ForestLPR: LiDAR Place Recognition in Forests Attentioning Multiple BEV Density Images
Supplementary Material

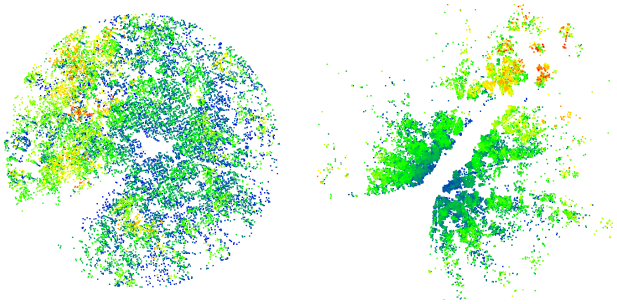
This supplementary material is structured as follows. In Section 1, we provide more details about the datasets and evaluation metrics. Section 2 provides the implementation details of our method and the compared baselines. In Section 3, we show some additional ablation studies. Finally, Section 4 contains additional quantitative results and qualitative results.

1. Dataset and Evaluation Metrics Details

Wild-Places. For both environments, Sequences 01 and 02 were collected on the same day, with Sequence 02 following the reverse route of Sequence 01. Sequence 03 was collected six months later and followed extended alternative routes, while Sequence 04 was collected 14 months after Sequence 01 and followed the same routes. Sequences 03 and 04 are reserved for intra-sequence loop closure detection, and all sequences are used for inter-sequence evaluation to test the challenges posed by long-term variations.

Along with accurate 6DoF ground truth, each submap corresponds to a .pcd file containing the x, y, and z values of its points. All submaps are sampled from the map within a one-second window at the corresponding timestamp. For training and validation, to more objectively evaluate the algorithm capabilities, we exclude the query neighboring frames that are temporally adjacent to the query from positive samples. This processing is similar to intra-sequence validation.

ANYmal Dataset. The dataset was gathered in forests by following a triangular route twice and once counterclockwise. Specifically, we first use Open3D SLAM [3] to generate the global poses and trajectory and remove frames with no or slight motion. For our collected 10 HZ scan data, every five frames are selected as keyframes, and the distance between them is about $0.5\,\mathrm{m}$. We also only sample points within a one second window of the corresponding timestamp for the submap. This means that a sequence of consecutive scans $\{\mathcal{P}_0, ..., \mathcal{P}_n\}$ are accumulated and transferred into the frame of middle scan $\mathcal{P}_{n/2}$, i.e., the keyframe. Finally, only the points with a diameter of $60\,\mathrm{m}$ are preserved. Compared to Wild-Places, the ANYmal dataset was collected by a quadruped robot with a lower LiDAR viewpoint, leading to greater LiDAR angular variations during robot walking. Furthermore, the LiDAR sensor is installed at the front-top of the robot, leading to a large blind area in scans (see Figure 1). The dataset contains a total of 1849 submaps (multi-frame registered scan samples) and 1239 loop closure-revisit pairs that meet the threshold, half of which are reverse revisits.



(a) Wild-Places (b) ANYmal Dataset

Figure 1. BEV visualization of point clouds in Wild-Places and ANYmal dataset. The points are colored by the height: red means the highest one, and blue means the lowest one. There are blind spots in the lower left area.

Table 1. Platform comparison.

Dataset	Platform	LiDAR mount condition
Wild-Places	handheld sensor payload	an angle of 45°, $1.5\,\mathrm{m}$ above the ground
ANYmal	quadrupedal robot	vertically, $0.7\,\mathrm{m}$ above the ground
Botanic	wheeled robot	vertically, $1.16\,\mathrm{m}$ above the ground

Botanic Dataset. We repurpose the *BotanicGarden* dataset [9], which is proposed for robot navigation in unstructured natural environments. Its submap generation is identical to the ANYmal dataset.

