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 and sets a new state-of-the-art on two challenging benchmarks for large-scale localization.[1]

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 [7, 29, 31, 5] and consumer devices with augmented reality features [30, 22], 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.

Robustness to such large variations 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.

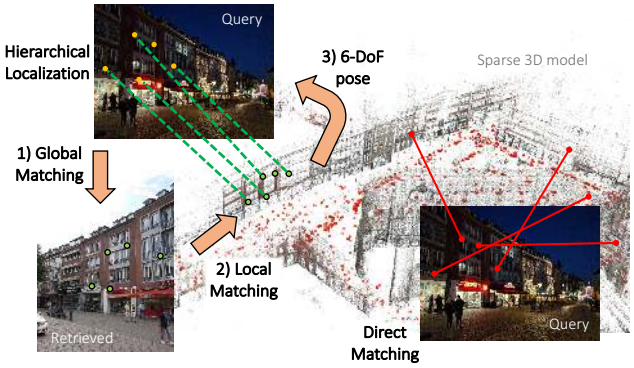

Figure 1. **Hierarchical localization.** A global search first retrieves candidate images, which are subsequently matched using powerful local features to estimate an accurate 6-DoF pose. This two-step process is both efficient and robust in challenging situations.

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 [48, 51, 41], or optimized for efficiency but fragile [27]. In both cases, the robustness of classical localization methods is limited by the poor invariance of hand-crafted local features [8, 26]. Recent features emerging from convolutional neural networks (CNN) exhibit unrivalled robustness at a low compute cost [12, 13, 32]. They have been, however, only recently [49] applied to the visual localization problem, and only in a dense, expensive manner. Learned sparse descriptors [12, 36] 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, but are not competitive in terms of accuracy. The benefits of an intermediate retrieval step have been demonstrated earlier [40], but fall short of reach-

[1]Code available at https://github.com/ethz-asl/hf_net

ing the scalability required by city-scale localization.

In this paper, we propose to leverage recent advances in learned features to bridge the gap between robustness and efficiency in the hierarchical localization paradigm. Similar to how humans localize, we employ a natural coarse-to-fine pose estimation process which leverages both global descriptors and local features, and scales well with large environments (Figure 1). We show that learned descriptors enable unrivaled robustness in challenging conditions, while learned keypoints improve the efficiency in terms of compute and memory thanks to their higher repeatability. To further improve the efficiency of this approach, we propose a Hierarchical Feature Network (HF-Net), a CNN that jointly estimates local and global features, and thus maximizes the sharing of computations. We show how such a compressed model can be trained in a flexible way using multitask distillation. By distilling multiple state-of-the-art predictors jointly into a single model, we obtain an incomparably fast, yet robust and accurate, localization. Such heterogenous distillation is applicable beyond visual localization to tasks that require both multimodal expensive predictions and computational efficiency. Overall, our contributions are as follows:

- We set a new state-of-the-art in several public benchmarks for large-scale localization with an outstanding robustness in particularly challenging conditions;
- We introduce HF-Net, a monolithic neural network which efficiently predicts hierarchical features for a fast and robust localization;
- We demonstrate the practical usefulness and effectiveness of multitask distillation to achieve runtime goals with heterogeneous predictors.

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 [48, 51, 41, 25, 49]. These methods are able to estimate accurate poses, but often rely on exhaustive matching and are thus compute intensive. 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 [42]. Some approaches directly regress the pose from a single image [6, 20], but are not competitive in term of accuracy [44]. Image-based methods are related to image retrieval [1, 52, 53] and are only able to provide an approximate pose up to the database discretization, which is not sufficiently precise for many applications [42, 49]. They are however significantly more robust than direct local matching as they rely on the global image-wide information. This 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 [8, 24, 37]. These improve the runtime on mobile devices but further impair the robustness of the localization, limiting their operations to stable conditions [27]. Hierarchical localization [19, 30, 40] takes a different approach by dividing the problem into a global, coarse search followed by a fine pose estimation. Recently, [40] 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 [10, 13, 35, 38] and localization [49, 42]. Matching dense features is however intractable with limited computing 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, 36, 16]. 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 [34] 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 [21, 9, 47], thus reducing the required network size. Distillation [18] can help to train a smaller network [39, 55, 56] from a larger one that is already trained, but is usually not applied in a multi-task setting.

