Map-free Visual Relocalization: Metric Pose Relative to a Single Image

Eduardo Arnold[1,2⋆], Jamie Wynn[1], Sara Vicente[1], Guillermo Garcia-Hernando[1], Áron Monszpart[1], Victor Adrian Prisacariu[1,3], Daniyar Turmukhambetov[1], and Eric Brachmann[1]

[1]Niantic [2]University of Warwick [3]University of Oxford

github.com/nianticlabs/map-free-reloc

Abstract. Can we relocalize in a scene represented by a single reference image? Standard visual relocalization requires hundreds of images and scale calibration to build a scene-specific 3D map. In contrast, we propose *Map-free Relocalization*, *i.e.*, using only one photo of a scene to enable instant, metric scaled relocalization. Existing datasets are not suitable to benchmark map-free relocalization, due to their focus on large scenes or their limited variability. Thus, we have constructed a new dataset of 655 small places of interest, such as sculptures, murals and fountains, collected worldwide. Each place comes with a reference image to serve as a relocalization anchor, and dozens of query images with known, metric camera poses. The dataset features changing conditions, stark viewpoint changes, high variability across places, and queries with low to no visual overlap with the reference image. We identify two viable families of existing methods to provide baseline results: relative pose regression, and feature matching combined with single-image depth prediction. While these methods show reasonable performance on some favorable scenes in our dataset, map-free relocalization proves to be a challenge that requires new, innovative solutions.

1 Introduction

Given not more than a single photograph we can imagine what a depicted place looks like, and where we, looking through the lens, would be standing relative to that place. Visual relocalization mimics the human capability to estimate a camera's position and orientation from a single query image. It is a well-researched task that enables exciting applications in augmented reality (AR) and robotic navigation. State-of-the-art relocalization methods surpass human rule-of-thumb estimates by a noticeable margin [10,33,53,54,55,58], allowing centimeter accurate predictions of a camera's pose. But this capability comes with a price: each scene has to be carefully pre-scanned and reconstructed. First, images need to be gathered from hundreds of distinct viewpoints, ideally spanning different times of day and even seasons. Then, the 3D orientation and position

⋆ Work done during internship at Niantic

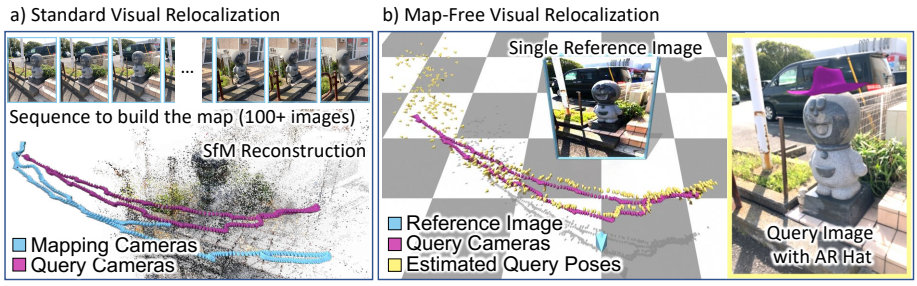

Fig. 1: Standard relocalization methods build a scene representation from hundreds of mapping images (a). For map-free relocalization (b), only a single photo (cyan) of the scene is available to relocalize queries. We show ground truth poses (purple) and estimated poses (yellow) in (b), and we use one estimate to render a virtual hat on the statue. We achieve these results with SuperGlue [54] feature matching and DPT [48] depth estimation.

of these images needs to be estimated, *e.g.*, by running structure-from-motion (SfM) [60,64,81,82] or simultaneous-localization-and-mapping (SLAM) [20,43] software. Oftentimes, accurate multi-camera calibration, alignment against LiDAR scans, high-definition maps or inertial sensor measurements are needed to recover poses in metric units, *e.g.*, [58,59]. Finally, images and their camera poses are fed to a relocalization pipeline. For traditional structure-based systems [33,53,55,57], the final scene representation consists of a point cloud triangulated from feature correspondences, and associated feature descriptors, see Fig. 1a).

The requirement for systematic pre-scanning and mapping restricts how visual relocalization can be used. For example, AR immersion might break if a user has to record an entire image sequence of an unseen environment first, gathering sufficient parallax by walking sideways, all in a potentially busy public space. Furthermore, depending on the relocalization system, the user then has to wait minutes or hours until the scene representation is built. We propose a new flavour of relocalization, termed *Map-free Relocalization*. We ask whether the mapping requirement can be relaxed to the point where a single reference image is enough to relocalize new queries in a metric coordinate system. Map-free relocalization enables instant AR capabilities at new locations: User A points their camera at a structure, takes a photo, and any user B can instantly relocalize w.r.t. user A. Map-free relocalization constitutes a systematic, task-oriented benchmark for two-frame relative pose estimation, namely between the reference image and a query image, see Fig. 1 b).

Relocalization by relative pose estimation is not new. For example, neural networks have been trained to regress metric relative poses directly from two images [4,42]. Thus far, such systems have been evaluated on standard relocalization benchmarks where structure-based methods rule supreme, to the extent where the accuracy of the ground truth is challenged [9]. We argue that we should

strive towards enabling new capabilities that traditional structure-based methods cannot provide. Based on a single photo, a scene cannot be reconstructed by SfM or SLAM. And while feature matching still allows to estimate the relative pose between two images, the reference and the query, there is no notion of absolute scale [32]. To recover a *metric* estimate, some heuristic or world knowledge has to be applied to resolve the scale ambiguity which we see as the key problem.

Next to pose regression networks, that predict metric poses by means of supervised learning, we recognize a second family of methods as suitable for map-free relocalization. We show that a combination of deep feature matching [54,66] and deep single-image depth prediction [48,40] currently achieves highest relative pose accuracy. To the best of our knowledge, this variant of relative pose estimation has not gained attention in relocalization literature thus far.

