S2DNet: Learning Image Features for Accurate Sparse-to-Dense Matching

Supplementary Material

Hugo Germain[1], Guillaume Bourmaud[2], and Vincent Lepetit[1]

[1] LIGM, cole des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-valle, France
[2] Laboratoire IMS, Universit de Bordeaux, France
hugo.germain, vincent.lepetit@enpc.fr
guillaume.bourmaud@u-bordeaux.fr

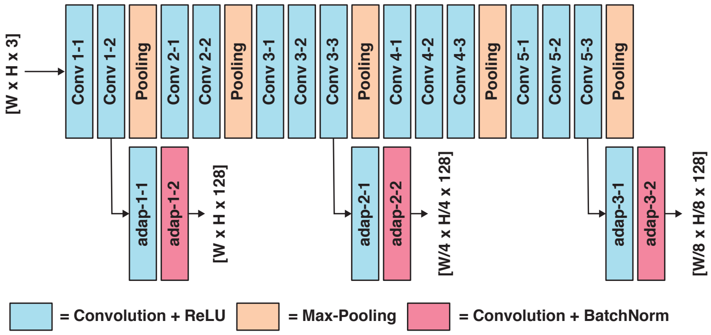

Fig. 1: S2DNet: Architecture overview. We feed images through a standard VGG-16 [14] backbone, and set three extraction points to process intermediate features. These features are sent to small, adaptation layers which help with the convergence and provide more condensed descriptors.

A Evaluation Details

In this section, we provide additional experiment details that were used to run our evaluations.

A.1 Cyclic Verification

As said in Section 4.2, we not only filter out correspondences using Eq. (4) (in the submitted version of the paper) but we also remove correspondences which do not pass the cyclic check of matching back on their source pixel. This is equivalent to performing a mutual nearest-neighbor verification as it is done with D2-Net [5]

ν	**MMA@1**	**MMA@2**	**MMA@3**	**MMA@10**
1.0	0.563	0.747	0.805	0.911
2.0	0.548	0.749	0.814	0.915
5.0	0.537	0.743	0.808	0.916
10.0	0.532	0.738	0.802	0.916

Table 1: Cyclic Verification. We report the MMA on HPatches [2] for several cyclic distance thresholds ν, using SuperPoint [4] detections and $\tau = 0.2$. We find that stricter thresholds improve the MMA at 1 pixel, while slightly damaging the coarser correspondences. In all our localization experiments, we use $\nu = 1.0$

and R2D2 [10]. To perform this verification, we measure the distance between a source keypoint $\mathbf{p}_{\mathrm{A}}^{n}$ and its cyclic correspondent after running the sparse-to-dense matching both ways and remove the correspondence if the following condition is not satisfied:

$$d_{\text{cyclic}} = \|\mathbf{p}_{\mathrm{A}}^{n} - \mathbf{p}_{\mathrm{A}}^{n\,*}\|_2 < \nu \,, \tag{1}$$

where

$$\mathbf{p}_{\mathrm{A}}^{n\,*} = \underset{\mathbf{p}\in\Omega}{\operatorname{argmax}}\; \mathtt{C}_{\mathbf{p}_{\mathrm{B}}^{n\,*}}^{\mathrm{B}\rightarrow\mathrm{A}}\,[\mathbf{p}] \tag{2}$$

and

$$\mathbf{p}_{\mathrm{B}}^{n\,*} = \underset{\mathbf{p}\in\Omega}{\operatorname{argmax}}\; \mathtt{C}_{\mathbf{p}_{\mathrm{A}}^{n}}^{\mathrm{A}\rightarrow\mathrm{B}}\,[\mathbf{p}] \,. \tag{3}$$

In our all experiments, we use a cyclic distance threshold of $\nu = 1$ pixel. In Table 1, we report the impact of this threshold on the mean matching accuracy.

