**
MPM14	ADD-S (%) ↑	73.46	95.49	26.70	76.36	46.19	**97.33**
	ADD (%) ↑	26.12	90.16	11.05	26.94	20.57	**94.52**
	CD (cm) ↓	6.50	94.99	12.52	52.84	6.18	**0.47**
SB11	ADD-S (%) ↑	63.39	94.44	58.72	30.06	9.67	**97.27**
	ADD (%) ↑	32.15	84.64	55.25	23.72	5.93	**94.39**
	CD (cm) ↓	84.72	75.83	3.01	81.73	20.19	**0.46**
SB13	ADD-S (%) ↑	91.88	95.66	32.15	36.05	47.73	**97.67**
	ADD (%) ↑	76.44	85.47	30.89	26.74	32.50	**95.24**
	CD (cm) ↓	3.15	2.49	21.39	32.91	9.60	**0.47**
SM1	ADD-S (%) ↑	67.86	84.94	30.88	10.65	71.19	**96.90**
	ADD (%) ↑	45.25	59.41	9.41	4.64	33.19	**94.24**
	CD (cm) ↓	4.21	2.04	13.95	26.05	6.39	**0.44**
Mean	ADD-S (%) ↑	64.64	92.39	25.81	22.29	35.88	**96.52**
	ADD (%) ↑	33.36	66.01	16.54	8.97	16.08	**92.62**
	CD (cm) ↓	30.84	52.05	15.49	52.57	9.65	**0.57**

Table 1. Per-video comparison on HO3D Dataset. ADD and ADD-S are AUC (0 to 0.1 m) percentage for pose evaluation. CD is the chamfer distance for shape reconstruction evaluation.

Object	Metric	MaskFusion* [50]	TEASER++* [78]	BundleTrack* [69]	BundleTrack [69]	DROID-SLAM [61]	KinectFusion [43]	NICE-SLAM [85]	SDF-2-SDF [53]	Ours
003_cracker_box	ADD-S (%) ↑	88.28	81.35	89.41	90.20	27.25	56.04	54.23	19.89	**90.63**
	ADD (%) ↑	79.74	63.24	85.07	85.08	19.73	42.73	24.92	12.13	**85.37**
	CD (cm) ↓	-	-	-	1.36	2.95	2.43	4.03	3.12	**0.76**
021_bleach_cleanser	ADD-S (%) ↑	43.31	82.45	94.72	**95.22**	27.13	53.98	17.96	30.63	94.28
	ADD (%) ↑	29.83	61.83	89.34	**89.34**	12.83	40.94	9.55	14.21	87.46
	CD (cm) ↓	-	-	-	1.31	2.43	1.99	9.40	3.87	**0.53**
004_sugar_box	ADD-S (%) ↑	45.62	81.42	90.22	90.68	53.87	45.20	14.56	24.57	**93.81**
	ADD (%) ↑	36.18	51.91	85.56	85.49	43.38	30.53	8.70	14.87	**88.62**
	CD (cm) ↓	-	-	-	2.25	2.41	2.56	7.75	1.70	**0.46**
005_tomato_soup_can	ADD-S (%) ↑	6.45	71.61	95.13	**95.24**	0.08	60.52	17.08	24.76	**95.24**
	ADD (%) ↑	5.65	41.36	**86.00**	85.78	0.08	45.64	11.45	10.89	83.10
	CD (cm) ↓	-	-	-	7.36	**0.99**	9.30	1.52	1.42	3.57
006_mustard_bottle	ADD-S (%) ↑	13.11	88.53	95.35	**95.84**	42.29	17.88	8.77	44.51	95.75
	ADD (%) ↑	11.55	71.92	**92.26**	92.15	15.10	16.01	7.33	18.23	89.87
	CD (cm) ↓	-	-	-	1.76	2.90	6.88	7.95	2.91	**0.45**
Mean	ADD-S (%) ↑	41.88	81.17	92.53	93.01	32.12	46.39	23.41	28.20	**93.77**
	ADD (%) ↑	35.07	57.91	**87.34**	87.26	20.39	34.68	12.70	14.04	86.95
	CD (cm) ↓	-	-	-	2.81	2.34	4.63	6.13	2.61	**1.16**

Table 2. Per-object comparison (following the same protocol as [69]) on YCBInEOAT Dataset. Results of MaskFusion* [50], TEASER++* [78] and BundleTrack* [69] are copied from the leaderboard in [69]. For BundleTrack, we re-run the algorithm with the same segmentation masks as ours for fair comparison, and we augment with TSDF Fusion [9, 83] for reconstruction evaluation. ADD and ADD-S are AUC (0 to 0.1 m) percentage for pose evaluation. CD is the chamfer distance for shape reconstruction evaluation.

Video	Metric	DROID-SLAM [61]	BundleTrack [69]	KinectFusion [43]	NICE-SLAM [85]	SDF-2-SDF [53]	Ours
Date03_Sub03_boxlarge.2	ADD-S (%) ↑	72.59	52.88	21.09	7.05	24.78	**92.63**
	ADD (%) ↑	21.04	13.00	11.00	3.02	7.97	**86.72**
	CD (cm) ↓	8.61	11.61	8.80	24.79	41.97	**1.46**
Date03_Sub03_boxlong.3	ADD-S (%) ↑	44.05	27.77	5.59	10.21	54.87	**77.0**
	ADD (%) ↑	14.06	20.31	1.83	1.58	13.75	**32.58**
	CD (cm) ↓	4.88	**1.61**	11.55	49.75	26.47	3.05
Date03_Sub03_boxmedium.2	ADD-S (%) ↑	75.98	86.25	11.84	12.60	5.86	**92.57**
	ADD (%) ↑	39.16	50.04	4.26	3.11	3.10	**85.24**
	CD (cm) ↓	14.49	3.28	3.23	49.73	44.36	**1.25**
Date03_Sub03_boxsmall.3	ADD-S (%) ↑	8.50	36.32	5.60	4.64	0.84	**70.83**
	ADD (%) ↑	5.40	20.93	4.09	2.84	0.78	**51.64**
	CD (cm) ↓	3.92	11.73	10.93	36.46	42.29	**2.92**
Date03_Sub03_boxtiny.3	ADD-S (%) ↑	41.70	52.40	9.94	6.80	19.97	**88.01**
	ADD (%) ↑	22.44	39.34	7.64	4.35	6.49	**74.24**
	CD (cm) ↓	3.27	13.31	14.64	46.00	26.47	**2.08**
Date03_Sub03_chairblack_hand.3	ADD-S (%) ↑	67.82	86.19	45.88	26.62	31.81	**95.52**
	ADD (%) ↑	10.89	70.83	6.86	3.26	0.66	**90.28**
	CD (cm) ↓	15.82	11.35	12.08	12.89	32.25	**2.4**
Date03_Sub03_chairblack_lift.1	ADD-S (%) ↑	32.98	27.85	21.15	**59.55**	14.29	28.03
	ADD (%) ↑	11.96	**15.61**	9.90	10.70	4.68	11.43
	CD (cm) ↓	51.95	20.86	8.15	31.12	34.36	**6.46**
Date03_Sub03_chairblack_sit.3	ADD-S (%) ↑	97.25	**98.87**	95.32	32.52	43.08	98.81
	ADD (%) ↑	94.49	**98.06**	91.39	13.08	16.75	97.95
	CD (cm) ↓	4.71	4.57	**3.48**	32.12	26.66	4.63
Date03_Sub03_chairblack_sitstand.3	ADD-S (%) ↑	92.75	**98.64**	97.26	78.19	34.57	98.56
	ADD (%) ↑	86.31	**97.73**	95.24	55.40	14.49	97.64
	CD (cm) ↓	4.25	2.49	**2.1**	27.56	37.39	4.75
Date03_Sub03_chairwood_hand.3	ADD-S (%) ↑	72.52	**98.24**	86.28	39.39	36.43	97.80
	ADD (%) ↑	34.60	**96.03**	56.04	9.64	10.83	94.75
	CD (cm) ↓	10.65	7.97	8.22	30.33	47.06	**0.92**
Date03_Sub03_chairwood_lift.3	ADD-S (%) ↑	60.58	60.24	7.83	19.90	11.50	**81.39**
	ADD (%) ↑	19.99	35.33	4.61	2.38	2.73	**48.19**
	CD (cm) ↓	13.52	10.30	8.09	44.33	49.34	**6.3**
Date03_Sub03_chairwood_sit.2	ADD-S (%) ↑	74.28	99.27	93.88	84.39	9.81	**99.31**
	ADD (%) ↑	52.26	98.92	85.09	68.02	8.78	**99.01**
	CD (cm) ↓	17.37	4.47	5.12	49.36	40.59	**3.87**
Date03_Sub03_monitor_move.1	ADD-S (%) ↑	9.14	32.37	15.03	**63.27**	30.76	51.38
	ADD (%) ↑	8.43	13.24	9.25	**32.85**	7.59	24.54
	CD (cm) ↓	2.83	19.83	**2.46**	40.97	28.32	3.09
Date03_Sub03_plasticcontainer.2	ADD-S (%) ↑	55.25	61.63	16.48	23.24	4.57	**84.65**
	ADD (%) ↑	13.84	44.28	8.37	4.15	2.14	**58.15**
	CD (cm) ↓	12.81	8.70	26.06	27.79	41.60	**5.62**
Date03_Sub03_stool_lift.2	ADD-S (%) ↑	73.65	18.15	19.38	23.13	8.30	**94.42**
	ADD (%) ↑	37.80	15.43	9.64	6.05	2.64	**82.53**
	CD (cm) ↓	10.56	26.86	5.73	45.05	50.46	**1.37**
Date03_Sub03_stool_sit.2	ADD-S (%) ↑	85.03	**98.68**	97.88	26.56	5.44	98.67
	ADD (%) ↑	69.71	96.64	91.98	16.93	4.03	**96.65**
	CD (cm) ↓	5.66	3.13	**1.54**	36.35	46.52	1.68
Date03_Sub03_suitcase_lift.0	ADD-S (%) ↑	69.41	81.78	16.46	35.64	49.11	**90.27**
	ADD (%) ↑	24.89	52.97	9.47	8.74	14.43	**76.77**
	CD (cm) ↓	10.22	6.95	14.05	13.97	33.97	**2.3**
Date03_Sub03_suitcase_move.0	ADD-S (%) ↑	71.95	35.00	22.58	37.79	77.35	**94.41**
	ADD (%) ↑	41.66	17.16	9.59	9.76	41.32	**79.25**
	CD (cm) ↓	9.34	17.34	3.35	27.54	26.47	**1.32**
Date03_Sub03_tablesmall_lean.3	ADD-S (%) ↑	52.26	**98.72**	93.40	50.70	32.52	98.55
	ADD (%) ↑	44.13	**96.52**	80.17	18.00	18.45	95.37
	CD (cm) ↓	**6.01**	8.13	8.50	37.76	32.43	14.38
Date03_Sub03_tablesmall_lift.2	ADD-S (%) ↑	46.86	48.88	12.70	45.02	15.03	**70.67**
	ADD (%) ↑	23.23	26.54	10.25	15.97	7.92	**44.03**
	CD (cm) ↓	11.10	44.79	10.56	40.56	46.03	**7.03**
Date03_Sub03_tablesmall_move.3	ADD-S (%) ↑	48.65	94.12	93.57	37.25	1.66	**98.31**
	ADD (%) ↑	28.78	84.67	75.58	12.00	1.64	**95.16**
	CD (cm) ↓	11.34	22.70	8.37	29.06	26.47	**5.22**
Date03_Sub03_tablesquare_lift.1	ADD-S (%) ↑	85.52	96.58	10.33	5.05	3.30	**97.02**
	ADD (%) ↑	50.60	91.95	4.79	1.52	2.25	**92.9**
	CD (cm) ↓	7.14	2.15	30.80	44.14	36.26	**0.68**
Date03_Sub03_tablesquare_move.2	ADD-S (%) ↑	97.09	**99.36**	99.21	15.44	41.38	99.35
	ADD (%) ↑	92.17	**98.98**	98.60	10.71	22.26	98.96
	CD (cm) ↓	4.22	2.86	**2.26**	43.09	50.02	2.31

