VXP: Voxel-Cross-Pixel Large-Scale Camera-LiDAR Place Recognition

Supplementary Material

In the supplementary we provide details on the used coordinate system convention in Sec. 1, evaluation procedure on the KITTI Odometry benchmark (Sec. 2), and further qualitative results in Sec. 4. Moreover, we visualize cross-modal local correspondences in the latent space in Sec. 5 and report few failure cases in Sec. 6.

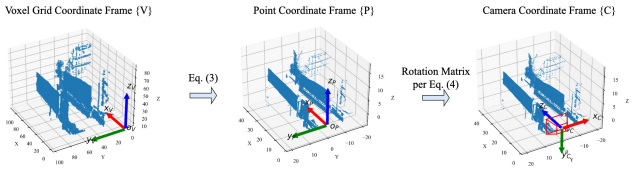


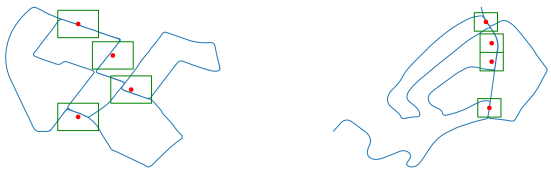
Figure 1. Illustrations of voxel grid coordinate frame $\{\mathcal{V}\}$ (left), intermediate point cloud (LiDAR) system $\{\mathcal{P}\}$ (middle), and target camera coordinate frame $\{\mathcal{C}\}$ (right). Note that once the points are transformed to $\{\mathcal{C}\}$, we can apply pinhole camera projection to project points to pixel coordinate frame.

1. Coordinate Frames

In Sec. 4.2 (main), we introduce Voxel-Pixel Projection module along with the associated coordinate transformations. Accordingly, in Fig. 1, we provide a comprehensive illustration of the operational coordinate frames $\{\mathcal{V}\}$, $\{\mathcal{P}\}$, and $\{\mathcal{C}\}$ and demonstrate the transformations between them as per Eq. (3) (main). Please note that camera parameters and relative transformation between sensors (camera and LiDAR) are known and provided as part of the datasets (Oxford, ViViD++, and KITTI). When performing pinhole camera projection as per Eq. (3) in the main paper, normalized intrinsics are used for adapting to different sizes of a feature map.

2. Training / Testing Setup (KITTI)

In Tab. 5 (main) we present the evaluation on the KITTI Odometry Benchmark [4]. Our evaluation protocol draws inspiration from the methodology employed in the Oxford RobotCar dataset [8]. As shown in Fig. 2, we define non-overlapping regions in KITTI sequences 00 and 02 and exclude samples from these regions during the training process. This ensures that the model is tested on unseen scenes. Throughout the training process, samples from the test regions in sequence 02 are utilized for validation, and we select the best model based on the validation set. During testing, both queries and database are sampled every 20 meters with the start offset of 5 meters. This ensures that samples from queries and database are not repeated, and the positive



(a) Test regions in KITTI sequence 00.

(b) Test regions in KITTI sequence 02.

Figure 2. We select 4 non-overlapping regions from KITTI (sequences 00, 02) and exclude their samples during training. Ability to perform accurate place retrieval in these regions is important to tackle localization drift of SLAM systems.

	Min.	Max.	Avg. # Positive
5m	0.3	19.9	5.5
10m	0.6	20.0	3.2
15m	0.4	19.9	2.5
20m	4.3	17.0	1.7

Table 1. Distance range [Min., Max.] of positive samples and their average number per query w.r.t different sampling intervals.

samples from the constructed query-database pair are in the range of [4.3m, 17.0m] to their corresponding queries as per Tab. 1. Notably, to mimic a real-world scenario of detecting loop candidates, we consider a "revisit" location by only keeping the positive samples (D_{t_i}) of a query (Q_{t_0}) such that $t_i < t_0$ and $t_0 - t_i > 10$. In other words, positive samples older than 10 seconds from the query timestamp are all "revisit" places. The 10-second threshold is determined empirically based on the sequences 00 and 02.