Here we summarize the LiDAR mounting configurations (e.g., positions and angles). Wild-Places is collected by a

handheld sensor payload, which includes a LiDAR mounted at an angle of 45°, 1.5 m above the ground. ANYmal dataset is collected by the quadrupedal robot, which contains a vertically mounted VLP-16, 0.7 m above the ground. Botanic dataset is collected by a wheeled robot Scout V1.0 from AgileX, which contains a vertically mounted LiDAR, 1.164 m above the ground.

Evaluation Metrics. Following the setting in [5], we consider a query successfully localized if the retrieved candidate is within 3 m.

For intra-sequence evaluation on all datasets, we use Recall@1 (R@1) and maximum F1 score (F1) as the metric. In particular, a query image is considered correctly localized if at least one of the top N ranked reference images is a positive candidate. In addition, previous entries adjacent to the query by less than t time difference are excluded from the search to avoid matching to the same/nearby instances. For Wild-Places, $t = 600$. For other datasets with a shorter sequence length, $t = 100$.

For inter-sequence evaluation on Wild-Places, Recall@1 and mean reciprocal rank (MRR) are used, where

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i},$$

and $rank_i$ is the rank of the first retrieved positive candidate to query submap i. If none of the top 25 candidates are correct, the reciprocal rank is 0. The final R@1 and MRR values are the means of the respective R@1 and MRR values overall evaluations.

2. More Implementation Details

PointPillar [7], TransLoc3D [13], MinkLoc3Dv2 [6], LoGG3D-Net [11], and BEVPlace [10] are all state-of-the-art learning-based methods and have saturated performance on the popular urban LiDAR datasets. ScanContext [4] and MapClosures [2] are handcrafted approaches, which first encode the highest z-value of bins or extract FAST and ORB features and use the key points for loop detection.

The implementation details of all compared baselines in Table 1 and 2 of the main paper are as follows:

- PointPillar[1]: Remove the classifcation head and add another GeM layer and L2Norm to generate the final features. Use pre-processed point cloud as input and Triplet loss to finetune the backbone.
- TransLoc3D[2]: Use *configs/transloc3d_baseline_cfg.py* and set quantization_size to 1.
- MinkLoc3Dv2[3]: Use *models/minkloc3dv2.txt*. Set coordinates to polar, quantization_step to (0.8, 0.15, 0.15), and normalize_embeddings to True.
- LoGG3D-Net[4]: Use the default setting in Wild-Places.
- ScanContext[5]: Use the default setting. To compare feature extraction modules fairly, we change the elevation BEV images generated from the whole point cloud to *density BEV images generated from pre-processed point cloud*, which is the same as ours.
- BEVPlace[6]: We use adjusted BEV density images as input and finetune the model with triplet loss, margin = 0.3.
- MapClosures[7]: Similarly, we use the adjusted BEV density images as input. Following the setting in the paper, we use a threshold of 50 bits on the Hamming distance for descriptor match. The threshold for the number of inliers obtained from RANSAC alignment is set to 10.
- Ours hyper-parameter selection: *1*. Multi-level feature extraction: we refer to the settings in [12]. *2*. Based on the distance threshold (3 m), we adjusted the overlap threshold for comparability. *3*. Backbone design: we adopted commonly used hyper-parameters in computer vision.

3. Additional Ablation Studies

First, we give more results to show the difference between density or height when generating BEV images for our model and Scan-Context. Then, we evaluate the model with different hyper-parameters. Finally, we present evaluation results for the model trained with a distance-based sample mining strategy and other similar overlap-based strategies [1, 2].

BEV Images Generation. We ablate our choice of density images over elevation images by modifying our pipeline and Scan-Context in intra- and inter-sequence evaluations of Wild-Places. Specifically, we use pre-processed point clouds as input and generate elevation images.

As shown in Table 2, for all three models, although BEV elevation images can also reflect the spatial distribution of trees, density is better than elevation in all tests. On the one hand, the pre-processed point cloud has removed points above 6 m, and the same elevation of many pixels causes the loss of features. On the other hand, the elevation image is also sensitive to the orientation of the sensor (pitch and roll), as the maximum height recorded varies with the distance and occlusions between the scanner and the object. More qualitative results refers to Section 4.2.

