Learning to Distill Convolutional Features into Compact Local Descriptors
-Supplementary material-

Method	Channel dim.	PF-PASCAL (τ_{img}) 0.05	0.1	0.15
COLD (Ours)	4	49.2	73.7	84.0
	8	56.5	78.2	87.5
	16	58.3	80.8	88.5
	32	66.4	84.0	90.0
	64	67.8	84.2	90.5
	128	71.2	86.8	92.1
	256	72.2	87.1	92.4
	512	73.3	88.2	92.9
	1024	75.7	88.2	92.9
	1536	76.6	88.8	93.6
	2048	75.8	**89.2**	**93.7**
	3072	**76.8**	89.0	93.4
NC-Net [7]	1024	54.3	78.9	86.0
SF-Net [4]	3072	53.6	81.9	90.6
DCC-Net [3]	1024	55.6	82.3	90.5
HPF [5]	6400	60.1	84.8	92.7

Table 1. Experiment of smaller channel dimension of local descriptors.

1. Extent of channel size

We study how far we can reduce our channel dimensions during feature transformation while minimizing the performance degradation, to see how more compact our descriptors can be. Table 1 illustrates the results evaluated on PF-PASCAL [2] dataset, showing the PCK values obtained for each final channel dimension of local descriptors. It can be seen that our model has comparable results with NC-Net [7] even when our final channel dimensions are as small as 8. COLD with 1024 channel size, which is same with NC-Net [7], performs 15.6%p, 9.3%p, and 6.9%p higher in threshold 0.05, 0.1, and 0.15, respectively. When channel dimensions are higher than 1024, there is a slight difference in their performance, revealing that the results of COLD with channel dimension more than 1024 is saturated in the PF-PASCAL dataset. Our model failed to converge when trying to output descriptors with channel dimensions below 4. We used ResNet-101 as our backbone network, and used the hyperparameter settings described in section 4 of the main paper for training our model in this experiment.

2. Qualitative results

Figure 1 and 2 show the qualitative results on Aachen day-night [8] and HPatches [1]. These datasets consist of rigid scenes with illumination and viewpoint changes. The visualized correspondences are found using nearest-neighbour search.

Figure 3 and 4 show the qualitative comparisons with the existing models [3, 5, 7] of PF-PASCAL [2] and SPair-71k [5] datasets. We visualize matches to transfer the key-points by prediction of each model.

