DiffLoc: Diffusion Model for Outdoor LiDAR Localization —Supplementary Material—

Wen Li[1] Yuyang Yang[1] Shangshu Yu[2] Guosheng Hu[3] Chenglu Wen[1] Ming Cheng[1] Cheng Wang[1*]
[1] Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University
[2] School of Computer Science and Engineering, Nanyang Technological University [3] Oosto

In this supplementary, we first describe more details of datasets (Sec. 1). Then, we describe the network architecture (Sec. 2). We further provide additional results (Sec. 3). Finally, we show more visualizations on the Oxford and NCLT datasets (Sec. 4).

1. Dataset Details

We evaluate the proposed DiffLoc for LiDAR localization on two large-scale outdoor benchmark datasets: Oxford Radar RobotCar [2] (Oxford) and NCLT [7] datasets.

The Oxford dataset is collected by sensors on an autonomous-capable Nissan LEAF platform, that contains over 32 repetitions traversals of a center Oxford route (about 10km, 200hm^2). The point cloud is scanned by dual Velodyne HDL-32E LiDAR. In this paper, we only use the point cloud from the left LiDAR.

The NCLT dataset is collected by sensors on a Segway robotic platform on the University of Michigan's North Campus. The dataset contains 27 traversals, where each traversal is nearly 5.5km and covers 45hm^2. The point cloud is scanned by a Velodyne HDL-32E LiDAR.

Both of the datasets are available online at:

- https://oxford-robotics-institute.github.io/radar-robotcar-dataset/
- https://robots.engin.umich.edu/nclt/

For each dataset, we list the corresponding data split as shown in Tab. 1 and Tab. 2. We also visualize the training and test trajectories, as shown in Fig. 1.

Sequence	Length	Tag	Training	Test
11-14-02-26	9.37km	sunny	✓	
14-12-05-52	9.22km	overcast	✓	
14-14-48-55	9.04km	overcast	✓	
18-15-20-12	9.04km	overcast	✓	
15-13-06-37	8.85km	overcast		✓
17-13-26-39	9.02km	sunny		✓
17-14-03-00	9.02km	sunny		✓
18-14-14-42	9.04km	overcast		✓

Table 1. Dataset details on the Oxford dataset.

Sequence	Length	Tag	Training	Test
2012-01-22	6.1km	overcast	✓	
2012-02-02	6.2km	sunny	✓	
2012-02-18	6.2km	sunny	✓	
2012-05-11	6.0km	sunny	✓	
2012-02-12	5.8km	sunny		✓
2012-02-19	6.2km	overcast		✓
2012-03-31	6.0km	overcast		✓
2012-05-26	6.3km	sunny		✓

Table 2. Dataset details on the NCLT dataset.

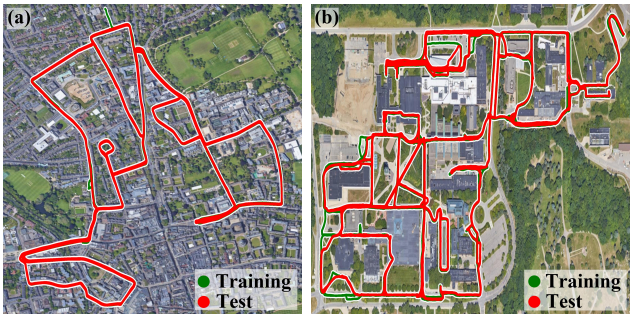

Figure 1. Visualization of the training and test trajectories on the (a) Oxford Radar RobotCar and (b) NCLT datasets.

2. Network Architecture

The detailed architecture of the proposed DiffLoc is illustrated in Fig. 2. Following [1], we use a convolutional stem to replace the patch embedding layer in the standard ViT. Specifically, for the input range image with size $5 \times H \times W$, the first 3 residual layers of the stem increase the dimension to 32. Subsequently, the last residual layers of the stem generate feature maps with the same resolution as the input image but with 64 dimensions. To ensure compatibility with ViT input requirements, we employ average pooling to reduce the size to $[H/P_H] \times [W/P_W]$, followed by a 1×1

*Corresponding author.

