Large-scale Localization Datasets in Crowded Indoor Spaces Supplementary Material

Donghwan Lee[1,*], Soohyun Ryu[1,*], Suyong Yeon[1,*], Yonghan Lee[1,*], Deokhwa Kim[1], Cheolho Han[1], Yohann Cabon[2], Philippe Weinzaepfel[2], Nicolas Guérin[2], Gabriela Csurka[2], and Martin Humenberger[2]

[1]NAVER LABS, [2]NAVER LABS Europe

[1,2]{donghwan.lee, soohyun.ryu, suyong.yeon, yh.l, deokhwa.kim, cheolho.han, yohann.cabon, philippe.weinzaepfel, nicolas.guerin, gabriela.csurka, martin.humenberger}@naverlabs.com

1. Introduction

In this supplementary material, in Section 2 we first provide more details about the structure-based methods we benchmark on our datasets. Then, in Section 3, we provide results on the test sets using Basler cameras as well as results on the validation sets using both, Basler and Galaxy cameras. Furthermore, for the Galaxy test sets we present plots with varying accuracy thresholds. Finally, to provide more insights on our datasets, we show the 4 best and 4 worst localized images.

2. Details of the structure-based methods

For our experiments, we chose a custom pipeline consisting of two stages: mapping and localization. For both, we used different selections of global and local representations. After extracting the local keypoints and robustly matching them using cross validation and geometric verification, we use the point_triangulator of COLMAP[1] to compute the 3D point locations for mapping and the image_registrator of COLMAP to compute the query camera pose from a set of 2D-3D matches. In Table 1, we show the COLMAP parameters we used (taken from [2]). We also evaluated a traditional pipeline, fully based on COLMAP [3], with SIFT [1] features and vocabulary tree based matching. Since we already obtained the camera intrinsic parameters during the calibration process described in Section 3 of the main paper, we fix them for the localization experiments. Figure 1 and 2 show SFM models and dense reconstructions of our datasets.

3. Further benchmark results

We denote the experiments as *GLOBAL+LOCAL*, where we use *GLOBAL* features to generate image pairs used during mapping and localization, and *LOCAL* features for keypoint matching (*e.g.* DenseVLAD+D2-Net). See Section 4.1 of the main paper for more details about the method.

In Table 4 of the main paper, we have shown the results obtained with various modern visual localization algorithms on the Galaxy images from the test set. In Table 2 of this supplementary material, we provide the results for the same methods on the Basler images from the test set, as well as Galaxy and Basler images from the validation set. As in the main paper, we report the percentages of successfully localized images for three thresholds, high (0.1m, 1°), medium (0.25m, 2°), and low (1m, 5°) accuracy.

In addition, for the Galaxy test sets, in Figure 3 of this supplementary material we show the percentages of successfully localized images for some of the methods when varying the thresholds for the positional error between 0 and 1m with an angular error threshold in degree varying as 10 times the positional error threshold in cm, *i.e.*, (10cm, 1°), (20cm, 2°), etc. These plots visually illustrate the comparison between different localization methods and confirm the observations drawn from the tables:

- The results for the Basler images are overall better than the ones for the Galaxy images both, on the test as well as on the validation sets. This is not surprising as the difference in quality and resolution between the images from the industrial Basler cameras and the ones from the smartphone cameras introduces a domain bias between image representations (local and global). Since for mapping we only use Basler cameras, this domain bias affects the localization results of all methods for all datasets.

- Concerning the comparison between different methods, in these tables we observe similar behaviour as in the table shown in the main paper. Hence they confirm the observations discussed in Section 4.3 of

*These authors contributed equally.

