Local Supports Global:
Deep Camera Relocalization with Sequence Enhancement
Supplementary Material

Fei Xue[1,2], Xin Wang[2,3], Zike Yan[2,3], Qiuyuan Wang[2,3], Junqiu Wang[4], and Hongbin Zha[2,3]

[1]UISEE Technology Inc.
[2]Key Laboratory of Machine Perception, Peking University
[3]PKU-SenseTime Machine Vision Joint Lab, Peking University
[4]Beijing Changcheng Aviation Measurement and Control Institute, AVIC
{feixue, xinwang_cis, zike.yan, wangqiuyuan}@pku.edu.cn
jerywangjq@foxmail.com, zha@cis.pku.edu.cn

1. Introduction

In the supplementary material, we first introduce the training and testing sequences adopted on the Oxford RobotCar dataset [5] in Table 1. The descriptions of corresponding sequences are also included.

Moreover, we perform an ablation study in Sec. 2. In Sec. 3, additional comparisons against previous methods on various challenging scenes of the Oxford RobotCar dataset are presented. We visualize the attention maps produced by PoseNet [2–4], MapNet [1] and our model in Sec. 4

2. Ablation Study

We perform an ablation study to evaluate the effectiveness of each part of our architecture in Table 2. Our basic model achieves the poorest performance because both the content augmentation and motion constraints are disabled. Both the translation and orientation errors are reduced by introducing the content augmentation. The performance is further enhanced by adding the motion constraints.

3. Robustness to Various Conditions

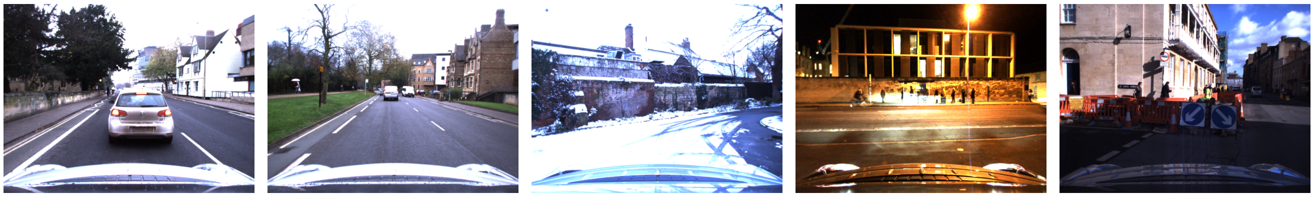
Since practical visual localization approaches need to deal with various conditions, we additionally compare the robustness of PoseNet [2–4], MapNet [1] and our method in handling situations including weather and seasonal variations, as well as day-night changes. Samples of images can be seen in Fig. 1. As Oxford RobotCar dataset [5] provides sequences fulfilling these conditions, we test the generalization ability of three models quantitatively and qualitatively. It's worthy to note that our model is trained on the same sequences as PoseNet and MapNet, without any fine-tuning.

Label	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: Training, testing sequences and corresponding descriptions on the Oxford RobotCar dataset [5]. We adopt the same train/test split as PoseNet [2–4] and MapNet [1].

3.1. Quantitative Comparison

Table 3 presents the adopted sequences, descriptions and quantitative results. Performances of both PoseNet and MapNet degrade a lot due to the challenging changes. We notice that MapNet achieves very close results with PoseNet. The possible reason is that MapNet takes only singe images to regress global camera poses, as PoseNet. Though MapNet employs motion constraints over several frames during training, motion constraints are incapable of mitigating visual ambiguities existing in challenging scenarios. In contrast, our content augmentation strategy copes with these problems effectively (see attention maps



(a) December, overcast (b) November, rain (c) February, snow (d) December, night (e) February, roadworks

Figure 1: Samples of images captured under different weather, season, illumination, and roadwork conditions on the Oxford RobotCar dataset [5].

