MeshLoc: Mesh-Based Visual Localization
-
Supplementary Material

Vojtech Panek[1,2], Zuzana Kukelova[3], and Torsten Sattler[2]

[1] Faculty of Electrical Engineering, Czech Technical University (CTU) in Prague
[2] Czech Institute of Informatics, Robotics and Cybernetics, CTU in Prague
[3] Visual Recognition Group, Faculty of Electrical Engineering, CTU in Prague

This supplementary provides more details on the experiments presented in the Sec. 4 of the main paper. In particular, Sec. 1 provides more implementation details. Source code for our approach and instructions on how to run it are available[4]. Sec. 2 shows renderings of the four Aachen models (AC13-C, AC13, AC14, AC15) used in the main paper and renderings of the 12 Scenes models provided by [24]. Sec. 2 furthermore provides an extended version of Tab. 1 in the main paper that includes details and rendering times for the 12 Scenes models. Sec. 3 provides a more detailed ablation study that includes results for different inlier thresholds for RANSAC and results for R2D2 [13] features and CAPS [26] descriptors. Sec. 4 shows more detailed results for the 12 Scenes dataset. Sec. 5 discusses the storage requirements and run-time overhead of our method compared to approaches that use on Structure-from-Motion . Sec. 6 contains discussion on the usage of neural scene representations with the MeshLoc pipeline.

We release the dense models for the Aachen Day-Night v1.1 [16,17,28] dataset as well as our renderings for the Aachen and the 12 Scenes [24] datasets[5].

1 Implementation Details

Feature extraction and matching. We use the image-matching-toolbox[6] for both feature extraction and matching.

Pose estimation. As mentioned in the main paper, we use the LO-RANSAC [3, 9] implementation from the PoseLib [8] library with a robust Cauchy loss for non-linear refinement. We run RANSAC for at leat 10k and at most 100k iterations. For position averaging (*cf.* Sec. 3 in the main paper), we use a volume side length of 2m and a step size of 0.25m for Aachen ($d_{\text{vol}} = 1$, $d_{\text{step}} = 0.25$) and a volume side length of 0.5m and a step size of 0.05m for 12 Scenes ($d_{\text{vol}} = 0.25$, $d_{\text{step}} = 0.05$). The step sizes were chosen based on the finest position thresholds used for evaluation on both datasets (0.25m respectively 0.05m). We did not tune any of the parameters involved in the position averaging.

[4] `https://github.com/tsattler/meshloc_release`
[5] `https://data.ciirc.cvut.cz/public/projects/2022MeshLoc/`
[6] `https://github.com/GrumpyZhou/image-matching-toolbox`

Tricolor rendering scheme. The tricolor lighting setup for uncolored models was inspired by the mesh visualization in the RealityCapture[7] software. It consists of three directional lights moving with the camera frame. One light is slightly blue and points into the direction of the camera's vertical axis. The other two lights are slightly yellow, with orientations parallel with the camera's horizontal plane, at 112° and −129° from the positive side of the optical axis. We did not tune the rendering style, but believe that this could be an interesting direction for future work.

Details on the query and database images. For the Aachen Day-Night v1.1 [16,17,28] dataset, we use undistorted database images (where Colmap [18, 19] was used to generate the undistorted images based on the calibration provided by the dataset). We did not undistort the query images but rather remove the distortion from the 2D match positions (where features were extracted from the distorted images) before pose estimatiom. To the best of our knowledge, the query and database images of the 12 Scenes [24] dataset are already undistorted.

2 Renderings

Tab. 1 is an extended version of Tab. 1 in the main paper. Besides details on the model sizes and rendering times for the Aachen Day-Night v1.1 [16, 17, 28] dataset (already presented in the main paper), we provide the same information for the 12 Scenes [24] dataset. Note that we only evaluated the 12 Scenes dataset in its original image resolution (1296×968 pixels).

Figs. 1, 2, 3, and 4 show the four Aachen models (AC13-C, AC13, AC14, AC15) from various views. The AC14 and AC15 models lead to sharper renderings compared to the AC13 model (the AC13-C model uses textures on top of a version of the AC13 model simplified using [7]). However, there is also more noise, in the form of floating blobs of geometry (*cf.* Fig. 3), especially for the AC14 model. These artifacts can reduce localization performance, as noted in the main paper. We also attempted to created textured versions of the AC13, AC14, and AC15 but ran out of memory when applying [25] on these significantly larger models.

Fig. 5 shows visualizations of the colored meshes for each of the scenes in the 12 Scenes dataset.

3 Experiments on Aachen Day-Night

This section provides a more detailed version of the results presented in Sec. 4 of the main paper. Note that we only provide results for the case where the images have a maximum side length of 800 pixels. While Tab. 2 in the paper also evaluates different features on full resolution images, the purpose of that experiment was to show that the MeshLoc pipeline can achieve a similar accuracy as

[7] `https://www.capturingreality.com/`

Table 1. Statistics for the 3D meshes used for experimental evaluation as well as rendering times for different rendering styles and resolutions

	Model	Style	Size [MB]	Vertices	Triangles	Render time [μs] 800 px	Render time [μs] full res.
Aachen v1.1	AC13-C	textured	645	$1.4 \cdot 10^6$	$2.4 \cdot 10^6$	1143	1187
	AC13-C	tricolor	47	$1.4 \cdot 10^6$	$2.4 \cdot 10^6$	115	219
	AC13	colored	617	$14.8 \cdot 10^6$	$29.3 \cdot 10^6$	92	140
	AC13	tricolor	558	$14.8 \cdot 10^6$	$29.3 \cdot 10^6$	97	152
	AC14	colored	1234	$29.4 \cdot 10^6$	$58.7 \cdot 10^6$	100	139
	AC14	tricolor	1116	$29.4 \cdot 10^6$	$58.7 \cdot 10^6$	93	205
	AC15	colored	2805	$66.8 \cdot 10^6$	$133.5 \cdot 10^6$	98	137
	AC15	tricolor	2538	$66.8 \cdot 10^6$	$133.5 \cdot 10^6$	97	160
12 Scenes	apt1/kitchen	colored	58	$1.4 \cdot 10^6$	$2.7 \cdot 10^6$	-	133
	apt1/kitchen	tricolor	52	$1.4 \cdot 10^6$	$2.7 \cdot 10^6$	-	106
	apt1/living	colored	99	$2.4 \cdot 10^6$	$4.7 \cdot 10^6$	-	146
	apt1/living	tricolor	90	$2.4 \cdot 10^6$	$4.7 \cdot 10^6$	-	107
	apt2/bed	colored	83	$2.0 \cdot 10^6$	$3.9 \cdot 10^6$	-	154
	apt2/bed	tricolor	75	$2.0 \cdot 10^6$	$3.9 \cdot 10^6$	-	225
	apt2/kitchen	colored	70	$1.7 \cdot 10^6$	$3.3 \cdot 10^6$	-	107
	apt2/kitchen	tricolor	63	$1.7 \cdot 10^6$	$3.3 \cdot 10^6$	-	135
	apt2/living	colored	136	$3.3 \cdot 10^6$	$6.4 \cdot 10^6$	-	134
	apt2/living	tricolor	123	$3.3 \cdot 10^6$	$6.4 \cdot 10^6$	-	137
	apt2/luke	colored	144	$3.5 \cdot 10^6$	$6.8 \cdot 10^6$	-	128
	apt2/luke	tricolor	130	$3.5 \cdot 10^6$	$6.8 \cdot 10^6$	-	132
	office1/gates362	colored	122	$3.0 \cdot 10^6$	$5.7 \cdot 10^6$	-	127
	office1/gates362	tricolor	110	$3.0 \cdot 10^6$	$5.7 \cdot 10^6$	-	131
	office1/gates381	colored	171	$4.1 \cdot 10^6$	$8.1 \cdot 10^6$	-	110
	office1/gates381	tricolor	155	$4.1 \cdot 10^6$	$8.1 \cdot 10^6$	-	103
	office1/lounge	colored	139	$3.4 \cdot 10^6$	$6.6 \cdot 10^6$	-	124
	office1/lounge	tricolor	126	$3.4 \cdot 10^6$	$6.6 \cdot 10^6$	-	133
	office1/manolis	colored	157	$3.8 \cdot 10^6$	$7.4 \cdot 10^6$	-	116
	office1/manolis	tricolor	142	$3.8 \cdot 10^6$	$7.4 \cdot 10^6$	-	119
	office2/5a	colored	122	$2.9 \cdot 10^6$	$5.8 \cdot 10^6$	-	117
	office2/5a	tricolor	110	$2.9 \cdot 10^6$	$5.8 \cdot 10^6$	-	106
	office2/5b	colored	170	$4.1 \cdot 10^6$	$8.0 \cdot 10^6$	-	108
	office2/5b	tricolor	153	$4.1 \cdot 10^6$	$8.0 \cdot 10^6$	-	104