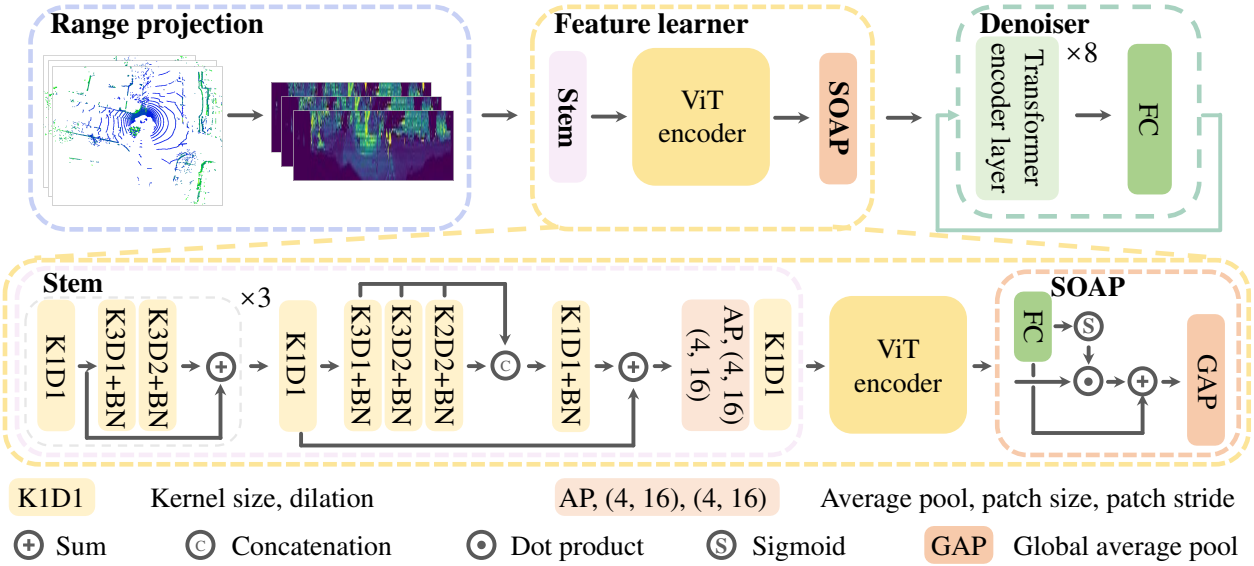

Figure 2. Overview of the proposed framework. The layer elements K, D, and BN represent the kernel size, dilation rate, and batch normalization, respectively. Here, each convolution layer is followed by Leaky Relu as the activation function.

Datasets	Rand	DINO	DINOv2
Oxford (m/°)	3.80/0.88	3.62/0.82	3.53/0.72
NCLT (m/°)	1.39/2.46	1.26/2.35	1.19/2.31

Table 3. Results with different foundation models on the Oxford and NCLT datasets. Rand: randomly initialized.

Metrics	16 × 16	8 × 16	8 × 8	4 × 16	4 × 8
Mean error (m/°)	3.52/0.97	3.66/0.81	3.44/0.86	3.53/0.72	3.57/0.78
Token number	64	128	256	256	512

Table 4. Results with different patch sizes on the Oxford dataset.

convolution layer producing C_{stem} output channels. In this paper, we set the values of $[H, W]$, $[P_H, P_W]$, and C_{stem} to $[32, 512]$, $[4, 16]$, and 384, respectively. As a result, the stem yields 256 visual tokens.

The output is then fed to a ViT, which is used by DINO [8], to get the feature map $F \in \mathbb{R}^{256\times384}$ (without classification token). The proposed static-object-aware pool (SOAP) module, employed to guide the feature reweighting, emphasizes features associated with robustness. For each range image, the output of the feature learner is a global feature $F_P \in \mathbb{R}^{384}$.

The feature F_P is used as a condition to the denoiser, achieving iteratively denoising from pure noise to the ground truth pose. Here, we use a transformer, which consists of 8 encoder layers with 4 attention heads for feature aggregation, to implement denoiser. The latent embedding dimension is set to 512. The final fully connected layers of the denoiser are configured as $[512, 64, 6]$.

3. Additional Results

3.1. Results with different foundation models

A recent study AnyLoc [5] indicates that joint embedding self-supervised foundation models (DINO [3], DINOv2 [8]) are more adept at learning long-range global patterns compared to contrastive learning methods (CLIP [9]) and masked autoencoding approaches (MAE [4]). This characteristic makes them well-suited for visual localization. Therefore, we explore the impact of DINO and DINOv2, both employing a ViT-S backbone, on LiDAR localization accuracy, as illustrated in Tab. 3.