[1]https://colmap.github.io

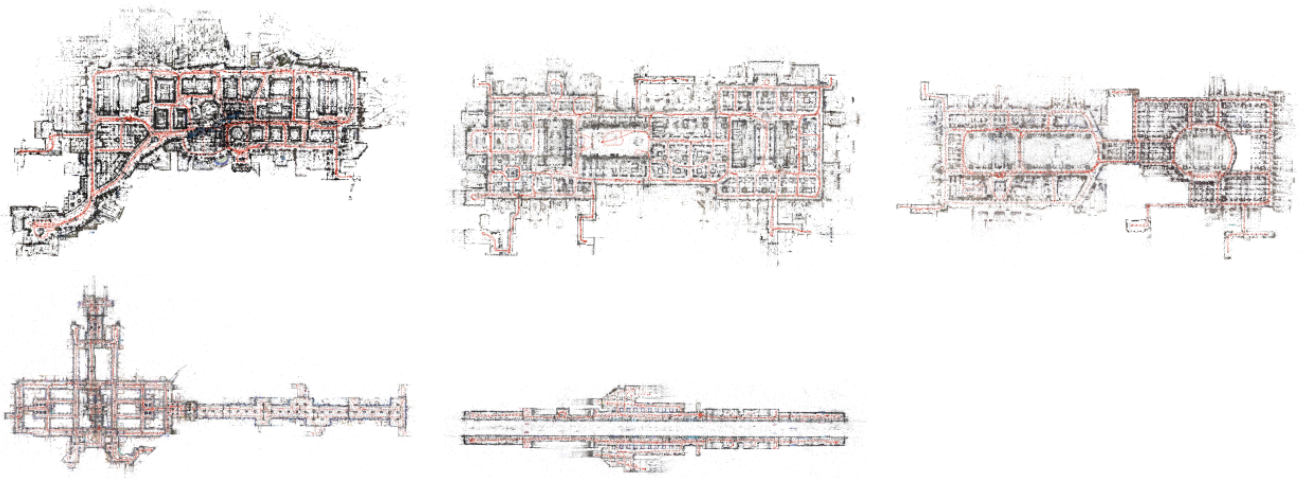

Figure 1. SFM reconstructions of our datasets (camera centers in red). From left to right: Top: Dept. B1, Dept. 1F, Dept. 4F, Bottom: Metro St. B1, Metro St. B2

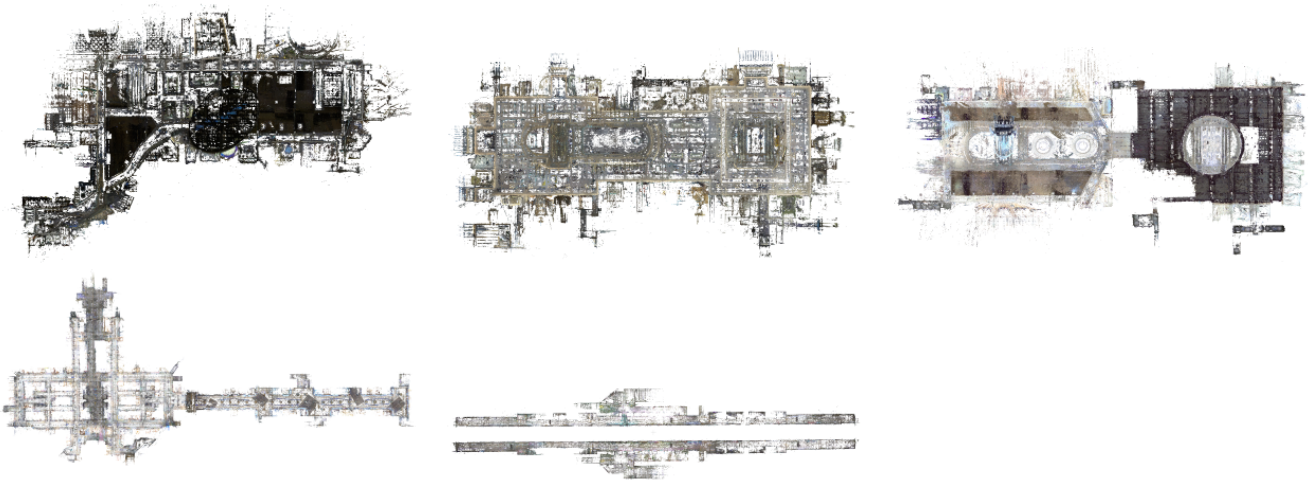

Figure 2. Dense reconstructions of our datasets. From left to right: Top: Dept. B1, Dept. 1F, Dept. 4F, Bottom: Metro St. B1, Metro St. B2

the main paper, namely that the structure-based methods significantly outperform ESAC, and that PoseNet completely fails to localize the query images. Furthermore, we have similar performance when varying the global image representation for retrieval, without having a clear winning representation. Finally, for local features, R2D2 slightly outperforms D2-Net and both yield much better localization results than using SIFT.