Table 3. Per-video comparison on BEHAVE Dataset. ADD and ADD-S are AUC (0 to 0.5 m) percentage for pose evaluation. CD is chamfer distance for shape reconstruction evaluation. Table continues on the next page. (This is part 1 of 4.)

Video	Metric	DROID-SLAM [61]	BundleTrack [69]	KinectFusion [43]	NICE-SLAM [85]	SDF-2-SDF [53]	Ours
Date03_Sub03_tablesquare_sit.3	ADD-S (%) ↑	81.23	99.09	98.97	64.13	57.54	**99.1**
	ADD (%) ↑	78.30	98.65	98.26	33.85	35.25	**98.71**
	CD (cm) ↓	3.04	1.49	**1.13**	37.66	36.43	2.22
Date03_Sub03_toolbox.3	ADD-S (%) ↑	0.08	26.69	2.50	5.96	9.01	**92.39**
	ADD (%) ↑	0.08	20.25	1.44	3.53	1.52	**68.97**
	CD (cm) ↓	**1.42**	34.63	22.42	44.52	26.47	1.70
Date03_Sub03_trashbin.1	ADD-S (%) ↑	72.44	30.27	52.37	24.45	5.90	**91.31**
	ADD (%) ↑	48.50	21.79	30.18	11.60	2.07	**73.23**
	CD (cm) ↓	8.67	15.10	14.71	47.01	42.50	**4.62**
Date03_Sub03_yogamat.2	ADD-S (%) ↑	45.99	17.04	17.27	14.54	69.35	**95.8**
	ADD (%) ↑	21.05	12.27	4.61	3.16	21.24	**73.06**
	CD (cm) ↓	9.66	15.32	11.58	57.95	26.47	**0.92**
Date03_Sub04_boxlarge.0	ADD-S (%) ↑	78.77	50.00	11.32	17.14	22.68	**90.81**
	ADD (%) ↑	39.96	44.56	8.91	7.66	6.57	**59.99**
	CD (cm) ↓	9.15	94.26	4.76	25.77	41.14	**2.55**
Date03_Sub04_boxlong.2	ADD-S (%) ↑	**30.54**	24.48	6.40	5.92	7.04	13.53
	ADD (%) ↑	8.48	**13.05**	4.60	2.60	2.49	5.37
	CD (cm) ↓	8.74	76.45	**8.43**	37.69	26.47	24.72
Date03_Sub04_boxmedium.0	ADD-S (%) ↑	5.05	29.29	5.40	14.67	6.06	**92.65**
	ADD (%) ↑	2.50	8.91	2.99	2.69	2.24	**30.34**
	CD (cm) ↓	4.12	69.32	5.83	26.99	26.47	**1.27**
Date03_Sub04_boxsmall.0	ADD-S (%) ↑	0.07	38.07	19.26	18.48	5.40	**88.35**
	ADD (%) ↑	0.07	23.81	11.46	10.55	2.98	**64.11**
	CD (cm) ↓	3.07	48.46	6.40	22.40	48.37	**2.78**
Date03_Sub04_boxtiny.0	ADD-S (%) ↑	1.36	12.90	2.92	5.57	11.97	**42.99**
	ADD (%) ↑	0.81	7.40	2.19	1.76	3.44	**28.52**
	CD (cm) ↓	34.18	68.38	**2.07**	29.79	26.47	3.54
Date03_Sub04_chairblack_hand.1	ADD-S (%) ↑	74.11	93.52	40.70	45.71	19.26	**96.61**
	ADD (%) ↑	20.40	86.55	15.73	10.10	2.03	**93.0**
	CD (cm) ↓	8.91	3.79	15.32	28.98	38.09	**1.35**
Date03_Sub04_chairblack_liftreal.1	ADD-S (%) ↑	47.82	**64.32**	11.18	6.90	1.37	40.10
	ADD (%) ↑	10.85	**20.65**	4.57	1.66	0.36	10.04
	CD (cm) ↓	81.37	17.57	**5.37**	25.04	26.47	7.95
Date03_Sub04_chairblack_sit.1	ADD-S (%) ↑	80.91	90.64	73.12	24.95	38.99	**97.69**
	ADD (%) ↑	56.35	83.45	46.21	11.76	23.92	**95.25**
	CD (cm) ↓	7.04	4.86	9.53	24.96	38.25	**3.61**
Date03_Sub04_chairwood_hand.0	ADD-S (%) ↑	61.54	68.00	4.54	30.96	37.45	**94.38**
	ADD (%) ↑	17.25	33.62	3.33	6.18	1.80	**86.84**
	CD (cm) ↓	12.81	11.76	31.75	27.24	26.47	**1.32**
Date03_Sub04_chairwood_lift.3	ADD-S (%) ↑	**64.87**	29.10	16.22	32.87	16.12	54.47
	ADD (%) ↑	**36.92**	10.57	7.70	9.45	5.79	12.13
	CD (cm) ↓	12.69	11.21	**6.22**	42.00	35.90	19.81
Date03_Sub04_chairwood_sit.1	ADD-S (%) ↑	76.25	**98.15**	71.86	56.97	31.97	98.14
	ADD (%) ↑	32.16	**95.67**	45.56	35.31	9.82	94.83
	CD (cm) ↓	10.16	6.93	13.44	30.31	34.57	**1.04**
Date03_Sub04_monitor_hand.3	ADD-S (%) ↑	98.21	**99.41**	98.81	60.24	12.56	99.38
	ADD (%) ↑	96.86	**99.24**	95.69	23.50	5.32	99.21
	CD (cm) ↓	4.13	4.35	**3.04**	14.61	38.55	3.30
Date03_Sub04_monitor_move.3	ADD-S (%) ↑	6.31	**16.72**	15.52	4.93	4.07	10.83
	ADD (%) ↑	4.62	**8.44**	6.47	4.10	2.31	5.52
	CD (cm) ↓	7.67	16.76	**2.16**	34.00	25.43	4.12
Date03_Sub04_plasticcontainer_lift.2	ADD-S (%) ↑	45.35	40.99	12.05	7.59	12.95	**73.63**
	ADD (%) ↑	12.91	23.34	8.37	2.86	6.86	**36.16**
	CD (cm) ↓	7.08	71.91	6.20	34.26	41.69	**5.76**
Date03_Sub04_stool_move.0	ADD-S (%) ↑	74.77	46.72	30.19	18.13	**76.73**	55.24
	ADD (%) ↑	**48.47**	27.65	21.74	7.14	44.05	31.78
	CD (cm) ↓	7.95	25.46	5.33	45.27	26.47	**1.25**
Date03_Sub04_stool_sit.0	ADD-S (%) ↑	0.51	**98.15**	97.56	41.58	9.88	98.14
	ADD (%) ↑	0.45	**95.57**	83.62	11.90	5.68	95.19
	CD (cm) ↓	4.30	3.67	2.87	28.76	33.65	**2.79**
Date03_Sub04_suitcase_ground.0	ADD-S (%) ↑	59.70	96.59	14.83	18.85	6.36	**96.93**
	ADD (%) ↑	20.56	92.75	12.21	8.12	5.23	**93.61**
	CD (cm) ↓	10.41	1.91	3.18	22.11	37.86	**1.17**
Date03_Sub04_suitcase_lift.2	ADD-S (%) ↑	34.95	31.68	25.40	29.32	11.21	**71.91**
	ADD (%) ↑	18.14	10.65	11.03	11.75	2.95	**64.51**
	CD (cm) ↓	5.53	58.81	8.84	49.20	47.01	**1.91**
Date03_Sub04_tablesmall_hand.0	ADD-S (%) ↑	61.21	29.93	16.53	39.31	21.32	**92.94**
	ADD (%) ↑	37.48	10.46	8.17	8.22	7.03	**85.62**
	CD (cm) ↓	9.09	9.89	24.59	35.72	42.35	**8.45**

