Feature Query Networks: Neural Surface Description for Camera Pose Refinement

Hugo Germain[1]*, Daniel DeTone[2], Geoffrey Pascoe[2], Tanner Schmidt[2], David Novotny[3],
Richard Newcombe[2], Chris Sweeney[2], Richard Szeliski[4], Vasileios Balntas[2]
[1]Ecole des Ponts ParisTech [2]Reality Labs, Meta
[3]Facebook AI Research [4]The University of Washington

Abstract

Accurate 6-DoF camera pose estimation in known environments can be a very challenging task, especially when the query image was captured at viewpoints strongly differing from the set of reference camera poses. While structure-based methods have proved to deliver accurate camera pose estimates, they rely on pre-computed 3D descriptors coming from reference images often misaligned with query images. This discrepancy can subsequently harm downstream camera pose estimation tasks. In this paper we introduce the Feature Query Network (FQN), a ray-based descriptor regressor that can be used to query descriptors at known 3D locations under novel viewpoints. We show that the FQN is able to model viewpoint-dependency of high-dimensional keypoint descriptors and bring significant relative improvements to structure-based visual localization baselines.

1. Introduction

Learning robust and invariant keypoint descriptors is an underpinning component to many computer vision applications, such as Structure-from-Motion (SfM) [26, 60, 62, 71] and visual localization [57, 69, 70]. These applications are in turn crucial backbones to Augmented and Virtual Reality or autonomous driving scenarios.

In the case of visual localization where a 3D model of the scene is available, keypoint descriptors can be used to perform camera pose estimation through direct 2D-to-3D matching followed by a Perspective-*n*-Pose (PnP) solver [10, 25, 31, 34] inside a RANdom SAmple Consensus approach (RANSAC) [20, 56] loop. When dense query feature maps are available, one can also either solve or refine the camera pose estimate using direct alignment methods [54, 68, 77]. Both approaches belong to *structure-based* localization methods which fully exploit the available 3D geometry, resulting in both lightweight and accurate relocalization pipelines [52, 55].

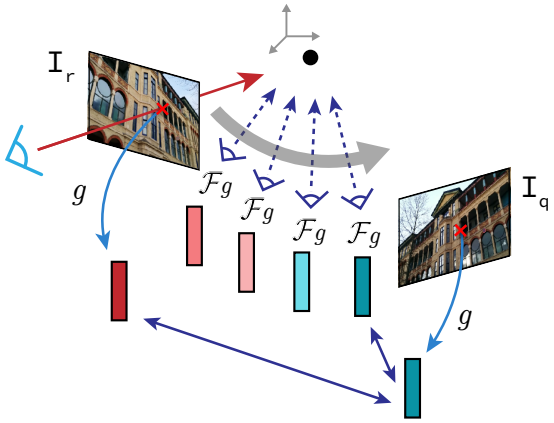

Figure 1. **Bridging viewpoint discrepancies using FQNs:** When performing structure-based localization, viewpoint discrepancies often occur between the reference image $\mathtt{I}_r$ and the query image $\mathtt{I}_q$, inducing increased descriptor distances for the same 3D point. In this paper, we introduce the Feature Query Network (FQN) $\mathcal{F}_g$, a ray-based MLP trained to model the variance of an arbitrary feature descriptor g at known 3D locations with respect to the camera pose. We demonstrate how this simple model can be used to improve structure-based localization by leveraging an image-free, continuous geometric modeling of descriptor space.

In structure-based localization however, the 3D descriptors come from reference images that are often geometrically misaligned with respect to the query image. When performing 2D-to-3D keypoint matching or feature-metric error minimization on query images captured under unseen viewpoints, a discrepancy naturally emerges between the pre-computed 3D descriptors and the 2D query features.

Alternative approaches to visual localization propose to perform end-to-end training of convolutional models to directly regress camera poses (in an absolute [28, 29, 78] or relative [1, 16, 33, 82] way) or scene coordinates [5–7, 9, 11, 12, 36, 79]. Such methods however fail to match

*Work done during an internship at Reality Labs.

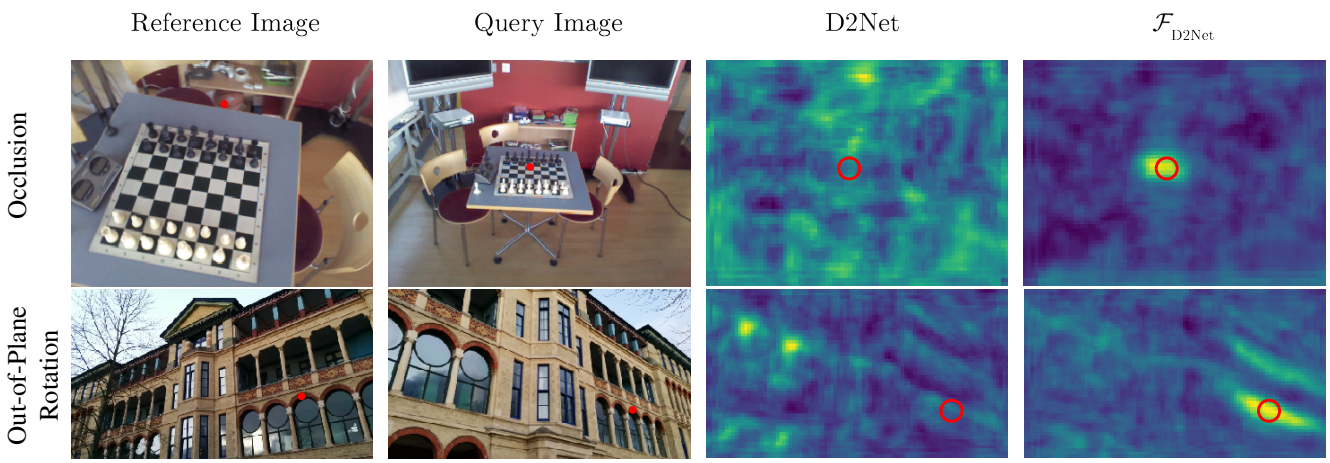

Figure 2. **Feature Query Networks for Neural Surface Description:** We report qualitative examples on scenes from [29, 63], obtained using FQN-regressed descriptors at unseen camera poses. We show in red the ground-truth reprojection of a given 3D keypoint in the scene and (from left to right) a pair of reference and query images, as well as dense correspondence maps obtained using image-based and FQN-regressed D2Net [17] descriptors (computed at the ground truth query camera pose). We find that our approach is able to produce much more accurate correspondence maps, and to bridge the viewpoint discrepancy between the reference and query image. This includes dealing with occlusions (upper row) and strong out-of-plane rotations (lower row).

the accuracy of structure-based localization [54], have little generalization abilities [58], and share the burden of requiring to retrain fairly heavy models on every scene, which can additionally be sometimes unstable [55, 61]. In comparison, our proposed approach is scene-specific but only requires a very lightweight model (about 10 to 50 times fewer parameters) with a very straightforward training procedure, and is compatible with structure-based approaches.

While recent advances in learning-based feature descriptors [2, 15, 17, 46, 49, 64, 65, 80], have strongly improved their robustness to illumination and viewpoint changes, regressing perfectly viewpoint-invariant keypoint descriptors remains an open problem. Instead of aiming to reach viewpoint invariance, we propose to instead learn to explicitly model the variance of existing keypoint descriptors w.r.t. geometry on specific scenes.