2.1. Sensitivity to Different Sampling (KITTI)

In Sec. 5.3 (main), we discussed the sensitivity of Lip-Loc [9] retrieval performance to variations in the sampling interval of queries and the database samples. To further investigate this observation, we perform additional studies in Fig. 3. As can be seen, our approach remains consistent across different sampling intervals. Lip-Loc, however, exhibits significant performance fluctuations when the sampling interval is changed. Notably, reducing the sampling interval results in more samples being classified as positives for each query, with these positive samples being spatially much closer to the query itself, as shown in Tab. 1. This sensitivity in Lip-Loc could be attributed to its training methodology, which uses N-pair batched contrastive loss. In their

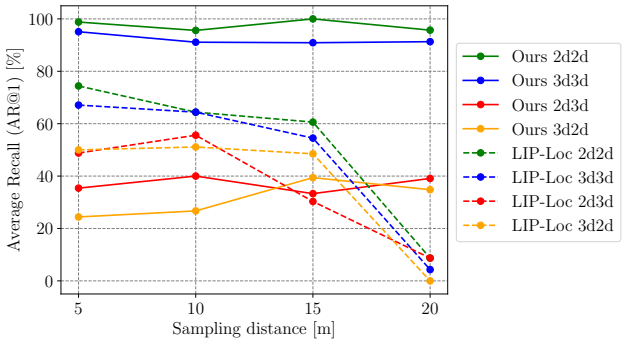


Figure 3. The impact of database-query sampling on VXP and LIP-Loc [9] retrieval accuracy. Our VXP shows consistent performance for all sampling ranges, while LIP-Loc results deteriorate rapidly.

approach, a pair of an image (I_{t_i}) and a LiDAR-scan (P_{t_j}) is considered a positive match only when $i = j$, while all the others (when $i \neq j$) are counted as negatives. Therefore, even a LiDAR-scan located 5 meters away from the image would still be labeled as negative, which does not allow the network to generalize to different sensor frequencies and setups. This limitation can be observed from the LIP-Loc performance on the Oxford RobotCar and ViViD++ benchmarks (Tab. 1 and Tab. 3 from main paper, respectively), where camera and LiDAR timestamps are not synchronized, negatively affecting the method's retrieval accuracy.

In contrast, VXP employs a training strategy that learns a shared embedding space by mimicking the output from an image network trained with a triplet loss function. In our approach, samples within 10 meters are considered positive, while those beyond 25 meters are labeled as negative as discussed in Sec. 5.1 (main). This training strategy enables VXP to robustly retrieve similar locations from the query, even under the more challenging conditions of a 20-meter query-database sampling interval.

3. Ablation Study of Image Network

We conducted extensive testing with various image encoders and pooling layers, and the results are summarized in Tab. 2. Overall, it is evident that DINO [2] stands out as the most favorable choice for the image encoder. Concerning the pooling layers, GeM [10] seems to perform slightly better than NetVLAD [1]. Based on the experiments, it is clear that the combination of DINO + GeM + FCN yields the most optimal results.

4. Qualitative Results in Challenging Illumination Conditions

Given that cameras are sensitive to changes in illumination conditions, robust visual place recognition at night poses significant challenges. In this experiment, we qualitatively demonstrate the differences in image-only retrieval (2D-2D) against cross-modal (2D-3D and 3D-2D) place recognition for the task of day-night and day-evening place recognition with reduced visibility.

2D-2D	Recall@1	Recall@1%
V16+N+L	80.4	91.6
V16+G+L	81.7	92.7
R18+N+L	77.2	90.3
R18+G+L	78.7	91.1
Dino+N+L	**85.3**	94.9
Dino+G+L (Ours)	**85.3**	**95.0**

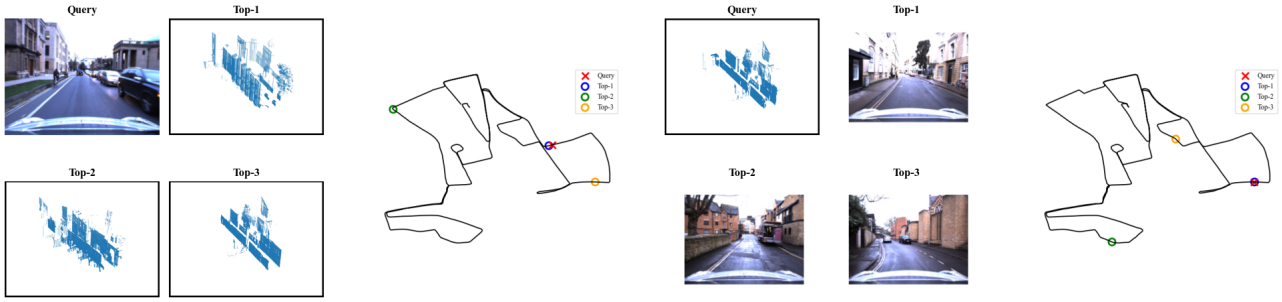
Table 2. The comparison of the different combinations of image encoder and pooling layer. V16 represents VGG16 [11], R18 is ResNet18 [5], Dino is DINO's ViTs-8 [2], N is NetVLAD [1], G refers to GeM [10], L means fully-connected layer. Our architectural design yeilds the best 2D-2D performance.