Table 4. Per-video comparison on BEHAVE Dataset, continued from previous page. (This is part 2 of 4.)

Video	Metric	DROID-SLAM [61]	BundleTrack [69]	KinectFusion [43]	NICE-SLAM [85]	SDF-2-SDF [53]	Ours
Date03_Sub04_tablesmall_lean.0	ADD-S (%) ↑	78.16	98.44	96.66	17.29	33.80	**98.49**
	ADD (%) ↑	66.19	95.09	87.52	14.88	18.06	**95.34**
	CD (cm) ↓	13.51	8.27	**7.89**	46.25	40.48	9.36
Date03_Sub04_tablesmall_lift.3	ADD-S (%) ↑	**43.3**	18.38	10.62	18.74	37.41	30.33
	ADD (%) ↑	**26.87**	8.81	6.95	7.78	9.85	11.81
	CD (cm) ↓	6.87	12.59	7.99	19.63	38.33	**5.53**
Date03_Sub04_tablesquare_hand.0	ADD-S (%) ↑	93.83	**98.95**	91.30	63.69	33.70	98.82
	ADD (%) ↑	46.35	**97.41**	72.31	19.86	24.62	96.69
	CD (cm) ↓	6.56	3.80	3.81	39.31	38.35	**1.63**
Date03_Sub04_tablesquare_lift.3	ADD-S (%) ↑	75.82	48.09	12.99	49.94	5.41	**96.13**
	ADD (%) ↑	26.25	16.71	7.92	3.48	3.08	**90.62**
	CD (cm) ↓	11.98	8.40	10.71	24.59	43.44	**0.7**
Date03_Sub04_tablesquare_sit.2	ADD-S (%) ↑	93.00	99.18	98.94	63.02	15.54	**99.25**
	ADD (%) ↑	82.80	98.94	97.97	35.62	11.70	**99.07**
	CD (cm) ↓	4.10	**2.27**	3.42	40.13	53.83	2.99
Date03_Sub04_toolbox.3	ADD-S (%) ↑	30.35	15.10	7.02	4.66	54.25	**80.91**
	ADD (%) ↑	17.38	9.44	4.37	3.70	29.63	**58.0**
	CD (cm) ↓	**2.47**	45.67	13.61	30.08	26.47	3.99
Date03_Sub04_trashbin.0	ADD-S (%) ↑	78.62	66.63	34.18	16.89	4.10	**95.62**
	ADD (%) ↑	54.54	34.15	21.41	8.34	3.14	**63.9**
	CD (cm) ↓	6.16	5.33	18.05	50.63	47.54	**1.05**
Date03_Sub04_yogamat.3	ADD-S (%) ↑	25.56	33.14	11.67	15.06	51.85	**85.55**
	ADD (%) ↑	4.67	8.74	6.92	3.53	5.65	**58.87**
	CD (cm) ↓	16.85	18.22	3.58	42.54	26.47	**2.4**
Date03_Sub05_boxlarge.1	ADD-S (%) ↑	66.41	42.28	19.60	9.10	34.96	**94.47**
	ADD (%) ↑	15.43	6.25	2.49	1.86	5.68	**20.02**
	CD (cm) ↓	11.90	15.68	8.48	38.84	32.29	**1.13**
Date03_Sub05_boxlong.3	ADD-S (%) ↑	3.26	35.26	2.70	5.87	16.01	**88.02**
	ADD (%) ↑	0.56	3.40	1.63	0.96	3.29	**59.52**
	CD (cm) ↓	**10.53**	67.09	27.56	38.73	37.49	36.11
Date03_Sub05_boxmedium.2	ADD-S (%) ↑	27.94	20.85	28.36	32.95	7.65	**84.52**
	ADD (%) ↑	18.57	12.96	13.51	15.80	3.60	**47.87**
	CD (cm) ↓	10.12	12.18	5.82	44.70	43.36	**2.51**
Date03_Sub05_boxsmall.3	ADD-S (%) ↑	73.21	4.39	5.12	37.15	77.48	**93.89**
	ADD (%) ↑	38.61	3.95	2.58	10.21	37.97	**75.64**
	CD (cm) ↓	5.23	12.72	2.86	27.07	26.47	**2.0**
Date03_Sub05_boxtiny.3	ADD-S (%) ↑	23.63	9.58	14.49	5.77	1.27	**54.23**
	ADD (%) ↑	12.80	5.32	5.57	2.81	0.87	**40.9**
	CD (cm) ↓	37.93	8.70	**2.3**	49.22	26.47	3.49
Date03_Sub05_chairblack.1	ADD-S (%) ↑	56.78	45.68	32.90	58.88	4.77	**69.13**
	ADD (%) ↑	30.54	39.49	27.99	18.43	2.22	**43.12**
	CD (cm) ↓	18.52	27.39	25.19	23.76	39.51	**8.36**
Date03_Sub05_chairwood.1	ADD-S (%) ↑	69.21	**92.03**	46.51	21.12	13.16	90.43
	ADD (%) ↑	30.01	**79.75**	28.99	11.83	6.28	75.08
	CD (cm) ↓	16.58	**5.14**	13.52	51.69	43.14	7.78
Date03_Sub05_monitor.1	ADD-S (%) ↑	67.18	64.86	73.46	71.04	15.64	**89.05**
	ADD (%) ↑	55.08	46.39	48.30	52.76	8.71	**75.96**
	CD (cm) ↓	7.28	52.23	**4.19**	32.80	31.32	7.37
Date03_Sub05_plasticcontainer.3	ADD-S (%) ↑	51.10	41.66	24.21	23.13	71.54	**76.33**
	ADD (%) ↑	16.60	15.12	9.06	4.40	29.34	**48.3**
	CD (cm) ↓	23.18	24.71	7.18	32.27	26.47	**6.34**
Date03_Sub05_stool.2	ADD-S (%) ↑	80.38	96.42	94.69	40.17	33.24	**98.27**
	ADD (%) ↑	60.87	86.80	75.13	27.59	9.88	**94.41**
	CD (cm) ↓	9.21	6.66	4.55	43.11	45.26	**4.13**
Date03_Sub05_suitcase.2	ADD-S (%) ↑	71.70	81.13	63.07	25.34	30.69	**97.39**
	ADD (%) ↑	30.48	68.68	26.68	7.04	4.30	**94.31**
	CD (cm) ↓	6.27	2.31	6.47	29.06	43.58	**0.96**
Date03_Sub05_tablesmall.1	ADD-S (%) ↑	50.53	56.23	51.31	23.32	59.44	**71.39**
	ADD (%) ↑	34.29	39.52	36.06	9.47	6.41	**55.86**
	CD (cm) ↓	12.63	27.61	**9.87**	56.51	32.95	17.35
Date03_Sub05_tablesquare.2	ADD-S (%) ↑	35.60	96.16	66.70	7.55	35.69	**97.93**
	ADD (%) ↑	23.15	87.43	53.86	5.28	25.12	**94.5**
	CD (cm) ↓	8.73	3.63	29.20	52.47	35.46	**1.27**

Table 5. Per-video comparison on BEHAVE Dataset, continued from previous page. (This is part 3 of 4.)

