Cross-Descriptor Visual Localization and Mapping Supplementary Material

Mihai Dusmanu [1] Ondrej Miksik [2] Johannes L. Schönberger [2] Marc Pollefeys [1, 2]
[1] Department of Computer Science, ETH Zürich [2] Microsoft MR & AI Lab, Zürich

This supplementary material provides the following information: first, we report additional experimental results on the Image Matching Workshop challenge [6], the HPatches benchmark [2], the InLoc Indoor Visual Localization dataset [16], and the Aachen Day-Night dataset [13]. Second, we provide further implementation details. Third, we explain the protocol for generating pseudo-ground-truth intrinsics and extrinsics for the Local Feature Evaluation benchmark [15]. Fourth, we show co-visibility statistics for the collaborative mapping experiments. Finally, we present an ablation study comparing different architectures and loss formulations.

1. Additional experimental results

In this section, we report additional experimental results. First, we evaluate our method on the Image Matching Workshop (IMW) challenge [6] as well as the HPatches descriptor evaluation benchmark [2]. Next, we provide a per-scene breakdown on the full sequences of the HPatches dataset [2]. Then, we study the impact of the joint embedding dimension in the scenario of localization to a collaborative map on the Aachen Day-Night dataset [13].

1.1. Image Matching Workshop challenge

We evaluate the performance of descriptor translation on the stereo and multi-view tasks of the IMW challenge [6]. Given the large number of methods to consider[1], we restrict the evaluation to the 3 validation scenes as follows: the smallest one (Reichstag) is used for parameter tuning (thresholds for the ratio test and RANSAC), while the other two (Sacre Coeur and Saint Peter's Square) are used for evaluation. We use the 2048 OpenCV SIFT keypoints with default parameters provided by the authors. For consistency, we retrain the encoder-decoder approach on patches extracted according to OpenCV SIFT keypoints on the same 3190 random internet images part of the Oxford-Paris revisited retrieval dataset distractors [11].

[1]The benchmark rules limit each team to a maximum of 2 submissions per week to avoid parameter tuning on the test set.

	Descriptor	Stereo AUC (%) 5°	Stereo AUC (%) 10°	Multi-view AUC (%) 5°	Multi-view AUC (%) 10°	Real.
Standard	BRIEF	35.3	41.8	31.9	36.5	✓
	SIFT	41.4	49.2	41.4	48.7	✓
	HardNet	51.4	59.9	55.9	63.5	✓
	SOSNet	51.4	60.1	58.6	66.2	✓
Directional	BRIEF → SIFT	25.2	31.5	14.9	17.6	✗
	BRIEF → HardNet	35.3	42.7	36.5	40.8	✗
	BRIEF → SOSNet	39.8	47.5	40.3	46.9	✗
	SIFT → HardNet	42.7	51.1	48.7	55.4	✗
	SIFT → SOSNet	45.1	53.5	47.3	55.2	✗
	HardNet → SOSNet	49.4	57.8	56.9	64.3	✗
Embed	BRIEF, SIFT, 1/2	39.5	47.1	41.6	48.1	✓
	BRIEF, HardNet, 1/2	42.4	50.3	46.2	52.8	✓
	BRIEF, SOSNet, 1/2	41.3	48.9	45.2	51.9	✓
	SIFT, HardNet, 1/2	46.8	55.1	53.4	61.3	✓
	SIFT, SOSNet, 1/2	46.2	54.4	49.9	57.6	✓
	HardNet, SOSNet, 1/2	50.4	58.9	57.6	64.9	✓
	All, 1/4	42.3	50.1	46.7	53.5	✓

Table 1: **Image Matching Workshop challenge.** We report results on the IMW challenge under two evaluation protocols: directional translation and collaborative mapping using the joint embedding.

We report results under two evaluation protocols in Table 1. First, we consider the case of directional translation. For a given direction ($A \rightarrow B$), in each image pair, we use the target description algorithm (B) in the first image and we translate source descriptors to target ones in the second image. Note that this does not correspond to a realistic scenario on the multi-view task, as the same image might use different descriptors in different image pairs. Second, we consider the case of collaborative mapping using the joint embedding. To this end, we randomly split the images of each dataset into balanced subsets, one for each description algorithm. Following the original evaluation protocol, we run each method three times and report the average over all runs. Once again, we notice an increase in performance when translating handcrafted to learned descriptors and matching them against natively extracted ones. Further,

Scenario	Database descriptor	Query descriptor	% localized queries DUC1 0.25m	DUC1 0.5m	DUC2 0.25m	DUC2 0.5m
Standard	BRIEF	BRIEF	25.3	36.4	22.9	41.2
	SIFT	SIFT	32.3	47.5	27.5	45.0
	HardNet	HardNet	36.4	52.5	30.5	54.2
	SOSNet	SOSNet	34.8	50.5	30.5	53.4
Cont. deployment	BRIEF →	SIFT	28.3	39.9	22.1	40.5
		HardNet	29.8	43.9	30.5	40.5
		SOSNet	31.8	43.4	23.7	40.5
	SIFT →	HardNet	36.4	50.0	31.3	50.4
		SOSNet	36.4	53.5	33.6	50.4
	HardNet →	SOSNet	33.3	48.5	30.5	55.7
Cross-device	SIFT ←	BRIEF	29.3	40.9	25.2	42.0
	HardNet ←	BRIEF	30.3	46.5	27.5	48.1
		SIFT	36.4	51.0	33.6	55.7
	SOSNet ←	BRIEF	29.8	44.9	29.0	45.0
		SIFT	34.8	51.0	33.6	53.4
		HardNet	37.4	50.5	29.0	49.6