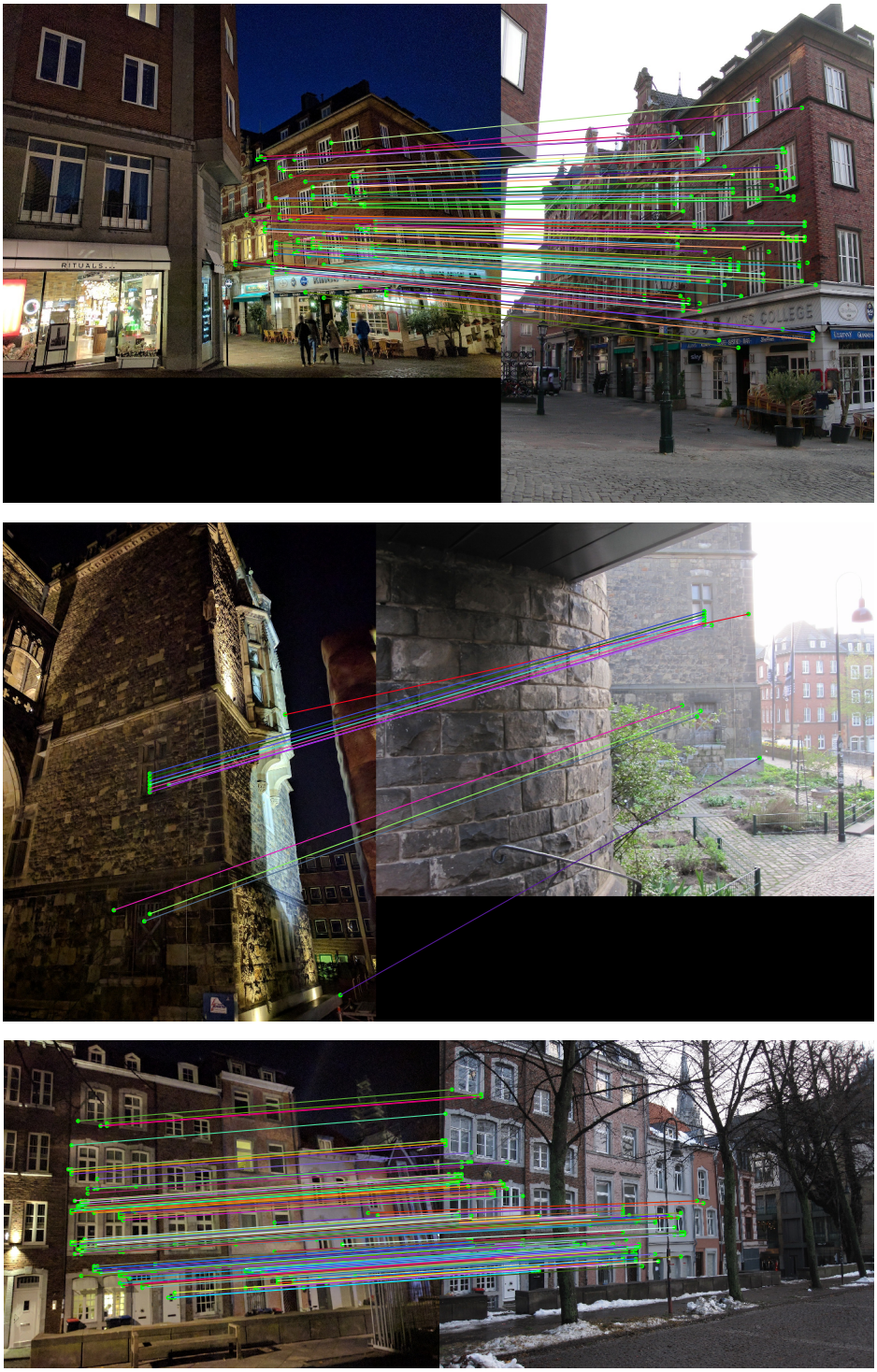


Figure 1. Qualitative results on Aachen day-night [8] dataset. Our model correctly infers corresponding points, even in occluded cases (row 2).

