Learning Multi-view Camera Relocalization with Graph Neural Networks
Supplementary Material

Fei Xue, Xin Wu, Shaojun Cai, Junqiu Wang

1. Implementation

Training Sequence. For the 7Scenes [8] and Cambridge datasets [6], we adopt the officially released training sequences to train our networks. While for the RobotCar benchmark [7], we follow the train/test split utilized in MapNet [1] and LsG [9]. The training and testing sequences along with their descriptions are demonstrated in Table 1. Although VidLoc [2] and ADPoseNet [3] also report results on the Oxford RobotCar dataset [7], they use the different sequences for training and testing. Therefore, we compare our method against PoseNet [6], MapNet [1], and LsG [9] on this dataset.

Network. The structure details of our GNN blocks for pose estimation, message generation, and node updating are shown in Fig. 1, Fig. 2, and Fig. 3, respectively. In order to simultaneously process the topological relationships of multiple frames and extract valuable visual clues from these images, we retain the spatial connections of image features, enabling the information of different frames to be propagated along edges effectively in the formulation of 3D tensors. Additionally, such design enhances the stability of GNNs by discarding fully-connected layers in the regular framework and facilitates the cooperation between GNNs and CNNs in processing unstructured inputs with high dimensions.

	Sequence	Tag	Train	Test
–	2014-06-26-08-53-56	overcast	✓	
–	2014-06-26-09-24-58	overcast	✓	
LOOP1	2014-06-23-15-41-25	sun		✓
LOOP2	2014-06-23-15-36-04	sun		✓
–	2014-11-28-12-07-13	overcast	✓	
–	2014-12-02-15-30-08	overcast	✓	
FULL1	2014-12-09-13-21-02	overcast		✓
FULL2	2014-12-12-10-45-15	overcast		✓

Table 1: Sequences and their descriptions on the Oxford RobotCar dataset [7]. We adopt the same train/test split as PoseNet [6, 4, 5], MapNet [1], and LsG [9].

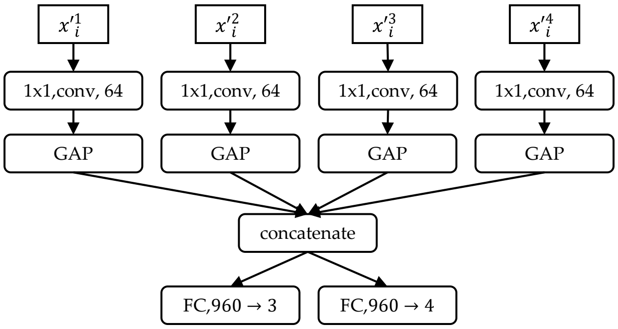


Figure 1: Pose estimator. Features at four levels are first passed through the 1x1 convolutional layer and GAP (global pooling layer) and then concatenated along the channel dimension and finally fed into two FC (fully-connected) layers to regress position and orientation, respectively.

2. Experimental Results

2.1. Influence of Sequence Length

The number of frames N_v have direct influence on both the accuracy and timing. Since images are processed highly parallel in CNNs and GNNs, our model achieves 130, 110, 100fps for 8, 10, and 15 frames on an 1080TI GPU. We also find the accuracy goes higher as N_v increases since more information is introduced to each view. While it has an upper bound as enough overlapped content is necessary. Studying the influence of N_v on both accuracy and timing is really interesting and deserves further exploration.

2.2. Results on the RobotCar Dataset

Robustness to Extremely Challenging Conditions. We further evaluate the performance of our system in handling day-night changes, various weather and season conditions, and dynamic objects (e.g., pedestrians, moving cars, and road works) on the Oxford RobotCar dataset [7] (sam-