In Fig. 9c we can observe that the top 3 places for 2D-2D retrieval are quite far from the query. This suggests that image-based place recognition struggles to retrieve the correct candidate when the images are under different illumination conditions. However, VXP successfully retrieves the closest candidates using 2D-3D cross-modal retrieval as illustrated in Fig. 9d. This capability could explain why VXP exhibits slightly better top-1 2D-3D recall performance than its 2D-2D counterpart in Tab. 3 (main). Notably, VXP stands out as the only model (compared to [3, 7]) emphasizing the practical advantage of cross-modal retrieval with respect to the uni-modal counterpart.

We demonstrate additional qualitative results of the cross-modal retrieval task on Oxford RobotCar and KITTI datasets in Fig. 4 and Fig. 6 respectively. These observations underscore the versatility and effectiveness of cross-modal retrieval approaches in challenging real-world scenarios.

5. Correspondences in Local Feature Space

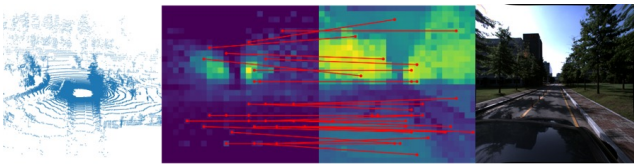
We visualize cross-modal matches from the ViViD++ dataset [6] in Fig. 5 established by picking the closest pairs of learned local descriptors. It is worth noting that LiDAR scans and images do not capture exactly the same information about the scene since LiDAR and camera are not synchronized. In addition, while we utilize data recorded by traversing the same route, the distance between samples corresponding to the same location can be significant among different traversals. For instance, in the ViViD++ dataset, the distance between an image and the corresponding point cloud averages about 7 meters according to calibration and GPS/INS poses.



(a) 2D-3D retrieval

(b) 3D-2D retrieval

Figure 4. Qualitative results for 2D-3D and 3D-2D cross-modal retrieval of our VXP on the Oxford RobotCar benchmark. We demonstrate the query and the top 3 closest places retrieved from a given map by our method. While the database samples traverse the same route, they are captured at different times when the environmental conditions vary. We can observe that the top 1 candidate is spatially close to the query, demonstrating our approach's effectiveness and accuracy.



LiDAR point cloud | Projected voxel feature map | Image feature map | Corresponding image

Figure 5. Example of local feature correspondences (red) between a projected voxel feature map and an image feature map on the ViViD++ dataset. Cross-matches are established by minimizing the cosine similarity distance between learned descriptors. Being important landmarks for the place recognition task, buildings and trees receive fairly accurate local matches.

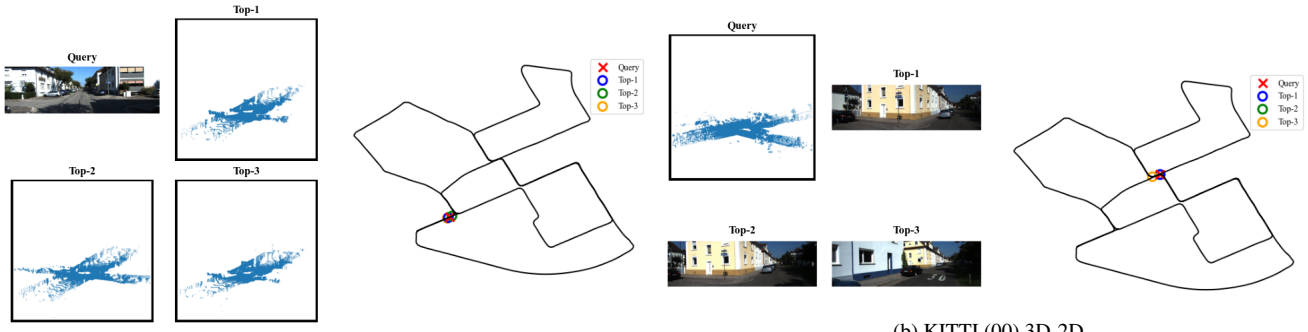
As it can be seen from Fig. 5, feature correspondences stemming from trees and building structures are accurately established. However, we encounter challenges in regions where voxel features are projected onto the ground region due to the ambiguous nature of respective image features. Therefore, capturing reliable correspondences within the ground regions appears to be a difficult task. Since the ground is constantly present in driving sequences, this misalignment only has minimal impact on the distinctiveness of the estimated global descriptors. Instead, it leverages correctly aligned correspondences from buildings and other static distinct objects in the scene to achieve state-of-the-art cross-modal performance.