While we provide evidence that existing methods can solve map-free relocalization with acceptable precision, such results are restricted to a narrow window of situations. To stimulate further research in map-free relocalization, we present a new benchmark and dataset. We have gathered images of 655 places of interest worldwide where each place can be represented well by a single reference image. All frames in each place of interest have metric ground truth poses. There are 522,921 frames for training, 36,998 query frames across 65 places for validation, and 14,778 query frames (subsampled from 73,902 frames) across 130 places in the test set. Following best practice in machine learning, we provide a public validation set while keeping the test ground truth private, accessed through an online evaluation service. This dataset can serve as a test bed for advances in relative pose estimation and associated sub-problems such as wide-baseline feature matching, robust estimation and single-image depth prediction.
We summarize our **contributions** as follows:

- Map-free relocalization, a new flavor of visual relocalization that dispenses with the need for creating explicit maps from extensive scans of a new environment. A single reference image is enough to enable relocalization.
- A dataset that provides reference and query images of over 600 places of interest worldwide, annotated with ground truth poses. The dataset includes challenges such as changing conditions, stark viewpoint changes, high variability across places, and queries with low to no visual overlap with the reference image.
- Baseline results for map-free relocalization using relative pose regression methods, and feature matching on top of single image-depth prediction. We expose the primary problems of current approaches to guide further research.
- Additional experiments and ablation studies on ScanNet and 7Scenes datasets, allowing comparisons to related, previous research on relative pose estimation and visual relocalization.

2 Related Work

Scene Representations in Visual Relocalization: In the introduction, we have discussed traditional structure-based relocalizers that represent a scene by

an explicit SfM or SLAM reconstruction. As an alternative, recent learning-based relocalizers encode the scene *implicitly* in the weights of their neural networks by training on posed mapping images. This is true for both scene coordinate regression [62,10,11,12,38,14] and absolute pose regression (APR) [37,36,77,15,61]. More related to our map-free scenario, some relative pose regression (RPR) methods avoid training scene specific networks [4,73,80]. Given a query, they use image retrieval [72,3,51,33,46,45,59] to look up the closest database image and its pose. A generic relative pose regression network estimates the pose between query and database images to obtain the absolute pose of the query. RPR methods claim to avoid creating costly scene-specific representations but ultimately these works do not discuss how posed database images would be obtained without running SfM or SLAM. ExReNet[80], a recent RPR method, shows that the database of posed images can be extremely sparse, keeping as little as four strategically placed reference images to cover an indoor room. Although only a few images make up the final representation, continuous pose tracking is required when recording them. In contrast, map-free relocalization means keeping only a single image to represent a scene without any need for pose tracking or pose reconstruction. The reference image has the identity pose.

Relative Pose by Matching Features: The pose between two images with known intrinsics can be recovered by decomposing the essential matrix [32]. This yields the relative rotation, and a *scaleless* translation vector. The essential matrix is classically estimated by matching local features, followed by robust estimation, such as using a 5-point solver [44] inside a RANSAC [26] loop. This basic formula has been improved by learning better features [41,52,21,74,8], better matching [54,66] and better robust estimators [47,85,49,13,5,6,67], and progress has been measured in wide-baseline feature matching challenges [35] and small overlap regimes.

In the relocalization literature, scaleless pairwise relative poses between the query and multiple reference images have been used to triangulate the scaled, metric pose of a query [86,87,80]. However, for map-free relocalization only two images (reference and query) are available at any time, making query camera pose triangulation impossible. Instead, we show that estimated depth can be used to resolve the scale ambiguity of poses recovered via feature matching.

Relative Pose Regression (RPR): Deep learning methods that predict the relative pose from two input images bypass explicit estimation of 2D correspondences [75,42,24,4,80,1]. Some methods recover pose up to a scale factor [42,80] and rely on pose triangulation, while others aim to estimate metric relative pose [4,24,1]. Both RelocNet [4] and ExReNet [80] show generalization of RPR across datasets by training on data different from the test dataset.

Recently, RPR was applied in scenarios that are challenging for correspondence-based approaches. Cai et al. [16] focus on estimating the relative rotation between two images in extreme cases, including when there is no overlap between the two images. Similarly, the method in [18] estimates scaleless relative pose for pairs of images with very low overlap. We take inspiration from the methods

above to create baselines and discuss in more detail the different architectures and output parameterizations in Section 3.2.

Single-image Depth Prediction: Advances in deep learning have allowed practical methods for single-image depth estimation, *e.g.*, [23,29]. There are two versions of the problem: relative and absolute depth prediction. Relative, also called scaleless, depth prediction aims at estimating depth maps up to an unknown linear or affine transformation, and can use scaless training data such as SfM reconstructions [39], uncalibrated stereo footage [83,50] or monocular videos [88,30]. Absolute depth prediction methods (*e.g.*, [48,40,29,79]) aim to predict depth in meters by training or fine-tuning on datasets that have absolute metric depth such as the KITTI [28], NYUv2 [63] and ScanNet [19] datasets. Generalizing between domains (*e.g.*, driving scenes vs. indoors) is challenging as collecting metric depth in various conditions can be expensive. Moreover, generalization of a single network that is robust to different input image resolutions, aspect ratios and camera focal lengths is also challenging [25].

Recently, single-image depth prediction was leveraged in some pose estimation problems. In [71], predicted depth maps are used to rectify planar surfaces before local feature computation for improved relative pose estimation under large viewpoint changes. However, that work did not use metric depth information to estimate the scale of relative poses. Depth prediction was incorporated into monocular SLAM [69,70] and Visual Odometry [84,17] pipelines to combat scale drift and improve camera pose estimation. Predicted depths were used as a soft constraint in multi-image problem, while we use depth estimates to scale relative pose between two images.

3 Map-free Relocalization

Our aim is to obtain the camera pose of a query image given a single RGB reference image of a scene. We assume intrinsics of both images are known, as they are generally reported by modern devices. The absolute pose of a query image Q is parameterized by $R \in SO(3)$, $t \in \mathbb{R}^3$, which maps a world point $\mathbf{y}$ to point $\mathbf{x}$ in the camera's local coordinate system as $\mathbf{x} = R\mathbf{y} + t$. Assuming the global coordinate system is anchored to the reference image, the problem of estimating the absolute pose of the query becomes one of estimating a scaled relative pose between two images. Next, we discuss different approaches for obtaining a metric relative pose between a pair of RGB images. The methods are split into two categories: methods based on feature matching with estimated depth, and methods based on direct relative pose regression.