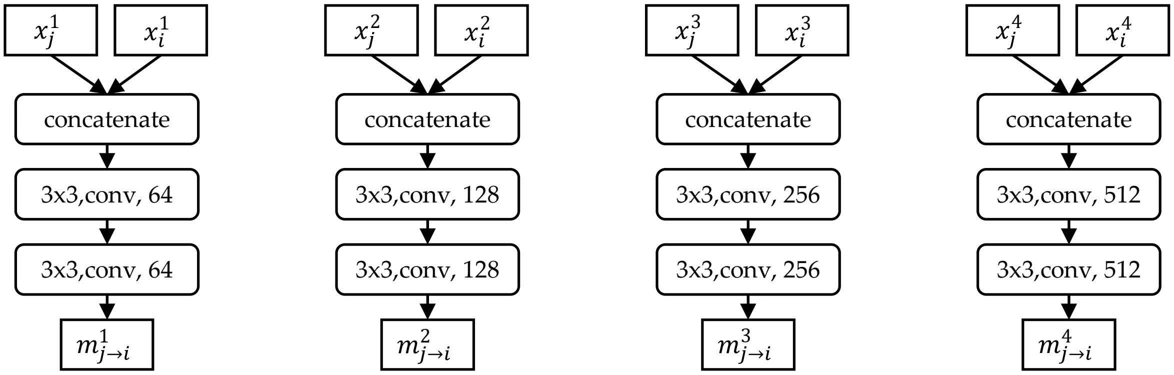


Figure 2: Multi-level message generation. The message generation functions incorporated in GNN blocks share the similar structure with individual parameters.

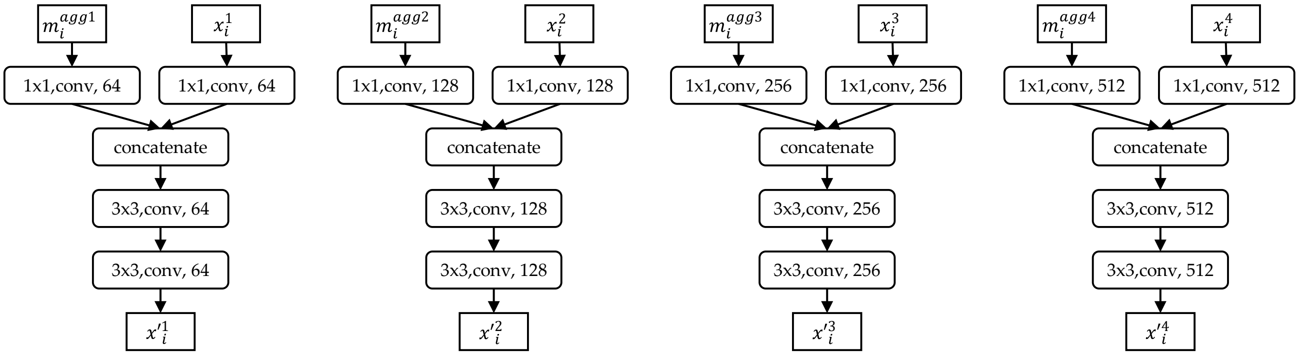
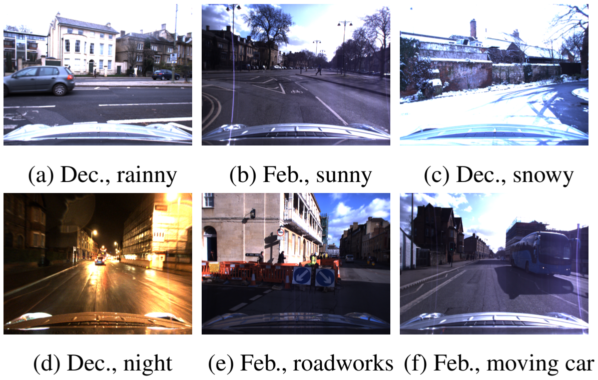


Figure 3: Multi-level node updating. Similar to message generation, the node updating functions in four GNN blocks share the same structures with individual parameters.

ple images can be seen in Fig. 4). Table 2 demonstrates that our network outperforms previous PoseNet [6], MapNet [1], and LsG [9] consistently. Fig. 5 and Fig. 6 show the accumulative position and orientation errors. The significant improvements suggest that our graph modeling exploits the benefit of multiple frames more effectively in enhancing relocalization accuracy, especially in dealing with challenging conditions.