	Method		
Scene	Ours (basic)	Ours (w/ content)	Ours (full)
LOOP1	19.39m, 7.56°	9.48m, 4.23°	**9.07**m, **3.31**°
LOOP2	21.07m, 9.42°	10.56m, 4.37°	**9.19**m, **3.53**°
FULL1	108.13m, 19.49°	59.83m, 10.97°	**31.65**m, **4.51**°
FULL2	109.73m, 19.01°	85.98m, 11.93°	**53.45**m, **8.60**°
Avg	64.58m, 13.87°	41.46m, 7.88°	**25.84**m, **4.99**°

Table 2: Mean translation and rotation errors of variations of our model on the Oxford RobotCar dataset [5]. **Ours (basic)** indicates the model without content augmentation and motion constraints. **Ours (w/ content)** indicates the model with content augmentation but without motion constraints. **Ours (full)** contains both the content augmentation and motion constraints. The best results are highlighted.

in Fig. 8) and the pose uncertainties are further alleviated by the motion constraints.

3.2. Qualitative Comparison

Fig. 2, 3, 4, 5 show the trajectories of PoseNet, MapNet and our model. To better visualize the comparison against PoseNet and MapNet, we plot the estimated poses of frames with translation errors within the range of 50m and 100m, respectively. Cumulative translation and rotation distribution errors of three models are illustrated as well. As can be seen obviously, our method gives much more accurate translation and rotation predictions, especially in scenes where PoseNet and MapNet produce lots of outliers.

3.3. Failure Cases

As shown in Fig. 6 and 7, all the three methods behave poorly in sequences captured at night (FULL7 and FULL8), although our methods gives better performance. The major reason is that pixel values are changed too much between day and night. The performance can be improved by finetuning on data with similar conditions. Semantic information can also be introduced in the future.

4. Feature Visualization

In Fig. 8, we visualize the attention maps of images for PoseNet, MapNet and our model. We observe that PoseNet and MapNet rely heavily on local regions including dynamic objects such cars (Fig. 8a, 8c, 8d). Even the front part of the moving car, which is used for data collection, is covered in the salient maps produced by PoseNet (Fig. 8c, 8d, 8e). These regions are either easily changed over time or sensitive to similar appearances, leading to severe localization uncertainties. In contrast, our model emphasizes on stable features such as buildings (Fig. 8b) and roads (Fig. 8c, 8d). Moreover, both local and global regions (Fig. 8b, 8e) are included. Benefited from the content augmentation, unstable features are suppressed while robust features are advocated.

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

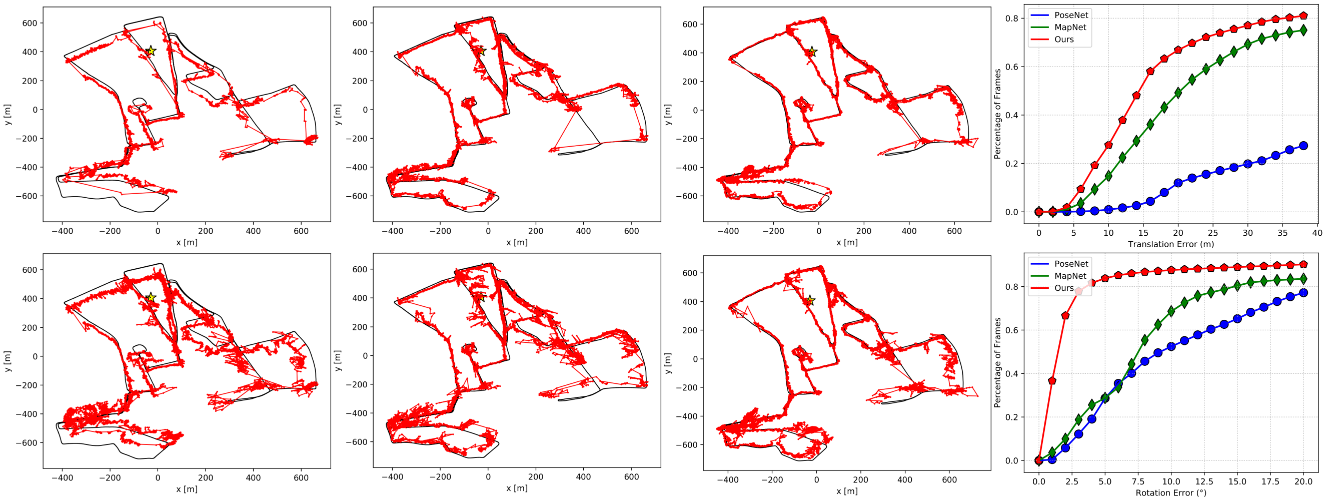
[3] Alex Kendall and Roberto Cipolla. Geometric Loss Functions for Camera Pose Regression with Deep Learning. In *CVPR*, 2017.

