FeatureBooster: Boosting Feature Descriptors with a Lightweight Neural Network
Supplemental Material

Xinjiang Wang[1,2] Zeyu Liu[1,2] Yu Hu[1,2] Wei Xi[3] Wenxian Yu[1,2] Danping Zou[1,2]*
[1]Shanghai Key Laboratory of Navigation and Location Based Services, Shanghai Jiao Tong University
[2]SJTU SEIEE · G60 Yun Zhi AI Innovation and Application Research Center
[3]Intelligent Perception Institute, Midea Corporate Research Center
{wangxj83,ribosomal,henryhuyu,wxyu,dpzou}@sjtu.edu.cn xiwei1@midea.com

This supplementary material provides the following additional information: Section A presents the result of indoor visual localization using NN search with the mutual check. Section B provides the result of our method in visual SLAM. Section C shows the efficiency of different Transformer modules for cross-boosting stage. Section D provides an ablation study of the loss function used to train our method. As mentioned in Section 5.2 in the paper, Section E details how we chose the threshold for Lowe's ratio test [8] or distance test used for the visual localization and 3D reconstructions. Section F shows more qualitative examples of the matching results of our approach (before and after boosting) on the Aachen Day-Night v1.1 [17] and InLoc [13] datasets.

A. Indoor visual localization

Method	InLoc [13] (0.25m,10°) / (0.50m,10°) / (5.0m,10°) ↑	
	DUC1	DUC2
ORB [11]	21.7 / 30.8 / 36.9	**24.4** / **30.5** / 35.9
ORB+Boost-B (Ours)	**25.3** / **36.4** / **43.4**	23.7 / 29.8 / **37.4**
SIFT [8]	23.2 / 35.9 / 46.0	13.0 / 22.1 / 28.2
SOSNet [14]	**31.8** / **44.4** / **54.0**	**23.7** / **39.7** / **48.1**
RootSIFT [1]	24.7 / 36.9 / 41.9	17.6 / 27.5 / 33.6
SIFT+Boost-F (Ours)	28.3 / 40.4 / 47.5	19.8 / 29.0 / 35.1
SIFT+Boost-B (Ours)	24.2 / 35.9 / 46.0	18.3 / 29.0 / 35.1
SuperPoint [4]	**33.3** / 49.5 / 61.1	33.6 / 51.9 / **61.8**
SuperPoint+Boost-F (Ours)	32.3 / **51.0** / **64.1**	**36.6** / 51.9 / 59.5
SuperPoint+Boost-B (Ours)	**33.3** / 49.0 / 60.1	35.1 / 51.9 / 59.5
ALIKE [18]	31.8 / 47.5 / 61.1	26.7 / 41.2 / **49.6**
ALIKE+Boost-F (Ours)	**33.8** / **53.0** / **68.2**	**31.3** / **42.0** / 48.1
ALIKE+Boost-B (Ours)	28.8 / 43.9 / 56.6	**31.3** / 39.7 / 45.8

Table 1. Visual localization results in indoor scenes (InLoc [13]). The **first** and second best result are highlighted. In this test, no ratio/distance test is used for feature matching.

*Corresponding Author: Danping Zou (dpzou@sjtu.edu.cn)

To further evaluate the performance of our method, we apply our method for visual localization on the InLoc dataset [13] using only NN search and a mutual check without using the ratio or distance tests.

As shown in Tab. 1, our method can also enhance the performance of all descriptors although a different matching strategy is used. The SIFT+Boost-B is better than both SIFT [8] and RootSIFT [1]. The SuperPoint+Boost-B shows considerable competitiveness compared with SuperPoint [4]. We can also see that our ORB+Boost-B performs worse compared with SuperPoint [4] and ALIKE [18] without distance tests. In comparison, the results in Section 5.2 in the paper show that our ORB+Boost-B can compete with SuperPoint and ALIKE when we adopt ratio or distance tests for matching.

