Supplementary Materials: CROSSLOC3D: Aerial-Ground Cross-Source 3D Place Recognition

Tianrui Guan[1] Aswath Muthuselvam[1] Montana Hoover[1] Xijun Wang [1]
Jing Liang [1] Adarsh Jagan Sathyamoorthy [1] Damon Conover [2] Dinesh Manocha[1]

[1]University of Maryland, College Park [2]DEVCOM Army Research Laboratory

1. More Details on CS-CAMPUS3D Collection

In Fig. 1, we show the routes for the ground LiDAR scan collection. We blur the satellite imagery of the collection location due to the anonymity policy. The total length of the collection is roughly 247 minutes, and we remove the LiDAR scan in the location where the covariance of the GPS location exceeds 10 m due to poor satellite signal.

The aerial data covers the entire region, while the ground data is sparsely distributed along the routes. One of our main goals is to find the correspondence between the features of the aerial (dense) and ground (sparse) datasets. During training, we include the aerial data as a database, which corresponds to the region in the training split of the ground data. During testing, all query points are taken from a disjoint set of ground data. In contrast to [4], we include all aerial data in the testing database, instead of a disjoint subset of the aerial data. (In practice, if we narrow down the query location to a small subset of the database like [4], it defeats the purpose of our data retrieval task.)

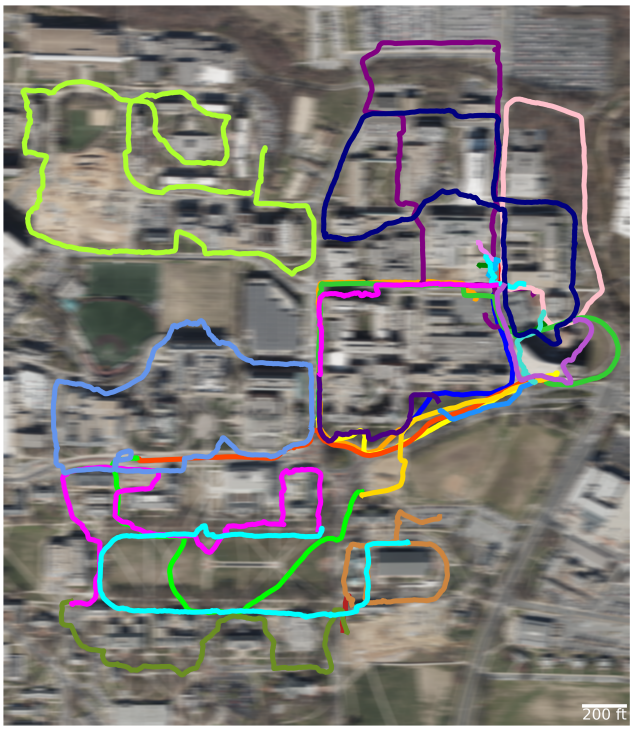

Figure 1: **CS-CAMPUS3D Routes.**

2. Implementation Details on CROSSLOC3D

In this section, we add more details about the implementation of CROSSLOC3D.

Backbone: For each voxel set V_i, we first perform a sparse convolution operation with a kernel size of 5 and a stride of 1, and another sparse convolution operation with a kernel size of 3 and a stride of 2, before multi-scale sparse convolution. The output feature size of both sparse convolutions is 64, and each sparse convolution is followed by a batch normalization, and a ReLU activation.

Iterative Refinement and NetVLAD: The EA block has a attention map that increases in linear complexity compared to the traditional quadratic complexity in transformer architecture. A single stream of the feature $\hat{F}$, which is also the query in traditional transformer, is passed as input. The self query is correlated with a pre-trained memory unit, which after training, contains the context of the entire point cloud dataset. The attention map's normalization allows scale invariant feature vectors. For EA blocks [1], we use 4 attention heads and a feature size of 64. We use two linear layers and one GELU activation to generate the time embedding. Following the iterative refinement, we use a 1D convolution with a feature size of 512 and a kernel size of 1, a batch normalization and a ReLU activation. For NetVLAD, the number of clusters K is 64.