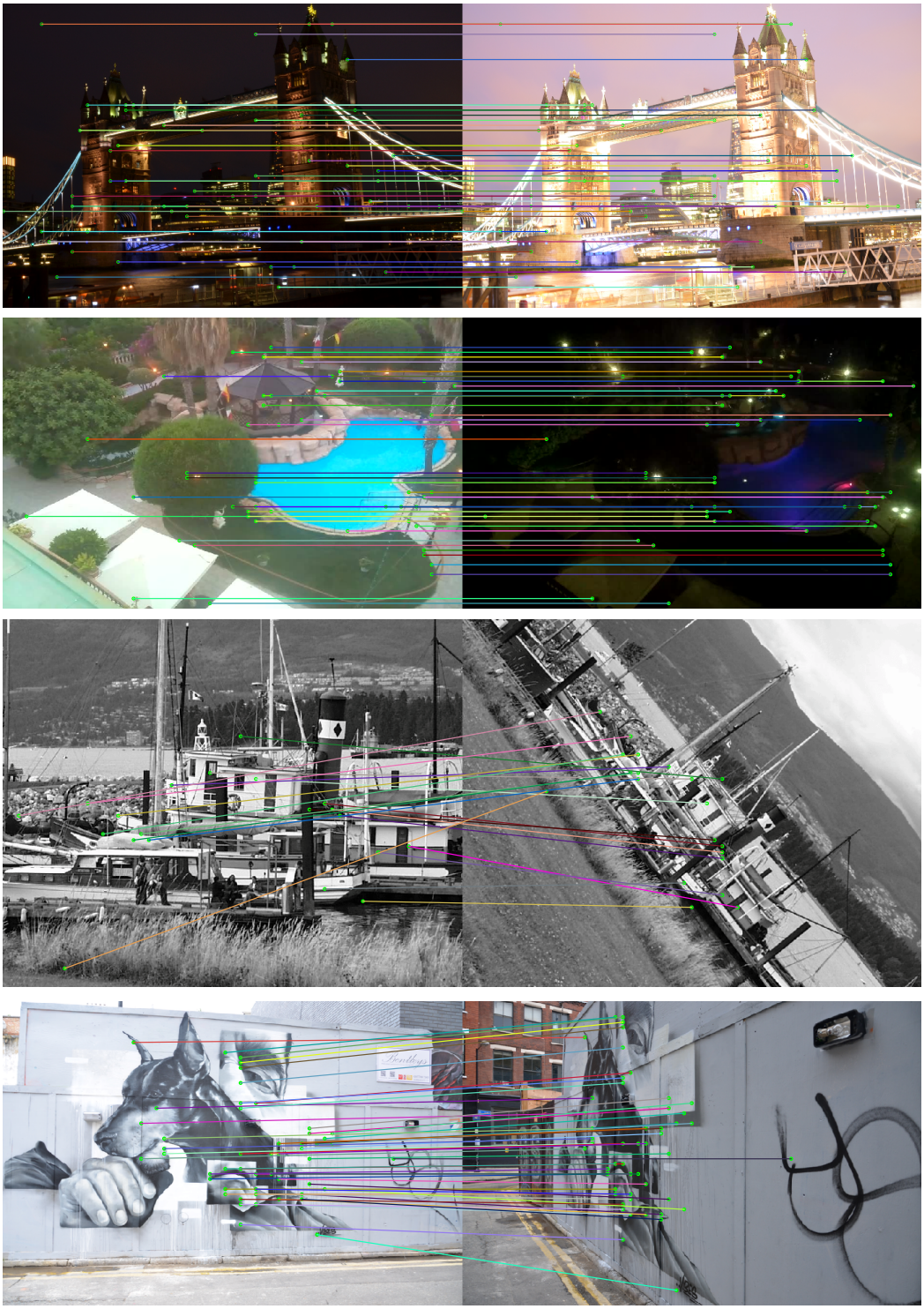


Figure 2. Qualitative results on HPatches [1] dataset. We select 50 random correspondences from the prediction of our model. These results show that our model is robust to illumination changes (row 1, 2), rotation changes (row 3), and viewpoint changes (row 4).

