From Coarse to Fine: Robust Hierarchical Localization at Large Scale

Paul-Edouard Sarlin[1] Cesar Cadena[1] Roland Siegwart[1] Marcin Dymczyk[1,2]
[1]Autonomous Systems Lab, ETH Zürich [2]Sevensense Robotics AG

Abstract

Robust and accurate visual localization is a fundamental capability for numerous applications, such as autonomous driving, mobile robotics, or augmented reality. It remains, however, a challenging task, particularly for large-scale environments and in presence of significant appearance changes. State-of-the-art methods not only struggle with such scenarios, but are often too resource intensive for certain real-time applications.

In this paper we propose HF-Net, a hierarchical localization approach based on a monolithic CNN that simultaneously predicts local features and global descriptors for accurate 6-DoF localization. We exploit the coarse-to-fine localization paradigm: we first perform a global retrieval to obtain location hypotheses and only later match local features within those candidate places. This hierarchical approach incurs significant runtime savings and makes our system suitable for real-time operation. By leveraging learned descriptors, our method achieves remarkable localization robustness across large variations of appearance. Consequently, we demonstrate new state-of-the-art performance on two challenging benchmarks for large-scale 6-DoF localization. The code of our method will be made publicly available.

1. Introduction

The precise 6-Degree-of-Freedom (DoF) localization of a camera within an existing 3D model is one of the core computer vision capabilities that unlocks a number of recent applications. These include autonomous driving in GPS-denied environments and consumer devices with augmented reality features, where a centimeter-accurate 6-DoF pose is crucial to guarantee reliable and safe operation and fully immersive experiences, respectively. More broadly, visual localization is a key component in computer vision tasks such as Structure-from-Motion (SfM) or SLAM. This growing range of applications of visual localization calls for reliable operation both indoors and outdoors, irrespective of the weather, illumination or seasonal changes.

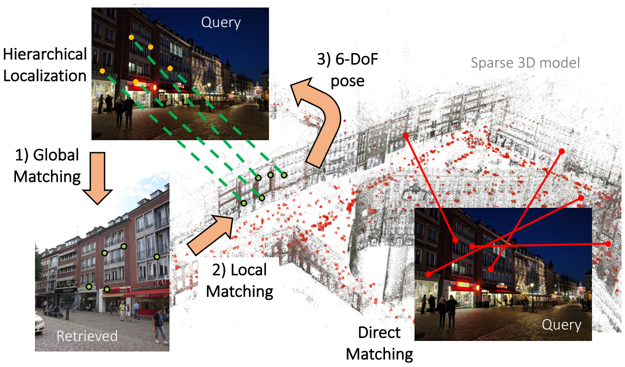

Figure 1. **HF-Net working principle.** We start by performing global matching, which helps us to remain computationally efficient and improves the robustness in challenging situations. Then, using powerful local features lets us establish reliable correspondences and estimate an accurate 6-DoF pose.

Robustness to large perceptual changes, in terms of illumination, viewpoint, or between weather conditions or seasons, is therefore critical, along with limited computational resources. Maintaining a model that allows accurate localization in multiple conditions, while remaining compact, is thus of utmost importance. In this work, we investigate whether it is actually possible to robustly localize in large-scale changing environments with constrained resources of mobile devices. More specifically, we aim at estimating the 6-DoF pose of a query image w.r.t. a given 3D model with the highest possible accuracy.

Current leading approaches mostly rely on estimating correspondences between 2D keypoints in the query and 3D points in a sparse model using local descriptors. This direct matching is either robust but intractable on mobile [40, 44], or optimized for efficiency but fragile [24]. In both cases, the robustness of classical localization methods is limited by the poor invariance of hand-crafted local features. Recent features emerging from convolutional neural networks (CNN) exhibit unrivalled robustness at a low compute cost. They have been, however, only recently [41] applied to the visual localization problem, and only in a dense, expensive manner. Learned sparse descriptors [12, 29] promise large benefits that remain yet unexplored in localization.

Alternative localization approaches based on image retrieval have recently shown promising results in terms of robustness and efficiency. As such, [32] demonstrated the benefits of an intermediate retrieval step at small-scale and using off-the-shelf components, thus not reaching the scalability required by city-scale localization.

In this paper, we propose a Hierarchical Feature Network (HF-Net) – an approach that bridges the gap between robustness and efficiency by leveraging recent advances in deep learning. The principle of our method is depicted in Figure 1. At its core, our approach is a CNN that jointly estimates local and global features to localize in a hierarchical manner. Similar to how humans localize, we employ a natural coarse-to-fine pose estimation process that is highly efficient and scales well with large environments. Global features first capture a wide context of the image, which provides robustness to perceptual aliasing while enabling a scalable coarse initial search. Learned local features are then used to estimate precise correspondences with the model for the accurate estimation of the pose. They carry a powerful representation of the visual information through fewer but more repeatable keypoints associated with highly matchable local descriptors. This enables a faster, yet precise, local search. The joint prediction of both scales within a single network maximizes the sharing of computation across the tasks for an efficient inference online. Overall, our contributions are as follow:

– We set a new state-of-the-art in several public benchmarks for large-scale localization with an outstanding robustness in particularly challenging condition, *e.g.* across seasons;

– We introduce HF-Net, a monolithic neural network which efficiently predicts hierarchical features for a fast and robust localization;

– We employ a novel multitask distillation procedure to train HF-Net in a flexible way that matches the accuracy of the teacher method with an unparalleled efficiency.

2. Related work

In this section we review other works that relate to different components of our approach, namely: visual localization, scalability, feature learning, and deployment on resource constrained devices.

6-DoF visual localization methods have traditionally been classified as either structure-based or image-based. The former perform direct matching of local descriptors between 2D keypoints of a query image and 3D points in a 3D SfM model. These methods are able to estimate accurate poses, but often rely on exhaustive matching and are thus compute intensive [40, 34]. As the model grows in size and perceptual aliasing arises, this matching becomes ambiguous, impairing the robustness of the localization, especially under strong appearance changes such as day-night. Another group of approaches attempts to directly regress the pose from the image [20, 18] or to classify it to a spatial bin [46]. Image-based methods are related to image retrieval and are only able to provide an approximate pose up to the database discretization, which is not sufficiently precise for many applications. They are however significantly more robust than direct local matching as they rely on the global image-wide information [2, 5]. Robustness comes at the cost of increased compute, as state-of-the-art image retrieval is based on large deep learning models.

Scalable localization often deals with the additional compute constrains by using features that are inexpensive to extract, store, and match together [7, 22, 30]. These improve the runtime on mobile devices but further impair the robustness of the localization, limiting their operations [24] to stable conditions. Hierarchical localization [17, 26, 32] takes a different approach by dividing the problem into a global, coarse search followed by a fine pose estimation. Recently, [32] proposed to search at the map level using image retrieval and localize by matching hand-crafted local features against retrieved 3D points. As we discuss further in Section 3, its robustness and efficiency are limited by the underlying local descriptors and heterogeneous structure.

Learned local features have recently been developed in attempt to replace hand-crafted descriptors. Dense pixel-wise features naturally emerge from CNNs and provide a powerful representation used for image matching [9] and localization [41]. Matching dense features is however intractable with limited computational power. Sparse learned features, composed of keypoints and descriptors, provide an attractive drop-in replacement to their hand-crafted counterparts and have recently shown outstanding performance [12, 29, 14]. They can easily be sampled from dense features, are fast to predict and thus suitable for mobile deployment. CNN keypoint detections have also been shown to outperform classical methods, although they are notably difficult to learn. SuperPoint [12] learns from self-supervision, while DELF [28] employs an attention mechanism to optimize for the landmark recognition task.

Deep learning on mobile. While learning some building blocks of the localization pipeline improves performance and robustness, deploying them on mobile devices is a non-trivial task. Recent advances in multi-task learning allow to efficiently share compute across tasks without manual tuning [19, 8, 39], thus reducing the required network size. Distillation [16] can help to train a smaller network [31] from a larger one that is already trained. It is however usually not applied in a multi-task setting.

To the best of our knowledge, our approach is the first of its kind that combines the developments in the aforementioned fields, notably scalable localization, learned features and multi-task learning, while retaining both efficiency and robustness. The method we propose seeks to leverage the synergies of those algorithms to deliver a competitive large-scale localization solution and bring this technology closer to real-time, online applications with constrained resources.

3. Hierarchical Localization

We aim at maximizing the robustness of the localization while retaining tractable computational requirements. Our method is loosely based on the hierarchical localization framework first introduced in [32]. For the sake of completeness we present its overview in this section.

Prior retrieval. We perform a coarse search at the map level by matching the query with the database images using their global descriptors. The k-nearest neighbors (NN), called prior frames, represent candidate locations in the map. This search is efficient given that there are far fewer database images than points in the 3D model.