To the best of our knowledge, our approach is the first of its kind that combines advances in the aforementioned fields to optimize for both efficiency and robustness. The proposed method seeks to leverage the synergies of these algorithms to deliver a competitive large-scale localization solution and bring this technology closer to real-time, online applications with constrained resources.

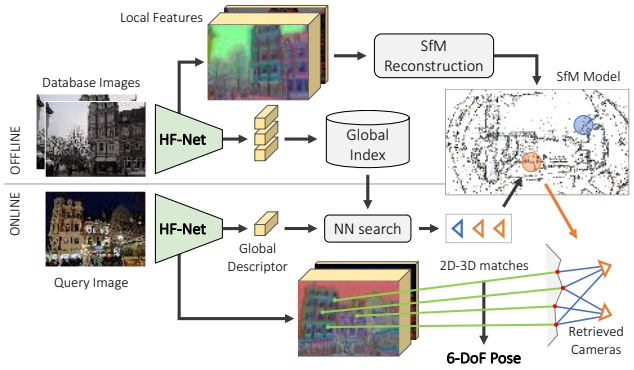

Figure 2. **The hierarchical localization with HF-Net** is significantly simpler than concurrent approaches [41, 48], yet more robust, accurate, and efficient.

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 [40], which we summarize here.

Prior retrieval. A coarse search at the map level is performed by matching the query with the database images using 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 SfM 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 [23] geometric consistency check within a RANSAC scheme [14]. This local search is also efficient as the number of 3D points considered is significantly lower in the place than in the whole model. The algorithm stops as soon as a valid pose is estimated.

Discussion. In the work of [40], a large state-of-the-art network for image retrieval, NetVLAD [1], is distilled into a smaller model, 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 [26], which is expensive to compute and generates a large number of features, making this step particularly expensive. While this method exhibits good performance in small-scale environments, it does not scale well to larger, denser models. Additionally, SIFT is not competitive with recent learned features, especially under large illumination changes [16, 36, 12, 32]. Lastly, a significant part of the computation of local and global descriptors is redundant, as they are both based on the image low-level clues. The heterogeneity of hand-crafted features and CNN image retrieval is thus computationally suboptimal and could be critical on resource-constrained platforms.

4. Proposed Approach

We now show how we address these issues and achieve improved robustness, scalability, and efficiency. We first motivate the use of learned features with a homogeneous network structure, and then detail the architecture in Section 4.1 and our novel training procedure in Section 4.2.

Learned features appear as a natural fit for the hierarchical localization framework. Recent methods like SuperPoint [12] have shown to outperform popular baseline like SIFT 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 networks is significantly faster than computing SIFT on GPU, it still remains a large computational bottleneck for the proposed localization system. With the goal of improving the ability to localize 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. We show in Figure 2 its application within the hierarchical localization framework.

4.1. HF-Net Architecture

Convolutional neural networks intrinsically exhibit a hierarchical structure. This paradigm fits well the joint predictions of local and global features and comes at low additional runtime costs. The HF-Net architecture (Figure 3) 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 computation 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 to predict local features.

The encoder of HF-Net is a MobileNet [39] backbone, a popular architecture optimized for mobile inference. Similarly to MNV [40], the global descriptor is computed by

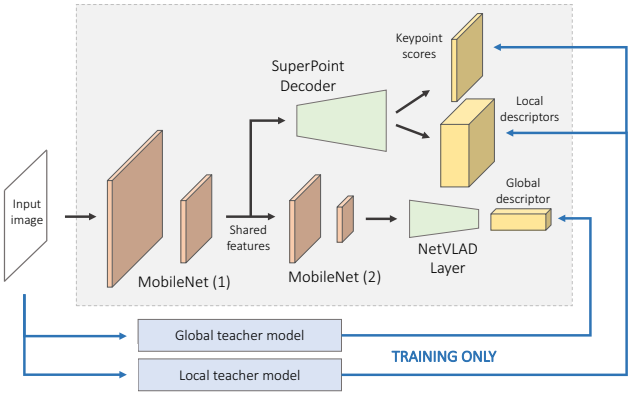

Figure 3. **HF-Net** generates three outputs from a single image: a global descriptor, a map of keypoint detection scores, and dense keypoint descriptors. All three heads are trained jointly with multi-task distillation from different teacher networks.

a NetVLAD layer [1] 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, resulting in a runtime independent from the number of detected keypoints. On the other hand, patch-based architectures like LF-Net [36] apply a Siamese network to image patches centered at all keypoint locations, 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. The local feature heads branch out from the MobileNet encoder at an earlier stage than the global head, as a higher spatial resolution is required to retain spatially discriminative features, local features are on a lower semantic level than image-wide descriptors [13].

