Supplementary Material
Occlusion-Robust Relative Pose Estimation for Multi-Robot Systems via Geometric-Aware Diffusion Matching

I. Experimental Result

As described in the main paper, the overall loss is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{diff} + \lambda_1 \mathcal{L}_{match} + \lambda_2 \mathcal{L}_{geom}. \tag{1}$$

To determine the appropriate weights, we evaluate average ATE across different combinations of (λ_1, λ_2), as shown in Fig. 1. This grid search–style analysis visualizes the impact of each hyperparameter choice, where each cell corresponds to a particular setting of (λ_1, λ_2) and lower values indicate better pose accuracy. To ensure robustness, we report both overall accuracy and per-occlusion performance, allowing us to identify hyperparameters that remain stable across varying occlusion rates. From these results, we select $\lambda_1 = 1$ and $\lambda_2 = 0.1$, which consistently yield low ATE without overemphasizing either the matching or geometric-consistency term.

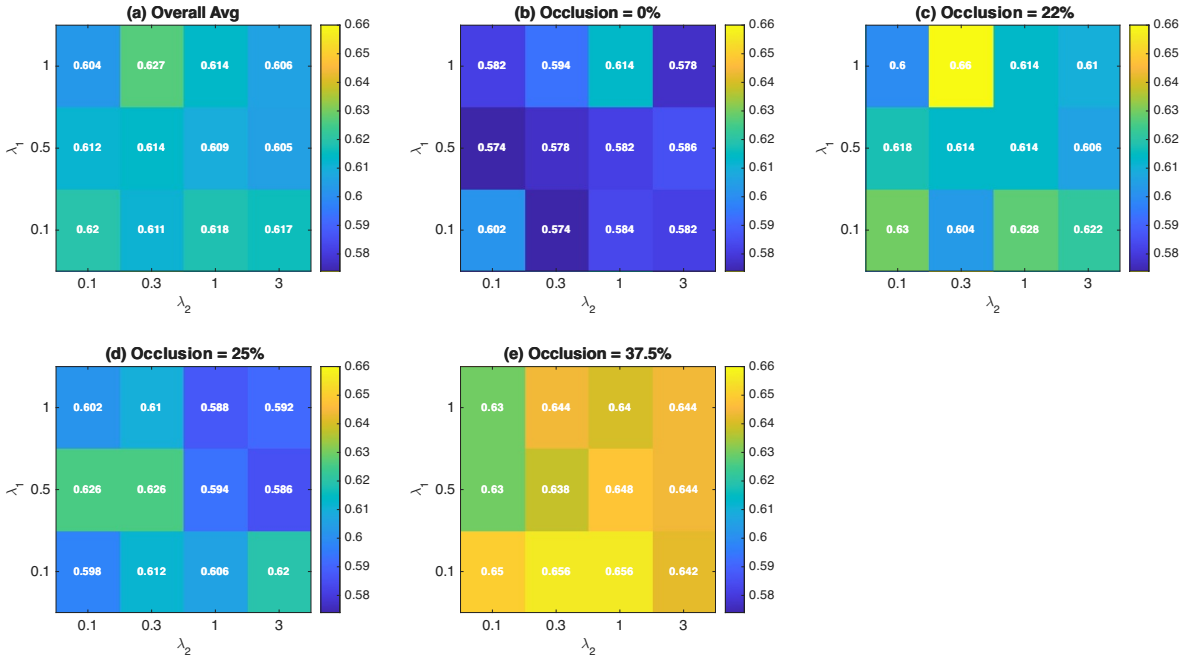


Fig. 1. ATE heatmaps across different (λ_1, λ_2) settings. Each cell represents a specific hyperparameter pair, with color intensity reflecting pose accuracy (lower is better). Both overall and per-occlusion results are reported to guide stable parameter selection under varying occlusion levels.