Covisibility clustering. The prior frames are clustered based on the 3D structure that they co-observe. This amounts to finding connected components, called places, in the covisibility graph that links database images to 3D points in the model.

Local feature matching. For each place, we successively match the 2D keypoints detected in the query image to the 3D points contained in the place, and attempt to estimate a 6-DoF pose with a PnP geometric consistency check within a RANSAC scheme. This local search is also efficient as the number of 3D points considered is significantly lower for the place than for the whole model. The algorithm stops as soon as a valid pose is estimated.

Discussion. In the work of [32], a large state-of-the-art network for image retrieval, NetVLAD, is distilled into a smaller model, coined MobileNetVLAD (MNV). This helps to achieve given runtime constraints while partly retaining the accuracy of the original model. The local matching step is however based on SIFT [23], which is expensive to compute and generates a large number of features, making this step particularly expensive. As such, while this method exhibits good performance in small-scale environments, it does not generalize well to large-scale denser models. Additionally, SIFT features have been shown to be outperformed by recent learned features, especially in case of large illumination changes [14, 29, 12, 27]. Lastly, a significant part of the computation of local and global descriptors is redundant, as they are both based on the image low-level features. The heterogeneity of hand-crafted features and CNN image retrieval is thus computationally suboptimal and could be critical on resource-constrained platforms. We address these issues and achieve improved robustness, scalability and efficiency.

4. Proposed Approach

In this section, we describe the proposed HF-Net architecture and training in more details. We start by motivating the use of learned features together with a homogeneous network structure for HF-Net. Then, we introduce in Section 4.1 the architecture and explain our design choices. Finally, Section 4.2 describes our novel training procedure.

We start by proposing learned features as a replacement for hand-crafted features like SIFT. Recent methods like SuperPoint [12] have shown to outperform this popular baseline in terms of keypoint repeatability and descriptor matching, which are both critical for localization. Some learned features are additionally significantly sparser than SIFT, thus reducing the number of keypoints to be matched and speeding up the matching step. We show in Section 5.1 that a combination of state-of-the-art networks in image retrieval and local features naturally achieves state-of-the-art localization. This approach particularly excels in extremely challenging conditions, such as night-time queries, outperforming competitive methods by a large margin along with a smaller 3D model size.

While the inference of such neural networks is significantly faster than computing SIFT on GPUs, it still remains a large computational bottleneck for the proposed hierarchical localization approach. With the goal of improving the ability of this method to run online on mobile devices, we introduce here a novel neural network for hierarchical features, HF-Net, enabling an efficient coarse-to-fine localization. It detects keypoints and computes local and global descriptors in a single shot, thus maximizing sharing of computations, but retaining performance of a larger baseline network.

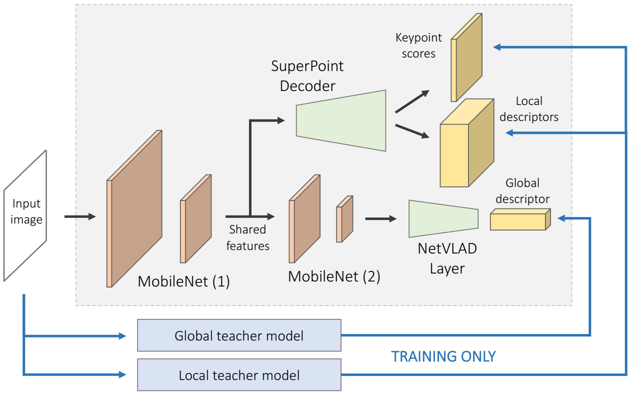

Figure 2. **The HF-Net architecture.** Our novel architecture takes image as an input and provides three outputs in a single shot. It computes a global descriptor using a NetVLAD layer. Then, it produces a map with keypoints detection scores and dense keypoints descriptors. All three heads are trained in a single training procedure with automatic loss weighting [19] from two teacher networks (marked in blue).

4.1. HF-Net architecture

Convolutional neural networks exhibit a hierarchical structure by design. This paradigm fits well the joint predictions of local and global features and comes at low additional runtime costs. The HF-Net architecture (Figure 2) is composed of a single encoder and three heads predicting: i) keypoint detection scores, ii) dense local descriptors and iii) a global image-wide descriptor. This sharing of compute is natural: in state-of-the-art image retrieval networks, the global descriptors are usually computed from the aggregation of local feature maps, which might be useful for the prediction of local features.

The encoder of HF-Net is a MobileNet [31] backbone, a popular architecture optimized for mobile inference. Similarly as MNV [32], the global descriptor is computed by a NetVLAD layer [2] on top of the last feature map of MobileNet. For the local features, the SuperPoint [12] architecture is appealing for its efficiency, as it decodes the keypoints and local descriptors in a fixed non-learned manner. This is much faster than applying transposed convolutions to upsample the features. It predicts dense descriptors which are fast to sample bilinearly and allows the runtime to be quasi-independent from the number of detected keypoints. On the other hand, patch-based architectures like LF-Net [29] apply a Siamese network to an image patch around each keypoint independently, resulting in a computational cost proportional to the number of detections.

For its efficiency and flexibility, we thus adopt the SuperPoint decoding scheme for keypoints and local descriptors. Note that this architecture is independent from the training process, and can be sparsely or densely supervised. The local feature heads branch out from the MobileNet encoder at an earlier stage than the global head, motivated by the requirement for a higher spatial resolution in order to retain spatially discriminative features. It also stems from the intuition that local features are on a lower semantic level than image-wide descriptors, and thus require fewer convolutions.

4.2. Training process

Data scarcity. Local and global descriptors are often trained with metric learning using ground truth positive and negative pairs of local patches and full images. These ground truth correspondences are particularly difficult to obtain at the scale required to train large CNNs. While global supervision naturally emerges from local correspondences, there is currently no such dataset that simultaneously i) exhibits a sufficient perceptual diversity at the global image level, *e.g.* with various conditions such as day, night, seasons, and ii) contains ground truth local correspondences between matching images. These correspondences are often recovered from the dense depth [29] computed from a SfM model [36, 38], which is intractable to build at the scale required by image retrieval.

Data augmentation. Self-supervised methods that do not rely on correspondences, such as SuperPoint, require heavy data augmentation, which is key for the local descriptor to learn a proper invariance. While data augmentation often captures well the variations in the real world at the local level, it can break the global consistency of the image and make the learning of the global descriptor very challenging.

Multi-task distillation is our solution to this data problem. We propose to employ distillation to learn the representation directly from an off-the-shelf trained teacher model. This alleviates the above issues, with a simpler and more flexible training setup that allows the use of arbitrary datasets, as infinite amount of labeled data can be obtained from the inference of the teacher network. Directly learning to predict the output of the teacher network additionally eases the learning task, allowing to directly train a smaller student network. We note an interesting similarity with SuperPoint, whose detector is training by bootstrapping, supervised by itself through the different training runs. This process could also be referred as self-distillation, and shows the effectiveness of distillation as a practical training scheme.

The supervision of local and global features can originate from different teacher networks, resulting in a multi-task distillation training that allows to leverage state-of-the-art teachers. Furthermore, recent advances in multi-task learning enable to train a student that accurately and optimally copies all teachers without any manual tuning of weights balancing the loss [19].

More generally, our formulation of the multi-task distillation can be applied to any applications that requires multiple predictions while remaining computationally efficient, particularly in settings where ground truth data for all tasks is expensive or impossible to collect.

5. Experiments

In this section, we present experimental evaluations of the building blocks of HF-Net and of the network as a whole. We want to prove its applicability to large-scale localization problems in challenging conditions while remaining computationally tractable. We first perform in Section 5.1 a thorough evaluation of current top-performing classical and learning-based methods for local feature detection and description. Our goal is to explain how these insights influenced the design choices of HF-Net presented in Section 5.2. We then evaluate in Section 5.3 our method on challenging large-scale localization benchmarks [34] and demonstrate the advantages of the coarse-to-fine localization paradigm. To address our real-time localization focus, we conclude with runtime considerations in Section 5.4.

5.1. Local features evaluation

We start our evaluation by investigating the performance of local matching methods under different settings on two datasets, HPatches [4] and SfM [29], that provide dense ground truth correspondences between image pairs for both 2D and 3D scenes.

Datasets. HPatches [4] contains 116 planar scenes containing illumination and viewpoint changes with 5 image pairs per scene and ground truth homographies. SfM is a dataset built by [29] composed of photo-tourism collections collected by [15, 42]. Ground truth correspondences are obtained from dense per-image depth maps and relative 6-DoF poses, computed using COLMAP [36]. We select 10 sequences for our evaluation and for each randomly sample 50 image pairs with a given minimum overlap. A metric scale cannot be recovered with SfM reconstruction but is important to compute localization metrics. We therefore manually label each SfM model using metric distances measured in Google Maps.

