MGSfM: Multi-Camera Geometry Driven Global Structure-from-Motion

Supplementary Material

Input	Only Trans				Only Track				Hybrid			
Bilinear	yes		no		yes		no		yes		no	
data	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_t$	$\bar{e}_t$	$\tilde{e}_t$	$\bar{e}_t$
00	**0.5**	**0.7**	**0.5**	0.8	**0.5**	**0.7**	0.6	0.7	0.6	0.9	**0.5**	**0.7**
01	0.9	2.7	1.1	1.6	0.8	31.7	**0.5**	23.4	**0.5**	1.2	0.6	**1.1**
02	**0.9**	1.4	**0.9**	1.4	1.3	1.8	**0.9**	1.4	1.0	**1.2**	**0.9**	1.4
03	**0.2**	**0.4**	**0.2**	**0.4**	**0.2**	**0.4**	**0.2**	**0.4**	**0.2**	**0.4**	**0.2**	**0.4**
04	**0.1**	**0.2**	**0.1**	**0.2**	**0.1**	**0.2**	**0.1**	**0.2**	**0.1**	**0.2**	**0.1**	**0.2**
05	**0.2**	**0.3**	**0.2**	**0.3**	**0.2**	**0.3**	**0.2**	**0.3**	**0.2**	**0.3**	**0.2**	**0.3**
06	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**
07	**0.2**	**0.3**	**0.2**	**0.3**	**0.2**	**0.3**	**0.2**	**0.3**	**0.2**	**0.3**	**0.2**	**0.3**
08	1.1	**1.4**	**0.9**	1.5	1.0	1.5	1.0	**1.4**	**0.9**	1.6	**0.9**	**1.4**
09	**0.5**	1.1	**0.5**	**1.0**	**0.5**	**1.0**	**0.5**	**1.0**	**0.5**	**1.0**	**0.5**	**1.0**
10	**0.6**	**1.3**	**0.6**	**1.3**	**0.6**	1.4	**0.6**	1.4	**0.6**	**1.3**	**0.6**	**1.3**

Table 1. Comparison of camera position accuracy estimated by six methods with different input image observations and different angle-based functions on KITTI Odometry dataset.

A. Ablation Study on KITTI Odometry

We analyze the impact of different input image observations and angle-based objective functions on the performance of multi-camera translation averaging on the KITTI Odometry benchmark [3], as shown in Tab. 1. For most data, all six methods yield comparable accuracy in camera positions. Since there is sufficient overlap in the field of view between the stereo cameras, four overlapping image pairs are typically formulated between two adjacent rigid units. Consequently, the relative scales between the rigid units can be accurately estimated using only relative translations, even in cases of collinear camera motion trajectories. In data 01, however, there are fewer feature points and a higher proportion of outliers between image pairs, causing methods that rely solely on feature tracks with random initialization to easily converge to incorrect local optima. In contrast, relative translations exhibit a higher inlier ratio, as they are robustly estimated from multiple feature matches. Given a reasonable initial solution for camera positions and 3D points, the unbiased angle-based refinement formulated from camera-to-point constraints exhibits higher robustness to outlier feature tracks and achieves higher accuracy.