II. Results on the Extended TUM-RGBD Dataset

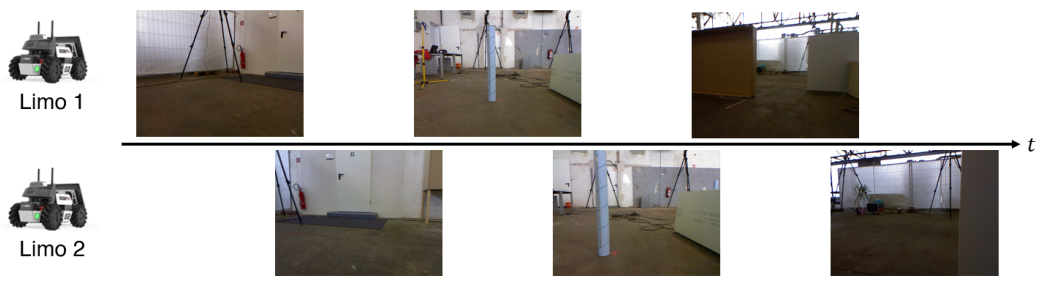


Fig. 2. Extension of the TUM-RGBD dataset to simulate the perception of a ground-robot pair for evaluating relative pose estimation.

TABLE I

EVALUATION RESULTS USING ATE RMSE (M) ON EXTENDED TUM DATASETS ACROSS DIFFERENT OCCLUSION RATES.

Occ Rate	Method	desk_with_person	large_no_loop	large_with_loop	pioneer_slam	pioneer_slam2	pioneer_slam3	long_office_household	structure_notexture_far	structure_texture_far
0%	CoViS-Net	0.57	1.29	1.25	1.55	1.47	1.45	0.93	0.71	0.5
	ORB-BF	0.06	0.45	0.3	0.35	0.22	0.19	0.06	0.18	0.03
	Ours	**0.03**	**0.17**	**0.15**	**0.08**	**0.06**	**0.08**	**0.04**	**0.03**	**0.02**
22%	CoViS-Net	0.67	1.39	1.37	1.54	1.52	1.51	0.96	0.5	0.62
	ORB-BF	0.06	0.43	0.28	0.22	0.2	0.23	0.09	0.04	0.04
	Ours	**0.03**	**0.18**	**0.17**	**0.07**	**0.08**	**0.09**	**0.05**	**0.02**	**0.02**
25%	CoViS-Net	0.68	1.46	1.26	1.55	1.54	1.53	0.98	0.59	0.58
	ORB-BF	0.07	0.48	0.31	0.25	0.2	0.27	0.11	0.14	0.04
	Ours	**0.03**	**0.2**	**0.15**	**0.08**	**0.08**	**0.1**	**0.05**	**0.02**	**0.02**
37.5%	CoViS-Net	1.02	1.63	1.52	1.58	1.58	1.63	1.04	0.55	0.69
	ORB-BF	0.07	0.56	0.28	0.24	0.22	0.33	0.11	0.14	0.06
	Ours	**0.03**	**0.19**	**0.17**	**0.08**	**0.09**	**0.1**	**0.05**	**0.03**	**0.02**

TABLE II

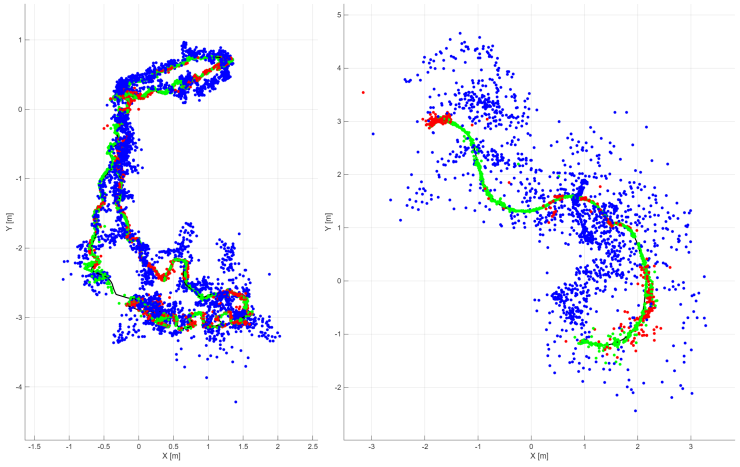
EVALUATION RESULTS USING ARE RMSE (DEG) ON EXTENDED TUM DATASETS ACROSS DIFFERENT OCCLUSION RATES.