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

[5] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *IJRR*, 2017.

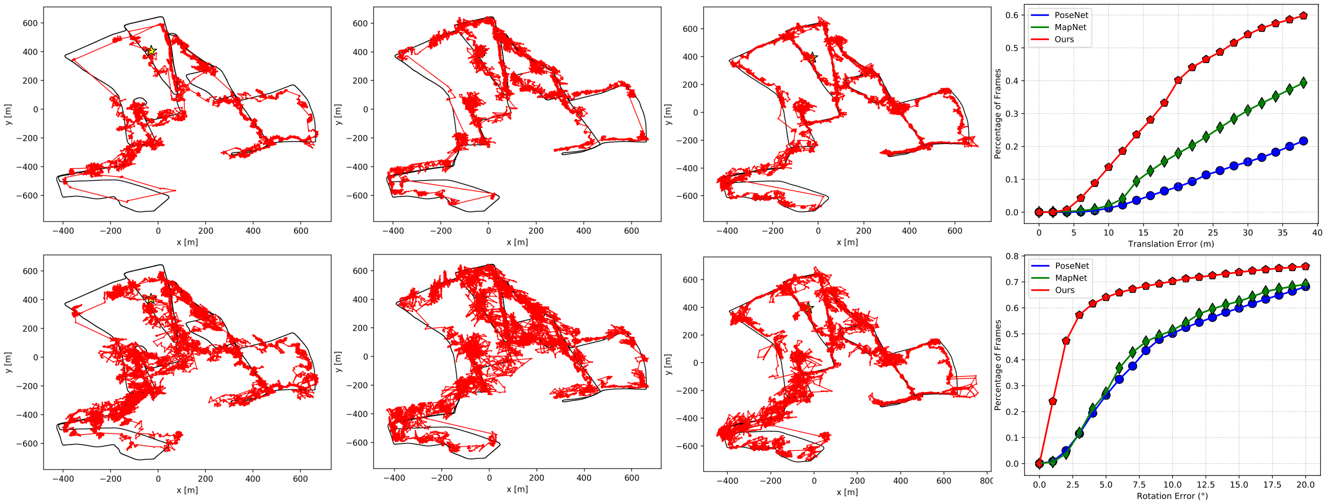
Description			Method		
Scene	Sequence	Tag	PoseNet [2–4]	MapNet [1]	**Ours**
FULL3	2014-12-05-11-09-10	overcast, rain	104.41m, 20.94°	73.74m, 21.06°	**57.54**m, **8.49**°
FULL4	2014-11-25-09-18-32	overcast, rain	151.24m, 34.70°	166.70m, 35.62°	**137.53**m, **23.23**°
FULL5	2015-02-03-08-45-10	snow	125.22m, 21.61°	139.75m, 29.02°	**71.42**m, **12.92**°
FULL6	2015-02-24-12-32-19	roadworks, sun	132.86m, 32.22°	157.64m, 33.88°	**81.92**m, **16.79**°
FULL7	2014-12-10-18-10-50	night	405.17m, 75.64°	397.80m, 81.40°	**385.58**m, **68.81**°
FULL8	2014-12-17-18-18-43	night, rain	471.89m, 82.11°	430.49m, 85.15°	**430.54**m, **72.35**°
Avg	–	–	231.80m, 44.54°	227.69m, 47.69°	**193.98**m, **33.77**°

