Supplementary Material of BEVPlace

Lun Luo, Shuhang Zheng, Yixuan Li, Yongzhi Fan, Beinan Yu, Si-Yuan Cao,
Junwei Li, Hui-Liang Shen
{luolun, zhengsh, yixuanli, tony_fan, yubeinan, cao_siyuan, lijunwei7788, shenhl}@zju.edu.cn

1. Implementation Details

The architectures of BEVPlace have been illustrated in the main paper. Here we provide a more detailed implementation of the BEV image generation module, the group convolution network, and the NetVLAD layer, as shown in Fig. 1.

In the BEV image generation stage, we project the point clouds into BEV images with a grid size of 0.4 meters. Since the point cloud has been cropped into a $[-20$ m, 20 m] cubic window, the BEV image is of size 100x100.

In the group convolution network, we first warp the input BEV image with rotation transforms sampled from the rotation group. In our implementation, we sample evenly in the rotation group with an interval of $45°$. Then, we fed the warped images into a vanilla CNN network. For each local group feature in the output feature maps, we use two group convolution branches and bilinear pooling to extract group equivariant local features. The configuration of the CNN and the group convolution layers are listed in Table 1. "Conv(output channels, kernel size, stride)" denotes a convolutional layer. "AvgPool(kernel size, stride)" denotes an average pooling layer. For more insights into the group feature design, please refer to [4].

Table 1. Group convolution Network Architecture.

Layer	Operations
CNN	Conv(16,5,1)-InstanceNorm-Relu Conv(32,5,1)-InstanceNorm-Relu-AvgPool(2,2) Conv(32,5,1)-InstanceNorm-Relu Conv(32,5,1)-InstanceNorm-L2Norm
Group Conv1	Conv(64,1,1)-Relu Conv(8,1,1)
Group Conv2	Conv(64,1,1)-Relu Conv(16,1,1)

In NetVALD, the global rotation invariant feature is obtained by aggregating the local features. In the inference stage, We apply principal component analysis (PCA) to reduce the feature dimension to 256.

2. Dataset Details

We evaluate the methods on the KITTI dataset [2], the ALITA dataset [7], and the benchmark dataset [1] in the main paper. The dataset partition of KITTI has been introduced in the main paper. We show the dataset details of the other two in the following.

ALITA dataset. The validation set of ALITA contains 6 sequences that have varying degrees of view change. Since each sequence has only a few point clouds, we merge all the sequences into one for a more challenging evaluation. After that, we obtain a validation set with 666 point clouds in the database and 1750 frames for query. The test set of ALITA contains 5623 point clouds. We generate all the global features and upload them to the website[1]. The recall rate at Top-1 calculated by the web server is used for performance comparison in the main text. Different from the KITTI dataset in which the point clouds are single frames, the point clouds of ALITA are submaps cropped from a global map. Thus, the point clouds are more evenly distributed in the 3D space.

Table 2. Number of queries for training and testing on the benchmark dataset.

	Train	Test
Oxford	21711	3030
U.S.	-	1972
R.A.	-	1579
B.D.	-	991

Benchmark dataset. The benchmark dataset contains four subsets including the Oxford RobotCar dataset, a university sector (U.S.), a residential area (R.A.), and a business district (B.D.). In our experiment, we only use the Oxford RobotCar dataset for training. We split the dataset following [1] and list the number of queries for training and testing in Table 2. Similar to ALITA, the point clouds of the Oxford RobotCar dataset are also submaps generated from multiple LiDAR scans.