Attention Maps. We visualize the attention maps of PoseNet, MapNet, LsG, and our model on image samples of the RobotCar dataset. Interestingly, PoseNet, MapNet, and LsG concentrate on small local areas of the road in front of the car. While our model perceives much larger areas especially distributed at regions with rich structures (Fig. 7a, 7c, and 7e). Additionally, our method can effectively deal with dynamic objects (Fig. 7b) and over-exposure (Fig. 7d) by focusing on more meaningful objects (e.g., buildings). Unfortunately, previous algorithms usually fail in these scenarios. We believe that the graph formulation boosts ability of the feature extractor in learning more global representation of the scene.



(a) Dec., rainny (b) Feb., sunny (c) Dec., snowy

(d) Dec., night (e) Feb., roadworks (f) Feb., moving car

Figure 4: Sample images captured under various season, weather, and illumination conditions in the Oxford RobotCar dataset [7].

2.3. Results on the 7Scenes Dataset

Fig. 8, 9, and 10 show the recovered trajectories and orientation error distributions of PoseNet [6, 4, 5], MapNet [1],

Description		Method			
Sequence	Tag	PoseNet [6, 4, 5]	MapNet [1]	LsG [9]	**Ours**
2014-12-05-11-09-10	overcast, rain	104.41m, 20.94°	73.74m, 21.06°	57.54m, 8.49°	**44.20**m, **6.88**°
2015-02-03-08-45-10	snow	125.22m, 21.61°	139.75m, 29.02°	71.42m, 12.92°	**51.42**m, **6.76**°
2015-02-24-12-32-19	roadworks, sun	132.86m, 32.22°	157.64m, 33.88°	81.92m, 16.79°	**60.96**m, **12.41**°
2014-12-17-18-18-43	night, rain	471.89m, 82.11°	430.49m, 85.15°	430.54m, 72.35°	**236.00**m, **44.02**°
Avg	–	208.60m, 34.22°	200.41m, 42.28°	160.36m, 27.64°	**98.15**m, **17.52**°

Table 2: Mean position and orientation errors of PoseNet [6, 4, 5], MapNet [1], LsG [9], and our method on the Oxford RobotCar dataset [7]. These sequences contain day-night changes, different weather and season conditions, and dynamic objects(e.g., pedestrians, moving cars, construction, and roadworks). The best results are highlighted.

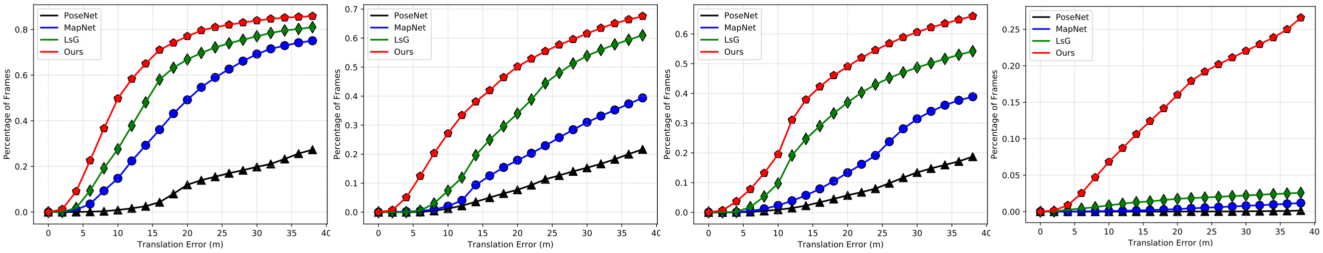


Figure 5: Cumulative distribution of the mean position errors of PoseNet [6], MapNet [1], LsG [9], and our method on the Oxford RobotCar [7] of sequences 2014-12-05-11-09-10, 2015-02-03-08-45-10, 2015-02-24-12-32-19, and 2014-12-17-18-18-43 (from left to right).

