Beyond Controlled Environments: 3D Camera Re-Localization in Changing Indoor Scenes

Supplementary Material

This supplementary material provides the following information: Sec. 1 shows the individual scenes of our benchmark dataset. Sec. 2 provides statistics about the changes that occur in our benchmark dataset (*c.f.* Sec. 4.1 in the main paper). Sec. 3 discusses the means by which we classify the difficulties of the test frames in our dataset (*c.f.* Sec. 5.1 in the main paper). Sec. 4 provides further details about our DCRE metric (*c.f.* Sec. 4.2 in the main paper). Sec. 5 discusses implementation details for HF-Net and the image retrieval methods we tested (*c.f.* Sec. 5.2 in the main paper), as well as the sequence-based approaches evaluated in the main paper. In addition to this document, we also provide a supplementary video summarizing our paper.

1 Benchmark Visualization

Figs. 6 – 15 show the 3D reconstructions of each of the individual scenes in the *RIO10* dataset. The scenes selected for *RIO10* are very diverse, and exhibit a wide variety of changes, including, but not limited to, complex illumination changes, and appearance variations mostly caused by human interactions, such as rigid object movements (e.g. the movement of major objects such as the bed and sofa in scenes 3 and 4, respectively) and non-rigid object deformations (e.g. the rearrangement of the blankets in scene 3). Our dataset provides 10 train, 10 validation and 54 test sequences, with 52 562 images in the train set, 34 415 images in the validation set and 165 744 in the test set.

2 Change Statistics

Per-scene change statistics corresponding to the change measures described in Sec. 4.1 of the main paper can be found in Fig. 1. It can be seen that on the one hand, scene 4 has the highest semantic (c) and geometric (d) change values (many objects, including a sofa, move in the rescans), whilst on the other hand, scenes 8 and 9 have a low normalised correlation coefficient (a) and a high normalised SSD (b), highlighting the visual differences that they contain (see also the original scans in Figs. 9, 13 and 14).

3 Classifying Frame Difficulty

Variance of Laplacian (VoL) As mentioned in the main paper, the Variance of Laplacian (VoL) measure captures both motion blur and a lack of texture

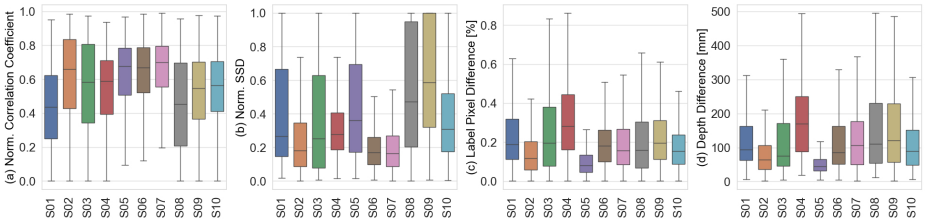

Fig. 1. Visual (a,b), semantic (c) and geometric (d) change statistics for each of the 10 different scenes in our *RIO10* dataset. These are computed by averaging the corresponding change measures over all frames from all test sequences for each scene.

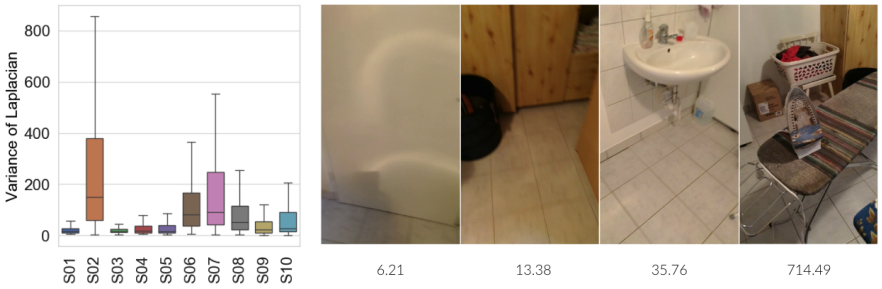

Fig. 2. Left: the average Variance of Laplacian (VoL) value for each test frame, for each scene in our *RIO10* dataset. Right: some example test frames from our dataset, and their VoL values. A low VoL value generally indicates that an image exhibits motion blur or a lack of texture (left two images). A high VoL value generally indicates the opposite (right two images).

in an image. Fig. 2 shows the average of this measure over all frames from all test sequences for each scene. Blurred images, or images with a lack of texture, such as the left two images in the figure, have a low VoL value and often lack features needed for localization, which can sometimes make them difficult even for humans to localize. By contrast, images with a higher VoL value, such as the right two images in the figure, often contain more descriptive features and are therefore expected to be easier for feature-based re-localization algorithms such as Active Search [11] to handle.