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 [36] computed from an SfM model [45, 46], 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 to the invariance of the local descriptor. 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 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. Recent advances [21] in multi-task learning enable a student s to optimally copy all teachers $t_{1,2,3}$ without any manual tuning of the weights that balance the loss:

$$\begin{aligned} L = {} & e^{-w_1}||\mathbf{d}_s^g - \mathbf{d}_{t_1}^g||_2^2 + e^{-w_2}||\mathbf{d}_s^l - \mathbf{d}_{t_2}^l||_2^2 \\ & + 2e^{-w_3}\text{CrossEntropy}(\mathbf{p}_s, \mathbf{p}_{t_3}) + \sum_i w_i, \end{aligned} \tag{1}$$

where $\mathbf{d}^g$ and $\mathbf{d}^l$ are global and local descriptors, $\mathbf{p}$ are keypoint scores, and $w_{1,2,3}$ are optimized variables.

More generally, our formulation of the multi-task distillation can be applied to any application that requires multiple predictions while remaining computationally efficient, particularly in settings where ground truth data for all tasks is expensive to collect. It could also be applied to some hand-crafted descriptors deemed too compute-intensive.

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 [42] 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 [3] and SfM [36], that provide dense ground truth correspondences between image pairs for both 2D and 3D scenes.

Datasets. HPatches [3] 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 [36] composed of photo-tourism collections collected by [17, 50]. Ground truth correspondences are obtained from dense per-image depth maps and relative 6-DoF poses, computed using COLMAP [45]. 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 [15] and the descriptor Root-SIFT [2]. For the learning-based methods, we evaluate the detections and descriptors of SuperPoint [12] and LF-Net [11]. We additionally evaluate a dense version of DOAP [16] and the feature map `conv3_3` of NetVLAD [1] and use SuperPoint detections for both. More details are provided in the supplementary material.

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.

	HPatches		SfM	
	Rep.	MLE	Rep.	MLE
DoG	0.307	0.94	0.284	1.20
Harris	0.535	1.14	0.510	1.46
SuperPoint	0.495	1.04	0.509	1.45
LF-Net	0.460	1.13	0.454	1.44

Table 1. **Evaluation of the keypoint detectors.** We report the repeatability (rep.) and mean localization error (MLE).

	HPatches			SfM		
(detector / descriptors)	Homography	MS	mAP	Pose	MS	mAP
Root-SIFT	0.681	0.307	0.651	0.700	0.199	0.236
LF-Net	0.629	0.305	0.572	0.676	0.221	0.207
SuperPoint	0.810	0.441	0.846	0.794	0.418	0.488
Harris / SuperPoint	0.669	0.448	0.737	0.684	0.404	0.397
SuperPoint / DOAP	-	-	-	0.838	0.448	0.554
SuperPoint / NetVLAD	0.788	0.419	0.798	0.800	0.374	0.423

Table 2. **Evaluation of the local descriptors.** 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.

Descriptors. DOAP outperforms SuperPoint on all metrics on the SfM dataset, but cannot be evaluated on HPatches as it was trained on this dataset. 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.

Interestingly, SuperPoint descriptors perform poorly when extracted from 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 sub-pixel detection and patch-based description, are outperformed by dense descriptors like DOAP and SuperPoint. A simple representation trained with the right supervision can thus be more effective than a complex and computational-heavy architecture. We note that SuperPoint requires significantly fewer keypoints to estimate a decent pose, which is highly beneficial 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. We evaluate the impact of the two best descriptors, DOAP and SuperPoint, on the localization in Section 5.3. Results show that the latter 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. The global head is supervised by NetVLAD.

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 [34], containing a wide variety of day-time urban scenes, and 37k images from the night and dawn sequences of the Berkeley Deep Drive dataset [54], 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. Sarlin *et al*. [40] identified the local 2D-3D matching as the bottleneck of the pipeline. Our system significantly improves on the efficiency of their approach: i) 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, similarly to [33], thus retaining more matches in highly covisible areas. ii) Learned global and local descriptors are normalized and matched with a single matrix multiplication on GPU. Additional implementation details and hyperparameters are provided in the supplementary material.