[1] https://www.aicrowd.com/challenges/icra2022-general-place-recognition-city-scale-ugv-localization

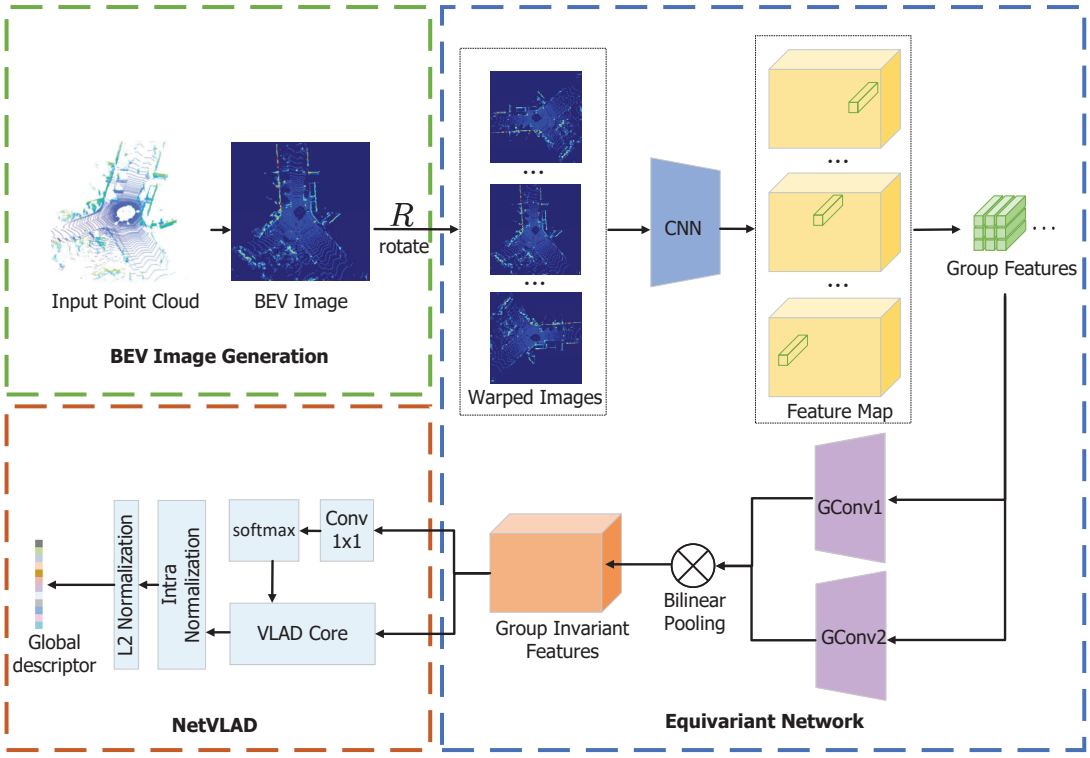

Figure 1. Detailed implementation of BEVPlace. We first generate the BEV image from the query point cloud. Then we extract group equivariant features through the group convolution network. The global rotation-invariant feature is obtained based on NetVLAD. Feature dimension reduction is achieved through PCA in the inference time.

3. Additional results on KITTI and discussions

Qualitative results of place retrieval. In addition to the quantitative results of place recognition in the main paper, we provide some qualitative examples in Fig. 2. We randomly select a query from each evaluation sequence of KITTI. Then, we display the query BEV image along with its Top-1, Top-5, Top-10, Top-15, and Top-20 retrieved results. For better visualization, we plot the location of each query in the reference map on the right. We use different colors to indicate the feature distance between the global feature of the query and the ones of all the other point clouds. Note that all the point clouds in the database have been randomly rotated and are sampled every 2 meters. It can be seen that, for each query, the Top-1 (red circle) and Top-5 (blue circle) matches correctly overlap with the target location (black circle), which demonstrates that our global feature has strong distinctiveness and good robustness to rotations.

More evaluation metrics of loop closure detection. In the main paper, we evaluate the precision-recall curve of loop closure detection on the KITTI dataset. Here, we further evaluate the average precision (AP) and F1 max score. Table. 3 shows that our method outperforms the compared methods with large margins in both metrics on all the sequences.

Correlation modeling between the geometry and the feature spaces. Fig. 3 shows the mapping relationship of different methods between the geometry and the feature distances on all the evaluation sequences of KITTI. We also plot the fitting curve using the modeling function, i.e. Eq. (5) in the main paper. It can be seen that our model can accurately depict the relationship between the geometry and the feature distances with specific parameters as denoted in each plot.