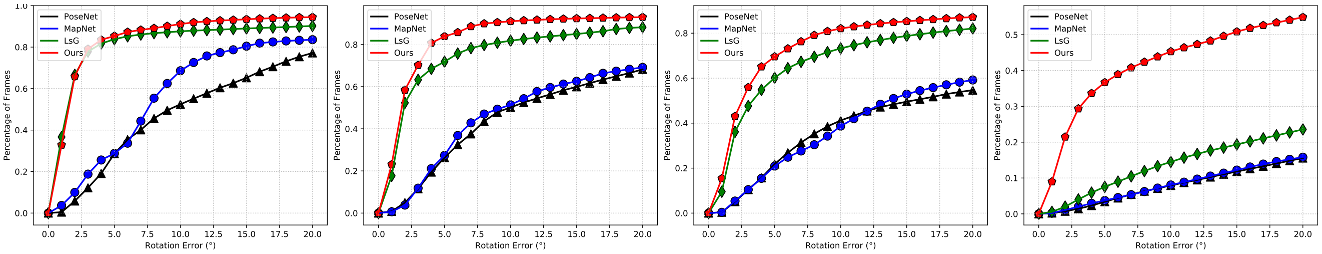


Figure 6: Cumulative distribution of the mean orientation errors of PoseNet [6], MapNet [1], LsG [9], and our method on the Oxford RobotCar [7] of sequences 2014-12-05-11-09-10, 2015-02-03-08-45-10, 2015-02-24-12-32-19, and 2014-12-17-18-18-43 (from left to right).

and our model on the 7Scenes dataset [8]. Samples from the test sequences are also plotted to demonstrate the challenging conditions of the dataset. From Fig. 8, 9, and 10, we can see that our model produces much more accurate trajectories and orientations than PoseNet owing to the advantages of multiple images in reducing visual ambiguities. Besides, our network yields much better orientations on most of the test sequences than MapNet because of proposed graph modeling for information propagation among different views.

References

[1] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. MapNet: Geometry-aware Learning of Maps for Camera Localization. In *CVPR*, 2018.

[2] Ronald Clark, Sen Wang, Andrew Markham, Niki Trigoni, and Hongkai Wen. Vidloc: A Deep Spatio-temporal Model for 6-DOF Video-clip Relocalization. In *CVPR*, 2017.

[3] Zhaoyang Huang, Yan Xu, Jianping Shi, Xiaowei Zhou, Hujun Bao, and Guofeng Zhang. Prior Guided Dropout for Robust Visual Localization in Dynamic Environments. In *ICCV*, 2019.

[4] Alex Kendall and Roberto Cipolla. Modelling Uncertainty in Deep Learning for Camera Relocalization. In *ICRA*, 2016.

[5] Alex Kendall and Roberto Cipolla. Geometric Loss Functions for Camera Pose Regression with Deep Learning. In *CVPR*, 2017.

[6] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A Convolutional Network for Real-time 6-DoF Camera Relocalization. In *ICCV*, 2015.

