On the Limits of Pseudo Ground Truth in Visual Camera Re-localisation

Eric Brachmann[1]   Martin Humenberger[2]   Carsten Rother[3]   Torsten Sattler[4]

[1]Niantic   [2]NAVER LABS Europe   [3]Visual Learning Lab, HCI/IWR, Heidelberg University

[4]Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague

Abstract

Benchmark datasets that measure camera pose accuracy have driven progress in visual re-localisation research. To obtain poses for thousands of images, it is common to use a reference algorithm to generate pseudo ground truth. Popular choices include Structure-from-Motion (SfM) and Simultaneous-Localisation-and-Mapping (SLAM) using additional sensors like depth cameras if available. Re-localisation benchmarks thus measure how well each method replicates the results of the reference algorithm. This begs the question whether the choice of the reference algorithm favours a certain family of re-localisation methods. This paper analyzes two widely used re-localisation datasets and shows that evaluation outcomes indeed vary with the choice of the reference algorithm. We thus question common beliefs in the re-localisation literature, namely that learning-based scene coordinate regression outperforms classical feature-based methods, and that RGB-D-based methods outperform RGB-based methods. We argue that any claims on ranking re-localisation methods should take the type of the reference algorithm, and the similarity of the methods to the reference algorithm, into account.

1. Introduction

The availability of benchmark datasets [14, 34, 38, 39, 43, 56, 64, 69, 72, 74, 77, 82] has been a driving factor for research on visual re-localisation, a core technology to make autonomous robots [40], self-driving cars [28], and augmented / mixed reality (AR / MR) systems [1, 10, 42] a reality. These benchmarks provide camera poses for a set of training and test images. The training images can be used to create a scene representation, and the test images serve as queries to determine the 3D position and 3D orientation (6DoF pose) of the camera with respect to the scene. Due to the challenge of jointly estimating the poses of thousands or more images, benchmark datasets are typically generated by a reference algorithm such as SfM or (RGB-)D SLAM [34, 38, 39, 64, 74]. As such, benchmarks measure how well visual re-localisation methods are able to replicate the results of the reference algorithm.

Ideally, the choice of reference algorithm should not

RGB-D SLAM Pseudo Ground Truth

Accuracy @ 1cm,1°

■ Pseudo Ground Truth
■ Active Search
■ DSAC* (RGB)
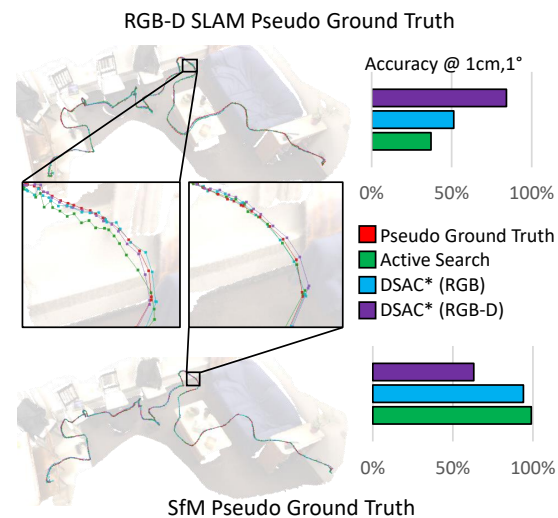■ DSAC* (RGB-D)

SfM Pseudo Ground Truth

Figure 1. **Visualisation of the same scene reconstructed by two different reference algorithms.** We show the trajectories of test images estimated by state-of-the-art visual localisation algorithms (left) and the percentage of images localised within 1cm and 1° of error w.r.t. the pseudo ground truth (right, higher is better). While the underlying 3D scene models are very similar, the relative ranking of methods is inverted. In this paper, we show that the reference algorithm used to create the pseudo ground truth has a significant impact on which method achieves the best results.

matter as long as it faithfully estimates the camera poses of the training and test images. In particular, the choice of reference algorithm should not affect the ranking of methods on a benchmark. In practice however, different reference algorithms optimise different cost functions, *e.g.*, reprojection errors of sparse point clouds for SfM [60, 80] or alignment errors in 3D space for depth-based SLAM methods [16, 30, 47, 62], leading to different local minima. We ask to what degree the choice of reference algorithm impacts the ranking of methods on a benchmark. This is an important question as it pertains to whether or not we can draw absolute conclusions, *e.g.*, algorithm A is better than algorithm B or using component C improves accuracy. Interestingly, to the best of our knowledge, this question has not received much attention in the re-localisation literature.

The main focus of this paper is to investigate how

the choice of reference algorithms impacts the measured performance of visual re-localisation algorithms. To this end, we compare two types of reference algorithms (depth-based SLAM and SfM) on two popular benchmark datasets [64, 74]. Detailed experiments with state-of-the-art re-localisation algorithms show that the choice of reference algorithm can have a profound impact on the ranking of methods. In particular, as illustrated in Fig. 1, we show that depending on the reference algorithm, a modern end-to-end-trainable approach [7] either outperforms or is outperformed by a classical, nearly 10 year-old baseline [54, 55]. Similarly, the choice of whether to use depth maps or SfM point clouds to represent a scene can improve or decrease performance depending on the reference algorithm. Our results show that we as a community should be careful when drawing conclusions from existing benchmarks. Instead, it is necessary to take into account that certain approaches more closely resemble the reference algorithm than others. The former are better able to replicate imperfections in a reference algorithm's *pseudo ground truth* (pGT). This natural advantage should be discussed when evaluating localisation results and designing new benchmarks.

In detail, this paper makes the following contributions:
1) we show that the choice of a reference algorithm for obtaining pGT poses can have a significant impact on the relative ranking of methods, to the extend that the rankings of methods can be (nearly) completely reversed. This implies that published results for visual re-localisation should always be considered under the aspect of which algorithm was used to create the pGT.

2) we provide a comparison of pGT generated by RGB-only SfM and (RGB-)D SLAM on the 7Scenes [64] and 12Scenes [74] datasets, which are widely used [2–6, 8, 11, 12, 26, 33, 34, 64, 74–76]. We show that none is clearly superior than the other. We show that commonly accepted results from the literature (RGB-D variants of re-localisation methods outperform their RGB-only counterparts; scene coordinate regression is more accurate than feature-based methods) are not absolute but depending on the pGT.

3) we are not aware of prior work aimed at evaluating the extent to which conclusions about localisation performance can be drawn from existing benchmarks. As such, this paper is the first to raise awareness that the limitations of the pGT for re-localisation need to be discussed in order to make valid comparisons across methods.

Our new pGT and our evaluation pipeline are available at github.com/tsattler/visloc_pseudo_gt_limitations/.

2. Related Work

The difficulty of obtaining ground truth varies for different tasks in computer visions. For tasks with a low dimensional structure, *e.g.* image classification and object detection, human annotations are effective [21] and can be scaled

via crowd sourcing [17]. For tasks with a more complex output, *e.g.* image segmentation or optical flow, annotation time quickly rises to a level which severely affects the scale and cost of associated datasets [15, 19, 25, 46].

The 6DoF camera pose estimation task comes with the added difficulty that humans are not skilled at directly annotating the camera poses. Instead they annotate image correspondences as input to an optimization problem to recover the pose indirectly [56, 69, 72]. Since many correspondences are required for a stable pose estimate, such an annotation approach does not scale beyond a few hundred images. Further, the annotations are usually only precise up to a few pixels, which, depending on the distance to the scene, can result in significant pose uncertainty [56, 72, 82].

