Supplementary Material
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

In the following pages, we present additional quantitative results, qualitative results and experimental details regarding Feature Query Networks. In particular, we present results for small-baseline FQN-based camera pose refinement, and find $\mathcal{F}_{\text{MobileNetv2}}$ achieves performance on par with the state-of-the-art method PixLoc [11].

A. Small-baseline FQN-based Direct Alignment

In the previous experiments, we showed FQN-based direct alignment results when the initial camera pose is far away from the ground truth query pose (*wide*-baseline). While this setup exhibits the power of Feature Query Networks for camera pose refinement, it is a significantly harder camera pose initialization to recover from and thus makes comparison to state-of-the-art localization methods difficult. In this section, we report results for *small*-baseline direct alignment on Cambridge Landmarks [5].

To do so, we first perform hierarchical localization (similar to HLoc [9]) on the top-50 nearest neighbours, and estimate the initial camera pose using PnP+RANSAC as done in Sec. 5.3. Then, we perform both standard and FQN-based direct alignment, and report the results in Tab. 1. We report the performance of other state-of-the-art methods including ActiveSearch [12], HLoc [9] (which uses SuperPoint [3] keypoints and matches them with SuperGlue [10]), DSAC* [1], HACNet [6], SANet [14] and PixLoc [11]. DSAC* results are from the RGB+3D setting. The train and test folds are identical for all methods in our experiments. Note that PixLoc is also a direct alignment method which uses learned scene-agnostic features tailored for camera pose refinement.

Overall we find that in a small-baseline setting, the FQN-based direct alignment brings consistent improvements over standard direct alignment when trained on MobileNetv2 [8], *i.e*. 128-dimensional descriptors. In fact, we find the achieved performance is on par if not better than other end-to-end learning-based competitors (including PixLoc [11]), despite our MobiletNetv2 [8] model being trained for keypoint matching as [7]. This is a strong indicator that modeling the variance of descriptors w.r.t. to viewpoint might be a promising direction for future research.

As previously discovered, we also find that the higher-dimensional descriptor D2Net [4] does not perform as well for FQN-based direct alignment. This could indicate $\mathcal{F}_{\text{D2Net}}$ is lacking capacity to accurately model such descriptors, especially around the global minimum. We believe this limitation of our approach is also an interesting path for future research. Illustrations of failure cases with $\mathcal{F}_{\text{D2Net}}$ are presented below.

B. Additional Qualitative Results

We report in Fig. 3 and Fig. 2 additional qualitative using the proposed iterative FQN-based PnP+RANSAC. These examples indirectly illustrate the ability of FQNs to perform descriptor regression from novel viewpoints.

A more direct visualization of this ability can be seen in Fig. 1, which displays correspondence maps obtained with both D2Net computed from reference images, and using FQNs $\mathcal{F}_{\text{D2Net}}$. We find FQNs help produce correspondence maps with modes much better aligned with ground truth correspondences.

Lastly we report in Fig. 4 and Fig. 5 failure cases, where FQNs fail to provide an improved descriptor regression. We hypothetize these failures come from the lack of generalization of FQNs to novel viewpoints due to a limited model capacity, given the high dimensionality of descriptors and scene scales. In the case of FQN-based iterative PnP+RANSAC, it is also likely that the algorithm gets stuck in local minima where the FQN descriptors are no longer updated significantly. This is an issue similar to the one encountered in gradient-based optimization.

*Work done during an internship at Reality Labs.