Table 2: **InLoc Indoor Visual Localization. Localization under continuous deployment.** A reference map is built using the database description algorithm. The descriptors of this map are translated to a target query descriptor. **Cross-device localization.** A reference map is built using the database description algorithm. The descriptors of query images are translated to be compatible with the map.

in the binary collaborative scenario, the results are generally in between the results of the individual descriptors.

1.2. HPatches benchmark

To analyze the raw matching performance between original and translated descriptors, we evaluate our method on the HPatches benchmark [2]. There are three different tasks, notably patch verification, image matching and patch retrieval. For the translated methods (denoted $A \rightarrow B$), we use the target description algorithm (B) directly in the reference patches and we translate source descriptors to target ones for all other patches. Results are reported in Figure 1. As in our previous experiment, we notice an improvement in performance when translating handcrafted to learned descriptor. Furthermore, while some translated descriptors achieve worse performance than the baselines (*e.g.*, HardNet $\rightarrow$ SOSNet), all three tasks are possible in this previously unfeasible cross-descriptor scenario.

1.3. InLoc Indoor Visual Localization dataset

We also evaluate descriptor translation on the challenging InLoc Indoor Visual Localization dataset [16]. We follow the regular evaluation protocol for local features [16]. For each query, we retrieve top 10 related images according to NetVLAD [1] global descriptors. 2D-2D matches are established between the query image and each retrieved image. Next, keypoints in database images are back-projected to 3D using the ground-truth LiDAR scans to obtain a set of 2D-3D matches. Finally, RANSAC pose estimation is ran for each set of 2D-3D matches and the pose with the highest number of inliers is selected. For this experiment, we use DoG keypoints extracted using COLMAP.

The results are shown in Table 2. As with previous experiments, we notice a significant uptick in performance when matching translated hand-crafted descriptors against natively extracted learned ones, notably for the lower localization threshold.

Scenario	Descriptor	% localized queries Day (824 images) 0.25m, 2°	Day (824 images) 0.5m, 5°	Night (98 images) 0.25m, 2°	Night (98 images) 0.5m, 5°
Standard	BRIEF	76.1	81.4	32.7	36.7
	SIFT	82.5	88.7	52.0	61.2
	HardNet	86.2	92.2	64.3	72.4
	SOSNet	86.4	92.7	65.3	75.5
Collaborative	Embed 256	84.6 -1.8	90.8 -1.9	58.2 -7.1	64.3 -11.2
	Embed 128	84.8 -1.6	90.9 -1.8	57.1 -8.2	60.2 -15.3
	Embed 64	83.7 -2.7	89.1 -3.6	56.1 -9.2	61.2 -14.3
	Embed 32	80.3 -6.1	86.8 -5.9	45.9 -19.4	49.0 -26.5
	Embed 16	75.1 -11.3	80.8 -11.9	30.6 -34.7	33.7 -41.8

Table 3: **Localization to collaborative maps – embedding dimension.** The database images are partitioned in 4 balanced subsets, one for each description algorithm. We use SOSNet for query images. Both database and query descriptors are translated to the joint embedding space.

1.4. Localization to collaborative maps

We train our encoder-decoder approach with varied joint embedding dimensions. We present results in the case of localization to crowd-sourced maps on the Aachen Day-Night dataset [13] in Table 3. To recall, in this scenario, the set of database images is split in 4 balanced subsets, one for each description algorithm. For query images, we extract SOSNet [17] descriptors. Both query and database features are then translated to the joint space for matching. Increasing the dimensionality past 128 does not have any benefits in terms of performance. Interestingly, the 64-dimensional variant performs better than SIFT [7] despite using a heterogeneous map. Finally, even the 32-dimensional variant drastically outperforms native BRIEF [4] localization.

1.5. HPatches sequences

We present a per-scene comparison between state-of-the-art descriptors on the full HPatches sequences [2] following the evaluation protocol of Dusmanu *et al*. [5]. We report the absolute difference between the area under the mean matching accuracy curve up to 5 pixels for different pairs of descriptors in Figure 2. Despite SOSNet [17] drastically out-