As an alternative, the recording camera can be tracked by an external tracking system [9, 62, 67]. While providing high precision poses, capturing a diverse set of scenes is challenging due to the complicated setup, *i.e.*, installation and calibration, ensuring good visibility of the sensors, *etc*. Similarly, industrial level LiDAR scanners have been used to produce high quality scans of landmarks, but the corresponding datasets provide only few scenes with limited spatial extent [35, 63, 66]. GPS-INS systems that combine GPS with inertial navigation systems (INS) have also been used to track camera poses on a large scale [25, 43]. Yet, post-processing is still required to obtain a higher accuracy [56].

Synthetic datasets come with *true* ground truth, but most current datasets, *e.g.*, Habitat [44], are limited in diversity of low-level noise, illumination conditions or specular reflections. Therefore, data association, which is at the core of re-localisation, can become too easy. An example are the very low errors reported in [58] for Active Search [54, 55] on a synthetic version of the Cambridge Landmarks dataset.

The vast majority of re-localisation benchmarks follow an automatic approach to ground truth recovery using a reference algorithm [14, 31, 34, 38, 39, 56, 64, 69, 72, 74, 77, 82]. Popular choices are SfM [60, 80], often for large-scale outdoor environments [14, 34, 38, 39, 56, 72], and depth-based SLAM [16, 30, 47, 62], often for small-scale indoor environments [64, 74, 77]. Hybrid solutions also exist, such as ICP-based registration of LiDAR scans followed by an SfM-based registration of RGB images [56, 69]. Some benchmarks use human visual inspection as a final quality control and verification stage [31, 56, 69, 82], and state-of-the-art reference algorithms are found to provide high-quality reconstructions and pose tracks. However, as shown in Fig. 1, subtle differences in the output of reference algorithms, unlikely to be recognized by visual checks, can have significant influence on the evaluation outcome of a benchmark.

Such evaluation artifacts have the potential to challenge some conclusions that have previously been drawn in the literature. Fig. 2 shows published results in re-localisation research on the popular indoor datasets 7Scenes [64] and
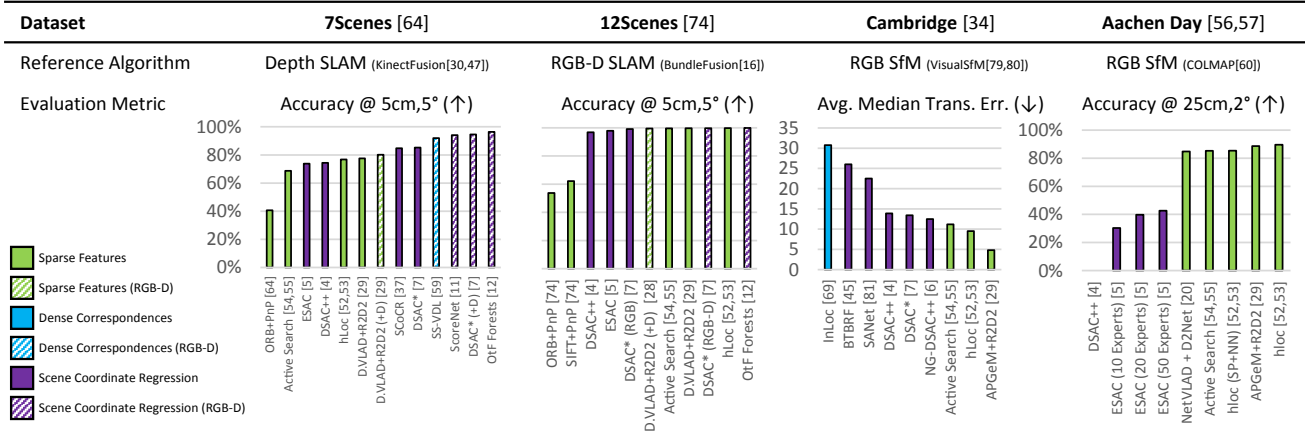
Dataset	7Scenes [64]	12Scenes [74]	Cambridge [34]	Aachen Day [56,57]
Reference Algorithm	Depth SLAM (KinectFusion[30,47])	RGB-D SLAM (BundleFusion[16])	RGB SfM (VisualSfM[79,80])	RGB SfM (COLMAP[60])
Evaluation Metric	Accuracy @ 5cm,5° (↑)	Accuracy @ 5cm,5° (↑)	Avg. Median Trans. Err. (↓)	Accuracy @ 25cm,2° (↑)

Legend:
- Sparse Features
- Sparse Features (RGB-D)
- Dense Correspondences
- Dense Correspondences (RGB-D)
- Scene Coordinate Regression
- Scene Coordinate Regression (RGB-D)

7Scenes methods: ORB+PnP [64], Active Search [54,55], ESAC [5], DSAC++ [5], hLoc [52,53], D.VLAD+R2D2 [29], D.VLAD+R2D2 (+D) [29], SCoCR [37], DSAC* [7], SS-VDL [59], ScoreNet [11], DSAC* (+D) [7], OtF Forests [12]

12Scenes methods: ORB+PnP [74], SIFt+PnP [74], DSAC++ [4], ESAC [5], DSAC* (RGB) [7], D.VLAD+R2D2 (+D) [28], Active Search [54,55], D.VLAD+R2D2 [29], DSAC* (RGB-D) [7], hLoc [52,53], OtF Forests [12]

Cambridge methods: InLoc [69], BTBRF [45], SANet [81], DSAC++ [4], DSAC* [7], NG-DSAC++ [6], Active Search [54,55], hLoc [52,53], APGeM+R2D2 [29]

Aachen Day methods: DSAC++ [4], ESAC (10 Experts) [5], ESAC (20 Experts) [5], ESAC (50 Experts) [5], NetVLAD + D2Net [20], Active Search [54,55], hLoc (SP+NN) [52,53], APGeM+R2D2 [29], hloc [52,53]

Figure 2. **State-of-the-art results for four benchmarks.** For indoor scenes (7Scenes) learned scene coordinate regression methods (purple) outperform feature-based approaches (green), and depth-based methods (dashed) outperform RGB-only methods. For outdoor scenes (Cambridge and Aachen) feature-based methods, in turn, outperform scene coordinate regression. We show that this behavior is coupled to the type of reference algorithm used to create the pGT. Results of 12Scenes are saturated under the common 5cm,5° threshold.

12Scenes [74], and the popular outdoor datasets Cambridge Landmarks [34] and Aachen Day [56,57]. We compare the dominant families of re-localisation methods, scene coordinate regression and sparse feature-based matching. Scene coordinate regression methods use a learned model, a neural network or a random forest, to predict dense image-to-scene correspondences [2–7,12,13,37,64,75,81]. RGB-D variants of scene coordinate regression methods dominate rankings for indoor re-localisation, which has been attributed to the inherent difficulty of the indoor scenario regarding texture-less surfaces and ambiguous structures that make it difficult to find and match sparse features [2,34,64,76]. For outdoor re-localisation, classical approaches, which match hand-crafted [55,64,74] or learned descriptors [18,29,52,53] at sparse feature locations to a 3D SfM reconstruction, achieve vastly superior results compared to scene coordinate regression. This has been attributed to an inability of scene coordinate regression to scale to spatially large scenes [56,70]. Our work offers a different explanation regarding the performance of re-localisers in different environments by taking the reference algorithms into account that were used to create the associated benchmarks.

