

Supplementary Material for ‘UniLGL: Learning Uniform Place Recognition for FOV-limited/Panoramic LiDAR Global Localization’

I. CAN FOUNDATION MODEL BRING GENERALIZATION ABILITY?

Foundation models in robotics aim to provide systems with broad generalization capabilities by leveraging large-scale pretraining on diverse data sources. To facilitate understanding of the role of introducing foundation models into LGL, we evaluate the LGL performance of three variants: UniLGL, *UniLGL w/o FM*, and UniLGL initialized with a foundation model but without fine-tuning (*UniLGL w/o FT*). To comprehensively evaluate the cross-model and cross-scene generalization capability of UniLGL, two benchmark datasets, MCD [1] and Garden [2], are utilized. Both datasets provide paired FoV-limited and panoramic LiDAR measurements, serving to assess the cross-model generalization performance. Moreover, their data are acquired from entirely distinct environments, thereby facilitating the evaluation of the cross-scene generalization capability.

A. Cross-model Generalization Ability

During the experiments, the FoV-limited LiDAR scans are used as queries, while the panoramic LiDAR scans serve as the database. The associated data sequences are referred to as NTU_CM_XX and Garden_CM_XX, respectively. It is worth noting that, to demonstrate the *zero-shot* generalization ability, no cross-modal training is performed. As shown in the t-SNE [3] visualization in Fig. 1, introducing a foundation model imparts a certain level of generalization ability to the LPR network, enabling *UniLGL w/o FT* to achieve better clustering performance than *UniLGL w/o FM*, without fine-tuning. By initializing the network with DINO [4] and fine-tuning with only a small amount of homogeneous LiDAR data, UniLGL learns highly discriminative global descriptors in the heterogeneous LPR task. In addition to the above qualitative analysis, we present quantitative results of place recognition and global localization performance in Table I. The results show that, thanks to the introduction of the foundation model, UniLGL achieves outstanding *zero-shot cross-modal generalization ability*. When compared to *UniLGL w/o FT* and *UniLGL w/o FM*, UniLGL achieves a 59.64%-82.24% improvement in recall and over 79% increase in LGL successful rate.

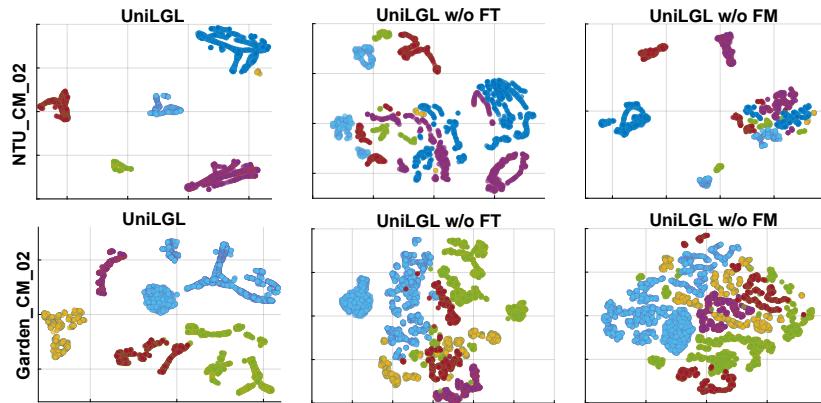


Fig. 1: t-SNE visualization of global descriptor encoded by UniLGL, UniLGL w/o FT, and UniLGL w/o FM. For each sequence, we select six distinct locations to visualize the discriminability of the global descriptors.

TABLE I: Zero-Shot Cross-model LPR (Recall (%)) at Top-1)/LGL (Success Rate (%)) Performance.

Sequence	UniLGL	UniLGL w/o FT	UniLGL w/o FM
NTU_CM_02	97.75/86.75	37.95/0.07	0.60/0.01
NTU_CM_10	92.95/78.19	32.61/0.13	3.64/0.01
NTU_CM_13	96.46/86.13	41.02/0.35	8.59/0.00
Garden_CM_01	93.19/75.77	26.02/0.25	3.33/0.00
Garden_CM_02	91.86/76.10	30.54/0.34	17.54/0.01
Garden_CM_03	94.37/76.07	31.32/0.41	19.71/0.00
Garden_CM_04	87.37/76.94	37.06/0.36	24.86/0.01
Average	93.42/79.42	33.78/0.27	11.18/0.01

¹ Best results are shown in **bold**.

TABLE II: Cross-scene LPR (Recall (%)) at Top-1)/LGL (Success Rate (%)) Performance.

Sequence	UniLGL	UniLGL w/o FT	UniLGL w/o FM
Mid_NTU_02	98.50/75.30	81.48/28.81	70.97/23.99
Mid_NTU_10	95.44/88.98	92.11/75.88	87.80/70.17
Mid_NTU_13	96.85/71.05	70.10/39.58	66.95/33.54
Garden_01	91.19/85.53	90.58/84.62	78.51/67.83
Garden_02	91.21/84.76	90.08/83.24	76.88/62.70
Garden_03	86.75/78.37	86.23/77.24	73.90/52.22
Garden_04	90.80/83.96	88.39/83.47	76.36/59.51
Average	92.96/81.14	85.57/67.57	75.91/52.85

¹ Best results are shown in **bold**.

B. Cross-scene Generalization Capability

To evaluate the cross-scene generalization capability, UniLGL and *UniLGL w/o FM* are retrained solely on the MCD dataset for 5 epochs, using *ntu_day_01* and *ntu_night_08* as the training sequences. Table II presents the performance of UniLGL, *UniLGL w/o FT*, and *UniLGL w/o FM* across two distinct environments — the large-scale campus scenes in the MCD dataset and the repetitive artificial vegetational scenes in the Garden dataset. From the results, *UniLGL w/o FT* consistently outperforms *UniLGL w/o FM*, and achieves a comparable performance with UniLGL on the Garden dataset. This indicates that the VFM enables the network to obtain preliminary place recognition ability even without LGL-specific fine-tuning. However, on the MCD dataset, *UniLGL w/o FT* exhibits a marked performance degradation. This is attributed to the presence of numerous low-overlap point cloud pairs induced by large viewpoint differences, which represents an LGL-specific condition that a task-agnostic pre-trained VFM is hard to handle adequately without fine-tuning. UniLGL fine-tunes the DINO [4] using a small amount of LiDAR data under LGL-specific supervision, enabling it to achieve consistent performance across distinct environments and thereby demonstrating *cross-scene generalization capability*. As the quantitative results shown in Table II, UniLGL achieves a 7.39-17.05% improvement in recall and a 13.57-28.29% increase in LGL successful rate.

REFERENCES

- [1] T.-M. Nguyen *et al.*, “McD: Diverse large-scale multi-campus dataset for robot perception,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 22 304–22 313.
- [2] Z. Wu *et al.*, “Mag-mm: Magnetic-enhanced multi-session mapping in repetitive environments,” *IEEE/ASME Trans. Mechatron.*, pp. 1–13, 2025.
- [3] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [4] M. Caron *et al.*, “Emerging properties in self-supervised vision transformers,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9630–9640.