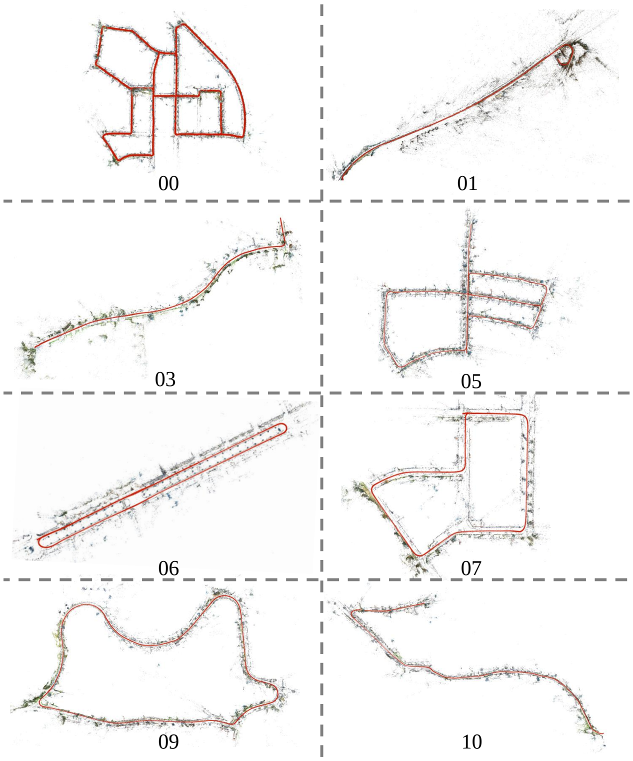

Figure 1. Our reconstruction results from the partial KITTI Odometry benchmark.

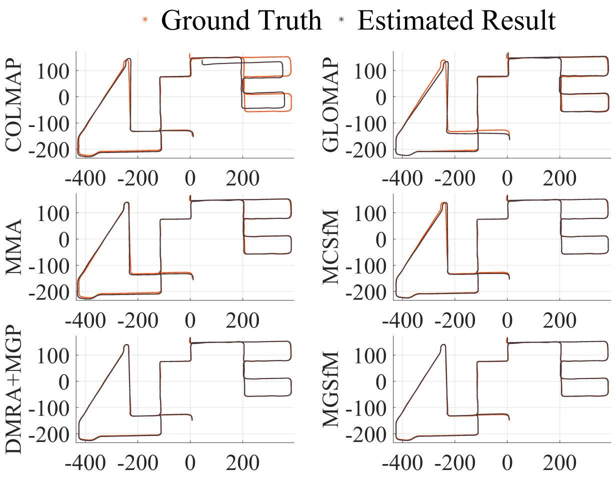

Figure 2. Comparison of camera motion trajectories on data 08 in KITTI Odometry benchmark [3]. The sample state-of-the-art SfM methods include COLMAP [6], GLOMAP [5], MMA [1], MCSfM [2] and our proposed DMRA+MGP and MGSfM.

B. Qualitative Results on KITTI Odometry

Fig. 1 presents additional reconstruction outcomes of MGSfM on the KITTI Odometry benchmark [3]. We compare the camera motion trajectories estimated by six state-of-the-art methods on the challenging large-scale sequence 08, which lacks complete loop closure. As depicted in Fig. 2, the red trajectory represents the ground truth camera positions, while the black trajectory represents the estimated camera

	MCSfM [2]						Median-RA			MGSfM					
Data	e_r^{r1}	e_r^{r2}	e_r^{r3}	e_t^{r1}	e_t^{r2}	e_t^{r3}	e_r^{r1}	e_r^{r2}	e_r^{r3}	e_r^{r1}	e_r^{r2}	e_r^{r3}	e_t^{r1}	e_t^{r1}	e_t^{r3}
0000	1.1	0.5	0.6	0.4	0.4	1.0	1.1	0.6	0.5	1.1	0.6	0.5	0.7	0.6	0.9
0002	1.1	0.5	0.6	0.4	0.4	1.0	1.1	0.6	0.5	1.1	0.6	0.5	0.6	0.3	1.2
0003	1.1	0.5	0.7	0.4	0.6	0.9	1.1	0.5	0.5	1.1	0.6	0.5	0.6	0.1	1.3
0004	1.1	0.5	0.6	0.4	0.4	1.0	1.1	0.6	0.6	1.1	0.6	0.5	0.5	0.3	1.2
0005	1.1	0.5	0.6	0.4	0.4	0.9	1.1	0.6	0.5	1.1	0.6	0.5	0.6	0.4	1.1
0006	1.1	0.5	0.6	0.4	0.4	1.0	1.1	0.6	0.5	1.1	0.6	0.5	0.6	0.2	1.2
0007	1.1	0.5	0.6	0.5	0.5	1.1	1.1	0.5	0.5	1.2	0.6	0.5	0.6	0.5	1.2
0009	1.1	0.5	0.6	0.4	0.4	1.0	1.2	0.6	0.5	1.2	0.7	0.5	0.8	0.5	0.8
0010	1.1	0.5	0.7	0.4	0.6	1.0	1.2	0.6	0.5	1.2	0.6	0.6	0.7	0.3	0.8

