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

Supplementary material

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 [13] encoder with depth multiplier 0.75. The local heads are identical to the original SuperPoint [6] 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 1.

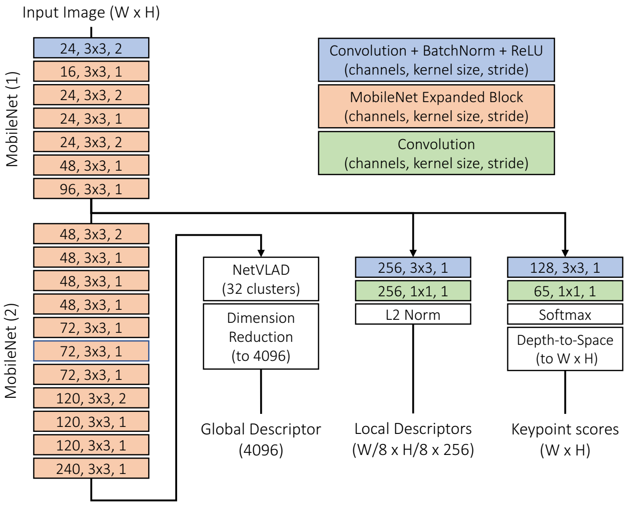

Figure 1. **Detail of the HF-Net architecture**, consisting of a MobiletNet encoder and three heads predicting a global descriptor, a dense local descriptor map, and keypoint scores.

A.2. Training Details

The images from both Google Landmarks [11] and Berkeley Deep Drive [21] 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, in a random order, 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.* [9].

The MobileNet layers are initialized with weights pretrained on ImageNet [5]. The network is implemented with Tensorflow [1] and trained for 85k iterations with the RMSProp optimizer [18] 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 [3] and SfM [12] datasets are resized so that their largest dimension is 640 pixels. The metrics are computed on image pairs and follow the definitions of [6, 12]. 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 the 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 [7], and SIFT [10], we use the implementations of OpenCV. For SuperPoint [6] and LF-Net [12], we use the implementations provided by the authors. For NetVLAD, we use the implementation of [4] and the original model trained on Pittsburgh30k. Dense descriptors are obtained by normalizing the feature map `conv3_3` before the ReLU activation. For DOAP [8], 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 2 and 3 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 1. We also report the track length, *i.e.* the number of observation per 3D point, as defined by [16]. 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**	
Statistics	SIFT	HF	SIFT	HF	SIFT	HF
# 3D points	1,900k	685k	6,869k	2,525k	961k	553k
# Keypoints per image	10,230	2,576	4,409	970	-	1,446
Ratio of matched keypoints [%]	18.8	33.8	39.4	59.9	-	45.3
Track length	5.85	5.87	5.34	4.71	4.11	4.95

Table 1. **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. Due to limits on the GPU memory, features are extracted on images downsampled such that their largest dimension is 960 pixels for Aachen and Robotcar, and 1024 for CMU. 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 [15], who also generated the plots shown in the main paper. For Active Search [14], City Scale Localization [17], DenseVLAD [20], and NetVLAD [2], we use the evaluation reported in the paper introducing the benchmark.

The evaluation of Semantic Match Consistency [19] (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 4, 5, and 6, respectively. We additionally show a comparison with NV+SIFT in Figure 7.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016. 1

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

[3] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, 2017. 1

[4] Titus Cieslewski, Siddharth Choudhary, and Davide Scaramuzza. Data-efficient decentralized visual SLAM. In *ICRA*, 2018. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[6] 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

[7] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244, 1988. 2

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

[9] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 1

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

[11] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *CVPR*, 2017. 1

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

[13] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: inverted residuals and linear bottlenecks. In *CVPR*, 2018. 1

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

[15] 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

[16] Johannes L Schönberger, Hans Hardmeier, Torsten Sattler, and Marc Pollefeys. Comparative evaluation of hand-crafted and learned local features. In *CVPR*, 2017. 2

[17] 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. 2

[18] Tijmen Tieleman and Geoffrey 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. 1