In this paper we introduce the Feature Query Network (FQN), a simple ray-based multilayer perceptron (MLP) designed to capture arbitrary descriptor variance with respect to the camera viewpoint on known geometry. Previous works on implicit neural representations [21, 27, 41, 47] have shown the outstanding ability of MLPs to represent various scene attributes from few posed RGB images, some leveraging the available 3D geometry. iNeRF [38] proposed to invert the NeRF [42] model to perform camera pose estimation from query images through photometric error minimization. In comparison, the FQN operates at a surface-level of a known 3D reconstruction and trades the need to estimate the scene geometry with the complexity of modeling high-dimensional scene descriptors.

As shown in Fig.1, the FQN can be used to dynamically update 3D descriptors at given camera pose estimates in structure-based visual localization methods. We experimentally demonstrate that this view-dependent regression ability can bring a significant relative improvement in camera pose estimation accuracy. While our approach is scene-specific by design, it is significantly more lightweight and straightforward to train than the aforementioned scene-specific visual localization methods. Our contributions are as follows:

- We introduce Feature Query Networks (FQNs), powerful neural networks trained to model the variance of a given descriptor w.r.t. the camera viewpoint on a given scene.
- We show how they can be parametrized to perform view-dependent descriptor regression in a known environment. We study the ability of FQNs to model learning-based descriptors, including a 512-dimensional one from [17].
- We demonstrate how the FQNs can be applied to structure-based camera pose estimation methods, and easily yield incremental improvements in accuracy.

2. Related Work

In this section we review existing approaches related to visual localization, as well as recent advances in implicit representation learning.

Structure-based visual localization. Assuming a known ground truth 3D model of the world and a set of posed reference images, structure-based localization leverages the available geometry to obtain accurate camera pose estimates on unseen query images. Direct 2D-to-3D matching methods try to identify explicit correspondences between the query image and the 3D model. These putative 2D-to-3D correspondences can subsequently be fed to a P*n*P solver [10, 25, 31, 34] inside a RANSAC loop [20, 56]. The recent progress in learning-based keypoint detection [32, 49, 59, 75, 76], description [2, 15, 17, 22, 23, 43, 46, 49, 64, 65, 73, 75, 80], matching [44, 50, 53, 81] and outlier rejection methods [3, 4, 8, 13] has improved structure-based localization performance significantly. Coupled with a hierarchical retrieval-based framework [52, 53], this approach has become competitive for large-scale image-based relocalization [52, 53, 55]. One can avoid the problem of explicit 2D-to-3D keypoint matching using Direct Alignment methods instead [14, 18, 19, 24, 54, 68, 77], which directly optimize for the camera pose through minimization of a photometric or learning-based cost function. The burden of structure-based localization however lies in the discrepancy between the 3D keypoint descriptors (assigned using reference images), and the 2D descriptors coming from the query image. Due to the frequent geometric and appearance domain gap between the reference and test set, the lack of viewpoint and illumination invariance of state-of-the-art feature descriptors introduces errors in both explicit 2D-to-3D matching and feature-metric camera pose optimization. The work of [74] proposes to pre-rectify images using piece-wise homographies adaptation, by assuming scene-planarity. We argue that this approach is limited in its warping fidelity and projects images to a canonical space which does not ensure alignment of descriptors. The purpose of this paper is instead to bridge the geometric domain gap by explicitly modeling the viewpoint-conditioned surface-level descriptor representation of the world.

Structure-free visual localization. A parallel line of work to perform visual localization ignores the underlying 3D geometry to rather focus on performing end-to-end learning-based camera pose estimation. This can be done using absolute [28, 29, 78] or relative [1, 16, 33, 82] camera pose regression, or by regressing scene coordinates [5–7, 9, 11, 12, 36, 79]. These methods however come with their set of drawbacks, namely the lack of generalization to novel viewpoints [58], training instabilities [55, 61], cumbersome per-scene retraining and set of weights, and overall reduced accuracy [54].

Implicit neural representations. Recent advances in implicit representation learning has demonstrated the power of coordinate-based multilayer perceptrons (MLPs) to map known world coordinates to signed distance fields [27, 47], occupancy grids [21, 41] or RGB values [72]. Other approaches relax the need for ground truth 3D geometry and apply differentiable rendering to perform novel-view synthesis from a set of posed images [42, 45, 67]. Closer to our work is iNeRF [38], which geometrically minimizes the photometric error using NeRF [42]-generated 3D reconstructions. In this paper however, we focus on leveraging available 3D geometry which yields considerably faster regression times, and directly operate in higher dimensional descriptor-space as opposed to RGB-space, which is a lot more robust to illumination changes.

3. Feature Query Networks

In this section, we introduce Feature Query Networks, simple ray-based MLPs that can be used to regress descriptors at known 3D locations for a given viewpoint. We propose a simple parametrization and show how they can be trained to learn to model descriptors at a surface-level.

3.1. Formalism

Given a set of M reference images $\{\mathtt{I}_j\}_{j=1}^{M}$ with camera poses $\{\mathtt{T}_j\}_{j=1}^{M}$, we consider a sparse 3D model of the world $\mathcal{M} = \{\mathbf{u}_i\}_{i=1}^{N}$ (*e.g.* built using SfM) where $\mathbf{u}_i \in \mathbb{R}^3$, $\mathtt{I}_i \in [0,1]^{3\times H\times W}$ and $\mathtt{T} \in \mathrm{SE}(3)$. Following [42], given a keypoint $\mathbf{u}_i$ and a camera pose $\mathtt{T}_j$ we define $(\theta,\phi)_i$ as the 2-dimensional viewing direction between $\mathbf{u}_i$ and the camera center of $\mathtt{T}_j$. In addition, let $l_{i,j}$ be the corresponding ray length, f_j the focal length of the camera and r_j the camera roll angle. For an arbitrary keypoint descriptor $g : \mathbb{R}^{3\times H\times W} \times \mathbb{R}^2 \to \mathbb{R}^D$, we model the descriptor-based representation of the world using the 8-dimensional vector-valued function:

$$\mathcal{F}_g(\mathbf{u},\theta,\phi,r,f,l) = \hat{\mathbf{d}} \in \mathbb{R}^D. \tag{1}$$

We refer to the Θ-parametrized MLP approximating $\mathcal{F}_g$ as the Feature Query Network, or FQN for short. An overview of the FQN parametrization can be found in Fig. 3.

3.2. Optimizing a Neural Descriptor Surface

Let $\mathbf{p}_i^j \in \mathbb{R}^2$ be the projection of $\mathbf{u}_i$ in the image plane of $\mathtt{I}_j$ defined by $\mathbf{p}_i^j = \omega(\mathbf{u}_i, \mathtt{T}_j, \mathtt{K}_j)$.[1] Writing the 3D descriptor $\mathbf{d}_i^j$ of $\mathbf{u}_i$ seen in $\mathtt{I}_j$ as $\mathbf{d}_i^j = g(\mathtt{I}_j, \mathbf{p}_i^j)$, we can train a Feature Query Network by minimizing the following loss function w.r.t. Θ:

[1] With a slight abuse of notation, we write $\mathbf{p}^j \in \mathbb{R}^{N\times 2}$ the stacked reprojected keypoints $\{\mathbf{p}_i^j\}_{i=1}^{N}$.

$$\mathcal{L}_g = \sum_{j=1}^{M} \sum_{i=1}^{N} \left\| \mathcal{F}_{g,\Theta}(\mathbf{u}_i, \theta_{i,j}, \phi_{i,j}, r_j, f_j, l_{i,j}) - \mathbf{d}_i^j \right\|_2^2 . \tag{2}$$

Compared to [38, 42], our FQN leverages the available 3D geometry to only regress descriptors at a surface level. Thus, it avoids the density estimation problem, which comes with its intricacies and a heavy computational cost. However, the FQN needs to properly model a descriptor-based representation of the world in a space of much higher dimension than the typical RGB space (in the case of NeRF [42]). As shown in our experiments (see Sec. 5.2), we find that our 8-dimensional parametrization of the FQN inputs is sufficient to encode the variance of various descriptors w.r.t. to scale, rotation, and translation. Notably we find the additional inputs l, f and r to be critical to properly encode the descriptor variance w.r.t. depth, 2D-scale and in-plane rotations respectively.