Table 2. Accuracy of internal camera poses estimated by MCSfM and MGSfM, as well as the median of internal rotations estimated via single-camera rotation averaging, on KITTI-360 benchmark. The left camera of the stereo pair in multi-camera system is selected as the reference. Here, e_r^r and e_t^r denote the angular errors of relative rotations and relative translations, respectively, in degrees.

positions obtained by the different methods. For incremental methods, compared to COLMAP, the incorporation of multi-camera constraints in MCSfM yields a camera trajectory that is noticeably closer to the ground truth, demonstrating the significance of multi-camera constraints in mitigating scale drift. Although GLOMAP jointly estimates all camera positions using camera-to-point constraints to achieve a uniform error distribution, its estimated camera trajectory still exhibits large errors relative to the ground truth. In contrast, MMA estimates camera positions using only relative translations but achieves higher accuracy than GLOMAP by fusing multi-camera constraints. Furthermore, global methods, such as MGSfM, that employ both camera-to-point and multi-camera constraints to jointly estimate all camera positions achieve a more uniform error distribution, resulting in trajectories that are closer to the ground truth than those produced by incremental methods such as MCSfM.

C. Internal Pose Estimation on KITTI-360

In this section, we compare the internal camera poses estimated by MCSfM [2] and MGSfM, as well as the median of internal camera rotations (denoted as “Median-RA”), on the KITTI-360 benchmark [4]. As shown in Tab. 2, the accuracy of the initial internal camera rotations from Median-RA is comparable to that achieved by MCSfM or to the results refined via BA in MGSfM, demonstrating the robustness of our decoupled rotation averaging method. Moreover, the accuracy of internal camera translations estimated by MCSfM and MGSfM is also comparable, indicating that MGSfM exhibits robustness on par with incremental methods.

D. Runtime of Ablation Study on KITTI-360

We report the runtime of six methods evaluated in an ablation study on the KITTI-360 dataset [4]. These methods respectively utilize only relative translations (denoted as “Only trans”), only feature tracks (denoted as “Only tracks”), or a hybrid of both (denoted as “Hybrid”), and employ either bilinear or non-bilinear angle-based objective functions. As shown in Fig. 3, except for methods with “Only Trans”, MGSfM (Hybrid-Non-Bilinear) is faster than all other approaches. Notably, by providing a reasonable initialization of camera positions and 3D points, MGSfM achieves significantly higher efficiency than the “Only-Track-Non-Bilinear” method, underscoring the importance of initialization.

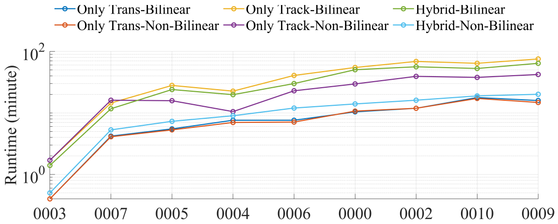

Figure 3. Runtime comparison (log scale) of six ablation study methods on the KITTI-360 dataset. Scenes are sorted by the ascending runtime of Hybrid-Non-Bilinear to facilitate visualization.

E. Qualitative Results on KITTI-360

We compare the reconstruction results from several state-of-the-art SfM methods on KITTI-360 benchmark [4], including COLMAP [6], GLOMAP [5], MMA [1], MCSfM [2] and our proposed MGSfM, as shown in Fig. 4.