Metrics. We compute and aggregate pairwise metrics defined by [12] over all pairs for each dataset. For the detectors, we report the repeatability and localization error of the keypoint locations. Both are important for visual localization as they can impact the number of inlier matches, the reliability of the matches, but also the quality of the 3D model. We compute nearest neighbor matches between descriptors and report the mean average precision and the matching score. The former reflects the ability of the method to reject spurious matches. The latter assesses the quality of the detector and the descriptor together. We also compute the recall of pose estimation, either a homography for HPatches or a 6-DoF pose for the SfM dataset, with thresholds of 3 pixels and 3 meters respectively.

Methods. We evaluate the classical detectors Difference of Gaussian (DoG) and Harris [13] and the descriptor Root-SIFT [3]. For the learning-based methods, we evaluate the detections and descriptors of SuperPoint [12] and LF-Net [10]. We additionally evaluate a dense version of DOAP [14] and the feature map `conv3_3` of NetVLAD [2] and use SuperPoint detections for both. More details are provided in the appendix.

Detectors. We report the results in Table 1. Harris exhibits the highest repeatability but also the highest localization error. Conversely, DoG is less repeatable but has the lowest error, likely due to the multi-scale detection and pixel refinement. SuperPoint seems to show the best trade-off between repeatability and error.

Descriptors. DOAP outperforms SuperPoint on most metrics. NetVLAD shows good pose estimation but poor matching precision on SfM, which is disadvantageous when the number of keypoints is limited or the inlier ratio important, *e.g.* for localization. Overall, it stands that learned features outperform hand-crafted ones.

	HPatches		**SfM**	
	Rep.	MLE	Rep.	MLE
DoG	0.438	1.00	0.284	1.20
Harris	0.531	1.18	0.511	1.46
SuperPoint	0.496	1.04	0.508	1.46
LF-Net	0.466	1.14	0.448	1.46

Table 1. **Evaluation of the keypoint detectors.** The repeatability (rep.) and mean localization error (MLE) are reported for the HPatches and SfM datasets.

	HPatches			**SfM**		
(keypoints / descriptors)	Homography	MS	mAP	Pose	MS	mAP
Root-SIFT	0.707	0.301	0.592	0.713	0.192	0.224
LF-Net	0.712	0.303	0.515	0.631	0.205	0.186
SuperPoint	0.822	0.456	0.797	0.744	0.337	0.403
Harris / SuperPoint	0.662	0.425	0.702	0.620	0.325	0.265
SuperPoint / DOAP	-	-	-	0.767	0.363	0.466
SuperPoint / NetVLAD	0.67	0.460	0.720	0.751	0.320	0.262

Table 2. **Evaluation of the local descriptors.** For both datasets, the matching score (MS) and mean Average Precision (mAP) are reported, in addition to the homography correctness for HPatches and the pose accuracy for the SfM dataset.

Interestingly, SuperPoint descriptors perform poorly when applied to Harris detections, although the latter is also a corner detector with high repeatability. This hints that learned descriptors can be highly coupled with the corresponding detections.

LF-Net and SIFT, both multi-scale approaches with subpixel detection and patch-based description, are outperformed by dense descriptors like DOAP and SuperPoint. A simple representation trained with the right supervision can thus have more effect than a complex and computational-heavy architecture. We note that SuperPoint requires significantly fewer keypoints to estimate a good quality pose, which brings considerable benefits for runtime-sensitive applications.

5.2. Implementation details

Motivated by the results presented in Section 5.1, this section briefly introduces the design and implementation of HF-Net. Below, we explain our choices of the distillation teacher models, training datasets and improvements to the baseline 2D-3D local matching.

Teacher models. In Section 5.1, we showed that both SuperPoint and DOAP deliver similarly high performance, superior to the conventional hand-crafted descriptors. We therefore evaluate their impact on the localization task in Section 5.3. Results show that the former is more robust to day-night appearance variations, as its training set included low-light data. We eventually chose it as the supervisor teacher network for the descriptor head of HF-Net.

Training data. In this work, we target urban environments in both day and night conditions. To maximize the performance of the student model on this data, we select training data that fits this distribution. We thus train on 185k images from the Google Landmarks dataset [28], containing a wide variety of day-time urban scenes, and 37k images from the night and dawn sequences of the Berkeley Deep Drive dataset [47], composed of road scenes with motion blur. We found the inclusion of night images in the training dataset to be critical for the generalization of the global retrieval head to night queries. For example, a network trained on day-time images only would easily confuse a night-time dark sky with a day-time dark tree. We also train with photometric data augmentation but use the targets predicted on the clean images.

Efficient hierarchical localization. In [32], the authors identified the local 2D-3D matching as the most demanding part of the pipeline. Our system significantly improves on the efficiency of their approach. As such, spurious local matches are filtered out using a modified ratio test that only applies if the first and second nearest neighbor descriptors correspond to observations of different 3D points, thus retaining more matches in highly covisible areas. In addition, the number of retrieved images in each cluster is truncated to a fixed value by discarding additional frames, reducing the 2D-3D matching runtime for easy queries without impacting difficult ones. We use efficient kd-trees for both global and local matching and the fast P3P-RANSAC implementation of Kneip *et al*. [21].

5.3. Large-scale localization

Now, taking into account the insights gathered in Sections 5.1 and 5.2, we present the evaluations of HF-Net on three challenging large-scale datasets introduced by [34].

Datasets. Each dataset is composed of a sparse SfM model built with a set of reference images. The Aachen Day-Night dataset based on [35] contains $4,328$ day-time database images from a European old town, and 824 and 98 queries taken in day and night conditions respectively. The RobotCar Seasons dataset [25] is a long-term urban road dataset that spans multiple city blocks. It is composed of $20,862$ overcast reference images and a total of $11,934$ query images taken in multiple conditions, such as sun, dusk, and night. Lastly, the CMU Seasons dataset [6] was recorded in urban and suburban environments over a course of 8.5 km. It contains $7,159$ reference images and $75,335$ query images recorded in different seasons. This dataset is of significantly lower scale as the queries are localized against isolated submodels containing around 400 images each.

Large scale model construction. SfM models built by COLMAP [36, 38] using RootSIFT are provided by the dataset authors. These are however not suitable when localizing with methods based on different feature detectors. We thus build new 3D models with keypoints detected by HF-Net. The process is as follows: i) We perform 2D-2D matching between reference frames using our features and an initial filtering ratio test, ii) The matches are further filtered using two-view geometry filtering within COLMAP, iii) 3D points are triangulated using the provided ground truth reference poses. Those steps result in a 3D model with the same scale and reference frame as the original one.

Comparison of model quality. The HF-Net Aachen model contains fewer 3D points ($684,990$ vs $1,899,775$ for SIFT) and fewer 2D keypoints per image ($2,576.0$ vs $10,229.5$ for SIFT). However, a larger ratio of the original 2D keypoints is matched (0.370 vs 0.249 for SIFT), and each 3D point is on average observed from more reference images. Matching a query keypoint against this model is thus more likely to succeed, showing that our feature network produces 3D models more suitable for localization.

Methods. We evaluate HF-Net combined with our hierarchical localization method. We also consider several localization baselines evaluated by the benchmark authors. Active Search (AS) [33] and City Scale Localization (CSL) [40] are both 2D-3D direct matching methods representing the current state-of-the-art in terms of accuracy. DenseVLAD [45] and NetVLAD [2] are image retrieval approaches. The recently-introduced Semantic Match Consistency (SMC) [44] relies on semantic segmentation for outlier rejection. It assumes known gravity direction and camera height and, for the RobotCar dataset, was trained on the evaluation data using ground truth semantic labels. We additionally introduce two baselines derived from our method. Both are too computationally-intensive for the applications that we target, although still faster than other baselines. They nevertheless illustrate the benefits of hierarchical localization. NV+SIFT uses NetVLAD for the global retrieval and RootSIFT as local features, and is an upper bound to the MNV+SIFT method of [32]. NV+SP uses NetVLAD retrieval and SuperPoint features and is an upper bound to our HF-Net.

Metrics used in this evaluation are defined by the benchmark [34]. We evaluate the pose recall at different position and orientation thresholds that depend on the sequences.

