MinkLoc3D: Point Cloud Based Large-Scale Place Recognition Supplementary Material

Jacek Komorowski
Warsaw University of Technology
Warsaw, Poland
jacek.komorowski@pw.edu.pl

1. Evaluation of different network architectures

This section contains results of experimental evaluation of different variants of MinkLoc3D network. In all cases the network is trained using the Baseline Dataset. The performance is reported as Average Recall@1% (AR@1%) on Oxford dataset and averaged AR@1% for three in-house datasets (U.S., R.A. and B.D.).

Table 1 shows impact of the kernel size in the first convolutional layer (Conv0) in MinkLoc3D network on the inference time and discriminability of the resultant descriptor. Larger kernels, 7x7x7 and above, significantly increase processing time (from 21 ms per cloud for 5x5x5 kernel to 31 ms for 7x7x7 kernel) and generalize worse (decreasing AR@1% on in-house datasets).

	Conv0 filter size	Oxford AR@1%	In-house AR@1%	Runtime per cloud
	1x1x1	97.9	93.0	18 ms
	3x3x3	98.1	93.0	19 ms
*	5x5x5	97.9	93.2	21 ms
	7x7x7	97.6	91.1	31 ms
	9x9x9	97.4	89.8	45 ms

Table 1. Impact of a kernel size of the first convolutional layer (Conv0), measured as Average Recall @1%. Networks are trained on the Baseline Dataset.

Tables 2 and 3 compare performance of deeper and wider architectures. Neither deeper networks with more bottom-up and top-down convolutional blocks, nor wider architecture with increased number of convolutional filters perform better than MinkLoc3D. They have similar performance on Oxford test subset, but noticeably worse on in-house datasets. With significantly larger number of trainable parameters, they tend to overfit more and generalize worse.

2. Visualization of nearest neighbour search results

Figure 1 visualizes nearest neighbour search results using our MinkLoc3D descriptor in Oxford evaluation subset. The leftmost column shows a query point cloud and other columns show its five nearest neighbours in the descriptor space. *dist* is an Euclidean distance in the descriptor space to the query element, TP indicates the correct match (true positive) and FP (false positive) is an incorrect match.

Figure 2 shows failure cases when searching for structurally similar point clouds using our MinkLoc3D descriptor in Oxford evaluation subset.