[19] 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. 2

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

[21] 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. 1

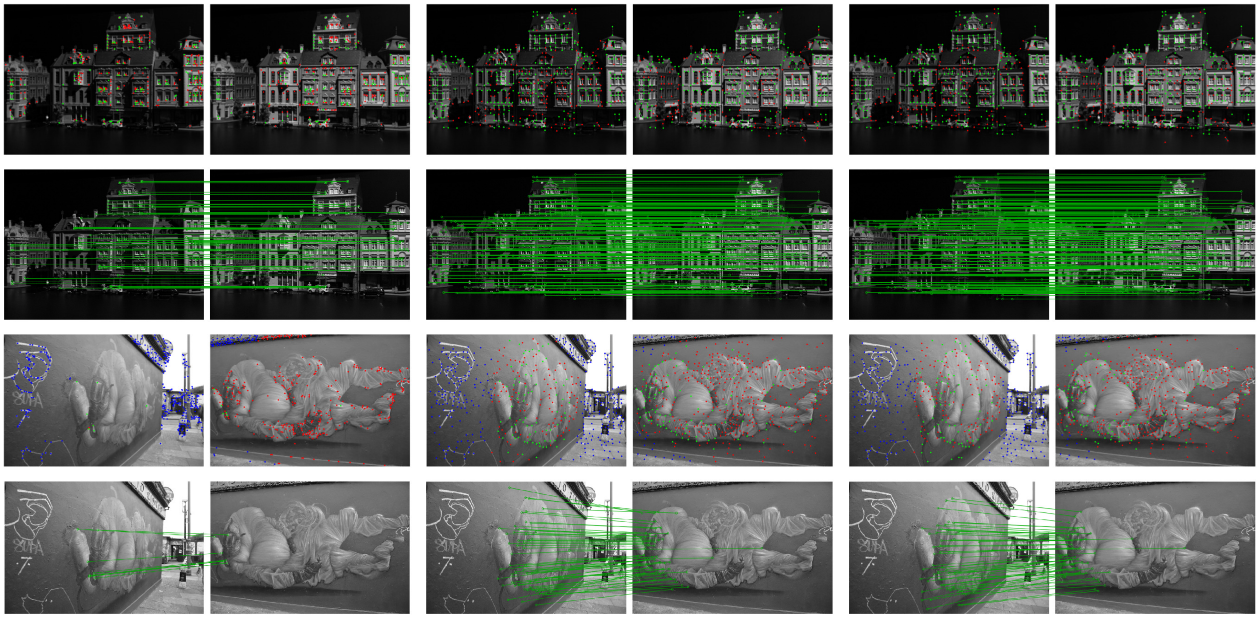

Figure 2. **Qualitative results on the HPatches dataset.** Keypoints (green if repeatable, red if not repeatable, blue if not visible in the other image) and inlier matches are shown for SIFT (left), SuperPoint (center) and HF-Net (right).

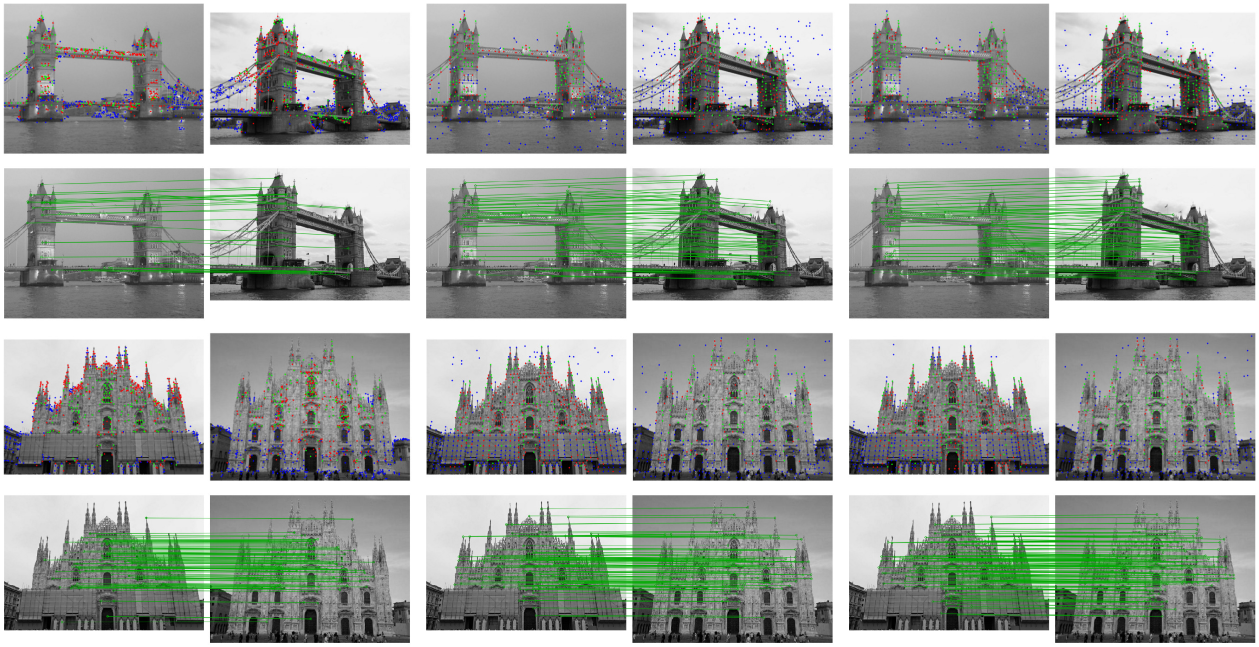

Figure 3. **Qualitative results on the SfM dataset** for SIFT (left), SuperPoint (center) and HF-Net (right).