SfM-based methods. Still, we do not consider experiments on full resolution images essential to the ablation study details presented in the following. Note that due to RANSAC's random nature and the fact that we re-ran the experiments for this more detailed ablation study, the results can differ (slightly) from those reported in the main paper.

Tab. 2 extends the results for reduced resolution images from Tab. 2 in the main paper by providing results for different inlier thresholds for RANSAC and results for R2D2 [13] features and CAPS [26] descriptors extracted around SuperPoint [5] (SP) features (denoted as CAPS+SP). We note that both R2D2 and CAPS+SP perform similarly well or slightly worse than the features evaluated in the main paper (which was the reason why their results were not shown in the main paper). In the main paper, we used the following inlier thresholds: 6 pixels for SuperGlue and LoFTR, 12 pixels for Patch2Pix+SuperGlue, and 20

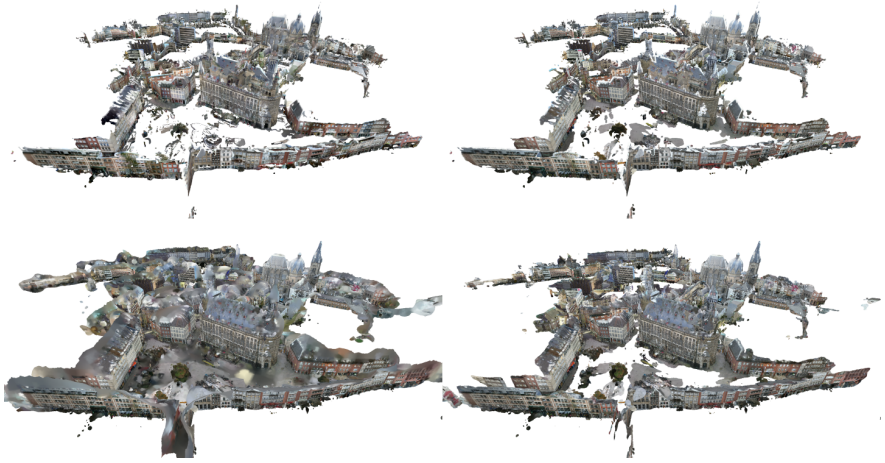

Fig. 1. Comparison of the used meshes generated from Aachen database images - bird view from market square. Top-left: AC13-C (textured), top-right: AC13 (colored per vertex), bottom-left: AC14 (colored per vertex), bottom-right: AC15 (colored per vertex)

pixels for Patch2Pix. As can be seen from Tab. 2, the choice of the threshold is not too critical for most features.

Tabs. 3 and 4 extend the results from Tab. 4 (*cf.* Sec. 4 in the main paper) by showing results obtained by the simple variant of MeshLoc, which uses all individual matches and position averaging without covisibility filtering, for rendered images. Besides the results for SuperGlue and Patch2Pix+SuperGlue (Patch2Pix+SG), we also show results for LoFTR, Patch2Pix, R2D2, and CAPS+SuperPoint (CAPS+SP). We further show results for varying inlier thresholds.[8] As can be seen from the tables, R2D2 essentially fails for the ambient occlusion and tricolor rendering styles, and also shows worse performance than the other features for colored and textured renderings. We further note that CAPS+SP and LoFTR typically perform worse than SuperGlue and both Patch2Pix variants for the two non-realistic rendering styles (ambient occlusion and tricolor). At the same time, LoFTR outperforms the other features on colored and textured renderings, reaching close to the same performance as on real images (78.5%/93.2%/99.5% on real images (*cf.* Tab. 2 in the main paper and Tab. 2 in this supp. mat.) vs. 78.0%/89.0%/95.8% for AC15 and an inlier threshold of 12 pixels (*cf.* Tab. 4)). This shows that there is limited room for improvement using more realistic rendering techniques such as NeRFs [11]. This result might indicate that LoFTR might focuse more on textures and color patterns than on

[8] For Tab. 4 in the main paper, we used the following thresholds: 12 pixels for SuperGlue on all models and for all rendering styles and 12 pixels for Patch2Pix+SuperGlue except for the colored / textured renderings, where we used a threshold of 6 pixels.

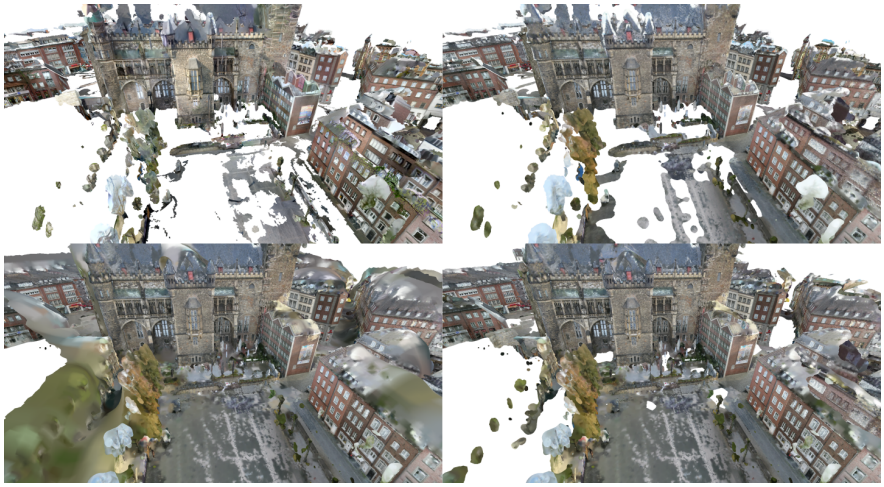

Fig. 2. Comparison of the used meshes generated from Aachen database images - view at the townhall. Top-left: AC13-C (textured), top-right: AC13 (colored per vertex), bottom-left: AC14 (colored per vertex), bottom-right: AC15 (colored per vertex)

shapes and counters (only the latter two are visible in the ambient occlusion and tricolor renderings). The best performance for the ambient occlusion and tricolor rendering styles is typically obtained by Patch2Pix and Patch2Pix+SuperGlue. We attribute this good performance to both methods using a backbone network pre-trained on ImageNet [4], which in our experience seems to lead to features that generalize quite well to unseen conditions. Overall, as mention in the main paper, a higher level of geometric detail (AC14 and AC15) leads to better results.