5.3. Large-scale Localization

Under the light of the local evaluation, we now evaluate our hierarchical localization on three challenging large-scale benchmarks introduced by [42].

Datasets. Each dataset is composed of a sparse SfM model built with a set of reference images. The Aachen Day-Night dataset [43] 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 [28] 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 [4] 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 with COLMAP [45, 46] 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 SuperPoint and 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 within COLMAP using two-view geometry; 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 (685k vs 1,899k for SIFT) and fewer 2D keypoints per image (2,576 vs 10,230 for SIFT). However, a larger ratio of the original 2D keypoints is matched (33.8% vs 18.8% 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 first evaluate our hierarchical localization based on learned features extracted by NetVLAD [1] and SuperPoint [12]. Named NV+SP, it uses the most powerful predictors available. We then evaluate a more efficient localization with global descriptors and local features computed by HF-Net. We also consider several localization baselines evaluated by the benchmark authors. Active Search (AS) [41] and City Scale Localization (CSL) [48] are both 2D-3D direct matching methods representing the current state-of-the-art in terms of accuracy. DenseVLAD [52] and NetVLAD [1] are image retrieval approaches that approximate the pose of the query by the pose of the top retrieved database image. The recently-introduced Semantic Match Consistency (SMC) [51] 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 introduce an additional baseline, NV+SIFT, that performs hierarchical localization with RootSIFT as local features, and is an upper bound to the MNV+SIFT method of [40].

Results. We report the pose recall at position and orientation thresholds different for each sequence, as defined by the benchmark [42]. Table 3 shows the localization results for the different methods. Cumulative plots for the three most challenging sequences are presented in Figure 4.

Localization with NV+SP. On the Aachen dataset, NV+SP is competitive on day-time queries and outperforms all methods for night-time queries, where the performance drop w.r.t. the day 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 better than direct matching approaches, but are far outperformed by

	Aachen		RobotCar				CMU	
	day	night	dusk	sun	night	night-rain	urban	suburban
distance [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
orient. [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
SMC	-	-	(53.8 / 83.0 / 97.7)	(46.7 / 74.6 / 95.9)	(6.2 / 18.5 / 44.3)	(8.0 / 26.4 / 46.4)	75.0 / 82.1 / 87.8	44.0 / 53.6 / 63.7
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
NV+SP (ours)	**79.7** / **88.0** / 93.7	**40.8** / **56.1** / **74.5**	54.8 / **83.0** / **96.2**	**51.7** / **73.9** / **92.4**	**6.6** / **17.1** / **32.2**	**5.2** / **17.0** / **26.6**	**91.7** / **94.6** / **97.7**	**74.6** / **81.6** / **91.4**
HF-Net (ours)	75.7 / 84.3 / 90.9	**40.8** / **55.1** / **72.4**	53.9 / 81.5 / 94.2	**48.5** / **69.1** / 85.7	2.7 / 6.6 / 15.8	**4.7** / **16.8** / 21.8	**90.4** / **93.1** / **96.1**	**71.8** / **78.2** / **87.1**

Table 3. **Evaluation of the localization** on the Aachen Day-Night, RobotCar Seasons, and CMU Seasons datasets. We report the recall [%] at different distance and orientation thresholds and highlight for each of them the **best** and **second-best** methods. X+Y denotes hierarchical localization with X (Y) as global (local) descriptors. SMC is excluded from the comparison for RobotCar as it uses extra semantic data.

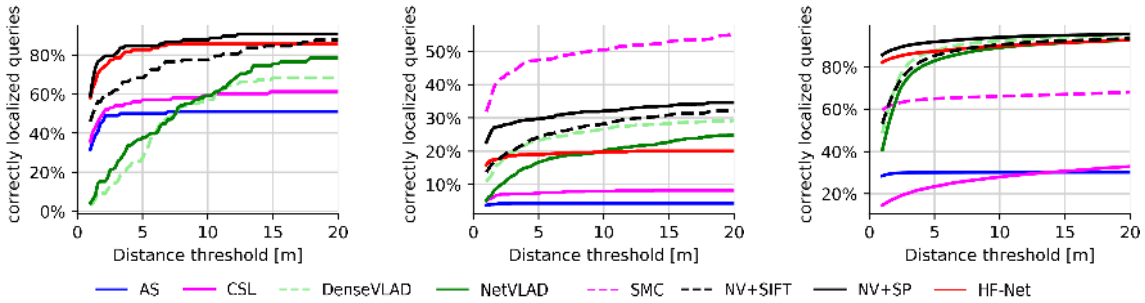

Figure 4. **Cumulative distribution of position errors** for the Aachen night (left), RobotCar night-all (center) and CMU suburban (right) datasets. On Aachen, HF-Net and NV+SP have similar performance and outperform approaches based on global retrieval and on feature matching. On RobotCar, HF+Net performs worse than NV+SP, which suggests a limitation of the distillation process. On CMU, the hierarchical localization shows a significant boost over other methods, particularly for small distance thresholds.

