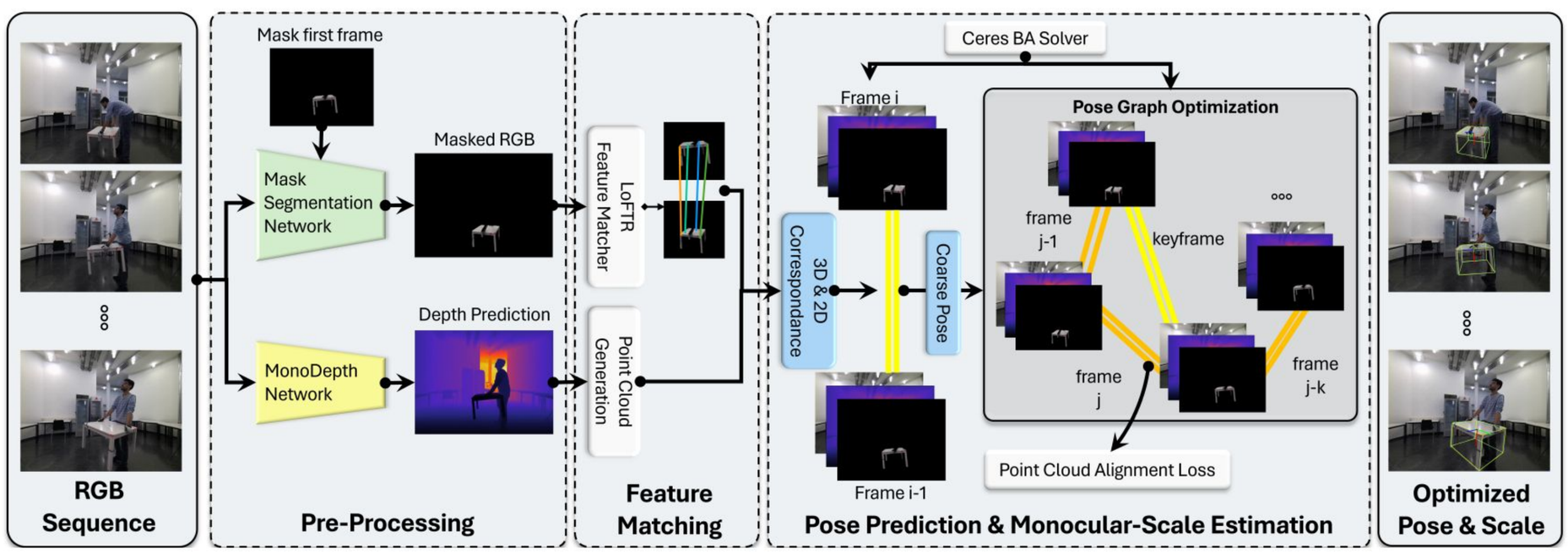
MonoTracker: Monocular RGB-Only 6D Tracking of Unknown Objects

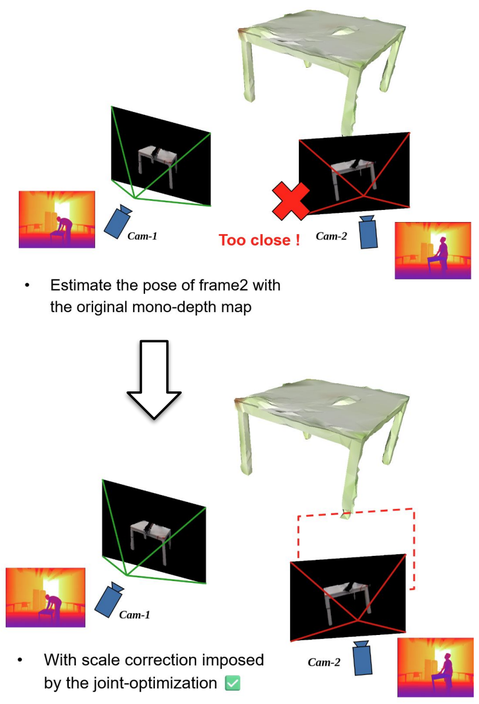
Zilong Deng[1,2], Shaochang Tan[1,2], Zuria Bauer[2], Dániel Baráth[2,3], Marc Pollefeys[2,4]
[1]University of Zurich; [2]ETH Zurich; [3]HUN-REN SZTAKI; [4]Microsoft

1 Introduction

Estimating the six degrees of freedom (6DoF) pose of unknown objects using only monocular RGB images is a challenging task, especially when dealing with textureless and small objects. Textureless surfaces hinder feature detection, while small motion complicates the estimation of geometric relationships. Monocular RGB inputs introduce scale ambiguity and inconsistencies across frames. Without depth information or pre-existing object models, traditional feature-based methods and SfM pipelines become ineffective. MonoTracker is a model-free, RGB-only, 6D detector that works on unseen objects, enabling zero-shot 6D pose estimation by jointly optimizing object poses and correcting scale inconsistencies in monocular depth predictions.