B. Visual SLAM

Our approach of reusing existing descriptors offers a cost-effective way to enhance the performance of established systems like visual SLAM. To demonstrate this, we integrated our ORB+Boost-B into ORB-SLAM2 [9].

The results of translation error in EuRoC dataset [3] for ORB-SLAM2 [9] using ORB and ORB+Boost-B are shown in Tab. 2. By boosting the original ORB [11] to ORB+Boost-B, ORB-SLAM2 provides more accurate estimate. Compared to other state-of-the-art local features, our method can improve the performance while introducing minimal additional time consumption (only 3.2ms on a desktop GPU and 27ms on an embedded GPU to process 2000 ORBs).

C. Transformer modules for cross-boosting

We compared the FeatureBooster using different Transformer modules for the cross-boosting stage. Specifically, we present the results of the vanilla transformer using MHA

Descriptor used (ORB-SLAM2 [9])	MH01	MH02	MH03	MH04	MH05	V101	V102	V103	V201	V202	V203
ORB [11]	0.0318	0.0215	0.0267	0.1282	0.0549	0.0349	**0.0211**	0.0486	0.0449	0.0270	**0.1716**
ORB+Boost-B (Ours)	**0.0304**	**0.0175**	**0.0252**	**0.0916**	**0.0470**	**0.0343**	0.0213	**0.0449**	**0.0379**	**0.0249**	0.2606

Table 2. Comparison of translation RMSE(m) in EuRoC dataset [3] for ORB-SLAM2 [9] using different descriptors. RMSE is the smaller the better and the **better** results are highlighted. The result shows that ORB+Boost-B improves the accuracy of ORB-SLAM2.

Descriptor	**Module used** (Cross-boosting)	**HPatches MMA**↑ @3 / @5	**RTX 3090 Runtime**(ms)↓ #500 / #1000 / #2000 / #4000 / #8000	**Jetson NX Runtime**(ms)↓ #500 / #1000 / #2000 / #4000 / #8000
ORB+Boost-B	Vanilla Transformer [15]	**0.437 / 0.500**	2.1 / 2.6 / 4.9 / 13.6 / 45.9	13.2 / 31.1 / 90.2 / 310.3 / ×
	Attention-Free Transformer [16]	0.436 / 0.495	**1.6 / 2.0 / 3.2 / 4.3 / 7.8**	**8.4 / 14.5 / 27.0 / 51.3 / 108.2**
SuperPoint+Boost-F	Vanilla Transformer [15]	**0.679 / 0.777**	2.8 / 3.9 / 8.7 / 27.1 / 96.8	23.5 / 60.0 / 185.0 / × / ×
	Attention-Free Transformer [16]	0.669 / 0.758	**1.9 / 2.1 / 3.3 / 5.4 / 10.2**	**13.2 / 23.5 / 44.1 / 87.3 / 194.2**

Table 3. The efficiency of using different Transformer modules. The table shows the mean matching accuracy (MMA) under thresholds 3 and 5 on HPatches dataset and the runtime for boosting different numbers of local features on RTX 3090 and Jetson Xavier NX 8GB. '×' indicates CUDA running out of memory.

Method		**Standard**		**Rotated**		**Average**	
		@3	**@5**	**@3**	**@5**	**@3**	**@5**
SIFT [8]	No boost	0.534	0.586	**0.505**	**0.559**	**0.519**	**0.572**
	No $\mathcal{L}_{BOOST}$	0.571	<u>0.644</u>	0.216	0.236	0.393	0.440
	$\lambda = 1$	**0.577**	**0.651**	0.263	0.287	0.420	0.469
	$\lambda = 10$	<u>0.573</u>	0.640	<u>0.391</u>	<u>0.428</u>	<u>0.482</u>	<u>0.534</u>
SuperPoint [4]	No boost	0.654	0.738	0.202	0.222	0.428	0.480
	No $\mathcal{L}_{BOOST}$	0.663	0.756	0.209	0.232	0.436	0.494
	$\lambda = 1$	**0.670**	**0.763**	**0.218**	**0.242**	**0.444**	**0.503**
	$\lambda = 10$	<u>0.669</u>	<u>0.758</u>	<u>0.213</u>	<u>0.235</u>	<u>0.441</u>	<u>0.497</u>