- Finally, there is little difference between the results on validation and test sets, showing that the difficulty between the two zones remains similar and hence makes the validation set appropriate to be used for algorithm development, parameter optimization, and model tuning.

Qualitative results. To further analyze the results of DELG+R2D2, one of the best methods according to our experiments, in Figure 4 (resp. Figure 5) we show the 4 images with the highest (resp. lowest) positional error for each of the 5 NAVER LABS localization datasets. We observe that many images that are not localized either lack relevant information, especially in Dept. 4F (which can be often seen in the low freq. score of these images), contain people occluding the images (which can be seen in the crowdedness

COLMAP parameters	triangulator	registrator
—Mapper.ba_refine_focal_length—	0	0
—Mapper.ba_refine_principal_point—	0	0
—Mapper.ba_refine_extra_params—	0	0
—Mapper.min_num_matches—	15	4
—Mapper.init_min_num_inliers—	100	4
—Mapper.abs_pose_min_num_inliers—	30	4
—Mapper.abs_pose_min_inlier_ratio—	0.25	0.05
—Mapper.ba_local_max_num_iterations—	25	50
—Mapper.abs_pose_max_error—	12	20
—Mapper.filter_max_reproj_error—	4	12

Table 1. The parameters we used for COLMAP.

Test set - *Basler* images	Dept. B1			Dept. 1F			Dept. 4F			Metro St. B1			Metro St. B2		
Algorithm / Accuracy th.	0.1m,1°	0.25m,2°	1m,5°	0.1m,1°	0.25m,2°	1m,5°	0.1m,1°	0.25m,2°	1m,5°	0.1m,1°	0.25m,2°	1m,5°	0.1m,1°	0.25m,2°	1m,5°
structure-based methods															
APGeM+SIFT	87.9	91.5	92.8	85.8	87.4	89.0	92.8	94.0	94.5	67.6	75.8	78.8	69.2	72.3	73.8
DELG+SIFT	88.6	92.0	93.3	83.7	85.4	86.6	92.7	93.8	94.1	67.0	74.7	77.8	69.0	71.9	73.8
DenseVLAD+SIFT	91.0	94.0	95.0	87.0	88.3	89.6	92.2	93.4	94.0	68.5	75.7	78.3	69.0	71.9	73.8
NetVLAD+SIFT	90.5	93.6	94.8	85.9	87.4	89.3	93.2	94.2	94.6	65.6	72.9	75.6	70.5	73.5	75.0
SIFT+vocab. tree (COLMAP)	77.1	84.0	85.9	80.1	82.1	83.4	86.0	87.1	87.6	63.6	71.8	75.8	65.8	69.1	70.5
APGeM+D2	92.4	95.6	96.2	91.6	92.3	92.7	92.9	94.0	94.4	70.8	78.8	81.3	67.2	71.2	73.1
DELG+D2	93.2	95.7	96.3	90.5	91.9	92.3	92.5	93.7	94.3	70.1	78.4	81.1	66.6	70.5	72.2
DenseVLAD+D2	94.7	96.6	97.0	90.8	91.9	92.5	92.2	93.6	94.2	70.5	77.9	79.8	66.8	70.1	71.5
NetVLAD+D2	94.6	96.7	97.1	91.7	93.1	93.5	92.8	93.9	94.5	68.5	75.7	78.0	68.7	72.4	73.8
APGeM+R2D2	93.8	96.2	96.7	93.1	94.1	94.8	93.8	94.8	**95.2**	73.2	**80.9**	**83.3**	71.6	73.9	75.5
DELG+R2D2	94.3	96.6	97.0	92.3	93.3	93.9	93.5	94.7	95.1	**73.3**	80.6	82.8	71.0	73.3	74.7
DenseVLAD+R2D2	**95.6**	97.2	97.6	91.8	93.0	93.4	93.5	94.5	94.9	73.1	80.2	82.1	71.1	73.3	74.6
NetVLAD+R2D2	95.4	**97.4**	**97.7**	**93.2**	**94.6**	**95.1**	**93.9**	**94.8**	95.1	71.1	78.1	80.3	**72.9**	**75.3**	**76.8**
ESAC															
1 expert	0.0	0.0	0.6	0.0	0.0	0.9	0.7	6.0	22.7	0.0	0.0	0.0	0.0	0.0	0.0
10 experts	0.6	4.1	13.4	2.9	8.8	17.5	64.8	82.5	88.7	30.6	55.1	68.9	30.8	50.9	65.3
20 experts	4.5	13.5	25.8	10.3	21.9	34.1	71.9	84.9	89.3	36.4	60.7	72.6	35.7	52.2	62.6
50 experts	12.9	25.2	37.0	16.6	30.6	40.3	76.8	86.1	90.0	47.3	67.5	75.5	35.6	53.7	63.0
PoseNet	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.2	0.0	0.0	0.0