Pose Novelty Fig. 3 shows a selection of test images, together with their nearest neighbours in the corresponding training sequences, as computed by using our novel DCRE measure as a pose similarity metric. Image pairs with higher DCREs broadly correspond to test images that were captured from more novel poses.

Field of View/Context The left side of Fig. 4 shows the average field of view/context for a test frame in each scene of our *RIO10* dataset, as per the description of this metric in Sec. 5.1 of the main paper. On the right side of Fig. 4, some example test images with a context of $> 10m^3$ are shown. In our

Fig. 3. Visualizing our pose novelty metric. Top row: test images; bottom row: nearest neighbour training images, as computed by using our novel DCRE measure as a pose similarity metric. The DCRE (in pixels), which is used to capture the pose novelty between each pair of images, is printed below them, as is the fraction of the image diagonal it represents in each case.

Fig. 4. Visualizing our field of view/context metric. Left: the average field of view/context value for each test frame, for each scene in our *RIO10* dataset. Right: some example test frames from our dataset that have particularly high field of view/context values ($> 10m^3$).

experiments in Table 3, it can be seen that methods struggle with low-context frames (compared to medium-context ones). Interestingly, our high-context frames proved more challenging than our medium-context ones on average, potentially due to a combination of factors such as motion blur, lack of texture and large scene element changes in some of our high-context frames.

4 Dense Correspondence Re-Projection Error

DCRE is a dense re-projection error of ground-truth correspondences applied in a novel context, namely the evaluation of camera re-localization. Compared to traditional applications (e.g. camera calibration or bundle adjustment), synthetic depth images are used, which gives us ground-truth correspondences. The DCRE

Table 1. Comparing the performance of the image retrieval methods NetVLAD and DenseVLAD with and without 20-NN interpolation (*c.f.* Table 3 in the main paper).

	Inlier				Outlier		
Method	$\mathcal{E}_{\mathrm{a}}(\mathbf{0.05m}, \mathbf{5}^\circ)$	$(\widetilde{\mathbf{\Delta t}}, \widetilde{\mathbf{\Delta}\theta})$	$\mathcal{E}_{\mathrm{f}}(\mathbf{0.05})$	$\mathcal{E}_{\mathrm{f}}(\mathbf{0.15})$	**N/A**	$\bar{\mathcal{E}}_{\mathrm{a}}(\mathbf{0.5m}, \mathbf{25}^\circ)$	$\bar{\mathcal{E}}_{\mathrm{f}}(\mathbf{0.5})$
NetVLAD [1]	0.0002	(0.93, 31.44)	0.006	0.125	0	0.798	0.452
NetVLAD Interpolation [1]	0.0003	(0.88, 38.36)	0.007	0.0999	0	0.840	0.531
DenseVLAD [14]	0.0003	(0.98, 32.26)	0.008	0.124	0.006	0.772	0.520
DenseVLAD Interpolation [14]	0.0002	(1.00, 50.26)	0.008	0.0967	0.006	0.827	0.612

will thus be 0 for the ground-truth pose. By contrast, the reprojection errors for SfM point clouds will generally be non-zero due to noise in the image measurements. The point clouds generated by rendering a 3D mesh further enable us to compute the metric densely over the whole image rather than being restricted to well-textured regions in the images where features are extracted. Despite its advantages, we are not aware of any re-localization benchmark that uses dense re-projection errors for evaluation.

Please note that we decided to normalize the DCRE error (see eq. 9). While normalization is not strictly necessary for this dataset, as all images have the same resolution, it helps future research to compare errors across datasets. 5px is not much in a 5K image, but might be significant at lower image resolutions.

5 Implementation Details

Details for HF-Net and Image Retrieval We trained HF-Net for $50k$ iterations on the $52k$ training images of *RIO10*. Image retrieval interpolation results (based on the top 20-NN) achieve very similar performance (see Fig. 5 and Table 1).

