SFD2: Semantic-guided Feature Detection and Description Supplementary Material

Fei Xue Ignas Budvytis Roberto Cipolla
University of Cambridge
{fx221, ib255, rc10001}@cam.ac.uk

In the supplementary material, we first show localization results on Extended CMU-Seasons dataset in Table 1. Next, we give more qualitative comparison of previous and our methods on feature extraction and matching in Sec. A. Then we provide a detailed ablation study of our approach in Sec. B. Finally, we introduce the process of generating global stability and the architecture of our network in Sec. C and Sec. D, respectively.

A. Analysis of Feature Detection and Matching

In this section, we show more qualitative results of keypoints detection and matching in comparison with previous popular local features including SuperPoint [2], D2Net [3], R2D2 [10], and ASLFeat [9].

A.1. Feature detection

For each method, we detect top 1k keypoints with highest scores from the query images of Aachen_v1.1 dataset [14, 15] at the original resolution and visualize these keypoints with different colors according to their scores (high→low: 1-250, 251-500, 501-750, 751-1000). As shown in Fig. 1, we can see that:

- D2Net [3] and ASLFeat [9] favor regions with rich textures especially objects such as tress and pedestrians, partially because D2Net and ASLFeat adopt the similar detection strategy: spatial locations with high values of the high-level features. As a result, they detect many keypoints from objects *e.g.* sky, tree, car, pedestrian, which are not useful for long-term localization.
- R2D2 [10] detects keypoints almost uniformly from the whole image due to its maximization of responses in a fixed sliding window with size of 16×16. Therefore, R2D2 [10] also detects a large number of keypoints from unstable objects.
- SuperPoint [2] is a good corner detector. As corners also exist in objects *e.g.* sky, tree, car, pedestrian, SuperPoint [2] detects many keypoints from the aforementioned unstable objects.
- Our detector is partially supersized by results of SuperPoint, so it favors corners as well. Because we rerank the corners with the *stability* of semantic labels, our method prefers to detect keypoints from stable objects *e.g.* building, with more red keypoints from buildings. Although we can see keypoints from unstable objects, their scores are relatively smaller (with blue or yellow colors).
- The distribution of keypoints detected by prior methods and our model indicates that without explicit semantic labels, previous approaches don't perform well of selecting globally reliable keypoints although they are trained to detect keypoints which have strong discriminative ability.

A.2. Feature matching

We detect 4k keypoints for SuperPoint [2], R2D2 [10], ASLFeat [9], and our method and visualize the inliers between query and reference images with illumination changes, season variations, dynamic objects in the Aachen_v1.1 dataset [14, 15]. From Fig. 2, we can see that:

- For image pairs with small illumination and season changes, almost all methods could give many inliers.
- For image pairs with season changes or occlusions from trees or dynamic objects *e.g.* car, SuperPoint, R2D2, and ASLFeat give fewer inliers than our model.
- For extremely challenging image pairs with illumination changes, season variations, and high occlusions of trees, almost all prior approaches fail to give enough inliers, resulting in the failure of localization. However, our method is still able to find enough inliers from robust regions. We analyze the reasons of improvements in Sec. B.