Validation set - *Galaxy* images	Dept. B1			Dept. 1F			Dept. 4F			Metro St. B1			Metro St. B2		
Algorithm / Accuracy th.	0.1m,1°	0.25m,2°	1m,5°	0.1m,1°	0.25m,2°	1m,5°	0.1m,1°	0.25m,2°	1m,5°	0.1m,1°	0.25m,2°	1m,5°	0.1m,1°	0.25m,2°	1m,5°
structure-based methods															
APGeM+SIFT	64.7	72.2	78.6	82.3	87.5	92.1	72.6	83.6	97.3	34.6	53.4	63.9	37.8	59.6	65.3
DELG+SIFT	64.0	70.1	77.3	83.5	89.2	94.5	71.9	84.8	98.1	35.2	53.0	64.2	39.2	60.9	66.2
DenseVLAD+SIFT	66.5	73.7	80.8	84.9	89.5	94.8	72.8	85.1	98.8	36.5	53.9	63.7	38.5	60.0	66.8
NetVLAD+SIFT	66.9	73.4	80.8	82.6	89.1	94.1	71.9	84.6	98.1	31.5	47.5	56.6	40.2	62.7	68.2
SIFT+vocab. tree (COLMAP)	64.2	71.6	77.3	82.7	87.1	93.5	72.6	84.9	98.5	33.0	49.3	59.6	31.4	50.7	55.8
APGeM+D2	70.2	78.0	86.1	83.2	89.2	94.5	72.1	85.3	98.5	40.9	61.6	71.2	37.3	60.1	66.6
DELG+D2	69.7	76.5	87.2	85.7	90.3	95.9	72.6	85.8	98.6	41.6	61.8	73.7	38.0	60.6	66.9
DenseVLAD+D2	70.7	77.2	87.1	85.0	89.8	95.1	73.6	**86.3**	98.6	42.6	61.8	71.6	37.1	57.4	63.1
NetVLAD+D2	72.5	**79.2**	88.5	86.0	90.2	95.5	**73.8**	86.0	99.0	36.1	54.1	65.2	38.3	60.5	67.1
APGeM+R2D2	71.6	78.0	86.0	85.8	89.9	94.4	72.6	84.6	98.3	43.1	62.2	72.6	39.4	62.7	67.5
DELG+R2D2	70.6	77.8	87.4	**86.4**	**90.9**	**96.9**	72.3	85.3	98.8	43.4	62.9	**73.7**	39.9	63.1	68.0
DenseVLAD+R2D2	71.9	77.8	87.9	85.8	90.5	96.5	73.0	85.8	**99.3**	**43.9**	**62.9**	72.8	40.1	59.6	64.6
NetVLAD+R2D2	**72.9**	79.0	**89.2**	85.7	90.6	95.9	73.3	84.8	99.0	39.0	56.4	66.8	**41.3**	**63.5**	**68.7**
ESAC															
1 expert	0.0	0.0	0.1	0.0	0.1	3.0	0.0	0.5	9.9	0.0	0.0	0.0	0.0	0.0	0.0
10 experts	0.6	1.8	6.0	17.3	46.7	73.2	28.1	60.5	86.8	1.1	6.6	17.2	4.4	14.8	22.9
20 experts	2.3	5.7	10.8	33.8	61.9	81.2	45.4	70.2	89.2	4.1	13.8	26.8	4.3	13.4	22.5
50 experts	5.4	9.1	14.2	49.7	71.5	84.1	45.2	69.9	85.1	7.9	20.3	32.7	6.0	16.1	24.6
PoseNet	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0