Table 4. Ablation study on the $\mathcal{L}_{BOOST}$. The table shows the mean matching accuracy (MMA) under thresholds 3 and 5 on the standard HPatches dataset, the rotated HPatches dataset, and the average performance using both datasets. We highlight the **first** and <u>second</u> best MMA values. The result shows the $\mathcal{L}_{BOOST}$ can help the boosted descriptors retain the performance of original descriptors in the cases where the training set does not include. Hence $\mathcal{L}_{BOOST}$ can improve the generalization ability of the trained model.

[15] and the attention-free transformer using AFT [16] in Tab. 3. The results show that the Attention-Free Transformer is much faster and consumes less GPU memory than the vanilla one, with a minor drop in matching performance.

D. Ablation study of the training loss

In this section, we study the impact of the training loss on our FeatureBooster. Our training loss consists of two term: $\mathcal{L}_{AP}$ and $\mathcal{L}_{BOOST}$, which are balanced using a weight λ. We use the HPatches [2] for the ablation study following the way in Section 5.1. To further evaluate the importance of $\mathcal{L}_{BOOST}$, we additionally use the rotated HPatches dataset [10] by applying random in-plane rotation of images from 0° to 360°, while our training set MegaDepth [7] does not contain large in-plane rotation cases.

Tab. 4 shows MMA (Mean Matching Accuracy) results under re-projection error thresholds of 3 and 5 pixels for three settings: standard, rotated, and average, which means using the standard HPatches dataset, the rotated HPatches dataset, and the average performance of using both datasets respectively. We can see that the original SIFT [8] achieves the best result under the rotated HPatches in the rotated and average settings. We believe the reason is that the training set (MegaDepth [7]) does not contain large-in-plane rotation cases. However, our $\mathcal{L}_{BOOST}$ can help the boosted SIFT retain the performance of SIFT on rotated HPatches when λ increases.

We also can see that the boosted SIFT and SuperPoint [4] can achieve better performance on Standard HPatches when $\lambda = 1$, but we set $\lambda = 10$ in the paper for a greater generalization of our method.

E. Threshold for ratio/distance test

It is known that using ratio or distance tests can reject many incorrect correspondences and improve the RANSAC [6] efficiency and the final matching results. The ratio test is to check if the ratio of the descriptor distance of the closest feature to that of the second closest one is smaller than a threshold. Distance tests simply check if the distance between two matched descriptors is within a threshold.

To find a suitable ratio/distance threshold for a fair comparison in the experiments, we compute the probability density functions (PDFs) of correct and incorrect matches following [8] and select thresholds for all descriptors according to the threshold criteria of their corresponding baselines. We use HPatches dataset [2] to compute the PDFs like D2-Net [5]. The PDFs are shown in Fig. 1.

We use ratio tests for matching DoG-based descriptors (*e.g.* RootSIFT [1], SOSNet [14] and our boosted SIFTs) like SIFT [8]. Specifically, we adopt Lowe's recommended threshold of 0.8 [8] for SIFT, RootSIFT and SIFT+Boost-B, while for SOSNet and SIFT+Boost-F we use a ratio threshold of 0.85.

We use distance tests instead of ratio tests for matching ORB [11] and ORB+Boosted-B descriptors since ratio tests do not work well for those descriptors. The selected distance thresholds are 45 and 50 respectively.

We use distance tests for matching SuperPoint [4] descriptors and use the same distance threshold of 0.7 as for HLoc [12]. We select 0.8 and 55 as the distance thresholds for matching SuperPoint+Boost-F and SuperPoint+Boost-B descriptors respectively.

