Raising the Ceiling: Conflict-Free Local Feature Matching with Dynamic View Switching

Supplementary Material

The supplementary material is summarized as follows: Section A presents additional qualitative and quantitative experiments. Section B offers additional details and insights into the design of networks.

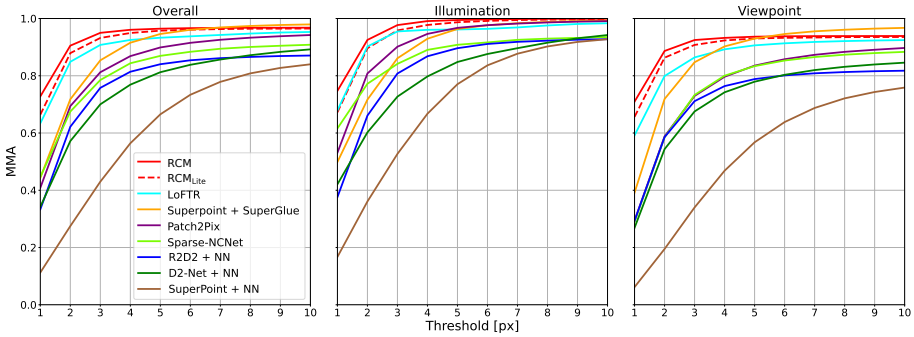

Fig. S1: Image matching on HPatches [1]. MMA curves are plotted by changing the reprojection error threshold.

A Additional Experiments

A.1 Image Matching

Dataset. Following the evaluation protocol introduced in D2-Net [3], we evaluate the performance of our method over 108 HPatches [1] sequences, which include 52 instances with illumination variations and 56 instances with viewpoint changes.

Metric. We compute the reprojection error of each match from the homographies provided by the HPatches dataset. The matching threshold is varied from 1 to 10 to visualize the mean matching accuracy (MMA), which is the average percentage of correct matches for each image.

Results. As shown in Fig. S1, our method RCM achieves the best accuracy at thresholds less than or equal to 6, and the RCM_{Lite} outperforms the dense method LoFTR at all thresholds. Dense and semi-sparse methods demonstrate significantly superior accuracy at lower matching thresholds. This advantage stems from their ability to produce precise matches at the sub-pixel level within the target image, independent of the imprecision of the keypoints. Compared to the dense method LoFTR, the semi-sparse matching paradigm with many-to-one matching and switcher can further improve the matching accuracy as it yields more precise matching points in the source image through detection.

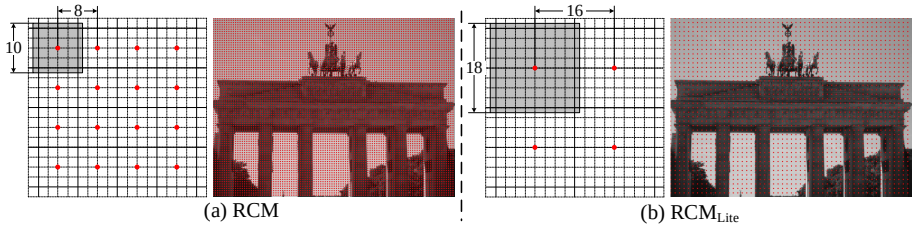

Fig. S2: Coarse grids (red points) and fine matching windows (gray windows) in the target image.

A.2 More Qualitative Results

Qualitative Ablation on Coarse Ground-Truth Matches. In Fig. S3 and Fig. S4, we illustrate how many-to-one matching and the switcher benefit both dense and semi-sparse matching paradigms by resolving matching conflicts in the target image and increasing matchable points in the source image. The substantial increase in ground-truth matches enhances the theoretical upper bound on the actual matching results produced by the matcher.

Qualitative Ablation on Actual Matching Results. Additional comparisons between one-to-one and many-to-one matching strategies are presented in Fig. S5. The many-to-one matching strategy substantially increases the number of matches and enhances accuracy by resolving conflicts in the coarse matching phase.