Table 3: Mean translation and rotation errors of PoseNet [2–4], MapNet [1] and our method on the Oxford RobotCar dataset [5]. Results of PoseNet and MapNet are generated from weights released by [1]. The sequences were captured at different times with day-night changes, as well as weather and seasonal variations. Moreover, changes of traffic, pedestrians, construction and roadworks are also included. The best results are highlighted.



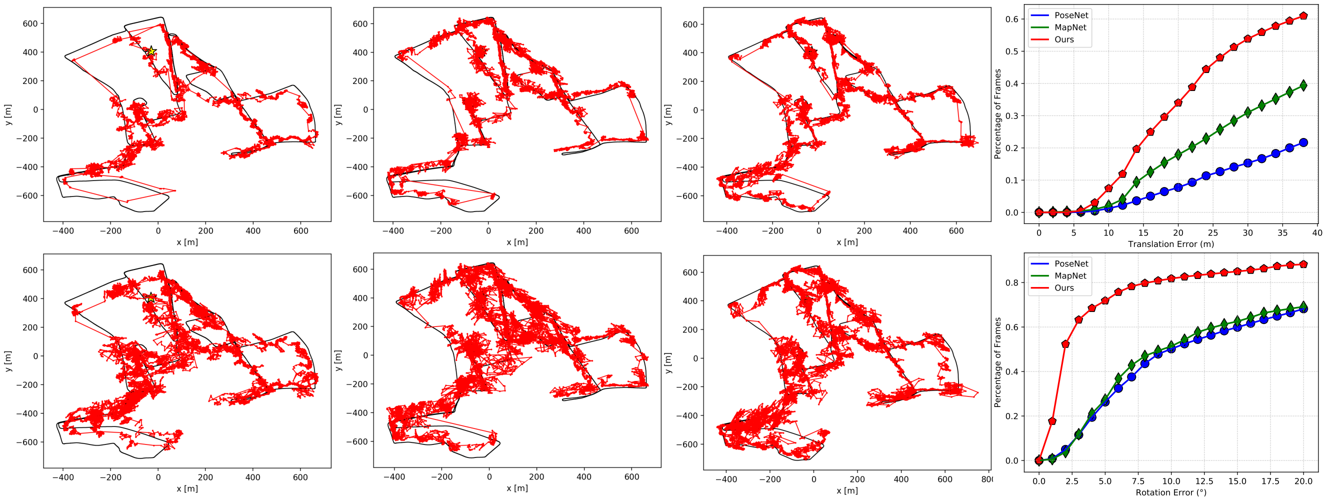
PoseNet [2–4] MapNet [1] Ours Cumulative errors

Figure 2: Results of PoseNet, MapNet and our model on the FULL3 scenes of the Oxford RobotCar dataset [5]. The red and black lines indicate predicted and ground truth poses respectively. The star represents the start point. The poses of PoseNet and MapNet are from [1]. To better visualize the trajectories, we select points with translation errors within 50m (top) and 100m (bottom). Cumulative translation (top right) and rotation (bottom right) errors are also illustrated.