Video	Metric	DROID-SLAM [61]	BundleTrack [69]	KinectFusion [43]	NICE-SLAM [85]	SDF-2-SDF [53]	Ours
Date03_Sub05_toolbox.1	ADD-S (%) ↑	55.27	24.17	23.30	13.54	52.24	**89.64**
	ADD (%) ↑	36.92	19.30	15.41	6.45	29.98	**71.47**
	CD (cm) ↓	7.80	17.49	4.94	38.37	26.47	**2.94**
Date03_Sub05_trashbin.3	ADD-S (%) ↑	78.78	56.89	24.44	32.29	40.09	**92.2**
	ADD (%) ↑	48.88	16.38	14.24	14.19	6.16	**56.67**
	CD (cm) ↓	7.46	8.89	5.58	36.88	26.47	**2.28**
Date03_Sub05_yogamat.3	ADD-S (%) ↑	62.56	66.92	8.02	25.46	17.43	**96.6**
	ADD (%) ↑	21.54	8.33	3.84	5.52	1.42	**78.41**
	CD (cm) ↓	8.92	12.68	2.99	40.50	26.47	**1.04**
Mean	ADD-S (%) ↑	56.14	59.06	38.37	28.80	25.71	**83.63**
	ADD (%) ↑	32.29	45.03	28.45	11.93	10.05	**67.52**
	CD (cm) ↓	11.24	19.27	9.36	36.03	35.99	**4.66**

Table 6. Per-video comparison on BEHAVE Dataset, continued from previous page. (This is part 4 of 4.)

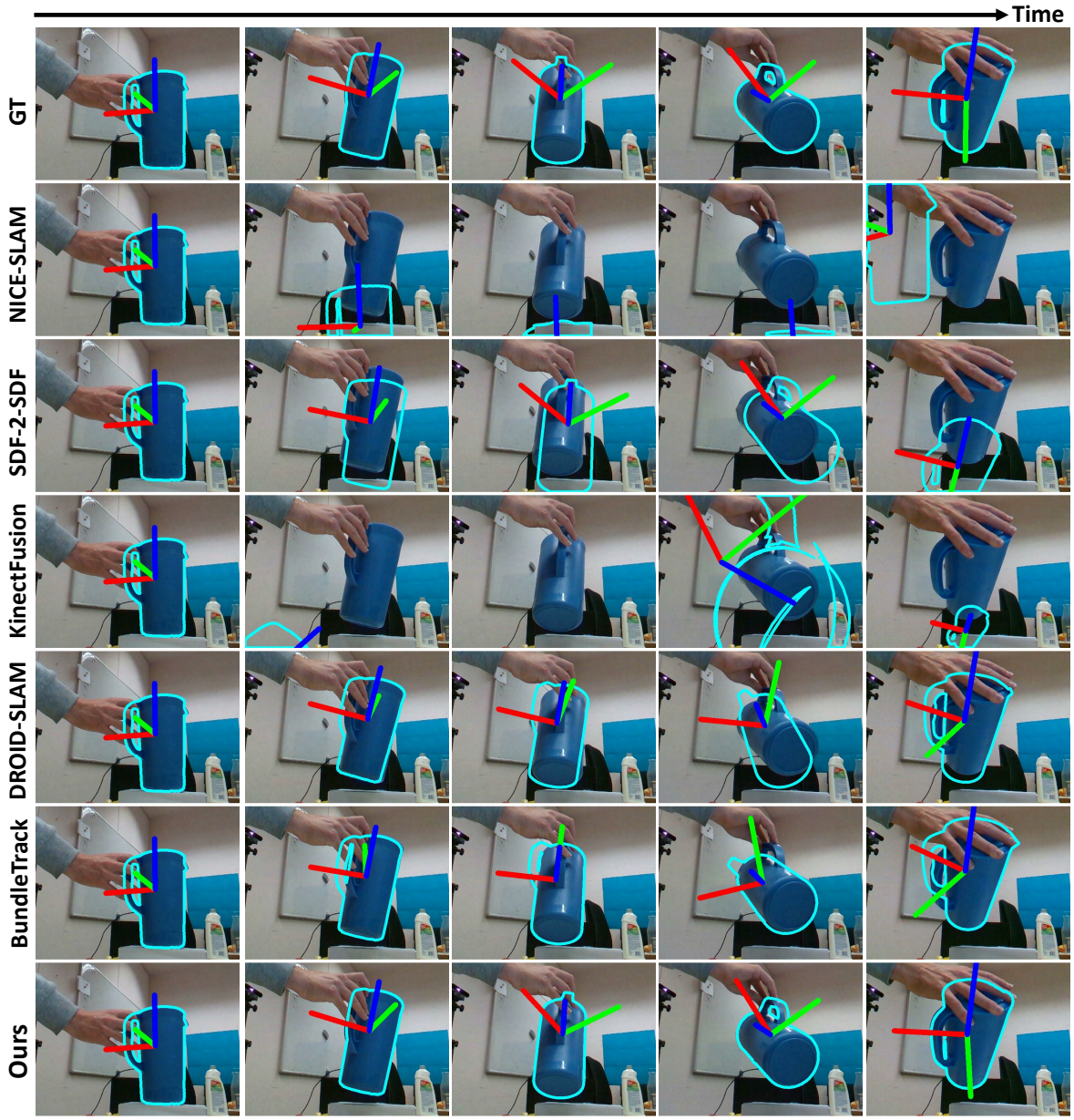

Figure 4. Qualitative comparison on HO3D video "AP13". Our method is robust to observations with little texture or geometric cues (large area of cylindrical surface), whereas comparison methods struggle.