In Fig. S5, further ablation comparisons for the switcher are provided. These comparisons consistently demonstrate that the switcher significantly increases the number of matches, consequently enhancing the performance of downstream tasks such as pose estimation.

Qualitative Comparisons of Sparse, Dense and Semi-Sparse Methods. In Fig. S6, we present additional qualitative comparisons of three matching paradigms in both outdoor and indoor scenes. In outdoor scenes, the superiority of our approach over sparse [5] and dense [7] methods in terms of accuracy and match quantity is evident from the first row. This advantage stems from the semi-sparse paradigm, which extracts precise keypoints from the source image and conducts a global search within the target image. The last two rows highlight that our proposed switcher, responsible for switching larger scale images to the source image, significantly enhances the number of matches. This improvement is attributed to our improved ability to detect more matchable keypoints within the overlapping region.

In indoor scenes, the proposed semi-sparse matching method, RCM, consistently produces superior results. In contrast to the sparse method, RCM achieves a significantly higher number of matches by mitigating reliance on keypoint repeatability, leading to improved pose estimation performance. Compared to the dense method, RCM excels in detecting keypoints at more discriminative positions, resulting in significantly higher matching precision.

Additional indoor and outdoor qualitative comparisons are presented in Fig. S7 and Fig. S8, where matched points are color-coded for clarity.

Matching Visualizations in More Datasets. Additional matching results of RCM on two distinct datasets, namely the Aachen Day-Night v1.1 dataset [8] and the HPatches dataset [1], are presented in Fig. S9 and Fig. S10.

Visualizations of the Dustbin and Attention Weights. Fig. S11(a) illustrates the role of the dustbin, designed to discard non-matchable points in non-overlapping regions, enabling RCM to effectively handle occlusions and viewpoint changes. Additionally, visualizations of self-attention weights and cross-attention weights are provided in Fig. S11(b) and (c), respectively.

Failure Cases. We present the failure cases of RCM in Fig. S12. These instances occur in outdoor scenes with severe scale changes and misclassification of the switcher. Failures also occur in indoor scenes featuring extensive texture-less areas and substantial viewpoint changes, where the detector struggles to produce discriminating keypoints.

3D Reconstruction Results. The HLoc pipeline [4] is employed for 3D reconstruction based on the matching results of RCM, followed by dense reconstruction using COLMAP [6]. The sparse and dense models of three landmarks are illustrated in Fig. S13.

B More Details

U-Net Feature Extraction. The encoder of the U-Net network inherits the SuperPoint [2] encoder, producing feature maps at resolutions of 1/2, 1/4, and 1/8. In RCM, we design a similar VGG-like structure for the decoder, progressively integrating information from the encoder to generate fine features at 1/2 resolution. RCM_{Lite} takes an additional step by incorporating 1/16 resolution encoding and decoding layers, accounting for its increased parameter count compared to RCM. RCM combines 1/2, 1/4, and 1/8 resolution decoder features linearly to form coarse features, while RCM_{Lite} includes 1/2, 1/4, 1/8, and 1/16 resolution decoder features to generate coarse features. The feature dimensions of the 1/2, 1/4, 1/8, and 1/16 resolution maps are 64, 128, 256, and 256, respectively.

Detection. In outdoor scenes, we detect keypoints with NMS radius of 4 pixels and keypoint threshold of 0.005, which are the default settings of SuperPoint [2]. In indoor scenes, we adjust the parameters to an NMS radius of 1 pixel and a keypoint threshold of 0.001 to enhance keypoint detection in low-texture areas.

Coarse Feature Resolution and Fine Matching Window. As illustrated in Fig. S2, RCM and RCM_{Lite} utilize coarse feature maps at resolutions of 1/8 and 1/16, respectively. Consequently, each coarse feature corresponds to an 8×8 pixel patch for RCM and a 16×16 pixel patch for RCM_{Lite}. To ensure complete coverage of the coarse feature patch by the fine matching window, window sizes of 10×10 and 18×18 pixels are designed for RCM and RCM_{Lite}, respectively. Given that the fine feature map is at 1/2 image resolution, the fine matching window sizes of RCM and RCM_{Lite} are $w = 5$ and $w = 9$, as discussed in the main text.