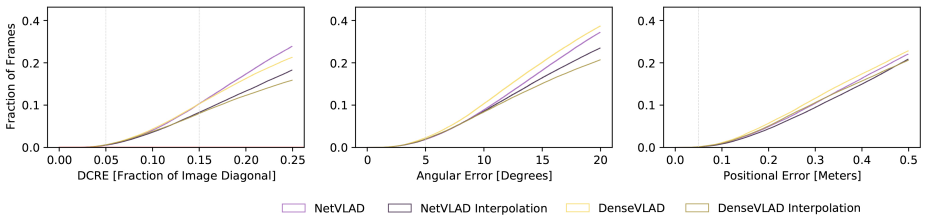

Fig. 5. Cumulative absolute pose recall and DCRE for image retrieval methods.

Implementation Details for Sequence-based Re-Localization The following provides details about the sequence-based re-localization experiments

presented in Fig. 9 of the main paper. We use two different approaches, one for RGB-only (Active Search and D2-Net) and one for RGB-D methods (Grove and Grove v2). For both, a sequence is defined as a consecutive set of frames with known *relative* poses. For all our experiments, we use relative poses defined by the ground truth absolute poses. Note that this does not provide sequence-based methods with any information about where in the scene the images were taken, but it eliminates the impact of pose tracking errors, *e.g.* due to drift in visual odometry or SLAM, from the localization process. As such, the experiments presented in the paper represent an upper bound on the performance of sequence-based approaches. Closing the gap between this upper bound obtained with "perfect" relative poses and relative poses computed by an existing odometry/SLAM system remains an open research question. However, Fig. 9 in the paper shows that considerable gains are possible, which should make practical implementation of sequence-based re-localization an interesting research topic.

For RGB-only methods, we model a sequence of images with known intrinsics and relative poses as a generalized camera [9], *i.e.* as a camera with multiple centers of projection. The 2D-3D matches found for each individual image then allow us to estimate the pose of the generalized camera (*i.e.* of all images in the sequence simultaneously) by applying a minimal solver for the generalized perspective-n-point pose (gPnP) problem [7, 8, 13, 15] inside a RANSAC [5] loop. More precisely, we use a gPnP+s solver [7] that estimates both the pose of the image trajectory and a scale factor, *i.e.*, our approach could account for scale differences between the global 3D model and the trajectory.

For RGB-D methods that process and relocalize each frame independently, we adopt a different approach. Specifically, for each sequence we want to evaluate, we first transform the relocalized pose for each frame (which denotes the estimated transformation from that frame's camera pose to the *origin of the reference scene*) into a pose expressed relative to the *last* frame in the sequence. This computation is done by combining the frames' *relative* poses with the relocalization output. We then cluster the transformed relocalized poses (each of which denotes a possible transformation between the *last frame's camera pose in the current scene* and the *origin of the reference scene*) using an iterative approach. Typically, as a result of the clustering, there will be a single large cluster of poses that are *mutually similar*, and a number of outliers. We return, as the relocalization result for the sequence, the centroid of the largest cluster (computed, for robustness, via dual-quaternion blending [6] of the corresponding poses).

We will release the code for both of these approaches, thus enabling researchers to more easily work on sequence-based localization.

Table 2. Filter setup for evaluation of different challenges on the test / validation images (see Table 3). In the following, ν is the field of view of a frame (as described in Sec. 5.1 of the main paper).

Filter	**# Images**	ρ_v	ζ_s	ζ_g	σ	ν	η
(1) no filter	200 159						
(2) default filter	161 282				> 7.2	$[0.2, 8]$	≤ 650
(3) well-textured	84 946				> 33	$[0.2, 8]$	≤ 650
(4) texture-less	84 704				≤ 33	$[0.2, 8]$	≤ 650
(5) high context	63 041				> 7.2	> 2.4	≤ 650
(6) medium context	62 264				> 7.2	$[0.9, 2.4]$	≤ 650
(7) low context	55 344				> 7.2	≤ 0.9	≤ 650
(8) novel poses	20 281				> 7.2	$[0.2, 8]$	> 500
(9) not novel poses	36 495				> 7.2	$[0.2, 8]$	≤ 150
(10) easy changes	5783	> 0.8	≤ 0.1	≤ 30	> 7.2	$[0.2, 8]$	≤ 650
(11) hard changes	13 363	≤ 0.7	> 0.4	> 30	> 7.2	$[0.2, 8]$	≤ 650

Table 3. Evaluation of the different camera re-localization methods with different setups (described in Table 2); the reported numbers are $\mathcal{E}_f(0.15)$.