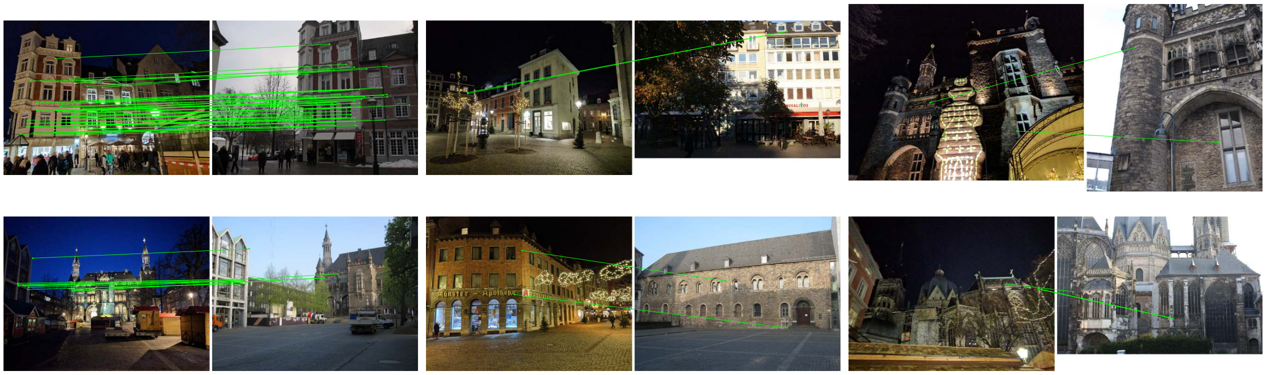

Figure 4. **Localization with HF-Net on Aachen night.** For each image pair, the left image is the query and the right image is the retrieved database image with the most inlier matches, as returned by PnP+RANSAC. We show challenging successful queries (left), failed queries due to an incorrect global retrieval (center), and failed queries due to incorrect or insufficient local matches (right).