Switcher. Initially, both sets of features undergo down-sampling to $(H, W) = (20, 20)$ using adaptive average pooling to optimize subsequent computations. The correlation map $C \in \mathbb{R}^{20\times20\times20\times20}$ is then computed by the inner product, capturing the similarity of individual patches between the two images. The correlation map is reshaped into $\hat{C} \in \mathbb{R}^{20\times20\times400}$ and processed by a lightweight CNN, which extracts features through two Conv-BN-ReLu-MaxPool layers. We subsequently reduce the spatial dimension to 1 with adaptive average pooling and the channel dimension to 2 with linear layer. Softmax is applied to compute the switching confidence VS, triggering feature switching when $VS > 1/2$.

References

1. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: Proceedings of the CVPR. pp. 5173–5182 (2017)
2. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-Supervised Interest Point Detection and Description. In: Proceedings of the CVPRW. pp. 224–236 (2018)
3. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-Net: A Trainable CNN for Joint Description and Detection of Local Features. In: Proceedings of the CVPR. pp. 8092–8101 (2019)
4. Sarlin, P., Cadena, C.and Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: Proceedings of the CVPR. pp. 12716–12725 (2019)
5. Sarlin, P., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning Feature Matching With Graph Neural Networks. In: Proceedings of the CVPR. pp. 4937–4946 (2020)
6. Schonberger, J.L., Frahm, J.: Structure-from-motion revisited. In: Proceedings of the CVPR. pp. 4104–4113 (2016)
7. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-Free Local Feature Matching With Transformers. In: Proceedings of the CVPR. pp. 8922–8931 (2021)
8. Zhang, Z., Sattler, T., Scaramuzza, D.: Reference pose generation for long-term visual localization via learned features and view synthesis. International Journal of Computer Vision pp. 821–844 (2021)

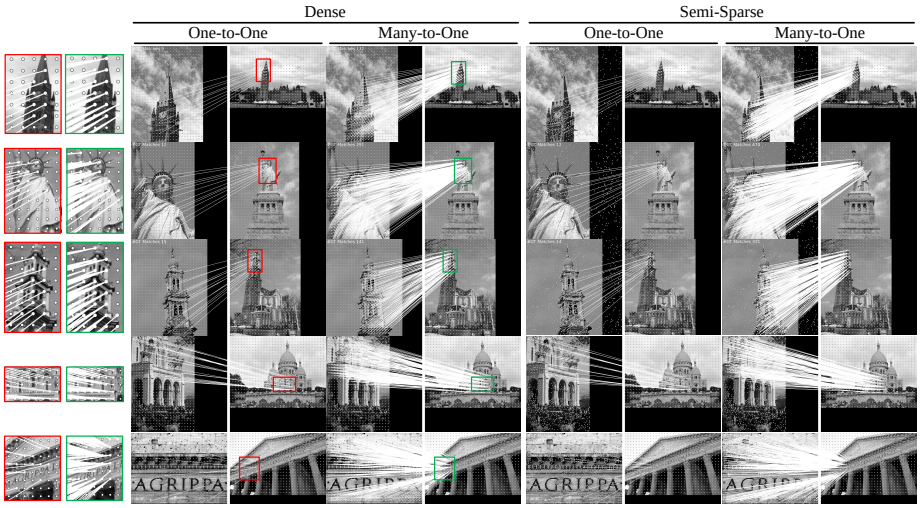

Fig. S3: Qualitative comparison of coarse ground-truth matches between ono-to-one and many-to-one matching. Many-to-one matching (green box) resolves the problem of matching conflicts in the target image, resulting in a greater number of ground-truth matches compared to one-to-one matching (red box). Note that we only show the ground-truth for the coarse matching stage.