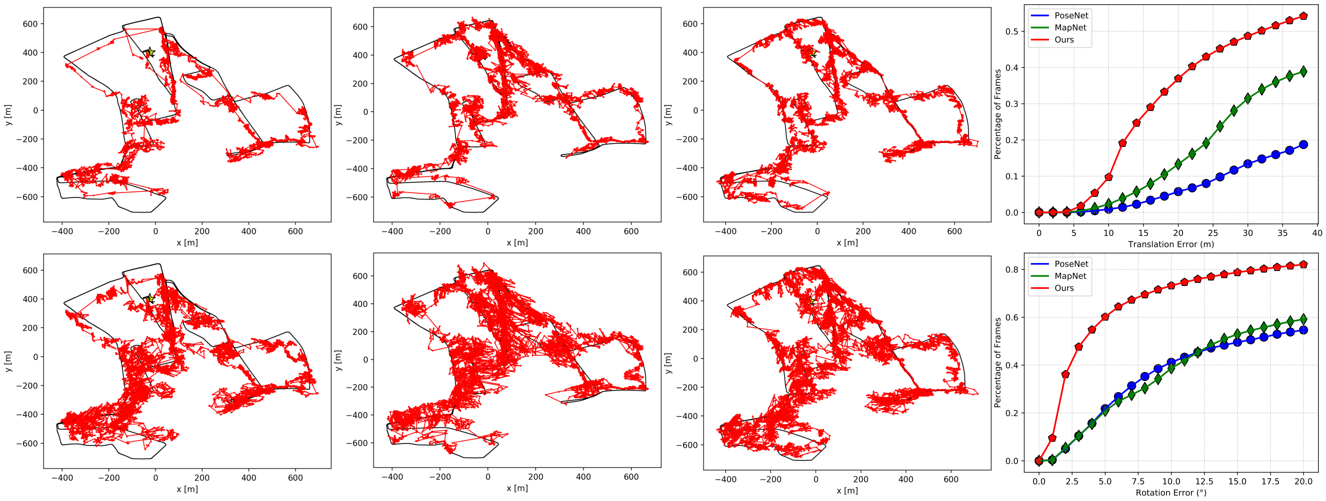
PoseNet [2–4] MapNet [1] Ours Cumulative errors

Figure 3: Results of PoseNet, MapNet and our model on the FULL4 scenes of the Oxford RobotCar dataset [5]. The red and black lines indicate predicted and ground truth poses respectively. The star represents the start point. The poses of PoseNet and MapNet are from [1]. To better visualize the trajectories, we select points with translation errors within 50m (top) and 100m (bottom). Cumulative translation (top right) and rotation (bottom right) errors are also illustrated.



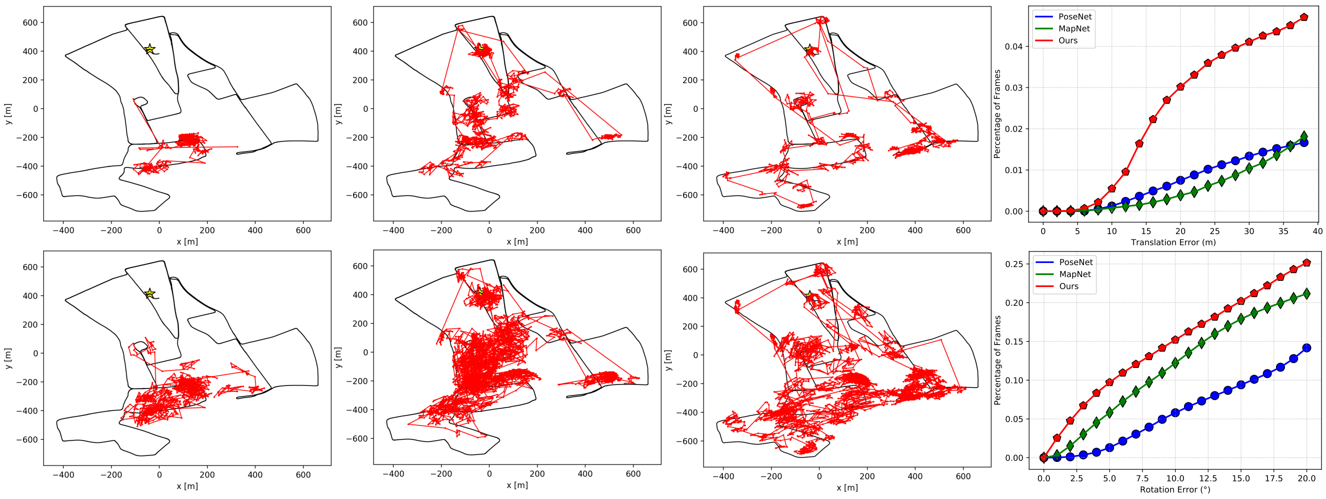
PoseNet [2–4] MapNet [1] Ours Cumulative errors

