Towards Visual Localization Interoperability: Cross-Feature for Collaborative Visual Localization and Mapping

Supplementary Material

This supplementary material provides information about additional ablation studies on different aspects for our proposed Cross-Feature method.

1. Ablation studies

In order to understand the impact on the performance of some architectural parameters, we carried different ablation studies. First, we studied the impact of the training strategy for interoperability Visual Localization. Then, we study the impact of some parameters in the training and matching performance: loss combination, patch size and number of features. For the latter, we used the HPatches dataset [1], measuring the MMA and the inlier number, in order to understand how the different parameters affect the Cross-Feature matching performance, so we limited the training to a per-pair embedding space with only SIFT and SuperPoint.

1.1. Training strategy

As discussed in the computational cost subsection from the main paper, due to the high computational load of our training method, we had to limit the number of training features to two. For that, we devised two different ways of training to reach a real common embedding space between different feature algorithms. First, by training it per-pair (CF^{emb*}) or second, by reaching a shared space for the four algorithms in an iterative way (the approach used in the whole experiment section of the main paper, denoted as CF^{emb}). In this ablation study, we measure the difference in performance between these two training strategies in the Aachen benchmark [2]. This is measured for ORB and DISK feature algorithms, as the first encoder for the second method was trained for SIFT-SuperPoint, and thus performance does not vary for these feature pair. Results of this ablation study are shown in Tab. 1, where we can see that performance is better for almost all combinations when using the per-pair training, except for SuperPoint and ORB, where the difference between the natures of the visual descriptors, i.e., learned and binary, may complicate to achieve good performance.

1.2. Training parameters

1.2.1. Loss type

For this ablation study, we studied the impact of the different losses (see Section 4.3) to evaluate how the information flow (which depends on the images and feature algorithms used for training) impacted the final matching performance of the feature embedding. For this, we trained the encoders for SIFT and SuperPoint under five different combinations of the losses: just homogeneous ($\mathcal{L}^{DS}$), combining homogeneous and the simpler Cross-Feature ($\mathcal{L}^{DS}+\mathcal{L}^{SD}$), combining homogeneous and the complex Cross-Feature loss ($\mathcal{L}^{DS}+\mathcal{L}^{DD}$), a combination of the two Cross-Feature losses without homogeneous ($\mathcal{L}^{SD}+\mathcal{L}^{DD}$) and the combination of the three losses (our proposed version). Results in Fig. 1 show that only using the homogeneous loss just allows for matching between same algorithms but impedes the heterogeneous case. However, its use is also important, as we can see a drop in the number of inliers in the case not using it. Additionally, results show that the combination of both Cross-Feature losses with the homogeneous is the one that rendered best results. We think that the simple Cross-Feature loss ($\mathcal{L}^{SD}$) acts as a bridge between the homogeneous and the complex Cross-Feature that provides a better consistency to the whole learning process.

Table 1. Comparison of the two different interoperability strategies in Aachen Day and Night benchmark [2].

Method	Algorithm		% localized queries					
			(0.25 m, 2°)	(0.5 m, 5°)	(5 m, 10°)			
	Map	Query	Day			Night		
CF^{emb}	SIFT	ORB	31.3	36.7	47.2	3.1	3.1	6.1
		DISK	33.5	40.9	54.1	4.1	6.1	10.2
CF^{emb*}	SIFT	ORB	35.7	39.9	48.3	1.0	3.1	5.1
		DISK	39.6	48.8	61.3	5.1	8.2	11.2
CF^{emb}	Superpoint	ORB	13.5	16.5	24.9	1.0	1.0	2.0
		DISK	45.1	53.2	65.4	8.2	12.2	14.3
CF^{emb*}	Superpoint	ORB	10.7	12.5	19.7	0.0	0.0	1.0
		DISK	69.4	78.8	85.7	35.7	46.9	55.1