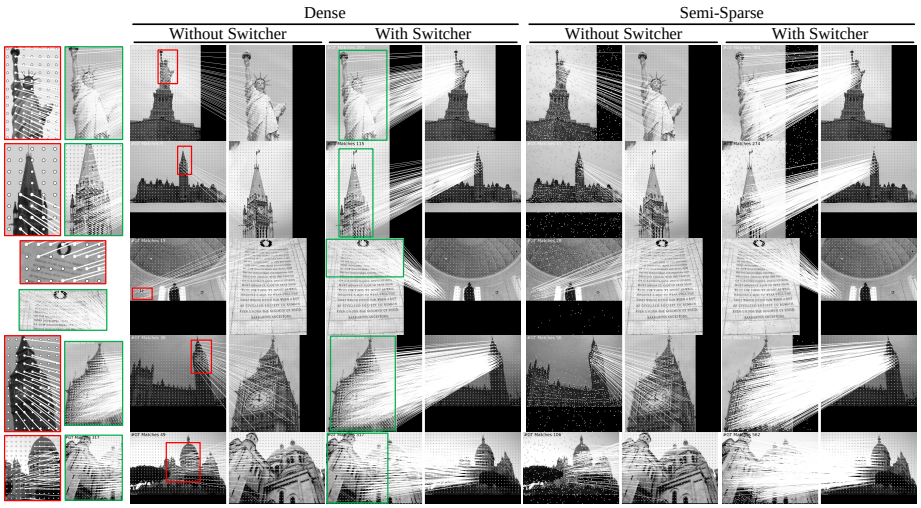

Fig. S4: Qualitative comparison of coarse ground-truth matches with and without switcher. The switcher resolves the problem of a shortage of matchable points (red box) in the source image, acquiring a significantly greater number of matchable points (green box) through the strategic switching of the two images.