2 Method Overview





3 Technical Details

MonoTracker takes as input a sequence of images and the object mask from the first frame.

Data Pre-processing: It simultaneously tracks the object mask in subsequent frames using XMem[1] and obtains metric monocular depth predictions via DepthAnything-v2[2].

Initial Pose Estimation: Feature matching within the masked area using LoFTR [3] between consecutive frames and previous keyframes to establish point correspondences. Following by a robust PnP solver to estimate initial pose.

Pose Graph Construction: The new frame is connected to the previous ones selected as keyframes in a pose graph.

Joint Pose-Scale Optimization: Solve an optimization problem that jointly refines the pose and scale of the depth maps, by minimizing point cloud alignment, disparity, and projection error-based losses.

• Keyframes in Pose Graph

Keyframes: with the closest estimated orientations within a window size of 15 frames. The **Greedy Rotation Similarity** strategy forms a robust pose graph with frames share similar viewpoints.

$$\Delta R_{ij} = \arccos\left(\left(\mathrm{trace}(\mathbf{R}_{\mathrm{pred}}^{i\,T}\mathbf{R}_{\mathrm{pred}}^{j}) - 1\right)/2\right)$$

• Joint Pose-Scale Optimization

We run a joint pose-scale optimization both for coarse pose estimation between neighboring frames and in a global Bundle Adjustment that runs every few frames.

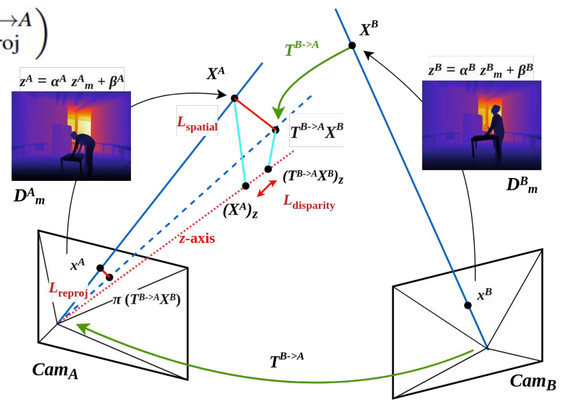
Optimized Variables: Poses $\boldsymbol{T^W_i}$ with respect to the World coordinate system and scale-shift parameters $\boldsymbol{\alpha^i}$, $\boldsymbol{\beta^i}$ for each frame are optimized variables.

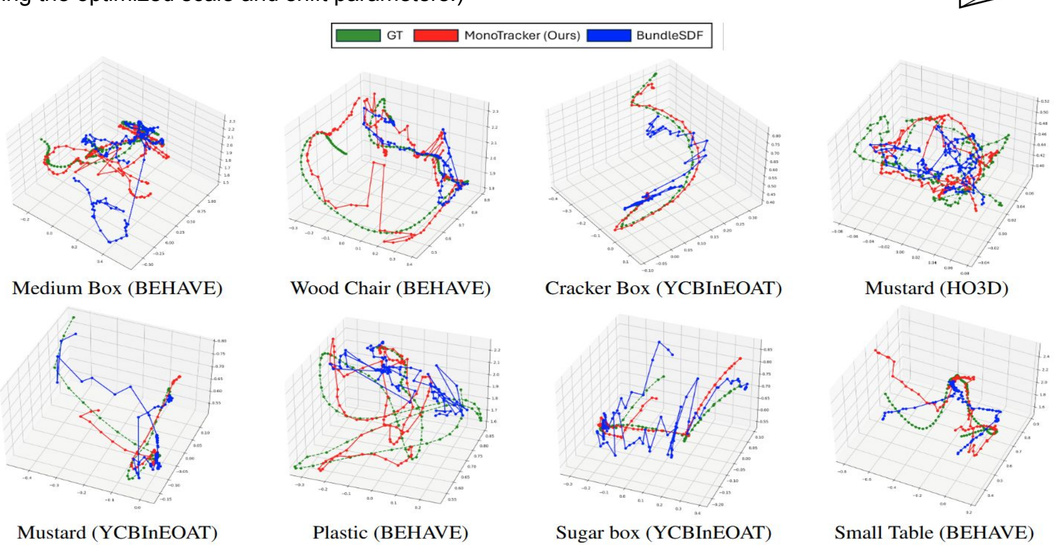
Loss Function: Weighted sum over all point correspondence $\mathcal{X}^{AB} = \{(\mathbf{x}_j^A, \mathbf{x}_j^B)\}_{j=0}^{n^{AB}-1}$