Group	Method	urban	suburban	park	overcast	sunny	foliage	mixed foliage	no foliage	low sun	cloudy	snow
C	SIFT [8]	56.9/63.9/70.2	37.8/45.3/55.4	20.0/24.4/31.7	36.1/42.6/50.5	30.9/36.3/43.6	32.7/38.2/45.7	35.5/42.2/51.4	59.5/67.5/74.7	43.7/50.8/59.2	43.0/49.6/58.3	46.1/54.2/63.1
	AS [12]	81.0/87.3/92.4	62.6/70.9/81.0	45.5/51.6/62.0	64.1/70.8/78.6	55.2/62.3/71.3	58.8/65.3/73.9	59.2/67.5/77.4	83.3/88.9/94.6	65.8/73.4/82.8	71.6/77.6/84.2	73.0/81.0/90.5
	CSL [17]	71.2/74.6/78.7	57.8/61.7/67.5	34.5/37.0/42.2	52.2/55.4/60.3	43.3/46.6/51.9	47.0/50.2/55.3	52.4/56.1/62.0	80.3/83.2/86.6	61.7/65.3/70.7	63.3/66.3/70.5	69.9/73.7/78.7
S	VLM [18]	17.3/42.5/89.0	5.8/19.4/76.1	6.6/23.1/73.0	11.5/30.8/80.8	9.7/27.1/76.1	9.5/26.7/77.4	10.3/28.4/79.0	9.4/30.3/84.6	9.3/27.6/79.2	9.4/28.0/83.7	7.6/27.6/75.9
	SSM [16]	88.8/93.6/96.3	78.0/83.8/89.2	63.6/70.3/77.3	79.1/84.9/89.7	69.2/75.4/81.3	73.4/79.1/84.2	75.1/81.8/87.9	90.9/94.5/97.1	78.5/84.5/90.1	86.4/90.5/92.9	84.1/89.8/94.6
L	SPP [2]	89.5/94.2/97.9	76.5/82.7/92.7	57.4/ 64.4/80.4	77.1/82.8/91.8	65.1/72.3/86.8	69.2/ 75.5/88.3	75.2/81.7/90.8	88.7/92.8/96.4	78.0/83.9/91.8	83.4/87.7/94.0	80.7/86.6/93.2
	D2Net(MS) [3]	82.6/94.8/98.4	75.9/86.8/93.8	66.6/82.6/ 88.6	76.3/89.0/94.1	68.2/83.8/92.0	70.4/85.2/92.5	75.8/88.6/93.8	86.2/94.4/96.7	78.6/89.9/94.4	79.1/90.7/95.1	82.0/91.1/93.8
	R2D2 [10]	89.7/96.6/98.3	76.1/83.8/89.0	64.4/72.1/76.5	79.9/87.0/90.6	70.3/78.3/83.2	74.1/81.2/85.6	75.7/84.1/87.9	86.6/93.3/95.3	77.8/85.7/89.3	84.1/90.0/92.5	79.8/87.6/91.1
M	PixLoc [20]	92.8/95.1/98.5	91.9/93.4/95.8	84.0 /85.8/90.9	90.3/92.2/96.2	85.3/88.8/94.0	87.1/89.9/94.7	90.5/91.9/95.1	95.1/95.7/96.8	91.2/92.3/94.8	93.9/94.8/97.4	91.6/92.3/94.0
	AHM [4]	65.7/82.7/91.0	66.5/82.6/92.9	54.3/71.6/84.1	62.8/78.8/89.4	56.6/74.5/87.2	58.5/75.7/87.8	62.9/79.6/89.4	72.0/87.7/94.5	64.0/81.0/90.2	69.4/84.4/92.8	61.7/80.6/90.3
	SPP+SPG [2, 11]	95.5/98.6/99.3	90.9/94.2/97.1	85.7/89.0/91.6	92.3/95.3/96.9	86.1/91.3/94.6	88.3/92.5/95.3	91.6/94.5/96.2	95.4/97.1/98.3	91.8/94.4/96.3	95.2/97.0/98.0	92.3/94.6/96.6
	Ours	95.0/97.5/98.6	90.5/92.7/95.3	86.4/89.1/91.2	92.1/94.0/95.8	86.3/90.3/93.4	87.9/91.0/93.9	91.9/94.0/95.5	95.3/96.6/97.6	92.4/94.4/95.8	93.3/94.7/96.3	92.9/94.6/96.0

Table 1. **Localization accuracy on the Extended CMU-Seasons dataset [13].** Results at error thresholds of $(0.25m, 2°), (0.5m, 5°), (5m, 10°)$ are reported.

B. Ablation Study of Feature Detection and Matching

In this section, we verify the efficacy of the proposed semantic-aware detection (SD), semantic-aware description (SS), and semantic-consistency (SF) losses by visualizing the detection and matching results. The base model is trained with results of SuperPoint [2] as supervision and a general ap loss [5] for descriptor learning as R2D2 [10]. Our full model comprises SD, SS, and SF three components.