Overall results. Table 3 shows the localization results for the different methods. On the Aachen dataset, our proposed HF-Net outperforms all the realistic methods for both fine- and coarse-precision regimes. It performs particularly well on night-time images (Figure 4), where the performance drop w.r.t day-time queries is significantly smaller than for direct matching methods, which suffer from the increased ambiguity of the matches. On the RobotCar dataset, it performs similarly to other methods on the dusk sequence where the accuracy tends to saturate. In the more challenging sequences, image retrieval methods tend to work best, while direct matching methods perform poorly. Hierarchical localization gives better results than both in the fine-

	Aachen		**RobotCar**				**CMU**	
	day	night	dusk	sun	night	night-rain	urban	suburban
m	.25/.50/5.0	0.5/1.0/5.0	.25/.50/5.0	.25/.50/5.0	.25/.50/5.0	.25/.50/5.0	.25/.50/5.0	.25/.50/5.0
deg	2/5/10	2/5/10	2/5/10	2/5/10	2/5/10	2/5/10	2/5/10	2/5/10
AS	57.3 / 83.7 / **96.6**	19.4 / 30.6 / 43.9	44.7 / 74.6 / **95.9**	25.0 / 46.5 / 69.1	0.5 / 1.1 / 3.4	**1.4** / 3.0 / 5.2	**55.2** / 60.3 / 65.1	20.7 / 25.9 / 29.9
CSL	52.3 / 80.0 / 94.3	24.5 / 33.7 / 49.0	**56.6 / 82.7 / 95.9**	**28.0** / 47.0 / 70.4	0.2 / 0.9 / 5.3	0.9 / 4.3 / 9.1	36.7 / 42.0 / 53.1	8.6 / 11.7 / 21.1
DenseVLAD	0.0 / 0.1 / 22.8	0.0 / 2.0 / 14.3	10.2 / 38.8 / 94.2	5.7 / 16.3 / 80.2	**0.9** / 3.4 / **19.9**	1.1 / 5.5 / **25.5**	22.2 / 48.7 / 92.8	9.9 / 26.6 / 85.2
NetVLAD	0.0 / 0.2 / 18.9	0.0 / 2.0 / 12.2	7.4 / 29.7 / 92.9	5.7 / 16.5 / **86.7**	0.2 / 1.8 / 15.5	0.5 / 2.7 / 16.4	17.4 / 40.3 / 93.2	7.7 / 21.0 / 80.5
HF-Net (ours)	**75.7** / **84.3** / 90.9	**40.8 /55.1 / 72.4**	22.1 / 69.0 / 94.4	26.5 / **58.5** / 86.1	0.7 / **4.6** / 14.8	**1.4** / **9.3** / 20.5	39.6 / **89.7 / 95.7**	**30.9 / 73.3 / 86.6**
NV+SP	79.7 / 88.0 / 93.7	40.8 / 56.1 / 74.5	22.8 / 70.1 / 96.2	26.3 / 66.1 / 92.6	1.8 / 11.9 / 31.3	1.8 / 12.3 / 26.6	40.4 / 90.6 / 97.5	31.5 / 75.6 / 91.0
NV+SIFT	82.8 / 88.1 / 93.1	30.6 / 43.9 / 58.2	55.6 / 83.5 / 95.3	46.3 / 67.4 / 90.9	4.1 / 9.1 / 24.4	2.3 / 10.2 / 20.5	63.9 / 71.9 / 92.8	28.7 / 39.0 / 82.1
SMC	-	-	53.8 / 83.0 / 97.7	46.5 / 74.6 / 95.9	6.2 / 18.5 / 44.3	8.0 / 26.4 / 46.4	75.2 / 82.1 / 87.7	44.6 / 53.9 / 63.5

Table 3. **Evaluation of the localization** on the Aachen Day-Night, RobotCar Seasons, and CMU Seasons datasets. We treat NV+SP as our upper bound for HF-Net. Additionally, we include SMC which is about two orders of magnitude slower and was fine-tuned on the RobotCar dataset.

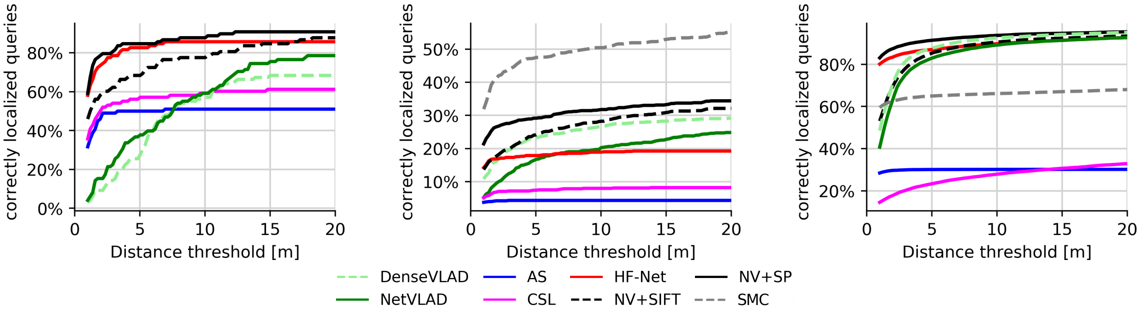

Figure 3. **Cumulative distribution of position errors** for the Aachen night (left), RobotCar night (center) and CMU suburban (right) datasets. On Aachen HF-Net performs close to our upper bound NV+SP and is superior than other global retrieval and matching based approaches. On RobotCar, HF+Net performs worse than NV+SP, which suggests a limitation of a distilled global descriptor. SMC performs particularly well, it was however fine-tuned using images from this dataset. On CMU, the hierarchical localization shows a significant boost over other methods, particularly for strict distance thresholds.

precision regime while maintaining a decent recall for larger error thresholds. On the CMU datasets, HF-Net exhibits the highest recall on the medium- and coarse-precision regimes, even compared to SMC. Cumulative plots for the three most challenging sequences are shown in Figure 3. Together, HF-Net and NV+SP set new state-of-the-art performance in the Aachen and CMU datasets for both efficient and expensive computation regimes.

Upper bounds We observe that NV+SIFT consistently outperforms AS, although both methods are based on RootSIFT features. This shows that our hierarchical approach with a coarse initial prior brings significant benefits, especially in challenging conditions where image-wide information helps disambiguate matches. Comparing NV+SP and NV+SIFT, SuperPoint brings significant performance gains over SIFT in the coarse regime as well as in challenging sequences. We however observe a drop in the fine-precision regime, which might be caused, as highlighted in Section 5.1, by the lower localization accuracy of the SuperPoint keypoints compared to DoG.

	Thresh	NV+SP	NV+HF-Net	NV+DOAP	HF-Net
Day	0.25m	79.7	81.2	80.0	75.7
	0.5m	88.0	88.2	88.5	84.3
	5m	93.7	94.2	93.3	90.9
Night	0.5m	40.8	40.8	34.7	40.8
	1m	56.1	56.1	52.0	55.1
	5m	74.5	76.5	72.4	72.4

Table 4. **Ablation study on the Aachen dataset.** We evaluate the localization with NetVLAD and different local features.

Ablation study permits us to understand the performance of each method better. The results are presented in Table 4. Interestingly, local HF-Net descriptors perform better than the SuperPoint model that was used to train them (compare NV+SP with NV+HF). This demonstrates the benefits of multi-task distillation. Comparing NV+DOAP and NV+SP shows that DOAP performs slightly better during the day, but it is significantly worse at night. Finally, comparison of HF-Net with NV+HF-Net shows that HF-Net global descriptors have a somewhat limited capacity compared to the original NetVLAD and are limiting the performance particularly during the day.

Failure cases Overall, HF-Net performs very similarly to NV+SP, except in the RobotCar night sequences, where the distilled global retrieval performs poorly on blurry low-quality images. This highlights a clear limitation of our approach: on large, self-similar environment, the model capacity of HF-Net becomes the limiting factor. A complete failure of the global retrieval directly translates into a failure of the hierarchical localization. This contrasts with [32], which reported limitations coming from local SIFT features, but with a model of smaller scale.

5.4. Runtime evaluation

The proposed HF-Net algorithm was developed keeping the computational constraints in mind. We thus provide in-

		HF-Net (ours)						AS			NV+SIFT			
		HF-Net	Global search	Covisibility	Local search	PnP+RANSAC	**Total**	SIFT	Loc.	**Total**	SIFT	NetVLAD	Loc.	**Total**
Aachen	Day	15	51	5	163	9	**243**	263	112	**375**	263	92	1264	**1356**
	Night	15	52	5	170	18	**260**	263	132	**395**	263	92	1563	**1655**
RobotCar	Dusk	19	53	1	58	4	**135**	189	283	**472**	189	91	294	**385**
	Night	19	55	1	103	38	**216**	189	1021	**1210**	189	91	554	**645**

Table 5. **Timings [ms]** of HF-Net, AS, and NV+SIFT. CNN inference (HF-Net and NetVLAD) and SIFT extraction are performed on the GPU, and all other operations on the CPU. The detailed runtime of HF-Net shows that most time is spent on the local search. In case of a low success rate, such as for RobotCar Night, the runtime of AS grows significantly. For NV+SIFT, the dense Aachen SIFT model returns a large number of points per cluster, which dramatically increases the execution time. SIFT extraction, required by both AS and NV+SIFT, is expensive even on GPU and intractable on CPU (exceeds 3500ms).