Validation set - *Basler* images	Dept. B1			Dept. 1F			Dept. 4F			Metro St. B1			Metro St. B2		
Algorithm / Accuracy th.	0.1m,1°	0.25m,2°	1m,5°	0.1m,1°	0.25m,2°	1m,5°	0.1m,1°	0.25m,2°	1m,5°	0.1m,1°	0.25m,2°	1m,5°	0.1m,1°	0.25m,2°	1m,5°
structure-based methods															
APGeM+SIFT	80.9	86.3	88.3	91.9	93.2	94.4	97.8	99.9	**100.0**	63.1	71.2	75.5	70.4	75.8	76.9
DELG+SIFT	79.3	84.4	86.2	90.0	91.5	92.5	97.8	99.9	**100.0**	61.6	69.2	73.8	69.0	74.6	75.6
DenseVLAD+SIFT	85.5	90.2	91.2	91.3	92.5	93.0	97.8	99.9	99.9	62.4	70.0	73.6	64.8	69.8	70.6
NetVLAD+SIFT	83.2	88.4	90.0	91.0	92.2	92.9	97.7	**100.0**	**100.0**	58.8	66.5	71.0	67.2	72.4	73.4
SIFT+vocab. tree (COLMAP)	69.8	77.7	80.3	89.5	91.0	92.2	97.6	99.8	99.8	56.2	65.6	70.3	54.9	59.2	60.6
APGeM+D2	87.0	91.8	92.9	95.8	96.6	97.0	97.5	**100.0**	**100.0**	68.0	77.8	81.4	68.7	75.2	76.5
DELG+D2	85.7	89.6	90.4	94.2	95.4	95.7	97.6	99.7	99.9	67.2	76.7	80.0	67.5	73.0	74.2
DenseVLAD+D2	91.5	94.3	94.9	92.9	93.9	94.1	97.3	99.3	99.3	68.4	76.0	79.2	61.9	67.2	68.5
NetVLAD+D2	89.7	93.1	94.1	93.9	95.2	95.7	97.5	99.8	99.9	64.0	72.4	75.9	65.3	70.8	71.5
APGeM+R2D2	89.6	92.6	93.4	**96.4**	**97.2**	**97.4**	**98.0**	**100.0**	**100.0**	70.0	**78.7**	**82.5**	**72.7**	**77.9**	**78.6**
DELG+R2D2	87.1	90.2	91.2	95.7	96.1	96.5	**98.0**	**100.0**	**100.0**	68.4	77.0	80.5	70.7	75.7	76.2
DenseVLAD+R2D2	**92.2**	**95.1**	**95.7**	95.4	96.2	96.3	97.8	99.9	99.9	**70.2**	78.0	81.0	65.5	69.8	70.5
NetVLAD+R2D2	91.1	93.9	94.8	95.6	96.4	97.0	97.8	**100.0**	**100.0**	65.8	73.6	77.2	68.4	72.8	73.3
ESAC															
1 expert	0.0	0.0	0.1	0.0	0.1	2.2	1.2	12.5	55.0	0.0	0.0	0.0	0.0	0.0	0.0
10 experts	0.0	0.6	4.2	19.4	36.4	52.4	86.9	98.2	99.8	10.8	30.1	49.0	24.3	41.2	55.4
20 experts	0.8	3.5	10.4	32.7	48.5	61.2	89.4	98.5	99.8	26.1	48.1	63.7	27.6	45.7	55.8
50 experts	5.4	13.3	23.7	44.8	60.5	69.9	89.7	99.2	99.9	35.8	56.7	68.8	26.7	42.5	52.5
PoseNet	0.0	0.0	0.1	0.0	0.0	0.4	0.0	0.0	0.5	0.0	0.0	0.2	0.0	0.0	0.1