3. Datasets and Reference Algorithms

In order to measure the impact different reference algorithms have on localisation performance, we consider pGT generated using (RGB-)D and sparse RGB-only data. We use the popular 7Scenes [64] and 12Scenes [74] datasets as they provide depth maps and pGT poses for both test and training images. This is in contrast to other common benchmarks [34, 38, 56, 69, 72, 77], which do not provide depth information for test and train images [34, 38, 56, 69, 72][1]

or do not make the poses of the test images publicly available [56,69,77]. In the following, we describe the datasets, their original pGT, and how we create an additional pGT for each dataset via RGB-only SfM. The purpose of this section is to familiarize the reader with the datasets and reference algorithms before evaluating the resulting pGT variants (Sec. 4) and measuring their impact on re-localisation performance (Sec. 5).

3.1. Incremental Depth SLAM

Camera poses can be tracked by incrementally registering dense depth measurements to a 3D scene representation. KinectFusion [30, 48], an early incarnation of such a system, uses a truncated signed distance function (TSDF) to represent the scene. The TSDF is updated by merging depth maps $\mathbf{D}_i$ of frames i into a weighted average

$$\min_{\mathbf{F}} \sum_i ||\mathbf{W}_{\mathbf{D}_i}\mathbf{F}_{\mathbf{D}_i} - \mathbf{F}|| \ ,$$

where $\mathbf{F}$ and $\mathbf{F}_{\mathbf{D}_i}$ denote the TSDF representations of the scene and of depth map $\mathbf{D}_i$, respectively. Weights $\mathbf{W}_{\mathbf{D}_i}$ capture the measurement uncertainty of the depth recording. For tracking the 6DoF camera pose of a new frame with rotation $\mathtt{R}_i$ and translation $\mathbf{t}_i$, KinectFusion minimises the point-plane distance between the measured depth and a depth rendering of the scene's TSDF volume

$$\min_{\mathtt{R}_i, \mathbf{t}_i} \sum_{\mathbf{x}} || \left(\mathbf{V}_i(\mathbf{x}) - [\mathtt{R}_i|\mathbf{t}_i] \, \hat{\mathbf{V}}_{i-1}^g(\hat{\mathbf{x}}) \right)^\top \hat{\mathbf{N}}_{i-1}^g(\hat{\mathbf{u}})|| \ . \quad (1)$$

The objective is minimised over 2D pixel positions $\mathbf{x}$. Measured depth and rendered depth are back-projected to 3D vertex maps $\mathbf{V}_i$ and $\hat{\mathbf{V}}_{i-1}^g$, respectively. Particularly, $\hat{\mathbf{V}}^g$ denotes the rendered vertex map of the scene in world (or *global*) coordinates and $\hat{\mathbf{N}}^g$ denotes rendered normals.

KinectFusion pGT for 7Scenes. Shotton *et al.* [64] created the 7Scenes dataset for re-localisation by scanning seven small-scale indoor environments with Kinect v1 and

---

[1]Depth maps could be estimated using motion stereo or single-view depth prediction. Yet, the former requires and is influenced by the pGT, while the latter offers limited quality and stability.

KinectFusion. Every scene was scanned multiple times by different users, and the resulting 3D scene models were registered using ICP [51]. No global optimization within a single scan or across multiple scans was performed, and any camera drift remains unaccounted for in the pGT of 7Scenes. In terms of RGB-D images, the 7Scenes dataset only provides uncalibrated output of the Kinect, *i.e.*, RGB images and depth maps are not registered, and the camera poses align with the depth sensor, not the RGB camera.

3.2. Globally Optimised RGB-D SLAM

To reduce camera drift during incremental scanning, more recent RGB-D SLAM systems like BundleFusion [16] jointly optimise all 6DoF camera poses. The parameter vector $\mathcal{X} = (R_0, t_0, ..., R_N, t_N)$ stacks rotations and translations of all frames recorded and BundleFusion optimises

$$\min_{\mathcal{X}} w_{\text{sprs}} E_{\text{sprs}}(\mathcal{X}) + w_{\text{pht}} E_{\text{pht}}(\mathcal{X}) + w_{\text{geo}} E_{\text{geo}}(\mathcal{X}) \ . \quad (2)$$

The term E_{sprs} minimises the Euclidean distance for sparse SIFT [41] feature matches across all images. Note that this term minimises a 3D distance, not a reprojection error, since the depth of image pixels is known. The term E_{pht} is a photometric loss that ensures a consistent gradient of image luminance across registered images. Finally, E_{geo} optimises a point-to-plane distance of depth maps with projective data association similar to KinectFusion, see Eq. 1.

BundleFusion pGT for 12Scenes. Valentin *et al.* [74] scanned twelve small-scale indoor environments for their 12Scenes dataset. They utilized a structure.io depth sensor mounted on an iPad that provided associated color images. Different from 7Scenes, 12Scenes comes with fully calibrated and synchronized color and depth images and depth is registered to the color images. Each room was scanned two times, once for training and once for testing, and both scans of each scene were registered manually.

3.3. Pseudo Ground Truth via SfM

A common approach to generate pGT [34, 38, 39, 56, 68, 72] is to use (incremental) SfM algorithms [60, 65, 80]. SfM methods rely on sparse local features such as SIFT [41] to establish feature matches between images, which are then used to recover camera poses and 3D scene structure. SfM is usually applied jointly on the test and training images to jointly recover the camera poses of all images [34, 39, 56].

SfM algorithms minimise the reprojection error between the estimated 3D points and their corresponding feature measurements in the images, optimising the problem

$$\min_{R_i, \theta_i, t_i, X_j} \sum_i \sum_j \delta_{ij} \rho \left(||\mathbf{x}_{ij} - \pi \left(R_i X_j + t_i, \theta_i \right) ||^2 \right) \quad (3)$$

during Bundle Adjustment (BA) [73]. Here, θ_i are the intrinsic camera parameters, X_j is the jth 3D point, $\delta_{ij} \in \{0, 1\}$ indicates whether the jth 3D point is visible in the ith image, $\mathbf{x}_{ij}$ is the corresponding 2D feature position of the

j^{th} 3D point in the ith image, π is the projection function, and ρ is a robust cost function [73]. SfM only reconstructs the scene up to an arbitrary scaling factor. Known 3D distances are used to recover the absolute scale of the model.

SfM pGT for 7Scenes and 12Scenes. As the basis for our analysis, we generate an alternative pGT for 7Scenes and 12Scenes. First, we reconstruct the scene with SfM using only the training images. Next, we continue the reconstruction process with the test images while keeping the training camera poses fixed. This strategy ensures that the training poses are not affected by the test images, as would be the case in practice. Finally, we recover the scale by robustly aligning the positions of all cameras to those of the original pGT. We implement this process with COLMAP [60], using the same camera intrinsics for all images in a scene.

This approach failed for the office2/5a and 5b datasets of 12Scenes. Both depict scenes with highly repetitive structures. As a result, the SfM reconstruction collapses, *i.e.*, visually similar but physically different parts of the scene are merged. Thus, for both scenes, we first triangulate 3D points using the original pGT. Next, we apply 10 iterations consisting of BA followed by merging and completing 3D points: nearby 3D points with matching features are merged and new features are added to 3D points if possible.

Some images of 12Scenes that were not registered by BundleFusion were reconstructed using COLMAP. Also, for the office2/5a and 5b scenes, we removed 61 images (out of 3,354 images contained in both scenes together) that we identified as obvious outliers via visual inspection.