B.1. Ablation study of detection

As in Sec. A.1, we visualize 1k keypoints with the highest scores and show them with different colors according to their scores (high→low: 1-250, 251-500, 501-750, 751-1000). As shown in Fig. 1, we can see the effectiveness of SD, SS, and SF in detail:

- Our base model performs closely to SuperPoint [2] (as shown in Fig. 1) with high response to corners as the detector is partially supervised with results of SuperPoint [2]. Meanwhile, the base model is also sensitive to unstable objects *e.g.* sky, tree, pedestrian, and car.
- The SD loss (W/ SD) is the key to rerank the keypopints. With SD loss, keypoints from unstable objects *e.g.* sky, car, pedestrian are suppressed. Keypoints from trees have lower score (with color of blue or yellow) and keypoints from stable objects *e.g.* building are favored (with color of red).
- The SS loss doesn't contribute to the detection process, so it shows the similar results as the base model, which again indicates that the importance of explicit semantic labels to detection as discussed in Sec. A.1,
- The full model with SF incorporated performs better than the model W/ SD, as it further enhances the ability of our model in learning semantic-aware features.

B.2. Ablation study of feature matching

We additionally visualize the effectiveness of SD, SS, SF losses in feature matching. From Fig. 5, we can see that:

- Benefiting from the corner detector and ap loss, the base model is already able to give promising performance in comparison with previous methods [2, 9, 10].
- The SD loss (W/ SD) marginally improves the matches possibly because those reranked keypoints from stable objects don't have strong discriminative ability by purely adopting ap loss over all keypoints.
- The SS loss (W/ SS) effectively solves the limitation of SD loss, as it augments the discriminative ability of descriptors with semantics.
- The full model gives the best performance because it combines the advantages of SD, SS, and SF losses.

C. Global stability map generation

During the training process, we utilize UperNet [1] with ConvNet [7] as encoder trained on ADE20k [19] dataset to provide semantic segmentation labels and high-level features for semantic-wise and feature-wise guidance, respectively. There are 150 labels in total which are categorized into 4 groups as shown in Table 2. Since large-scale localization happens mainly in outdoor environments, only several objects such as sky, water, pedestrian, car, tree, plant, and building are frequently used.

D. Network

Alike to SuperPoint [2], we adopt 8 times downsampling to reduce the resolution of high-dimension features, making the model efficient at test time. To increase the representation ability of our model, we introduce 3 ResBlocks [6]. Details of the network are shown in Fig. 3.