We report in Fig. 2 qualitative results obtained with FQNs trained on D2Net [17] descriptors. In particular, we report dense correspondence maps computed as per [23], and find that when provided with ground truth camera poses FQN effectively bridges the viewpoint gap between the reference and query image set. More details are provided in Sec. 5.

3.3. Implementation Details

Our implementation of the FQN architecture closely follows [42]. We use an MLP architecture consisting of 8 fully-connected layers (with ReLU activations, 256 channels per layer). The final layer of the model outputs D channels, depending on the choice of g. We employ a 3-dimensional unit-norm viewing direction to encode the viewing direction, and apply Fourier Basis functions [42, 72] for positional encoding on every input to the FQN independently. For stability the ray length l is first transformed using $l' = 1/l^2$, and we apply L2-normalization on both $\mathbf{d}$ and $\mathcal{F}_{g,\Theta}$ in Eq. 2. Please refer to Sec. 5 for additional details regarding our training procedure.

4. FQNs for Visual Localization

In this section, we will show how the Feature Query Networks can be leveraged to improve upon existing structure-based visual localization methods.

4.1. Application to 2D-to-3D Matching

We first present how the FQNs can be used in direct 2D-to-3D matching methods for visual localization.

Reprojection Error Minimization. A popular approach to structure-based localization consists in establishing putative 2D-to-3D correspondences between the query image and the 3D model of the scene. Considering a subset of the 3D model $\mathcal{N} \subset \mathcal{M}$, let us write $(\mathbf{q}, \mathbf{v}) \in \mathbb{R}^2 \times \mathcal{N}$ a single 2D-to-3D correspondence. For a given query image $\mathtt{I}_q$, $\mathcal{N}$ can be defined as the set of 3D keypoints visible in a nearest-neighbour image (*e.g.* obtained using image retrieval as [52]). Recovering 2D-to-3D correspondences is then often done using detection and description in $\mathtt{I}_q$, followed by direct 2D-to-3D matching using the offline-computed 3D descriptors.

To estimate the query camera pose $\mathtt{T}_q = (\mathtt{R}_q|\mathbf{t}_q) \in \mathrm{SE}(3)$, one can then minimize the 2D reprojection error such that:

$$\hat{\mathtt{R}_q}, \hat{\mathbf{t}_q} = \underset{\mathtt{T}=(\mathtt{R}|\mathbf{t})}{\arg\min} \sum_{i=1}^{N} \left\| \mathbf{q}_i - \omega\left(\mathbf{v}_i, \mathtt{T}, \mathtt{K}_q\right) \right\|_2^2 . \tag{3}$$

While solving this minimization is well studied [10, 25, 31, 34], the heart of the problem lies in finding accurate 2D-to-3D correspondences. This becomes particularly difficult when the query image viewpoint differs strongly from the set of reference images. Indeed, the viewpoint discrepancy in $\mathtt{I}_q$ has direct repercussions on the extracted 2D descriptors subsequently used in the keypoint matching stage. We argue that Feature Query Networks can be used in this case to help bridge this gap.

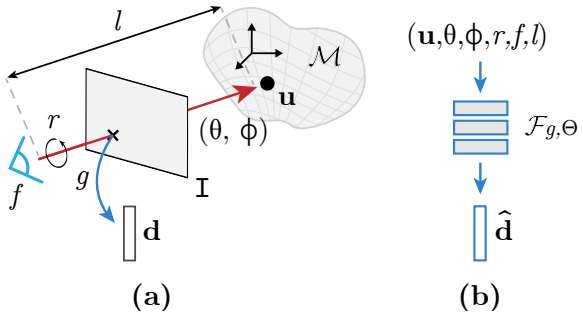

Figure 3. **Feature Query Network parametrization:** (a) Given a posed reference image $\mathtt{I}$, a pre-computed ground truth scene geometry $\mathcal{M}$ and an arbitrary feature descriptor g, we train (b) a simple MLP $\mathcal{F}_{g,\Theta}$ to optimize an 8-dimensional neural surface-level descriptor representation. We denote the focal length of the camera by f, the camera roll by r, the 2D viewing direction by (θ, ϕ) and the D-dimensional descriptor by $\mathbf{d}$.

FQN-based Iterative P*n*P Solving. In order to leverage the view-dependent regression power of FQNs, we propose the following simple iterative algorithm for P*n*P-based camera pose estimation. We first perform direct 2D-to-3D matching following the aforementioned steps, using image-based 3D descriptors (*i.e.* interpolated at their 2D reprojections in reference images). We write $\mathtt{T}_{q,0}$ this initially estimated query

Algorithm 1 FQN-based Iterative P*n*P+RANSAC

Given the descriptor g, keypoint matcher m, FQN $\mathcal{F}_{g,\Theta}$
procedure FQNITERATIVEPNP($\mathtt{I}_q$, $\mathtt{I}_j$, $\mathtt{T}_j$, $\mathtt{K}_q$, $\mathtt{K}_j$, $\mathcal{N}$)
  $(\mathbf{p}^q, \mathbf{d}^q) \leftarrow$ detectAndDescribe($\mathtt{I}_q, g$) ▷ Perform detection and description in $\mathtt{I}_q$
  $\mathbf{d}_i^j \leftarrow g(\mathtt{I}_j, \omega(\mathbf{v}_i, \mathtt{T}_j, \mathtt{K}_j))\ \forall\ \mathbf{v}_i \in \mathcal{N}$ ▷ Compute Image-based 3D Descriptors
  $\mathtt{T}_{q,0} \leftarrow$ PnPRANSAC($\mathbf{p}_q$, $\mathbf{v}$, $\mathtt{K}_q$, $m(\mathbf{d}^q, \mathbf{d}^j)$) ▷ 2D-to-3D Matching + PnP+RANSAC
  for $k = 0$ to $K - 1$ **do**
    $\hat{\mathbf{d}}_i^k \leftarrow \mathcal{F}_{g,\Theta}(\mathbf{v}_i, \theta_{i,k}, \phi_{i,k}, r_k, f_k, l_{i,k})\ \forall\ \mathbf{v}_i \in \mathcal{N}$ ▷ Update descriptors using $\mathcal{F}_{g,\Theta}$
    $(\mathtt{T}_{q,k+1}, n_{inliers}[k+1]) \leftarrow$ PnPRANSAC($\mathbf{p}_q$, $\mathbf{v}$, $\mathtt{K}_q$, $m(\mathbf{d}^q, \hat{\mathbf{d}}^k)$) ▷ 2D-to-3D Matching + PnP+RANSAC
  return $\mathtt{T}_{q,argmax(n_{inliers})+1}$

pose. Then, for K iterations, we (i) recompute the 3D descriptors at the previously estimated camera pose $\mathtt{T}_k$ using the FQN, (ii) perform direct 2D-to-3D matching against the query image and, lastly, (iii) solve for Eq. 3 using the updated set of keypoint correspondences. From this set of K predicted camera poses, we propose to select the best refinement prediction based on the highest number of RANSAC inliers. We describe an overview of our algorithm in Alg. 1.