Table 2. Results of various visual localization methods on the 5 NAVER LABS datasets, with the percentages of successfully localized test images within three thresholds for each datasets. The best method is shown in bold, the second best is underlined. We report the results for the Basler images of the test set (top), the Galaxy images of the validation set (middle) and the Basler images of the validation (bottom).

score), or contain large changing elements (*e.g.* large screen with varying content). In contrast, images that are well localized contain lot of high frequency and relevant information. They also contain some dynamic elements but these are not dominant in the images. Note that these observations cannot be generalized because the localization performance also depends on the content of the images. An image with little low frequency content, for example, can still be localized precisely if the combination of visual information can be uniquely described in the dataset and robustly recovered during the localization process.

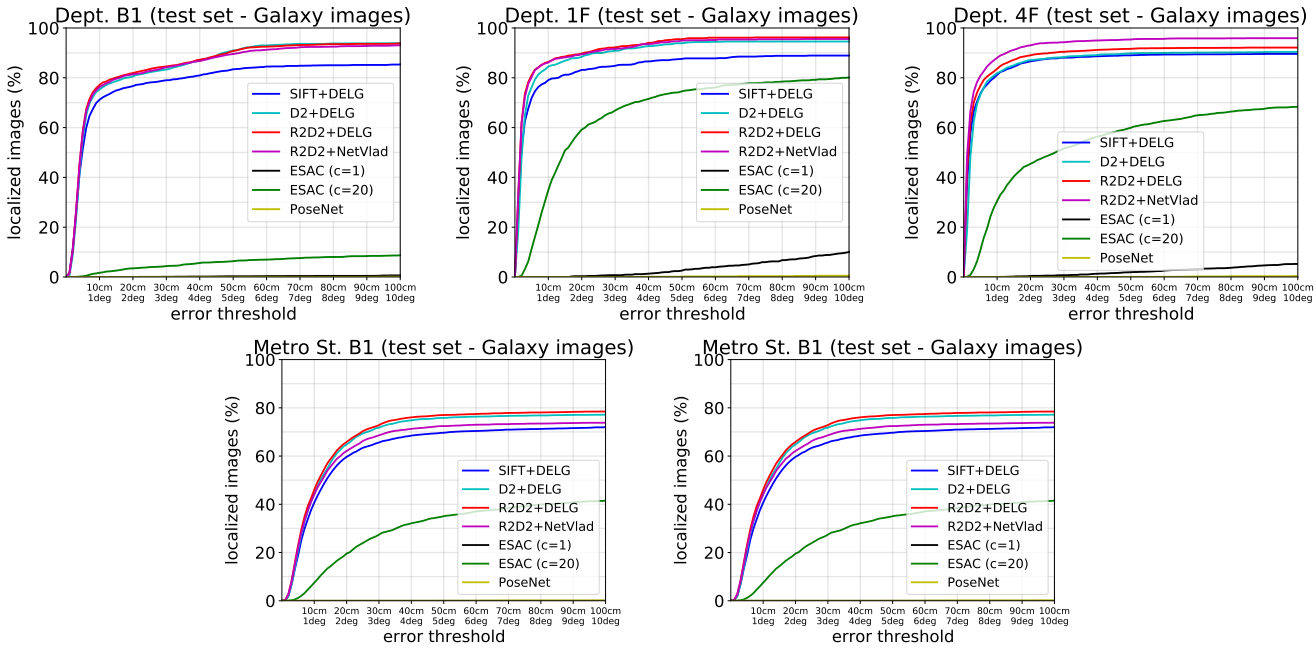

Figure 3. Results with varying error threshold on Galaxy images for all 5 NAVER LABS localization datasets. The angular error threshold in degree varies as 10 times the positional error threshold in cm.