6. Failure Cases

Although VXP achieves state-of-the-art performance on cross-modal place recognition task and scores well in many challenging conditions, it fails in some cases. Primarily we have noticed that repetitive structures such as highway roads cause confusion for our method and lead to incorrect retrievals. Notably, uni-modal methods also fail in such cases as shown in Fig. 7. We believe that integrating sequential information as part of the mobile robot localization system would facilitate the task and remain part of future work.

Secondly, sparse representation of places where the environment contains a lot of empty space and few, far-away structures, is not effective and lacks distinctive information. We demonstrate some failure examples in Fig. 8. One possible reason for worse performance lies in the projective nature of supervisory signal for our VXP. It is infeasible to learn a meaningful shared latent space and successfully perform cross-modal retrieval without establishing sufficient number of correspondences between voxels and pixels. Inspired by CASSPR [12], it might be beneficial to incorporate a point branch and enhance point cloud encoding to tackle such challenging cases.

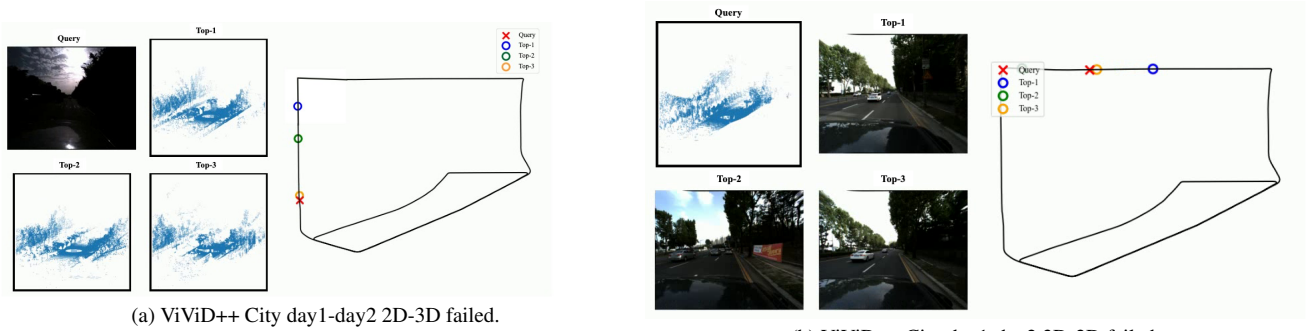
Lastly, illumination conditions are crucial in cross-modal retrieval where images are used as queries. Therefore, an image feature map generated from a poorly illuminated scene would obtain a bad-quality feature map, and leveraging such images for cross-modal retrievals becomes considerably challenging, as shown in Fig. 10. We believe that localizing with LiDAR scans (3D-2D retrieval) that are not affected by the light conditions would be more robust in such extreme cases.