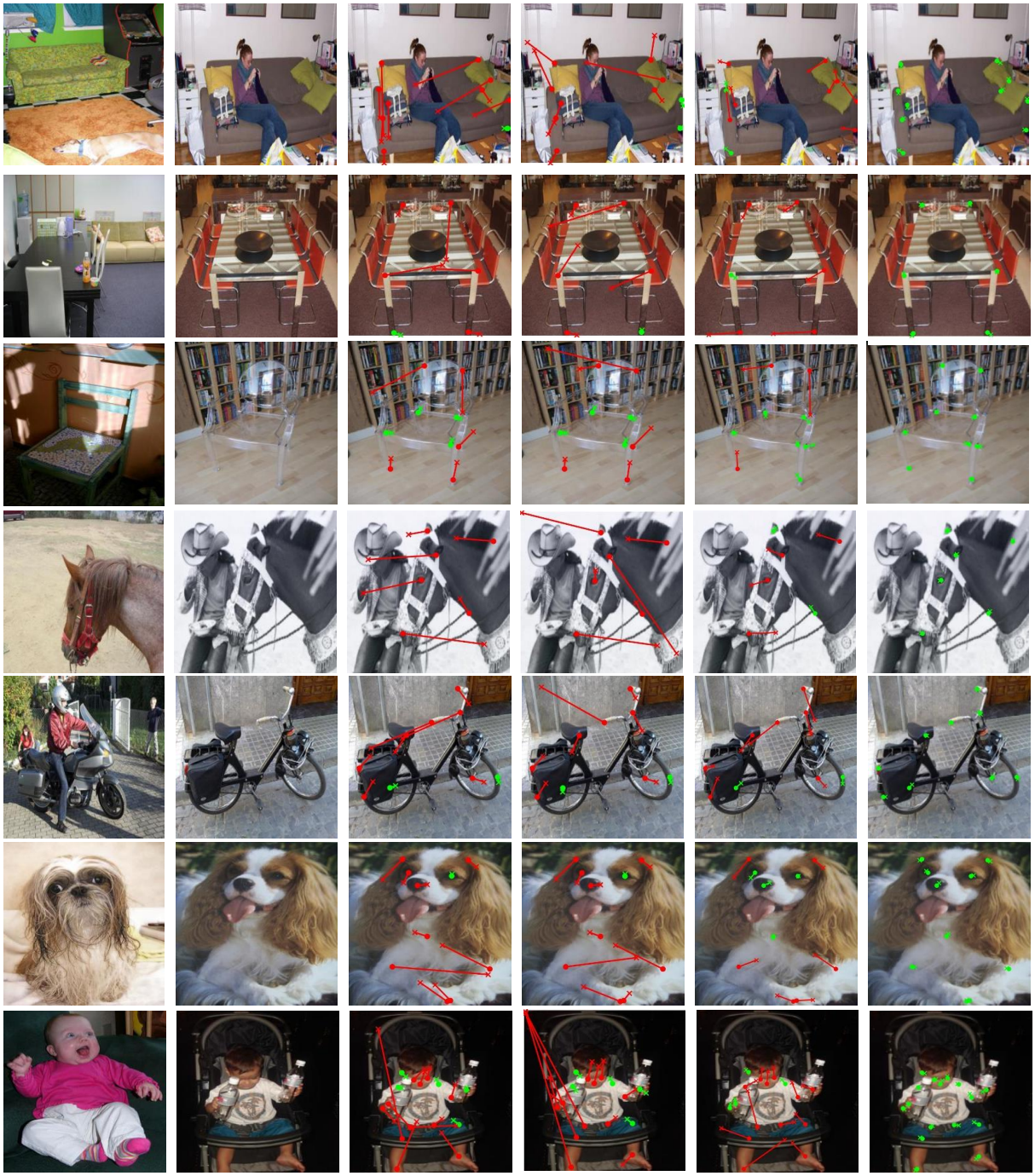
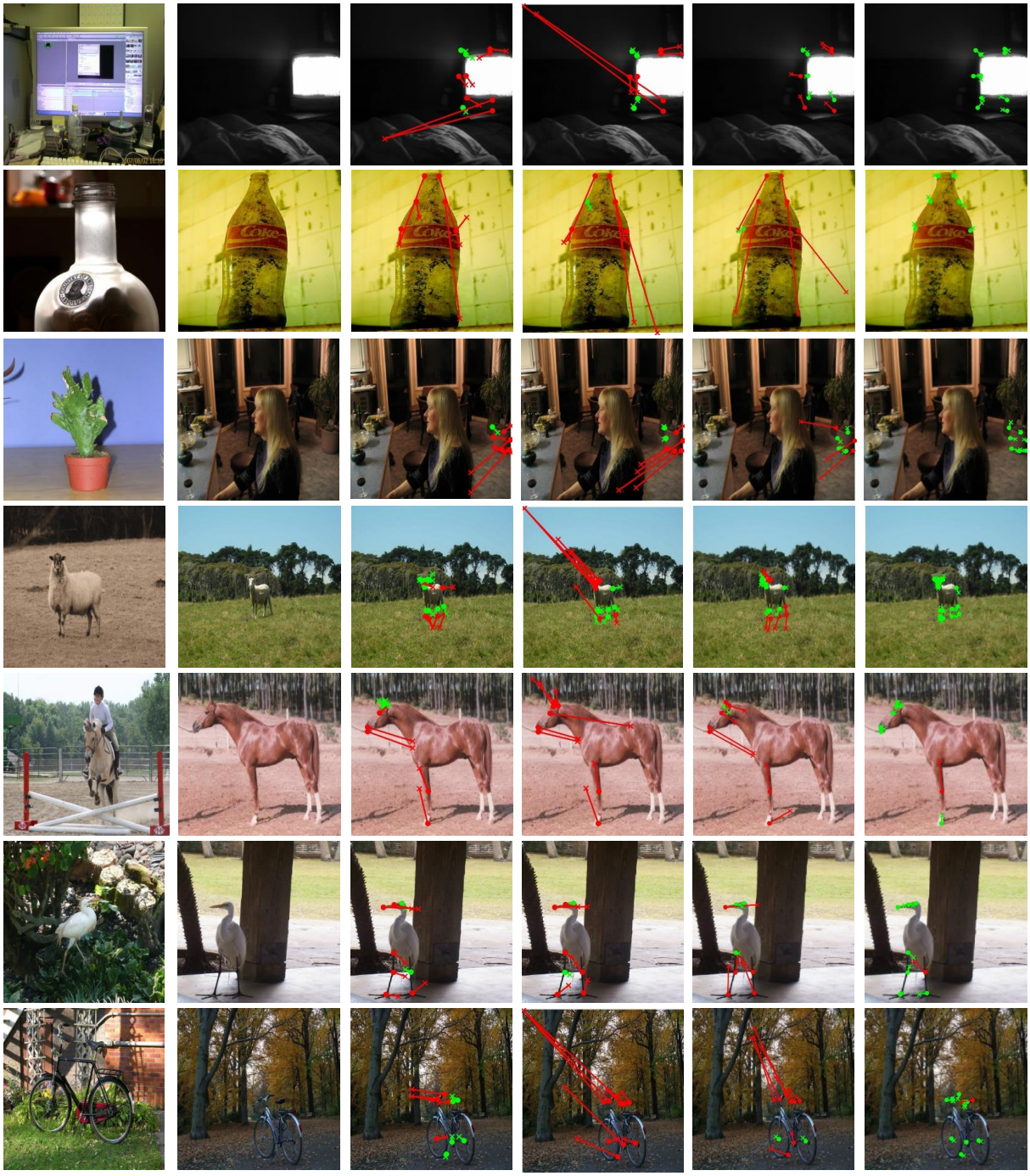