A.2 Local Features Evaluation

The local features benchmark [13] couples the localization task with a multiview 3D reconstruction task. As discussed in the paper, the nature of sparse-to-dense matching in S2DNet does not guarantee the uniqueness of detections across multiple images. Thus, performing 3D reconstruction with S2DNet would result in a very high number of triangulated landmarks with low track lengths. Therefore, we perform the preliminary 3D reconstruction step with an off-the-shelf feature detector in a sparse-to-sparse fashion instead. Since we are dealing with daytime image pairs which are easier to match, we find that this is sufficient to obtain an accurate triangulation. We use the SURF [3] detector as we found it provided the best results. Indeed, SuperPoint [4] detects fewer keypoints which harms the performance under strong changes of scale. We then relocalize query images using S2DNet adopting this time a sparse-to-dense approach, and using triangulated keypoints as source detections.

A.3 Hierarchical Localization

In the day-night visual localization benchmark [13], we use S2DNet to perform hierarchical localization. We first perform image retrieval using DenseVLAD [18] global image descriptors to fetch the top-20 nearest neighbours of both daytime and nighttime queries. Similar to [11], we compute a covisibility graph on the retrieved database images to cluster 3D points, leading to a reduced set of places. For each landmark, we pre-compute sparse descriptors using S2DNet and perform sparse-to-dense matching on the query image to find its correspondent. The subsequent 3D-2D correspondences are then fed to a Perspective-n-Point (PnP) solver [8] inside a RANSAC [6] loop.

We compare our method to several baselines provided by the benchmark authors. Active Search (AS) [12] and City Scale Localization (CSL) [15] are both 2D-3D direct matching methods representing the current state-of-the-art in terms of accuracy. Semantic Match Consistency (SMC) [17] applies a semantic segmentation-based match rejection to improve the predicted poses. We report the results of pure image retrieval-based approaches using DenseVLAD [18] and NetVLAD [1]. For these methods, the query pose is approximated by the pose of the top-1 retrieved database image.

For hierarchical approaches, S2DHM [7] is the closest to ours. This method also performs sparse-to-dense matching, but is trained with weak supervision and computes downsampled correspondence maps. One main advantage of this method is that it is pre-trained on RobotCar daytime and nighttime images for the task of image retrieval. These training images are separate from the reference and evaluation set but still very similar visually to RobotCar evaluation images. We report the results of HF-Net [11], which performs hierarchical localization with NetVLAD and SuperPoint [4]. Lastly, we report the performance of D2-Net [5] provided by the authors.

A.4 InLoc evaluation

Since S2DNet was trained on outdoor images, we found that the confidence scores are overall lower when applied indoors, and confidence thresholdings $\tau > 0$ result in very few correspondences and damages the overall localization results. Thus, for the InLoc experiments, we use $\tau = 0$. InLoc query images are also of very high resolution (3024×4032 pixels). To speed up the matching process, we downscale all images to a maximum width or length of 1200 pixels.

B Qualitative Results

We report in Figure 2 example correspondence maps for InLoc [16] and RobotCar nighttime query images. We show the intermediate correspondence maps, as well as the final aggregated map and the retrieved correspondent. We report in Figure 3 inliers for the Sparse PE pipeline of InLoc [16], and show in Figure 4 two failure cases.