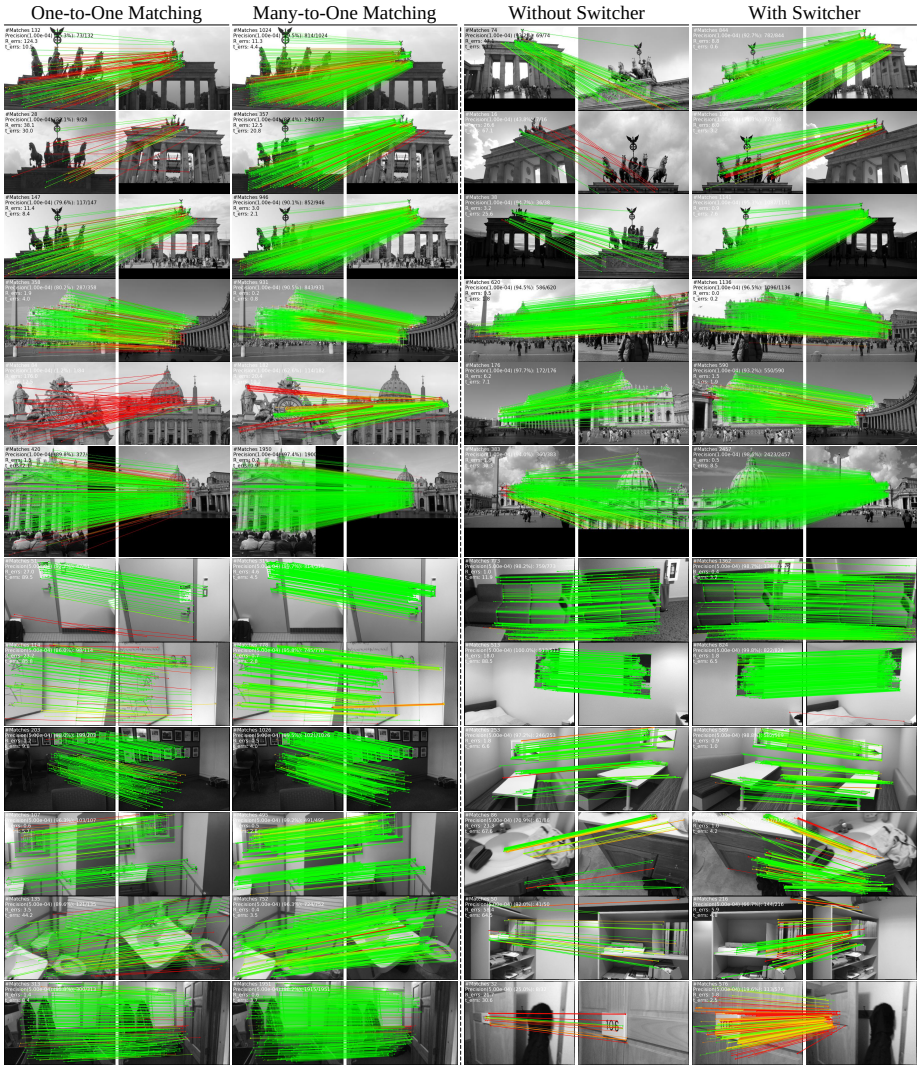

Fig. S5: Qualitatively results of many-to-one matching and switcher.

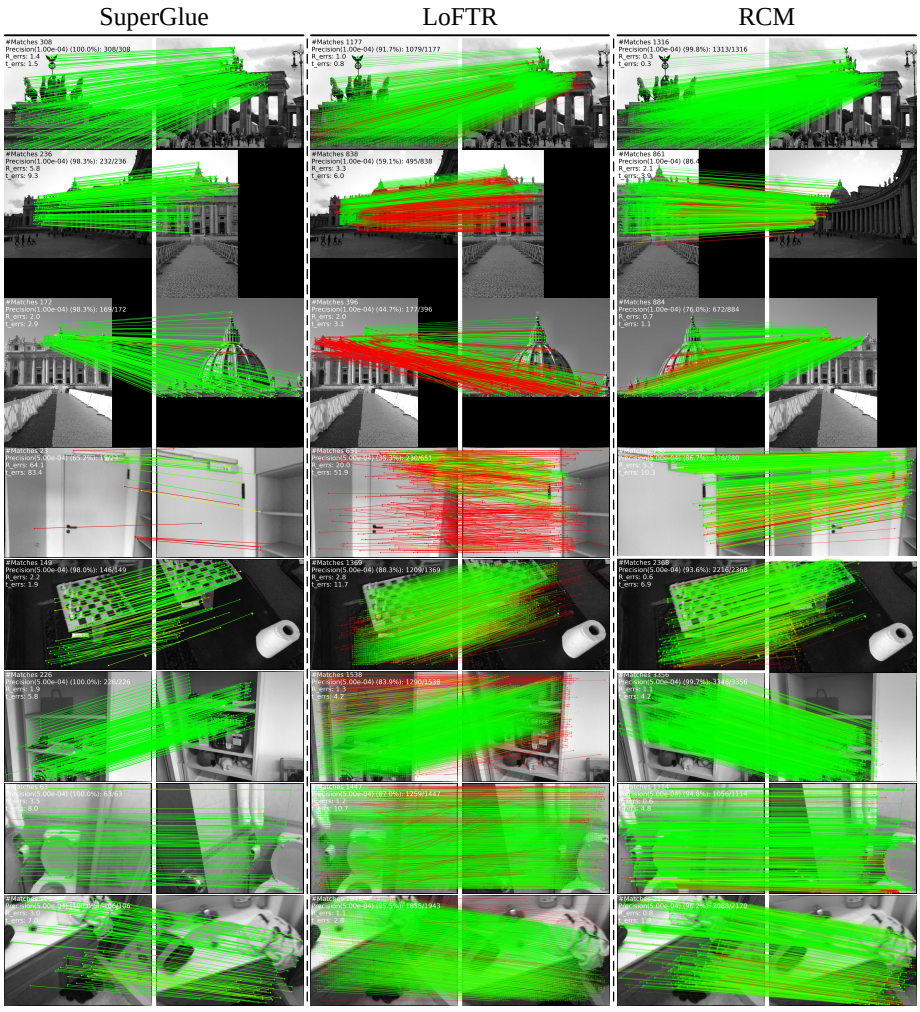

Fig. S6: Qualitative comparison in outdoor scenes. The semi-sparse matching method RCM consistently generates superior matches in both qualitative and quantitative aspects.