4. Comparison of Pseudo Ground Truths

Given the two versions of pGT for each scene, estimated using (RGB-)D SLAM respectively SfM, a natural question is whether one version is more precise than the other. In this section, we quantitatively and qualitatively show that no version of the pGT is clearly preferably over the other: we first show that the SfM pGT outperforms the (RGB-)D SLAM version according to metrics that are optimised during the SfM process. We then show that the (RGB-)D SLAM pGT in turn outperforms the SfM version in terms of dense 3D point alignment, *i.e.*, the metrics optimised by depth-based methods. Thus, both versions can be considered as valid pGT for re-localisation experiments. Note that our analysis is focused on the two particular datasets. For a more general analysis of various reference algorithms, *e.g.*, about the influence of calibration accuracy, we refer to [62].

Evaluation based on SfM metrics. The first experiment focuses on standard metrics used to evaluate SfM reconstructions [61]. We measure the number of 3D points (#3D), the number of feature observations (#feat.) used to triangulate the 3D points, the average track length (track), *i.e.*, the average number of features used to triangulate a 3D point, and the average reprojection error (err.). For the same num-

		Chess			Fire			Heads			Office			Pumpkin			Red Kitchen			Stairs		
	Ref. Algo.	orig.	+BA	SfM	orig.	+BA	SfM	orig.	+BA	SfM	orig.	+BA	SfM	orig.	+BA	SfM	orig.	+BA	SfM	orig.	+BA	SfM
7Scenes	#3D	433k	204k	190k	628k	316k	296k	104k	73k	70k	515k	261k	249k	282k	131k	150k	1.0M	455k	472k	178k	119k	132k
	#feat.	7.4M	7.7M	**7.9M**	9.9M	10.2M	**10.4M**	1.4M	1.4M	**1.5M**	7.4M	7.8M	8.0M	4.5M	5.0M	**5.2M**	14.1M	15.8M	**16.5M**	2.3M	2.4M	**2.6M**
	track	17.0	37.7	**41.6**	15.7	32.3	**35.0**	13.4	19.8	**20.9**	14.3	29.9	32.1	15.9	**38.2**	35.2	13.5	34.7	**35.0**	13.0	**20.2**	19.5
	err. [px]	1.74	1.40	**1.25**	1.54	0.95	**0.88**	1.49	1.10	**1.01**	1.68	1.25	1.12	1.76	1.40	**1.24**	1.73	1.28	**1.13**	1.62	1.25	**1.10**

	apt1/kitchen			apt1/living			apt2/bed			apt2/kitchen			apt2/living			apt2/luke			office1/gates362		
#3D	146k	106k	104k	166k	112k	120k	245k	201k	171k	208k	121k	119k	148k	116k	121k	201k	135k	140k	658k	424k	419k
#feat.	1.3M	1.3M	**1.4M**	1.4M	1.5M	**1.6M**	1.9M	2.0M	**2.2M**	2.8M	2.9M	**3.0M**	1.2M	1.2M	**1.3M**	1.5M	**1.7M**	**1.7M**	9.7M	10.1M	**10.3M**
track	8.6	12.6	**13.0**	8.3	**13.6**	13.3	7.8	11.9	**12.7**	13.3	24.0	**24.9**	7.8	10.4	10.4	7.3	12.4	**12.5**	14.7	23.8	**24.7**
err. [px]	1.63	1.33	**1.25**	1.72	1.38	**1.28**	1.58	1.03	**0.97**	1.72	1.21	**1.07**	1.59	1.18	**1.12**	1.75	1.42	**1.33**	1.69	1.31	**1.18**

	office1/gates381			office1/lounge			office1/manolis			office2/5a		office2/5b	
#3D	695k	447k	471k	161k	116k	120k	364k	275k	273k	261k	202k	607k	580k
#feat.	6.8M	7.5M	**7.9M**	1.4M	1.4M	**1.5M**	3.7M	3.8M	**3.9M**	1.7M	**2.0M**	4.0M	**4.6M**
track	9.9	16.8	16.7	8.5	12.4	**12.5**	10.0	13.7	**14.2**	6.3	9.7	6.7	7.9
err. [px]	1.60	1.17	**1.09**	1.69	1.27	**1.19**	1.60	1.27	**1.19**	1.57	0.83	1.43	0.80

Table 1. **SfM statistics for different pseudo ground truth (pGT) versions.** We report the number of 3D points (#3D), the number of triangulated features (#feat.), the average track length (track), and the mean reprojection error (err.) in px. Longer tracks generated from more features and a lower reprojection error indicate more accurate poses. We show results for the original [64, 74] training and test poses generated by (RGB-)D SLAM (orig.), the original poses iteratively refined by alternating bundle adjustment and point merging (+BA), and the SfM pGT (SfM). For office2/5a and office2/5b, SfM from scratch failed and the SfM pGT is generated using the +BA strategy.
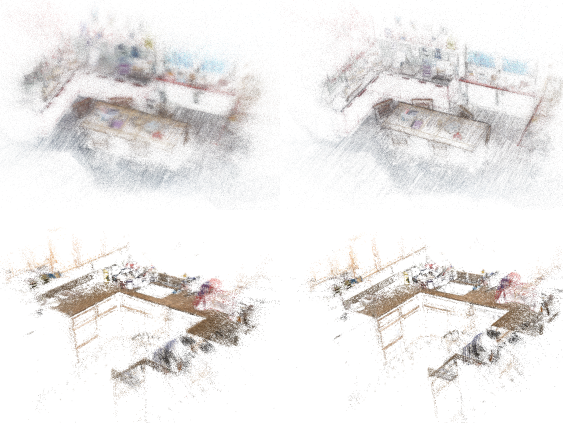

Figure 3. **Visualisation of SfM point clouds** for Red Kitchen (7Scenes) and apt2/kitchen (12Scenes) obtained using the original (left) and the SfM pGT (right) poses. The SfM pGT results in less noisy models where structures are more clearly visible.

ber of images in a 3D model, more observations and longer tracks, especially in combination with a lower reproj. error, indicate higher camera pose accuracy. Shorter tracks, *i.e.*, more 3D points, indicate that a single physical 3D point is represented by multiple SfM points: due to pose inaccuracies, no single SfM point projects within the error threshold used for robust triangulation [60] for all its measurements.

We compare the SfM pGT with point clouds obtained by triangulating the scenes from the original (RGB-)D pGT. For 7Scenes, we adjust the original pGT to account for the offset between RGB camera and depth sensor using the calibration from [78]. We use the same set of matches and the same COLMAP parameters for both pGT versions and use training and test images to calculate the statistics.

Tab. 1 shows the SfM metrics for both datasets. The SfM pGT clearly outperforms the original (RGB-)D SLAM pGT in the number of observations, track length, and reprojection error, especially on 7Scenes. We attribute this to the fact that KinectFusion, in contrast to BundleFusion, does not perform global optimisation and is thus suscepti-

ble to drift [74]. Fig. 3 qualitatively compares the SfM point clouds obtained with both versions of the pGT, showing that the SfM pGT leads to significantly less noisy SfM points.

As way to measure the similarity of the local optima found by the different pGT algorithms, we generate an "intermediate" pGT, denoted as +BA in Tab. 1: starting from the original pGT, we alternate between BA of the triangulated 3D model and merging and completing 3D points. As for office2/5a and office2/5b, we repeat this process for 10 iterations. In case that the local minima found by the (RGB-)D SLAM and SfM algorithms are close-by, we expect this process to result in a similar local optimum for the SfM metrics.[2] As can be seen in Tab. 1, the "intermediate" pGT results in similar or slightly worse statistics compared to the SfM pGT for both datasets. This indicates that the difference between poses is not large enough for bundle adjustment to result in significantly different local minima.