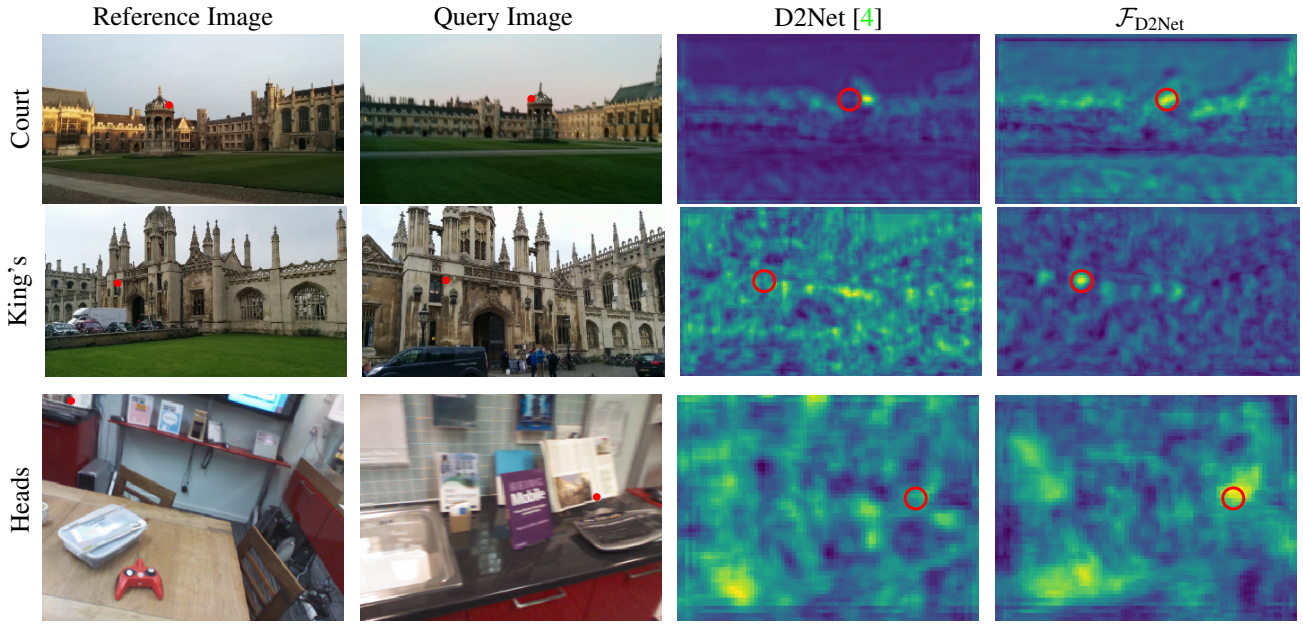

Figure 1. **FQN-based Correspondence Maps**: We show in red the ground-truth reprojection of a given 3D keypoint in the scene and (from left to right) a pair of reference and query images, as well as dense correspondence maps obtained using image-based and FQN-regressed D2Net [4] descriptors (computed at the ground truth query camera pose). We find that our approach is able to produce much more accurate correspondence maps on wide-baseline image pairs.

C. Additional Experiments Details

In this section we report experimental details regarding our application of FQNs to camera pose refinement.

FQN-based Iterative PnP+RANSAC. To perform this study, we retrieve for every query image of every scene the top-1 nearest-neighbour image and identify its set of visible 3D keypoints. To increase robustness to scale changes, we compute multi-scale image-based descriptors using both D2Net [4] and our trained MobileNetv2 model. We also regress multi-scale descriptors using FQNs through the focal length f parametrization. We use the OpenCV [2] RANSAC implementation and tune the threshold for optimal performance.

FQN-based Direct Alignment. For FQN-based direct alignement, we employ a similar multi-scale strategy and average multi-scale descriptors, as we find it brings an increased robustness to wide baseline camera pose initializations. We use $\lambda_{\text{min}} = 10^{-8}$ and $\lambda_{\text{max}} = 10^{7}$.

D. Computational Cost

We experimentally find that FQNs are not only light in memory (about 2Mb), but they also bring a small additional cost to camera pose refinement. On a single NVIDIA GTX 3070 we find performing a batched forward pass on $10,000$ keypoints takes only 11.2ms on average. In comparison a multi-scale inference on a model like D2Net [4] which takes several hundred milliseconds, this cost is neglictible and allows for multiple FQN-based descriptor updates per query image.

References

[1] 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, 7

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 2

[3] D. Detone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-Supervised Interest Point Detection and Description. In *CVPR*, 2018. 1

[4] 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. 1, 2, 3, 4, 5, 7

[5] 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, 3, 7

[6] Xiaotian Li, Shuzhe Wang, Yi Zhao, J. Verbeek, and Juho Kannala. Hierarchical scene coordinate classification and regression for visual localization. *2020 IEEE/CVF Conference*

Figure 2. **Qualitative FQN-based Iterative PnP+RANSAC results on Cambridge Landmarks [5]:** We show reprojected keypoints at the estimated camera pose using image-based PnP+RANSAC with D2Net [4] 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), and in green the reprojection of 3D keypoints at the estimated pose.