4 Experiments on 12 Scenes

Sec. 4 of the main paper only provides results for the average number of images localized within a given error threshold for the 12 Scenes [24] dataset. Tabs. 5, 6, and 7 provided detailed measurements per scene, as well as the average number of localized images, for different error thresholds. We use the evaluation toolkit provided by [1] and the original pseudo ground truth provided by the 12 Scenes dataset. We compare the MeshLoc results obtained with SuperGlue with the baseline methods used in [1]: Active Search [15] and DenseVLAD+R2D2 [6, 13, 22] (DVLAD+R2D2) both use a SfM-based scene representation. While Active Search uses SIFT [10] features, DVLAD+R2D2 uses R2D2 [13] features. DVLAD+R2D2 uses the same top-20 images retrieved by DenseVLAD [22] as our method. DVLAD+R2D2 (+D) is a variant that uses the depth images provided by the dataset to obtain 3D points instead of a SfM-based model. DSAC* [2] is a state-of-the-art scene coordinate regressor that predicts a 2D-3D match for each pixel in an image. We compare against an RGB-only

Fig. 3. Comparison of the used meshes generated from Aachen database images - view at houses at market square. Top-left: AC13-C (textured), top-right: AC13 (colored per vertex), bottom-left: AC14 (colored per vertex), bottom-right: AC15 (colored per vertex)

version (DSAC*) and a version based on RGB-D query images (DSAC* (+D)). Please see [1] for more details on the methods.

We notice that on many scenes, MeshLoc performs close to the baselines when using real images instead of renderings. However, there are some scenes, *e.g.*, apt1/living, apt2/luke, office1/lounge, and office2/5b, for which MeshLoc performs noticeably worse than the baselines. As can be seen in Fig. 6, we notice that for these scenes, the alignment between the RGB database images and the scene geometry is not very accurate. As a result, obtaining 3D points from depth maps by rendering the scene model leads to shifted 3D points, which ultimately leads to less accurate poses. However, as can be seen in Tab. 7, MeshLoc is still able to localize nearly all query images at reasonable precision when using either real database images or colored renderings.

5 Run-time and storage consumption

SfM-based methods store 3D point positions and visibility information. Based on numbers from the authors, this is 192MB for LoFTR and 84MB for SuperGlue for Aachen, while our tricolor AC13-C model only requires 47MB (plus an additional 0.18MB for storing camera poses). Storing the descriptors of the SfM points is often more expensive than storing the original images (7.36GB *vs.* between 4.5 and 5GB for Aachen). Rendering images rather than storing them further reduces memory requirements: the colored AC15 model uses 2.8GB (*cf.* Tab. 1) at a similar pose accuracy (*cf.* Tab. 4). Using dense meshes can thus reduce memory consumption.

Fig. 4. Comparison of the used meshes generated from Aachen database images - detail view of one of the houses. Top-left: AC13-C (textured), top-right: AC13 (colored per vertex), bottom-left: AC14 (colored per vertex), bottom-right: AC15 (colored per vertex)

For each of the top-k retrieved images, MeshLoc renders a depth map (and potentially an image) in time T_{R}, extracts features from the retrieved database image in time T_{db}, and matches these features against features extracted from the query image in time T_{M}. Ignoring the retrieval stage (which is shared by methods based on SfM point clouds), this results in an overall time of $T_{\mathrm{q}} + k \cdot (T_{\mathrm{R}} + T_{\mathrm{db}} + T_{\mathrm{M}})$, where T_{q} is the time needed for query feature extraction. SfM-based methods need either $T_{\mathrm{q}} + k \cdot T_{\mathrm{M}}$ when using pre-extracted features or $T_{\mathrm{q}} + k \cdot (T_{\mathrm{db}} + T_{\mathrm{M}})$ when computing features on the fly to save memory. LoFTR and Patch2Pix(+SG) extract their features as part of the matching process as both are based on densely extracted features, resulting in times $k \cdot (T_{\mathrm{R}} + T_{\mathrm{M}})$ (MeshLoc) respectively $k \cdot T_{\mathrm{M}}$ (SfM). For Aachen, we have $k = 50$ and $T_{\mathrm{R}} \approx 0.2$ms. Using Patch2Pix+SG and AC15, MeshLoc thus requires less memory than SfM-based methods at an overhead of only 10ms, while performing similar to using the orig. images (*cf.* Tab. 4).

6 Using neural-based rendering techniques

Our implementation uses an OpenGL rendering pipeline, which is a well-matured technology optimized for real-time performance and use of GPUs, allowing us to render on-the-fly. The MeshLoc pipeline can readily be used with any rendering technique that provides images and depth maps, such as Neural Radience Field (NeRF). Preliminary experiments with a recent NeRF implementation [12] resulted in realistic renderings for the 12 Scenes dataset [24]. We did not obtain good depth maps using NeRF, but upcoming work, *e.g.*, [20], promises to solve

Table 2. Ablation study on the Aachen Day-Night v1.1 dataset [16, 17, 28] using real images at reduced (max. side length 800 px) resolution, and depth maps rendered using the AC13 model. We evaluate different strategies for obtaining 2D-3D matches (using all individual matches (I), merging matches (M), or triangulation (T)), with and without covisibility filtering (C), and with and without position averaging (PA) for various local features. We vary the inlier threshold t used in RANSAC (in pixels). We report the percentage of nighttime query images localized within 0.25m and 2° / 0.5m and 5° / 5m and 10° of the ground truth pose

