ViSTA-SLAM: Visual SLAM with Symmetric Two-view Association

Supplementary Material

Method	Metric	chess	fire	heads	office	pumpkin	redkitchen	stairs	Avg
CUT3R	Accuracy	0.274	0.102	0.106	0.326	0.452	0.325	0.345	0.276
	Completeness	0.303	0.081	0.093	0.441	0.459	0.358	0.389	0.303
	Chamfer	0.289	0.091	0.100	0.384	0.456	0.342	0.367	0.290
	ATE	0.743	0.226	0.363	0.664	0.546	0.381	0.413	0.477
SLAM3R	Accuracy	0.043	0.022	0.020	0.035	0.072	0.062	0.116	0.053
	Completeness	0.030	0.013	0.015	0.030	0.055	0.061	0.209	<u>0.059</u>
	Chamfer	0.037	0.018	0.017	0.033	0.064	0.061	0.162	<u>0.056</u>
	ATE	0.089	0.048	0.036	0.088	0.196	0.102	0.126	0.098
MASt3R-SLAM	Accuracy	0.090	0.037	0.027	0.047	0.097	0.070	0.045	0.059
	Completeness	0.055	0.024	0.021	0.053	0.054	0.036	0.149	**0.056**
	Chamfer	0.073	0.031	0.024	0.050	0.075	0.053	0.097	0.057
	ATE	0.063	0.046	0.029	0.103	0.112	0.074	0.032	<u>0.066</u>
VGGT-SLAM	Accuracy	0.029	0.014	0.031	0.041	0.128	0.036	0.087	<u>0.052</u>
	Completeness	0.052	0.064	0.021	0.066	0.054	0.057	0.110	0.061
	Chamfer	0.040	0.039	0.026	0.054	0.091	0.047	0.098	<u>0.056</u>
	ATE	0.037	0.026	0.022	0.103	0.147	0.063	0.095	0.070
Ours	Accuracy	0.065	0.015	0.031	0.036	0.061	0.035	0.074	**0.045**
	Completeness	0.063	0.022	0.040	0.048	0.037	0.030	0.154	**0.056**
	Chamfer	0.064	0.019	0.035	0.042	0.049	0.033	0.114	**0.051**
	ATE	0.073	0.035	0.028	0.055	0.129	0.035	0.029	**0.055**

Table 7. **Per Scene Evaluation on 7-Scenes [43].** Comparison of accuracy, completeness, Chamfer distance, and trajectory error on the 7-Scenes dataset. Lower is better. Best results are **bold**, second best are <u>underlined</u>.

5. Relative Scale in Pose Graph

As mentioned in Sec. 2.3, the *scale edges* connect nodes corresponding to the same view but obtained from different forwarding passes. Since the training supervision of the frontend STA model uses only normalized pointmaps, the scales of the same view across different passes are not consistent. Therefore, estimating the relative scale is crucial for pose graph construction. Given two pointmaps $\boldsymbol{P}_i^j$ and $\boldsymbol{P}_i^k$ of the same view i (obtained from forwarding passes with input view i j and view i k, respectively), along with their confidence maps $\boldsymbol{C}_i^j$ and $\boldsymbol{C}_i^k$, we first get the confidence score w_x of the point pair for pixel $\boldsymbol{x}$,

$$w_x = \boldsymbol{C}_i^j(\boldsymbol{x}) \cdot \boldsymbol{C}_i^k(\boldsymbol{x}),$$

then, the relative scale s_i^{jk} can be computed as,

$$\begin{aligned} s_i^{jk} &= \min_s \sum_{\boldsymbol{x}} w_x \|\boldsymbol{P}_i^j(\boldsymbol{x}) - s\boldsymbol{P}_i^k(\boldsymbol{x})\|^2 \\ &= \frac{\sum_{\boldsymbol{x}} w_x (\boldsymbol{P}_i^j(\boldsymbol{x}) \cdot \boldsymbol{P}_i^k(\boldsymbol{x}))}{\sum_{\boldsymbol{x}} w_x \|\boldsymbol{P}_i^k(\boldsymbol{x})\|^2}. \end{aligned} \tag{11}$$