F. Test on Indoor ETH3D-SLAM.

As shown in table below, MGSfM achieves the best accuracy and efficiency (T in seconds) of translation averaging.

Data		GLOMAP			DMRA+MGP			MGSfM		
Name	N	AUC@		T	AUC@		T	AUC@		T
		0.1m	0.5m		0.1m	0.5m		0.1m	0.5m	
ceiling_1	3190	18.5	66.2	240	32.8	78.0	154	**59.7**	**87.4**	**34**
desk_3	4132	86.4	97.3	462	95.3	99.1	467	**95.8**	**99.2**	**89**
large_loop_1	3022	70.0	92.7	250	81.1	96.0	166	**87.9**	**97.6**	**30**
motion_1	4732	25.7	70.3	885	22.4	70.3	678	**46.5**	**87.6**	**158**
reflective_1	9202	79.1	95.7	3239	87.5	96.3	670	**91.3**	**97.2**	**335**
repetitive	3932	66.9	93.2	90	83.0	96.2	104	**91.4**	**97.8**	**23**

G. Results on Self-Collected Datasets

We compare our method with several state-of-the-art SfM methods, including COLMAP [6], GLOMAP [5], and MCSfM [2], on self-collected datasets. The CAMPUS dataset comprises more than 29,000 images covering an area of approximately 520,000 m^2. Qualitative results for the CAMPUS dataset are presented in Fig. 5. The runtime of MGSfM is 66 minutes, compared to approximately 1588 minutes for COLMAP, 580 minutes for GLOMAP, 401 minutes for MCSfM and 51 minutes for MMA. Due to the lack of multi-camera constraints in COLMAP and GLOMAP, both methods are sensitive to outlier feature matches. Although MCSfM incorporates multi-camera constraints, its inherent error accumulation limits its accuracy.

The STREET dataset, which is captured by a four-camera system, comprises more than 12,000 images covering an area of approximately 500,000 m^2. Qualitative results for the STREET dataset are shown in Fig. 6. Only MGSfM accurately reconstructs the tracks of the roads.

References

[1] Hainan Cui and Shuhan Shen. MMA: Multi-camera based global motion averaging. In *AAAI Conference on Artificial Intelligence*, pages 490–498, 2022. 1, 2, 4, 5, 6

[2] Hainan Cui, Xiang Gao, and Shuhan Shen. MCSfM: Multi-camera-based incremental structure-from-motion. *IEEE Transactions on Image Processing*, 32:6441–6456, 2023. 1, 2, 4, 5, 6

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research (IJRR)*, 32(11): 1231–1237, 2013. 1

[4] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(3):3292–3310, 2022. 2, 4

[5] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, pages 58–77. Springer, 2024. 1, 2, 4, 5, 6