Regarding the ALIKE-based descriptor, the distinctions between correct and incorrect matches in the PDF curves are unclear. We heuristically use a ratio threshold of 0.9 for both ALIKE [18] and our ALIKE+Boost-F, and a threshold of 0.88 for our ALIKE+Boost-B, which can retain 77.3%/77.6%/77.4% correct matches while filtering out 94.4%/94.6%/91.3% incorrect matches.

F. Qualitative examples

Fig. 2 and Fig. 3 show some matching results using different descriptors on Aachen Day-Night v1.1 [17] and InLoc [13].

References

[1] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, pages 2911–2918, 2012. 1, 2

[2] 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. 2, 4

[3] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The EuRoC micro aerial vehicle datasets. *Int. J. Robot. Res.*, 35(10):1157–1163, 2016. 1, 2

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, pages 224–236, 2018. 1, 2, 3

[5] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *CVPR*, pages 8092–8101, 2019. 2

[6] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[7] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018. 2

[8] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1, 2

[9] Raul Mur-Artal and Juan D Tardós. ORB-SLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017. 1, 2

[10] Udit Singh Parihar, Aniket Gujarathi, Kinal Mehta, Satyajit Tourani, Sourav Garg, Michael Milford, and K Madhava Krishna. RoRD: Rotation-robust descriptors and orthographic views for local feature matching. In *IROS*, pages 1593–1600, 2021. 2

[11] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *ICCV*, pages 2564–2571, 2011. 1, 2, 3

[12] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *CVPR*, 2019. 3

[13] 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 *CVPR*, pages 7199–7209, 2018. 1, 3, 5, 6

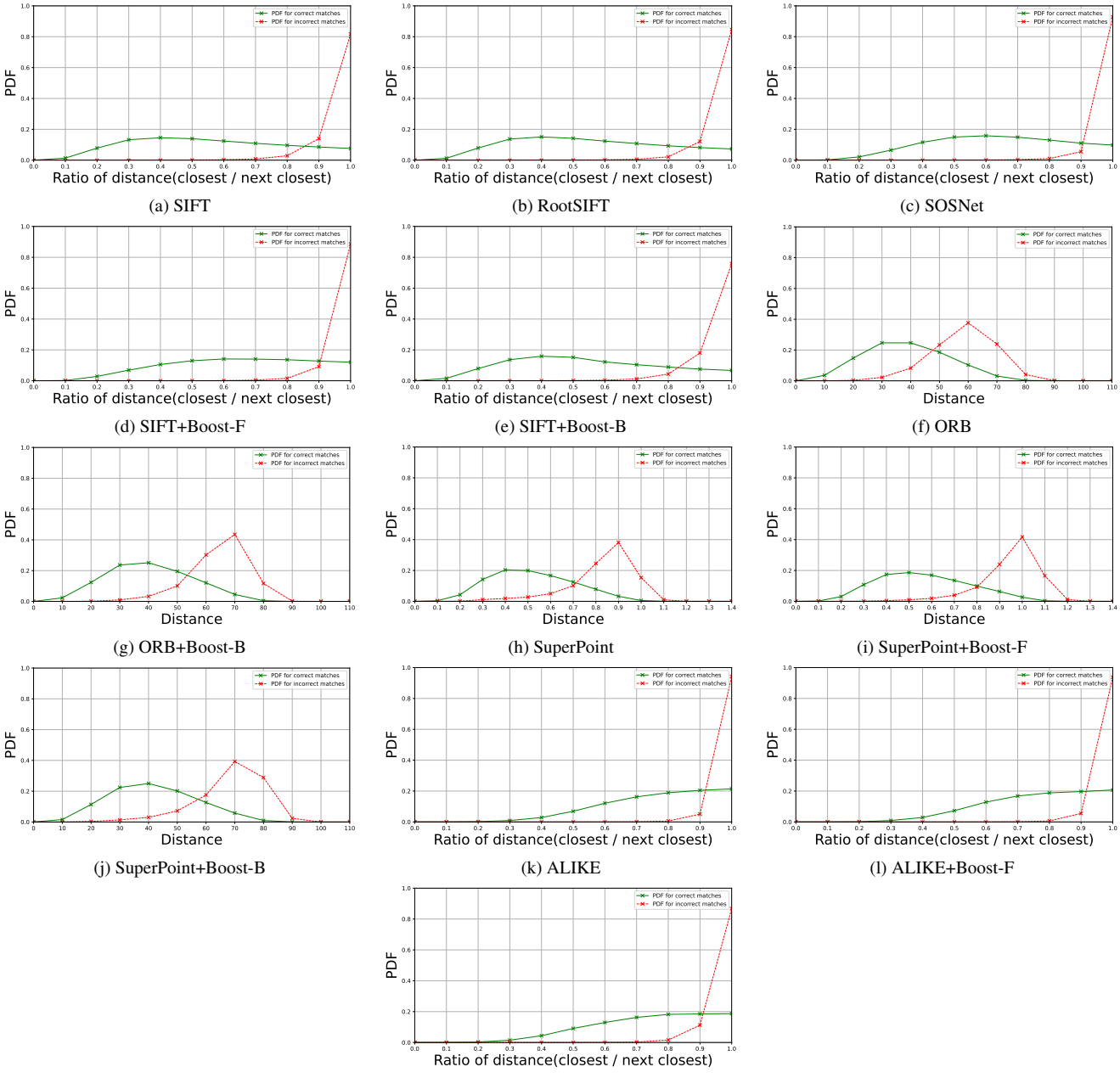
[14] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. SOSNet: Second order similarity regularization for local descriptor learning. In *CVPR*, pages 11016–11025, 2019. 1, 2