Figure 4: Results of PoseNet, MapNet and our model on the FULL5 scenes of the Oxford RobotCar dataset [5]. The red and black lines indicate predicted and ground truth poses respectively. The star represents the start point. The poses of PoseNet and MapNet are from [1]. To better visualize the trajectories, we select points with translation errors within 50m (top) and 100m (bottom). Cumulative translation (top right) and rotation (bottom right) errors are also illustrated.



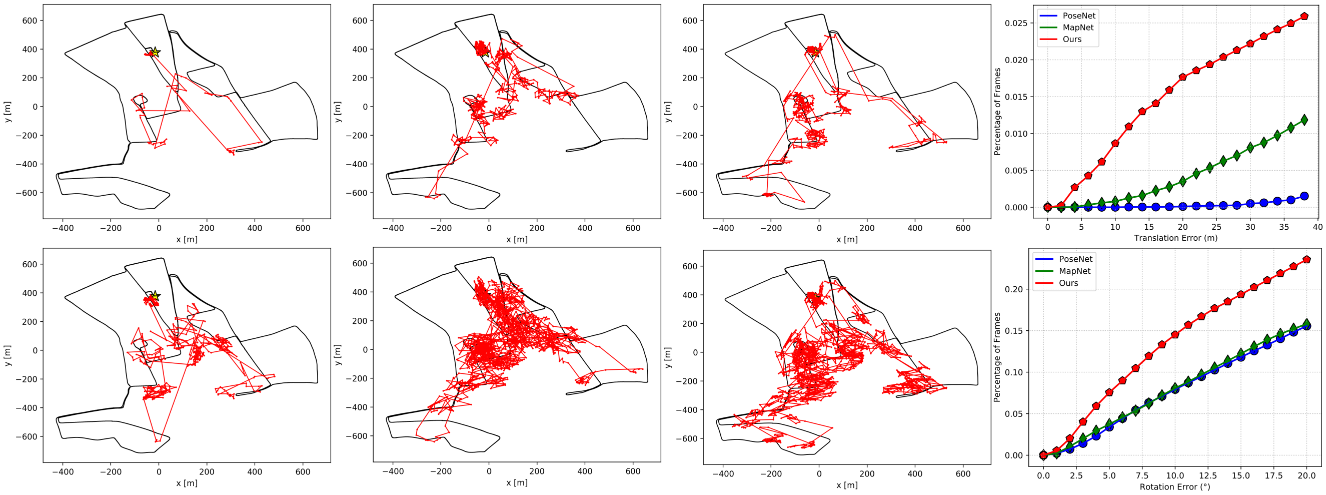
PoseNet [2–4] MapNet [1] Ours Cumulative errors

Figure 5: Results of PoseNet, MapNet and our model on the FULL6 scenes of the Oxford RobotCar dataset [5]. The red and black lines indicate predicted and ground truth poses respectively. The star represents the start point. The poses of PoseNet and MapNet are from [1]. To better visualize the trajectories, we select points with translation errors within 50m (top) and 100m (bottom). Cumulative translation (top right) and rotation (bottom right) errors are also illustrated.