Distance estimation error on more sequences. Fig. 4 shows the fitted distribution of the distance errors of the methods on KITTI. Our method has lower errors on all the sequences and consequently achieves more accurate position estimation performance as shown in Fig. 7 of the main paper.

References

[1] Mikaela Angelina Uy and Gim Hee Lee. PointNetVLAD: Deep point cloud based retrieval for large-scale place recog-

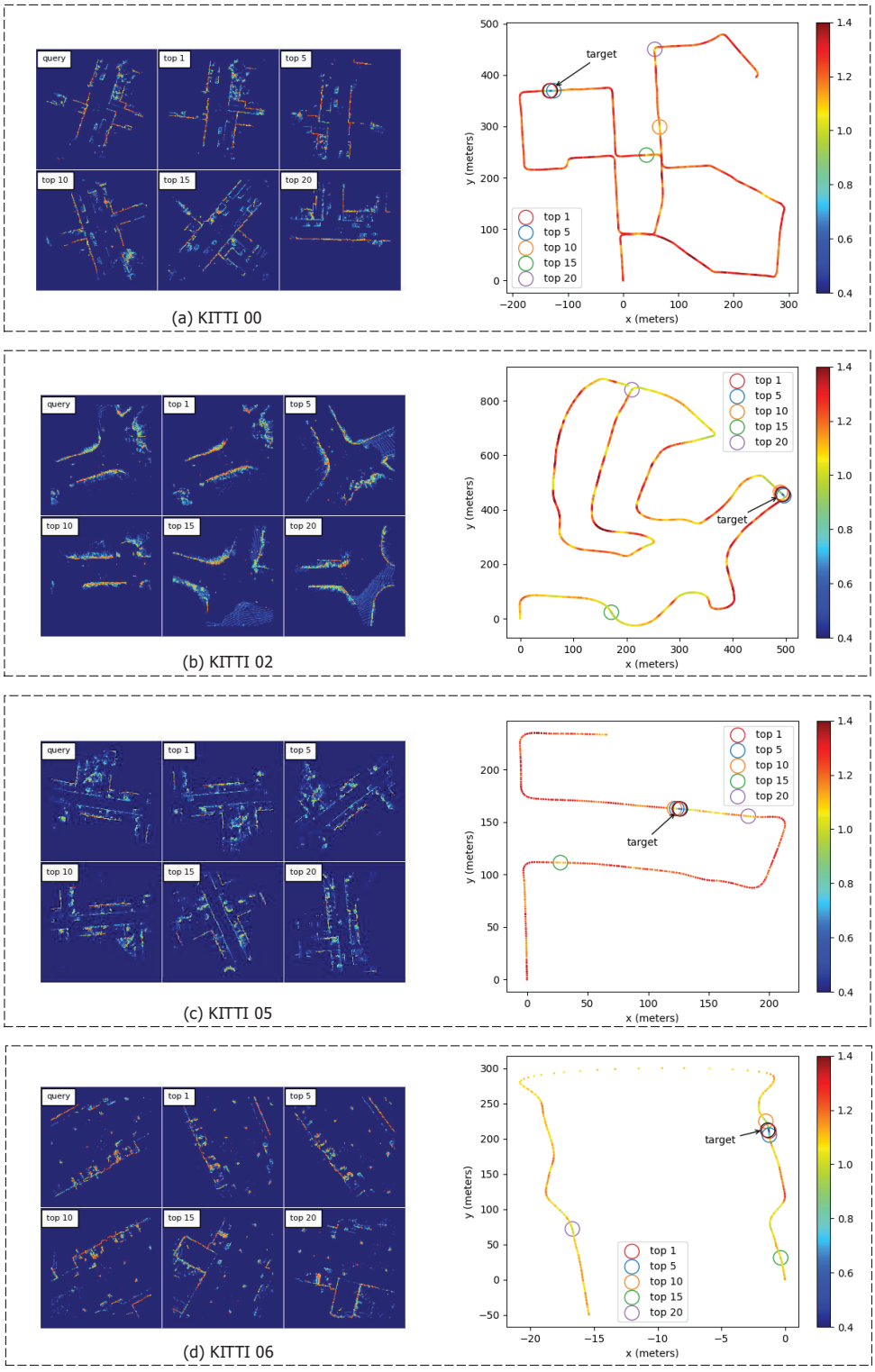

Figure 2. Visualizations of example retrieval results of BEVPlace on the KITTI dataset. For each sequence, we randomly choose a query and then display the query BEV image and the Top-1, Top-5, Top-10, Top-15, Top-20 matches. We also indicate the 2D position of the query and the matches in the associated reference map. The feature distance between the query and all the other point clouds are color-coded. All the retrievals are performed under random rotations to demonstrate the robustness of our method.

Table 3. Loop closure detection performance on the KITTI dataset.

	KITTI 00		KITTI 02		KITTI 05		KITTI 06	
	AP	F1 Max	AP	F1 Max	AP	F1 Max	AP	F1 Max
PointNetVLAD [1]	0.788	0.751	0.187	0.346	0.517	0.532	0.317	0.407
LPD-Net [5]	0.816	0.802	0.489	0.557	0.586	0.591	0.267	0.410
SOE-Net [8]	0.878	0.860	0.582	0.626	0.719	0.718	0.483	0.544
MickLoc3D-V2 [3]	0.864	0.868	0.407	0.517	0.761	0.758	0.326	0.459
OverlapTransformer [6]	0.934	0.895	0.755	0.790	0.824	0.792	0.939	0.874
BEVPlace	**0.986**	**0.983**	**0.848**	**0.830**	**0.978**	**0.953**	**0.996**	**0.994**