References

[1] David G. Lowe. Distinctive Image Features from Scale-invariant Keypoints. *IJCV*, 2004. 1

[2] Torsten Sattler. Local feature evaluation. https://github.com/tsattler/visuallocalizationbenchmark/tree/master/local_feature_evaluation. 1

[3] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

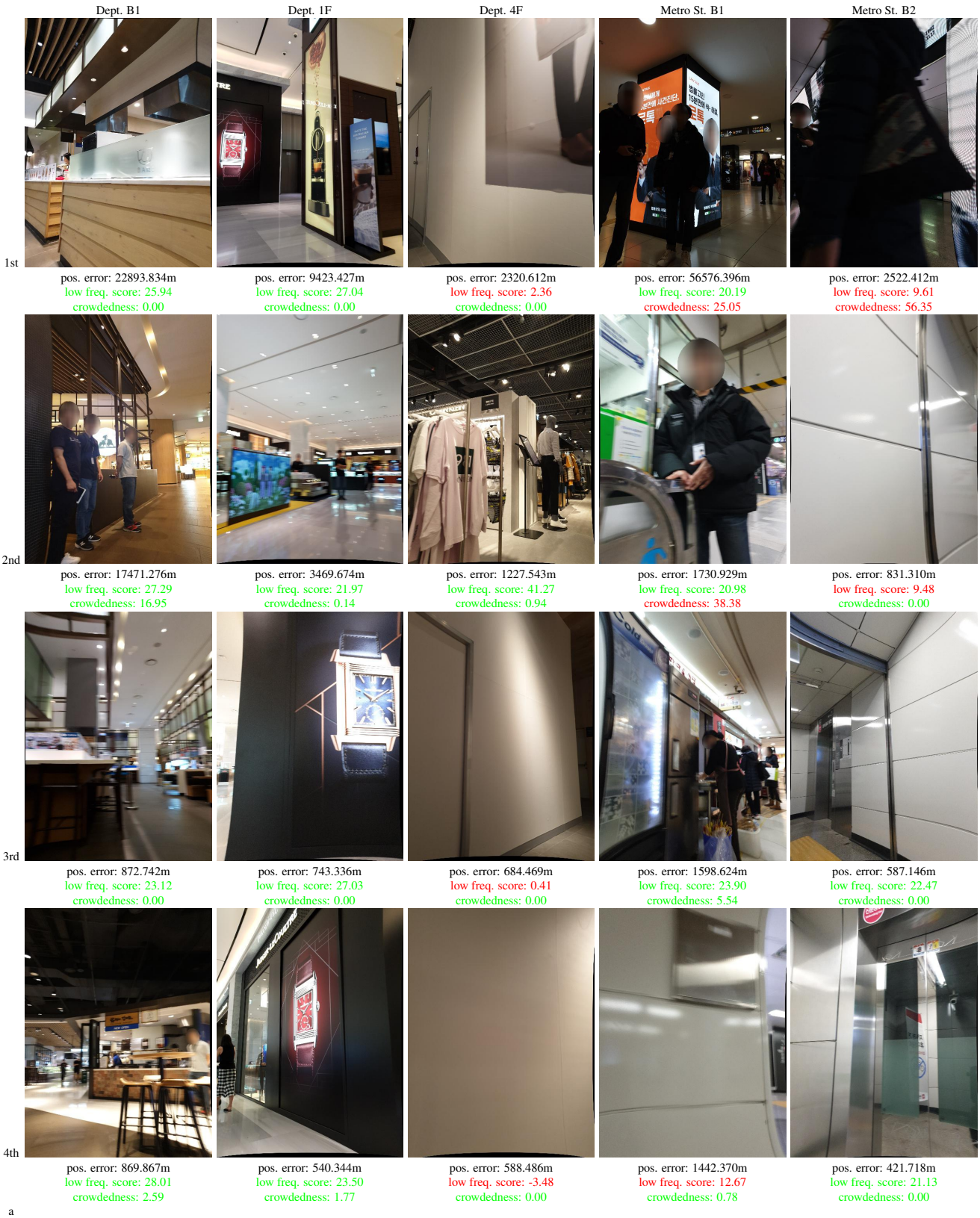

Figure 4. Worst localized images according to the positional error for DELG+R2D2 on the 5 NAVER LABS localization datasets (test set - Galaxy images). Red: low freq. score below 20, crowdedness above 20