[6] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 1, 2, 4, 5, 6

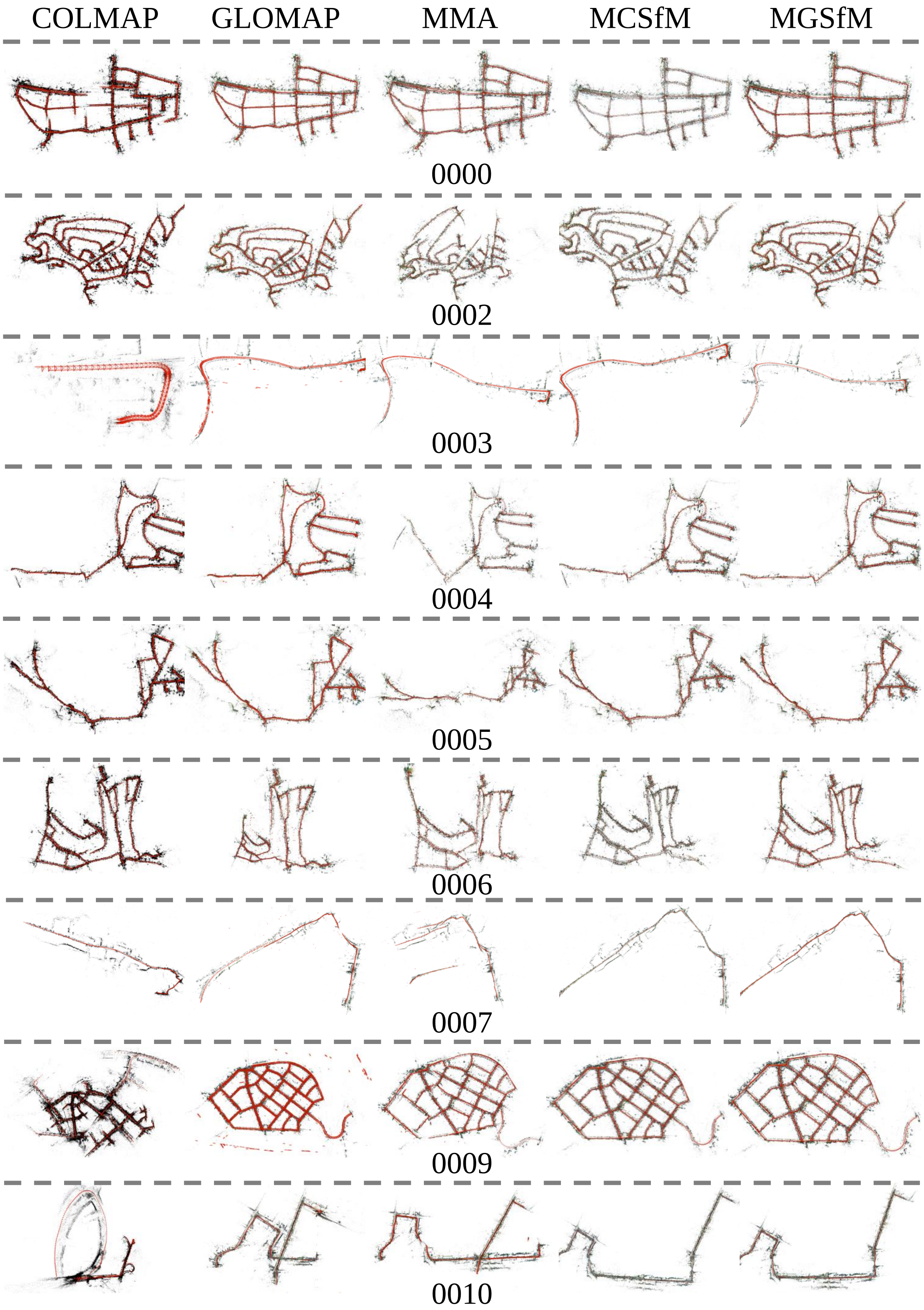

Figure 4. Comparison of reconstruction results on KITTI-360 benchmark [4]. From the qualitative comparison, our method demonstrates superior robustness compared to COLMAP [6], GLOMAP [5] and MMA [1]. As highlighted in our main manuscript, our system achieves a computational efficiency approximately 10 times faster than that of MCSfM [2].