t	2D-3D	C	PA	SuperGlue (SG) [14]	LoFTR [21]	Patch2Pix (P2P) [29]	P2P + SG [29]	R2D2 [13]	CAP+SP [26]
6.0	I			72.8/92.1/99.0	77.5/92.1/99.5	68.6/88.5/96.3	74.9/91.6/100.0	69.1/81.7/92.1	71.7/90.1/96.3
	I		✓	73.8/92.7/99.0	78.5/92.7/99.5	70.2/89.5/96.3	74.3/91.6/100.0	70.7/82.2/92.1	69.6/90.1/96.3
	M		✓	72.3/92.7/99.5	77.0/92.7/99.0	70.2/87.4/96.3	73.3/92.7/100.0	67.5/83.8/91.6	68.1/88.0/96.3
	M	✓	✓	71.7/92.7/99.5	75.9/91.6/99.5	69.6/88.0/97.9	72.8/91.6/99.5	67.0/88.5/99.0	70.7/90.1/97.9
	T		✓	70.7/89.0/97.4	74.3/90.6/98.4	62.3/81.2/95.3	73.3/89.5/97.9	63.9/80.6/94.8	62.8/85.3/95.8
	T	✓	✓	70.2/89.0/98.4	74.3/91.1/99.0	63.4/81.7/95.8	73.3/89.5/97.9	63.4/80.6/95.3	63.4/85.3/96.9
12.0	I			72.3/92.7/99.0	77.5/92.1/99.5	72.3/89.0/96.3	72.3/91.6/100.0	70.7/84.3/92.1	69.6/89.0/97.4
	I		✓	72.8/93.2/99.0	77.5/92.7/99.5	74.9/90.1/96.3	73.3/92.1/100.0	71.7/84.3/92.1	69.6/89.5/97.4
	M		✓	73.3/92.1/99.5	77.0/92.1/99.5	69.1/88.0/96.3	73.8/92.7/99.5	68.6/85.3/93.7	66.0/88.5/96.9
	M	✓	✓	72.8/92.7/99.5	76.4/91.6/99.5	69.6/90.1/97.9	72.8/93.2/100.0	68.6/85.9/99.0	67.0/89.0/97.4
	T		✓	70.2/89.5/97.9	75.4/92.1/98.4	63.4/82.7/94.8	70.7/90.6/97.4	62.8/80.1/92.7	62.8/85.9/95.3
	T	✓	✓	69.6/89.5/99.5	74.9/92.1/98.4	62.3/82.7/95.8	71.2/90.6/97.4	62.3/80.1/93.2	62.3/83.2/96.9
20.0	I			72.8/92.1/99.0	78.0/92.7/99.5	73.3/90.1/96.3	72.3/91.1/99.5	68.6/84.3/91.6	69.1/89.0/96.9
	I		✓	72.8/92.1/99.0	77.5/92.7/99.5	76.4/90.6/96.3	72.3/90.6/99.5	69.1/84.3/91.6	69.6/88.5/96.9
	M		✓	74.3/91.6/99.5	77.5/92.1/99.5	68.1/89.0/95.8	71.2/92.7/99.5	67.5/84.3/92.7	62.8/88.5/96.9
	M	✓	✓	72.8/91.1/99.0	77.5/92.1/99.5	69.1/91.1/97.9	70.2/92.1/99.5	68.1/86.4/97.9	63.9/88.5/97.9
	T		✓	69.1/89.5/97.4	72.8/91.6/98.4	60.7/81.7/93.7	70.7/89.0/97.9	61.3/78.5/92.1	58.6/83.8/95.8
	T	✓	✓	68.6/89.5/97.4	73.8/92.7/98.4	61.8/81.7/94.8	70.2/89.0/97.4	59.7/79.6/92.1	59.2/82.7/96.9
24.0	I			73.3/92.1/99.5	78.5/92.7/99.5	72.3/91.1/96.3	71.7/90.6/99.5	69.6/83.8/92.1	68.6/89.0/96.3
	I		✓	73.3/92.7/99.5	77.5/92.7/99.5	75.4/91.1/96.3	72.3/91.6/99.5	69.6/83.8/92.1	68.6/88.5/96.3
	M		✓	72.8/91.1/99.5	78.0/92.1/99.5	69.1/89.0/97.4	71.2/93.2/99.5	66.5/85.9/93.7	60.7/88.0/96.3
	M	✓	✓	71.2/91.6/99.0	77.5/92.1/99.5	67.0/90.6/97.9	71.2/92.7/99.5	66.5/85.9/96.9	62.8/88.5/97.9
	T		✓	68.1/87.4/97.4	73.8/91.1/98.4	58.6/81.7/92.7	70.2/90.1/97.4	62.8/78.0/90.6	57.1/82.7/96.3
	T	✓	✓	67.5/89.0/97.4	74.3/92.7/98.4	59.7/82.7/94.8	70.2/89.0/97.4	61.3/76.4/91.6	58.1/81.7/95.8
32.0	I			72.8/91.1/99.0	78.5/92.1/99.0	70.7/91.6/96.3	71.2/90.6/99.0	68.1/83.2/91.6	67.0/90.1/95.8
	I		✓	72.8/91.1/99.0	77.5/92.1/99.0	73.8/92.1/96.3	72.8/91.1/99.0	68.1/83.2/91.6	67.5/90.1/95.8
	M		✓	71.2/89.5/99.5	77.0/91.6/99.5	68.1/89.0/96.9	71.2/92.1/99.5	66.0/83.8/93.7	60.7/89.0/96.9
	M	✓	✓	70.7/90.6/99.5	77.0/91.6/99.5	68.1/90.1/97.9	70.2/91.1/99.5	66.0/83.2/95.8	59.7/89.0/97.4
	T		✓	70.2/85.9/97.4	72.8/89.0/98.4	57.6/80.6/93.2	69.1/89.0/96.9	61.3/75.9/90.1	56.0/81.2/95.3
	T	✓	✓	69.1/86.9/97.9	73.3/91.1/98.4	58.1/81.2/93.7	68.6/90.1/97.4	60.7/76.4/90.6	55.5/81.7/94.2
48.0	I			72.3/91.1/97.4	78.5/92.1/99.0	69.6/89.0/94.8	71.2/91.1/99.0	67.0/80.6/89.5	67.0/88.5/94.2
	I		✓	72.8/91.1/97.4	78.0/92.1/99.0	71.7/89.0/94.8	71.2/91.1/99.0	67.5/80.6/89.5	66.5/88.5/94.2
	M		✓	71.2/88.0/99.0	77.0/92.7/99.5	62.8/86.4/95.3	72.3/92.7/99.5	65.4/80.6/91.6	58.1/87.4/95.3
	M	✓	✓	70.7/89.0/99.5	76.4/92.7/99.0	64.4/89.5/96.9	71.7/92.7/99.5	63.9/83.8/94.2	56.5/87.4/95.3
	T		✓	66.5/84.3/97.4	72.3/89.5/97.4	57.1/78.5/92.1	65.4/86.4/97.4	55.5/69.1/86.4	52.4/76.4/91.1
	T	✓	✓	67.5/84.8/96.3	72.8/92.7/97.4	57.6/80.1/92.7	66.5/86.9/97.4	58.1/71.2/88.0	52.4/77.0/92.1

this issue. Another issue is the scalability of the neural scene representations, which is being targeted in several recent publications, *e.g.*, [23, 27]. We're excited by the recent progress of Neural Rendering and we believe, that the current interest of the community will push the performance of Neural Rendering at the level of standard pipelines in terms of quality, speed and will surpass the standard SfM meshes in terms of memory efficiency in the coming years. We're watching the state of Neural Rendering and want to pursue their use for the task in following publications.