Method	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**
Active Search [12]	0.258	0.285	0.388	0.156	0.296	0.303	0.218	0.236	0.442	0.405	0.113
Grove [3]	0.395	0.416	0.471	0.345	0.447	0.423	0.349	0.327	0.631	0.616	0.078
Grove v2 [2]	0.487	0.509	0.570	0.430	0.559	0.514	0.425	0.413	0.715	0.714	0.112
HF-Net [10]	0.103	0.113	0.162	0.054	0.129	0.132	0.063	0.074	0.239	0.226	0.022
HF-Net Trained [10]	0.295	0.320	0.404	0.214	0.354	0.343	0.223	0.229	0.577	0.468	0.115
D2-Net [4]	0.513	0.544	0.608	0.448	0.630	0.559	0.406	0.407	0.775	0.735	0.244
NetVLAD [1]	0.128	0.139	0.162	0.107	0.124	0.156	0.118	0.097	0.299	0.242	0.056
DenseVLAD [14]	0.126	0.136	0.160	0.102	0.135	0.153	0.105	0.110	0.307	0.237	0.049

Fig. 6. 3D reconstructions of scene 1 of our benchmark dataset.

Fig. 7. 3D reconstructions of scene 2 of our benchmark dataset.

Fig. 8. 3D reconstructions of scene 3 of our benchmark dataset.

Fig. 9. 3D reconstructions of scene 4 of our benchmark dataset.

Fig. 10. 3D reconstructions of scene 5 of our benchmark dataset.

Fig. 11. 3D reconstructions of scene 6 of our benchmark dataset.

Fig. 12. 3D reconstructions of scene 7 of our benchmark dataset.

Fig. 13. 3D reconstructions of scene 8 of our benchmark dataset.

Fig. 14. 3D reconstructions of scene 9 of our benchmark dataset.

Fig. 15. 3D reconstructions of scene 10 of our benchmark dataset.

References

1. Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
2. Cavallari*, T., Golodetz*, S., Lord*, N.A., Valentin*, J., Prisacariu, V.A., Stefano, L.D., Torr, P.H.S.: Real-Time RGB-D Camera Pose Estimation in Novel Scenes using a Relocalisation Cascade. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
3. Cavallari, T., Golodetz, S., Lord, N.A., Valentin, J.P.C., di Stefano, L., Torr, P.H.S.: On-the-Fly Adaptation of Regression Forests for Online Camera Relocalisation. IEEE Conference on Computer Vision and Pattern Recognition (2017)
4. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
5. Fischler, M., Bolles, R.: Random Sampling Consensus: A Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography. Communications of the ACM **24**, 381–395 (1981)
6. Kavan, L., Collins, S., O'Sullivan, C., Zara, J.: Dual Quaternions for Rigid Transformation Blending. Tech. Rep. TCD-CS-2006-46, Trinity College Dublin (2006)
7. Kukelova, Z., Heller, J., Fitzgibbon, A.: Efficient Intersection of Three Quadrics and Applications in Computer Vision. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
8. Lee, G.H., Li, B., Pollefeys, M., Fraundorfer, F.: Minimal solutions for the multi-camera pose estimation problem. International Journal of Robotics Research **34**(7), 837–848 (2015)
9. Pless, R.: Using Many Cameras as One. In: IEEE Conference on Computer Vision and Pattern Recognition (2003)
10. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
11. Sattler, T., Leibe, B., Kobbelt, L.: Fast Image-Based Localization using Direct 2D-to-3D Matching. In: International Conference on Computer Vision. pp. 667–674 (2011)
12. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. IEEE Transactions on Pattern Analysis and Machine Intelligence **9** (2017)
13. Sweeney, C., Fragoso, V., Höllerer, T., Turk, M.: gDLS: A Scalable Solution to the Generalized Pose and Scale Problem. In: European Conference on Computer Vision (2014)
14. Torii, A., Arandjelović, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
15. Ventura, J., Arth, C., Reitmayr, G., Schmalstieg, D.: A Minimal Solution to the Generalized Pose-and-Scale Problem. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)