NV+SP in both fine- and coarse-precision regimes. On the difficult CMU dataset, NV+SP achieves an outstanding robustness compared to all baselines, including the most recent SMC. Overall, NV+SP sets a new state-of-the-art on the CMU dataset and on the challenging sequences of the Aachen and RobotCar datasets. The superior performance in both fine- and coarse-precision regimes shows that our approach is both more accurate and more robust.

Comparison with NV+SIFT. We observe that NV+SIFT consistently outperforms AS and CSL, although all methods are based on the same 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. It thus provides a better outlier rejection than complex domain-specific heuristics used in AS and CSL. The superiority of NV+SP highlights the simple gain of learned features like SuperPoint. On the Aachen night and RobotCar dusk sequences, which are the easiest ones, NV+SIFT performs marginally better than NV+SP for the fine threshold. This is likely due to the lower localization accuracy of the SuperPoint keypoints, as highlighted in Section 5.1, since DoG performs a subpixel refinement.

Localization with HF-Net. On most sequences, HF-Net performs similarly to its upper bound NV+SP, with a recall drop of 2.6% on average. We show qualitative results in Figure 5. In the RobotCar night sequences, HF-Net is significantly worse than NV+SP. We attribute this to the poor performance of the distilled global descriptors 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.

	Distance 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 Day-Night dataset. We report the recall [%] of the hierarchical localization with diffrent gloabal descriptors (NetVLAD and HF-Net) and local features (SuperPoint, DOAP, and HF-Net).

Ablation study. In Table 4, we evaluate the influence of different predictors within the hierarchical localization framework. Comparing NV+SP with NV+HF, we note that local HF-Net features perform better than the SuperPoint model that was used to train them. This demonstrates the benefits of multi-task distillation, where the supervision signal from the global teacher can improve intermediate features and help local descriptors. We also observe that the localization with DOAP is significantly worse at night, which

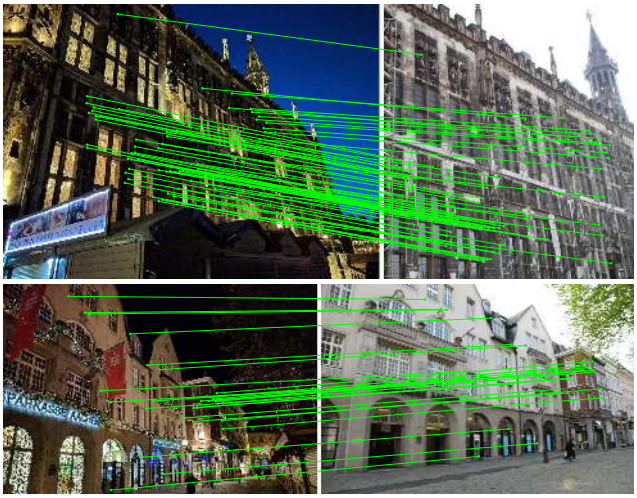

Figure 5. **Successful localization with HF-Net** on the Aachen Day-Night dataset. We show two queries (left) and the retrieved database images with the most inlier matches (right).

might be due to the complex augmentation schemes SuperPoint is based on. Finally, the comparison of HF-Net with NV+HF-Net reveals that HF-Net global descriptors have a somewhat limited capacity compared to the original NetVLAD and are limiting the performance.

5.4. Runtime Evaluation

As our propose localization solution was developed keeping the computational constraints in mind, we analyze its runtime and compare it with baselines presented in Section 5.3. These were measured on a PC equipped with an Intel Core i7-7820X CPU (3.60GHz) CPU, 32GB of RAM and an NVIDIA GeForce GTX 1080 GPU. Table 5 presents the detailed timings.