3.1 Feature Matching and Scale from Estimated Depth

The relative pose from 2D correspondences is estimated up to scale via the Essential matrix [32]. We consider SIFT [41] as a traditional baseline as well as more recent learning-based matchers such as SuperPoint + SuperGlue [54] and

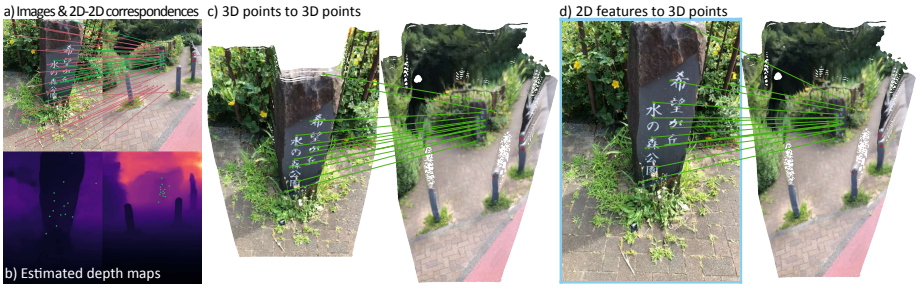

Fig. 2: Given the reference and query images, we obtain 2D-2D correspondences using the feature matching method in [54] (a). Inlier correspondences for the robust RANSAC-based essential matrix computation are visualized in green and outlier correspondences in red. Estimated monocular depth maps using [48] are shown in (b). The depth maps can be coupled with the 2D-2D correspondences to obtain 3D-3D correspondences (c) or 2D-3D correspondences (d), which are used in the geometric methods discussed in Section 3.1.

LoFTR [66]. To recover the missing scale, we utilize monocular depth estimation. For indoors, we experimented with DPT [48] fine-tuned on the NYUv2 dataset [63] and PlaneRCNN [40], which was trained on ScanNet [19]. For outdoors, we use DPT [48] fine-tuned on KITTI [28]. Given estimated depth and 2D correspondences we compute scaled relative poses in the following variants. See also Fig. 2 for an illustration.

(2D-2D) Essential matrix + depth scale (*Ess.Mat. + D.Scale)*: We compute the Essential matrix using a 5-point solver [44] with MAGSAC++ [6] and decompose it into a rotation and a unitary translation vector. We back-project MAGSAC inlier correspondences to 3D using the estimated depth. Each 3D-3D correspondence provides one scale estimate for the translation vector, and we select the scale estimate with maximum consensus across correspondences, see the supplemental material for details.

(2D-3D) Perspective-n-Point (PnP): Using estimated depth, we back-project one of the two images to 3D, giving 2D-3D correspondences. This allows us to use a PnP solver [27] to recover a metric pose. We use PnP within RANSAC [26] and refine the final estimate using all inliers. We use 2D features from the query image and 3D points from the reference image.

(3D-3D) Procrustes: Using estimated depth, we back-project both images to 3D, giving 3D-3D correspondences. We compute the relative pose using Orthogonal Procrustes [22] inside a RANSAC loop [26]. Optionally, we can refine the relative pose using ICP [7] on the full 3D point clouds. This variant performs significantly worse compared to the previous two, so we report its results in the supplemental material.

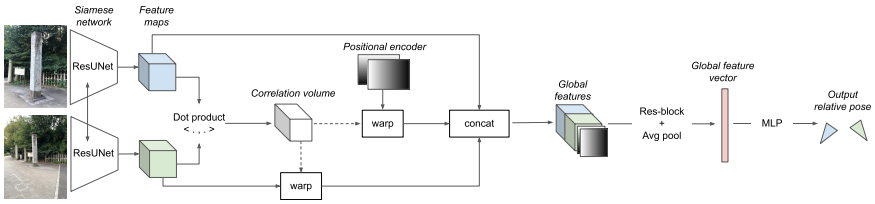

Fig. 3: Overview of the network architecture for RPR. We use a Siamese network (ResUNet [16]) to extract features from the two input images. Following [16,80], we compute a 4D correlation volume to mimic soft feature matching. The correlation volume is used to warp the features of the second image and a regular grid of coordinates (positional encoding). These are concatenated channel-wise with the first image's feature map to create the global feature map. The global volume is fed to four ResNet blocks followed by global average pooling, resulting in a single 512-dimensional global feature vector. Finally, an MLP generates the output poses. See supplement for details.

3.2 Relative Pose Regression

Relative pose regression (RPR) networks learn to predict metric relative poses in a forward pass. We implement a baseline architecture following best practices reported in the literature [87,80,89] – see Fig. 3, and the supplement for more details. In the following, we focus on the different output parameterizations and leave a discussion about losses and other design choices to the supplement.

RPR networks often parameterize rotations as **quaternions** [24,42,80] (denoted as $R(q)$). [89] argues that a **6D parameterization** of rotation avoids discontinuities of other representations: the network predicts two 3D vectors and creates an orthogonal basis through a partial Gram-Schmidt process (denoted as $R(6D)$). Finally, for rotation, we experiment with **Discrete Euler angles** [16], denoted as $R(\alpha, \beta, \gamma)$. Following [16], we use 360 discrete values for yaw and roll, and 180 discrete values for the pitch angle. For the translation vector we investigate three parameterization options: predicting the **scaled translation** (denoted as t), predicting a **scale and unitary translation** separately (denoted as $s \cdot \hat{t}$), and **scale and discretized unitary translation**. For the latter we predict translation in spherical coordinates ϕ, θ with quantized bins of 1deg as well as a 1D scale (denoted as $s \cdot \hat{t}(\phi, \theta)$). As an alternative which model rotation and translation jointly, we adapt the method of [68] which predicts 3D-3D correspondences for predefined keypoints of specific object classes. Here, we let the network predict **three 3D-3D correspondences** (denoted as $[3D - 3D]$). We compute the transformation that aligns these two sets of point triplets using Procrustes, which gives the relative rotation and translation between the two images. The models are trained end-to-end until convergence by supervising the output pose with the ground truth relative pose. We experimented with different loss functions and weighting between rotation and translation losses. For details, see supplemental material.

4 Map-free Relocalization Datasets

In this section, we first discuss popular relocalization datasets and their limitations for map-free relocalization. Then, we introduce the Niantic map-free relocalization dataset which was collected specifically for the task. Finally, we define evaluation metrics used to benchmark baseline methods.