We observe that using foundation models pretrained on RGB images consistently outperforms training from scratch on LiDAR data (entry Rand). Specifically, on the Oxford and NCLT datasets, employing DINO and DINOv2 leads to average improvements of 7.1%/5.7% and 10.8%/12.2%, respectively. Moreover, we empirically find that using DINOv2 for feature learning surpasses the performance of using DINO, achieving a mean error of 3.53m/0.72° vs. 3.62m/0.82° and 1.19m/2.31° vs. 1.26m/2.35° on the Oxford and NCLT datasets, respectively. This experiment demonstrates that (1) the foundation model trained on RGB images can effectively improve localization accuracy, even with large domain differences, and (2) the foundation model trained on a larger dataset can bring a more significant accuracy improvement. Therefore, in this paper, we use DINOv2 for feature learning.

3.2. Results with different patch sizes

We investigate the impact of different patch sizes on Localization results, as shown in Tab. 4. Intuitively, reducing the patch size results in a finer representation. We observe

Methods	SGLoc	STCLoc	NIDALoc	HypLiLoc	DiffLoc
Params	105M	9M	8M	52M	40M
Runtime	57ms	97ms	120ms	21ms	39ms

Table 5. Parameter count and runtime of different methods.

Layer number	2	4	6	8
Params	27M	32M	36M	40M
Results	6.75m/11.88°	1.30m/2.60°	1.26m/2.43°	1.19m/2.31°

Table 6. Localization results with different parameter quantity.

that (1) the position accuracy is robust to patch sizes, (2) the orientation accuracy tends to increase as the patch size reduces, and (3) the optimal result is obtained with $[4, 16]$. This means the smaller patches enable the learning of more fine-grained information, which benefits orientation prediction. Taking both position and orientation accuracy into account, in this paper, we implement the patch size as $[4, 16]$.

3.3. Results about the parameter count

Tab. 5 shows the parameter count and runtime comparison of DiffLoc with the existing competitive methods. DiffLoc achieves SOTA accuracy with a moderate 40M parameter and 39ms runtime. Note that the parameter of STCLoc and NIDALoc is less than 10M due to their simple network design. Thus, the overhead of DiffLoc is very competitive with these methods, but the accuracy of DiffLoc is much better.

In addition, we investigate the relationship between the parameter count of APR and performance. APR captures scene information by an encoder and the map scene to pose by a regressor. As demonstrated in the literature [12], (1) the regressor is primarily responsible for memorization, and (2) increasing the number of parameters of the regressor enhances its memorization capacity. To ensure localization accuracy, they recommend the regressor parameter should be increased with larger scene sizes. We investigate the relationship of DiffLoc performance with parameter quantity by modifying the number of Transformer layers n in the denoiser on the NCLT dataset. As shown in Tab. 6, when $n = 2$, the performance is greatly reduced. The performance is relatively accurate when $n = 4$. These results show the same trend as the literature [12]. In this paper, to preserve the localization accuracy and real-time performance, we set $n = 8$.

4. Visualization

We show more visualization results of the top 4 methods in the main paper (NIDALoc [13], HypLiLoc [11], SGLoc [6], and the proposed DiffLoc) in Fig. 3 and Fig. 4 on the Oxford and NCLT datasets, respectively.

Clearly, the trajectories predicted by DiffLoc closely overlap with the ground truth, with fewer outliers compared with other methods, demonstrating its great effectiveness. It is important to note that certain regions (approximately frames 4300 to 4500) on the 2012-05-26 trajectory of the NCLT dataset, absent in the training data, contribute to the presence of numerous outliers, as depicted in the last row of Fig. 4. Previous studies [10, 12] indicate that regression-based localization methods are not guaranteed to generalize from training data in practical scenarios. In this paper, the proposed DiffLoc can efficiently identify these regions with uncertain results, as shown in Fig. 6 of the main paper, which bridges the gaps in practical applications.

References

[1] Angelika Ando, Spyros Gidaris, Andrei Bursuc, Gilles Puy, Alexandre Boulch, and Renaud Marlet. Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. In *CVPR*, pages 5240–5250, 2023. 1

[2] Dan Barnes, Matthew Gadd, Paul Murcutt, Paul Newman, and Ingmar Posner. The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset. In *ICRA*, pages 6433–6438, 2020. 1, 4

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2

[5] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *arXiv preprint arXiv:2308.00688*, 2023. 2

[6] Wen Li, Shangshu Yu, Cheng Wang, Guosheng Hu, Siqi Shen, and Chenglu Wen. Sgloc: Scene geometry encoding for outdoor lidar localization. In *CVPR*, pages 9286–9295, 2023. 3

[7] Carlevaris-Bianco Nicholas, K. Ushani Arash, and M. Eustice Ryan. University of michigan north campus long-term vision and lidar dataset. *IJRR*, 35:545–565, 2015. 1, 5

[8] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2

[10] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *CVPR*, pages 3302–3312, 2019. 3