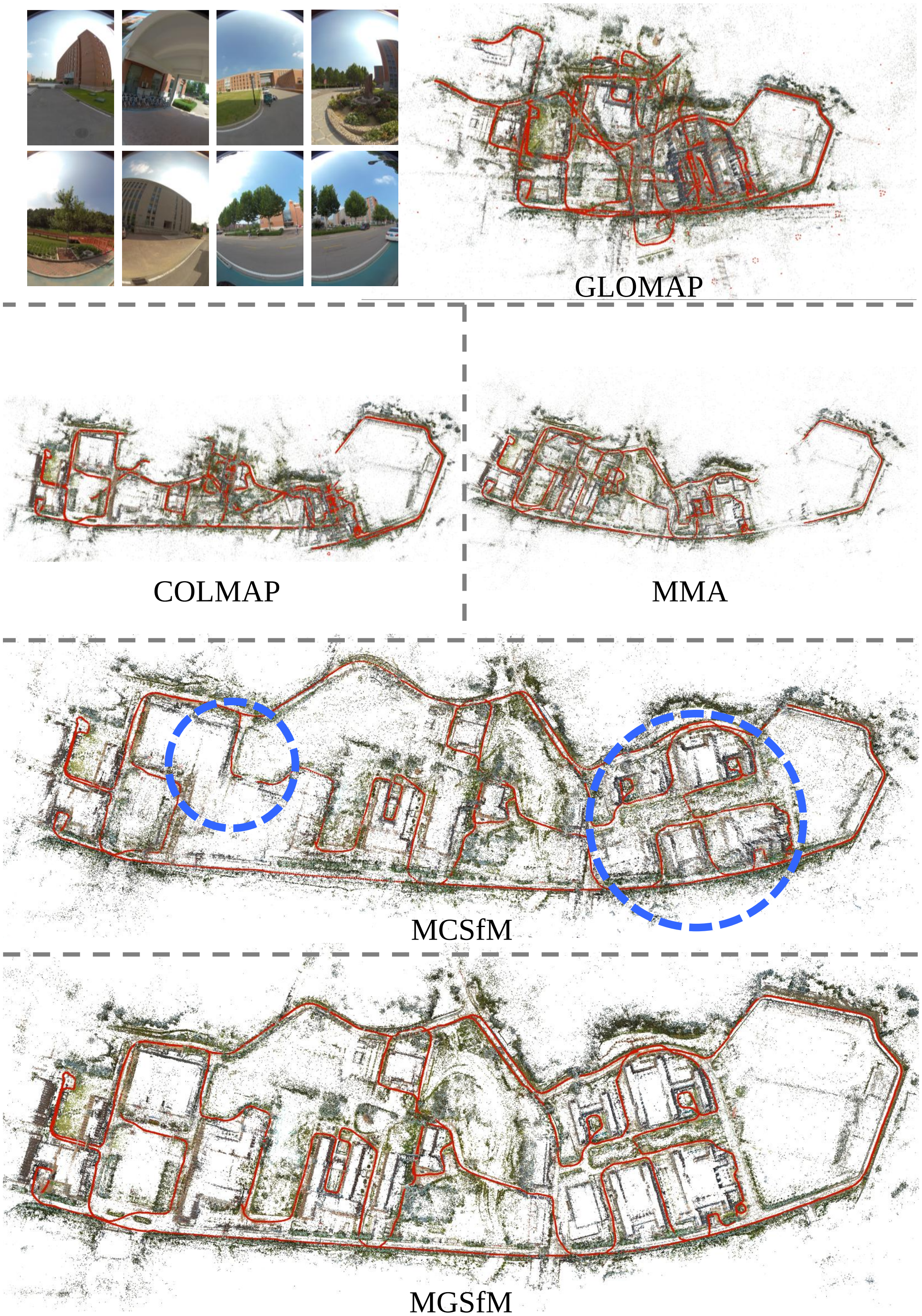

Figure 5. Comparison of reconstruction results on the self-collected CAMPUS dataset. The state-of-the-art SfM methods compared include COLMAP [6], GLOMAP [5], MCSfM [2], MMA [1] and our proposed MGSfM. The results produced by COLMAP, GLOMAP and MMA are wrong. For the result produced by MCSfM, the area enclosed by the blue circle indicates an incorrect reconstruction structure.