4.1 Existing Relocalization Datasets

One of the most commonly used datasets for visual relocalization is 7Scenes [62], consisting of seven small rooms scanned with KinectFusion [34]. 12Scenes [76] provides a few more, and slightly larger environments, while RIO10 [78] provides 10 scenes focusing on condition changes between mapping and query images. For outdoor relocalization, Cambridge Landmarks [37] and Aachen Day-Night [58], both consisting of large SfM reconstructions, are popular choices.

We find existing datasets poorly suited to benchmark map-free relocalization. Firstly, their scenes are not well captured by a single image which holds true for both indoor rooms and large-scale outdoor reconstructions. Secondly, the variability across scenes is extremely limited, with 1-12 distinct scenes in each single dataset. For comparison, our proposed dataset captures 655 distinct outdoor places of interest with 130 reserved for testing alone. Despite these issues, we have adapted the 7Scenes dataset to our map-free relocalization task.

Regarding relative pose estimation, ScanNet [19] and MegaDepth [39] have become popular test beds, *e.g.*, for learning-based 2D correspondence methods such as SuperGlue [54] and LoFTR [66]. However, both datasets do not feature distinctive mapping and query sequences as basis for a relocalization benchmark. Furthermore, MegaDepth camera poses do not have metric scale. In our experiments, we use ScanNet [19] as a training set for scene-agnostic relocalization methods to be tested on 7Scenes. In the supplemental material, we also provide ablation studies on metric relative pose accuracy on ScanNet.

4.2 Niantic Map-free Relocalization Dataset

We introduce a new dataset for development and evaluation of map-free relocalization. The dataset consists of 655 outdoor scenes, each containing a small 'place of interest' such as a sculpture, sign, mural, etc, such that the place can be well-captured by a single image. Scenes of the dataset are shown in Fig. 4.

The scenes are split into 460 training scenes, 65 validation scenes, and 130 test scenes. Each training scene has two sequences of images, corresponding to two different scans of the scene. We provide the absolute pose of each training image, which allows determining the relative pose between any pair of training images. We also provide overlap scores between any pair of images (intra- and inter-sequence), which can be used to sample training pairs. For validation and test scenes, we provide a single reference image obtained from one scan and a sequence of query images and absolute poses from a different scan. Camera intrinsics are provided for all images in the dataset.

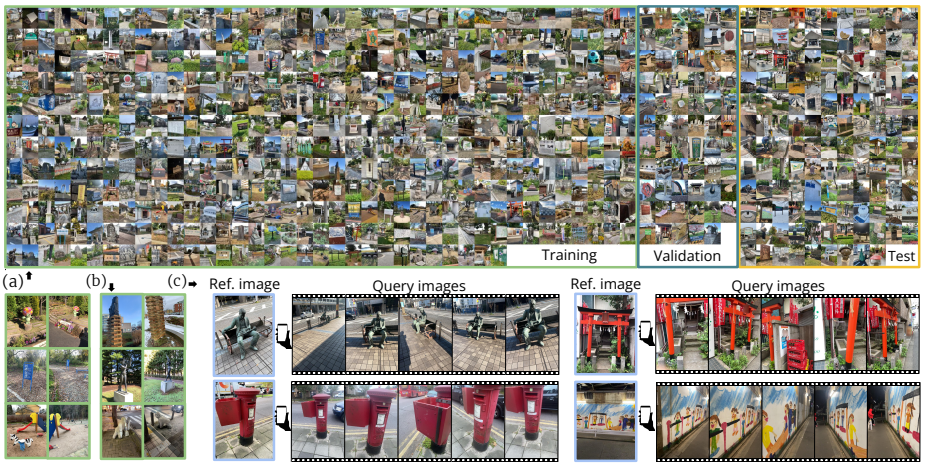

Fig. 4: Niantic map-free relocalisation dataset. (a) Dataset overview. Training (460 scenes), validation (65) and test (130) thumbnails. Better seen in color and magnified in electronic format. (b) Examples of training pairs sampled from training scenes. (c) Reference frame (enclosed in blue) and an example of query images. Query sequences have been sampled at relative temporal frames: 0%, 25%, 50%, 75% and 100% of the sequence duration.

The Niantic map-free dataset was crowdsourced from members of the public who scanned places of interest using their mobile phones. Each scan contains video frames, intrinsics and (metric) poses estimated by ARKit (iOS) [2] or ARCore (Android) [31] frameworks and respective underlying implementations of Visual-Inertial Odometry and Visual-Inertial SLAM. We use automatic anonymization software to detect and blur faces and car license plates in frames. Scans were registered to each other using COLMAP [60]. First, we bundle adjust the scans individually by initializing from raw ARKit/ARCore poses. Then, the two 3D reconstructions are merged into a single reconstruction by matching features between scans, robustly aligning the two scans and bundle adjusting all frames jointly. We then compute a scale factor for each scan, so that the frames of the 3D reconstructions of each scan would (robustly) align to the raw ARKit/ARCore poses. Finally, the 3D reconstruction is rescaled using the average scale factor of the two scans. Further details are provided in supplemental material. Poses obtained via SfM constitute only a *pseudo* ground truth, and estimating their uncertainty bounds has recently been identified as an open problem in relocalization research [9]. However, as we will discuss below, given the challenging nature of map-free relocalization, we evaluate at much coarser error threshold than standard relocalization works. Thus, we expect our results to be less susceptible to inaccuracies in SfM pose optimization.