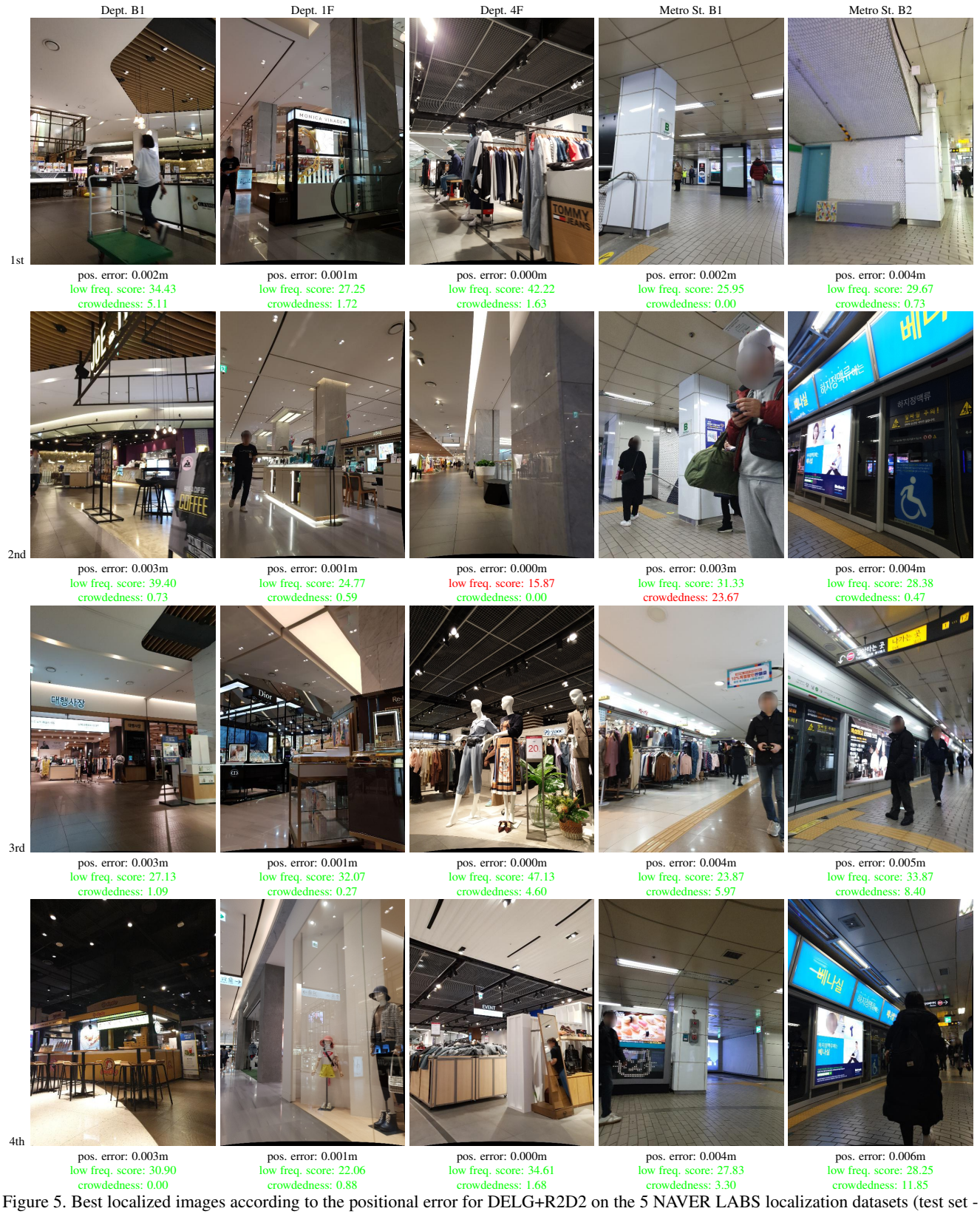

Figure 5. Best localized images according to the positional error for DELG+R2D2 on the 5 NAVER LABS localization datasets (test set - Galaxy images). Red: low freq. score below 20, crowdedness above 20