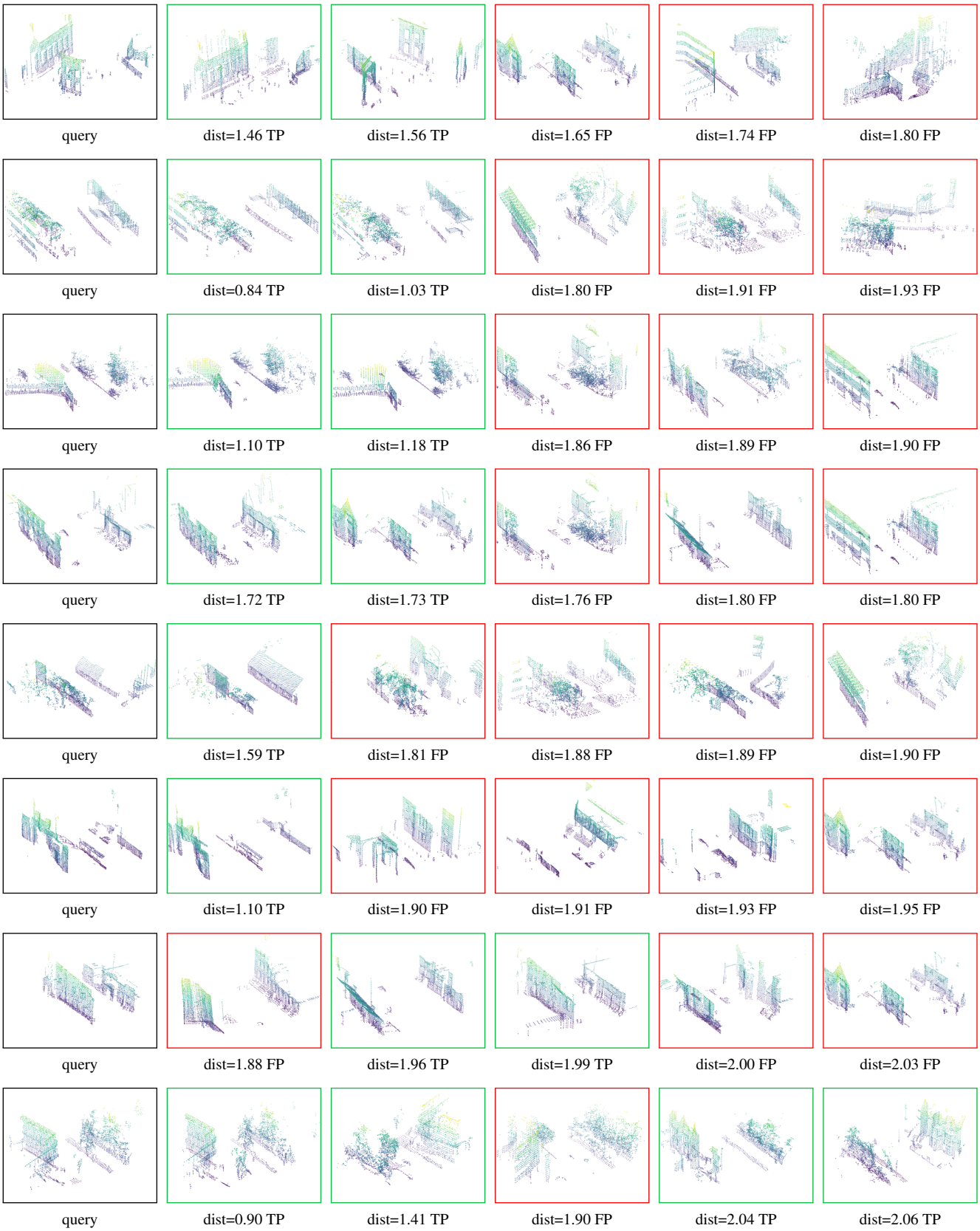


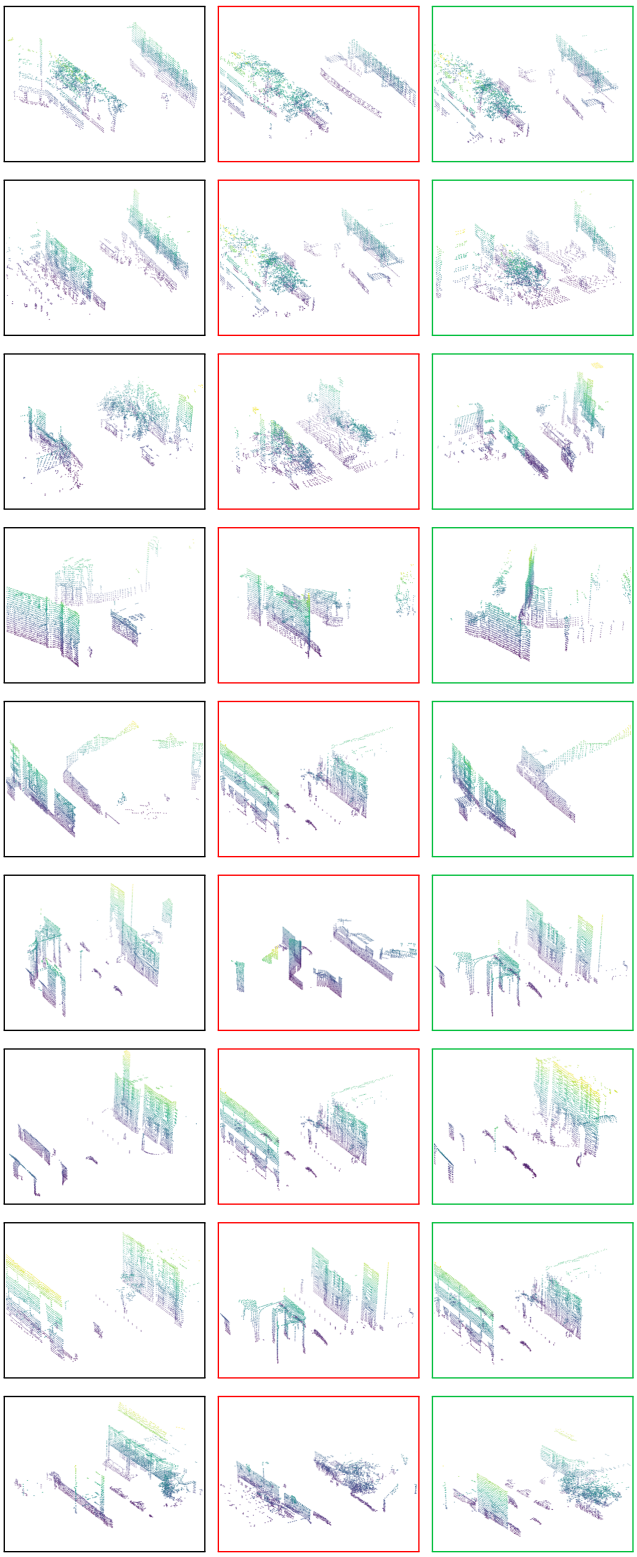
Figure 1. Nearest neighbours search results in the global descriptor space. The leftmost column shows a query point cloud. Other columns show its five nearest neighbours. *dist* is an Euclidean distance in the descriptor space. *TP* indicates correct match (true positive) and *FP* incorrect match (false positive).