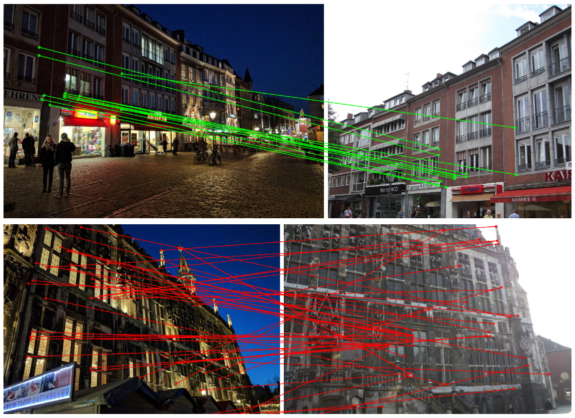

Figure 4. **Successful (top) and failed (bottom) HF-Net queries**. The top query was successfully retrieved by a HF-Net global descriptor despite the appearance change. Then, it was successfully matched by HF-Net local descriptors. SIFT descriptors failed in this case. The bottom query failed as local matching did not find enough consistent correspondences because of visual aliasing.

sights into the runtime of our method and compare it with baselines presented in Section 5.3. The evaluations were performed on a PC equipped with an Intel Core i7-7820X CPU (3.60GHz) CPU, 32GB of RAM and an NVIDIA GeForce GTX 1080 GPU.

HF-Net and baseline methods The results presented in Table 5 demonstrate that our hierarchical coarse-to-fine approach scales well on large-scale datasets. The global retrieval step permits to successfully narrow down the set of potential candidate correspondences, which enables tractable 2D-3D matching. The subsequent geometric verification of matches is fast, as HF-Net local descriptors yield a high inlier ratio and, consequently, few RANSAC iterations. Overall, HF-Net achieves the frequency of about 4 Hz, which makes it suitable for real-time applications.

When compared to baseline methods, HF-Net delivers lower runtimes, particularly for large-scale models (*e.g.* RobotCar), where global retrieval cuts the cost of search in the entire model. Furthermore, we can notice that the time of 2D-3D matching strongly depends on the dataset – the denser the covisibility graph is, the more 3D points are retrieved and matched per prior frame, which increases the runtime. The sparser HF-Net models yield clusters with fewer 3D points, whereas NV+SIFT struggles with a dense Aachen model. Furthermore, a large failure rates causes the AS runtime to grow quickly, while HF-Net is less affected.

Finally, we compare the inference time of HF-Net (less than 20ms) with the NV+SP baseline (about 90ms for NetVLAD and 25ms for SuperPoint). The distilled and homogeneous structure of HF-Net saves computation time even compared to the original SuperPoint detector and descriptor alone.

Accuracy-computation trade-off Large-scale evaluations of HF-Net show that NN searches, particularly the local 2D-3D matching, consume the majority of time. We want, however, to emphasize that there exists an easily-tunable trade-off between runtime and accuracy. The options include pruning of the local graph to limit the 2D-3D matching time or using approximate kd-trees. They permit to achieve further speed-ups with a marginal performance loss.

6. Conclusion

In this paper, we have presented HF-Net – a hierarchical localization approach that is designed for robust but computationally tractable localization. Our method follows a coarse-to-fine localization paradigm. First, it performs a global image retrieval to obtain a set of prior frames. The retrieved frames are clustered into places using the covisibility graph of a 3D database model. Then, we perform local 2D-3D matching within the candidate places to obtain an accurate 6-DoF estimate of the camera pose.

Our method is centered around HF-Net, a novel network architecture that permits to predict feature locations, local features and a global descriptor in a single shot. We propose to train it using a multi-task distillation procedure, with close-to-optimal loss weighting and carefully selected teacher networks. Our localization solution outperforms state-of-the-art methods on several large-scale 6-DoF localization benchmarks, that include day-night queries and substantial appearance variations across weather conditions and seasons. Furthermore, the lightweight architecture of HF-Net reduces its computational requirements and makes it suitable for real-time operation. Those results validate our approach and demonstrate the usefulness of a monolithic CNN to perform hierarchical localization.

Appendix

We provide here additional experiment details and qualitative results.

A. HF-Net Implementation

A.1. Network Architecture

HF-Net is built on top of a MobileNetV2 [31] encoder with depth multiplier 0.75. The local heads are identical to the original SuperPoint [12] and branch off at the layer 7. The global head is composed of a NetVLAD layer [2] and a dimensionality reduction, implemented as a multiplication with a learnable matrix, in order to match the dimension of the target teacher descriptor. The global head is appended to the MobileNet layer 18. The detailed architecture is shown in Figure 5.

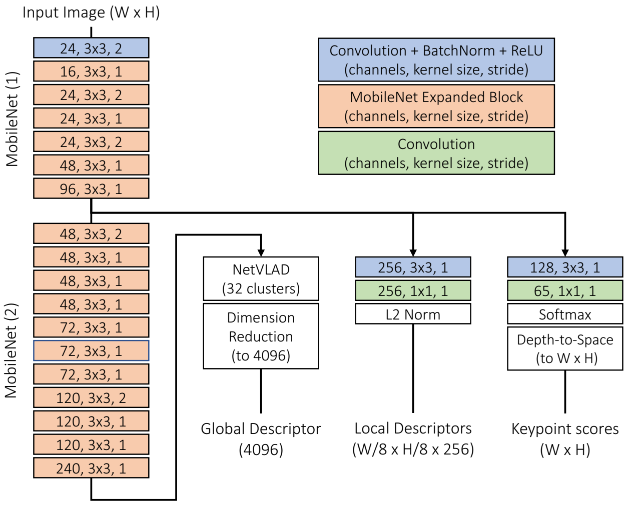

Figure 5. **Detail of the HF-Net architecture**, consisting of a MobiletNet encoder and three heads predicting a global descriptor, a dense local descriptor map, and keypoint scores.

A.2. Training details

The images from both Google Landmarks [28] and Berkeley Deep Drive [47] are resized to 640×480 and converted to grayscale. We found RGB to be detrimental to the performance of the local feature heads, most likely because of the limited bandwidth of the encoder. As photometric data augmentation, we apply Gaussian noise, motion blur in random directions, and random brightness and contrast changes.

The losses of the global and local descriptors are the L2 distances with their targets. For the keypoints, we apply the cross-entropy with the target probabilities (soft labels). We found hard labels to perform poorly, likely due to their sparsity and the smaller size of the student network. The three losses are aggregated using the multi-task learning scheme of Kendall *et al.* [19].

The MobileNet layers are initialized with weights pretrained on ImageNet [11]. The network is implemented with Tensorflow [1] and trained for 85k iterations with the RMSProp optimizer [43] and a batch size of 32. We use an initial learning rate of 10^{-3}, which is successively divided by ten at iterations 60k and 80k.

B. Local Feature Evaluation

B.1. Setup

The images of both HPatches [4] and SfM [29] datasets are resized so that their largest dimension is 640 pixels. The metrics are computed on image pairs and follow the definitions of [12, 29]. A keypoint k_1 in an image is deemed correct if its reprojection $\hat{k}_1$ in a second image lies within a given distance threshold ϵ to a second detected keypoint k_2. Additionally, k_1 is matched correctly if it is correct and if k_2 is its nearest neighbor in descriptor space.

For HPatches, we detect 300 keypoints for both keypoint and descriptor evaluations, and set $\epsilon = 3$ pixels. The homography is estimated using the OpenCV function `findHomography` and considered accurate if the average reprojection error of the image corners is lower than 3 pixels. For the SfM dataset, due to the extensive texture, 1000 keypoints are detected. The keypoint and descriptor metrics use correctness thresholds ϵ of 3 and 5, respectively. The 6-DoF pose is estimated with the function `solvePnPRansac`, and deemed correct if its ground truth is within distance and orientation thresholds of 3 m and 1°, respectively.