Hyper-parameters Sensitivity. Considering that S (the number of slices) and $\triangle h$ (height interval of slices) setting in height cropping are the most critical hyper-parameters for our work, we perform ablation experiments on them by setting $\triangle h = 0.5\,\mathrm{m}, 2.5\,\mathrm{m}$ respectively.

For setting different $\triangle h$, the results show that 1 m is a better choice. When digging into the samples, we find that

[1] github.com/zhulf0804/PointPillars
[2] github.com/slothfulxtx/TransLoc3D
[3] github.com/jac99/MinkLoc3Dv2
[4] github.com/csiro-robotics/Wild-Places/scripts/eval/logg3d
[5] github.com/csiro-robotics/Wild-Places/scripts/eval/scancontext
[6] github.com/zjuluolun/BEVPlace
[7] github.com/PRBonn/MapClosures

Table 2. Ablation studies on BEV generation. "elevation" denotes using maximum height to generate BEV images, and the results with "density" are the same as those in Table 1 and Table 2 of the main paper.

Method	Setting	V-03		V-04		K-03		K-04		Inter-V		Inter-K	
		F1	R@1	F1	R@1	F1	R@1	F1	R@1	R@1	MRR	R@1	MRR
Scan Context [4]	elevation	8.54	34.73	23.28	50.05	15.40	35.03	42.56	55.91	36.37	41.39	24.88	30.44
	density	37.66	54.80	64.49	75.99	34.33	50.23	66.85	63.48	57.23	58.11	52.81	55.82
Single BEV	elevation	42.35	46.64	32.80	43.01	39.42	51.56	54.36	61.93	34.20	55.84	35.07	56.73
	density	52.17	53.70	68.07	64.46	58.85	54.72	67.93	70.03	52.59	69.87	53.45	69.71
Ours	elevation	44.81	63.09	39.20	58.47	43.40	59.51	56.77	68.08	45.76	64.34	46.81	65.78
	density	64.15	76.53	78.62	82.33	65.01	74.89	81.97	76.73	77.14	84.26	79.02	83.87

Table 3. Ablation studies on slice number when generating multiple BEV density images from pre-processed point clouds. * represents the results already given in our main paper.

Slice Number	$\triangle h$	V-03		V-04		K-03		K-04		Inter-V		Inter-K		ANYmal		Botanic-03		Botanic-06	
		F1	R@1	F1	R@1	F1	R@1	F1	R@1	R@1	MRR	R@1	MRR	F1	R@1	F1	R@1	F1	R@1
10	0.5	57.63	70.81	73.15	77.83	60.35	68.12	77.48	70.54	71.59	76.34	73.86	74.19	75.02	70.43	71.87	79.36	73.07	76.21
5	1*	64.15	76.53	78.62	82.33	65.01	74.89	81.97	76.73	77.14	84.26	79.02	83.87	81.45	71.87	78.21	84.81	78.82	82.00
2	2.5	55.68	67.91	70.86	75.19	58.71	63.78	72.57	65.93	64.71	72.48	70.36	71.25	74.31	70.82	70.34	77.56	71.94	75.92

0.5m is too small to obtain effective density information and 2m is too large, causing confusion between different heights. So, we select $1\,\mathrm{m}$ as the final hyper-parameter in our main paper based on experience.