Figure 3. Qualitative comparisons on PF-PASCAL [2] dataset. Correct matches are colored as green and incorrect matches as red. The distance threshold for correctness was set to 20 pixels. We evaluated our model under images with partial occlusion (rows 1-3), extremely variant semantics (rows 4-5), and deformable objects (rows 6-7). The first two columns show source and target images, and the remaining columns show results from NCNet [7], DCCNet [3], HPF [5], and ours, respectively.



(a) Source (b) Target (c) NCNet (d) DCCNet (e) HPF (f) Ours

Figure 4. Qualitative comparisons on SPair-71k [6] dataset. Correct matches are colored as green and incorrect matches as red. The distance threshold for correctness was set to 20 pixels. We evaluated on image pairs under various changes - illumination, viewpoint, scale, occlusion, and truncation - where each image pair may be under multiple condition changes. For example, we can observe not only illumination changes but also scale changes in row 1. The first two columns show source and target images, and the remaining columns show results from NCNet [7], DCCNet [3], HPF [5], and ours, respectively.

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5173–5182, 2017.

[2] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(7):1711–1725, 2018.

[3] Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, and Xuming He. Dynamic context correspondence network for semantic alignment. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

[4] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

[6] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019.

[7] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Proc. Neural Information Processing Systems (NeurIPS)*, pages 1656–1667, 2018.

[8] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8601–8610, 2018.