Fig. 5. Comparison of the used meshes generated from 12 Scenes dataset database images. From top-left: apt1/kitchen, apt1/living (on right), apt2/bed, apt2/kitchen, apt2/living, apt2/luke, office1/gates362, office1/gates381, office1/lounge, office1/manolis, office2/5a, office2/5b

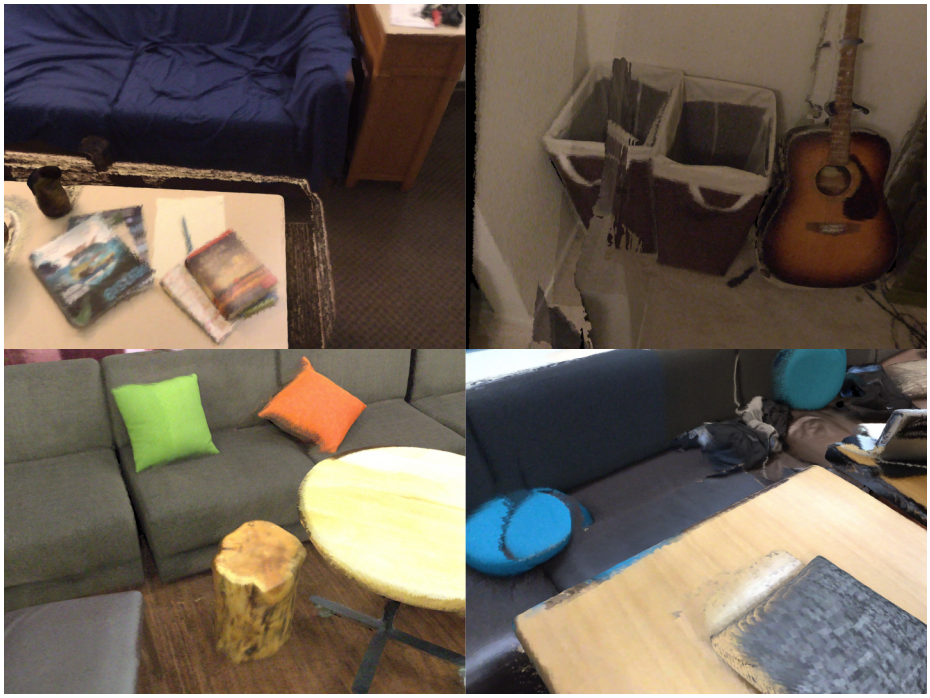

Fig. 6. For some scenes of the 12 Scenes [24] dataset, there are (slight) misalignments between the RGB images and the scene geometry. We show these misalignments for colored renderings from the apt1/living, apt2/luke, office1/lounge, and office2/5b scenes.

Table 3. Ablation study on the Aachen Day-Night v1.1 dataset [16,17,28] using images rendered at reduced resolution (max. 800 px) from 3D meshes of different levels of detail (*cf.* Tab. 1) and different rendering types (textured / colored, raw geometry with ambient occlusion (AO), raw geometry with tricolor shading (tricolor)). We report results for a MeshLoc variant that uses individual matches and position averaging, but no covisibility filtering. We vary the inlier threshold t used in RANSAC

AC13-C:	t	textured	AO	tricolor
SuperGlue [14]	6.0	71.7/91.1/99.0	0.5/1.0/15.7	6.3/19.4/39.8
	12.0	71.2/92.1/99.0	1.0/2.1/17.3	5.8/20.4/45.0
	20.0	70.2/92.7/99.0	1.0/2.1/17.3	5.2/22.0/44.5
	24.0	70.2/92.1/99.0	0.5/2.1/17.3	6.3/23.0/43.5
	32.0	70.2/92.1/99.0	0.5/1.6/14.7	7.3/19.4/39.8
	48.0	69.6/92.1/98.4	0.5/1.0/10.5	4.2/15.2/37.7
LoFTR [21]	6.0	74.9/90.1/99.0	0.0/0.0/3.7	1.6/11.0/37.7
	12.0	74.3/91.1/99.0	0.0/0.0/4.7	1.6/12.6/41.9
	20.0	73.8/91.1/98.4	0.0/0.0/6.8	2.6/13.6/41.9
	24.0	73.8/91.6/98.4	0.0/0.0/5.8	2.1/14.1/38.7
	32.0	74.3/91.1/98.4	0.0/0.0/4.2	1.6/13.1/36.6
	48.0	74.3/90.6/97.4	0.0/0.0/2.1	1.0/7.9/30.9
Patch2Pix [29]	6.0	65.4/87.4/93.2	1.6/4.7/25.1	4.7/21.5/59.2
	12.0	66.5/87.4/94.2	2.1/6.8/30.4	9.4/28.3/63.4
	20.0	64.9/85.9/93.7	2.6/7.9/30.4	8.4/31.9/65.4
	24.0	65.4/85.3/94.2	4.2/10.5/30.4	8.4/31.9/66.5
	32.0	62.8/84.8/93.7	1.6/7.3/28.3	7.9/27.7/63.4
	48.0	62.8/83.2/93.2	1.0/6.3/20.4	6.3/22.5/52.9
Patch2Pix+SG [14, 29]	6.0	69.6/91.1/99.5	1.6/2.6/19.9	6.8/24.1/53.4
	12.0	68.1/92.1/99.5	1.0/2.1/22.5	6.8/26.2/53.9
	20.0	67.5/92.1/99.5	0.5/2.1/22.0	8.9/25.1/53.4
	24.0	67.0/91.6/99.0	0.5/2.6/22.0	7.9/25.1/52.9
	32.0	67.5/91.6/98.4	0.5/2.6/20.9	8.9/24.1/49.7
	48.0	68.1/91.6/99.0	0.5/1.6/19.4	7.9/23.0/46.1
R2D2 [13]	6.0	58.1/73.8/83.8	0.0/0.0/0.0	0.0/0.0/0.0
	12.0	59.2/76.4/85.9	0.0/0.0/0.0	0.0/0.0/0.0
	20.0	58.6/75.9/85.9	0.0/0.0/0.0	0.0/0.0/0.5
	24.0	58.1/74.9/84.8	0.0/0.0/0.0	0.0/0.0/0.5
	32.0	56.5/73.8/83.2	0.0/0.0/0.0	0.0/0.0/1.0
	48.0	56.5/70.2/79.1	0.0/0.0/0.0	0.0/0.0/0.5
CAPS+SP [5, 26]	6.0	64.9/89.0/96.9	1.0/7.3/40.3	3.7/17.8/64.4
	12.0	67.0/88.5/97.4	3.1/7.3/50.3	3.1/18.3/70.7
	20.0	66.5/89.0/96.9	2.6/9.4/51.3	3.1/18.3/70.2
	24.0	67.0/88.0/96.3	2.1/8.4/50.8	2.6/17.8/69.6
	32.0	66.5/88.0/96.9	3.7/9.4/46.6	2.6/18.8/67.5
	48.0	64.4/85.3/93.2	2.6/6.8/40.3	4.2/19.9/64.4