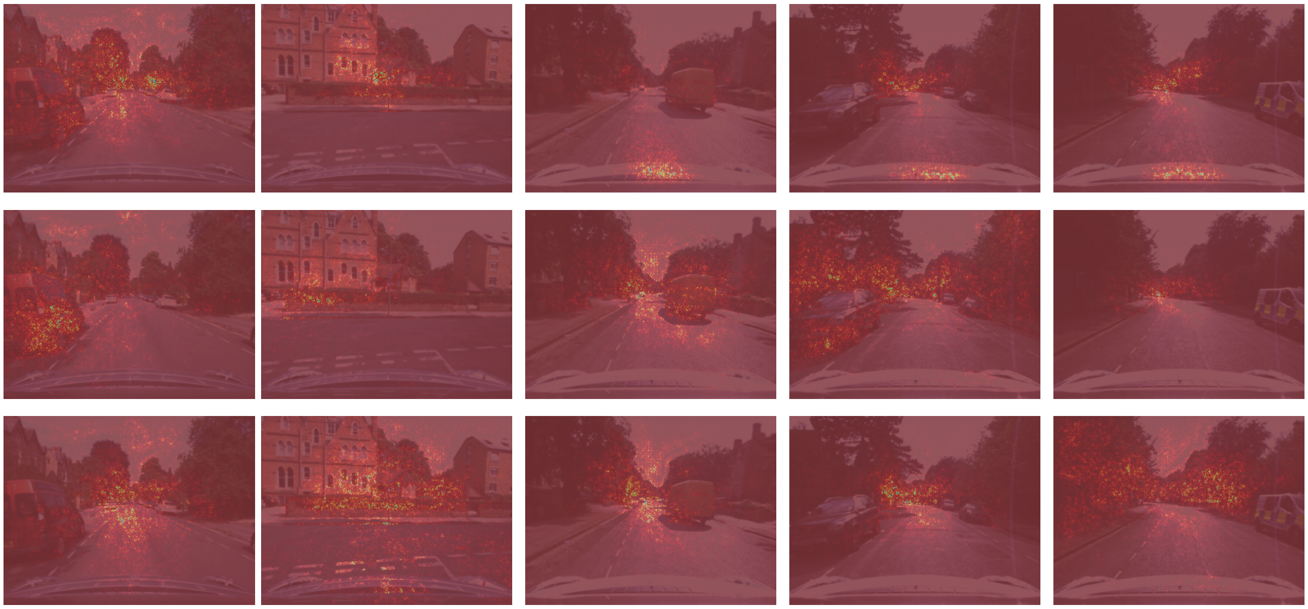
PoseNet [2–4] MapNet [1] Ours Cumulative errors

Figure 6: Results of PoseNet, MapNet and our model on the FULL7 scenes of the Oxford RobotCar dataset [5]. The red and black lines indicate predicted and ground truth poses respectively. The star represents the start point. The poses of PoseNet and MapNet are from [1]. To better visualize the trajectories, we select points with translation errors within 50m (top) and 100m (bottom). Cumulative translation (top right) and rotation (bottom right) errors are also illustrated.



PoseNet [2–4] MapNet [1] Ours Cumulative errors

Figure 7: Results of PoseNet, MapNet and our model on the FULL8 scenes of the Oxford RobotCar dataset [5]. The red and black lines indicate predicted and ground truth poses respectively. The star represents the start point. The poses of PoseNet and MapNet are from [1]. To better visualize the trajectories, we select points with translation errors within 50m (top) and 100m (bottom). Cumulative translation (top right) and rotation (bottom right) errors are also illustrated.



(a) (b) (c) (d) (e)

Figure 8: Attention maps of example images for PoseNet [2–4] (top), MapNet [1] (middle) and our model (bottom) on the Oxford RobotCar dataset [5]. Compared with PoseNet and MapNet, our model focuses more on static objects and regions with geometric meanings. Both local and global are concentrated in our method to mitigate local similar appearances.