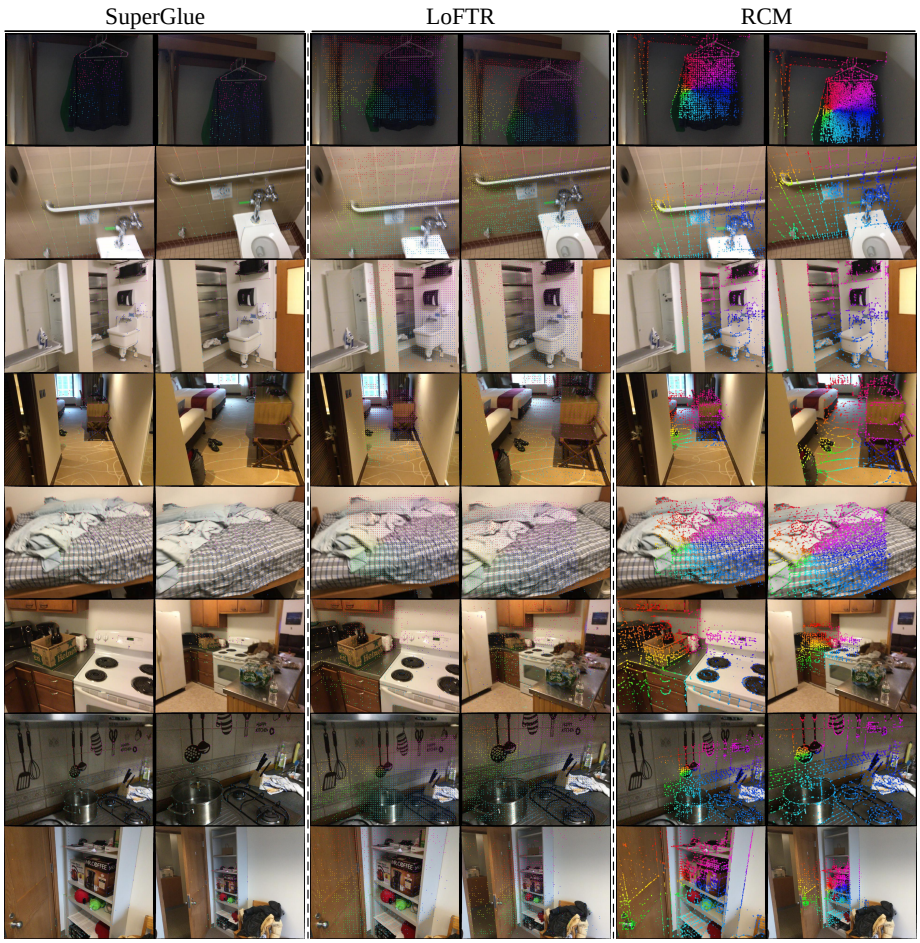

Fig. S7: More qualitative comparisons in indoor scenes. The matched features are visualized as the same color.