The places of interest in the Niantic map-free dataset are drawn from a wide variety of locations around the world and captured by a large number of

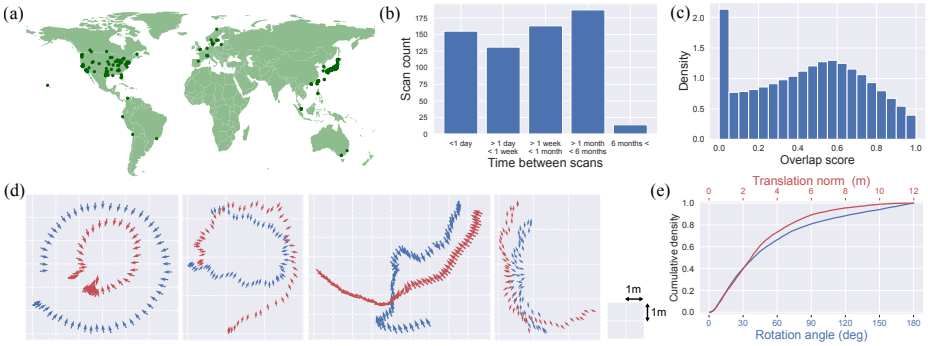

Fig. 5: Niantic map-free dataset statistics. (a) Geographic location of scans. (b) Time elapsed between two different scans from the same scene. (c) Visual overlap between training frames estimated using co-visible SfM points, inspired by [54]. (d) Sample of different dataset trajectories seen from above. Each plot represents one scene and shows two different trajectories corresponding to two different scans, one in each color. The direction of the arrows represent the camera viewing direction. Each trajectory has been subsampled for visualization. (e) Relative pose distribution between reference image and query images in the test set.

people. This leads to a number of interesting challenges, such as variations in the capture time, illumination, weather, season, and cameras, and even the geometry of the scene; and variations in the amount of overlap between the scans. Fig. 5 summarizes these variations.

4.3 Evaluation Protocol

Our evaluation protocol consists of rotation, translation and reprojection errors computed using ground truth and estimated relative poses that are predicted for each query and reference image pair. Given estimated (R, t) and ground truth $(R_{\text{gt}}, t_{\text{gt}})$ poses, we compute the rotation error as the angle (in degrees) between predicted and ground truth rotations, $\measuredangle(R, R_{\text{gt}})$. We measure the translation error as the Euclidean distance between predicted c and ground truth c_{gt} camera centers in world coordinate space, where $c = -R^T t$.

Our proposed reprojection error provides an intuitive measure of AR content misalignment. We were inspired by the Dense Correspondence Reprojection Error (DCRE) [78] which measures the average Euclidean distance between corresponding original pixel positions and reprojected pixel positions obtained via back-projecting depth maps. As our dataset does not contain depth maps we cannot compute the DCRE. Hence, we propose a Virtual Correspondence Reprojection Error (VCRE): ground truth and estimated transformations are used to project virtual 3D points, located in the query camera's local coordinate

system. VCRE is the average Euclidean distance of the reprojection errors:

$$\text{VCRE} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{v} \in \mathcal{V}} \left\| \pi(\mathbf{v}) - \pi(T T_{\text{gt}}^{-1} \mathbf{v}) \right\|_2 \quad \text{with} \quad T = [R|t], \tag{1}$$

where π is the image projection function, and $\mathcal{V}$ is a set of 3D points in camera space representing virtual objects. For convenience of notation, we assume all entities are in homogeneous coordinates. To simulate an arbitrary placement of AR content, we use a 3D grid of points for $\mathcal{V}$ (4 in height, 7 in width, 7 in depth) with equal spacing of 30 cm and with an offset of 1.8m along the camera axis. See supplemental material for a video visualisation, and an ablation showing that DCRE and VCRE are well-aligned. In standard relocalization, best methods achieve a DCRE below a few pixels [9]. However, map-free relocalization is more challenging, relying on learned heuristics to resolve the scale ambiguity. Thus, we apply more generous VCRE thresholds for accepting a pose, namely 5% and 10% of the image diagonal. While a 10% offset means a noticeable displacement of AR content, we argue that it can still yield an acceptable AR experience.

Our evaluation protocol also considers the confidence of pose estimates. Confidence enables the relocalization system to flag and potentially reject unreliable predictions. This is a crucial capability for a map-free relocalization system to be practical since a user might record query images without any visual overlap with the reference frame. A confidence can be estimated as the number of inlier correspondences in feature matching baselines. Given a confidence threshold, we can compute the ratio of query images with confidence greater-or-equal to the threshold, *i.e.*, the ratio of confident estimates or the ratio of non-rejected samples. Similarly, we compute the precision as the ratio of non-rejected query images for which the pose error (translation, rotation) or the reprojection error is acceptable (below a given threshold). Each confidence threshold provides a different trade-off between the number of images with an estimate and their precision. Models that are incapable of estimating a confidence will have a flat precision curve.

5 Experiments

We first report experiments on the 7Scenes [62] dataset, demonstrating that our baselines are competitive with the state of the art when a large number of mapping images is available. We also show that as the number of mapping images reduces, map-free suitable methods degrade more gracefully than traditional approaches. Additional relative pose estimation experiments on ScanNet [19] are reported in the supplement, to allow comparison of our baselines against previous methods. Finally, we report performance on the new Niantic map-free relocalization dataset and identify areas for improvement.

5.1 7Scenes

First, we compare methods described in Sec. 3.1 and Sec. 3.2 against traditional methods when all mapping frames are available. Fig. 6a shows impressive scores

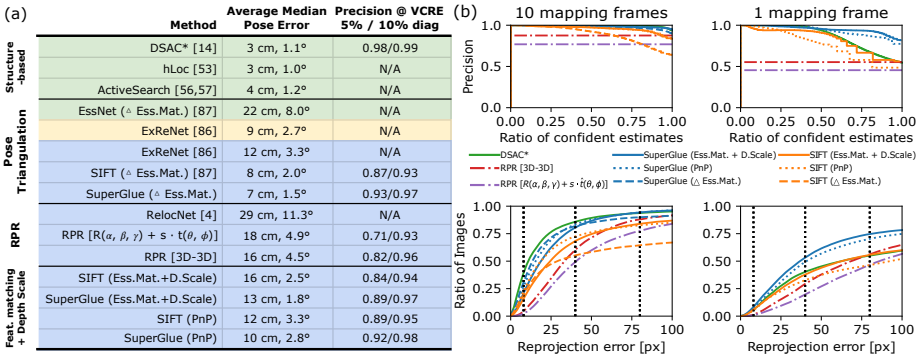

Fig. 6: 7Scenes results. (a) Using all mapping frames. Dataset-specific (7Scenes) methods in green, trained on SUNCG [65] in yellow, and trained on Scannet [19] in blue. (b) 10 and 1 mapping frame scenarios: precision curves (top), cumulative density of reprojection error (bottom). Dashed vertical lines indicate 1%, 5% and 10% of the image diagonal, correspondingly 8px, 40px and 80px.

of structure-based DSAC* [14] (trained with depth from PlaneRCNN [40]), hLoc [53] and ActiveSearch [56,57]. When 5 reference frames can be retrieved for each query using DenseVLAD [72] (following [87]), triangulation-based relative pose methods are competitive with structure-based methods, especially in average median rotation error. See results for EssNet[87], ExReNet[86] and our feature matching and triangulation baselines, denoted by △.