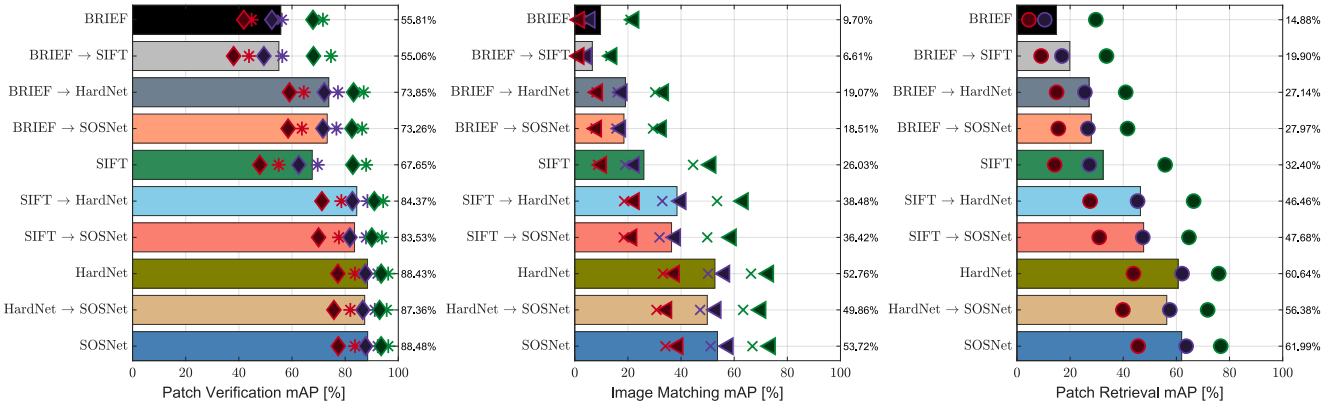

Figure 1: **HPatches.** We report verification, matching, and retrieval results on the HPatches dataset. Color of the marker indicates Easy, Hard, and Tough noise. For the patch verification task, diamonds and stars show results with negatives from the same sequence and from different sequences, respectively. For the image matching task, crosses and triangles denote illumination and viewpoint results, respectively.

performing BRIEF [4] and SIFT [7] overall, the handcrafted descriptors achieve better matching accuracy on a significant number of scenes. Similarly, while HardNet [9] and SOSNet have a similar overall performance, there are still small scene-to-scene variations. Thus, it is unclear whether existing learned descriptors (*e.g.*, SOSNet) are the best under all conditions.

2. Further implementation details

To encourage and facilitate future research on the topic of collaborative localization and mapping from heterogeneous devices, the code of our method and the evaluation protocols will be released as open source at `https://github.com/mihaidusmanu/cross-descriptor-vis-loc-map`.

The architectures used throughout our experiments are detailed in Table 4. Our approach was implemented in Python using PyTorch [10] and Kornia [12]. For the learned descriptors, we use the official Liberty [18] pre-trained weights released by the authors. Training the encoder-decoder approach for all 4 description algorithms takes around 30 minutes on a single NVIDIA RTX 2080Ti.

3. Pseudo-ground-truth generation

Similar to other datasets [13, 6], we generate pseudo-ground-truth intrinsics and extrinsics for the Local Feature Evaluation benchmark [15] via an initial Structure-from-Motion process. For each dataset, there are four steps:

- We extract SOSNet [17] descriptors around DoG [7] keypoints obtained using COLMAP [14] with default parameters. We exhaustively match all images using a mutual nearest neighbors matcher enforcing the ratio test [7] with a threshold of 0.9.
- We run COLMAP [14] for geometric verification and mapping. All images with less than 100 3D points are not considered during the next steps.
- We run geometric verification and mapping again on the remaining images – this time all intrinsics are fixed to the estimates from the previous step.
- We rescale the final model with respect to Google Maps by manual correspondence clicking to obtain final pseudo-ground-truth metric poses.

4. Collaborative mapping – co-visibility

Figures 9 and 10 report additional co-visibility statistics for our approaches to collaborative mapping from heterogeneous descriptors on the benchmark of Schönberger *et al.* [15]. The "Embed" approach translating everything to the joint embedding space generally manages to have more balanced models. This is especially noticeable in the percentage of 3D points containing at least one BRIEF descriptor in their tracks. However, this comes at a cost, as learned features (*i.e.*, HardNet, SOSNet) are less represented than in the "Progressive" approach.

We show a qualitative comparison of point-clouds in Figure 11. We compare the real-world point-clouds (*i.e.*, where each description algorithm only has access to a quarter of images) with the proposed crowd-sourced reconstruction. Our method is able to successfully match descriptors of different types yielding significantly more complete 3D models. On the most difficult landmark containing strongly symmetric structures and multiple night images (Gendarmenmarkt), we notice that some individual reconstructions are unable to recover the correct ground-truth scene geom-

BRIEF			
Layer	Batch norm.	Activation	Output dim.
Encoder			
input			512
hidden1	✓	ReLU	1024
hidden2	✓	ReLU	1024
embed			128
Decoder			
embed			128
hidden1	✓	ReLU	1024
hidden2	✓	ReLU	1024
output		Sigmoid	512

SIFT			
Layer	Batch norm.	Activation	Output dim.
Encoder			
input			128
hidden1	✓	ReLU	1024
hidden2	✓	ReLU	1024
embed			128
Decoder			
embed			128
hidden1	✓	ReLU	1024
hidden2	✓	ReLU	1024
output		ReLU	128

HardNet / SOSNet			
Layer	Batch norm.	Activation	Output dim.
Encoder			
input			128
hidden1	✓	ReLU	256
hidden2	✓	ReLU	256
embed			128
Decoder			
embed			128
hidden1	✓	ReLU	256
hidden2	✓	ReLU	256
output			128

Table 4: **Architectures.** We use shallow MLPs with 2 hidden layers for all methods. For the handcrafted algorithms (BRIEF [4], SIFT [7]) we use larger hidden layers as these descriptors encode lower level image structures. The joint embedding is ℓ_2 normalized and so is the output if required (*i.e.*, in the case of SIFT [7], HardNet [9], SOSNet [17]).

etry (notably BRIEF and SIFT).