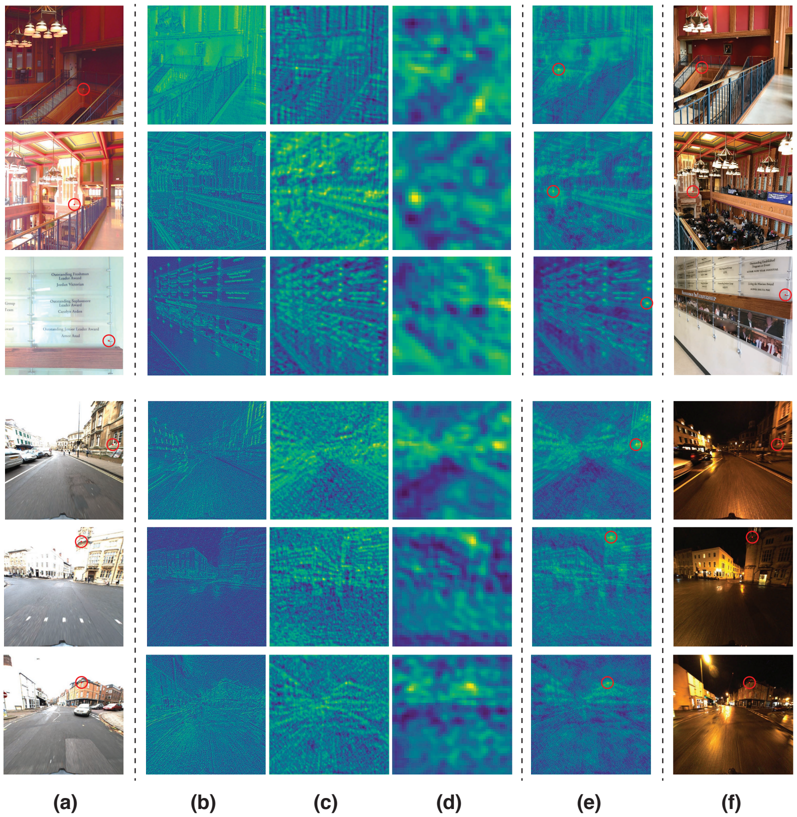

Fig. 2: Correspondence maps examples. From left to right: Reference image with a keypoint detection (a), intermediate correspondence maps predicted by S2DNet (b, c, d), aggregated pre-softmax correspondence map (e) and retrieved correspondent in the query image (f). The top three images are from InLoc [16] and the bottom three from RobotCar Seasons [9].

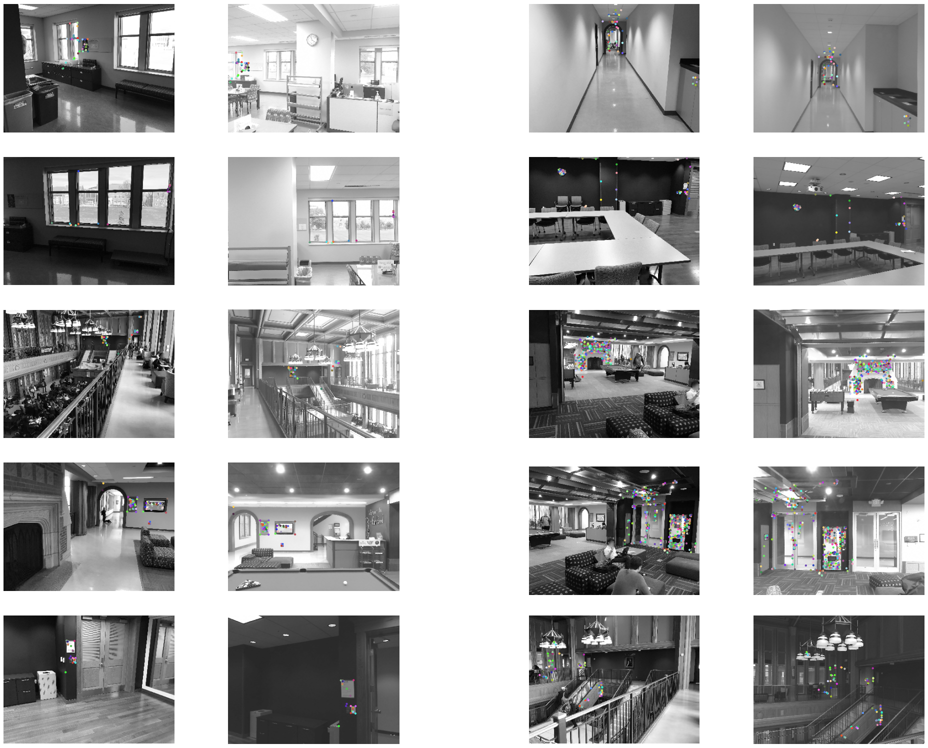

Fig. 3: Inlier Correspondences on InLoc [16]. Despite strong changes in scale, illumination and the large scale of the database, S2DNet manages to build robust and accurate correspondences.