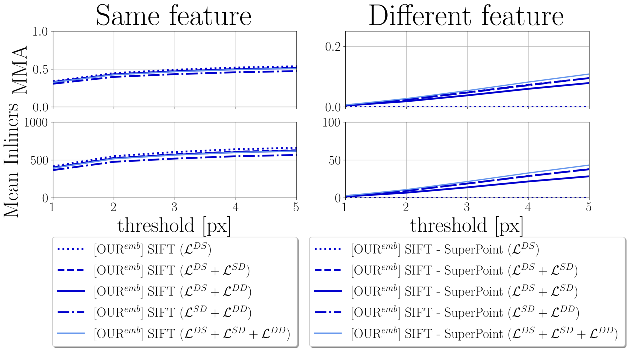

Figure 1. Performance on the HPatches dataset for different combinations of the loss functions. We average results on MMA and inliers for all sequences, comparing separately when projecting features from the same (left) and different (right) algorithms. Note the difference in scale.

1.2.2. Patch size

For this ablation study, we studied the impact of the patch window size T_w (used to fill the feature map) in the final matchability of the embedded feature in the heterogeneous scenario. The bigger this window size is, the higher number of overlaps between features. We varied this parameter from 7 to 15 and studied the matching performance, as we observed that it did not affect the self-consistency of the projections (the projection of the features is straightforward in the homogeneous case). Results can be seen in the Fig. 2, showing that the performance grows substantially when $T_w > 9$, due to an increment of the available information. The matching performance do not increase much between $T_w = 13$ and $T_w = 15$, so we theorize that it will not be much more informative for bigger window sizes, while it takes substantially more training time.

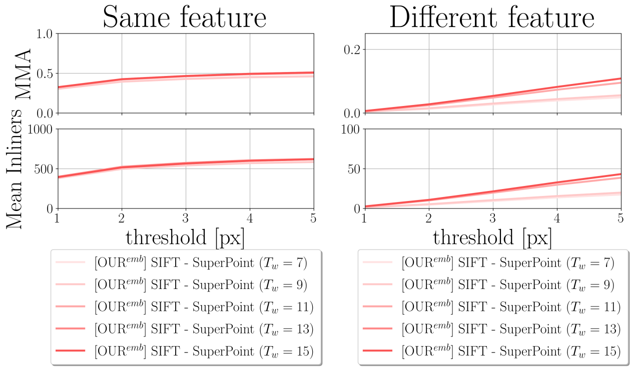

Figure 2. Performance on the HPatches dataset for varying patch size. We average results on MMA and inliers for all sequences, comparing separately when projecting from different features.

1.2.3. Numer of features

For this ablation, we studied the impact of the number of extracted features per patch on the model performance. We observed that, in the default configuration, the number of features per patch was $n \sim 300$ by average. For the experiment, we trained the encoders varying the number of sampled features (by setting a maximum number of detections in the feature extractors). For the three experiments, we used the same number of extracted features in evaluation. Fig. 3 demonstrates that the number of features heavily impacts its matching performance, as a higher number of features ensures a higher amount of available information, especially for heterogeneous correspondences.

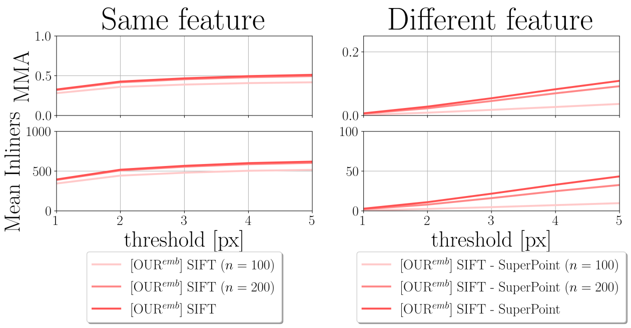

Figure 3. Performance on the HPatches dataset for varying number of features. We average results on MMA and inliers for all sequences, comparing separately when projecting features from the same (left) and different (right) algorithms. Note the difference in scale.

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, pages 5173–5182, 2017. 1

[2] 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 *CVPR*, pages 8601–8610, 2018. 1