[11] Sijie Wang, Qiyu Kang, Rui She, Wei Wang, Kai Zhao, Yang Song, and Wee Peng Tay. Hypliloc: Towards effective li-

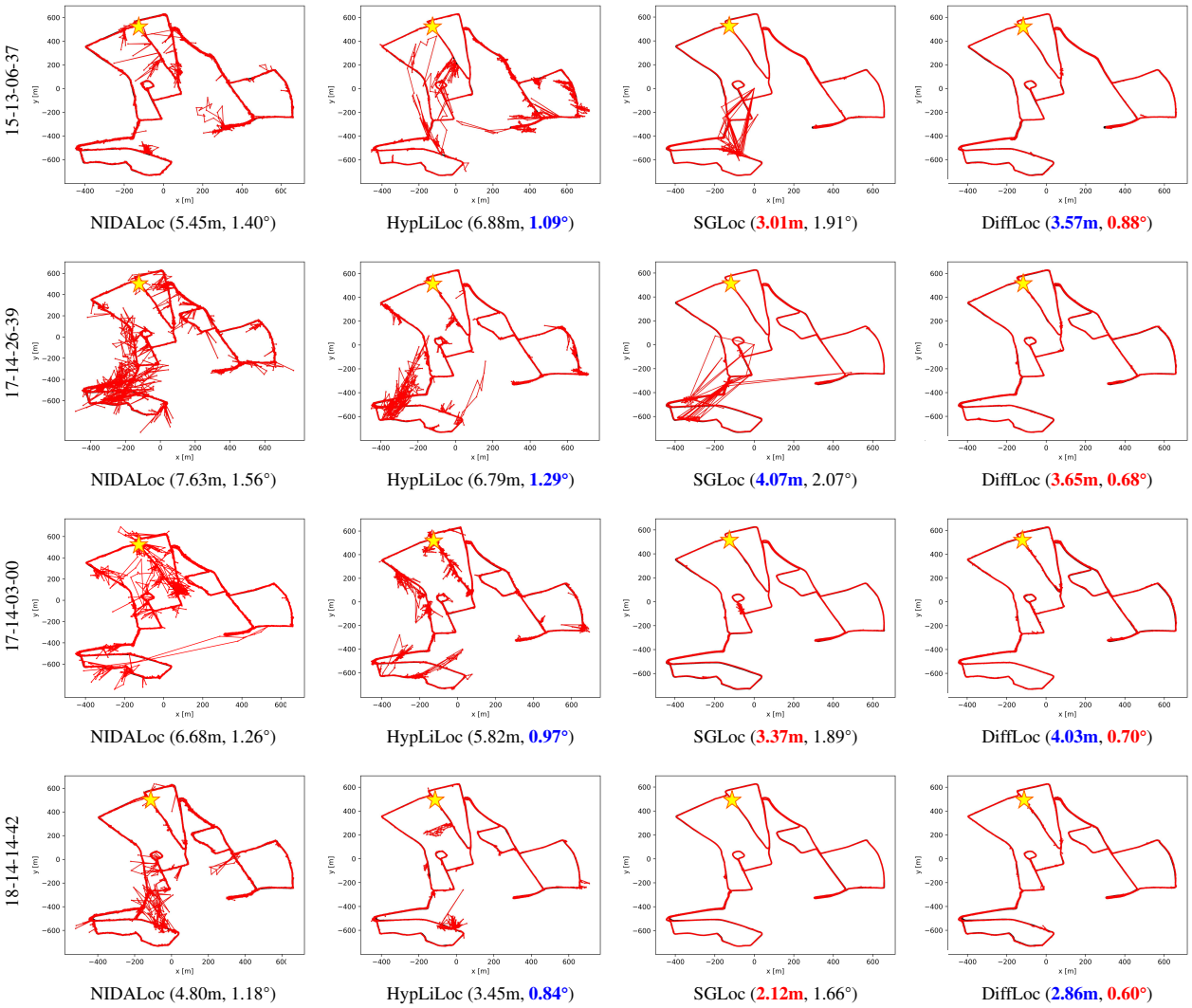

Figure 3. LiDAR localization results of on the Oxford [2] dataset. The ground truth and prediction are black and red lines, respectively. The star denotes the first frame. The caption of each subfigure shows the mean position error (m) and orientation error (°). For each trajectory, we highlight the **best** and **second-best** results.

dar pose regression with hyperbolic fusion. In *CVPR*, pages 5176–5185, 2023. 3

[12] Shangshu Yu, Cheng Wang, Chenglu Wen, Ming Cheng, Minghao Liu, Zhihong Zhang, and Xin Li. Lidar-based localization using universal encoding and memory-aware regression. *PR*, 128:108915, 2022. 3