[7] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *IJRR*,

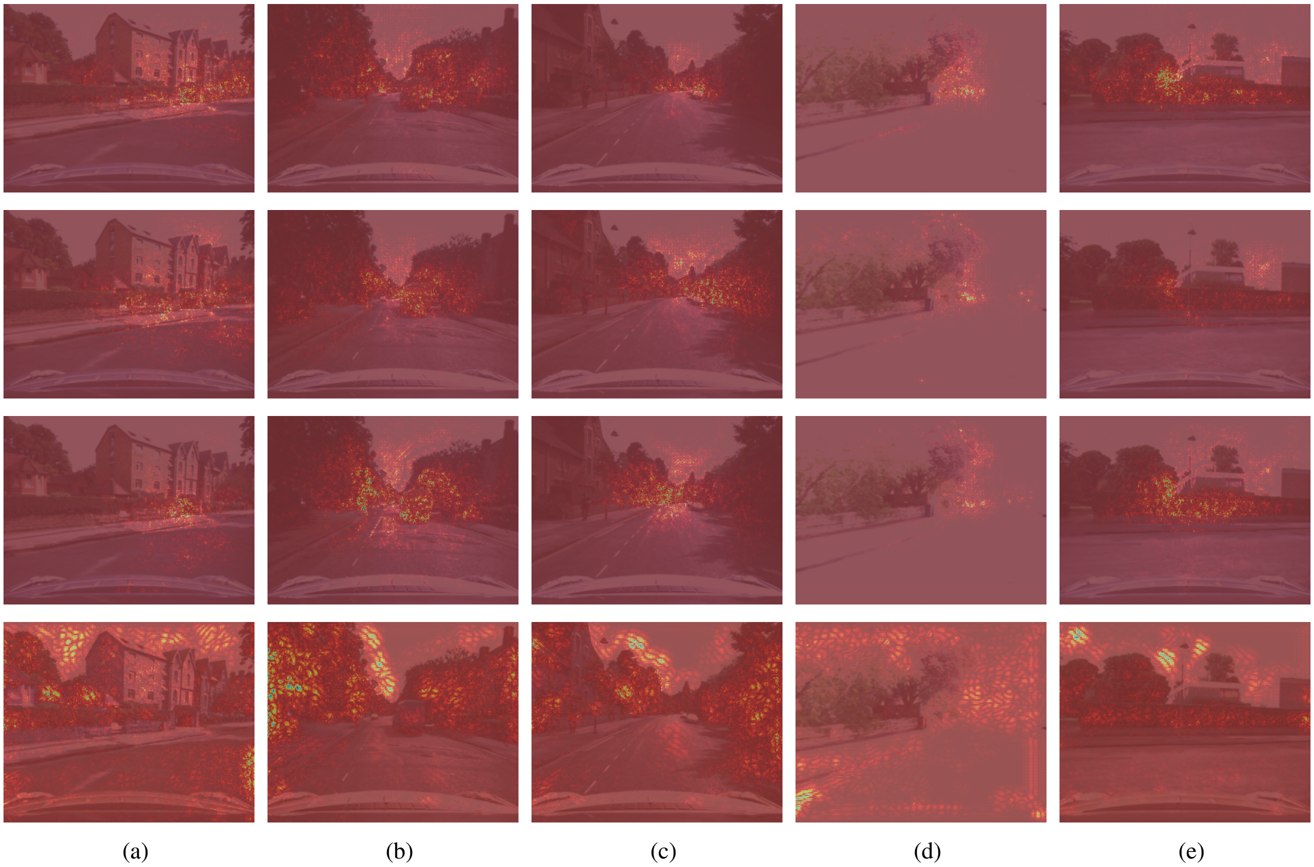


Figure 7: Attention maps for PoseNet [6, 4, 5] (1st row), MapNet [1] (2nd row), LsG [9] (3rd row), and our model (4th row) on the Oxford RobotCar dataset [7]. Compared with PoseNet, MapNet, and LsG, which concentrate on a local area in front of the car, our method focus mainly on global structure of the whole scene, leading to larger perception areas. Moreover, our model performs more effectively in handling challenging conditions, e.g., dynamic objects and over-exposure.

2017.

[8] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In *CVPR*, 2013.

[9] Fei Xue, Xin Wang, Zike Yan, Qiuyuan Wang, Junqiu Wang, and Hongbin Zha. Local Supports Global: Deep Camera Relocalization with Sequence Enhancement. In *ICCV*, 2019.