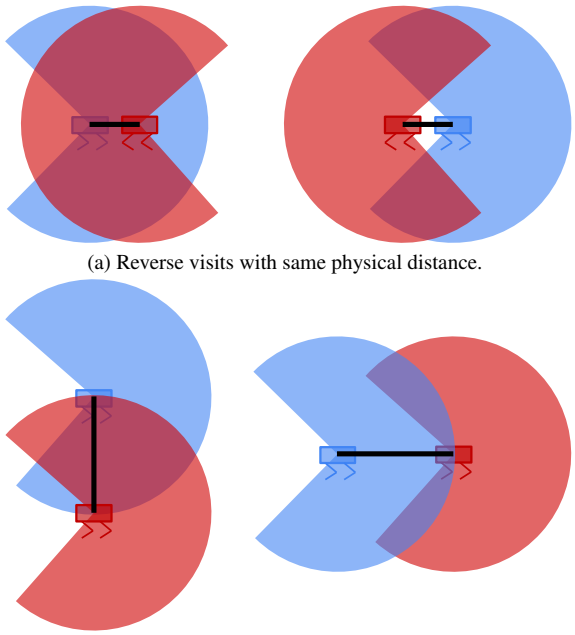
Positive Sample Mining Strategy. In this experiment, we adopt the same global features from our model and compare our method based on different ground truth definitions (i.e., positive sample mining strategies). "Distance" uses the ground truth based on physical distance 12.5m as the threshold for positive examples and 50m as the threshold for negative examples), and "Overlap" denotes the similar overlap-based method [1, 2] (0.9 as the threshold for positives and 0.5 as the threshold for negatives), which is similar to [8] for visual place recognition.

The results in Table 4 indicate that overlap-based mining is more helpful than physical distance-based mining, especially for V-03, which contains a large number of reverse revisits. That's because the overlap calculation directly corresponds to the degree of common vision. As shown in Figure 2, for data with blind spots, physical distance and degree of common vision are no longer proportional, and it's more reasonable to consider perspective simultaneously. In addition, our overlap calculation is:

$$o\left(\mathcal{V}_q, \mathcal{V}_p\right) = \frac{\left|\mathcal{V}_q \cap \mathcal{V}_p\right|}{\left|\mathcal{V}_q \cup \mathcal{V}_p\right|}, \tag{1}$$

which is more suitable for handling blind spots than those works [1, 2] that utilize $min(\cdot)$ as denominator.

In some extreme cases, the valid points within the scan range of one sample may completely cover another. Then, the similarity calculated by [1, 2] is 1, which is unreasonable, while Eq. (1) can better ensure the rationality of sample mining.



(a) Reverse visits with same physical distance.

(b) Visits with same physical distance.

Figure 2. Illustration of the effect of blind spots on the degree of common vision between two frames of point clouds.

Table 4. Ablation studies on positive sample mining strategy. All ablations use our final backbone and are trained on Wild-Places.

Strategy	V-03		V-04		K-03		K-04		Inter-V		Inter-K		ANYmal		Bo.-03		Bo.-06	
	F1	R@1	F1	R@1	F1	R@1	F1	R@1	R@1	MRR	R@1	MRR	F1	R@1	F1	R@1	F1	R@1
Distance	57.58	70.14	75.86	78.12	60.95	70.81	78.34	72.98	72.56	76.85	75.41	76.48	76.46	67.59	74.31	80.86	75.49	79.63
Overlap[1, 2]	61.04	74.82	77.91	80.26	63.25	73.59	80.31	74.92	75.08	78.31	77.63	78.52	80.03	70.51	75.96	81.98	77.38	80.14
Ours	64.15	76.53	78.62	82.33	65.01	74.89	81.97	76.73	77.14	84.26	79.02	83.87	81.45	71.87	78.21	84.81	78.82	82.00

Table 5. Additional studies on removing ground and tree top from point clouds. Here we show the results of the methods with cropping (pre-processed point clouds) and without cropping (raw point clouds).