Evaluation based on 3D alignment metrics. We next evaluate how accurately the two pGT versions align the depth maps available for each image. For a pair of images (A, B) in a scene, we use the pGT poses to transform their depth maps to 3D point clouds in scene coordinates. For each 3D point in A's depth map, we find the nearest point in B's depth map. We report the root mean square error (RMSE) of all point correspondences below a 5cm outlier threshold.[3] This cost function, implemented in Open3D [83], measures the 3D alignment of the two point clouds and replicates the metric minimised by algorithms such as KinectFusion [30, 47] and BundleFusion [16].

We select image pairs for evaluation based on visual overlap in the SfM pGT [49]: Let $|P_{AB}|$ be the number of 3D points jointly observed by images A and B and $|P_A|$ and $|P_B|$ be the number of 3D points seen in A respectively

---

[2]We compare the "intermediate" pGT to the other pGT based on SfM metrics rather than comparing pose errors. The alignment between SfM and SLAM pGT introduces a potential error that we cannot easily remove.

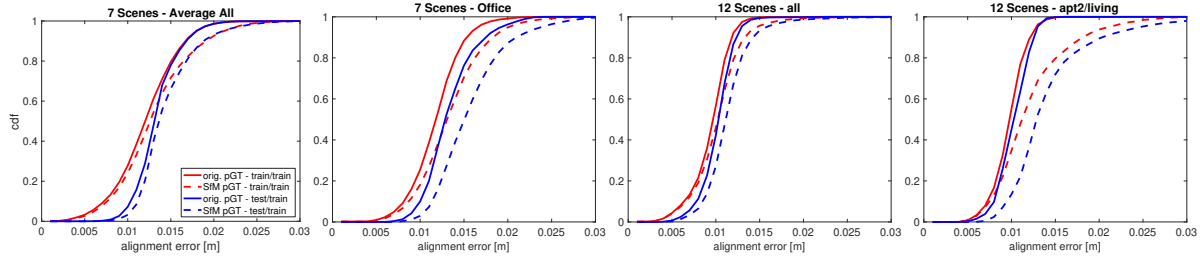[3]We did not observe image pairs with no correspondences within 5cm.

Figure 4. **3D alignment statistics** for the 7Scenes and 12Scenes datasets. We show cumulative distributions (cdfs) of the 3D alignment errors between the depth maps of train/train and test/train image pairs with a visual overlap of at least 30% for the original (RGB-)D SLAM and the SfM pseudo GT. As can be seen, the original pseudo GT produces more accurate alignments of the depth maps.

B. We consider a pair if $|P_{AB}|/\max(|P_A|,|P_B|) \geq 0.3$.

Fig. 4 shows cumulative histograms over the alignment errors for both pGT versions. We separately show curves for pairs of training images and pairs containing one test and one training image. The former measures the consistency between the training images and the latter measures how well the test images align with the training images. Since images are taken in continuous sequences, there are smaller changes in viewpoint between pairs of training images than for pairs containing training and test images. As such, there is a larger error for test/train pairs than for train/train pairs.

Fig. 4 shows smaller alignment errors for the orig. pGT. We also show dense point clouds obtained by fusing individual depth maps using the different pGT poses in Fig. 5 and Fig. 6. While the SfM pGT leads to globally more consistent geometry with less drift[4], fine details of foreground objects are better recovered with the original pGT. This confirms the results from Fig. 4, which show a more precise relative alignment of depth maps for the orig. pGT.

5. Re-localisation Evaluation

Sec. 4 showed that neither the original (RGB-)D SLAM pGT nor the SfM pGT is clearly better than the other. Thus, both pGT versions are valid choices for evaluating re-localisation algorithms. Their differences are in the order of centimeters. However, this is typically the range used to measure localisation accuracy. This section thus investigates how different pGT versions affect the performance of re-localisers. We show that RGB-D baselines fare better on pGT generated with (RGB-)D SLAM. Baselines that minimise a reprojection error perform better on the SfM pGT.

Evaluation measures. We report the percentage of images localised within X cm and X° of the respective pGT [64, 74]. We also report the Dense Correspondence Reprojection Error (DCRE) [77]: for each test image, we back-project the depth map into a 3D point cloud using its pGT pose. We project each 3D point into the image using the estimated and the pGT pose and measure the 2D distance between both projections. We report the maximum DCRE per test image below and the mean DCRE per test image in the supp. mat.

---

[4]Note that global consistency might not always be necessary, *e.g.*, AR applications where a user observes only a small part of the scene.

Baselines. We evaluate classical, feature-based as well as learning-based re-localisers on the two versions of the pGT. Learning-based methods are re-trained on each pGT, and feature-based methods use each version of the pGT to create their map. Please see the supp. mat. for details.

DSAC* [2,4,7] is a learning-based scene coordinate regression approach, where a neural network predicts for each pixel the corresponding 3D point in scene space. DSAC* uses a PnP [24] solver and RANSAC [22] on top of the 2D-3D matches. Its RGB-D variant, **DSAC* (+D)**, uses image depth to establish 3D-3D matches and a Kabsch [32] solver. **hLoc** [52] combines image retrieval with SuperPoint [18] features and SuperGlue [53] for matching, followed by P3P+RANSAC-based pose estimation. **DenseVLAD+R2D2** [29,50,71] uses DenseVLAD [71] for retrieving image pairs and R2D2 features for matching. The training images and poses are used to construct a 3D SfM map, and test images are localised using 2D-3D matches and P3P+RANSAC. Instead of triangulating point matches, **DenseVLAD+R2D2 (+D)** constructs the 3D map by projecting R2D2 keypoints to 3D space using depth maps. **Active Search (AS)** [54, 55] is a classical feature-based approach that establishes 2D-3D correspondences based on prioritized SIFT [41] matching. AS estimates the camera pose with a P3P solver [27,36] inside a RANSAC loop [22].

Results. Tab. 2 reports the percentage of test images localised within 5cm and 5° of the pGT for the 7Scenes dataset. For the original pGT, depth-based *DSAC* (+D)* clearly outperforms all other methods. Depth-based *DenseVLAD+R2D2 (+D)* achieves the best results among all sparse feature-based methods. *AS*, using classical SIFT features, achieves the lowest accuracy using the original pGT.

The ranking changes drastically using the SfM pGT. *AS* jumps from last to first place with an absolute difference of +29.8 in pose accuracy, outperforming all learning-based and depth-based competitors. Particularly notable are the results on Pumpkin and Red Kitchen, where AS improves from localising less than 50% within the 5cm, 5° threshold to localizing more than 99% of the images. For both scenes, Tab. 1 shows a significant difference in the SfM statistics between the two pGT versions. The previously-leading depth-based *DSAC* (+D)* and *DenseVLAD+R2D2 (+D)* drop to
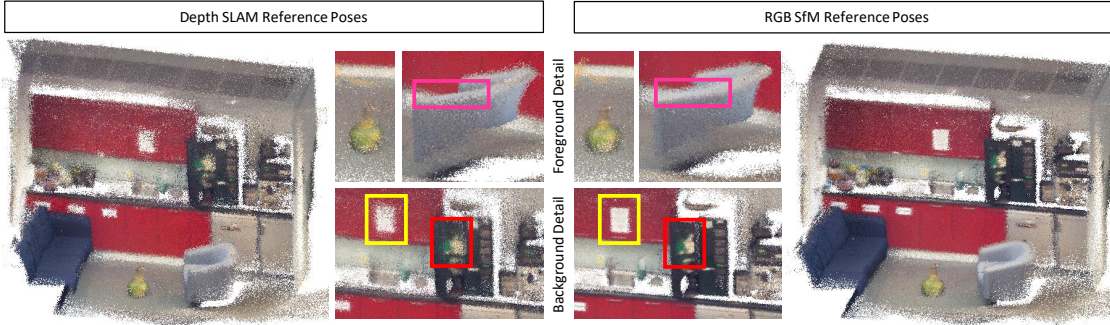
Figure 5. **Consistency of depth map fusion for 7Scenes [64]**. We reproject depth maps of the *Pumpkin* scene and accumulate the resulting 3D point clouds in world space using the pGT reference poses. While the depth SLAM pGT leads to sharper reconstructions of foreground objects like the pumpkin and chair, SfM pGT leads to better alignment of distant objects like the paper note and coffee machine.