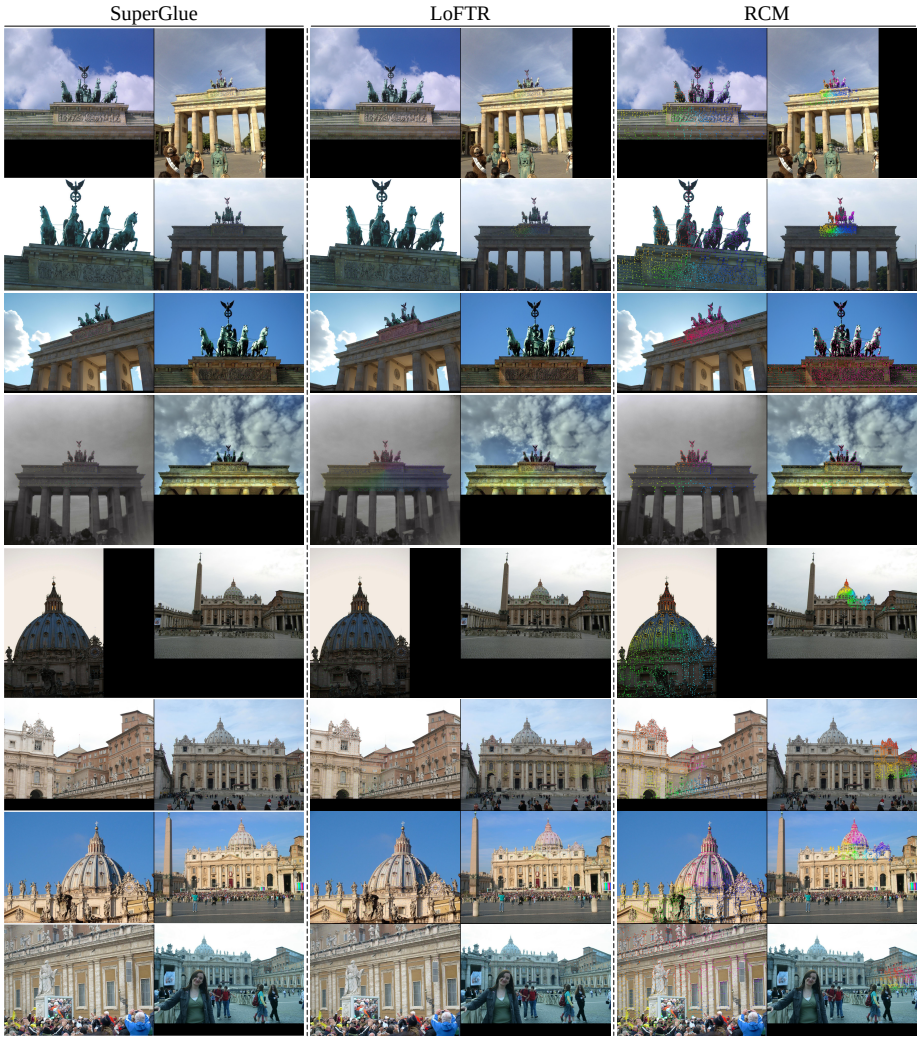

Fig. S8: More qualitative comparisons in outdoor scenes. The matched features are visualized as the same color.