Blocks	Oxford AR@1%	In-house AR@1%	Parameters	Runtime per cloud
$\uparrow_3 (32, 32, 64) \downarrow_0 256$	97.0	92.4	0.3M	16 ms
* $\uparrow_4 (32, 32, 64, 64) \downarrow_1 256$	97.9	93.2	1.1M	21 ms
$\uparrow_5 (32, 32, 64, 64, 128) \downarrow_2 256$	98.0	91.8	2.3M	25 ms
$\uparrow_6 (32, 32, 64, 64, 128, 128) \downarrow_3 256$	98.0	92.4	3.9M	26 ms
$\uparrow_7 (32, 32, 64, 64, 128, 128, 256) \downarrow_4 256$	98.0	92.1	7.3M	31 ms

Table 2. Impact of the network depth measured as Average Recall @1%. $\uparrow_i (c_0, c_1, \ldots c_{i-1}) \downarrow_j 256$ denotes an architecture with i bottom-up blocks outputting feature maps with $c_0, c_1, \ldots c_{i-1}$ channels respectively and j top-down blocks each producing 256-channel output. * indicates MinkLoc3D architecture. Networks are trained on the Baseline Dataset.

Blocks	Oxford AR@1%	In-house AR@1%	Parameters	Runtime per cloud
$\uparrow_4 (4, 4, 8, 8) \downarrow_1 256$	93.2	83.4	0.5M	19 ms
$\uparrow_4 (8, 8, 16, 16) \downarrow_1 256$	96.5	89.9	0.6M	20 ms
$\uparrow_4 (16, 16, 32, 32) \downarrow_1 256$	97.2	91.0	0.7M	20 ms
* $\uparrow_4 (32, 32, 64, 64) \downarrow_1 256$	97.9	93.2	1.1M	21 ms
$\uparrow_4 (64, 64, 128, 128) \downarrow_1 256$	98.0	91.8	2.6M	25 ms

Table 3. Impact of the network capacity measured as Average Recall @1%. $\uparrow_i (c_0, c_1, \ldots c_{i-1}) \downarrow_j 256$ denotes an architecture with i bottom-up blocks outputting feature maps with $c_0, c_1, \ldots c_{i-1}$ channels respectively and j top-down blocks each producing 256-channel output. * indicates MinkLoc3D architecture. Networks are trained on the Baseline Dataset.



(a) (b) (c)

Figure 2. Failure cases: Examples of unsuccessful retrieval results using our network. (a) is the query point cloud, (b) incorrect match to the query and (c) the closest true match.