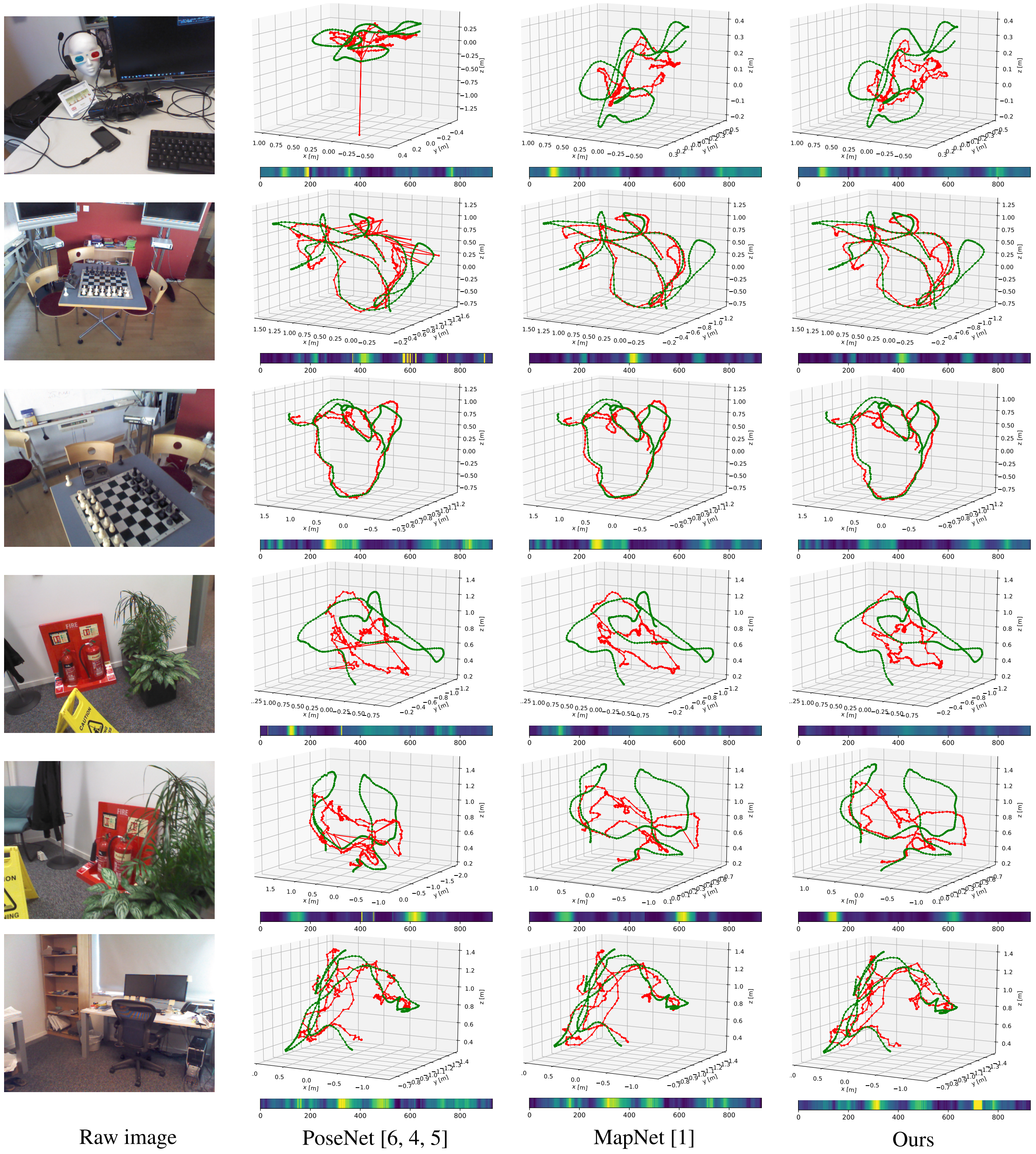


Figure 8: Raw images and recovered trajectories (red line) of PoseNet [6, 4, 5], MapNet [1], and our method on the 7Scenes dataset [8]. The green lines are the ground-truth trajectories. These sequences (from top to bottom) are heads-01, chess-03, chess-05, fire-03, fire-04, and office-02.