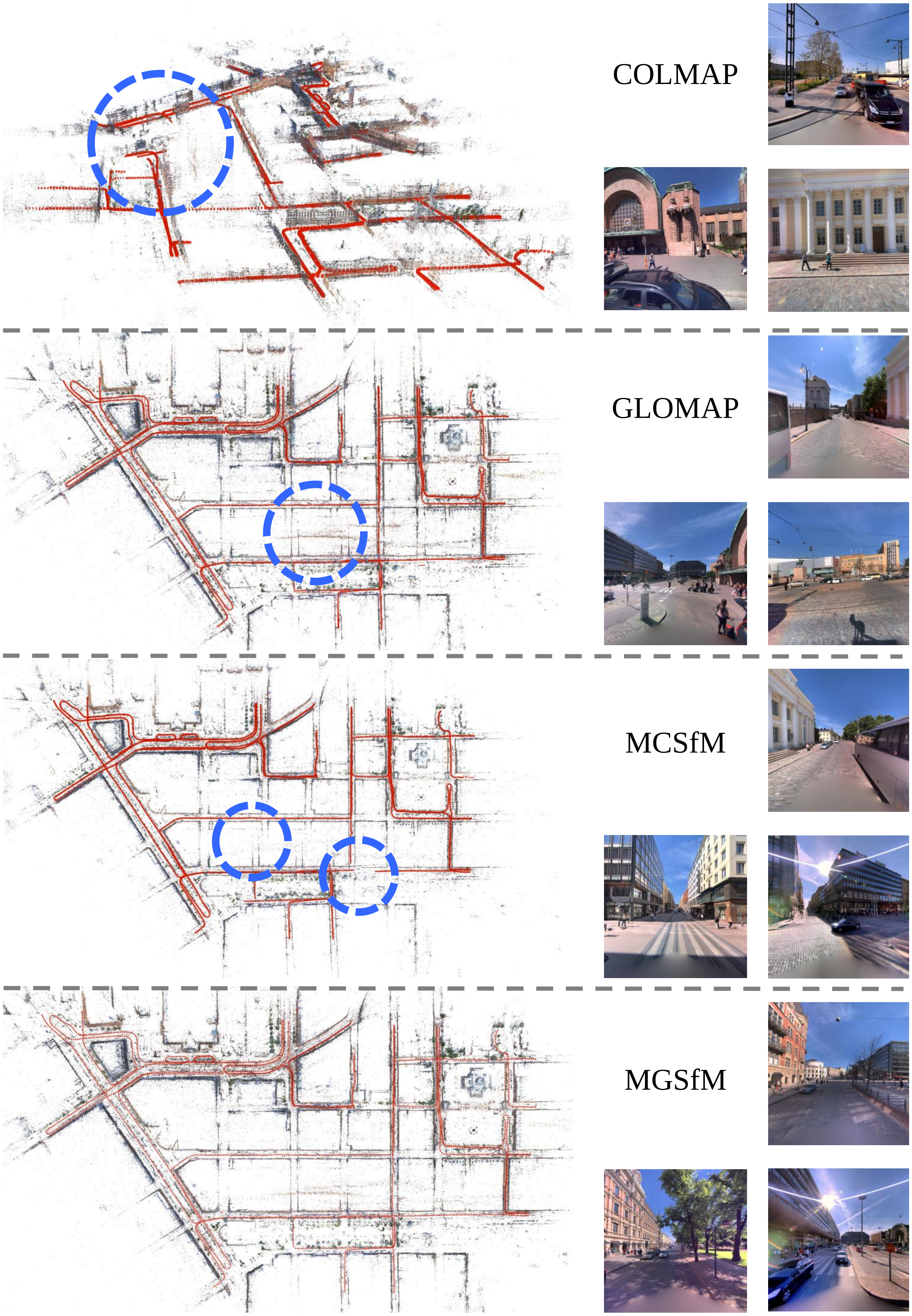

Figure 6. Comparison of reconstruction results on the self-collected STREET dataset. Some sample images are shown near the name of the compared method. The state-of-the-art SfM methods compared include COLMAP [6], GLOMAP [5], MCSfM [2], MMA [1], and our proposed MGSfM. The area enclosed by the blue circle indicates an incorrect scene structure in the reconstruction.