Figure 6. **Consistency of depth map fusion for 12Scenes [74]. Left:** RGB-D SLAM pseudo GT leads to a significant warp of the global scene layout, which is clearly reduced by the SfM pseudo GT. **Right:** Foreground objects appear blurred for SfM pseudo GT poses compared to the sharper depth alignment generated by RGB-D SLAM pGT.
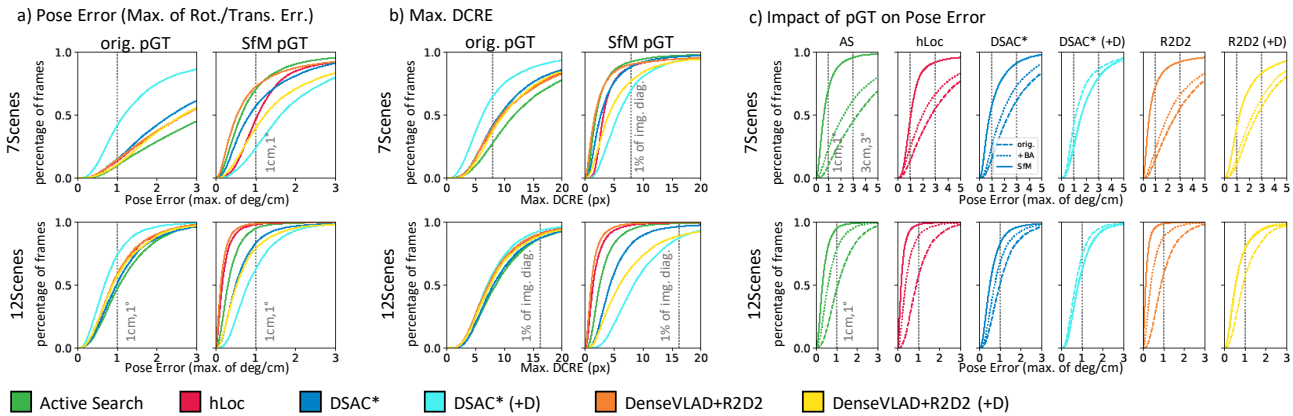

Figure 7. **Pose error and Dense Correspondence Reprojection Error (DCRE) [77]. a)** Cum. distributions of pose error (max. of rotation and position error) for 7Scenes (top row) and 12Scenes (bottom row), averaged across scenes. **b)** Cum. distribution of the maximum DCRE per test image. We mark 1% of the image diagonal in the DCRE plots. The relative ordering of the methods changes when switching between pGTs. **c)** Changes in the cum. pose error per method and pGT (orig. pGT, locally bundle-adjusted pGT (+BA), and SfM pGT).

the last places. Particularly, both methods are outperformed by their RGB-only counterparts when using the SfM pGT.

We can correlate these observations with each method's similarity to the respective reference algorithm (*c.f.* column 1 of Tab. 2 for a coarse classification). We regard methods that optimise a reprojection error over sparse features as *similar to SfM* and methods that optimise a dense 3D-3D error as *similar to (RGB-)D SLAM*. The RGB

	Chess			Fire			Heads			Office			Pumpkin			Red Kitchen			Stairs			**Average**		
Pseudo GT	orig.	+BA	SfM	orig.	+BA	SfM	orig.	+BA	SfM	orig.	+BA	SfM	orig.	+BA	SfM	orig.	+BA	SfM	orig.	+BA	SfM	orig.	+BA	SfM
Active Search	86.4	98.6	99.9	86.3	96.6	**99.8**	95.7	96.5	**100**	65.6	76.7	98.6	34.1	44.9	99.6	45.1	61.1	**99.8**	67.8	82.0	91.9	68.7	79.5	**98.5**
hLoc	94.2	99.6	**100**	93.7	97.0	99.4	99.7	99.9	**100**	83.2	88.7	**100**	55.2	65.5	**100**	61.9	72.7	98.6	49.4	58.1	72.0	76.8	83.1	95.7
DVLAD+R2D2	94.0	97.6	**100**	95.3	94.8	99.1	95.6	95.6	97.0	78.8	84.1	99.7	59.2	69.5	98.8	61.2	69.6	98.4	59.2	69.6	76.9	77.6	83.0	95.7
DVLAD+R2D2(+D)	93.1	98.2	**100**	91.3	95.3	98.5	96.5	96.4	96.3	81.2	86.1	96.6	58.5	68.0	91.6	72.6	78.5	97.8	68.0	72.8	69.5	80.2	85.0	92.9
DSAC*	97.8	99.2	99.9	94.5	**98.7**	98.9	98.8	99.6	99.8	83.9	89.8	98.1	62.0	73.9	99.0	65.5	79.1	97.0	77.7	**91.5**	92.0	82.9	90.3	97.8
DSAC*(+D)	**99.4**	**99.7**	99.6	**98.9**	98.4	96.9	**99.9**	**100**	99.5	**98.9**	**96.4**	95.3	**80.9**	**77.6**	90.9	**92.4**	**94.9**	96.4	**92.6**	85.5	88.4	**94.7**	**93.2**	95.3

Table 2. **Re-localisation results on 7Scenes [64]** as the percentage of images localised within 5cm and 5° of the original pGT generated by Depth SLAM (*orig.*), local SfM minima obtained by fixing the orig. pGT training poses and optimising the test poses (*+BA*; see Sec. 5 for details), and the SfM pGT (*SfM*). We visualise the ranking of methods from best to worst depending on the pGT. We color-code methods based on their similarity to the reference algorithm: similar to SfM , similar to D-SLAM , or intermediary .

variant of *DSAC** optimises a dense reprojection error. *DVLAD+R2D2(+D)* optimises a sparse reprojection error but incorporates depth when building the 3D map. Thus, we classify those two methods as *intermediary*. Among methods *similar to SfM*, AS shows the largest improvement under the SfM pGT as it re-uses the SIFT features from SfM.

Fig. 7(a) shows cumulative distributions over the fraction of images localised within Xcm, X° of the pGT for tighter thresholds than used before. This is particularly interesting for 12Scenes, where the accuracy of all methods saturates under the 5cm, 5° threshold. Poses predicted by *DSAC* (+D)* better align with the original (RGB-)D SLAM pGT than with the SfM pGT. At the same time, poses predicted by RGB-based methods better align with the SfM pGT. There are larger differences between the methods for finer thresholds. For 12Scenes, *hLoc* and *DenseVLAD+R2D2* achieve the highest accuracy under a 1cm, 1° threshold.