Figure 3. **Qualitative FQN-based Iterative PnP+RANSAC results on 7Scenes [13]:** We show reprojected keypoints at the estimated camera pose using image-based PnP+RANSAC with D2Net [4] 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), and in green the reprojection of 3D keypoints at the estimated pose.

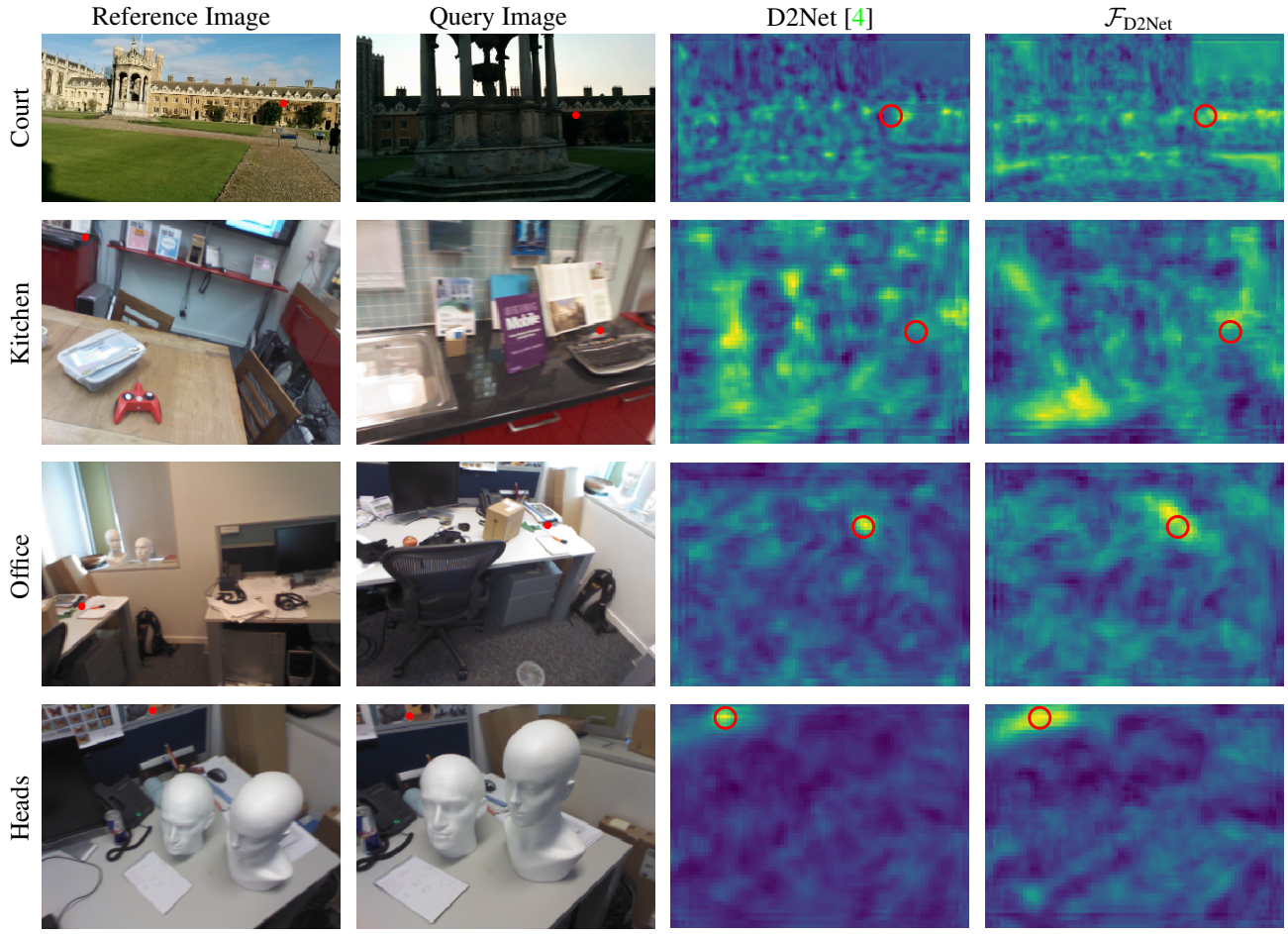

Figure 4. **FQN-based descriptor regression failure cases**: We show in red the ground-truth reprojection of a given 3D keypoint in the scene and (from left to right) a pair of reference and query images, as well as dense correspondence maps obtained using image-based and FQN-regressed D2Net [4] descriptors. We find that on very wide baselines FQN descriptors are do not manage to produce accurate correspondence maps. This could be explained by the novel query viewpoint being too far away from the reference image set (*e.g.* first three rows). We also find that when images are better aligned, the FQN leads to overall less peaky correspondence maps which translates a limitation in regression fidelity, perhaps linked to a limited model capacity.

on Computer Vision and Pattern Recognition (CVPR), pages 11980–11989, 2020. 1, 7

[7] 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. 1

[8] 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. 1

[9] 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, 7

[10] Paul-Edouard Sarlin, Daniel Detone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning Feature Matching with Graph Neural Networks. In *CVPR*, 2020. 1

[11] 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, 7

[12] Torsten Sattler, B. Leibe, and L. Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1744–1756, 2017. 1, 7

[13] 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. 4

[14] Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and P. Tan. Sanet: Scene agnostic network for camera localization. *2019 IEEE/CVF International Con-*