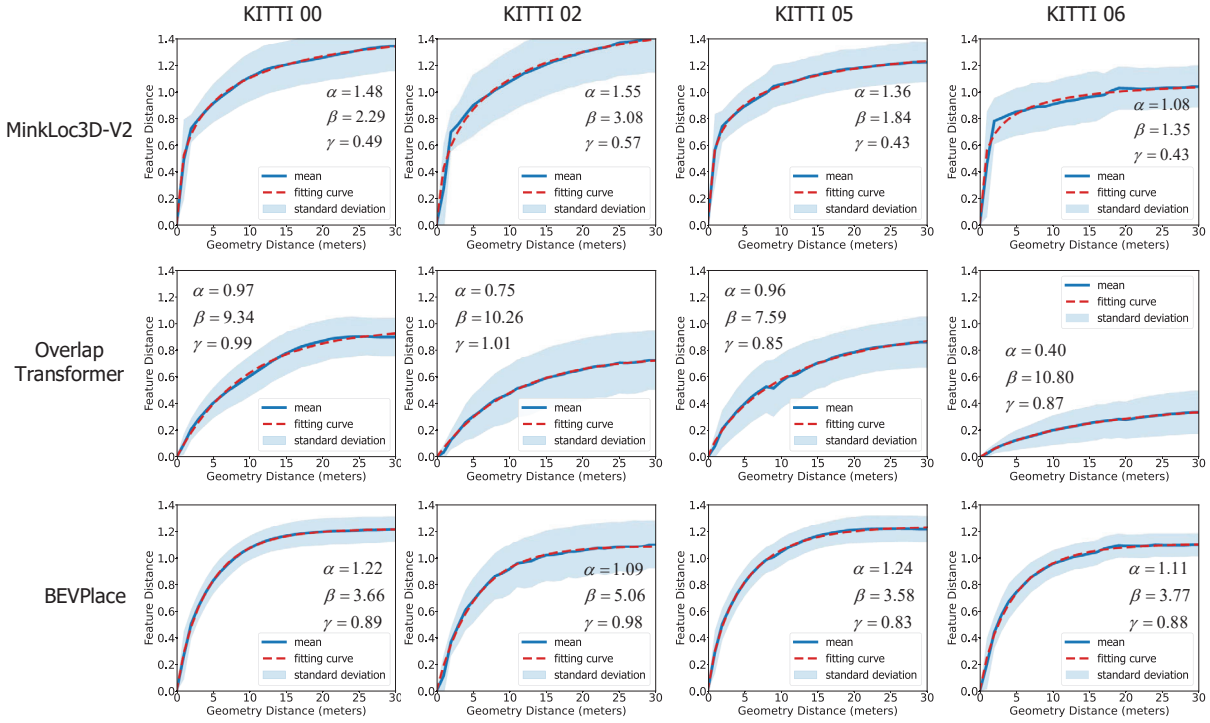

Figure 3. Geometry distance and feature distance relationship of the point clouds in the evaluation sequences of KITTI. We also plot the fitting curve based on the mapping model and give the fitting parameters α, β, and γ.

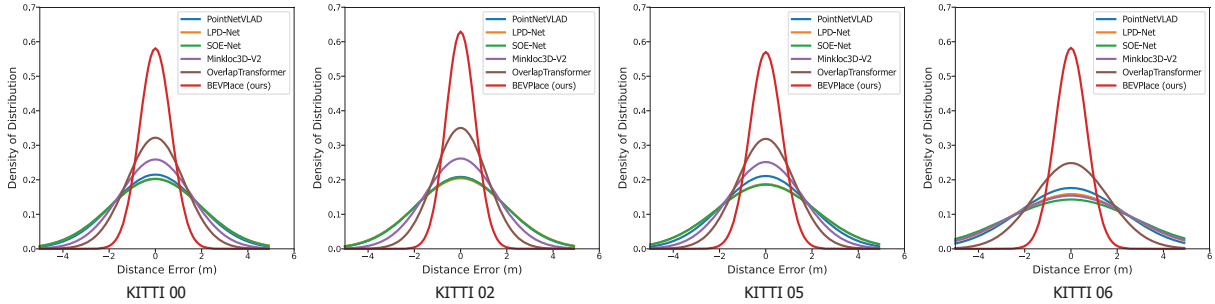

Figure 4. Distance estimation error distribution of the evaluation sequences on the KITTI.

nition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4470–4479. IEEE, 2018. 1, 4

[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 1

[3] Jacek Komorowski. Improving point cloud based place recognition with ranking-based loss and large batch training. In *IEEE International Conference on Pattern Recognition*, 2022. 4

[4] Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. GIFT: Learning transformation-invariant dense visual descriptors via group cnns. In *Conference and Workshop on Neural Information Processing Systems*, 2019. 1

[5] Zhe Liu, Shunbo Zhou, Chuanzhe Suo, Peng Yin, Wen Chen, Hesheng Wang, Haoang Li, and Yun-Hui Liu. LPD-Net: 3D point cloud learning for large-scale place recognition and environment analysis. In *IEEE International Conference on Computer Vision*, pages 2831–2840. IEEE, 2019. 4

[6] Junyi Ma, Jun Zhang, Jintao Xu, Rui Ai, Weihao Gu, and Xieyuanli Chen. OverlapTransformer: An efficient and yaw-angle-invariant transformer network for LiDAR-based place recognition. *IEEE Robotics and Automation Letters*, 7(3):6958–6965, 2022. 4

[7] Yin Peng, Zhao Shiqi, Ge Ruohai, Cisneros Ivan, Fu Ruijie, Zhang Ji, Choset Howie, and A. Scherer Sebastian. ALITA: A large-scale incremental dataset for long-term autonomy. In *arXiv preprint arXiv:2105.11605*, 2022. 1

[8] Yan Xia, Yusheng Xu, Shuang Li, Rui Wang, Juan Du, Daniel Cremers, and Uwe Stilla. SOE-Net: A self-attention and orientation encoding network for point cloud based place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11343–11352, 2021. 4