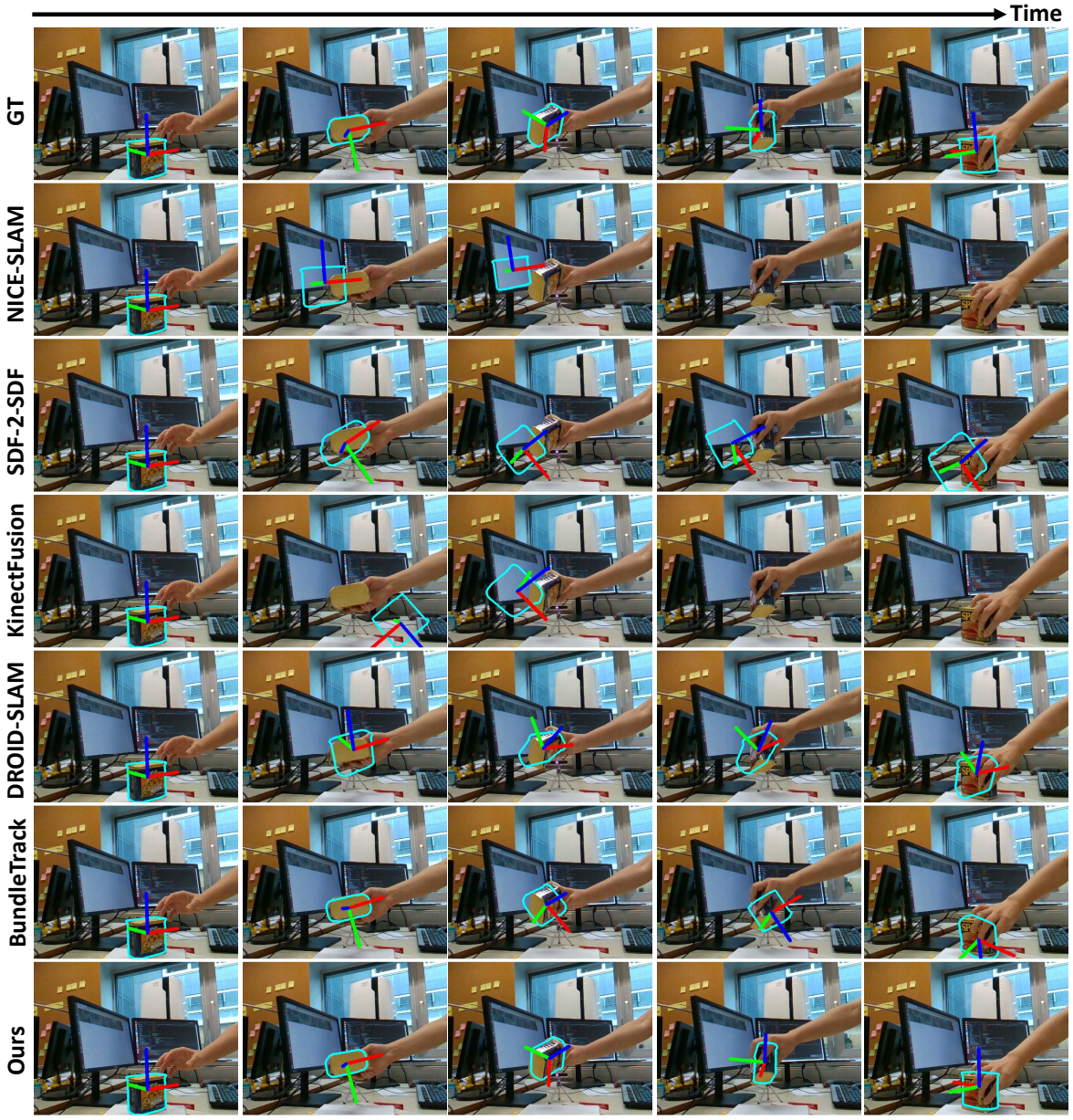

Figure 5. Qualitative comparison on HO3D video "MPM13". Note that our pose tracking at times appears to be slightly more accurate than the ground-truth as shown in the rightmost column.

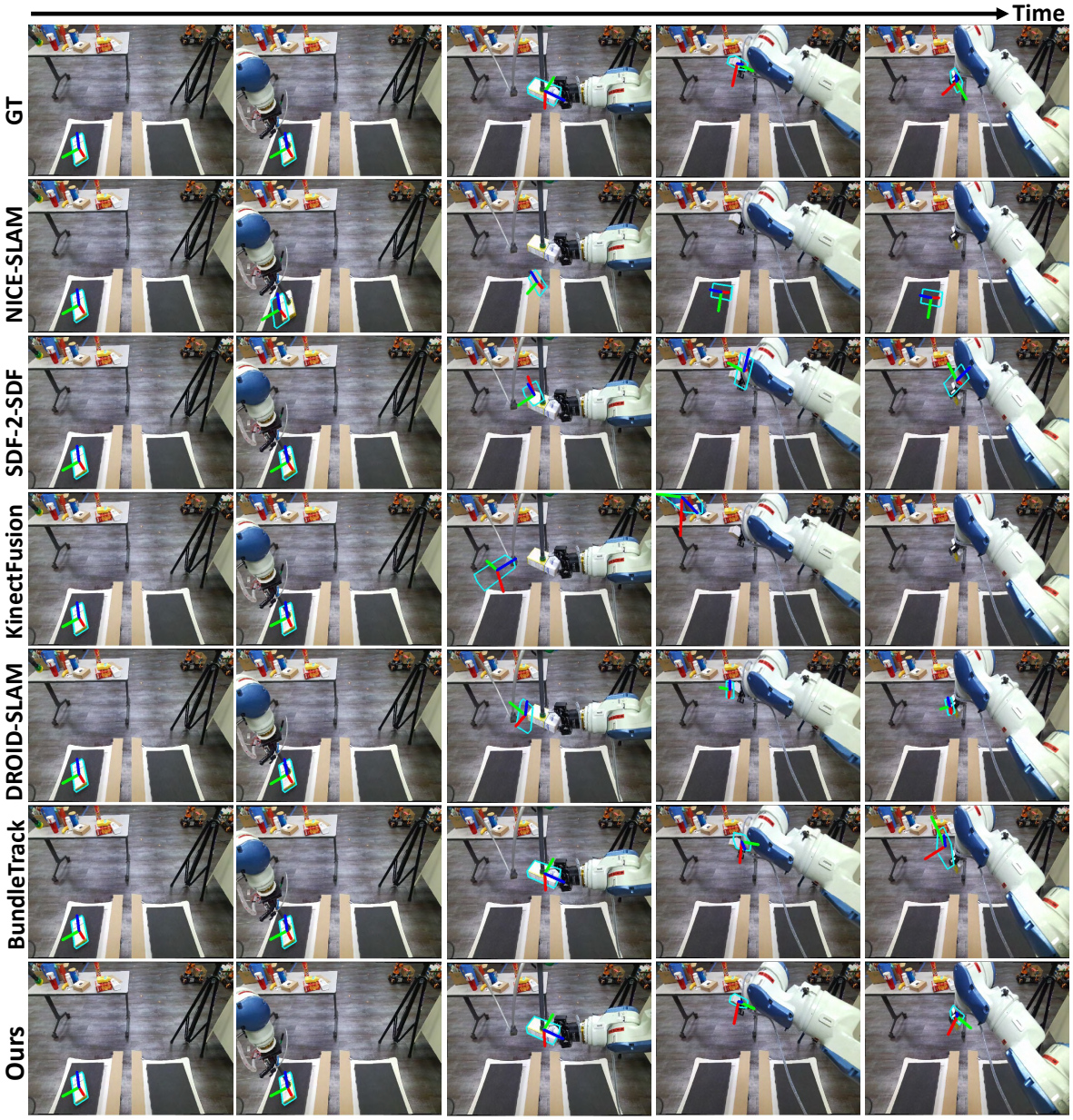

Figure 6. Qualitative comparison on YCBInEOAT video “sugar_box1".

Figure 7. Qualitative comparison on BEHAVE video "Date03_Sub03_chairblack_hand.3". Our method is robust to severe and even complete occlusions (3rd and last column).

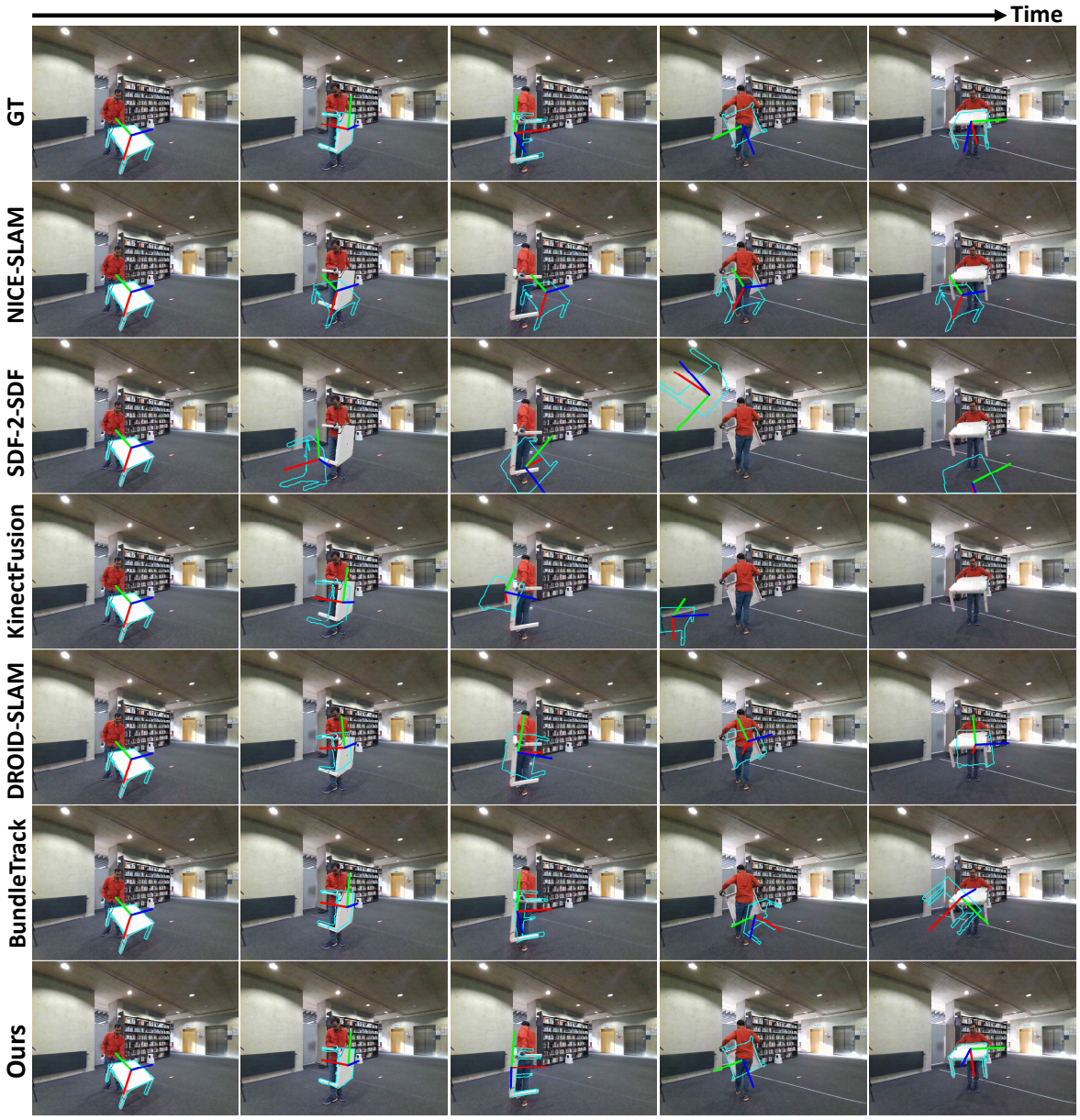

Figure 8. Qualitative comparison on BEHAVE video "Date03_Sub04_tablesquare_lift.3". Our method is sometimes even more accurate than ground-truth (3rd and last column). It is also robust to severe occlusions (4th column).