Method	Crop	V-03		V-04		K-03		K-04		inter-V		inter-K	
		F1	R@1	F1	R@1	F1	R@1	F1	R@1	R@1	MRR	R@1	MRR
Scan-Context [4]	×	4.77	11.03	36.19	43.01	30.65	45.81	50.58	60.33	46.76	48.87	56.40	59.57
	✓	37.66	54.80	64.49	75.99	34.33	50.23	66.85	63.48	57.23	58.11	52.81	55.82
BEVPlace [10]	×	0.73	2.32	28.65	48.13	26.83	34.68	60.59	73.61	33.93	56.16	34.99	58.51
	✓	6.79	24.04	43.67	63.50	32.87	40.62	60.82	81.74	51.91	67.68	41.40	61.59
MapClosures [2]	×	12.99	6.95	29.44	14.94	25.05	14.35	65.38	55.05	26.68	27.44	19.48	20.22
	✓	38.47	23.82	49.95	25.61	21.35	11.95	70.62	54.59	34.31	37.83	19.35	21.16

4. Additional Results

4.1. Point Cloud Cropping

As mentioned in the main paper, the results of BEV-based methods in Tables 1 and 2 are also computed from pre-processed point clouds.

To prove that ground and tree top removal is general for single BEV-based baselines in forests, we give more results in Table 5. For single BEV-based baselines without cropping, we use whole point clouds to generate BEV density images and re-train BEVPlace on Wild-Places. It is worth noting that we have tried using pre-processed point clouds as input and training LoGG3D-Net on Wild-Places, but the performance was poor. We speculate that it is a problem with the hyper-parameter settings during training. For fairness, we do not give the results here for comparative analysis.

As shown in Table 5, using the whole point cloud has a detrimental impact on all single BEV-based methods, with performance drops by up to more than 40% across the board. It demonstrates the importance of removing ground and tree tops for single BEV-based place recognition methods in forests.

As the special case in Table 5, Scan-Context without cropping achieves better performance on inter-K than the cropped one. MapClosures also shows some opposite situations on K-03/04 and inter-K. Both of these two methods are non-learning based. This may be because trees in Karawatha are not as tall as those in Venman, and more discriminative density information from the tree top can be captured during scanning. (more visualization examples are shown in 4.2.)

In a word, the results provide solid insights about removing ground and tree tops in forests to guide future related works.

4.2. Qualitative Results

As shown in Figure 3, BEV images from the cropped point cloud are more recognizable, and the ones using the density value are easier to identify. The BEV elevation images are sensitive to poses because the coordinate values of points can severely vary when the robot moves. In contrast, the BEV density images are more robust because point cloud density does not depend on specific points.

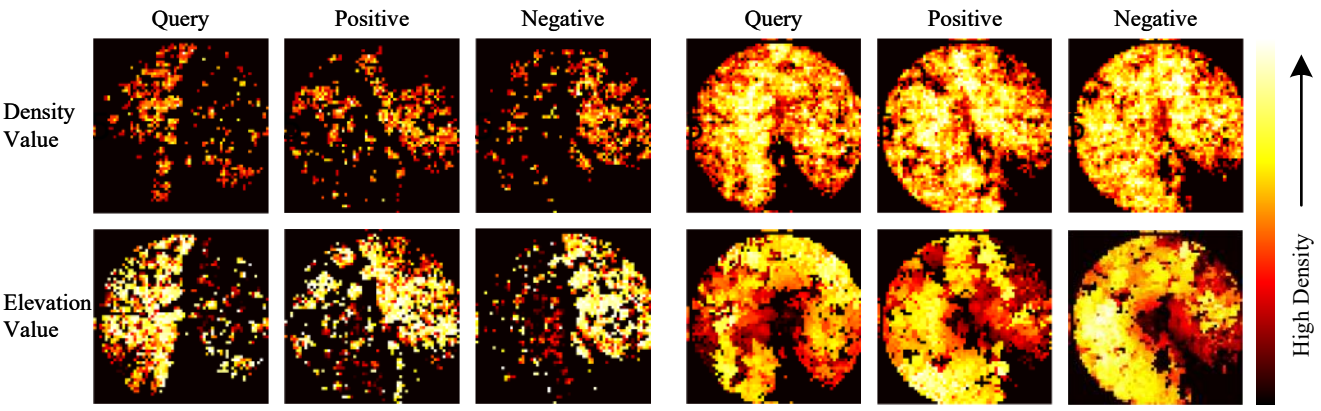
Figure 4 gives the height histograms of random samples in the Karawatha-03 and Venman-03 sequence, illustrating the height difference of trees in the Karawatha and Venman datasets. It supports our speculations that some algorithms present different performance levels on different sequences due to different scan patterns.