4.2. Application to Direct Alignment

We now present a second applications of FQNs, to improve the performance of direct alignment methods.

Direct Alignment. Considering a calibrated query image $\mathtt{I}_q$ (unseen at training time), we aim at estimating its unknown camera pose $\mathtt{T}_q = (\mathtt{R}_q|\mathbf{t}_q) \in \mathrm{SE}(3)$. The commonly used approach to solve for the camera pose using direct alignment consists in minimizing the sum of feature-metric errors between the query and a set of partially covisible 3D keypoints such that:

$$\hat{\mathtt{R}_q}, \hat{\mathbf{t}_q} = \underset{\mathtt{T}=(\mathtt{R}|\mathbf{t})}{\arg\min} \sum_{j=1}^{M} \sum_{i=1}^{N} \rho(\left\| \mathbf{r}_i^j \right\|_2^2), \tag{4}$$

where ρ is a robust parametric kernel, and $\mathbf{r}_i$ the vector of residuals of the i-th keypoint visible in I_j defined by:

$$\mathbf{r}_i^j = g(\mathtt{I}_q, \mathbf{p}_i) - \mathbf{d}_i^j \in \mathbb{R}^D. \tag{5}$$

Note that g here can simply return the interpolated RGB value at $\mathbf{p}_i$ which results in a simple photometric error minimization, however learning-based features have proved to deliver much more robust and accurate results [54, 68, 77].

Starting from an initial estimate $\mathtt{T}_0$, this nonlinear least-squares cost function is usually iteratively minimized using the Levenberg-Marquardt (LM) algorithm [35, 40]. Let the residual vector at the k-th iteration of the LM algorithm be $\mathbf{r}_i^{j,k} = g(\mathtt{I}_q, \omega(\mathbf{u}_i, \mathtt{T}, \mathtt{K}_q)) - \mathbf{d}_i^j$. We parametrize the camera pose increments in $\mathrm{se}(3)$, resulting in a 6-dimensional update pose vector $\boldsymbol{\delta} \in \mathbb{R}^6$. We write $\mathtt{J} \in \mathbb{R}^{N \times 6}$ the Jacobian of the residual vectors w.r.t. the pose, $\mathtt{W}$ the diagonal matrix of the robust kernel derivatives, and lastly the Hessian $\mathtt{H} = \mathtt{J}^T\mathtt{W}\mathtt{J}$. At every iteration we compute all the residual vectors $\mathbf{r}$, $\mathtt{J}$, $\mathtt{H}$, $\mathtt{W}$ and solve for the camera pose update $\boldsymbol{\delta} \in \mathbb{R}^6$ using:

$$\boldsymbol{\delta} = -(\mathtt{H} + \lambda \mathrm{diag}(\mathtt{H}))^{-1} \mathtt{J}^T \mathtt{W} \mathbf{r}. \tag{6}$$

We finally obtain the camera pose estimate by applying the exponential map over the skew-symmetric matrix of $\boldsymbol{\delta}$ [18]. In all our experiments we initialize the damping factor λ at λ_{min}, multiply it by 2 when the camera pose update increases the feature-metric error, and divide it by half otherwise. We stop iterating when λ reaches λ_{max}.

FQN-based Direct Alignment. With a Feature Query Network $\mathcal{F}_g$, we can now rewrite the residual vector of the i-th keypoint at the k-th iteration of the LM optimization as:

$$\mathbf{r}_i^{j,k} = g(\mathtt{I}_q, \omega(\mathbf{u}_i, \mathtt{T}_k, \mathtt{K}_q)) - \mathcal{F}_g(\mathbf{u}_i, \theta_{i,k}, \phi_{i,k}, r_k, f_q, l_{i,k}). \tag{7}$$

With this novel formulation residual vectors are now independent of j. We no longer rely on reference images $\mathtt{I}_j$ and can now dynamically query and update the previously fixed 3D descriptors at every iteration of the LM algorithm. In a wide-baseline scenario where $\mathtt{I}_q$ is far away from the reference images $\{\mathtt{I}_j\}_{j=1}^M$ but our initial pose estimate $\mathtt{T}_0$ is close to the global minimum, we can now rely on $\mathcal{F}_g$ to generate descriptors much more geometrically aligned w.r.t. $\mathtt{I}_q$. This comes at a minimal computational cost as batched-forward passes on Feature Query Networks can be efficiently parallelized (see supp. mat.).

5. Experiments

In this section, we describe the experiments to demonstrate the power of Feature Query Networks for camera pose refinement. We first run an ablation study on the model parametrization, and subsequently demonstrate its ability to bridge viewpoint gaps in structure-based localization approaches.

5.1. Evaluation Details

Datasets. We evaluate our approach on 7Scenes [63] and Cambridge Landmarks [29], two popular datasets used in evaluation of learning-based relocalization methods. The former consists of 7 indoor scenes with posed RGB-D reference sequences, and RGB query images captured with different trajectories in the same environments. The latter consists 5 outdoor scenes, containing posed RGB reference images, on top of which S*f*M was run using COLMAP [60, 62] to obtain a SIFT [39]-based sparse 3D reconstruction. We train our FQN models on reference images, using dense 3D data for 7Scenes, and sparse 3D point cloud for Cambridge Landmarks.

Choice of g. To evaluate the ability of FQNs to model high-dimensional feature descriptors coming from different architectures, we choose 2 different CNN-based models pre-trained for feature matching. The first one is D2Net [17], a feature descriptor trained on Megadepth [37] with a VGG16 architecture [66], producing 512-dimensional descriptors. The second is a MobileNet-v2 [51] model trained on Megadepth [37] following [49] to produce more compact, 128-dimensional descriptors. We refer to each model as $\mathcal{F}_{\text{D2Net},\Theta}$ and $\mathcal{F}_{\text{MobileNetv2},\Theta}$.

Training details. We train a separate FQN model for every each of the previously mentioned 12 scenes and 2 descriptors, resulting in a total of 24 models weighting about 2Mb each. Training is done using Eq. 2 for 400k iterations with the Adam [30] optimizer. We set the initial learning rate to 1×10^{-4} and apply an exponential decay following [42]. Ground truth 3D descriptors are extracted using bilinear interpolation at the reprojected keypoint locations in reference images. For models using the focal length f as input, we apply a random image resizing (between 25% and 100%) to enable multi-scale inference, and randomly subsample a maximum of 2048 3D keypoint per sample. Training a single model takes about a day on an NVIDIA V100 GPU.

5.2. Ablation Study

To evaluate the importance of every parametrization term in Eq. 1, we perform an ablation study with $\mathcal{F}_{\text{D2Net},\Theta}$ on 7Scenes [63]. We train one model per scene with different input parameters. For every unseen test image and set of visible 3D coordinates in that image, we compute the FQN-based descriptors at the ground truth query pose, as well as the image-based descriptors (*i.e.* the interpolated descriptors at the 3D keypoint reprojections in query images). To study the impact of image scales on CNN-based descriptors, we also apply random image downsizing as done at training time. Finally we report the average L2-error over the 512-dimensional L2-normalized descriptors in Table 1.