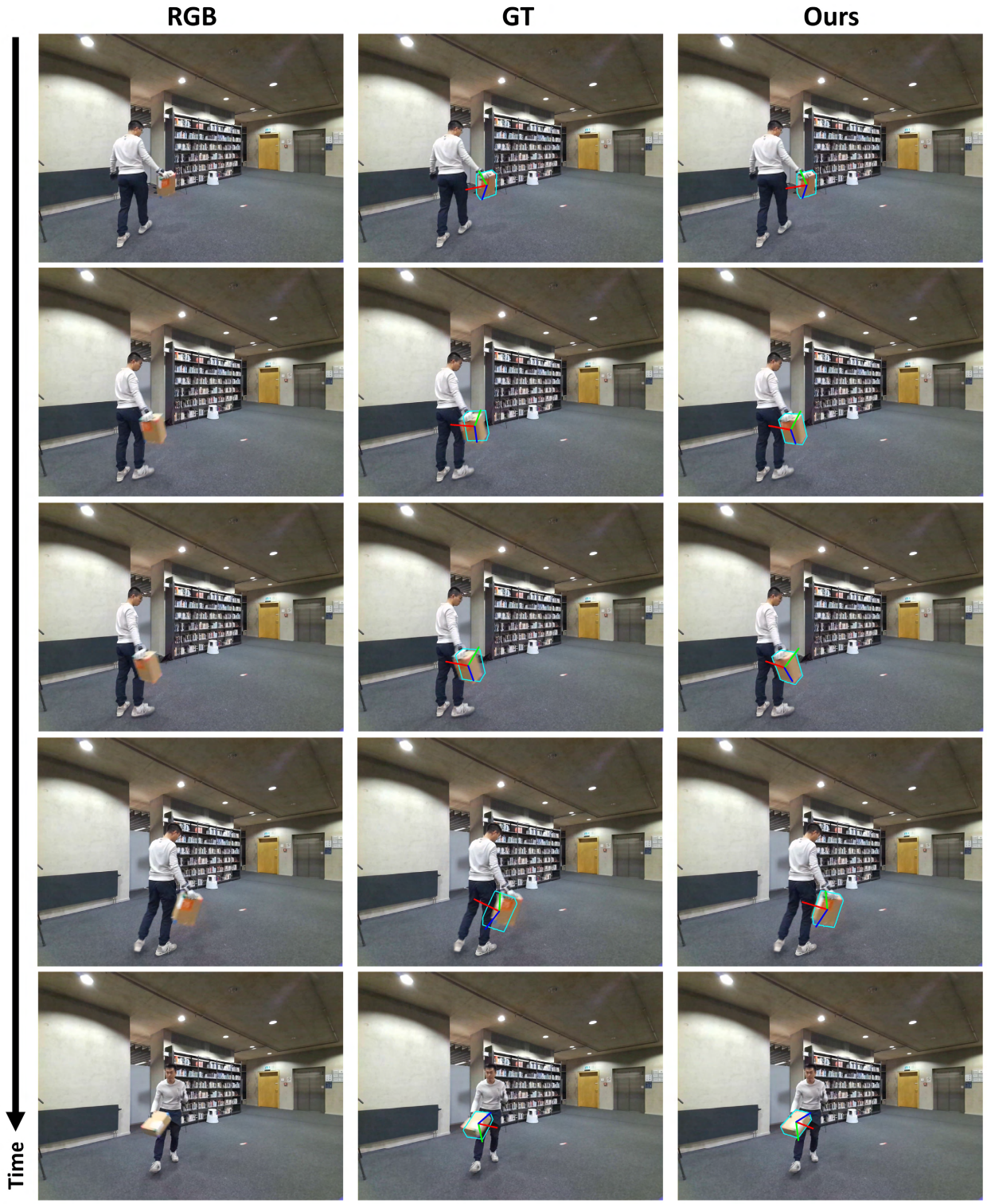

Figure 9. Despite fast object pose change and motion blur, our approach produces even more accurate pose than ground-truth. Image is best viewed by zooming in.

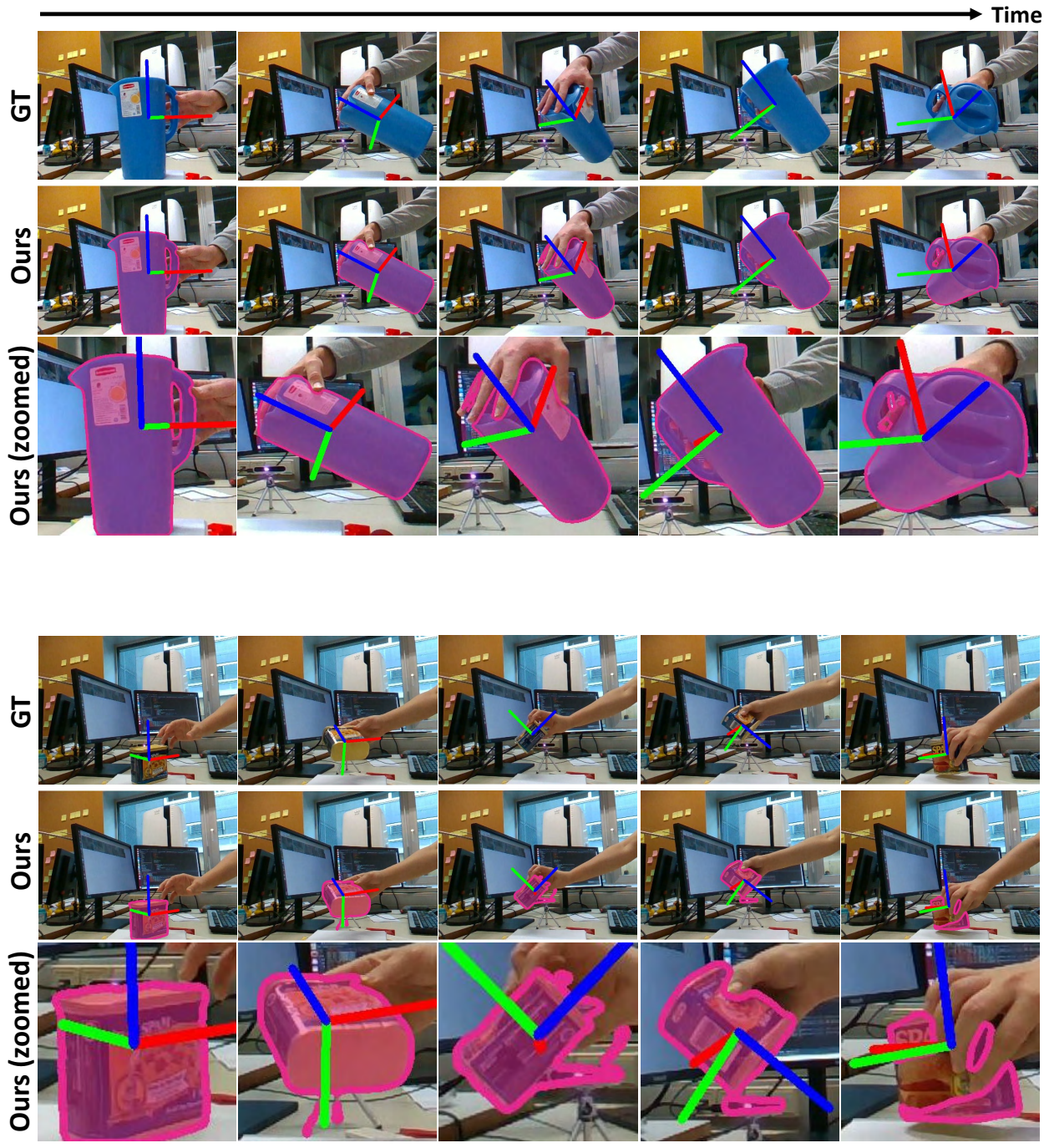

Figure 10. Example of noisy masks (purple) from the video segmentation network, showing both false positive and false negative predictions. The first column visualizes the first frame's mask that initializes tracking. Our method is robust to noisy segmentation and maintains accurate tracking despite such noise. Figure is continued on the next page. (Part 1 of 2.)