CROSSLOC3D for Oxford RobotCar: We use a different configuration setting for the single-source benchmark, Oxford RobotCar [4], compared to the configuration for the proposed CS-CAMPUS3D. We use a quantization size of $[0.01, 0.12, 0.2]$ with a sub-step size of 2. In addition, the

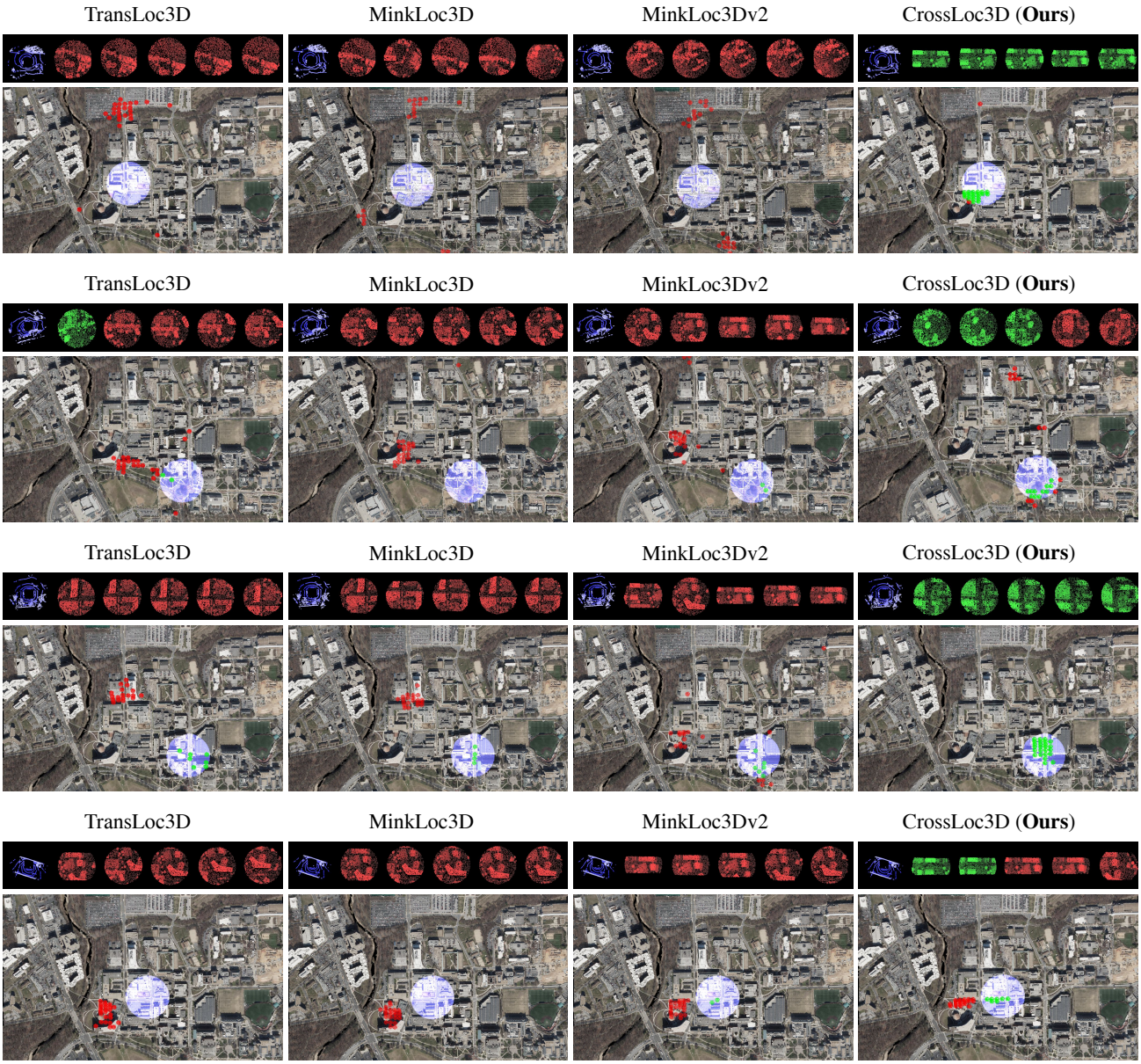

Figure 2: **More qualitative comparisons between CROSSLOC3D and other SOTA methods [5, 2, 3]: For each row, top:** Input ground query (blue) and top 5 retrievals (ranked from left to right) from the database of aerial data (true neighbor – green, false neighbor – red). **For each row, bottom:** Distributions of ground query and top 25 retrievals.

feature dimension of the NetVLAD is 512.