4.3. Inter-sequence Evaluation

As a supplement to TABLE 2 in the main paper, we provide the complete results of all compared methods on inter-sequence evaluations in Table 6, including R@1 and MRR scores.

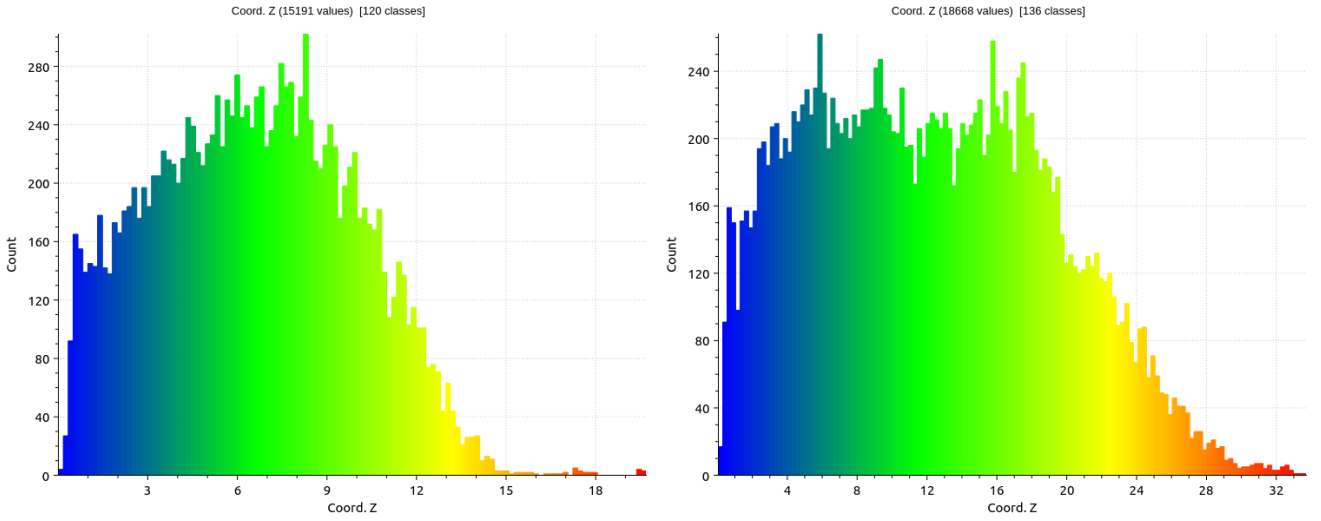
References

[1] Xieyuanli Chen, Thomas Läbe, Andres Milioto, Timo Röhling, Olga Vysotska, Alexandre Haag, Jens Behley, and Cyrill Stachniss. Overlapnet: Loop closing for lidar-based slam. *arXiv preprint arXiv:2105.11344*, 2021. 2, 3, 4



(a) BEV Images from Pre-processed Point Clouds (b) BEV Images from Raw Point Clouds

Figure 3. Generated single BEV images from point clouds. (a) shows the BEV images from pre-processed point clouds, and (b) shows ones from raw point clouds. It can be shown that the single BEV density image generated from pre-processed point clouds contains clearer spatial information than others, although it is still insufficient for place recognition in forests.



(a) Height histogram of a submap in K-03 (b) Height histogram of a submap in V-03

Figure 4. Height histogram of submaps in Karawatha-03 and Venman-03 sequences. It can be seen that the growth of vegetation in K-03 is shorter than that in V-03.