5. Ablation study

In this section, we study the impact of architecture and losses on our data-driven translation approach. For this purpose, we consider the full sequences of the well-known HPatches dataset [2]. Following the protocol introduced by Dusmanu *et al*. [5], we report the mean matching accuracy of a mutual nearest-neighbor matcher for different thresholds up to which a match is considered correct. In each sequence, the first image is treated as query and matched against the other five. For translation experiments, the query descriptors are translated from a source description algorithm (*e.g.*, SIFT [7]) to a target one (*e.g.*, HardNet [9])) and matched against natively extracted descriptors (*e.g.*, HardNet) in the other images.

5.1. Naive matching

We first try naively matching different descriptors by running nearest neighbor search from one descriptor space to the other. Results are reported in Figure 4. BRIEF cannot be matched against SOSNet due to the different dimensionality. SIFT does not yield any correct matches when matched directly against SOSNet. This is also valid for HardNet, despite using the same backbone architecture and training data as SOSNet. Thus, it is impossible to naively match different descriptors and, without cross-descriptor matching, the final 3D models would be disconnected.

5.2. Pair vs. encoder-decoder

We compare a pair network trained specifically for SIFT to HardNet translation with an encoder-decoder network trained for all 4 description algorithms. We use the same dataset and hyper-parameters. We set the number of weights of the pair network equal to that of the encoder of SIFT concatenated with the decoder of HardNet. Results are reported in Figure 3. The performance of both approaches is similar. However, the encoder-decoder network can be trained once no matter the number of description algorithms and has the advantage of a joint embedding.

5.3. Number of description algorithms

In Figure 5, we show an ablation based on the number of different description algorithm used during training. We report the matching performance when matching HardNet to SOSNet features in the joint space. We consider 3 variants of the encoder-decoder architecture trained with different algorithm subsets: 4 was trained with all descriptors (BRIEF, SIFT, HardNet, SOSNet), 3 with SIFT, HardNet, SOSNet, and 2 only with HardNet and SOSNet. As can be seen, the performance gain is marginal when training exclusively with the learned methods. We believe the performance loss when compared to raw descriptors is due to enforcing consistency between different methods.

5.4. Loss

We investigate the effect of different losses on the encoder-decoder approach.

Matching loss. We first study the usefulness of the matching loss. For this purpose, we randomly select 512 patches from our training dataset. We extract the 4 descriptors from each patch and map them to the joint space using their associated encoders. Finally, we use t-SNE [8] for visualization. For clarity, we only plot 128 descriptors of each type in Figure 7. Training without a matching loss yields a representation that cannot be used for cross-descriptor matching. However, HardNet and SOSNet seem coherent suggesting that learned descriptors focus on similar informa-