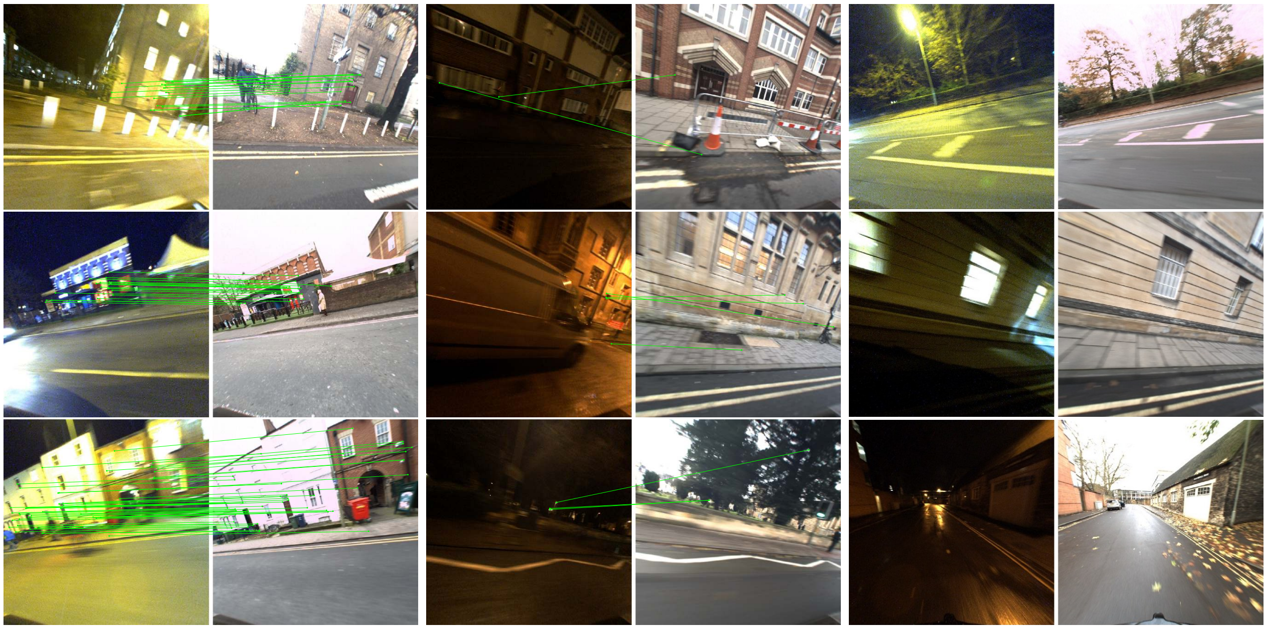

Figure 5. **Localization with HF-Net on RobotCar night and night-rain.** For each image pair, the left image is the query and the right image is the retrieved database image with the most inlier matches, as returned by PnP+RANSAC. We show challenging successful queries (left), failed queries due to an incorrect global retrieval (center), and failed queries due to insufficient local matches (right).

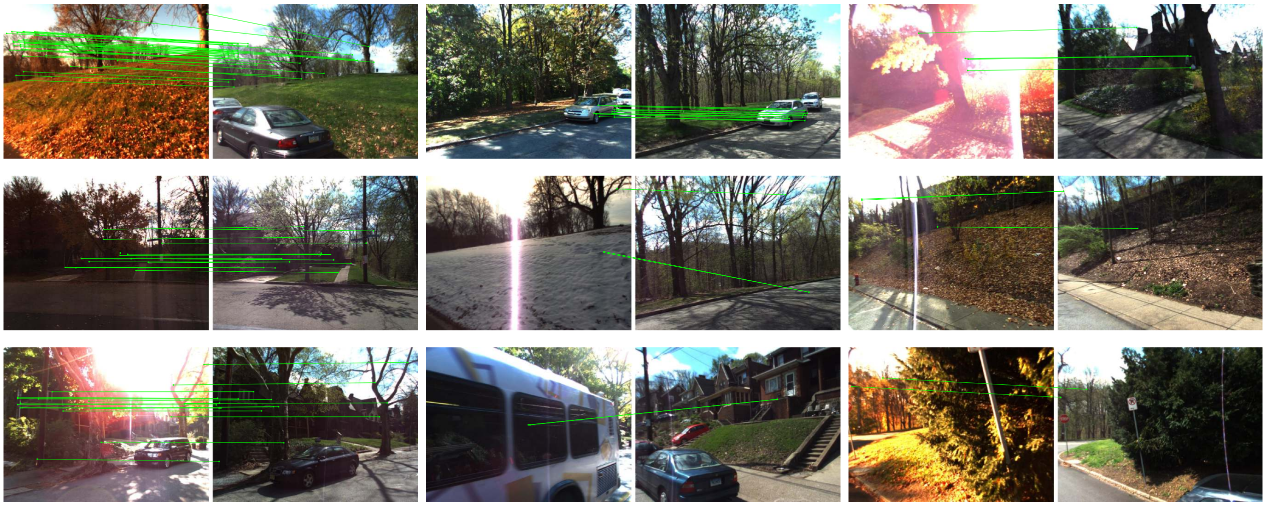

Figure 6. **Localization with HF-Net on CMU suburban.** For each image pair, the left image is the query and the right image is the retrieved database image with the most inlier matches, as returned by PnP+RANSAC. We show challenging successful queries (left), failed queries due to an incorrect global retrieval (center), and failed queries due to insufficient local matches (right).

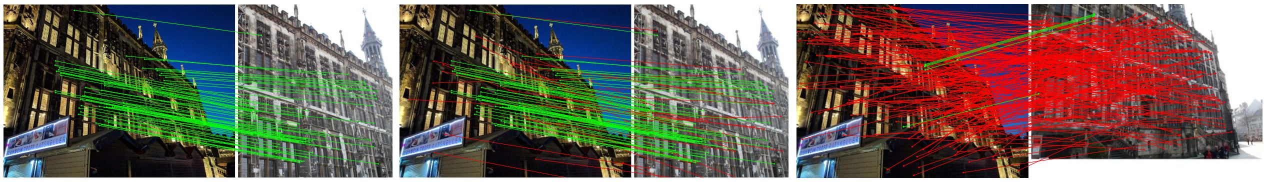

Figure 7. **Comparison between HF-Net and NV+SIFT on Aachen night,** with one query for which HF-Net returns the correct location but NV+SIFT fails. We show the matches with one retrieved database image, labeled by PnP+RANSAC as inliers (green) and outliers (red). We show the inliers of HF-Net (left), 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.