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2

[16] Shuangfei Zhai, Walter Talbott, Nitish Srivastava, Chen Huang, Hanlin Goh, Ruixiang Zhang, and Josh Susskind. An attention free transformer. *arXiv preprint arXiv:2105.14103*, 2021. 2

[17] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference pose generation for long-term visual localization via learned features and view synthesis. *IJCV*, 129(4):821–844, 2021. 1, 3, 5

[18] Xiaoming Zhao, Xingming Wu, Jinyu Miao, Weihai Chen, Peter CY Chen, and Zhengguo Li. ALIKE: Accurate and Lightweight Keypoint Detection and Descriptor Extraction. *IEEE TMM*, 2022. 1, 3



(a) SIFT (b) RootSIFT (c) SOSNet

(d) SIFT+Boost-F (e) SIFT+Boost-B (f) ORB

(g) ORB+Boost-B (h) SuperPoint (i) SuperPoint+Boost-F

(j) SuperPoint+Boost-B (k) ALIKE (l) ALIKE+Boost-F

(m) ALIKE+Boost-B

Figure 1. Ratio or distance PDFs for different descriptors. We compute the PDFs for all the descriptors using HPatches dataset [2]. For correct matches, the distance between the warp points and the keypoints is below 4 pixels. For incorrect matches, the distance is greater than 10 pixels.