For DoG, Harris [13], and SIFT [23], we use the implementations of OpenCV. For SuperPoint [12] and LF-Net [29], we use the implementations provided by the authors. For NetVLAD, we use the implementation of [10] and the original model trained on Pittsburgh30k. Dense descriptors are obtained by normalizing the feature map `conv3_3` before the ReLU activation. For DOAP [14], we use the trained model provided by the authors. As we are mostly interested in dense descriptors for run-time efficiency, we disable the spatial transformer and enable padding in the last layer, thus producing a feature map four times smaller than the input image. We found the model trained on HPatches with spatial transformer to give the best results and thus only evaluate DOAP on the SfM dataset. As a post-processing, we apply Non-Maximum Suppression (NMS) with a radius of 4 to both Harris and SuperPoint. Sparse descriptors are sampled from the dense maps of SuperPoint, NetVLAD, and DOAP using bilinear interpolation.

B.2. Qualitative Results

We show in Figures 6 and 7 detected keypoints and their corresponding matches on the HPatches and SfM datasets, respectively.

C. Large-scale Localization

C.1. Model Quality

Extended statistics of models built with SIFT and HF-Net for the Aachen Day-Night, RobotCar Seasons, and CMU Seasons datasets, are provided in Table 6. We also report the track length, i.e. number of observation per 3D point, as defined by [37]. The metrics for the CMU dataset are aggregated over the models of the slices corresponding to the urban and suburban environments. For SIFT, some metrics cannot be computed on the CMU model as the keypoints that are not matched were not provided.

	Aachen		**RobotCar**		**CMU**	
	SIFT	HF	SIFT	HF	SIFT	HF
# 3D points	1,900K	685K	6,869K	2,956K	961K	597K
# Keypoints per image	10,230	2,576	4,409	970	-	1,446
Ratio of matched keypoints [%]	18.8	33.8	39.4	59.4	-	41.2
Track length	5.85	5.87	5.34	4.04	-	4.19

Table 6. **Statistics of 3D models** built with SIFT and HF-Net.

C.2. Implementation Details

We now provide additional details regarding the implementation of our hierarchical localization pipeline. For all datasets, we reduce the dimensionality of the global descriptors predicted by both NetVLAD and HF-Net to 1024 dimensions using PCA, whose parameters are learned on the reference images, independently for each dataset. A total of 10 prior frames are retrieved and clustered. The size of each cluster is limited to 5, additional frames are discarded. For both SuperPoint and HF-Net, NMS with radius 4 is applied to the detected keypoints in the query image and 2k of them are retained. When performing local matching, our modified ratio test uses a threshold of 0.9. PnP+RANSAC uses a threshold on the reprojection error of 10 pixels for Aachen, 5 pixels for CMU (due to the lower image size), and 12 pixels for RobotCar (due to the lower keypoint localization accuracy of SuperPoint and HF-Net). The estimated pose is deemed correct when the number of inliers is larger than a threshold, whose value is 12 for Aachen and CMU, and 15 for Robotcar.

C.3. Evaluation Process

The method and baselines introduced in this work are evaluated on all three datasets by the benchmark's authors [34], who also generated the plots shown in the main paper. For Active Search [33], City Scale Localization [40], DenseVLAD [45], and NetVLAD [2], we use the evaluation reported in the paper introducing the benchmark.

The evaluation of Semantic Match Consistency [44] (SMC) is the one reported in the original paper. We do not directly compare this method to the ones introduced in the present work, nor to the benchmark baselines, as SMC assumes a known camera height, and, more importantly, relies on a semantic segmentation CNN which was trained on the evaluation dataset of RobotCar. We emphasize that our HF-Net never encountered any test data during training, and that it was evaluated on the three datasets using the same trained model.

C.4. Qualitative Results

Visual results of HF-Net on the Aachen Day-Night, RobotCar Seasons, and CMU Seasons datasets are shown in Figures 8, 9, and 10, respectively. We additionally show comparisons with NV+SIFT in Figures 11 and 12.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016. 9

[2] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 4, 5, 6, 9, 10

[3] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 5

[4] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5, 9

[5] V. Balntas, S. Li, and V. Prisacariu. RelocNet: Continuous metric learning relocalisation using neural nets. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[6] A. Bansal, H. Badino, and D. Huber. Understanding how camera configuration and environmental conditions affect appearance-based localization. In *IEEE Intelligent Vehicles (IV)*, 2014. 6

[7] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF: Binary robust independent elementary features. In *European Conference on Computer Vision (ECCV)*, 2010. 2

[8] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning (ICML)*, 2018. 2

[9] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker. Universal Correspondence Network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2

[10] T. Cieslewski, S. Choudhary, and D. Scaramuzza. Data-efficient decentralized visual SLAM. In *International Conference on Robotics and Automation (ICRA)*, 2018. 5, 9

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 9

[12] D. DeTone, T. Malisiewicz, and A. Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Workshop on Deep Learning for Visual SLAM at Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 4, 5, 9

[13] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988. 5, 9

[14] K. He, Y. Lu, and S. Sclaroff. Local descriptors optimized for average precision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 5, 9

[15] J. Heinly, J. L. Schonberger, E. Dunn, and J.-M. Frahm. Reconstructing the World in Six Days. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5

[16] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015. 2

[17] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[18] A. Kendall, R. Cipolla, et al. Geometric loss functions for camera pose regression with deep learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[19] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 4, 9

[20] A. Kendall, M. Grimes, and R. Cipolla. PoseNet: a convolutional network for real-time 6-DOF camera relocalization. In *International Conference on Computer Vision (ICCV)*, 2015. 2

[21] L. Kneip, D. Scaramuzza, and R. Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 6

[22] S. Leutenegger, M. Chli, and R. Y. Siegwart. BRISK: Binary robust invariant scalable keypoints. In *International Conference on Computer Vision (ICCV)*, 2011. 2

[23] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 3, 9

[24] S. Lynen, T. Sattler, M. Bosse, J. A. Hesch, M. Pollefeys, and R. Siegwart. Get out of my lab: Large-scale, real-time visual-inertial localization. In *Robotics: Science and Systems (RSS)*, 2015. 1, 2

[25] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017. 6

[26] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt. Scalable 6-DOF localization on mobile devices. In *European Conference on Computer Vision (ECCV)*, 2014. 2

[27] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[28] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6, 9

[29] Y. Ono, E. Trulls, P. Fua, and K. M. Yi. LF-Net: Learning local features from images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1, 2, 3, 4, 5, 9

[30] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 2

[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. MobileNetV2: inverted residuals and linear bottlenecks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4, 9

[32] P.-E. Sarlin, F. Debraine, M. Dymczyk, R. Siegwart, and C. Cadena. Leveraging deep visual descriptors for hierarchical efficient localization. In *Conference on Robot Learning (CoRL)*, 2018. 1, 2, 3, 4, 6, 7

[33] T. Sattler, B. Leibe, and L. Kobbelt. Improving image-based localization by active correspondence search. In *European Conference on Computer Vision (ECCV)*, 2012. 6, 10

[34] T. Sattler, W. Maddern, A. Torii, J. Sivic, T. Pajdla, M. Pollefeys, and M. Okutomi. Benchmarking 6DOF urban visual localization in changing conditions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4, 6, 10

[35] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt. Image retrieval for image-based localization revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 6

[36] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 5, 6

[37] J. L. Schönberger, H. Hardmeier, T. Sattler, and M. Pollefeys. Comparative evaluation of hand-crafted and learned local features. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 10

[38] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 4, 6

[39] O. Sener and V. Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[40] L. Svärm, O. Enqvist, F. Kahl, and M. Oskarsson. City-scale localization for cameras with known vertical direction. *IEEE PAMI*, 39(7):1455–1461, 2017. 1, 2, 6, 10

[41] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii. InLoc: Indoor visual localization with dense matching and view synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[42] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. YFCC100M: The new data in multimedia research. *arXiv:1503.01817*, 2015. 5

[43] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012. 9

[44] C. Toft, E. Stenborg, L. Hammarstrand, L. Brynte, M. Pollefeys, T. Sattler, and F. Kahl. Semantic match consistency for long-term visual localization. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 6, 10

[45] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6, 10

[46] T. Weyand, I. Kostrikov, and J. Philbin. PlaNet - photo geolocation with convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[47] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv:1805.04687*, 2018. 6, 9