					Average Per-Channel L2 Error	
	(θ, ϕ)	l	r	f	w/o random resizing ($\times 10^{-3}$)	w/ random resizing ($\times 10^{-3}$)
7Scenes [63]	-	-	-	-	1.508 ± 0.296	1.712 ± 0.261
	✓	-	-	-	1.506 ± 0.297	1.705 ± 0.264
	✓	✓	-	-	1.496 ± 0.304	1.710 ± 0.266
	✓	✓	✓	-	**1.491** ± 0.308	1.708 ± 0.268
	✓	✓	✓	✓	1.533 ± 0.270	**1.505** ± 0.236

Table 1. **FQN Parametrization Study:** We report the average per-channel L2 error between descriptors computed using $\mathcal{F}_{\text{D2Net},\Theta}$ and D2Net [17] on all 7Scenes [63] test images (lower is better). We find that as [42] the viewing direction (θ, ϕ) is a crucial parameter to encode descriptor variance. It is however not sufficient to handle changes caused by moving camera distance w.r.t. geometry or in-plane rotations, which are encoded by l and r respectively. Lastly to enable image-scale dependency we parametrize our model with the focal length f, and report errors on randomly resized images (last column).

We find that beyond the required 3D location input $\mathbf{u}$ of the keypoint to regress in the scene, every other additional input parameter contributes to a better modeling of the descriptor variance w.r.t. the camera viewpoint. In particular we find the camera distance to the geometry l and its roll angle r play an important role as deep descriptors lack invariance to these parameters. The same conclusion can be drawn for 2D image resizing, which is modeled by our model using f. Interestingly however, training with random image sizes coupled with the focal length parametrization damages the results on full-resolution images, indicating possible capacity limitations of our model.

5.3. Camera Pose Refinement

We now evaluate the application of FQNs to structure-based localization. More specifically, we run the algorithms proposed in Sec. 4.1 and Sec. 4.2 to perform camera pose refinement in wide-baseline configurations.

FQN-based Iterative PnP+RANSAC. In order to evaluate our proposed Algorithm 1, we consider the popular hierarchical localization framework [52], in which image retrieval w.r.t. the reference image set is used to identify a candidate set of 3D keypoints covisible with the query image $\mathtt{I}_q$. To exhibit the ability of FQNs to reduce the viewpoint discrepancy between reference and query images, we voluntarily choose to relocalize using only the top-1 nearest-neighbour. This also comes with the benefit of reducing the overall computational cost of hierarchical localization.

In our experiments, we perform image retrieval by computing global image descriptors using [48]. Matching is done with a simple mutual nearest-neighbour algorithm. To improve robustness to changes in scale w.r.t. the query image, we average FQN descriptors regressed at 25%, 50%

	Method	*7Scenes [63] (Indoor)*							*Cambridge Landmarks [29] (Outdoor)*				
		Chess	Office	Pumpkin	Heads	Fire	Kitchen	Stairs	StMary's	Court	Hospital	King's	Shop Facade
$\mathcal{F}_{\text{D2Net}}$	$K=0$	6.90/2.48	25.30/7.35	14.93/4.10	**4.03/2.83**	**6.36/2.54**	26.19/6.59	**109.18/27.20**	**44.45/1.72**	4685.66/95.82	105.66/1.56	35.69/0.58	17.41/0.77
	$K=1$	6.76/2.18	20.15/5.65	10.13/2.88	5.04/3.65	7.37/2.72	31.95/7.83	141.66/52.24	134.37/4.55	*4653.13/89.12*	77.80/1.28	40.61/0.58	*14.40/0.70*
	$K=5$	*6.13/1.95*	*15.17/4.31*	*9.57/2.70*	4.87/3.59	6.71/2.46	*20.41/5.37*	141.98/53.64	100.44/3.63	4673.10/89.16	*68.08/1.06*	*32.57/0.50*	14.84/0.67
	$K=30$	**5.96/1.87**	**14.16/4.11**	**9.53/2.66**	*4.86/3.51*	*6.58/2.42*	**18.37/4.79**	*140.76/53.02*	*92.82/3.54*	**4381.71/74.25**	**64.62/0.91**	**32.40/0.47**	**14.19/0.61**
$\mathcal{F}_{\text{MobileNetv2}}$	$K=0$	7.54/2.62	87.52/26.85	19.31/5.23	**3.11/2.15**	5.09/2.09	121.97/30.61	169.08/43.88	329.34/11.66	7679.41/99.43	468.21/9.15	35.84/0.44	31.14/1.39
	$K=1$	5.33/1.73	27.46/7.70	10.75/2.87	3.90/2.61	5.48/2.05	82.41/21.11	183.88/59.91	545.44/13.53	6976.24/68.84	144.43/2.83	32.19/0.45	24.74/1.08
	$K=5$	*4.49/1.40*	*12.31/3.49*	*9.48/2.51*	3.65/2.44	*4.76/1.80*	*19.25/5.25*	*168.08/50.71*	*120.44/4.49*	*5601.26/54.13*	*61.66/0.87*	*28.54/0.41*	**12.54/0.56**
	$K=30$	**4.13/1.31**	**10.47/2.97**	**9.24/2.45**	*3.55/2.36*	**4.62/1.76**	**16.12/4.42**	**139.50/34.67**	**57.95/2.00**	**4253.12/39.16**	**53.95/0.82**	**28.25/0.38**	*13.02/0.63*

Table 2. **FQN-based Iterative PnP+RANSAC Quantitative Results:** We report the median translation and rotation errors (in $cm/°$, lower is better) on both indoor and outdoor scenes using Alg. 1, for different values of K. We write in **bold** and *italic* the first and second lowest error respectively for every scene and descriptor. $K=0$ corresponds to the standard image-based PnP+RANSAC approach.