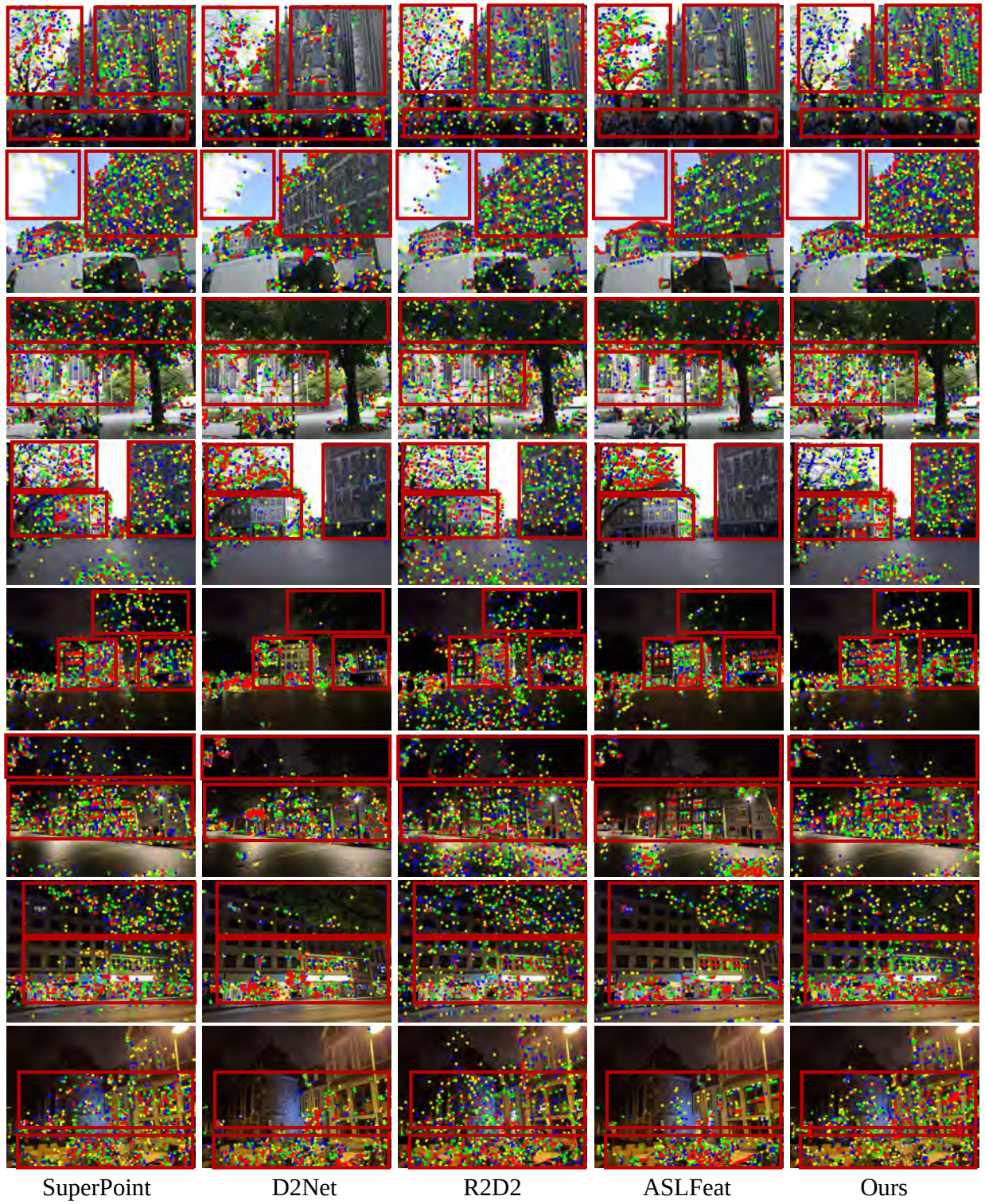

Figure 1. **Comparison of feature detection.** We show top 1k keypoints with highest scores (high→low: 1-250, 251-500, 501-750, 751-1000) of prior SOTA local features including SuperPoint [2], D2Net [3], R2D2 [10], and ASLFeat [9]. They are more sensitive to regions with rich textures even those from objects *e.g.* sky, tree, pedestrian, car, which are unstable for long-term localization. By introducing the semantics for reranking keypoints, our model prefers keypoints from stable objects *e.g.* building.