$$\mathcal{L}^{B\to A} = \sum_{j=0}^{n^{AB}-1}\left(\lambda_s\mathcal{L}_{\mathrm{spatial}}^{j,B\to A} + \lambda_d\mathcal{L}_{\mathrm{disparity}}^{j,B\to A} + \lambda_r\mathcal{L}_{\mathrm{reproj}}^{j,B\to A}\right)$$

$$\begin{cases} \mathcal{L}_{\mathrm{spatial}}^{j,B\to A} = \left\|\mathbf{X}_j^A - \mathbf{T}^{B\to A}\mathbf{X}_j^B\right\|_2^2 \\ \mathcal{L}_{\mathrm{disparity}}^{j,B\to A} = \left|\dfrac{1}{(\mathbf{X}_j^A)_z} - \dfrac{1}{(\mathbf{T}^{B\to A}\mathbf{X}_j^B)_z}\right| \\ \mathcal{L}_{\mathrm{reproj}}^{j,B\to A} = \left\|\mathbf{x}_j^A - \pi\left(\mathbf{K}_A, \mathbf{T}^{B\to A}\mathbf{X}_j^B\right)\right\|_2^2 \end{cases}$$

(where $\mathbf{X}_j^A$ and $\mathbf{X}_j^B$ are 3D points back-projected using the optimized scale and shift parameters.)