(a) KITTI (00) 2D-3D

(b) KITTI (00) 3D-2D

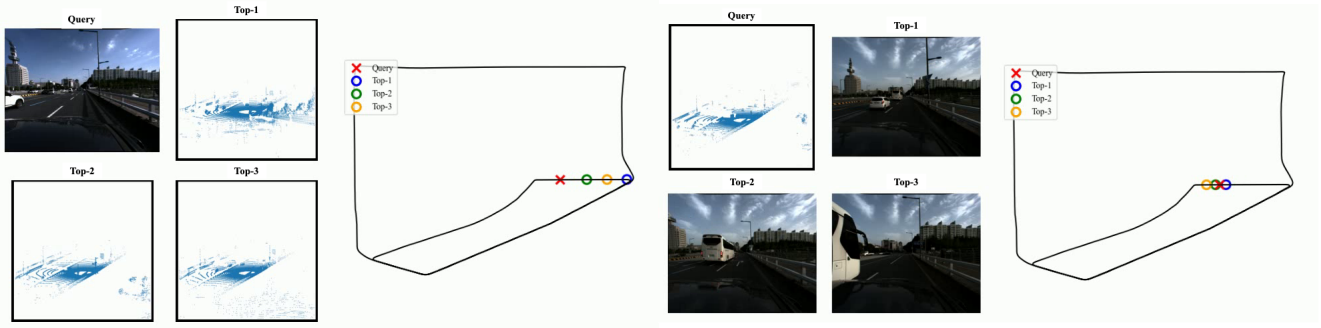
Figure 6. Qualitative results for 2D-3D and 3D-2D cross-modal retrieval of our VXP on the KITTI Odometry benchmark. We demonstrate the query and the top 3 closest places retrieved from a given map by our method. As described in Sec. 2 test queries are taken from the region unseen during training. We can observe that all top 3 candidates are spatially close to the query, demonstrating our approach's effectiveness and accuracy.



(a) ViViD++ City day1-day2 2D-3D failed.

(b) ViViD++ City day1-day2 3D-2D failed.

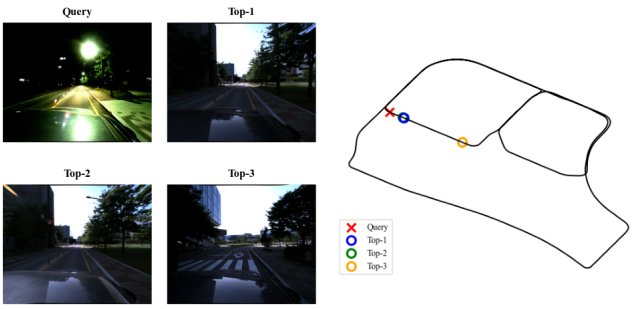
Figure 7. Failure cases of 2D-3D and 3D-2D cross-modal retrieval with our VXP due to challenging and repetitive scenes from the ViViD++ benchmark. For each retrieval, the query and its top 3 retrievals are shown. Although the correct candidate is obtained within the top 3 places, the top 1 is not the closest in the latent embedding space, and thus, the performance is negatively affected.



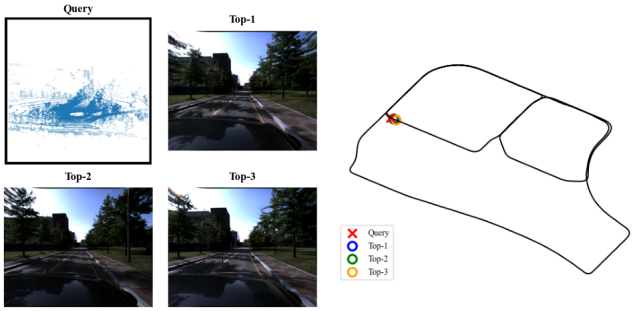
(a) ViViD++ City day1-day2 2D-3D failed.

(b) ViViD++ City day1-day2 3D-2D failed.

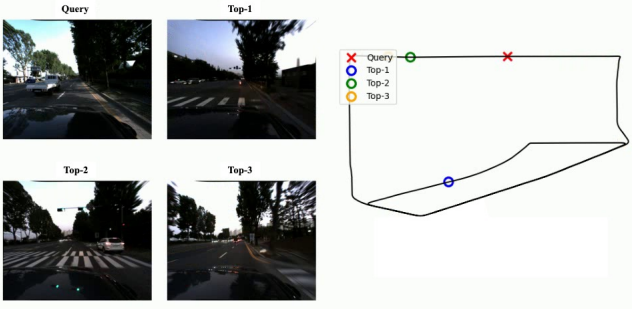
Figure 8. Failure cases of 2D-3D and 3D-2D cross-modal retrieval with our VXP due to sparse point cloud and lack of meaningful structures the ViViD++ benchmark. For each retrieval, the query and its top 3 retrievals are shown. Although the correct candidate is obtained within the top 3 places, the top 1 is not the closest in the latent embedding space, and thus, the performance is impaired.