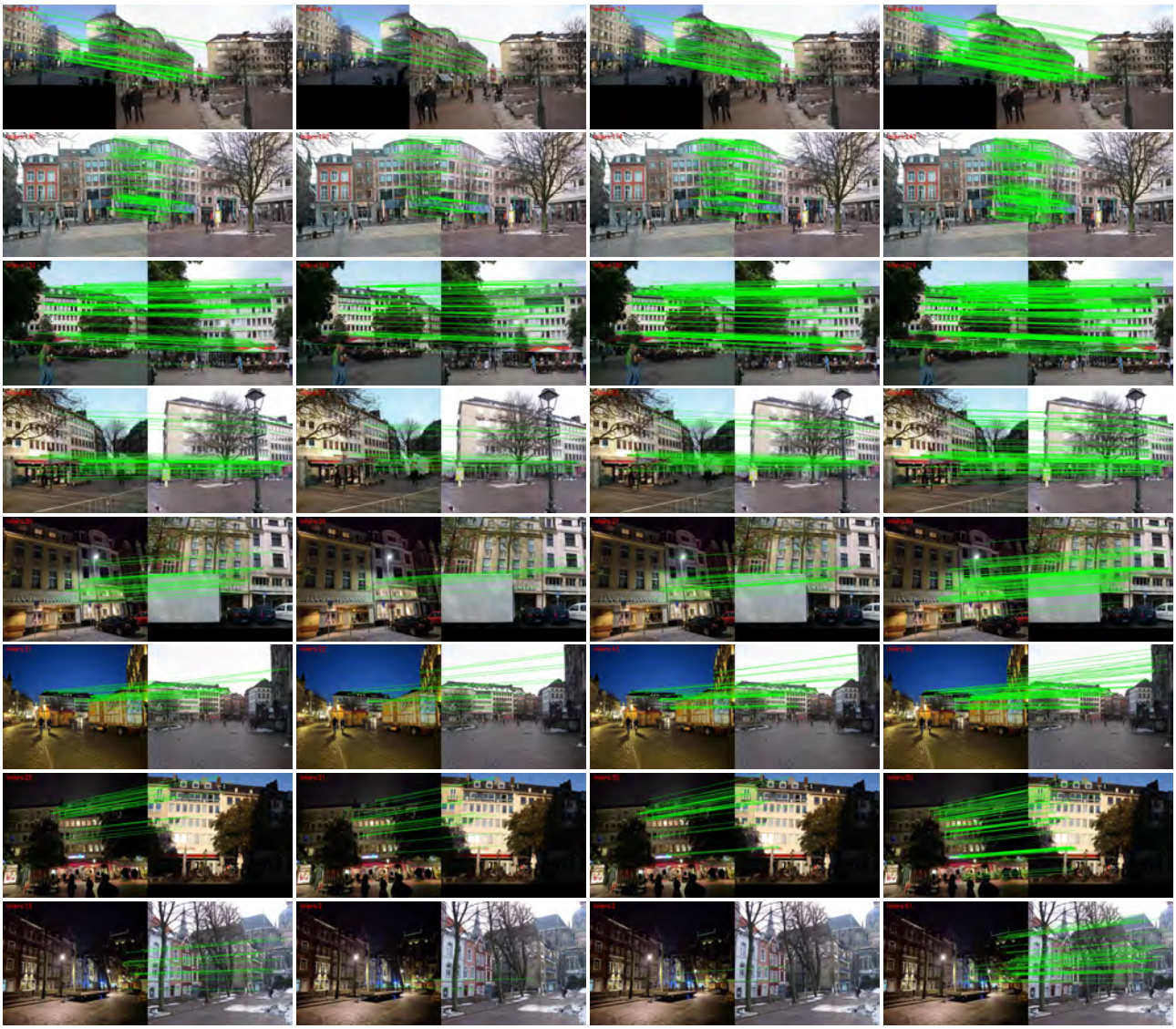

Figure 2. **Comparison of feature matching.** We show inliers between query and reference images from the Aachen_v1.1 [14, 15] dataset under challenges of illumination changes, season variations, and dynamic objects. Results of SuperPoint [2], R2D2 [10], ASLFeat [9], and our model are visualized. Compared with prior methods, our model is able to produce more inliers even under extremely challenging conditions when others fail to give enough inliers to guarantee the success of localization.

Category	Semantics
Volatile	sky, mountain, curtain, water, sea, mirror, rug, field, bathtub, stand, sand, sink, river, hill, bench,light,dirt,land,fountain, swimming pool, waterfall, lake
Dynamic	person, automobile, boat, truck
Short-term	tree, grass, plant, flower, palm, airplane, van, ship, minibike, bike, shower
Long-term	wall, building, floor, ceiling, road, bed, window, cabinet, sidewalk, ground, door, chair, painting, sofa, shelf, house, armchair, seat, fence, rock, wardrobe, lamp, rail, cushion, box, pillar, signboard, chest, counter, skyscraper, fireplace, grandstand, path, stairs, runway, case, table, pillow, screen, stairway, bridge, bookcase, toilet, book, countertop, stove, kitchen, computer, swivel, bar, arcade, hovel, tower, chandelier, sunshade, streetlight, booth, television, clothes, pole, bannister, escalator, ottoman, bottle, buffet, poster, stage, conveyer, canopy, washer, toy, stool, cask, basket, tent, bag, cradle, oven, ball, food, step, tank, trade name, pot, dishwasher, screen, blanket, sculpture, hood, sconce, vae, traffic light, tray, dustbin, plate, monitor, bulletin, glass, clock, flag

Table 2. **Stability map of different labels**. All semantic labels are categorized into four groups denoted as *Volatile*, *Dynamic*, *Short-term*, and *Long-term* according to their reliability in the visual localization task.