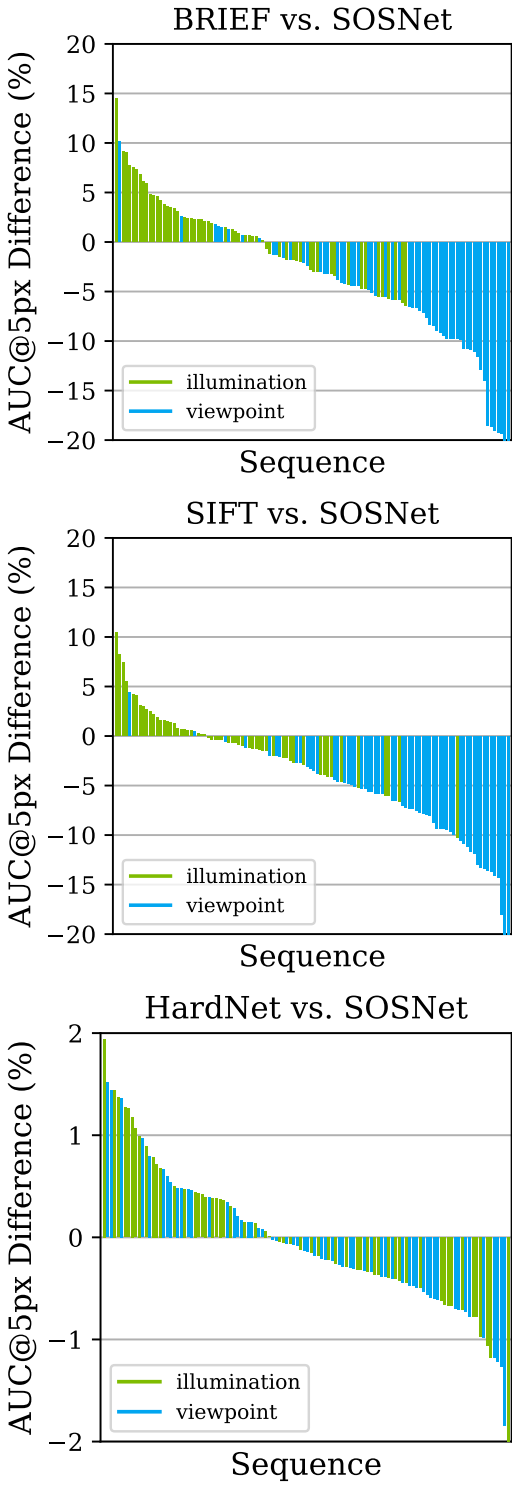

Figure 2: **HPatches sequences breakdown.** We report the per-scene absolute difference in the area under the mean matching accuracy curve up to 5 pixels between different descriptors. While SOSNet has a better overall performance, it does not outperform BRIEF or SIFT on all scenes. Similarly, while smaller, there are still scene-to-scene differences between the two learned descriptors.

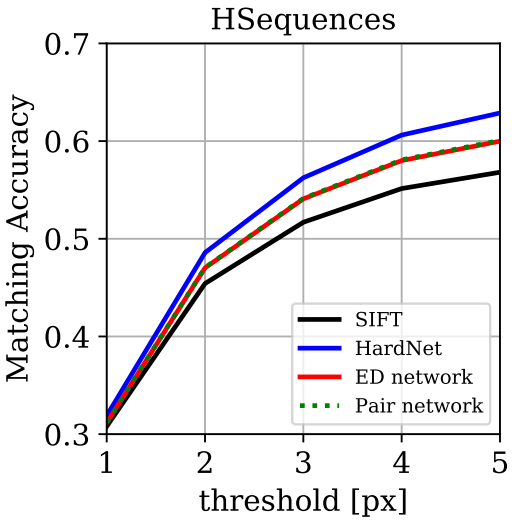

Figure 3: **Pair vs. encoder-decoder.** We report the performance of SIFT to HardNet translation on the full sequences from the HPatches dataset. The encoder-decoder (ED) network performs on par with the pair network despite being trained for 4 description algorithms at once.

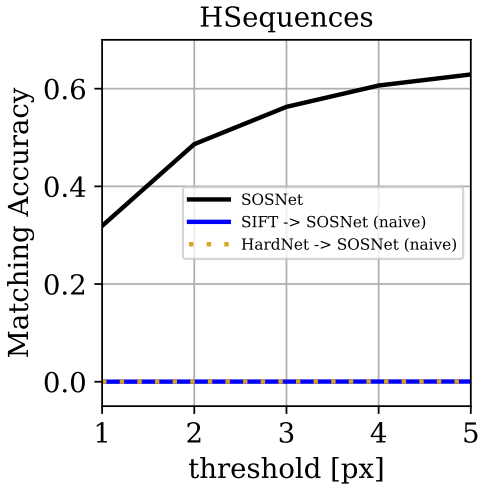

Figure 4: **Naive matching.** Matching different descriptors by running nearest neighbor search from one descriptor space to the other does not yield any correct matches.

tion. When leveraging the matching loss, all descriptors are well aligned. Furthermore, as shown by Figure 6, enforcing matchability in the joint space does not have a significant impact on the pair-wise translation.

Final loss. We study three variations of the final loss used for training. First, we consider the formulation presented in the main paper which takes into account interactions between all encoders and decoders:

$$\mathcal{L}^T_{\text{quadratic}} = \frac{1}{|\mathcal{A}|^2} \sum_{A_i, A_j \in \mathcal{A}^2} \mathcal{L}^T_{i \to j} \ , \qquad (1)$$

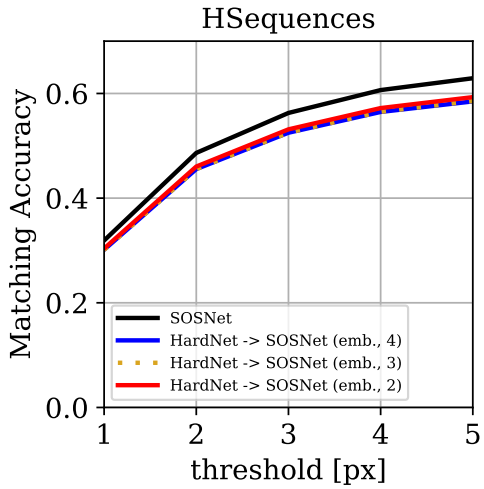

Figure 5: **Number of description algorithms.** The performance gain is minimal when training the encoder-decoder approach using the learned descriptors only.