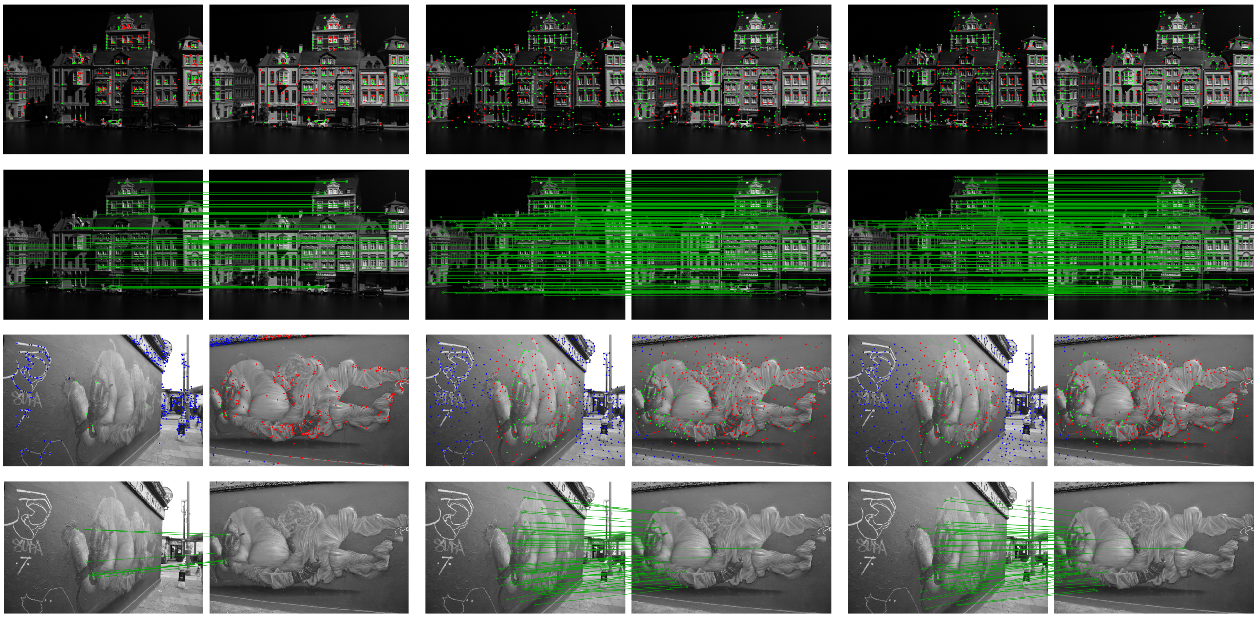

Figure 6. **Qualitative results on the HPatches dataset.** Keypoints (green if repeatable, red if not repeatable, blue if not visible in the other image) and inlier matches are shown for SIFT (left), SuperPoint (center) and HF-Net (right).

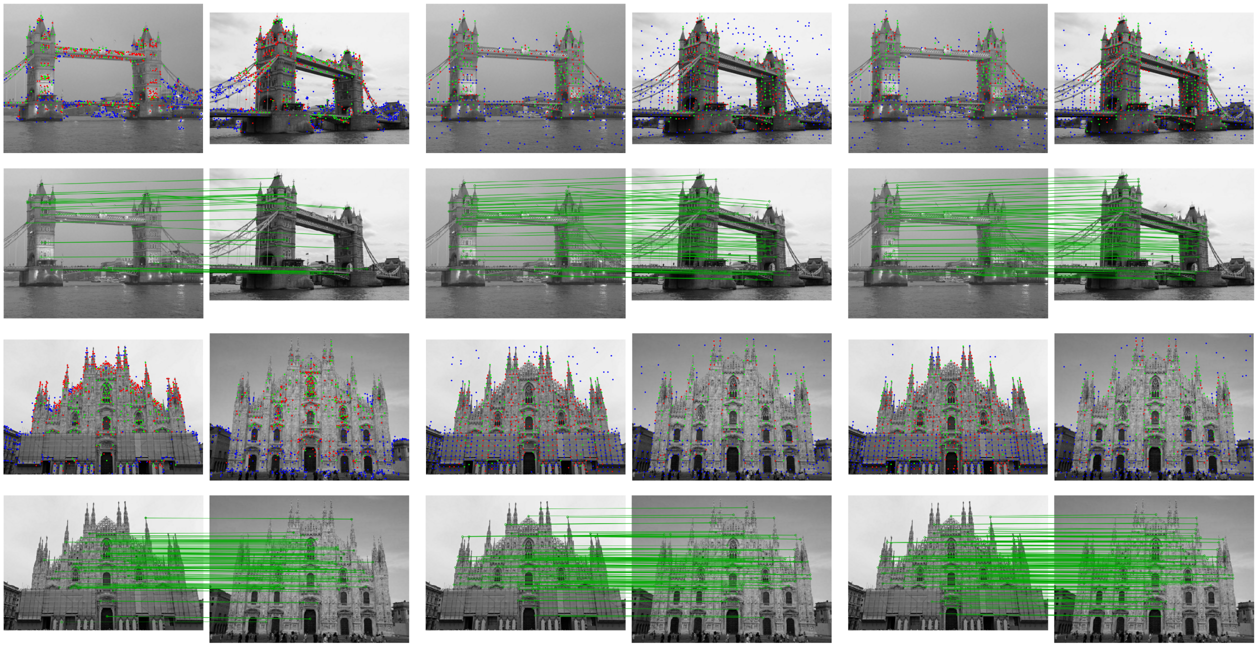

Figure 7. **Qualitative results on the SfM dataset** for SIFT (left), SuperPoint (center) and HF-Net (right).

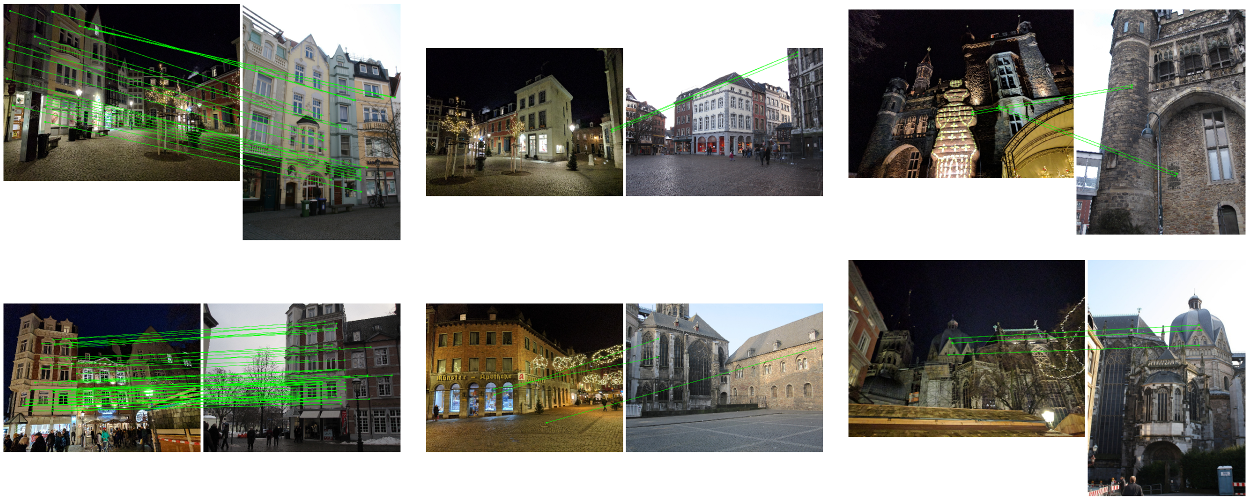

Figure 8. **Localization with HF-Net on Aachen night.** For each image pair, the left image is the query and the right image is the retrieved database image with the most inlier matches, as returned by PnP+RANSAC. We show challenging successful queries (left), failed queries due to an incorrect global retrieval (center), and failed queries due to incorrect or insufficient local matches (right).