AC13:	t	colored	AO	tricolor
SuperGlue [14]	6.0	67.5/88.0/96.9	1.6/13.1/30.9	19.4/45.5/68.1
	12.0	70.7/90.1/97.9	3.1/13.6/35.1	23.0/49.2/69.1
	20.0	69.6/89.5/97.4	2.6/11.0/34.0	21.5/49.7/69.6
	24.0	68.6/90.1/97.9	3.7/12.0/35.6	23.0/46.6/68.1
	32.0	67.5/89.0/97.9	4.7/12.6/34.6	21.5/45.5/68.6
	48.0	66.5/88.0/96.3	3.7/9.9/29.8	18.3/35.1/61.8
LoFTR [21]	6.0	69.1/88.0/94.8	2.6/6.3/29.3	9.9/28.8/64.4
	12.0	72.3/86.9/94.8	1.6/6.8/32.5	15.7/36.1/63.4
	20.0	72.3/87.4/94.2	3.1/8.4/32.5	13.1/36.1/61.8
	24.0	71.7/87.4/93.7	3.1/8.9/29.8	12.6/34.6/60.2
	32.0	71.2/86.9/92.7	2.6/7.9/28.3	13.6/35.1/58.1
	48.0	68.1/84.3/89.0	1.6/3.1/18.8	8.9/27.7/49.2
Patch2Pix [29]	6.0	64.4/81.7/90.1	5.8/20.4/49.2	20.9/39.8/75.4
	12.0	61.3/82.7/90.6	8.9/24.6/55.5	26.2/50.8/82.2
	20.0	60.2/81.2/92.1	11.0/28.3/58.6	30.4/58.6/80.1
	24.0	59.2/82.2/93.2	12.6/28.3/57.1	31.9/58.1/81.2
	32.0	58.6/82.7/91.6	13.6/27.7/57.1	29.8/53.4/77.5
	48.0	57.1/80.1/89.0	10.5/23.6/49.2	26.7/46.6/72.3
Patch2Pix+SG [14, 29]	6.0	71.2/92.1/97.4	4.7/17.8/41.9	22.0/47.6/72.8
	12.0	69.6/91.6/97.9	5.8/21.5/45.0	23.0/50.3/73.8
	20.0	67.5/91.6/97.9	5.2/23.0/45.5	22.0/45.5/73.8
	24.0	68.6/91.6/97.4	5.2/22.0/45.0	20.9/46.1/73.8
	32.0	69.1/91.1/96.9	5.2/21.5/44.5	20.4/43.5/72.3
	48.0	69.6/90.6/96.3	5.8/18.8/43.5	19.9/42.9/69.1
R2D2 [13]	6.0	55.5/69.6/82.7	0.0/0.0/0.0	0.0/0.5/1.6
	12.0	57.6/71.2/83.2	0.0/0.0/0.0	1.0/1.0/2.6
	20.0	56.5/71.2/83.2	0.0/0.0/0.0	0.5/0.5/2.6
	24.0	56.0/71.2/82.7	0.0/0.0/0.0	0.5/0.5/3.1
	32.0	55.0/69.1/81.7	0.0/0.0/0.0	0.0/0.5/2.6
	48.0	52.9/66.0/75.9	0.0/0.0/0.0	0.0/0.5/2.1
CAPS+SP [5, 26]	6.0	61.3/82.7/94.2	9.4/25.7/70.7	13.6/39.3/81.2
	12.0	60.7/82.7/94.8	9.4/30.9/75.4	16.8/47.1/85.9
	20.0	60.2/81.7/95.3	7.9/31.9/75.4	14.7/45.5/85.9
	24.0	58.6/81.7/95.3	7.9/31.9/73.8	14.1/45.5/83.8
	32.0	57.6/80.6/92.7	8.9/32.5/70.7	13.1/42.4/79.6
	48.0	57.1/79.1/91.1	7.3/27.7/64.9	10.5/37.2/74.3

Table 4. Ablation study on the Aachen Day-Night v1.1 dataset [16,17,28] using images rendered at reduced resolution (max. 800 px) from 3D meshes of different levels of detail (*cf.* Tab. 1) and different rendering types (textured / colored, raw geometry with ambient occlusion (AO), raw geometry with tricolor shading (tricolor)). We report results for a MeshLoc variant that uses individual matches and position averaging, but no covisibility filtering. We vary the inlier threshold t used in RANSAC

AC14:	t	colored	AO	tricolor
SuperGlue [14]	6.0	69.6/88.5/95.8	18.3/37.2/55.5	32.5/58.1/72.3
	12.0	69.1/88.5/95.8	21.5/40.3/58.6	32.5/63.4/74.9
	20.0	69.1/88.0/95.8	21.5/39.3/57.1	33.0/63.4/74.3
	24.0	69.1/88.0/95.8	19.4/36.1/53.4	32.5/62.3/74.3
	32.0	68.6/88.0/95.8	19.9/33.0/51.3	31.9/59.7/72.8
	48.0	69.6/88.0/95.8	20.4/31.4/46.6	29.3/53.4/70.2
LoFTR [21]	6.0	74.3/87.4/95.8	15.7/39.3/62.8	26.2/58.6/78.0
	12.0	76.4/87.4/95.3	19.4/42.9/66.0	27.7/62.3/78.5
	20.0	73.8/86.9/93.7	18.3/42.4/66.0	28.3/60.2/77.0
	24.0	73.3/85.3/93.2	16.2/41.4/65.4	29.3/60.7/75.9
	32.0	73.3/85.9/93.2	14.7/38.7/62.8	29.8/57.1/73.8
	48.0	72.8/84.3/92.1	13.6/35.1/57.6	27.2/52.9/66.5
Patch2Pix [29]	6.0	62.8/81.2/92.1	18.8/36.1/70.2	30.9/58.1/80.6
	12.0	62.3/84.3/94.2	26.2/48.7/74.9	38.2/67.5/83.2
	20.0	63.9/84.3/93.2	26.2/51.8/74.3	38.2/67.5/84.3
	24.0	62.8/83.8/93.2	26.2/53.4/75.4	38.7/65.4/83.2
	32.0	63.4/85.9/92.7	24.1/47.6/72.8	37.2/66.0/83.8
	48.0	62.3/81.7/88.5	22.0/41.4/64.9	31.9/59.7/77.0
Patch2Pix+SG [14, 29]	6.0	72.3/90.6/96.9	19.9/40.8/63.4	35.6/64.9/78.5
	12.0	70.7/89.0/96.9	20.4/39.8/64.9	33.5/64.4/78.5
	20.0	70.2/89.5/96.9	19.4/40.3/62.8	32.5/61.8/78.5
	24.0	70.7/89.5/96.9	19.9/38.2/61.8	33.0/61.8/78.0
	32.0	70.2/89.0/96.9	19.9/37.7/59.7	33.5/61.8/78.0
	48.0	69.1/89.0/96.3	18.8/36.6/57.6	30.9/60.2/75.4
R2D2 [13]	6.0	57.6/67.5/81.2	0.0/0.0/1.0	0.5/0.5/1.6
	12.0	56.5/69.6/83.8	0.0/0.0/1.6	0.5/1.0/3.1
	20.0	56.5/70.2/84.3	0.0/0.0/2.1	0.5/0.5/2.1
	24.0	55.5/69.1/82.7	0.0/0.0/2.6	1.0/1.0/2.6
	32.0	54.5/69.6/80.1	0.0/0.0/3.1	1.0/1.0/2.1
	48.0	52.9/67.0/77.5	0.0/0.0/2.6	0.0/0.5/1.0
CAPS+SP [5, 26]	6.0	56.5/82.7/94.2	24.1/53.4/84.8	33.0/64.4/86.4
	12.0	60.7/84.8/95.3	25.1/62.8/84.8	30.9/65.4/90.1
	20.0	61.3/83.2/94.8	24.6/57.6/82.7	30.4/68.1/87.4
	24.0	61.3/83.2/93.7	24.6/55.5/82.7	31.4/68.6/88.0
	32.0	61.8/82.7/92.7	23.6/55.5/81.2	29.8/66.0/85.3
	48.0	59.2/79.1/89.0	22.5/55.5/76.4	25.1/62.3/80.1