Closer to map-free relocalization, if for each query frame, we retrieve a single reference frame from the set of mapping images, the accuracy of metric relative pose estimation becomes more important, see the sections for relative pose regression (*RPR*) and *Feature Matching+D.Scale* in Fig. 6a. Unsurprisingly, methods in both families slightly degrade in performance, with Feature Matching + D.Scale methods beating RPR methods. However, all baselines remain competitive, despite depth [40] and RPR networks being trained on ScanNet [19] and evaluated on 7Scenes. High scores for all methods in Fig. 6a are partially explainable by the power of image retrieval and good coverage of the scene.

In map-free relocalization, the query and reference images could be far from each other. Thus, we evaluate the baselines on heavily sparsified maps, where metric relative pose accuracy is more important. We find the K most representative reference images of each scene by K-means clustering over DenseVLAD [72] descriptors of the mapping sequence. In Fig. 6b we show results for $K = 10$ and $K = 1$, where $K = 1$ corresponds to map-free relocalization. We show precision curves using a pose acceptance threshold of VCRE $< 10\%$ of the image diagonal (*i.e.*, 80px). We also plot the cumulative density of the VCRE. Unsurprisingly, pose triangulation methods fare well even when $K = 10$ but cannot provide estimates when $K = 1$. For $K = 1$, Feature Matching+D.Scale outperforms the competition. Specifically, SuperGlue (Ess.Mat.+ D.Scale) recovers more than 50% of query images with a reprojection error below 40px.

DSAC* remains competitive in sparse regimes, but it requires training per scene, while the other baselines were trained on ScanNet. Both ScanNet and 7Scenes show very similar indoor scenes. Yet, single-image depth prediction seems to generalize better across datasets compared to RPR methods, as Feature Matching+D.Scale methods outperform RPR baselines both with $K = 10$ and $K = 1$ scenarios. RPR methods perform relatively well for larger accuracy thresholds but they perform poorly in terms of precision curves due to their lack of estimated confidence. Further details on all baselines, qualitative results and additional ablation studies can be found in the supplement.

5.2 Niantic map-free relocalization dataset

Fig. 7 shows our main results on the Niantic map-free dataset. As seen in Fig 7 a, b and c, this dataset is much more challenging than 7Scenes for all methods. This is due to multiple factors: low overlap between query and reference images; the quality of feature matching is affected by variations in lighting, season, etc; and the use of single-image depth prediction networks trained on KITTI for non-driving outdoor images.

In Fig. 7d and 7e we show results of the best methods in each family of baselines: RPR with 6D rotation and scaled translation parameterization and SuperGlue (Ess.Mat.+D.Scale). SuperGlue (Ess.Mat.+D.Scale) in Fig. 7e reports a median angular rotation below 10° for a large number of scenes. In these cases, the high variance of the median translation error is partly due to the variance of depth estimates. Further improvement of depth prediction methods in outdoor scenes should improve the metric accuracy of the translation error. Qualitative examples in Fig 7f shows where depth improvements could produce better results: both the angular pose and the absolute scale in the first row are accurate, while the second row has good angular pose and bad absolute scale.

The RPR method in Fig. 7d exhibits a different behavior: the average angular error is lower than for Feature Matching+D.Scale baselines, yet it rarely achieves high accuracy. This is also evident in Fig. 7 c, where Feature Matching+D.Scale methods outperform RPR methods for stricter thresholds, but degrade for broader thresholds. Indeed, when the geometric optimization fails due to poor feature matches, the estimated scaleless pose can be arbitrarily far from the ground truth. In contrast, RPR methods fail more gracefully due to adhering to the learned distribution of relative poses. For example, in Fig. 7c allowing for a coarser VCRE threshold of 10% of the image diagonal, the $[3D - 3D]$ and $[R(6D) + t]$ variants overtake all methods, including feature matching-based methods. Hence, RPR methods can be more accurate than feature matching at broad thresholds, but they offer lower precision in VCRE at practical thresholds.

RPR methods currently do not predict a confidence which prevents detecting spurious pose estimates, *e.g.*, when there is no visual overlap between images, as illustrated in the supplement. Although feature matching methods can estimate the confidence based on the number of inliers, the precision curves in Fig. 7a show that these confidences are not always reliable. Further research in modeling confidence of both families of methods could allow to combine their advantages.

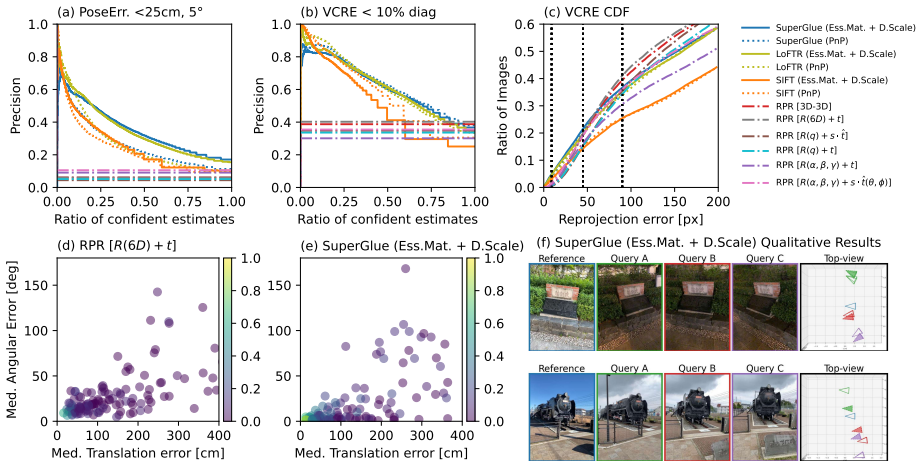

Fig. 7: Our dataset results. (a,b) Precision plots using pose error (a) and VCRE (b) thresholds. (c) VCRE CDF, vertical lines indicate 1, 5 and 10% of the image diagonal, corresp. 9px, 45px and 90px.(d,e) Scatter plot of median angular vs translation error for each scene, estimated using RPR $[R(6D)+t]$ (d) and SG [54] Ess.Mat.+D.scale (e). Each point represents a scene, and the colormap shows precision for pose error threshold 25cm, 5°. (f) Qualitative results: the reference frame and three queries are shown for two scenes. The top view shows the ground truth (solid line, no fill) and estimated poses (dashed line, filled).