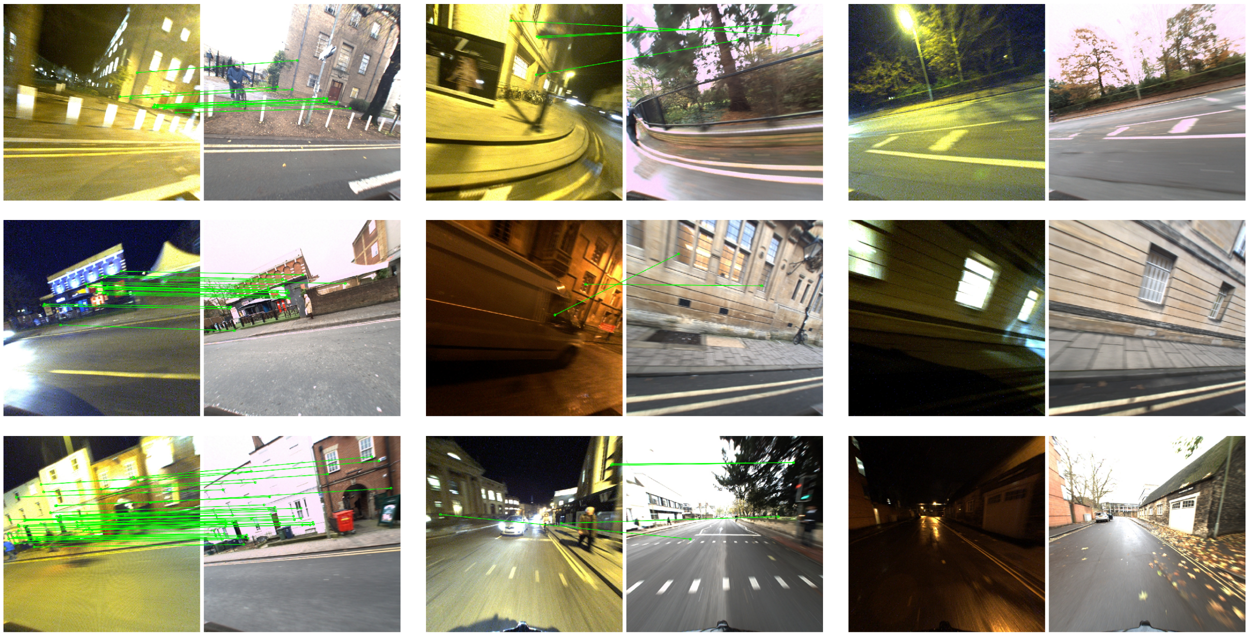

Figure 9. **Localization with HF-Net on RobotCar night and night-rain.** For each image pair, the left image is the query and the right image is the retrieved database image with the most inlier matches, as returned by PnP+RANSAC. We show challenging successful queries (left), failed queries due to an incorrect global retrieval (center), and failed queries due to insufficient local matches (right).

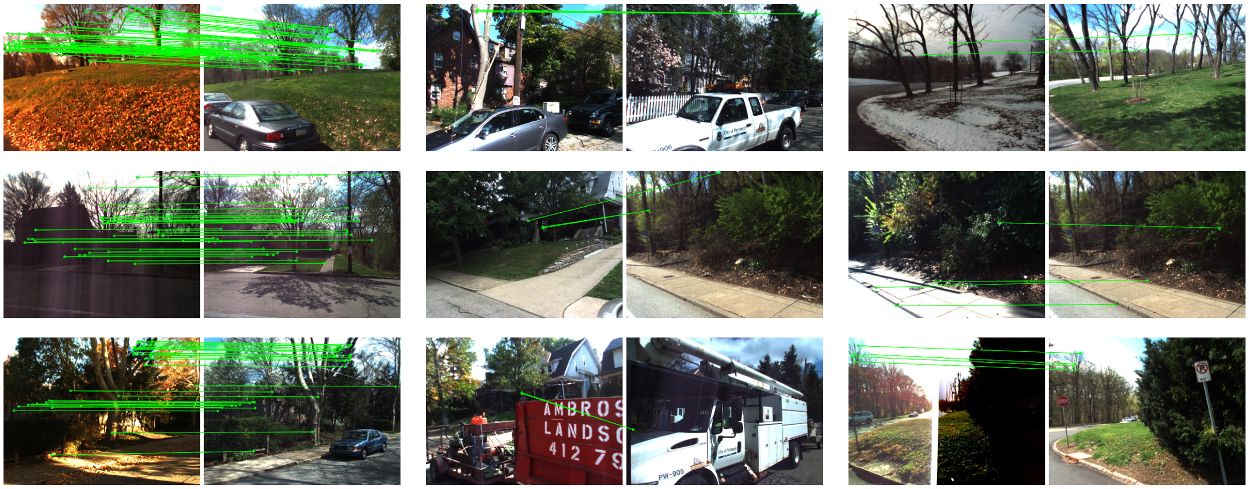

Figure 10. **Localization with HF-Net on CMU suburban.** For each image pair, the left image is the query and the right image is the retrieved database image with the most inlier matches, as returned by PnP+RANSAC. We show challenging successful queries (left), failed queries due to an incorrect global retrieval (center), and failed queries due to insufficient local matches (right).

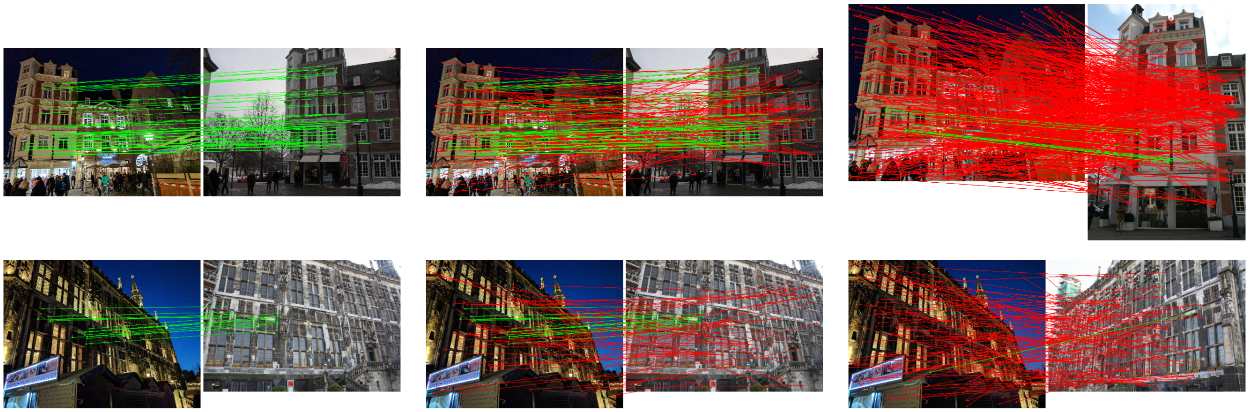

Figure 11. **Comparison between HF-Net and NV+SIFT on Aachen night.** Each row corresponds to one query for which HF-Net returns the correct location but NV+SIFT fails. We show the matches with one retrieved database image, labeled by PnP+RANSAC as inliers (green) and outliers (red). We show the inliers of HF-Net (left), all the matches of HF-Net (center), and all the matches of NV+SIFT (right). HF-Net generates significantly fewer matches than SIFT, thus reducing the computational footprint of the local matching. At the same time, more of its matches are inliers, increasing the robustness of the localization. The higher inlier ratio reduces the number of required RANSAC iterations.

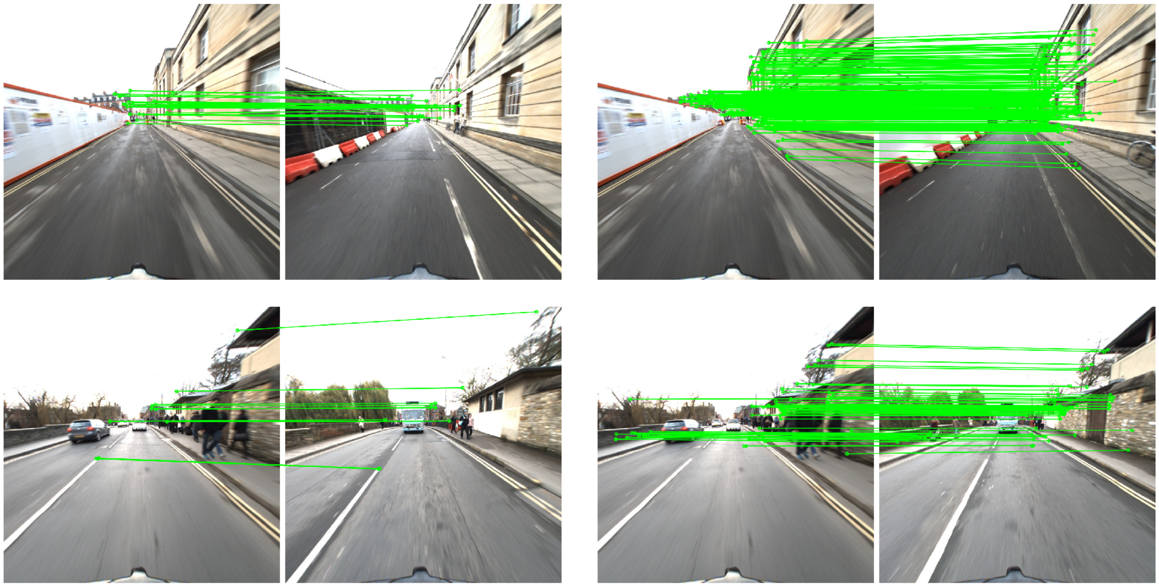

Figure 12. **Comparison between HF-Net and NV+SIFT on RobotCar dusk.** We show inlier matches of HF-Net (left) and NV+SIFT (right) on two queries (top and bottom) for which the pose estimated by HF-Net is less accurate than the one returned by NV+SIFT. In both cases, the global retrieval of HF-Net is less accurate than that of NV and local features and consequently the local matching is less robust to the extensive radial blur. This translates in matches with 3D points that are further away from the camera. The estimated pose is thus less constrained and less accurate than when estimated by NV+SIFT. This partially explains the poor performance of HF-Net in the fine-precision regime.