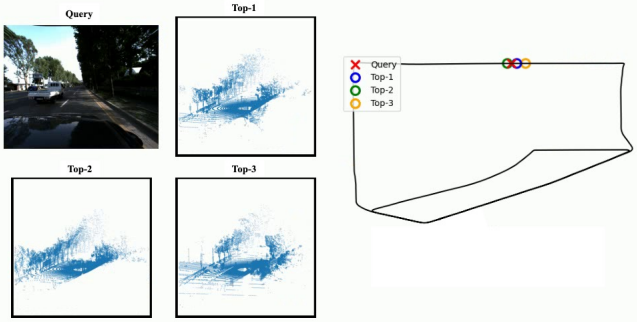
(a) ViViD++ campus night-day2 2D-2D failed.



(b) ViViD++ campus night-day2 3D-2D succeeded.

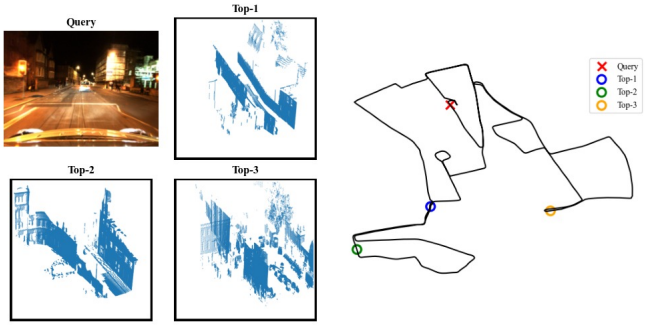


(c) ViViD++ city day1-evening 2D-2D failed.

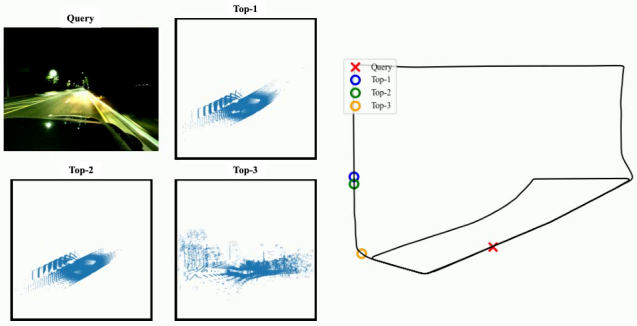


(d) ViViD++ city day1-evening 2D-3D succeeded.

Figure 9. Qualitative results of uni-modal (left) and cross-modal (right) on challenging illumination conditions such as evening and night sequences from the ViViD++ dataset. While 2D-2D place recognition is impaired by poor illumination conditions, integration of LiDAR scans, which remain unaffected, mitigates the issue and allows accurate cross-modal retrieval.



(a) Oxford RobotCar night-overcast 2D-3D failed.



(b) ViViD++ City night-day2 2D-3D failed.

Figure 10. Failure cases of 2D-3D cross-modal retrieval with our VXP due to poor illumination conditions in the night sequences from the Oxford RobotCar and ViViD++ datasets.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[3] Daniele Cattaneo, Matteo Vaghi, Simone Fontana, Augusto Luis Ballardini, and Domenico G Sorrenti. Global visual localization in lidar-maps through shared 2d-3d embedding space. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4365–4371. IEEE, 2020. 2

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[6] Alex Junho Lee, Younggun Cho, Young-sik Shin, Ayoung Kim, and Hyun Myung. Vivid++: Vision for visibility dataset. *IEEE Robotics and Automation Letters*, 7(3):6282–6289, 2022. 2

[7] Alex Junho Lee, Seungwon Song, Hyungtae Lim, Woojoo Lee, and Hyun Myung. Lc^2: Lidar-camera loop constraints for cross-modal place recognition. *IEEE Robotics and Automation Letters*, 8(6):3589–3596, 2023. 2

[8] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 1

[9] Sai Shubodh Puligilla, Mohammad Omama, Husain Zaidi, Udit Singh Parihar, and Madhava Krishna. Lip-loc: Lidar image pretraining for cross-modal localization. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 939–948. IEEE, 2024. 1, 2

[10] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 2

[11] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[12] Yan Xia, Mariia Gladkova, Rui Wang, Qianyun Li, Uwe Stilla, Joao F Henriques, and Daniel Cremers. Casspr: Cross attention single scan place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8461–8472, 2023. 3