AC15:	t	colored	AO	tricolor
SuperGlue [14]	6.0	71.2/89.0/96.9	21.5/38.2/62.3	37.2/55.5/70.2
	12.0	73.3/90.1/97.9	23.6/45.0/63.4	35.6/59.7/75.9
	20.0	73.3/89.5/97.4	24.6/40.8/59.7	35.6/59.7/74.9
	24.0	72.3/89.5/97.9	20.9/40.3/58.1	35.6/60.2/74.3
	32.0	71.2/88.5/96.9	20.4/37.2/55.5	35.1/54.5/70.7
	48.0	70.2/87.4/96.3	17.3/36.6/52.9	31.4/51.3/68.1
LoFTR [21]	6.0	75.9/88.5/95.8	19.9/40.8/62.8	32.5/53.9/74.9
	12.0	78.0/89.0/95.8	19.9/42.9/65.4	34.0/60.7/77.0
	20.0	75.9/88.0/94.2	19.9/41.4/61.8	32.5/59.7/76.4
	24.0	74.9/88.0/94.2	18.8/39.3/59.7	31.4/57.1/73.8
	32.0	74.3/87.4/94.2	17.3/39.3/57.1	30.9/52.9/69.6
	48.0	73.8/86.9/92.7	15.7/32.5/52.4	26.7/47.6/63.4
Patch2Pix [29]	6.0	63.9/81.2/93.7	22.0/46.1/73.3	26.7/53.9/77.5
	12.0	63.9/84.3/93.7	29.8/54.5/77.5	35.6/60.7/80.1
	20.0	65.4/84.8/94.2	27.7/55.5/77.5	36.1/64.4/81.7
	24.0	65.4/83.8/93.7	26.7/53.4/76.4	38.2/62.8/82.2
	32.0	64.4/83.8/93.7	27.2/52.9/75.9	36.6/63.9/81.2
	48.0	63.4/81.7/89.5	24.6/45.5/66.0	35.1/57.1/74.9
Patch2Pix+SG [14, 29]	6.0	72.8/90.1/98.4	24.6/47.1/65.4	37.2/60.7/78.0
	12.0	72.3/91.6/98.4	24.6/49.2/67.0	39.3/62.3/79.6
	20.0	72.8/91.1/97.9	22.5/46.1/66.0	36.6/62.3/80.1
	24.0	71.7/91.1/97.9	20.4/46.1/65.4	35.6/61.3/81.7
	32.0	69.6/91.1/98.4	22.0/46.1/63.9	35.1/59.2/80.1
	48.0	68.1/90.6/97.9	21.5/43.5/63.4	35.1/59.7/78.0
R2D2 [13]	6.0	56.5/72.8/82.7	0.0/0.0/1.0	0.0/0.0/1.6
	12.0	59.2/74.9/84.3	0.0/0.0/1.0	0.0/0.5/2.1
	20.0	59.2/72.8/84.3	0.0/0.0/1.0	0.0/0.5/1.6
	24.0	59.2/72.8/83.8	0.0/0.0/0.5	0.0/0.5/1.6
	32.0	57.6/70.7/83.8	0.0/0.0/0.5	0.0/0.0/1.6
	48.0	52.4/68.1/79.1	0.0/0.0/0.0	0.0/0.5/1.6
CAPS+SP [5, 26]	6.0	61.8/86.9/95.3	28.8/58.6/83.8	34.0/66.0/89.0
	12.0	67.5/85.9/96.3	25.7/58.1/89.0	36.1/70.2/91.1
	20.0	67.0/85.3/95.8	27.7/59.7/86.4	34.0/69.1/89.5
	24.0	67.0/85.3/95.3	26.7/59.2/85.3	32.5/68.6/89.0
	32.0	65.4/84.8/94.2	27.2/57.6/84.3	31.4/67.0/88.5
	48.0	64.9/82.2/89.5	23.6/51.3/75.9	29.8/62.3/79.6

Table 5. Results on the 12 Scenes [24] dataset using real images (r), images rendered from colored models (c), and images rendered using the tricolor scheme (t). We report results for a MeshLoc variant that uses all individual SuperGlue matches and position averaging, but no covisibility filtering. We use an inlier threshold of 6 pixels used in RANSAC. We report the percentage of images localized within **5 cm and 5°** of the ground truth and compare against results reported in [1]

	apt1/		apt2/				office1/				office2/		
Method	kitchen	living	bed	kitchen	living	luke	gates362	gates381	lounge	manolis	5a	5b	Average
MeshLoc (SG) (r)	98.0	88.2	100.0	97.6	97.4	92.9	97.7	93.4	87.8	95.5	91.5	88.4	94.0
MeshLoc (SG) (c)	63.6	31.0	86.8	92.9	68.8	68.8	68.1	78.3	60.2	75.7	35.8	60.0	65.8
MeshLoc (SG) (t)	17.4	15.8	14.7	21.9	5.4	5.4	17.6	15.3	12.5	23.3	5.6	14.6	14.1
Active Search	100.0	99.8	100.0	100.0	100.0	99.2	100.0	98.2	100.0	100.0	98.0	100.0	99.6
HLoc	100.0	100.0	100.0	100.0	100.0	98.9	100.0	99.1	100.0	100.0	99.6	100.0	99.8
DVLAD+R2D2	100.0	100.0	100.0	100.0	100.0	99.4	100.0	98.4	100.0	100.0	99.2	100.0	99.7
DVLAD+R2D2 (+D)	98.0	100.0	100.0	100.0	100.0	99.4	100.0	98.4	100.0	99.5	96.8	100.0	99.3
DSAC*	100.0	100.0	100.0	100.0	100.0	97.4	100.0	99.3	100.0	97.6	97.2	98.8	99.2
DSAC* (+D)	100.0	100.0	100.0	100.0	100.0	97.8	100.0	97.8	100.0	100.0	100.0	100.0	99.6