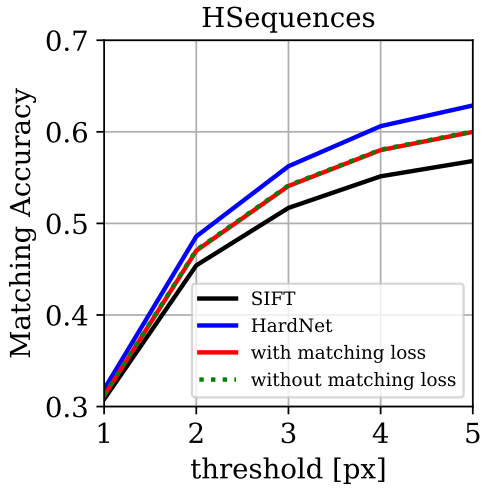

Figure 6: **Matching loss.** We report the performance of SIFT to HardNet translation on the full sequences from the HPatches dataset. The matching loss makes the joint space suitable for establishing correspondences and does not have a negative impact on pair-wise translation.

$$\mathcal{L}^M_{\text{quadratic}} = \frac{1}{|\mathcal{A}|^2} \sum_{A_i, A_j \in \mathcal{A}^2} \mathcal{L}^M_{i \to j}. \tag{2}$$

Second, the translation loss can be replaced by the traditional auto-encoder loss [3] defined as:

$$\mathcal{L}^T_{\text{auto-encoder}} = \frac{1}{|\mathcal{A}|} \sum_{A_i \in \mathcal{A}} \mathcal{L}^T_{i \to i} \ , \tag{3}$$

while the matching loss is kept as is:

$$\mathcal{L}^M_{\text{auto-encoder}} = \frac{1}{|\mathcal{A}|^2} \sum_{A_i, A_j \in \mathcal{A}^2} \mathcal{L}^M_{i \to j} \ . \tag{4}$$

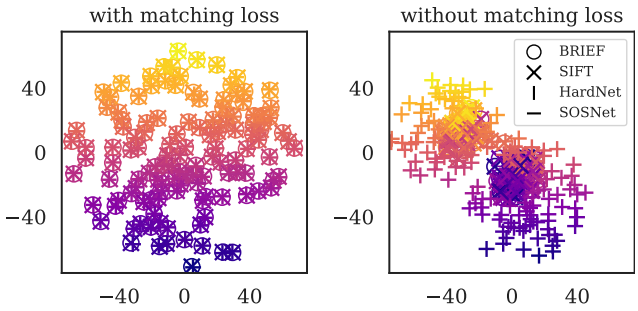

Figure 7: **t-SNE visualization of the joint space.** We visualize the embedding of 128 training patches with different description algorithms. Without matching loss, the handcrafted and learned descriptors are not coherent.

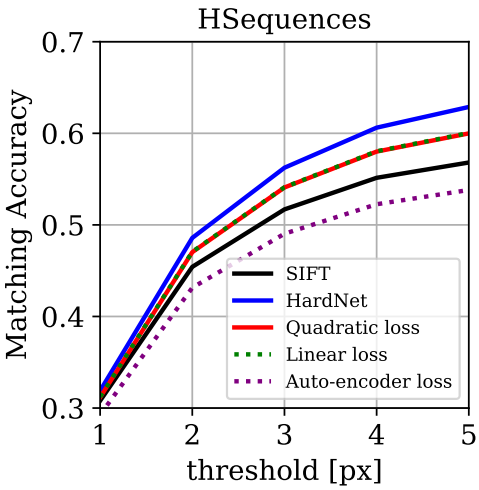

Figure 8: **Loss ablation.** We report the performance of SIFT to HardNet translation on the full sequences from the HPatches dataset. We train our encoder-decoder approach with different losses. Taking into account the interaction between encoders and decoders of different description algorithms is required for better performance.

Third, we propose a linear relaxation of our losses as:

$$\mathcal{L}^T_{\text{linear}} = \frac{1}{|\mathcal{A}|} \sum_{A_i \in \mathcal{A}} \mathcal{L}^T_{i \to \sigma(i)} \ , \tag{5}$$

$$\mathcal{L}^M_{\text{linear}} = \frac{1}{|\mathcal{A}|} \sum_{A_i \in \mathcal{A}} \mathcal{L}^M_{i \to \sigma(i)} \ , \tag{6}$$

with σ a permutation of $\{1, \ldots, |\mathcal{A}|\}$ chosen randomly at every optimization iteration. In each case, the final loss is a weighted sum of the translation and matching losses:

$$\mathcal{L}_* = \mathcal{L}^T_* + \alpha \mathcal{L}^M_* \ . \tag{7}$$

We train the encoder-decoder approach for all 4 description algorithms (*i.e.*, BRIEF, SIFT, HardNet, SOSNet) with

the same architecture and hyper-parameters using each loss independently. We report the results for SIFT to HardNet translation in Figure 8. The auto-encoder loss performs poorly as it does not consider the interaction between the encoders and decoders of different description algorithms. To speed up the training process (especially for larger collections of algorithms), one can use the linear variant of our losses as it yields similar performance.