6 Conclusion and Future Work

We have proposed map-free relocalization, a new relocalization task. Through extensive experiments we demonstrate how existing methods for single-image depth prediction and relative pose regression can be used to address the task with some success. Our results suggest some directions for future research: improve the scale estimates by improving depth estimation in outdoor scenes; improve the accuracy of metric RPR methods; and derive a confidence for their estimates.

To facilitate further research, we have presented the Niantic map-free relocalization dataset and benchmark with a large number of diverse places of interest. We define an evaluation protocol to closely match AR use cases, and make the dataset and an evaluation service publicly available.

As methods for this task improve, we hope to evaluate at stricter pose error thresholds corresponding to visually more pleasing results. A version of map-free relocalization could use a burst of query frames rather than a single query frame to match some practical scenarios. Our dataset is already suitable for this variant of the task, so we hope to explore baselines for it in the future.

Acknowledgements We thank Pooja Srinivas, Amy Duxbury, Camille François, Brian McClendon and their teams, for help with validating and anonymizing the dataset and Galen Han for his help in building the bundle adjustment pipeline. We also thank players of Niantic games for scanning places of interests.

References

1. Abouelnaga, Y., Bui, M., Ilic, S.: DistillPose: Lightweight camera localization using auxiliary learning. In: IROS (2021) 4
2. Apple: ARKit, https://developer.apple.com/documentation/arkit/configuration_objects/understanding_world_tracking, Accessed: 6 March 2022 9
3. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: CVPR (2016) 4
4. Balntas, V., Li, S., Prisacariu, V.: RelocNet: Continuous metric learning relocalisation using neural nets. In: ECCV (2018) 2, 4
5. Barath, D., Matas, J., Noskova, J.: MAGSAC: marginalizing sample consensus. In: CVPR (2019) 4
6. Barath, D., Noskova, J., Ivashechkin, M., Matas, J.: MAGSAC++, a fast, reliable and accurate robust estimator. In: CVPR (2020) 4, 6
7. Besl, P., McKay, N.D.: A method for registration of 3-D shapes. IEEE TPAMI (1992) 6
8. Bhowmik, A., Gumhold, S., Rother, C., Brachmann, E.: Reinforced feature points: Optimizing feature detection and description for a high-level task. In: CVPR (2020) 4
9. Brachmann, E., Humenberger, M., Rother, C., Sattler, T.: On the limits of pseudo ground truth in visual camera re-localisation. In: ICCV (2021) 2, 9, 11
10. Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C.: DSAC-differentiable ransac for camera localization. In: CVPR (2017) 1, 4
11. Brachmann, E., Rother, C.: Learning Less is More - 6D Camera Localization via 3D Surface Regression. In: CVPR (2018) 4
12. Brachmann, E., Rother, C.: Expert sample consensus applied to camera re-localization. In: ICCV (2019) 4
13. Brachmann, E., Rother, C.: Neural-guided RANSAC: Learning where to sample model hypotheses. In: ICCV (2019) 4
14. Brachmann, E., Rother, C.: Visual camera re-localization from RGB and RGB-D images using DSAC. IEEE TPAMI (2021) 4, 12
15. Brahmbhatt, S., Gu, J., Kim, K., Hays, J., Kautz, J.: Geometry-aware learning of maps for camera localization. In: CVPR (2018) 4
16. Cai, R., Hariharan, B., Snavely, N., Averbuch-Elor, H.: Extreme rotation estimation using dense correlation volumes. In: CVPR (2021) 4, 7
17. Campos, C., Tardós, J.D.: Scale-aware direct monocular odometry. In: ICUS (2021) 5
18. Chen, K., Snavely, N., Makadia, A.: Wide-baseline relative camera pose estimation with directional learning. In: CVPR (2021) 4
19. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: CVPR (2017) 5, 6, 8, 11, 12
20. Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C.: Bundlefusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. ACM TOG (2017) 2
21. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable CNN for joint detection and description of local features. In: CVPR (2019) 4

22. Eggert, D.W., Lorusso, A., Fisher, R.B.: Estimating 3-D rigid body transformations: A comparison of four major algorithms. Mach. Vision Appl. (1997) 6
23. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NeurIPS (2014) 5
24. En, S., Lechervy, A., Jurie, F.: RPNet: An end-to-end network for relative camera pose estimation. In: ECCVW (2018) 4, 7
25. Facil, J.M., Ummenhofer, B., Zhou, H., Montesano, L., Brox, T., Civera, J.: CAM-Convs: Camera-aware multi-scale convolutions for single-view depth. In: CVPR (2019) 5
26. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM (1981) 4, 6
27. Gao, X.S., Hou, X.R., Tang, J., Cheng, H.F.: Complete solution classification for the perspective-three-point problem. IEEE TPAMI (2003) 6
28. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: CVPR (2012) 5, 6
29. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017) 5
30. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth prediction. In: ICCV (2019) 5
31. Google: ARCore, https://developers.google.com/ar/develop/fundamentals, Accessed: 6 March 2022 9
32. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003) 3, 4, 5
33. 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 (2020) 1, 2, 4
34. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., et al.: Kinectfusion: real-time 3D reconstruction and interaction using a moving depth camera. In: ACM UIST (2011) 8
35. Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M., Trulls, E.: Image Matching across Wide Baselines: From Paper to Practice. IJCV (2020) 4
36. Kendall, A., Cipolla, R.: Geometric loss functions for camera pose regression with deep learning. CVPR (2017) 4
37. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-DOF camera relocalization. In: CVPR (2015) 4, 8
38. Li, X., Wang, S., Zhao, Y., Verbeek, J., Kannala, J.: Hierarchical scene coordinate classification and regression for visual localization. In: CVPR (2020) 4
39. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: CVPR (2018) 5, 8
40. Liu, C., Kim, K., Gu, J., Furukawa, Y., Kautz, J.: PlaneRCNN: 3D plane detection and reconstruction from a single image. In: CVPR (2019) 3, 5, 6, 12
41. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV (2004) 4, 5
42. Melekhov, I., Ylioinas, J., Kannala, J., Rahtu, E.: Relative camera pose estimation using convolutional neural networks. In: International Conference on Advanced Concepts for Intelligent Vision Systems (2017) 2, 4, 7
43. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A.: KinectFusion: Real-time dense surface mapping and tracking. In: ISMAR (2011) 2