Datasets		Methods	Features	Global	Covis.	Local	PnP	**Total**
Aachen	Day	AS	263	-	-	— 112 —		**375**
		NV+SIFT	92+263	7	8	1220	29	**1356**
		NV+SP	92+26	7	5	9	9	**148**
		HF-Net	15	7	5	9	9	**45**
	Night	AS	263	-	-	— 132 —		**395**
		NV+SIFT	92+263	7	8	1492	56	**1655**
		NV+SP	92+26	7	5	10	18	**158**
		HF-Net	15	7	5	10	18	**55**
RobotCar	Dusk	AS	189	-	-	— 283 —		**472**
		NV+SIFT	92+189	13	3	264	14	**575**
		NV+SP	92+26	13	1	3	4	**139**
		HF-Net	15	13	1	3	4	**36**
	Night	AS	189	-	-	— 1021 —		**1210**
		NV+SIFT	92+189	13	3	389	149	**835**
		NV+SP	92+26	13	1	6	38	**176**
		HF-Net	15	13	1	6	38	**73**

Table 5. **Timings [ms]** for the different steps of hierarchical localization: feature extraction, global search, covisibility clustering, local matching, and pose estimation with PnP. Feature extraction with SIFT or CNN and matching of learned descriptors are performed on the GPU, and other operations on the CPU. We highlight the **fastest** method for each sequence. Localizing with HF-Net is 10 times faster than with AS, the fastest method available.

Hierarchical localization. Timings of NV+SP and HF-Net show that our coarse-to-fine approach scales well to large environments. The global search is fast, and only depends on the number of images used to build the model. It successfully reduces the set of potential candidate correspondences and enables a tractable 2D-3D matching. This strongly depends on the SfM model – the denser the covisibility graph is, the more 3D points are retrieved and matched per prior frame, which increases the runtime. NV+SIFT is therefore prohibitively slow, as its SfM model is much denser, especially on Aachen. NV+SP significantly improves on it, as the sparser SfM model yield clusters with fewer 3D points. The inference of NetVLAD and SuperPoint however accounts for 75% of its runtime, and is thus, as previously mentioned, the bottleneck. HF-Net mitigates this issues with an inference 7 times faster.

Existing approaches. CSL and SMC are not listed in Table 5 as they both require several tens of seconds per query, and are thus three orders of magnitude slower than our fastest method. AS improves on this, but is still slower, especially in case of a low success rate, such as on RobotCar night. Overall, our localization system based on HF-Net can run at 20 FPS on very large-scale environments. It is 10 times faster than AS, designed for efficiency, and is much more accurate on all datasets.

6. Conclusion

In this paper, we have presented a method for visual localization that is at the same time robust, accurate, and runs in real-time. Our system follows a coarse-to-fine localization paradigm. First, it performs a global image retrieval to obtain a set of database images, which are subsequently clustered into places using the covisibility graph of a 3D SfM model. We then perform local 2D-3D matching within the candidate places to obtain an accurate 6-DoF estimate of the camera pose.

A version of our method is based on existing neural networks for image retrieval and feature matching. It outperforms state-of-the-art localization approaches on several large-scale benchmarks that include day-night queries and substantial appearance variations across weather conditions and seasons. We then improve its efficiency by proposing HF-Net, a novel CNN that computes keypoints as well as global and local descriptors in a single shot. We demonstrate the effectiveness of multitask distillation to train it in a flexible manner while retaining the original performance. The resulting localization systems runs at more than 20 FPS at large scale and offers an unparalleled robustness in challenging conditions.

Acknowledgements. We thank the reviewers for their valuable comments, Torsten Sattler for helping to evaluate the localization, and Eduard Trulls for providing support for the SfM dataset.

References

[1] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, 2016. 2, 3, 4, 5, 6

[2] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012. 5

[3] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, 2017. 5

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

[5] Ioan Andrei Barsan, Shenlong Wang, Andrei Pokrovsky, and Raquel Urtasun. Learning to localize using a lidar intensity map. In *Conference on Robot Learning (CoRL)*, 2018. 1

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

[7] Mathias Bürki, Lukas Schaupp, Marcin Dymczyk, Renaud Dubé, Cesar Cadena, Roland Siegwart, and Juan Nieto. Vizard: Reliable visual localization for autonomous vehicles in urban outdoor environments. *arXiv:1902.04343*, 2019. 1

[8] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. BRIEF: Binary robust independent elementary features. In *ECCV*, 2010. 1, 2