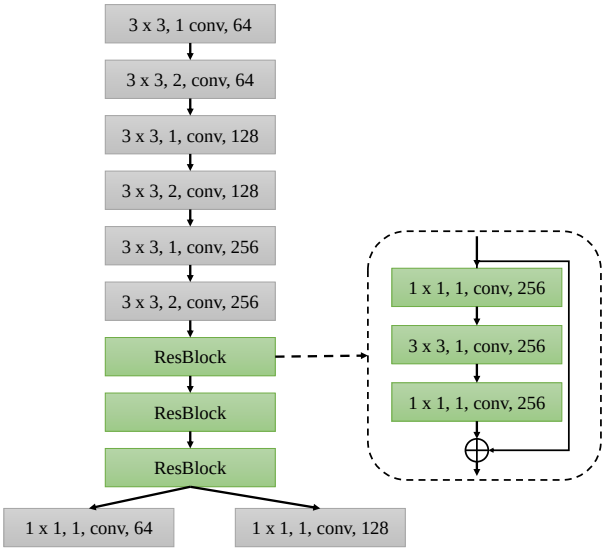

Figure 3. **Architecture of the network.** We adopt 6 Convolution layers with kernel size of 3×3 to generate high-level features with $8\times$ downsampling (implementation by using stride of 2). Then 3 ResBlocks [6] are followed to further enhance the ability of the model.

Figure 4. **Ablation study of feature detection.** We show top 1k keypoints with the highest scores (high→low: 1-250, 251-500, 501-750, 751-1000) of our base model, model with SD loss (W/ SD), SS loss (W/ SS) and the full model (with SD, SS, SF). The base model is more sensitive to regions with rich corners as SuperPoint [2]. SD loss effectively mitigates this problem by introducing semantic-aware detection loss. SS loss focus mainly on descriptor learning, so it gives similar results to the base model. The full model additionally introduces SF loss, which further enhances the detection process.