Table 6. Results on the 12 Scenes [24] dataset using real images (r), images rendered from colored models (c), and images rendered using the tricolor scheme (t). We report results for a MeshLoc variant that uses all individual SuperGlue matches and position averaging, but no covisibility filtering. We use an inlier threshold of 6 pixels used in RANSAC. We report the percentage of images localized within **7 cm and 7°** of the ground truth and compare against results reported in [1]

	apt1/		apt2/				office1/				office2/		
Method	kitchen	living	bed	kitchen	living	luke	gates362	gates381	lounge	manolis	5a	5b	Average
MeshLoc (SG) (r)	100.0	100.0	100.0	99.0	100.0	99.7	100.0	98.7	100.0	99.9	98.6	99.8	99.6
MeshLoc (SG) (c)	95.8	83.0	100.0	98.6	98.6	96.2	96.9	97.6	95.1	98.0	69.2	83.2	92.7
MeshLoc (SG) (c)	42.6	40.4	37.7	56.2	15.5	21.8	46.4	36.2	28.1	64.6	12.5	29.9	36.0
Active Search	100.0	99.8	100.0	100.0	100.0	99.8	100.0	99.0	100.0	100.0	98.6	100.0	99.8
HLoc	100.0	100.0	100.0	100.0	100.0	99.8	100.0	99.9	100.0	100.0	100.0	100.0	100.0
DVLAD+R2D2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.1	100.0	100.0	99.6	100.0	99.9
DVLAD+R2D2 (+D)	98.0	100.0	100.0	100.0	100.0	100.0	100.0	98.7	100.0	99.7	97.4	100.0	99.5
DSAC*	100.0	100.0	100.0	100.0	100.0	98.1	100.0	99.7	100.0	99.1	100.0	100.0	99.7
DSAC* (+D)	100.0	100.0	100.0	100.0	100.0	97.9	100.0	99.0	100.0	100.0	100.0	100.0	99.7

Table 7. Results on the 12 Scenes [24] dataset using real images (r), images rendered from colored models (c), and images rendered using the tricolor scheme (t). We report results for a MeshLoc variant that uses all individual SuperGlue matches and position averaging, but no covisibility filtering. We use an inlier threshold of 6 pixels used in RANSAC. We report the percentage of images localized within **10 cm and 10°** of the ground truth and compare against results reported in [1]

	apt1/		apt2/				office1/				office2/		
Method	kitchen	living	bed	kitchen	living	luke	gates362	gates381	lounge	manolis	5a	5b	Average
MeshLoc (SG) (r)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.7	100.0	100.0	99.8	100.0	100.0
MeshLoc (SG) (c)	100.0	99.8	100.0	100.0	100.0	99.8	100.0	99.9	99.7	100.0	92.2	97.8	99.1
MeshLoc (SG) (t)	69.2	60.6	49.0	76.7	26.6	40.5	75.1	52.2	47.1	84.8	20.3	47.9	54.2
Active Search	100.0	100.0	100.0	100.0	100.0	99.8	100.0	99.2	100.0	100.0	99.4	100.0	99.9
HLoc	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
DVLAD+R2D2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.1	100.0	100.0	100.0	100.0	99.9
DVLAD+R2D2 (+D)	98.0	100.0	100.0	100.0	100.0	100.0	100.0	98.7	100.0	99.7	98.4	100.0	99.6
DSAC*	100.0	100.0	100.0	100.0	100.0	98.4	100.0	99.9	100.0	99.6	100.0	100.0	99.8
DSAC* (+D)	100.0	100.0	100.0	100.0	100.0	98.1	100.0	100.0	100.0	100.0	100.0	100.0	99.8

References

1. Brachmann, E., Humenberger, M., Rother, C., Sattler, T.: On the Limits of Pseudo Ground Truth in Visual Camera Re-localisation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6218–6228 (2021)
2. Brachmann, E., Rother, C.: Visual camera re-localization from RGB and RGB-D images using DSAC. TPAMI (2021)
3. Chum, O., Matas, J.: Randomized RANSAC with $T_{d,d}$ Test. In: British Machine Vision Conference (BMVC) (2002)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009)
5. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-Supervised Interest Point Detection and Description. In: CVPR Workshops (2018)
6. Humenberger, M., Cabon, Y., Guerin, N., Morat, J., Revaud, J., Rerole, P., Pion, N., de Souza, C., Leroy, V., Csurka, G.: Robust Image Retrieval-based Visual Localization using Kapture. arXiv:2007.13867 (2020)
7. Jakob, W., Tarini, M., Panozzo, D., Sorkine-Hornung, O.: Instant field-aligned meshes. ACM Trans. Graph. **34**(6), 189–1 (2015)
8. Larsson, V.: PoseLib - Minimal Solvers for Camera Pose Estimation (2020), `https://github.com/vlarsson/PoseLib`
9. Lebeda, K., Matas, J.E.S., Chum, O.: Fixing the Locally Optimized RANSAC. In: BMVC (2012)
10. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. IJCV (2004)
11. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020)
12. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4), 102:1–102:15 (Jul 2022). https://doi.org/10.1145/3528223.3530127, `https://doi.org/10.1145/3528223.3530127`
13. Revaud, J., Weinzaepfel, P., de Souza, C.R., Humenberger, M.: R2D2: repeatable and reliable detector and descriptor. In: NeurIPS (2019)
14. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning Feature Matching with Graph Neural Networks. In: CVPR (2020)
15. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. PAMI (2017)
16. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., Pajdla, T.: Benchmarking 6DOF Urban Visual Localization in Changing Conditions. In: CVPR (2018)
17. Sattler, T., Weyand, T., Leibe, B., Kobbelt, L.: Image Retrieval for Image-Based Localization Revisited. In: BMVC (2012)
18. Schönberger, J.L., Frahm, J.M.: Structure-From-Motion Revisited. In: CVPR (2016)
19. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise View Selection for Unstructured Multi-View Stereo. In: European Conference on Computer Vision (ECCV) (2016)
20. Sun, J., Chen, X., Wang, Q., Li, Z., Averbuch-Elor, H., Zhou, X., Snavely, N.: Neural 3D Reconstruction in the Wild. In: SIGGRAPH Conference Proceedings (2022)

21. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
22. Torii, A., Arandjelović, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: CVPR (2015)
23. Turki, H., Ramanan, D., Satyanarayanan, M.: Mega-NeRF: Scalable Construction of Large-Scale NeRFs for Virtual Fly-Throughs (2021)
24. Valentin, J., Dai, A., Niessner, M., Kohli, P., Torr, P., Izadi, S., Keskin, C.: Learning to Navigate the Energy Landscape. In: 3DV (2016)
25. Waechter, M., Moehrle, N., Goesele, M.: Let there be color! Large-scale texturing of 3D reconstructions. In: European conference on computer vision. pp. 836–850. Springer (2014)
26. Wang, Q., Zhou, X., Hariharan, B., Snavely, N.: Learning Feature Descriptors using Camera Pose Supervision. In: The European Conference on Computer Vision (ECCV) (2020)
27. Xiangli, Y., Xu, L., Pan, X., Zhao, N., Rao, A., Theobalt, C., Dai, B., Lin, D.: Citynerf: Building nerf at city scale (Dec 2021), http://arxiv.org/abs/2112.05504v2
28. Zhang, Z., Sattler, T., Scaramuzza, D.: Reference Pose Generation for Long-term Visual Localization via Learned Features and View Synthesis. IJCV (2020)
29. Zhou, Q., Sattler, T., Leal-Taixe, L.: Patch2pix: Epipolar-guided pixel-level correspondences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)