Occ Rate	Method	desk_with_person	large_no_loop	large_with_loop	pioneer_slam	pioneer_slam2	pioneer_slam3	long_office_household	structure_notexture_far	structure_texture_far
0%	CoViS-Net	18.97	31.93	28.03	50.46	42.61	45.01	31.65	37.63	11.71
	ORB-BF	1.56	5.76	2.42	6.03	2.91	2.57	1.43	3.88	0.68
	Ours	**0.68**	**2.26**	**1.42**	**1.57**	**1.23**	**1.66**	**1.09**	**0.74**	**0.56**
22%	CoViS-Net	25	33.23	33.54	45.56	44.5	39.84	32.09	18.77	12.16
	ORB-BF	1.56	5.76	2.42	6.03	2.91	2.57	1.43	3.88	0.68
	Ours	**0.74**	**2.45**	**1.56**	**1.26**	**1.39**	**1.61**	**1.2**	**0.52**	**0.61**
25%	CoViS-Net	26.42	36.98	25.16	46.73	44.71	45.05	32.66	27.68	10.82
	ORB-BF	1.56	6.2	2.57	3.2	2.69	3.45	2.56	2.48	0.8
	Ours	**0.72**	**2.7**	**1.46**	**1.32**	**1.66**	**1.69**	**1.27**	**0.53**	**0.58**
37.5%	CoViS-Net	42.03	44.27	28.85	44.59	44.88	44.85	33.58	20.62	11.49
	ORB-BF	1.84	7.2	2.19	3.09	2.72	4.38	2.23	2.49	1.11
	Ours	**0.75**	**2.63**	**1.46**	**1.28**	**1.48**	**1.79**	**1.08**	**0.59**	**0.62**

We additionally evaluate our approach on the widely used TUM-RGBD dataset [1], which features full 3D motion captured with a handheld RGB-D camera across diverse environments, including textureless surfaces and dynamic objects. To adapt this single-camera SLAM dataset for multi-robot relative pose estimation, we construct paired sequences by introducing a 1-second temporal offset between trajectories. Fig. 2 illustrates the evaluation setup.

Tables I and II summarize the translation error (ATE) and rotation error (ARE), respectively. Our method achieves consistently higher accuracy than both CoViS-Net and ORB-BF across occlusion rates. At 0% occlusion, our approach achieves an average ATE of 0.07 m compared to 1.08 m for CoViS-Net and 0.20 m for ORB-BF. Even at 37.5% occlusion, our ATE remains stable at 0.08 m, while CoViS-Net degrades to 1.25 m and ORB-BF to 0.22 m. Similarly, for rotation, our ARE is 1.24° at 0% occlusion versus 33.11° for CoViS-Net and 3.03° for ORB-BF. At 37.5% occlusion, our ARE remains nearly unchanged at 1.30°, while CoViS-Net worsens to 35.02° and ORB-BF stays at 3.03°.

We visualize the performance of different methods in Fig. 3 (a) and (b) using sample sequences containing dynamic obstacles and full 3D handheld motion. Our method maintains consistent alignment with ground truth despite occlusions and scene clutter. ORB-BF frequently fails under such conditions, which can be seen as many missing pose estimates. On the other hand, CoViS-Net occasionally achieves reasonable accuracy but suffers from significant drift in the presence of dynamic objects in the scene.



(a) desk with person (b) long office household

Fig. 3. Visualization of map-free pose estimation shown in the global coordinate frame, obtained from offline evaluation of TUM-RGBD dataset.

REFERENCES

[1] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.