[9] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018. 2

[10] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal Correspondence Network. In *NIPS*, 2016. 2

[11] Titus Cieslewski, Siddharth Choudhary, and Davide Scaramuzza. Data-efficient decentralized visual SLAM. In *ICRA*, 2018. 5

[12] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Workshop on Deep Learning for Visual SLAM at CVPR*, 2018. 1, 2, 3, 4, 5, 6

[13] Mohammed E. Fathy, Quoc-Huy Tran, M. Zeeshan Zia, Paul Vernaza, and Manmohan Chandraker. Hierarchical metric learning and matching for 2d and 3d geometric correspondences. In *ECCV*, 2018. 1, 2, 4

[14] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[15] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244, 1988. 5

[16] Kun He, Yan Lu, and Stan Sclaroff. Local descriptors optimized for average precision. In *CVPR*, 2018. 2, 3, 5

[17] Jared Heinly, Johannes L Schonberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the World in Six Days. In *CVPR*, 2015. 5

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015. 2

[19] Arnold Irschara, Christopher Zach, Jan-Michael Frahm, and Horst Bischof. From structure-from-motion point clouds to fast location recognition. In *CVPR*, 2009. 2

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

[21] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 2, 4

[22] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *International Symposium on Mixed and Augmented Reality*. IEEE Computer Society, 2007. 1

[23] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *CVPR*, 2011. 3

[24] Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. BRISK: Binary robust invariant scalable keypoints. In *ICCV*, 2011. 2

[25] Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *CVPR*, 2017. 2

[26] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 1, 3

[27] Simon Lynen, Torsten Sattler, Michael Bosse, Joel A Hesch, Marc Pollefeys, and Roland Siegwart. Get out of my lab: Large-scale, real-time visual-inertial localization. In *RSS*, 2015. 1, 2

[28] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *International Journal of Robotics Research*, 36(1):3–15, 2017. 6

[29] Colin McManus, Winston Churchill, Will Maddern, Alexander D Stewart, and Paul Newman. Shady dealings: Robust, long-term visual localisation using illumination invariance. In *ICRA*, 2014. 1

[30] Sven Middelberg, Torsten Sattler, Ole Untzelmann, and Leif Kobbelt. Scalable 6-DOF localization on mobile devices. In *ECCV*, 2014. 1, 2

[31] Michael J Milford and Gordon F Wyeth. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *ICRA*, 2012. 1

[32] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *NIPS*, 2017. 1, 3

[33] Dmytro Mishkin, Jiri Matas, and Michal Perdoch. Mods: Fast and robust method for two-view matching. *Computer Vision and Image Understanding*, 141:81–93, 2015. 6

[34] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *CVPR*, 2017. 2, 6

[35] David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Self-supervised learning of geometrically stable features through probabilistic introspection. In *CVPR*, 2018. 2

[36] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. LF-Net: Learning local features from images. In *NeurIPS*, 2018. 1, 2, 3, 4, 5

[37] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008. 2

[38] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *NeurIPS*, 2018. 2

[39] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: inverted residuals and linear bottlenecks. In *CVPR*, 2018. 2, 3

[40] 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, 6

[41] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE PAMI*, 39(9):1744–1756, 2017. 1, 2, 3, 6

[42] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF outdoor visual localization in changing conditions. In *CVPR*, 2018. 2, 4, 6

[43] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image retrieval for image-based localization revisited. In *CVPR*, 2008. 6

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

[45] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 4, 5, 6

[46] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 4, 6

[47] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *NeurIPS*, 2018. 2

[48] Linus Svärm, Olof Enqvist, Fredrik Kahl, and Magnus Oskarsson. City-scale localization for cameras with known vertical direction. *IEEE PAMI*, 39(7):1455–1461, 2017. 1, 2, 3, 6

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

[50] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *arXiv:1503.01817*, 2015. 5

[51] Carl Toft, Erik Stenborg, Lars Hammarstrand, Lucas Brynte, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl. Semantic match consistency for long-term visual localization. In *ECCV*, 2018. 1, 2, 6

[52] Akihiko Torii, Relja Arandjelović, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, 2015. 2, 6

[53] Tobias Weyand, Ilya Kostrikov, and James Philbin. PlaNet - photo geolocation with convolutional neural networks. In *ECCV*, 2016. 2

[54] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv:1805.04687*, 2018. 6

[55] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, 2018. 2

[56] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *CVPR*, 2018. 2