44. Nister, D.: An efficient solution to the five-point relative pose problem. IEEE TPAMI (2004) 4, 6
45. Pion, N., Humenberger, M., Csurka, G., Cabon, Y., Sattler, T.: Benchmarking image retrieval for visual localization. In: 3DV (2020) 4
46. Radenović, F., Tolias, G., Chum, O.: CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In: ECCV (2016) 4
47. Raguram, R., Frahm, J.M., Pollefeys, M.: A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus. In: ECCV (2008) 4
48. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021) 2, 3, 5, 6
49. Ranftl, R., Koltun, V.: Deep fundamental matrix estimation. In: ECCV (2018) 4
50. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE TPAMI (2020) 5
51. Rau, A., Garcia-Hernando, G., Stoyanov, D., Brostow, G.J., Turmukhambetov, D.: Predicting visual overlap of images through interpretable non-metric box embeddings. In: ECCV (2020) 4
52. Revaud, J., Weinzaepfel, P., de Souza, C.R., Humenberger, M.: R2D2: repeatable and reliable detector and descriptor. In: NeurIPS (2019) 4
53. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: CVPR (2019) 1, 2, 12
54. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020) 1, 2, 3, 4, 5, 6, 8, 10, 14
55. Sattler, T., Leibe, B., Kobbelt, L.: Fast image-based localization using direct 2D-to-3D matching. In: ICCV (2011) 1, 2
56. Sattler, T., Leibe, B., Kobbelt, L.: Improving image-based localization by active correspondence search. In: ECCV (2012) 12
57. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. IEEE TPAMI (2017) 2, 12
58. 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 outdoor visual localization in changing conditions. In: CVPR (2018) 1, 2, 8
59. Sattler, T., Weyand, T., Leibe, B., Kobbelt, L.: Image retrieval for image-based localization revisited. In: BMVC (2012) 2, 4
60. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016) 2, 9
61. Shavit, Y., Ferens, R., Keller, Y.: Learning multi-scene absolute pose regression with transformers. In: ICCV (2021) 4
62. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene coordinate regression forests for camera relocalization in RGB-D images. In: CVPR (2013) 4, 8, 11
63. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: ECCV (2012) 5, 6
64. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3D. In: ACM SIGGRAPH (2006) 2
65. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. CVPR (2017) 12
66. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. In: CVPR (2021) 3, 4, 6, 8

67. Sun, W., Jiang, W., Trulls, E., Tagliasacchi, A., Yi, K.M.: ACNe: Attentive context normalization for robust permutation-equivariant learning. In: CVPR (2020) 4
68. Suwajanakorn, S., Snavely, N., Tompson, J., Norouzi, M.: Discovery of latent 3D keypoints via end-to-end geometric reasoning. NeurIPS (2018) 7
69. Tateno, K., Tombari, F., Laina, I., Navab, N.: CNN-SLAM: Real-time dense monocular slam with learned depth prediction. In: CVPR (2017) 5
70. Tiwari, L., Ji, P., Tran, Q.H., Zhuang, B., Anand, S., Chandraker, M.: Pseudo rgb-d for self-improving monocular slam and depth prediction. In: ECCV (2020) 5
71. Toft, C., Turmukhambetov, D., Sattler, T., Kahl, F., Brostow, G.J.: Single-image depth prediction makes feature matching easier. In: ECCV (2020) 5
72. Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: CVPR (2015) 4, 12
73. Türkoğlu, M.Ö., Brachmann, E., Schindler, K., Brostow, G., Monszpart, A.: Visual Camera Re-Localization Using Graph Neural Networks and Relative Pose Supervision. In: 3DV (2021) 4
74. Tyszkiewicz, M., Fua, P., Trulls, E.: Disk: Learning local features with policy gradient. In: NeurIPS (2020) 4
75. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: Demon: Depth and motion network for learning monocular stereo. In: CVPR (2017) 4
76. Valentin, J., Dai, A., Nießner, M., Kohli, P., Torr, P., Izadi, S., Keskin, C.: Learning to navigate the energy landscape. In: 3DV (2016) 8
77. Walch, F., Hazirbas, C., Leal-Taixé, L., Sattler, T., Hilsenbeck, S., Cremers, D.: Image-based localization using LSTMs for structured feature correlation. In: ICCV (2017) 4
78. Wald, J., Sattler, T., Golodetz, S., Cavallari, T., Tombari, F.: Beyond controlled environments: 3D camera re-localization in changing indoor scenes. In: ECCV (2020) 8, 10
79. Watson, J., Firman, M., Brostow, G.J., Turmukhambetov, D.: Self-supervised monocular depth hints. In: ICCV (2019) 5
80. Winkelbauer, D., Denninger, M., Triebel, R.: Learning to localize in new environments from synthetic training data. In: ICRA (2021) 4, 7
81. Wu, C.: VisualSFM: A visual structure from motion system (2011), http://ccwu.me/vsfm/ 2
82. Wu, C.: Towards linear-time incremental structure from motion. In: 3DV (2013) 2
83. Xian, K., Shen, C., Cao, Z., Lu, H., Xiao, Y., Li, R., Luo, Z.: Monocular relative depth perception with web stereo data supervision. In: CVPR (2018) 5
84. Yang, N., Wang, R., Stuckler, J., Cremers, D.: Deep Virtual Stereo Odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In: ECCV (2018) 5
85. Yi, K.M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P.: Learning to find good correspondences. In: CVPR (2018) 4
86. Zhang, W., Kosecka, J.: Image based localization in urban environments. In: 3DV (2006) 4, 12
87. Zhou, Q., Sattler, T., Pollefeys, M., Leal-Taixe, L.: To learn or not to learn: Visual localization from essential matrices. In: ICRA (2020) 4, 7, 12
88. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017) 5
89. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR (2019) 7