References

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

[2] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proc. CVPR*, 2017. 1, 2, 4

[3] Hervé Bourlard and Yves Kamp. Auto-Association by Multilayer Perceptrons and Singular Value Decomposition. *Biological Cybernetics*, 1988. 6

[4] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. BRIEF: Binary Robust Independent Elementary Features. In *Proc. ECCV*, 2010. 2, 3, 4

[5] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In *Proc. CVPR*, 2019. 2, 4

[6] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image Matching across Wide Baselines: From Paper to Practice. *arXiv preprint arXiv:2003.01587*, 2020. 1, 3

[7] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2, 3, 4

[8] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 2008. 4

[9] Anastasiya Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in NeurIPS*, 2017. 3, 4

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in NeurIPS*, 2019. 3

[11] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. In *Proc. CVPR*, 2018. 1

[12] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an Open Source Differentiable Computer Vision Library for PyTorch. In *Proc. WACV*, 2020. 3

[13] 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 *Proc. CVPR*, 2018. 1, 2, 3

[14] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Proc. CVPR*, 2016. 3

[15] Johannes L. Schönberger, Hans Hardmeier, Torsten Sattler, and Marc Pollefeys. Comparative Evaluation of Hand-Crafted and Learned Local Features. In *Proc. CVPR*, 2017. 1, 3

[16] 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 *Proc. CVPR*, 2018. 1, 2

[17] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. SOSNet: Second Order Similarity Regularization for Local Descriptor Learning. In *Proc. CVPR*, 2019. 2, 3, 4

[18] Simon A. J. Winder and Matthew Brown. Learning Local Image Descriptors. In *Proc. CVPR*, 2007. 3

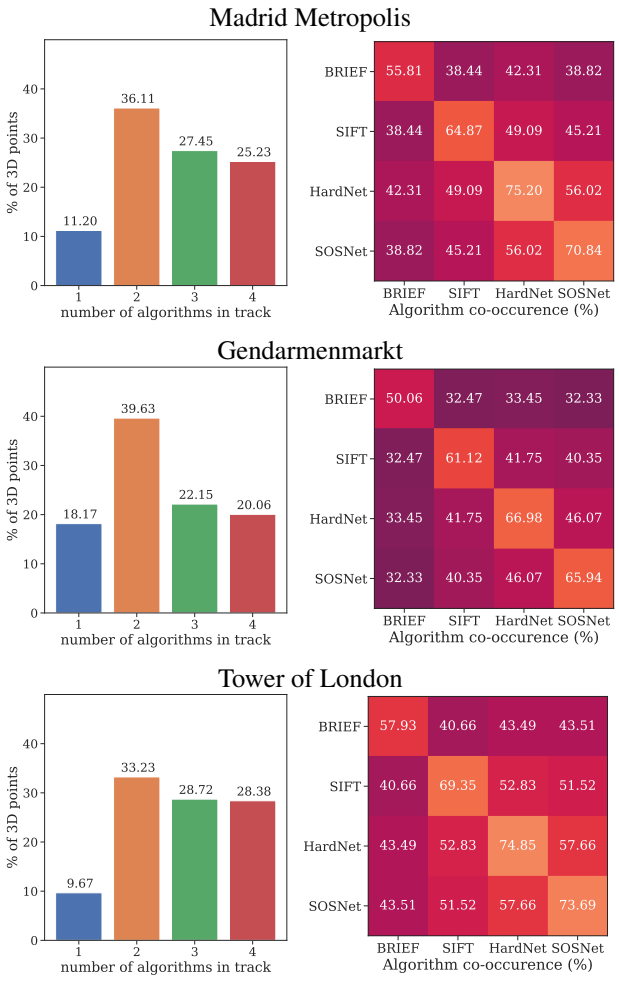

Figure 9: **Co-visibility statistics – "Embed".** For the "Embed" approach, we report the % of 3D points containing $1-4$ distinct algorithms in their tracks on the left. On the right, we visualize the co-occurence, *i.e.*, the percentage of 3D points containing descriptors originating from 2 given description algorithms in their tracks.