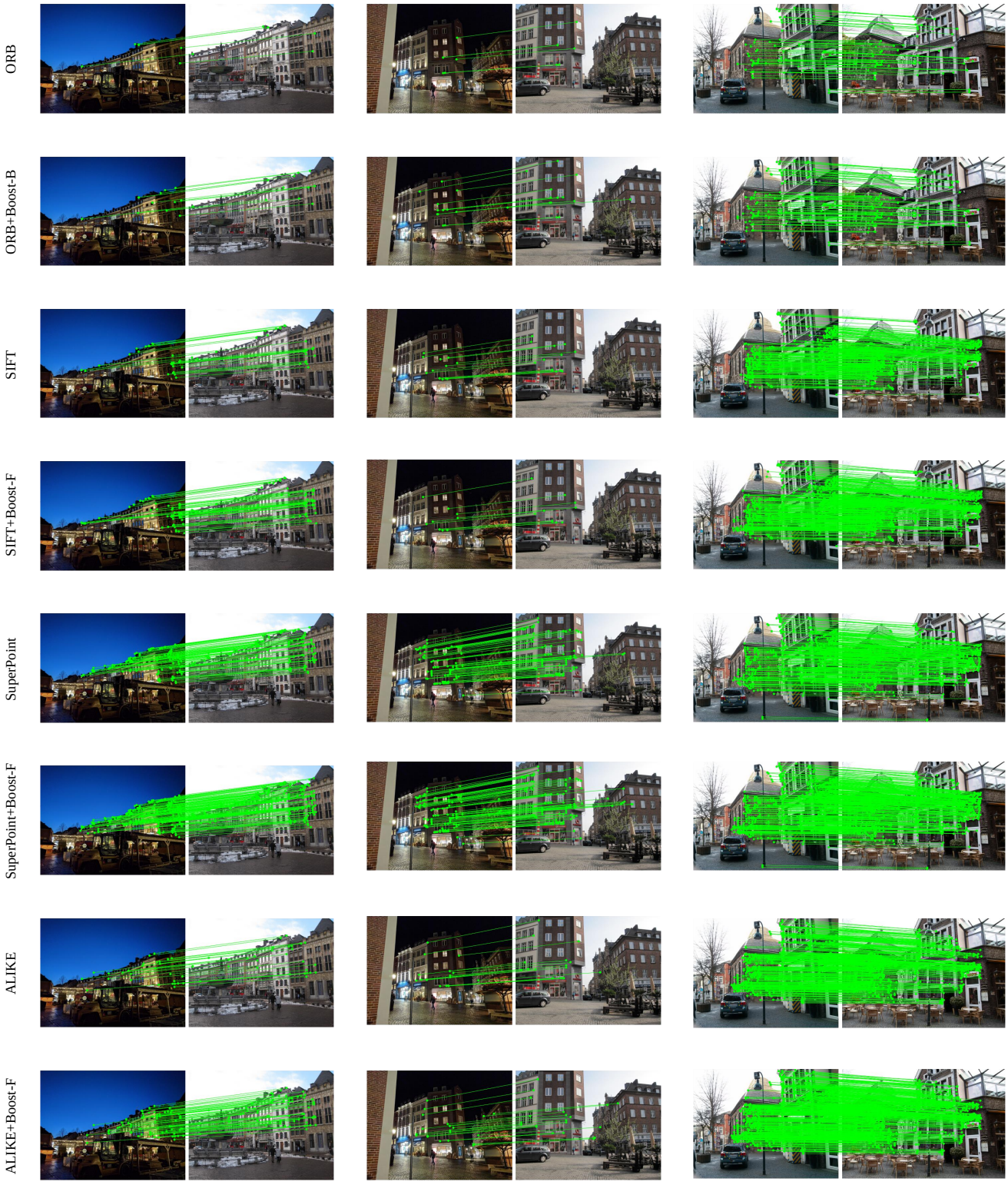


Figure 2. Matching results of using different descriptors on Aachen Day-Night v1.1 [17]. By boosting the original descriptors, our methods (represented by 'xxx+Boost-x') can produce more correct matches under significant changes in viewpoint and illumination. More results on InLoc [13] are shown in Fig. 3.