3. Inference Visualizations

We give some visualization results on CS-CAMPUS3D in Fig. 2. In Fig. 3, we show two challenging queries in the CS-CAMPUS3D benchmark. In the example on the top, the scan is captured in an open area with limited ground features, which causes some issues for retrieval. In the second example, the top 25 recall is low due to the lack of distinctive features on the ground, except for two parallel walls.

References

[1] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shimin Hu. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2021. 1

[2] Jacek Komorowski. Minkloc3d: Point cloud based large-scale place recognition. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1789–1798, 2021. 2, 3

[3] J. Komorowski. Improving point cloud based place recog-

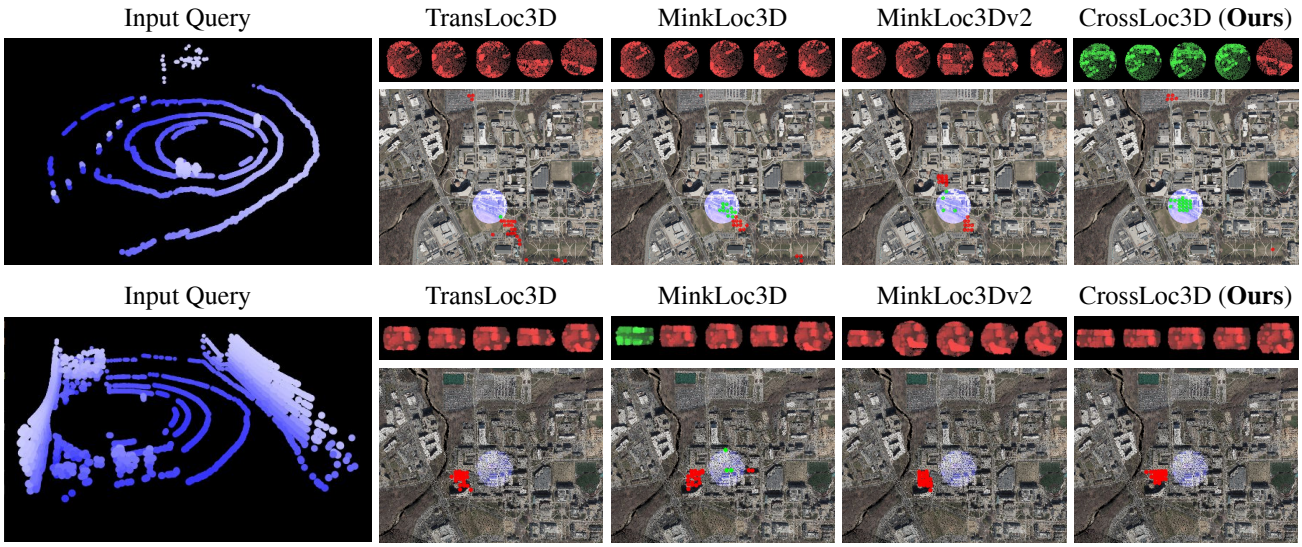

Figure 3: **Two Challenging cases for CROSSLOC3D and other SOTA methods [5, 2, 3]: For each row, left:** Input ground query (blue). **For each row, top:** Top 5 retrievals (ranked from left to right) from the database of aerial data (true neighbor – green, false neighbor – red). **For each row, bottom:** Distributions of ground query and top 25 retrievals.

nition with ranking-based loss and large batch training. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3699–3705, Los Alamitos, CA, USA, aug 2022. IEEE Computer Society. 2, 3

[4] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017. 1

[5] 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, 2021. 2, 3