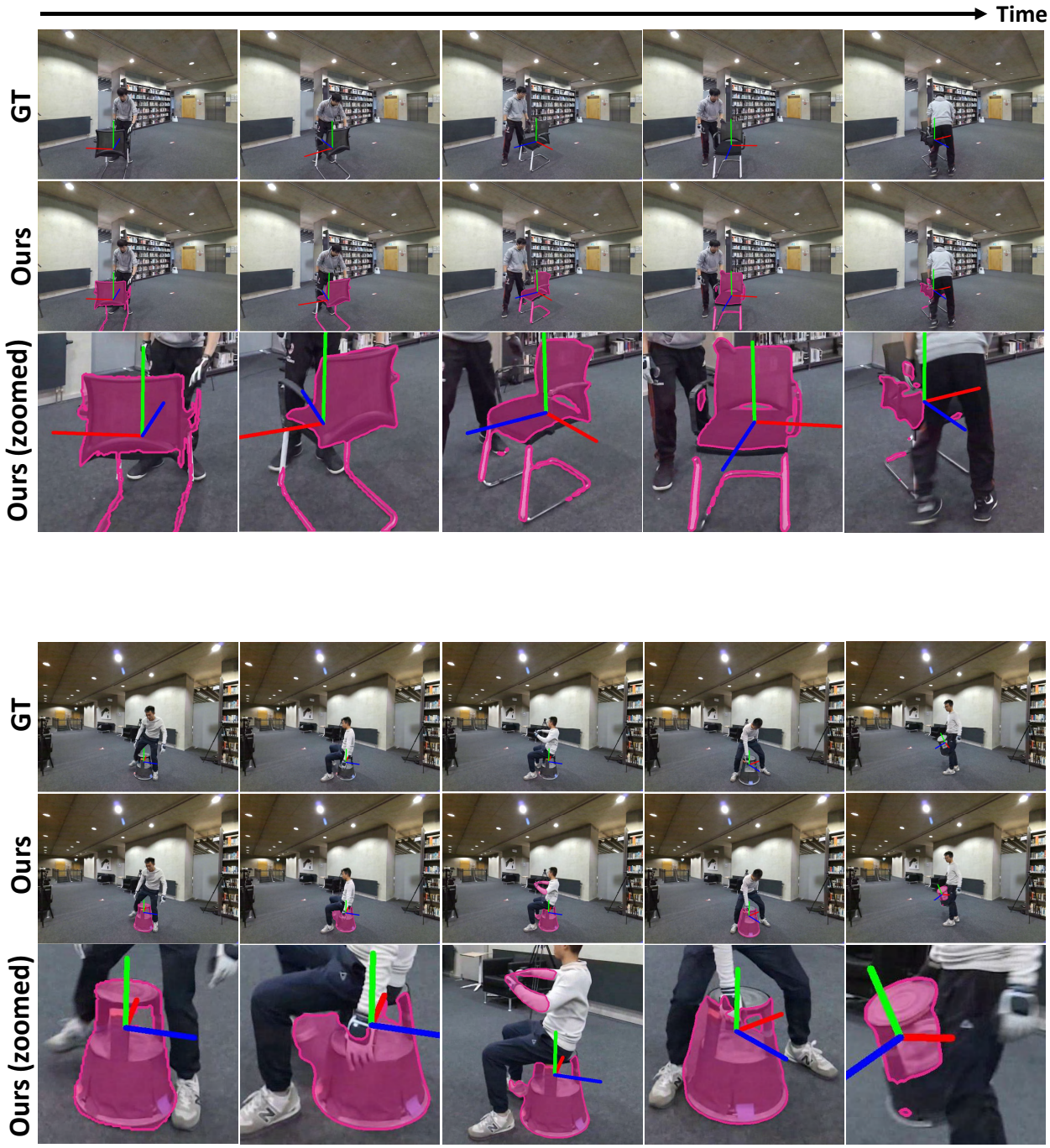

Figure 11. Example of noisy masks (purple) from the video segmentation network. Continued from previous figure. (Part 2 of 2.)

Figure 12. Example of noisy depth from BEHAVE video "Date03_Sub04_tablesquare_lift.3". **Left:** Fused point cloud using ground-truth pose and masks from the video segmentation network. **Right:** Final reconstruction from our approach without any trimming.

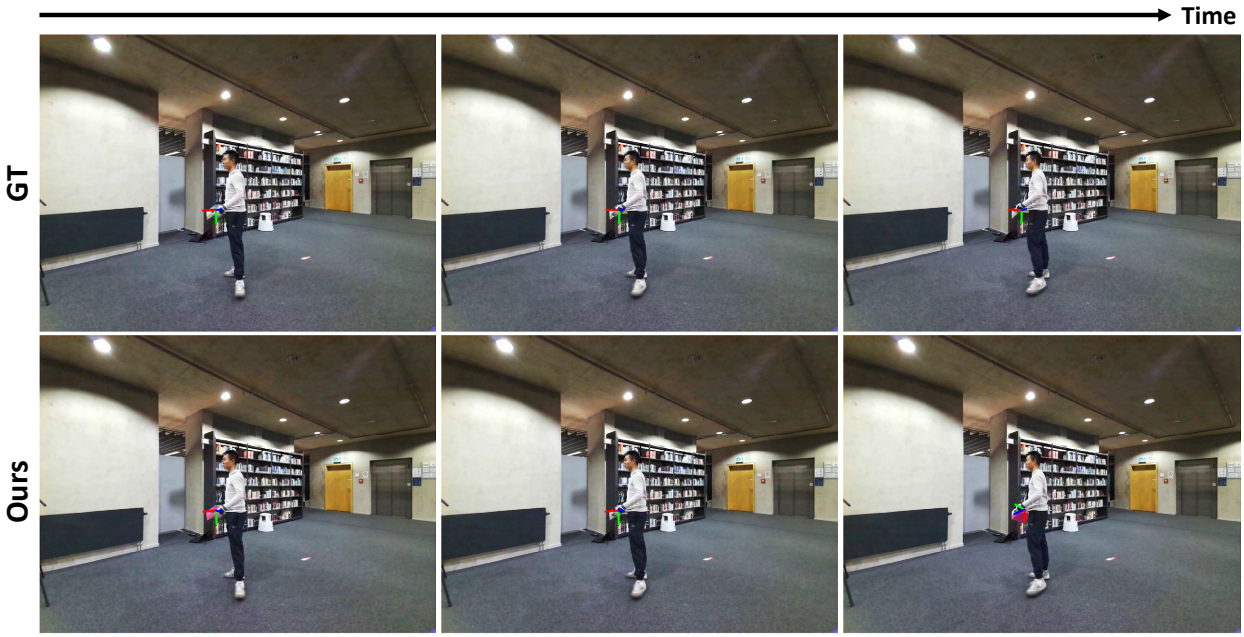

Figure 13. Failure case. The occurrence of severe occlusion, segmentation error, dearth of texture or geometric cues together lead to tracking failure. When the object re-appears, the recovered pose is affected by symmetric geometry.

References

[1] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

[2] K Somani Arun, Thomas S Huang, and Steven D Blostein. Least-squares fitting of two 3-D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 9(5):698–700, 1987.

[3] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6290–6301, 2022.

[4] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022. 3

[5] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar. Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set. *IEEE Robotics and Automation Magazine*, 22(3), Sept. 2015.

[6] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021.

[7] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6D object pose and size estimation. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 11973–11982, 2020.

[8] Ho Kei Cheng and Alexander G Schwing. XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. In *ECCV*, 2022.

[9] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pages 303–312, 1996. 6

[10] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017.

[11] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[12] Clement Fuji Tsang, Maria Shugrina, Jean Francois Lafleche, Towaki Takikawa, Jiehan Wang, Charles Loop, Wenzheng Chen, Krishna Murthy Jatavallabhula, Edward Smith, Artem Rozantsev, Or Perel, Tianchang Shen, Jun Gao, Sanja Fidler, Gavriel State, Jason Gorski, Tommy Xiang, Jianing Li, Michael Li, and Rev Lebaredian. Kaolin: A PyTorch library for accelerating 3D deep learning research. https://github.com/NVIDIAGameWorks/kaolin, 2022.

[13] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *International Conference on Machine Learning (ICML)*, pages 3789–3799, 2020.

[14] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020.