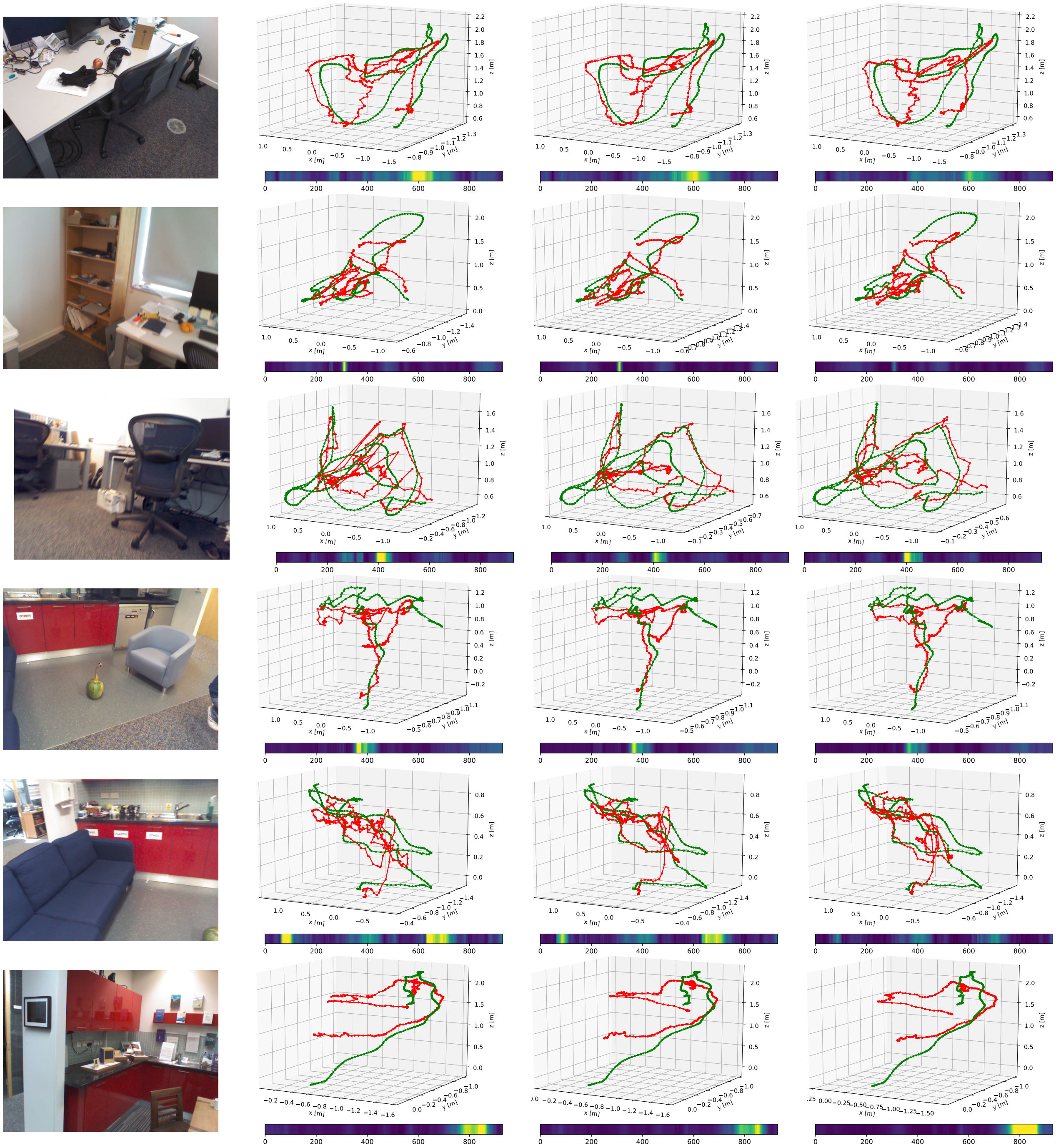


Figure 9: Raw images and recovered trajectories (red line) of PoseNet [6, 4, 5], MapNet [1], and our method. The green lines are the ground-truth trajectories. These sequences (from top to bottom) are office-06, office-07, office-09, pumpkin-01, pumpkin-07, and redkitchen-03.