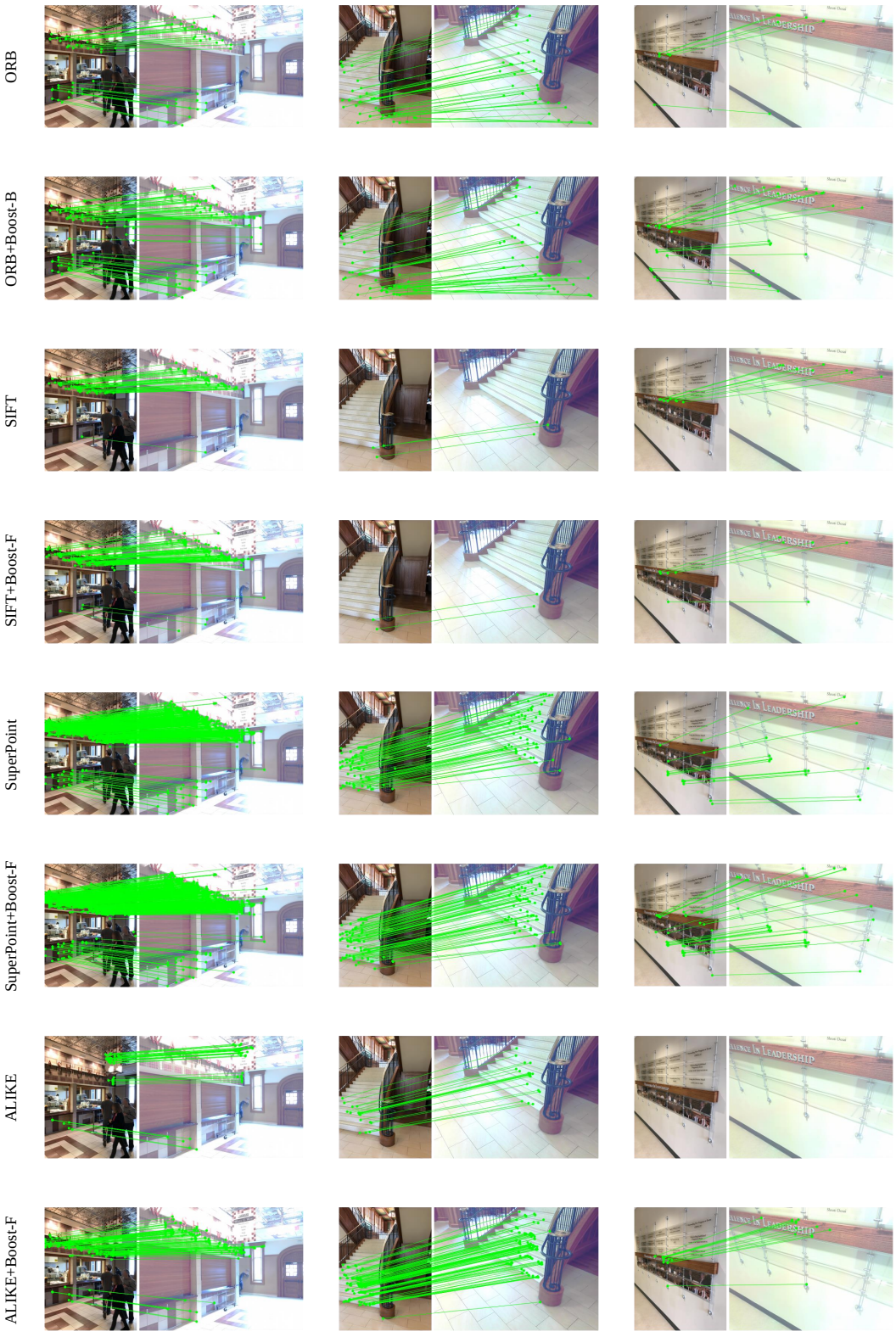


Figure 3. Matching results on InLoc dataset [13]. Our methods can boost the performance of descriptors under significant changes in viewpoint and texture-less areas. We also can see the failure case, where SIFT even performs worse than ORB and SIFT+Boost-F cannot improve the performance in those indoor scenes.