6. Additional Quantitative Results

6.1. Per Scene Evaluation Results on 7-Scenes

In Sec. 3.2, only the average reconstruction evaluation is provided. In Tab. 7, we present more detailed per-scene results on 7-Scenes [43] to offer deeper insights. The pure regression method SLAM3R [27] performs well in scenes where the camera primarily focuses on a single corner,

such as chess, fire, and heads. However, in scenes involving longer camera trajectories, like pumpkin and redkitchen, its performance degrades due to difficulties in accurately registering points. Our method, ViSTA-SLAM, achieves the best performance in average across all four metrics.

6.2. Additional Trajectory Results on TUM-RGBD

In Sec. 3.1, to align with previous methods, only results for the *freiburg1* partition of the TUM RGB-D dataset [45] are reported. In Tab. 8, we also report results for the *freiburg2* and *freiburg3* partitions.

Sequence	ATE RMSE (m)
freiburg2_360_hemisphere	0.2037
freiburg2_360_kidnap	0.4617
freiburg2_desk	0.0577
freiburg2_large_with_loop	0.2170
freiburg2_rpy	0.0222
freiburg2_xyz	0.0155
freiburg3_cabinet	0.3869
freiburg3_large_cabinet	0.1334
freiburg3_long_office_household	0.1013
freiburg3_teddy	0.0789

Table 8. **Trajectory ATE results on the *freiburg2* and *freiburg3* partitions of the TUM RGB-D dataset [45].**

6.3. Additional Evaluation on More Datasets

In Tab. 9 and Tab. 10, we additionally report camera trajectory evaluation results on Replica [44] and ScanNet [8], as well as reconstruction evaluation results on Replica for several commonly used SLAM testing scenes.

7. Additional Qualitative Results

7.1. Pose Graph Optimization

In Fig. 7, we compare the reconstruction and trajectory estimation results with and without pose graph optimization on ScanNet [8] scene0000_00. Pose graph optimization effectively corrects misaligned areas and averages out the errors from the frontend.

7.2. Wrong Loop Filtering

In Sec. 2.3, we describe feeding each loop candidate pair into our STA model to verify their spatial proximity. This is necessary because Bag of Words loop detection can produce false positives, which may significantly degrade performance by introducing misleading edges into the pose graph. As shown in Fig. 8, rejecting incorrect loop candidates using the relative pose confidence score provided by STA results in a much more stable performance.

Scene	ATE	Acc.	Comp.	Chamfer
office0	0.0744	0.0595	0.0226	0.0410
office1	0.1934	0.2614	0.1833	0.2223
office2	0.1177	0.0914	0.0316	0.0615
office3	0.0485	0.0623	0.0221	0.0422
office4	0.1302	0.1338	0.0688	0.1013
room0	0.0688	0.0766	0.0209	0.0488
room1	0.0934	0.1105	0.0552	0.0828
room2	0.1363	0.1194	0.0223	0.0709