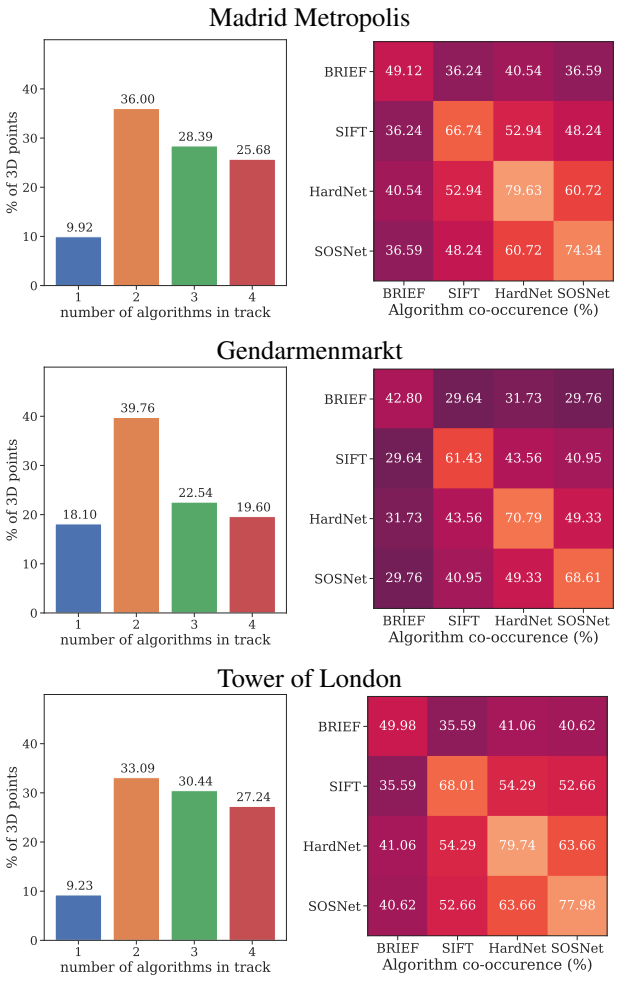

Figure 10: **Co-visibility statistics – "Progressive".** For the "Progressive" approach, we report the % of 3D points containing $1-4$ distinct algorithms in their tracks on the left. On the right, we visualize the co-occurence, *i.e.*, the percentage of 3D points containing descriptors originating from 2 given description algorithms in their tracks.

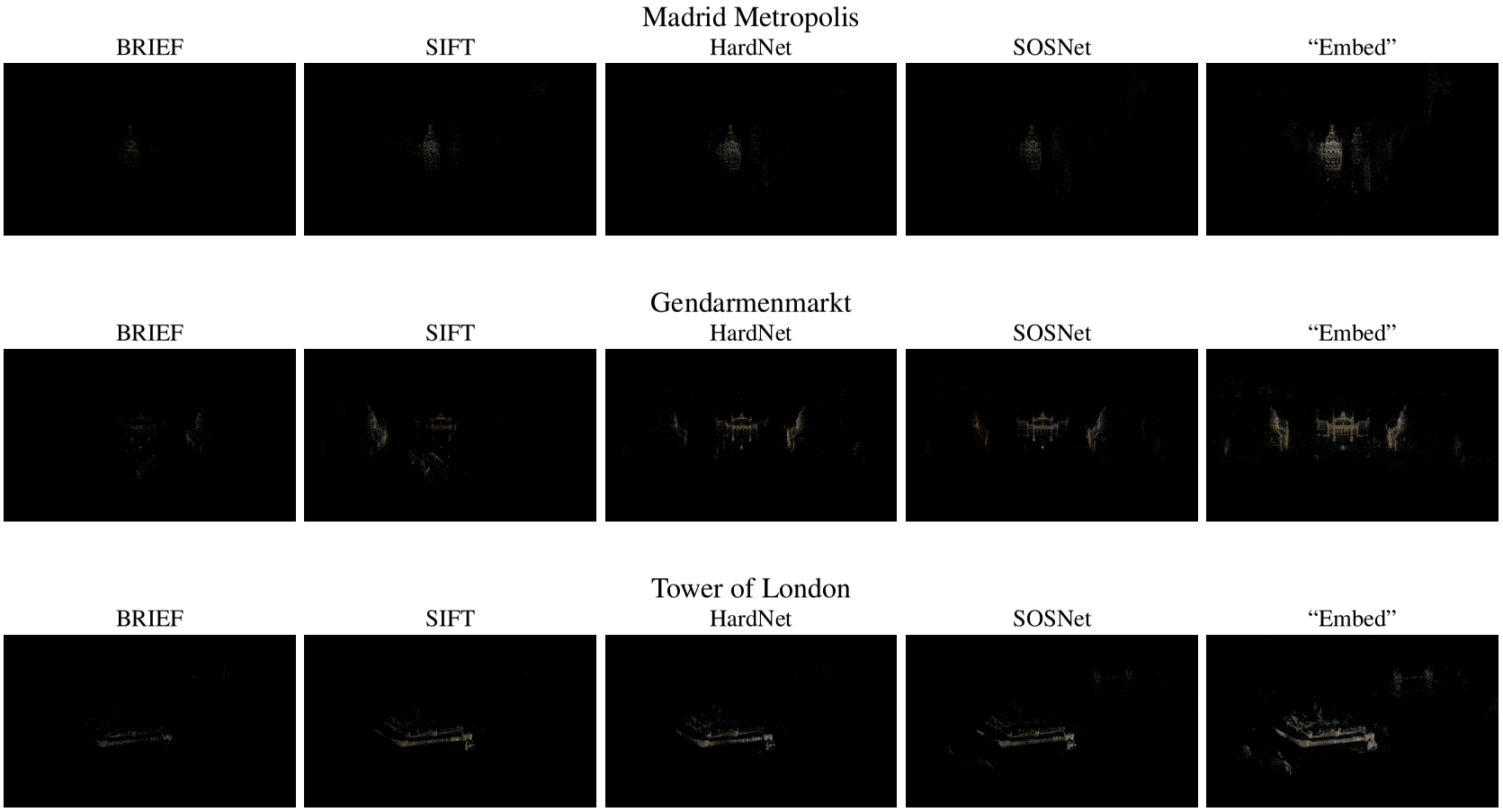

Figure 11: **Point cloud visualisations.** We show a qualitative comparison of the point-clouds obtained in the real-world setup (*i.e.*, where each description algorithm only has access to a quarter of images) and the ones obtained using our method to match different descriptors.