[2] S. Gupta, T. Guadagnino, B. Mersch, I. Vizzo, and C. Stachniss. Effectively Detecting Loop Closures using Point Cloud Density Maps. In *IEEE Int. Conf. Robot. Autom.*, 2024. 2, 3, 4, 6

[3] Edo Jelavic, Julian Nubert, and Marco Hutter. Open3d slam: Point cloud based mapping and localization for education. In *Robotic Perception and Mapping: Emerging Techniques, ICRA 2022 Workshop*, page 24, 2022. 1

[4] Giseop Kim and Ayoung Kim. Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map. In *IEEE Int. Conf. Intell. Robots Syst.*, pages 4802–4809. IEEE, 2018. 2, 3, 4, 6

[5] Joshua Knights, Kavisha Vidanapathirana, Milad Ramezani, Sridha Sridharan, Clinton Fookes, and Peyman Moghadam. Wild-places: A large-scale dataset for lidar place recognition in unstructured natural environments. In *IEEE Int. Conf. Robot. Autom.*, pages 11322–11328, 2023. 2

[6] Jacek Komorowski. Improving point cloud based place recognition with ranking-based loss and large batch training. In *ICPR*, pages 3699–3705. IEEE, 2022. 2, 6

[7] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds.

Table 6. Comparison with SOTA methods on Wild-Places. Validation refers to the testing dataset in training, and Inter-Venman/ Karawatha refers to Inter-sequence evaluations. Point-based methods are evaluated on raw point clouds. PointPillar and BEV-based methods are marked with *, showing the finetuning results on preprocessed point clouds and density images for fair comparison.

Method	Validation	Inter-V		Inter-K	
	R@1	R@1	MRR	R@1	MRR
PointPillar* [7]	35.49	60.34	70.68	52.19	66.72
TransLoc3D [13]	37.75	62.85	74.92	54.32	69.08
MinkLoc3Dv2 [6]	51.40	75.77	<u>84.87</u>	67.82	79.21
LoGG3D-Net [11]	<u>54.34</u>	**79.84**	**87.33**	<u>74.67</u>	<u>83.68</u>
Scan-Context* [4]	-	57.23	58.11	52.81	55.82
BEVPlace* [10]	26.25	51.91	67.68	41.40	61.59
MapClosures* [2]	-	34.31	37.83	19.35	21.16
Ours	**80.03**	<u>77.14</u>	84.26	**79.02**	**83.87**

In *CVPR*, pages 12697–12705, 2019. 2, 6

[8] María Leyva-Vallina, Nicola Strisciuglio, and Nicolai Petkov. Data-efficient large scale place recognition with graded similarity supervision. In *CVPR*, pages 23487–23496, 2023. 3

[9] Yuanzhi Liu, Yujia Fu, Minghui Qin, Yufeng Xu, Baoxin Xu, Fengdong Chen, Bart Goossens, Poly Z.H. Sun, Hongwei Yu, Chun Liu, Long Chen, Wei Tao, and Hui Zhao. Botanicgarden: A high-quality dataset for robot navigation in unstructured natural environments. *Robot. Autom. lett.*, 9(3):2798–2805, 2024. 1

[10] Lun Luo, Shuhang Zheng, Yixuan Li, Yongzhi Fan, Beinan Yu, Si-Yuan Cao, Junwei Li, and Hui-Liang Shen. Bevplace: Learning lidar-based place recognition using bird's eye view images. In *ICCV*, pages 8700–8709, 2023. 2, 4, 6

[11] Kavisha Vidanapathirana, Milad Ramezani, Peyman Moghadam, Sridha Sridharan, and Clinton Fookes. Logg3d-net: Locally guided global descriptor learning for 3d place recognition. In *IEEE Int. Conf. Robot. Autom.*, pages 2215–2221. IEEE, 2022. 2, 6

[12] Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, and Nanning Zheng. Transvpr: Transformer-based place recognition with multi-level attention aggregation. In *CVPR*, pages 13648–13657, 2022. 2

[13] Tian-Xing Xu, Yuan-Chen Guo, Zhiqiang Li, Ge Yu, Yu-Kun Lai, and Song-Hai Zhang. Transloc3d: Point cloud based large-scale place recognition using adaptive receptive fields. *arXiv preprint arXiv:2105.11605*, 2021. 2, 6