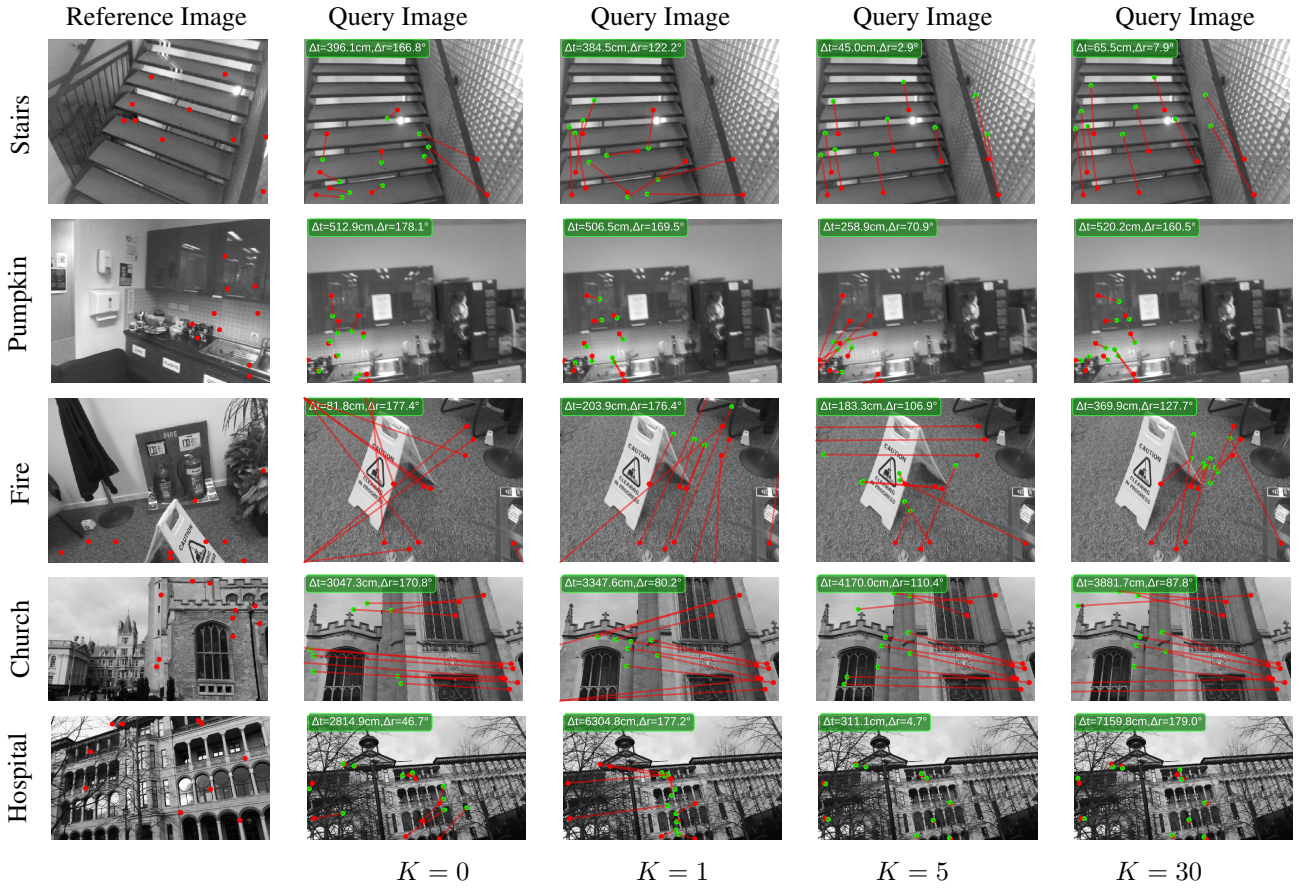

Figure 5. **FQN-based Iterative PnP+RANSAC Failure Cases:** We report in this figure failure cases for our proposed camera refinement method. We find that our method sometimes diverges, leading to high or increased pose errors. This is most likely due to regression errors in the FQN descriptor, but also possibly to the method falling in local minima, akin to gradient-based optimization methods.

ference on Computer Vision (ICCV), pages 42–51, 2019. 1, 7

	Method	*Cambridge Landmarks [5] (Outdoor)*				
		StMary's	Court	Hospital	King's	Shop Facade
	AS [12]	8/0.25	24/0.13	20/0.36	13/0.22	4/0.21
	HLoc [9]	7/0.21	16/0.11	15/0.30	12/0.20	4/0.20
	DSAC* [1]	13/0.4	49/0.3	21/0.4	15/0.3	5/0.3
	HACNet [6]	9/0.3	28/0.2	19/0.3	18/0.3	6/0.3
	SANet [14]	16/0.57	328/1.95	32/0.53	32/0.54	10/0.47
	PixLoc [11]	10/0.34	30/0.14	16/0.32	14/0.24	5/0.23
$\mathcal{F}_{\text{D2Net}}$	$K = 0$	18.94/0.64	115.53/0.74	48.21/0.81	33.46/0.55	11.32/0.51
	Eq.5	**9.38/0.34**	**42.17/0.29**	*20.29*/**0.35**	**14.52**/*0.26*	**5.40/0.26**
	Eq.7 - *Static*	12.24/0.42	70.04/*0.38*	20.84/*0.38*	15.50/**0.24**	6.64/0.34
	Eq.7 - *Dynamic*	*11.83/0.40*	*68.98/0.38*	**18.67**/*0.38*	*15.21*/**0.24**	*6.62/0.33*