[13] Shangshu Yu, Xiaotian Sun, Wen Li, Chenglu Wen, Yunuo Yang, Bailu Si, Guosheng Hu, and Cheng Wang. Nidaloc: Neurobiologically inspired deep lidar localization. *IEEE TITS*, 2023. 3

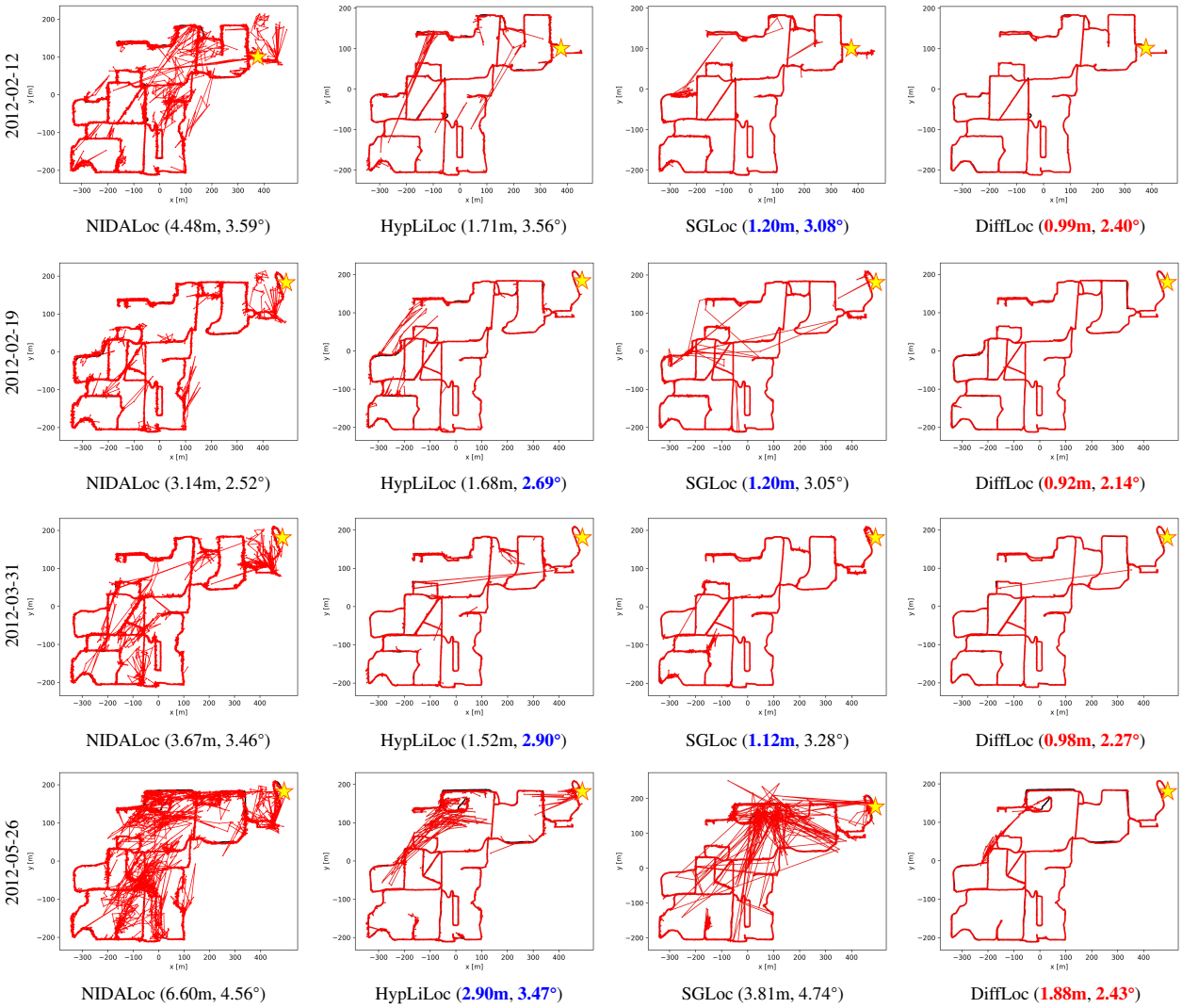

Figure 4. LiDAR localization results on the NCLT dataset [7]. The ground truth and prediction are black and red lines, respectively. The star denotes the first frame. The caption of each subfigure shows the mean position error (m) and orientation error (°). For each trajectory, we highlight the **best** and **second-best** results.