Fig. 7(b) shows cumulative distributions for the max. DCRE. Since the DCRE depends on the pose accuracy, we observe the same behavior as before, *i.e.*, methods more similar to SfM outperform depth-based methods on the SfM pGT while performing worse on the original (RGB-D) pGT. Yet, this does not necessarily imply that such methods are superior. They closely resemble the SfM ref. algorithm, and they use the 3D points triangulated by the SfM pipeline from the training images for pose estimation. Thus, it seems likely that feature-based methods "overfit" to the SfM pGT by being able to closely replicate SfM behavior. To further illustrate the issue, we created an "intermediate" pGT: starting with the original pGT poses, we triangulate the scene and use bundle adjustment followed by point merging to optimise the test poses while keeping the training poses fixed. Intuitively, the resulting poses, denoted as "+BA", approximate the "optimal" test poses for the original training image pGT under the reprojection error metric. Fig. 7(c) and Tab. 2 show results obtained using the +BA test poses. The +BA poses significantly improve the evaluation scores of RGB-based methods such as AS, but less so for depth-based methods such as DSAC* (+D) or DenseVLAD+R2D2 (+D). The closer the pGT is to the cost function optimised by both methods, the better they perform. In contrast, depth-based methods typically either perform similar or worse under these pGT poses. Our results

indicate that learning-based methods might have some capacity to adjust to the pGT since DSAC* ranks well across all pGT versions. Still, DSAC* is always outperformed by methods more similar to the ref. algorithm.

6. Conclusion

Re-localisation benchmarks usually rely on a reference algorithm to create pseudo ground truth for evaluation. As such, they do not measure absolute pose accuracy but rather how well a given method is able to reproduce the reference output. Our paper points out an important implication: different cost functions optimised by reference algorithms lead to different local minima. This affects re-localisation evaluation as methods that optimise a similar cost function as the reference algorithm better replicate the local minima and imperfections of the pGT, to a degree that relative rankings can be (nearly) completely inverted. This issue is fundamental, and we do not see a solution to this problem. However, there are ways to address the issue, as shown in Sec. 5: new benchmark datasets could provide multiple pGT versions to enable a more concise evaluation that takes the impact of the pGT into account. *E.g.*, although DSAC* does not perform *best* under any pGT, it performs *well* under all pGT versions. If multiple pGT versions are not available, localisation algorithms can be grouped based on their similarity to the reference algorithm (*c.f.* color-coding in Tab. 2) and only be compared within but not between groups. Another approach is to choose evaluation thresholds that are large enough that the difference in pGT will not affect the measured performance, *e.g.*, 5cm, 5° for 12Scenes. Such an approach will likely have to explicitly account for the uncertainties in the estimated poses, which itself is a complex problem [23]. Still, knowledge about pose uncertainties would allow us to determine when a dataset is solved. Another direction is a task-specific evaluation of re-localisation methods, *e.g.*, measuring their performance in the context of AR, robotic navigation, *etc*. Again, understanding the impact of the pGT on such evaluations is an interesting and open problem.

References

[1] Clemens Arth, Manfred Klopschitz, Gerhard Reitmayr, and Dieter Schmalstieg. Real-Time Self-Localization from Panoramic Images on Mobile Devices. In *ISMAR*, 2011. 1

[2] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. DSAC - Differentiable RANSAC for Camera Localization. In *CVPR*, 2017. 2, 3, 6

[3] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In *CVPR*, 2016. 2, 3

[4] Eric Brachmann and Carsten Rother. Learning Less is More - 6D Camera Localization via 3D Surface Regression. In *CVPR*, 2018. 2, 3, 6

[5] Eric Brachmann and Carsten Rother. Expert sample consensus applied to camera re-localization. In *ICCV*, 2019. 2, 3

[6] Eric Brachmann and Carsten Rother. Neural-guided RANSAC: Learning where to sample model hypotheses. In *ICCV*, 2019. 2, 3

[7] Eric Brachmann and Carsten Rother. Visual camera relocalization from RGB and RGB-D images using DSAC. *TPAMI*, 2021. 2, 3, 6

[8] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *CVPR*, 2018. 2

[9] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 2016. 2

[10] Robert Castle, Georg Klein, and David W. Murray. Videorate localization in multiple maps for wearable augmented reality. In *ISWC*, 2008. 1

[11] Tommaso Cavallari, Luca Bertinetto, Jishnu Mukhoti, Philip Torr, and Stuart Golodetz. Let's take this online: Adapting scene coordinate regression network predictions for online RGB-D camera relocalisation. In *3DV*, 2019. 2, 3

[12] Tommaso Cavallari, Stuart Golodetz, Nicholas A. Lord, Julien Valentin, Luigi Di Stefano, and Philip H. S. Torr. On-The-Fly Adaptation of Regression Forests for Online Camera Relocalisation. In *CVPR*, 2017. 2, 3

[13] Tommaso Cavallari, Stuart Golodetz, Nicholas A. Lord, Julien Valentin, Victor A. Prisacariu, Luigi Di Stefano, and Philip H. S. Torr. Real-time RGB-D camera pose estimation in novel scenes using a relocalisation cascade. *TPAMI*, 2019. 3

[14] David M. Chen, Georges Baatz, Kevin Köser, Sam S. Tsai, Ramakrishna Vedantham, Timo Pylvänäinen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, Bernd Girod, and Radek Grzeszczuk. City-Scale Landmark Identification on Mobile Devices. In *CVPR*, 2011. 1, 2

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2

[16] Angela Dai, Matthias Nießner, Michael Zollöfer, Shahram Izadi, and Christian Theobalt. BundleFusion: Real-time Globally Consistent 3D Reconstruction using On-the-fly Surface Re-integration. *ACM Transactions on Graphics 2017 (TOG)*, 2017. 1, 2, 3, 4, 5

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 2

[18] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. 3, 6

[19] Axel Donath and Daniel Kondermann. Is crowdsourcing for optical flow ground truth generation feasible? In Mei Chen, Bastian Leibe, and Bernd Neumann, editors, *ICVS*, 2013. 2

[20] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A trainable CNN for joint detection and description of local features. In *CVPR*, 2019. 3

[21] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 2

[22] Martin A. Fischler and Robert C. Bolles. Random Sampling Consensus: A Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography. *CACM*, 24:381–395, 1981. 6

[23] Wolfgang Förstner and Bernhard P. Wrobel. *Photogrammetric Computer Vision: Statistics, Geometry, Orientation and Reconstruction*. Springer Publishing Company, Incorporated, 1st edition, 2016. 8

[24] Xiao-Shan Gao, Xiao-Rong Hou, Jianliang Tang, and Hang-Fei Cheng. Complete solution classification for the perspective-three-point problem. *TPAMI*, 25(8):930–943, 2003. 6

[25] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *IJRR*, 32(11):1231–1237, 2013. 2

[26] Abner Guzman-Rivera, Pushmeet Kohli, Ben Glocker, Jamie Shotton, Toby Sharp, Andrew Fitzgibbon, and Shahram Izadi. Multi-output learning for camera relocalization. In *CVPR*, 2014. 2

[27] Bert M. Haralick, Chung-Nan Lee, Karsten Ottenberg, and Michael Nölle. Review and analysis of solutions of the three point perspective pose estimation problem. *IJCV*, 13(3):331–356, 1994. 6

[28] Lionel Heng, Benjamin Choi, Zhaopeng Cui, Marcel Geppert, Sixing Hu, Benson Kuan, Peidong Liu, Rang Nguyen, Ye Chuan Yeo, Andreas Geiger, Gim Hee Lee, Marc Pollefeys, and Torsten Sattler. Project AutoVision: Localization and 3D Scene Perception for an Autonomous Vehicle with a Multi-Camera System. In *ICRA*, 2019. 1

[29] Martin Humenberger, Yohann Cabon, Nicolas Guerin, Julien Morat, Jérôme Revaud, Philippe Rerole, Noé Pion, Cesar

de Souza, Vincent Leroy, and Gabriela Csurka. Robust Image Retrieval-based Visual Localization using Kapture. arXiv:2007.13867, 2020. 3, 6