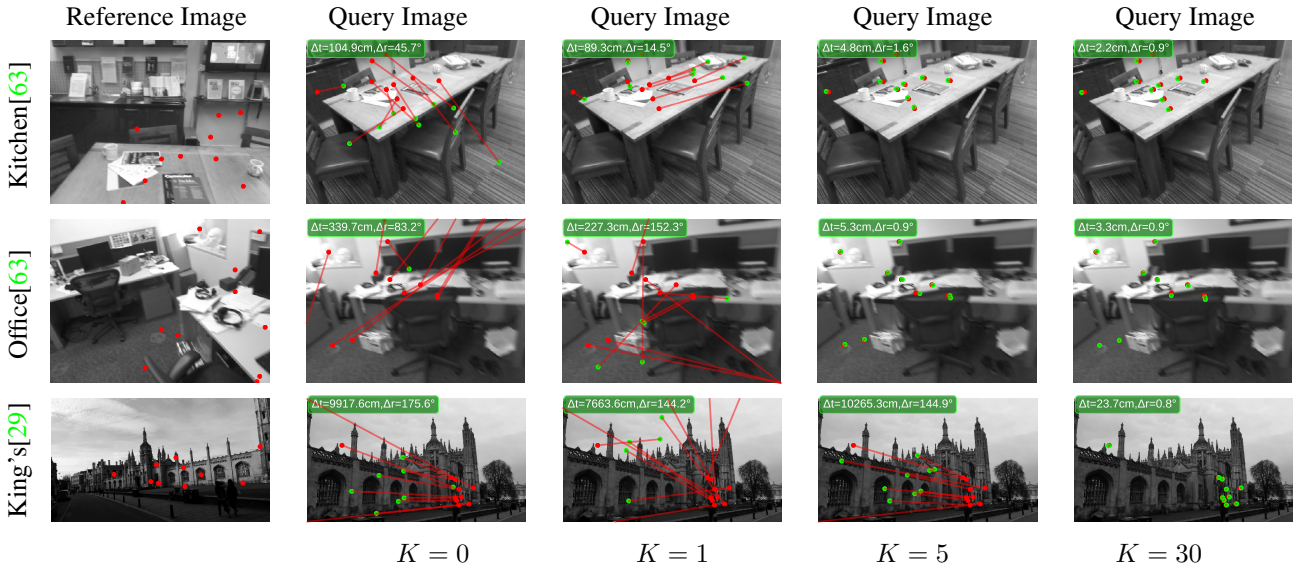

Figure 4. **Qualitative FQN-based Iterative PnP+RANSAC results:** We show reprojected keypoints at the estimated camera pose using image-based PnP+RANSAC with D2Net [17] descriptors ($K=0$), and FQN-based descriptors using $\mathcal{F}_{\text{D2Net}}$ ($K \geq 1$). We report in red the ground truth reprojection of a random subset of the 3D keypoints visible in the top-1 retrieved reference image (left column) and query image (other columns). We show in green the reprojection of 3D keypoints at the estimated camera pose. We find the geometric modeling of descriptors by FQNs helps recovering accurate camera pose estimates even for wide baselines.

and 100% original input size (through the focal length parameter f). We report in Tab. 2 the translation and rotation error of the estimated camera poses as a function of K. Note that $K=0$ corresponds to the standard 2D-to-3D matching followed by P*n*P+RANSAC, where the 3D descriptors are interpolated at their 2D reprojections in the reference image. We find that on most scenes, $\mathcal{F}_{\text{MobileNetv2}}$ is able to significantly reduce the initial camera pose error yielded by the classic P*n*P+RANSAC. On the higher-dimensional model $\mathcal{F}_{\text{D2Net}}$ and on larger scenes (*e.g.* St Mary's Church) improvements are less consistent, which is another hint that our model might be suffering from a limited capacity. We report qualitative results in Fig. 4.

In a second experiment, we deplete the set of reference camera poses in the scenes by randomly sampling 10 poses per scene and report the results in Fig. 5. This leads to even wider baselines between the query and reference images. We find that although the initial error at $K=0$ is fairly high, the FQN is able to reduce it significantly, halving the rotation error on average on Cambridge Landmarks [29]. Both experiments demonstrate the ability of FQNs to bridge viewpoint gaps for wide-baseline structure-based relocalization.

FQN-based Direct Alignment. We now evaluate the application of FQNs to wide-baseline direct alignment, as per Sec 4.2. We use as camera pose initialization the results of the previous PnP+RANSAC-based estimate, when $K=0$ (*i.e.* a standard P*n*P+RANSAC). The wide-baseline nature of the initial camera poses w.r.t. the ground truth query

	Method	*7Scenes [63] (Indoor)*							*Cambridge Landmarks [29] (Outdoor)*				
		Chess	Office	Pumpkin	Heads	Fire	Kitchen	Stairs	StMary's	Court	Hospital	King's	Shop Facade
$\mathcal{F}_{\text{D2Net}}$	$K = 0$	*6.90/2.48*	25.30/7.35	14.93/4.10	**4.03/2.83**	**6.36/2.54**	26.19/6.59	109.18/27.20	44.45/*1.72*	*4685.66*/95.82	105.66/1.56	*35.69*/0.58	17.41/**0.77**
	Eq. 5	8.24/3.12	31.89/8.56	23.48/5.86	7.99/6.07	15.11/5.81	24.53/6.35	**95.24/18.58**	**51.96/1.62**	**4671.24/90.78**	**89.60**/*1.27*	44.47/0.73	17.85/*0.79*
	Eq. 7 - *Static*	7.28/2.78	**21.08/5.99**	**14.08/3.73**	6.33/4.89	*9.59/3.86*	**23.47**/*5.60*	100.75/20.73	*53.78*/1.74	4942.98/*92.95*	*93.14*/**1.11**	39.66/*0.53*	**16.20**/0.87
	Eq. 7 - *Dynamic*	**6.19/2.3**	*19.97/5.94*	*14.53/3.72*	*6.27/4.63*	10.75/4.35	*23.76*/**5.50**	*97.33/20.00*	60.14/1.74	4861.03/95.45	105.14/1.28	**35.45/0.47**	*16.59*/0.85
$\mathcal{F}_{\text{MobileNetv2}}$	$K = 0$	7.54/2.62	87.52/26.85	19.31/5.23	**3.11/2.15**	5.09/2.09	121.97/30.61	169.08/43.88	329.34/11.66	7679.41/99.43	468.21/9.15	35.84/0.44	31.14/1.39
	Eq. 5	6.32/2.42	77.87/**24.77**	13.67/3.91	*3.26/2.73*	**3.94/1.78**	*116.15*/**28.95**	*166.85/43.02*	**65.73**/*3.41*	7735.02/99.58	81.56/1.04	31.92/0.43	12.11/0.63
	Eq. 7 - *Static*	*5.34/2.05*	**76.91**/*24.81*	*11.55/3.14*	3.54/2.94	*4.72/1.93*	116.46/29.52	167.09/43.32	91.05/3.66	*7709.67/97.98*	*65.09/0.94*	*26.06/0.34*	**9.69/0.50**
	Eq. 7 - *Dynamic*	**5.18/1.99**	*77.05*/25.30	**11.12/3.10**	3.59/2.91	4.84/1.97	**115.76**/*29.33*	**163.76/41.32**	*68.78*/**2.88**	7782.13/**96.69**	**65.01/0.87**	**25.29/0.33**	*9.91/0.55*

Table 3. **FQN-based Direct Alignment Quantitative Results:** We report the median translation and rotation errors (in $cm/^\circ$, lower is better) on both indoor and outdoor scenes using FQN-based direct alignment. We report results using standard direct alignment (Eq. 5), as well as FQN-based residuals (Eq. 7), which we either use at initialization (*Static*) or continuously (*Dynamic*). We write in **bold** and *italic* the first and second lowest error respectively for every scene and descriptor. We find FQNs work best with medium-sized descriptors and scenes, and can sometimes bring significant improvements compared to standard direct alignment approaches.