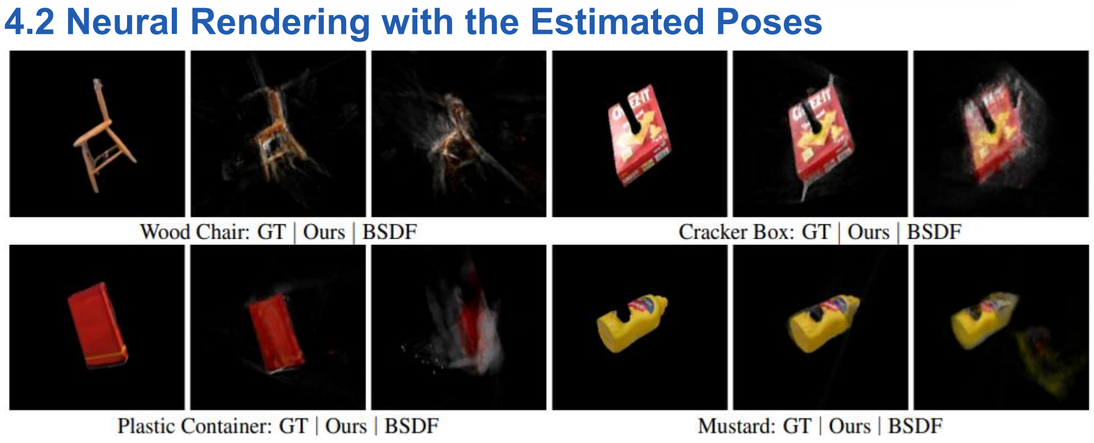


4 Results and Metrics

• Estimating 6D object trajectories on three datasets

YCBInEOAT	COLMAP	DSfM	ACE0	RayDiffusion	MonoGS	BundleSDF	Ours
Pos. Error (cm) ↓	34.76	37.91	--	12.43	7.39	5.84	3.79
Accuracy @ 5cm ↑	0.14	0.08	--	0.52	0.61	0.61	0.82
Accuracy @ 10cm ↑	0.32	0.28	--	0.73	0.81	0.86	0.95
AUC @ 5cm ↑	0.06	0.03	--	0.28	0.32	0.30	0.46
AUC @ 10cm ↑	0.15	0.10	--	0.47	0.53	0.55	0.69
ADD-S @ 10cm ↑	0.02	0.17	--	0.32	0.39	0.53	0.63
ADD @ 10cm ↑	0.01	0.00	--	0.11	0.15	0.27	0.50
ADD-S @ 30cm ↑	0.23	0.22	--	0.65	0.74	0.82	0.86
ADD @ 30cm ↑	0.23	0.21	--	0.56	0.62	0.71	0.81
Med. Rel. Rot. Err. (°) ↓	12.33	12.10	--	8.20	1.39	2.85	1.49
Mean Rel. Rot. Err. (°) ↓	13.29	13.05	--	11.21	3.12	4.54	2.68
Rel. Rot. Err. AUC@5° ↑	0.21	0.24	--	0.18	0.60	0.46	0.61
BEHAVE	**COLMAP**	**DSfM**	**ACE0**	**RayDiffusion**	**MonoGS**	**BundleSDF**	**Ours**
Pos. Error (cm) ↓	43.14	--	42.83	30.14	34.92	24.26	19.37
Accuracy @ 5cm ↑	0.11	--	0.06	0.14	0.20	0.30	0.36
Accuracy @ 10cm ↑	0.20	--	0.15	0.34	0.43	0.46	0.63
AUC @ 5cm ↑	0.05	--	0.03	0.06	0.10	0.12	0.14
AUC @ 10cm ↑	0.11	--	0.07	0.16	0.22	0.26	0.33
ADD-S @ 10cm ↑	0.00	--	0.00	0.06	0.00	0.01	0.10
ADD @ 10cm ↑	0.00	--	0.00	0.00	0.00	0.00	0.01
ADD-S @ 30cm ↑	0.00	--	0.01	0.17	0.06	0.12	0.25
ADD @ 30cm ↑	0.00	--	0.01	0.00	0.04	0.07	0.11
Med. Rel. Rot. Err. (°) ↓	11.42	--	27.16	7.34	3.89	4.48	2.89
Mean Rel. Rot. Err. (°) ↓	11.80	--	32.11	14.48	4.51	6.42	3.57
Rel. Rot. Err. AUC@5° ↑	0.20	--	0.05	0.16	0.34	0.30	0.44
HO3D	**COLMAP**	**DSfM**	**ACE0**	**RayDiffusion**	**MonoGS**	**BundleSDF**	**Ours**
Pos. Error (cm) ↓	--	3.37	3.96	3.14	2.93	2.67	2.65
Accuracy @ 5cm ↑	--	0.69	0.68	0.88	0.92	0.96	0.97
Accuracy @ 10cm ↑	--	0.77	0.89	1.00	1.00	1.00	1.00
AUC @ 5cm ↑	--	0.31	0.27	0.41	0.45	0.48	0.48
AUC @ 10cm ↑	--	0.53	0.54	0.69	0.71	0.74	0.74
ADD-S @ 10cm ↑	--	0.15	0.14	0.35	0.25	0.38	0.42
ADD @ 10cm ↑	--	0.09	0.13	0.22	0.22	0.27	0.28
ADD-S @ 30cm ↑	--	0.57	0.64	0.78	0.75	0.81	0.79
ADD @ 30cm ↑	--	0.56	0.63	0.73	0.74	0.76	0.76
Med. Rel. Rot. Err. (°) ↓	--	6.07	21.29	14.43	7.08	5.44	4.69
Mean Rel. Rot. Err. (°) ↓	--	9.62	28.85	24.20	7.91	6.76	5.79
Rel. Rot. Err. AUC@5° ↑	--	0.17	0.07	0.07	0.17	0.24	0.30

4.2 Neural Rendering with the Estimated Poses



References

1. Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In ECCV, pages 640–658, 2022.
2. Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. arXiv preprint arXiv:2406.09414, 2024.
3. Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In IEEE CVPR, pages 8922–8931, 2021.
4. Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In IEEE Conf. Comput. Vis. Pattern Recog., 2016.
5. Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In IEEE CVPR, pages 21594–21603, 2024.
6. Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In ECCV, pages 421–440, 2024.
7. Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In IEEE CVPR, pages 18039–18048, 2024.
8. Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. arXiv preprint arXiv:2402.14817, 2024.
9. Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In IEEE CVPR, pages 606–617, 2023.