[30] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and Andrew Fitzgibbon. KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera. In *ACM Symposium on User Interface Software and Technology (UIST)*, 2011. 1, 2, 3, 5

[31] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image Matching across Wide Baselines: From Paper to Practice. *IJCV*, 129:517–547, 2021. 2

[32] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 1976. 6

[33] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *CVPR*, 2017. 2

[34] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *ICCV*, 2015. 1, 2, 3, 4

[35] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics (TOG)*, 36(4), 2017. 2

[36] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. A Novel Parametrization of the Perspective-Three-Point Problem for a Direct Computation of Absolute Camera Position and Orientation. In *CVPR*, 2011. 6

[37] Xiaotian Li, Shuzhe Wang, Yi Zhao, Jakob Verbeek, and Juho Kannala. Hierarchical scene coordinate classification and regression for visual localization. In *CVPR*, 2020. 3

[38] Yunpeng Li, Noah Snavely, and Dan P. Huttenlocher. Location Recognition using Prioritized Feature Matching. In *ECCV*, 2010. 1, 2, 3, 4

[39] Yunpeng Li, Noag Snavely, Dan P. Huttenlocher, and Pascal Fua. Worldwide Pose Estimation Using 3D Point Clouds. In *ECCV*, 2012. 1, 2, 4

[40] Hyon Lim, Sudipta N. Sinha, Michael F. Cohen, and Matthew Uyttendaele. Real-Time Image-Based 6-DOF Localization in Large-Scale Environments. In *CVPR*, 2012. 1

[41] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2), 2004. 4, 6

[42] Simon Lynen, Torsten Sattler, Mike Bosse, Joel Hesch, Marc Pollefeys, and Roland Siegwart. Get Out of My Lab: Large-scale, Real-Time Visual-Inertial Localization. In *RSS*, 2015. 1

[43] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *IJRR*, 36(1):3–15, 2017. 1, 2

[44] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *ICCV*, 2019. 2

[45] Lili Meng, Jianhui Chen, Frederick Tung, James J. Little, Julien Valentin, and Clarence W. de Silva. Backtracking regression forests for accurate camera relocalization. In *IROS*, 2017. 3

[46] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 2

[47] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *IEEE ISMAR*, 2011. 1, 2, 3, 5

[48] Richard A. Newcombe, Steven J. Lovegrove, and Andrew J. Davison. DTAM: Dense tracking and mapping in real-time. In *ICCV*, 2011. 3

[49] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-Tuning CNN Image Retrieval with No Human Annotation. *TPAMI*, 41(7):1655–1668, 2019. 5

[50] Jerome Revaud, Philippe Weinzaepfel, César Roberto de Souza, and Martin Humenberger. R2D2: repeatable and reliable detector and descriptor. In *NeurIPS*, 2019. 6

[51] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the ICP algorithm. In *Proceedings Third International Conference on 3-D Digital Imaging and Modeling*, 2001. 4

[52] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *CVPR*, 2019. 3, 6

[53] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning Feature Matching with Graph Neural Networks. In *CVPR*, 2020. 3, 6

[54] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Improving Image-Based Localization by Active Correspondence Search. In *ECCV*, 2012. 2, 6

[55] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. *PAMI*, 39(9):1744–1756, 2017. 2, 3, 6

[56] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF Urban Visual Localization in Changing Conditions. In *CVPR*, 2018. 1, 2, 3, 4

[57] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image Retrieval for Image-Based Localization Revisited. In *BMVC*, 2012. 3

[58] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixé. Understanding the limitations of cnn-based absolute camera pose regression. In *CVPR*, 2019. 2

[59] Tanner Schmidt, Richard A. Newcombe, and Dieter Fox. Self-supervised visual descriptor learning for dense correspondence. *IEEE Robotics and Automation Letters*, 2(2):420–427, 2017. 3

[60] Johannes L. Schönberger and Jan-Michael Frahm. Structure-From-Motion Revisited. In *CVPR*, 2016. 1, 2, 3, 4, 5

[61] Johannes Lutz Schönberger, Hans Hardmeier, Torsten Sattler, and Marc Pollefeys. Comparative Evaluation of Hand-Crafted and Learned Local Features. In *CVPR*, 2017. 4

[62] Thomas Schöps, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle Adjusted Direct RGB-D SLAM. In *CVPR*, 2019. 1, 2, 4

[63] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A Multi-View Stereo Benchmark With High-Resolution Images and Multi-Camera Videos. In *CVPR*, 2017. 2

[64] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In *CVPR*, 2013. 1, 2, 3, 5, 6, 7, 8

[65] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Modeling the World from Internet Photo Collections. *IJCV*, 80(2):189–210, 2008. 4

[66] Christoph Strecha, Wolfgang von Hansen, Luc Van Gool, Pascal Fua, and Ulrich Thönnessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR*, 2008. 2

[67] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IROS*, 2012. 2

[68] Xun Sun, Yuanfan Xie, Pei Luo, and Liang Wang. A Dataset for Benchmarking Image-Based Localization. In *CVPR*, 2017. 4

[69] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. InLoc: Indoor visual localization with dense matching and view synthesis. In *CVPR*, 2018. 1, 2, 3

[70] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. *TPAMI*, 43(4):1293–1307, 2021. 3

[71] Akihiko Torii, Relja Arandjelović, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 Place Recognition by View Synthesis. In *CVPR*, 2015. 6

[72] Akihiko Torii, Hajime Taira, Josef Sivic, Marc Pollefeys, Masatoshi Okutomi, Tomas Pajdla, and Torsten Sattler. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? *TPAMI*, 43(3):814–829, 2021. 1, 2, 3, 4

[73] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle Adjustment — A Modern Synthesis. In *Vision Algorithms: Theory and Practice*, 2000. 4

[74] Julien Valentin, Angela Dai, Matthias Niessner, Pushmeet Kohli, Philip Torr, Shahram Izadi, and Cem Keskin. Learning to Navigate the Energy Landscape. In *3DV*, 2016. 1, 2, 3, 4, 5, 6, 7

[75] Julien Valentin, Matthias Nießner, Jamie Shotton, Andrew Fitzgibbon, Shahram Izadi, and Philip Torr. Exploiting Uncertainty in Regression Forests for Accurate Camera Relocalization. In *CVPR*, 2015. 2, 3

[76] Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-Based Localization Using LSTMs for Structured Feature Correlation. In *ICCV*, 2017. 2, 3

[77] Johanna Wald, Torsten Sattler, Stuart Golodetz, Tommaso Cavallari, and Federico Tombari. Beyond Controlled Environments: 3D Camera Re-Localization in Changing Indoor Scenes. In *ECCV*, 2020. 1, 2, 3, 6, 7

[78] Christian Wolf, Eric Lombardi, Julien Mille, Oya Celiktutan, Mingyuan Jiu, Emre Dogan, Gonen Eren, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot, Christophe Garcia, and Bülent Sankur. Evaluation of video activity localizations integrating quality and quantity measurements. *Comput. Vis. Image Underst.*, 127:14–30, 2014. 5

[79] Changchang Wu. VisualSFM: A visual structure from motion system. *URL: http://homes.cs.washington.edu/~ccwu/vsfm*, 9, 2011. 3

[80] Changchang Wu. Towards Linear-time Incremental Structure From Motion. In *3DV*, 2013. 1, 2, 4

[81] Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and Ping Tan. SANet: Scene agnostic network for camera localization. In *ICCV*, 2019. 3

[82] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference Pose Generation for Long-term Visual Localization via Learned Features and View Synthesis. *IJCV*, 2020. 1, 2

[83] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 5