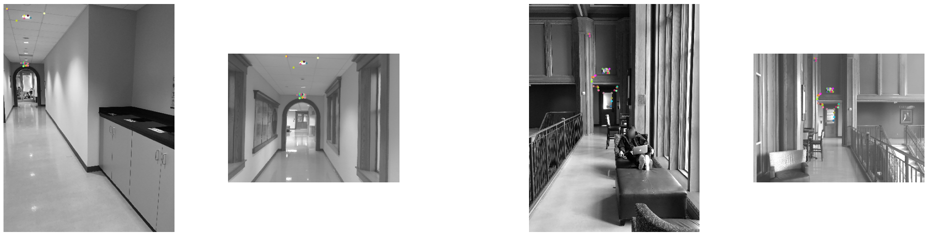

Fig. 4: Failure Cases Examples on InLoc [16]. Due to the repetitive structures present in the dataset, we find failure cases where such structures are matched despite being from two different places. In InLoc however, such cases are typically discarded when performing the dense pose verification (Dense PV) step.

References

1. Arandjelovic, R., Gronát, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In: Conference on Computer Vision and Pattern Recognition (2016) 3
2. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: HPatches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors. In: Conference on Computer Vision and Pattern Recognition (2017) 2
3. Bay, H., Tuytelaars, T., Gool, L.V.: SURF: Speeded Up Robust Features. In: European Conference on Computer Vision (2006) 2
4. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-Supervised Interest Point Detection and Description. In: CVPR Workshop (2018) 2, 3
5. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-Net: A Trainable CNN for Joint Description and Detection of Local Features. In: Conference on Computer Vision and Pattern Recognition (2019) 1, 3
6. Fischler, M.A., Bolles, R.C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Commun. ACM **24** (1981) 3
7. Germain, H., Bourmaud, G., Lepetit, V.: Sparse-To-Dense Hypercolumn Matching for Long-Term Visual Localization. In: International Conference on 3D Vision (2019) 3
8. Kneip, L., Scaramuzza, D., Siegwart, R.: 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 (2011) 3
9. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 Year, 1000 Km: The Oxford Robotcar Dataset. I. J. Robotics Res. **36** (2017) 4
10. Revaud, J., Weinzaepfel, P., de Souza, C.R., Pion, N., Csurka, G., Cabon, Y., Humenberger, M.: R2D2: Repeatable and Reliable Detector and Descriptor. In: Advances in Neural Information Processing Systems (2019) 2
11. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In: Conference on Computer Vision and Pattern Recognition (2019) 3
12. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. IEEE Transactions on Pattern Analysis and Machine Intelligence **39** (2017) 3
13. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., Pajdla, T.: Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In: Conference on Computer Vision and Pattern Recognition (2018) 2, 3
14. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR **abs/1409.1556** (2014) 1
15. Svärm, L., Enqvist, O., Kahl, F., Oskarsson, M.: City-Scale Localization for Cameras with Known Vertical Direction. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(7) (2017) 3
16. Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A.: Inloc: Indoor Visual Localization with Dense Matching and View Synthesis. CoRR **abs/1803.10368** (2018) 3, 4, 5
17. Toft, C., Stenborg, E., Hammarstrand, L., Brynte, L., Pollefeys, M., Sattler, T., Kahl, F.: Semantic Match Consistency for Long-Term Visual Localization. In: European Conference on Computer Vision (2018) 3

18. Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 Place Recognition by View Synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence **40** (2015) 3