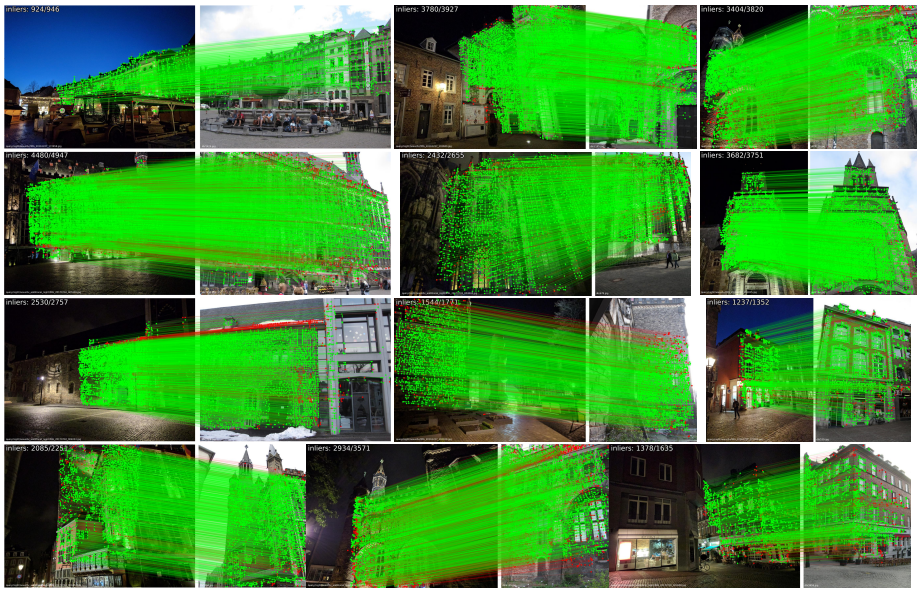

Fig. S9: Qualitative results of RCM in Aachen Day-Night dataset [8].

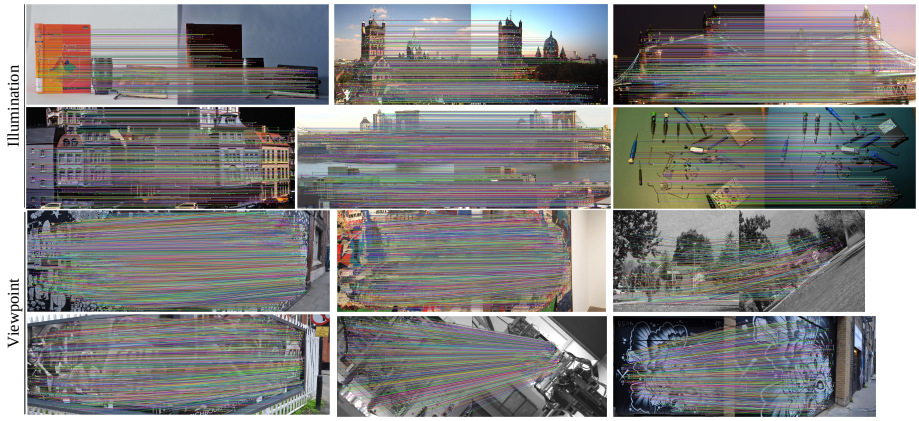

Fig. S10: Qualitative results of RCM in HPatches dataset [1].

Fig. S11: Visualizations of the dustbin and attention weights.

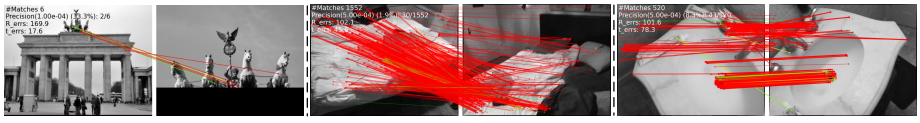

Fig. S12: Failure cases.

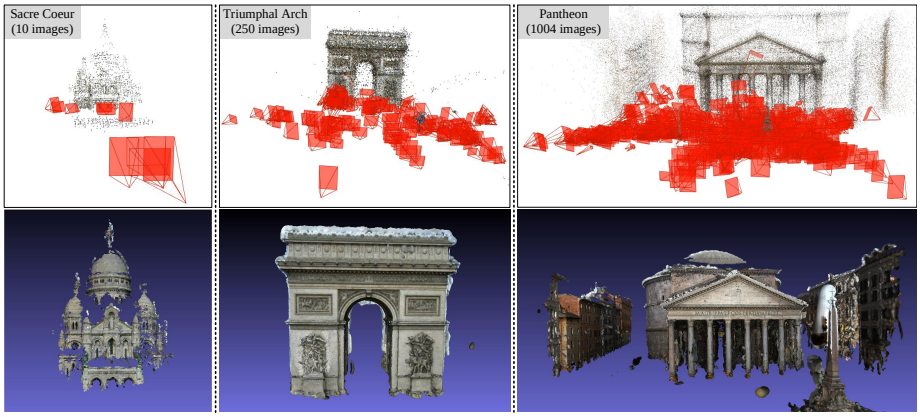

Fig. S13: 3D reconstruction results based on RCM.