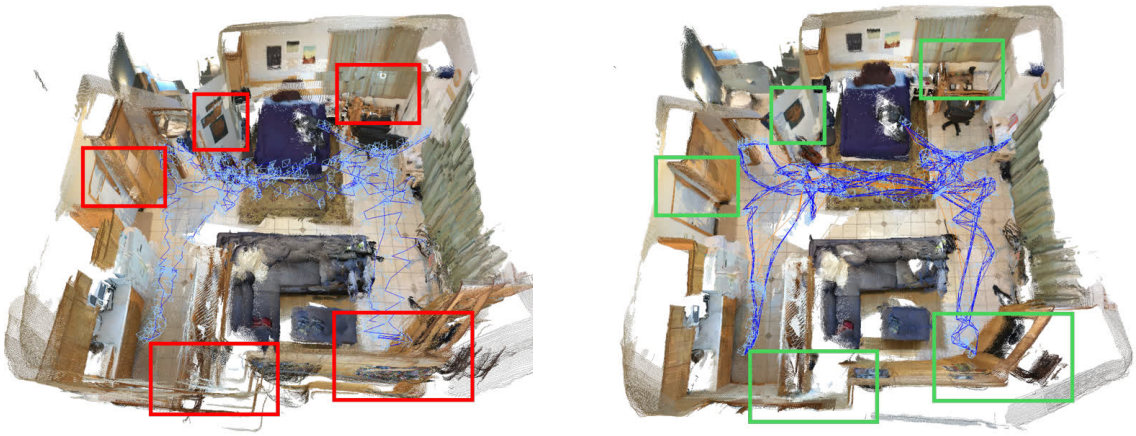
Table 9. **Per-scene evaluation results on Replica [44].**

Sequence	ATE RMSE (m)
scene0000_00	0.0483
scene0059_00	0.0391
scene0106_00	0.0559
scene0169_00	0.0526
scene0181_00	0.0520
scene0207_00	0.0479

Table 10. **Per-scene camera trajectory evaluation results on ScanNet [8].**

7.3. More Results

In Fig. 9, we present the results of ViSTA-SLAM across various datasets. ViSTA-SLAM demonstrates stable performance despite differing camera motions in these scenes. As before, light blue frustums represent camera poses, blue lines connect neighboring views, while orange lines indicate loop closures.



result *w/o* PGO result *w/* PGO

Figure 7. **Qualitative Comparison for Pose Graph Optimization.** Red boxes highlight regions with misalignments, while green boxes indicate areas where these misalignments have been corrected after pose graph optimization.

Figure 8. **Qualitative Comparison of Wrong Loop Filtering.** Keeping wrong loop candidates decreases the performance a lot.

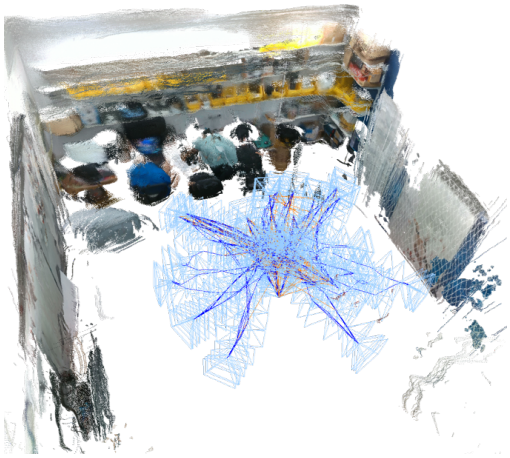
BundleFusion [9] `apt2`

TUM-RGBD [45] `floor`

BundleFusion [9] `apt0`

7-Scenes [43] `pumpkin`

M2DGR [55] `room_03`



M2DGR [55] `room_01`

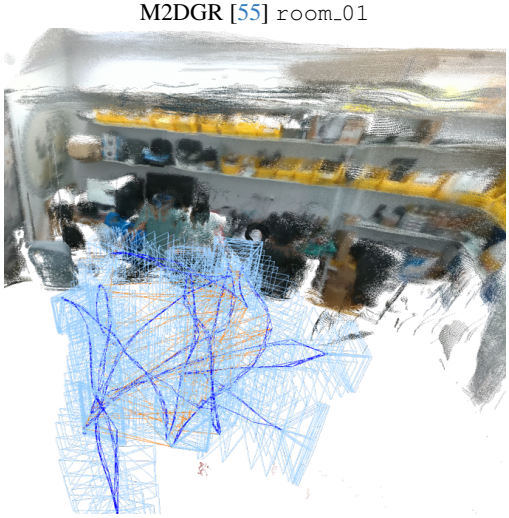


Figure 9. **More Qualitative Results.** Reconstructions and camera trajectories from different datasets.