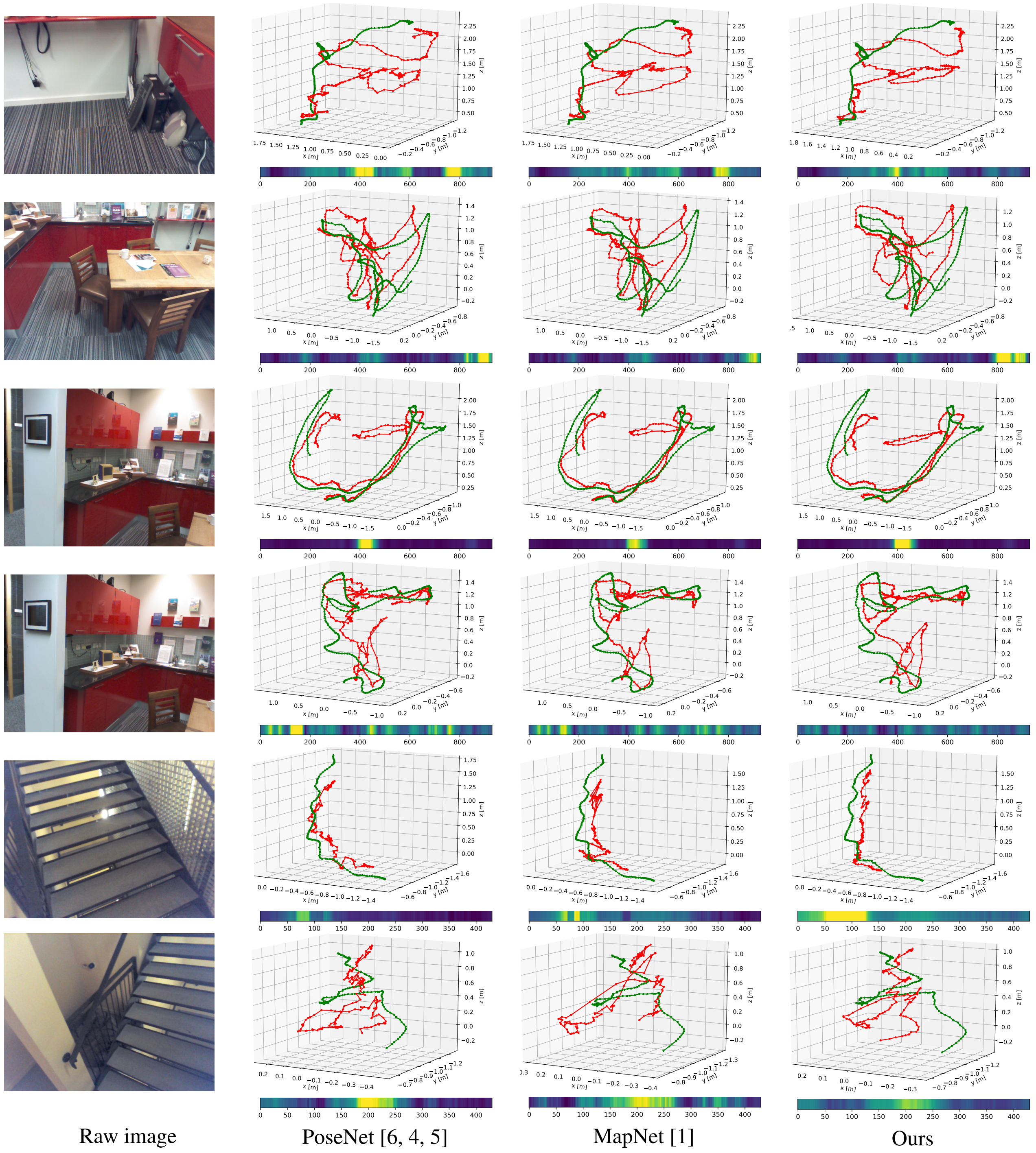


Figure 10: Raw images and recovered trajectories (red line) of PoseNet [6, 4, 5], MapNet [1], and our method. The green lines are the ground-truth trajectories. These sequences (from top to bottom) are redkitchen-04, redkitchen-06, redkitchen-12, redkitchen-14, stairs-01, and stairs-04.