$\mathcal{F}_{\text{MobileNetv2}}$	$K = 0$	11.98/0.40	59.06/0.35	36.44/0.55	17.05/0.29	5.25/0.24
	Eq.5	11.99/*0.38*	37.99/*0.24*	24.62/*0.35*	15.59/*0.26*	6.15/*0.33*
	Eq.7 - *Static*	*10.55*/**0.30**	**36.43/0.13**	*17.28*/**0.26**	*11.71*/**0.21**	**4.85/0.23**
	Eq.7 - *Dynamic*	**10.49/0.30**	*36.49*/**0.13**	**17.09/0.26**	**11.58/0.21**	*4.87*/**0.23**

Table 1. **Small-baseline FQN-based Direct Alignment:** We report the median translation and rotation errors (in $cm/^\circ$, lower is better) on Cambridge Landmarks [5] scenes using FQN-based direct alignment. In this setup, we use the top-50 nearest neighbours which naturally leads to better initial camera pose estimates ($K = 0$) and smaller baselines using the subsequent direct alignments. We report results using standard direct alignment, as well as FQN-based residuals, 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 $\mathcal{F}_{\text{MobiletNetv2}}$ is able to achieve performance on par with state-of-the-art learning-based direct alignement method PixLoc [11]. This study also corroborates the limitations of FQNs in modeling high-dimensional descriptors like D2Net [4].