[15] Ankur Handa, Arthur Allshire, Viktor Makoviychuk, Aleksei Petrenko, Ritvik Singh, Jingzhou Liu, Denys Makoviichuk, Karl Van Wyk, Alexander Zhurkevich, Balakumar Sundaralingam, et al. DeXtreme: Transfer of agile in-hand manipulation from simulation to reality. *arXiv preprint arXiv:2210.13702*, 2022.

[16] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 571–580, 2020.

[17] Yisheng He, Yao Wang, Haoqiang Fan, Jian Sun, and Qifeng Chen. FS6D: Few-shot 6D pose estimation of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6814–6824, 2022.

[18] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *German Conference on Pattern Recognition (GCPR)*, volume 13485 of *Lecture Notes in Computer Science*, pages 281–299, 2022.

[19] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

[20] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 362–379. Springer, 2016.

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

[22] Daniel Kappler, Franziska Meier, Jan Issac, Jim Mainprice, Cristina Garcia Cifuentes, Manuel Wüthrich, Vincent Berenz, Stefan Schaal, Nathan Ratliff, and Jeannette Bohg. Real-time perception meets reactive motion generation. *IEEE Robotics and Automation Letters*, 3(3):1864–1871, 2018.

[23] Michael Krainin, Peter Henry, Xiaofeng Ren, and Dieter Fox. Manipulator and object tracking for in-hand 3d object modeling. *The International Journal of Robotics Research*, 30(11):1311–1327, 2011.

[24] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. CosyPose: Consistent multi-view multi-object 6D

pose estimation. In *European Conference on Computer Vision*, pages 574–591, 2020.

[25] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. MegaPose: 6D pose estimation of novel objects via render & compare. In *6th Annual Conference on Robot Learning (CoRL)*, 2022.

[26] Jiahui Lei, Srinath Sridhar, Paul Guerrero, Minhyuk Sung, Niloy Mitra, and Leonidas J Guibas. Pix2Surf: Learning parametric 3D surface models of objects from images. In *European Conference on Computer Vision (ECCV)*, pages 121–138, 2020.

[27] Xiaolong Li, Yijia Weng, Li Yi, Leonidas Guibas, A. Lynn Abbott, Shuran Song, and He Wang. Leveraging SE(3) equivariance for self-supervised category-level object pose estimation from point clouds. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:15370–15381, 2021.

[28] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIM: Deep iterative matching for 6D pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018.

[29] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2041–2050, 2018.

[30] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A Vela, and Stan Birchfield. Keypoint-based category-level object pose tracking from an RGB sequence with uncertainty estimation. In *International Conference on Robotics and Automation (ICRA)*, 2022.

[31] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3D hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14687–14697, 2021.

[32] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. *ECCV*, 2022.

[33] Lu Ma, Mahsa Ghafarianzadeh, David Coleman, Nikolaus Correll, and Gabe Sibley. Simultaneous localization, mapping, and manipulation for unsupervised object discovery. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[34] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: A hands-on survey. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(12):2633–2651, 2015.

[35] John McCormac, Ronald Clark, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Fusion++: Volumetric object-level SLAM. In *International Conference on 3D Vision (3DV)*, pages 32–41, 2018.

[36] Nathaniel Merrill, Yuliang Guo, Xingxing Zuo, Xinyu Huang, Stefan Leutenegger, Xi Peng, Liu Ren, and Guoquan Huang. Symmetry and uncertainty-aware object SLAM for 6DoF object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14901–14910, 2022.

[37] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[38] Norman Müller, Yu-Shiang Wong, Niloy J Mitra, Angela Dai, and Matthias Nießner. Seeing behind objects for 3D multi-object tracking in RGB-D sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6071–6080, 2021.

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

[40] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8280–8290, 2022.

[41] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

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

[43] 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. 2, 5, 6, 7, 8, 9, 10

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

[45] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. LatentFusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10710–10719, 2020.

[46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 1

[47] Timothy Patten, Kiru Park, Markus Leitner, Kevin Wolfram, and Markus Vincze. Object learning for 6D pose estimation and grasping from RGB-D videos of in-hand manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4831–4838, 2021.

[48] Carl Yuheng Ren, Victor Prisacariu, David Murray, and Ian Reid. STAR3D: Simultaneous tracking and reconstruction

of 3D objects using RGB-D data. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1561–1568, 2013.

[49] Martin Rünz and Lourdes Agapito. Co-Fusion: Real-time segmentation, tracking and fusion of multiple objects. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4471–4478. IEEE, 2017.

[50] Martin Runz, Maud Buffier, and Lourdes Agapito. MaskFusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 10–20, 2018. 6

[51] Renato F. Salas-Moreno, Richard A. Newcombe, Hauke Strasdat, Paul H. J. Kelly, and Andrew J. Davison. SLAM++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1352–1359, 2013.

[52] Akash Sharma, Wei Dong, and Michael Kaess. Compositional and scalable object SLAM. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 11626–11632, 2021.

[53] Miroslava Slavcheva, Wadim Kehl, Nassir Navab, and Slobodan Ilic. SDF-2-SDF registration for real-time 3D reconstruction from RGB-D data. *International Journal of Computer Vision (IJCV)*, 126(6):615–636, 2018. 2, 5, 6, 7, 8, 9, 10

[54] Manuel Stoiber, Martin Sundermeyer, and Rudolph Triebel. Iterative corresponding geometry: Fusing region and depth for highly efficient 3d tracking of textureless objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6855–6865, 2022.

[55] Michael Strecke and Jörg Stückler. EM-fusion: Dynamic object-level SLAM with probabilistic data association. In *Proceedings IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[56] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. iMAP: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6229–6238, 2021.

[57] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[58] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022.

[59] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021.

[60] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3D orientation learning for 6D object detection from RGB images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 699–715, 2018.

[61] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual slam for monocular, stereo, and RGB-D cameras. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:16558–16569, 2021. 1, 2, 3, 5, 6, 7, 8, 9, 10

[62] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6D object pose and size estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 530–546, 2020.

[63] Henning Tjaden, Ulrich Schwanecke, and Elmar Schomer. Real-time monocular pose estimation of 3D objects using temporally consistent local color histograms. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 124–132, 2017.

[64] Kentaro Wada, Edgar Sucar, Stephen James, Daniel Lenton, and Andrew J Davison. Morefusion: Multi-object reasoning for 6d pose estimation from volumetric fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14540–14549, 2020.

[65] Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu. 6-PACK: Category-level 6D pose tracker with anchor-based keypoints. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 10059–10066, 2020.

[66] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. GDR-Net: Geometry-guided direct regression network for monocular 6D object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16611–16621, 2021.

[67] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6D object pose and size estimation. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 2642–2651, 2019.

[68] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[69] Bowen Wen and Kostas Bekris. BundleTrack: 6D pose tracking for novel objects without instance or category-level 3D models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8067–8074, 2021. 1, 2, 3, 5, 6, 7, 8, 9, 10

[70] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. Catgrasp: Learning category-level task-relevant grasping in clutter from simulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6401–6408. IEEE, 2022.

[71] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. You only demonstrate once: Category-level manipulation from single visual demonstration. *RSS*, 2022.

[72] Bowen Wen, Chaitanya Mitash, Baozhang Ren, and Kostas E Bekris. se(3)-TrackNet: Data-driven 6D pose tracking

by calibrating image residuals in synthetic domains. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10367–10373, 2020. 3

[73] Bowen Wen, Chaitanya Mitash, Sruthi Soorian, Andrew Kimmel, Avishai Sintov, and Kostas E Bekris. Robust, occlusion-aware pose estimation for objects grasped by adaptive hands. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6210–6217. IEEE, 2020.

[74] Yijia Weng, He Wang, Qiang Zhou, Yuzhe Qin, Yueqi Duan, Qingnan Fan, Baoquan Chen, Hao Su, and Leonidas J Guibas. CAPTRA: Category-level pose tracking for rigid and articulated objects from point clouds. *ICCV*, 2021.

[75] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018.

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

[77] Binbin Xu and et al. Mid-fusion: Octree-based object-level multi-instance dynamic slam. In *ICRA*, 2019.

[78] H. Yang, J. Shi, and L. Carlone. TEASER: Fast and certifiable point cloud registration. *IEEE Trans. Robotics*, 37(2):314–333, Apr. 2021. 6

[79] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11097–11106, 2021.

[80] Zhenpei Yang, Zhile Ren, Miguel Angel Bautista, Zaiwei Zhang, Qi Shan, and Qixing Huang. FvOR: Robust joint shape and pose optimization for few-view object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2497–2507, 2022.

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

[82] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:2492–2502, 2020.

[83] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3DMatch: Learning local geometric descriptors from RGB-D reconstructions. In *CVPR*, 2017. 6

[84] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3D human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020.

[85] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. NICE-SLAM: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 2, 3, 5, 6, 7, 8, 9, 10