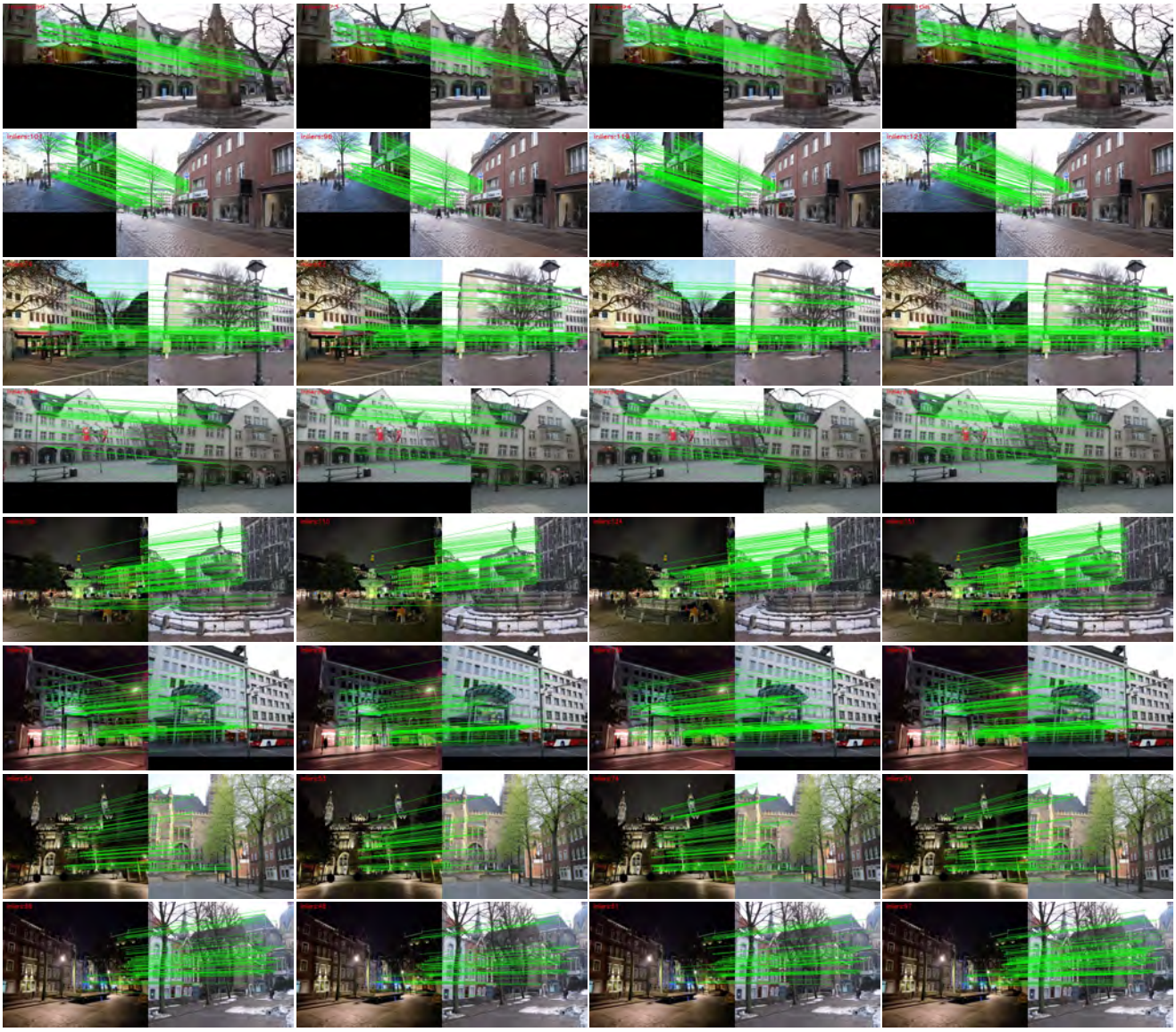

Figure 5. **Ablation study of feature matching.** We show the inliers between query and reference images from the Aachen_v1.1 [14, 15] dataset under challenges of illumination changes, season variations and dynamic objects. Results of the base model, with SD loss (W/ SD), with SS loss (W/ SS) and the full model (with SD, SS, SF) are visualized. SD loss slightly improves the matching as it focus mainly the detection process. SS loss effectively augments the matching accuracy by introducing semantic labels. Results of SS loss are further improved by the full model, which has an additional SF loss to enhance the model's ability of learning semantic-aware features.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2

[2] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, 2018. 1, 2, 3, 4, 6

[3] 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*, 2019. 1, 2, 3

[4] Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. Sparse-to-dense hypercolumn matching for long-term visual localization. In *3DV*, 2019. 2

[5] Kun He, Yan Lu, and Stan Sclaroff. Local descriptors optimized for average precision. In *CVPR*, 2018. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 5

[7] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 2

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

[9] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. ASLFeat: Learning local features of accurate shape and localization. In *CVPR*, 2020. 1, 2, 3, 4

[10] Jerome Revaud, Philippe Weinzaepfel, César Roberto de Souza, and Martin Humenberger. R2D2: Repeatable and reliable detector and descriptor. In *NeurIPS*, 2019. 1, 2, 3, 4

[11] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2

[12] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized for large-scale image-based localization. *TPAMI*, 2016. 2

[13] 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*, 2018. 2

[14] 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*, 2018. 1, 4, 7

[15] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image retrieval for image-based localization revisited. In *BMVC*, 2012. 1, 4, 7

[16] Tianxin Shi, Shuhan Shen, Xiang Gao, and Lingjie Zhu. Visual localization using sparse semantic 3D map. In *ICIP*, 2019. 2

[17] Linus Svärm, Olof Enqvist, Fredrik Kahl, and Magnus Oskarsson. City-scale localization for cameras with known vertical direction. *TPAMI*, 2016. 2

[18] Zhe Xin, Yinghao Cai, Tao Lu, Xiaoxia Xing, Shaojun Cai, Jixiang Zhang, Yiping Yang, and Yanqing Wang. Localizing discriminative visual landmarks for place recognition. In *ICRA*, 2019. 2

[19] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019. 2

[20] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2pix: Epipolar-guided pixel-level correspondences. In *CVPR*, 2021. 2