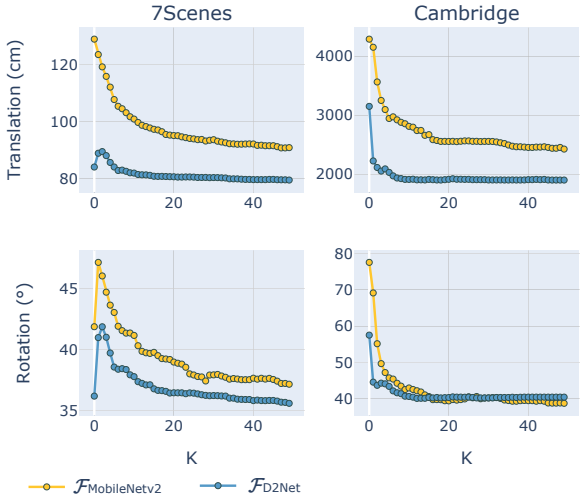

Figure 5. **FQNs for few-reference image HLoc:** To exhibit the ability of FQNs to bridge viewpoint gaps we run Alg. 1 using only 10 retrievable reference camera poses per scene. This sparsified reference set naturally leads to wider baselines when relocalizing, and thus a higher error at $K = 0$. We report the averaged median error on all scenes from both 7Scenes [63] and Cambridge Landmarks [29] and find the FQN helps reduce the initial pose errors.

pose makes this problem especially challenging for a direct alignment method, which is prone to falling in local minima and requires wide and accurate convergence basins. We report the results in Tab. 3 of our feature-metric error minimization algorithm using both standard image-based 3D descriptors (Eq. 5), and using updated FQN-based residuals (Eq. 7). We report results using two variants for this FQN-based formulation, one where we only regress 3D descriptors at initialization which is referred to as *static* (*i.e.* setting $\mathbf{r}_i^{j,k} = \mathbf{r}_i^{j,0}\ \forall\ k > 1$), and one where the 3D descriptors are continuously updated, which we refer to as *dynamic*.

We find the FQN-based direct alignment to provide consistent improvements over standard direct alignment methods in cases where the scene is both relatively small (*e.g.* Old Hospital, Shop Facade) and where descriptors are more compact ($\mathcal{F}_{\text{MobileNetv2}}$). Interestingly, the *dynamic* update of descriptors does not necessarily imply an improvement over *static* descriptors, which could indicate a lack of accuracy in descriptor regression around the global optimum. We find that on larger scenes such as St Mary's Church or Great Court, as well as on high-dimensional descriptors, FQN-based residuals can damage performance.

6. Discussion

As shown in this paper, attempting to reach descriptor invariance can be circumvented by rather explicitly modeling the variance of such descriptors w.r.t. the camera viewpoint. This initial formulation however comes with some limitations. Much like other implicit representation learning methods [42], Feature Query Networks seem to exhibit a limited scaling ability when applied to large scenes or very high-dimensional descriptor. Local conditioning of the model or increased model capacity might help tempering those issues. Our model also lacks a proper modeling of appearance variations in descriptor (mainly due to the lack of available training data), which could be important for long-term relocalization. Incorporating density estimation in FQNs could also enable potentially help deal with noisy 3D reconstructions. We believe these research paths could make for interesting future work.

7. Conclusion

In this paper we introduce Feature Query Networks, simple surface-level MLPs designed to model the variance of a given descriptor in a scene w.r.t. the camera viewpoint. Rather than trying to force invariance in descriptors, we model it directly with powerful neural networks. We showed their capacity to regress high-dimensional descriptor under novel viewpoints on specific scenes, as well as applications to wide-baseline structure-based visual localization for improved camera pose estimation.

References

[1] Vassileios Balntas, Shuda Li, and V. Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In *ECCV*, 2018. 1, 3

[2] Vassileios Balntas, Edgar Riba, D. Ponsa, and K. Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *BMVC*, 2016. 2, 3

[3] Daniel Barath, Jiri Matas, and Jana Noskova. MAGSAC: Marginalizing Sample Consensus. In *CVPR*, 2019. 3

[4] D. Baráth, J. Noskova, Maksym Ivashechkin, and Jiri Matas. Magsac++, a fast, reliable and accurate robust estimator. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1301–1309, 2020. 3

[5] Eric Brachmann, Alexander Krull, S. Nowozin, J. Shotton, Frank Michel, S. Gumhold, and C. Rother. Dsac — differentiable ransac for camera localization. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2492–2500, 2017. 1, 3

[6] Eric Brachmann and C. Rother. Learning less is more - 6d camera localization via 3d surface regression. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4654–4662, 2018.

[7] Eric Brachmann and C. Rother. Expert sample consensus applied to camera re-localization. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7524–7533, 2019. 1, 3

[8] Eric Brachmann and Carsten Rother. Neural- Guided RANSAC: Learning Where to Sample Model Hypotheses. In *ICCV*, 2019. 3

[9] Eric Brachmann and C. Rother. Visual camera relocalization from rgb and rgb-d images using dsac. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2021. 1, 3

[10] Martin Bujnak, Z. Kukelova, and T. Pajdla. New efficient solution to the absolute pose problem for camera with unknown focal length and radial distortion. In *ACCV*, 2010. 1, 3, 4

[11] Tommaso Cavallari, Luca Bertinetto, Jishnu Mukhoti, Philip H. S. Torr, and S. Golodetz. Let's take this online: Adapting scene coordinate regression network predictions for online rgb-d camera relocalisation. *2019 International Conference on 3D Vision (3DV)*, pages 564–573, 2019. 1, 3

[12] Tommaso Cavallari, S. Golodetz, N. Lord, Julien P. C. Valentin, L. D. Stefano, and Philip H. S. Torr. On-the-fly adaptation of regression forests for online camera relocalisation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 218–227, 2017. 1, 3

[13] O. Chum, Tomás Werner, and Jiri Matas. Two-view geometry estimation unaffected by a dominant plane. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:772–779 vol. 1, 2005. 3

[14] J. Czarnowski, Stefan Leutenegger, and A. Davison. Semantic texture for robust dense tracking. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 851–859, 2017. 3

[15] D. Detone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-Supervised Interest Point Detection and Description. In *CVPR*, 2018. 2, 3

[16] Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, and P. Luo. Camnet: Coarse-to-fine retrieval for camera relocalization. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2871–2880, 2019. 1, 3

[17] Mihai Dusmanu, Ignacio Rocco, T. Pajdla, M. Pollefeys, Josef Sivic, A. Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8084–8093, 2019. 2, 3, 4, 6, 7

[18] Jakob Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:611–625, 2018. 3, 5

[19] Jakob Engel, Thomas Schöps, and D. Cremers. Lsd-slam: Large-scale direct monocular slam. In *ECCV*, 2014. 3

[20] M. A. Fischler and R. C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM*, 24, 1981. 1, 3

[21] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and T. Funkhouser. Local deep implicit functions for 3d shape. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4856–4865, 2020. 2, 3

[22] H. Germain, G. Bourmaud, and V. Lepetit. Sparse-To-Dense Hypercolumn Matching for Long-Term Visual Localization. In *International Conference on 3D Vision*, 2019. 3

[23] Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. S2DNet: Learning Image Features for Accurate Sparse-to-Dense Matching. In *ECCV*, 2020. 3, 4

[24] Hugo Germain, Vincent Lepetit, and Guillaume Bourmaud. Neural reprojection error: Merging feature learning and camera pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 414–423, June 2021. 3

[25] R. M. Haralick, C.-N. Lee, K. Ottenberg, and M. Nölle. Review and Analysis of Solutions of the Three Point Perspective Pose Estimation Problem. *IJCV*, 13, 1994. 1, 3, 4

[26] Jared Heinly, J. Schonberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the world* in six days *(as captured by the yahoo 100 million image dataset). In *CVPR 2015*, 2015. 1

[27] C. Jiang, Avneesh Sud, A. Makadia, Jingwei Huang, Matthias Nießner, and T. Funkhouser. Local implicit grid representations for 3d scenes. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6000–6009, 2020. 2, 3

[28] Alex Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6555–6564, 2017. 1, 3

[29] Alex Kendall, M. Grimes, and R. Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2938–2946, 2015. 1, 2, 3, 6, 7, 8

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 6

[31] Z. Kukelova, Martin Bujnak, and T. Pajdla. Real-time solution to the absolute pose problem with unknown radial distortion and focal length. *2013 IEEE International Conference*

on Computer Vision, pages 2816–2823, 2013. 1, 3, 4

[32] Axel Barroso Laguna, Edgar Riba, D. Ponsa, and K. Mikolajczyk. Key.net: Keypoint detection by handcrafted and learned cnn filters. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5835–5843, 2019. 3

[33] Zakaria Laskar, I. Melekhov, S. Kalia, and Juho Kannala. Camera relocalization by computing pairwise relative poses using convolutional neural network. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 920–929, 2017. 1, 3

[34] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. *International Journal of Computer Vision*, 81:155–166, 2008. 1, 3, 4

[35] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2:164–168, 1944. 5

[36] Xiaotian Li, Shuzhe Wang, Yi Zhao, J. Verbeek, and Juho Kannala. Hierarchical scene coordinate classification and regression for visual localization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11980–11989, 2020. 1, 3

[37] Z. Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. *CVPR*, pages 2041–2050, 2018. 6

[38] Yen-Chen Lin, Peter R. Florence, J. T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. *ArXiv*, abs/2012.05877, 2020. 2, 3, 4

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

[40] K. Madsen, H. B. Nielsen, and O. Tingleff. Methods for non-linear least squares problems (2nd ed.). 2004. 5

[41] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, S. Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4455–4465, 2019. 2, 3

[42] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, J. T. Barron, R. Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3, 4, 6, 8

[43] Anastasiya Mishchuk, Dmytro Mishkin, Filip Radenović, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *NIPS*, 2017. 3

[44] Kwang Moo yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to Find Good Correspondences. In *CVPR*, pages 2666–2674, 2018. 3

[45] Michael Niemeyer, Lars M. Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3512, 2020. 3

[46] Y. Ono, Eduard Trulls, P. Fua, and K. M. Yi. Lf-net: Learning local features from images. In *NeurIPS*, 2018. 2, 3

[47] Jeong Joon Park, Peter R. Florence, J. Straub, Richard A. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019. 2, 3

[48] F. Radenovic, G. Tolias, and O. Chum. Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE TPAMI*, 2018. 6

[49] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. In *NeurIPS*, volume 32, pages 12405–12415. Curran Associates, Inc., 2019. 2, 3, 6

[50] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic. Neighbourhood Consensus Networks. In *NeurIPS*, 2018. 3

[51] M. Sandler, Andrew G. Howard, Menglong Zhu, A. Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 6

[52] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *CVPR*, pages 12716–12725, 2019. 1, 3, 4, 6

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

[54] Paul-Edouard Sarlin, Ajaykumar Unagar, Maans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl, and Torsten Sattler. Back to the feature: Learning robust camera localization from pixels to pose. *ArXiv*, abs/2103.09213, 2021. 1, 2, 3, 5

[55] Torsten Sattler, William P. Maddern, Carl Toft, A. Torii, L. Hammarstrand, Erik Stenborg, Daniel Safari, M. Okutomi, M. Pollefeys, Josef Sivic, F. Kahl, and T. Pajdla. Benchmarking 6dof outdoor visual localization in changing conditions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8601–8610, 2018. 1, 2, 3

[56] T. Sattler, C. Sweeney, and M. Pollefeys. On Sampling Focal Length Values to Solve the Absolute Pose Problem. In *ECCV*, 2014. 1, 3

[57] T. Sattler, A. Torii, J. Sivic, M. Pollefeys, H. Taira, M. Okutomi, and T. Pajdla. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? In *CVPR*, 2017. 1

[58] Torsten Sattler, Qunjie Zhou, M. Pollefeys, and L. Leal-Taixé. Understanding the limitations of cnn-based absolute camera pose regression. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3297–3307, 2019. 2, 3

[59] N. Savinov, A. Seki, L. Ladicky, T. Sattler, and M. Pollefeys. Quad-Networks: Unsupervised Learning to Rank for Interest Point Detection. In *CVPR*, 2016. 3

[60] J. L. Schönberger and J.-M. Frahm. Structure-From-Motion Revisited. In *CVPR*, 2016. 1, 6

[61] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler. Semantic Visual Localization. *CoRR*, abs/1712.05773, 2017. 2, 3

[62] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *ECCV*, 2016. 1, 6

[63] J. Shotton, Ben Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–

2937, 2013. 2, 6, 7, 8

[64] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative Learning of Deep Convolutional Feature Point Descriptors. In *ICCV*, 2015. 2, 3

[65] K. Simonyan, A. Vedaldi, and A. Zisserman. Learning Local Feature Descriptors Using Convex Optimisation. *IEEE TPAMI*, 36, 2014. 2, 3

[66] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556, 2014. 6

[67] V. Sitzmann, Michael Zollhoefer, and G. Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *ArXiv*, abs/1906.01618, 2019. 3

[68] L. Stumberg, P. Wenzel, Nan Yang, and D. Cremers. Lm-reloc: Levenberg-marquardt based direct visual relocalization. *2020 International Conference on 3D Vision (3DV)*, pages 968–977, 2020. 1, 3, 5

[69] Linus Svärm, O. Enqvist, F. Kahl, and M. Oskarsson. City-scale localization for cameras with known vertical direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1455–1461, 2017. 1

[70] L. Svärm, O. Enqvist, M. Oskarsson, and F. Kahl. Accurate Localization and Pose Estimation for Large 3D Models. In *CVPR*, 2014. 1

[71] C. Sweeney, V. Fragoso, T. Höllerer, and M. Turk. Large Scale SfM with the Distributed Camera Model. In *International Conference on 3D Vision*, 2016. 1

[72] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, R. Ramamoorthi, J. T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *ArXiv*, abs/2006.10739, 2020. 3, 4

[73] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas. SOSNet: Second Order Similarity Regularization for Local Descriptor Learning. In *CVPR*, 2019. 3

[74] Carl Toft, Daniyar Turmukhambetov, Torsten Sattler, F. Kahl, and G. Brostow. Single-image depth prediction makes feature matching easier. In *ECCV*, 2020. 3

[75] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning Local Features with Policy Gradient. In *NeurIPS*, 2020. 3

[76] Yannick Verdie, K. M. Yi, P. Fua, and V. Lepetit. Tilde: A temporally invariant learned detector. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5279–5288, 2015. 3

[77] Lukas von Stumberg, P. Wenzel, Q. Khan, and D. Cremers. Gn-net: The gauss-newton loss for multi-weather relocalization. *IEEE Robotics and Automation Letters*, 5:890–897, 2020. 1, 3, 5

[78] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers. Image-Based Localization Using LSTMs for Structured Feature Correlation. In *ICCV*, 2017. 1, 3

[79] Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and P. Tan. Sanet: Scene agnostic network for camera localization. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 42–51, 2019. 1, 3

[80] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. LIFT: Learned Invariant Feature Transform. In *ECCV*, 2016. 2, 3

[81] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning Two-View Correspondences and Geometry Using Order-Aware Network. In *ICCV*, 2019. 3

[82] Qunjie Zhou, Torsten Sattler, M. Pollefeys, and L. Leal-Taixé. To learn or not to learn: Visual localization from essential matrices. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3319–3326, 2020. 1, 3