IMP: Iterative Matching and Pose Estimation with Adaptive Pooling Supplementary Material

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

In the supplementary material, we provide additional implementation details, qualitative results and analysis in Sec. A and Sec. B, respectively.

A. Implementation

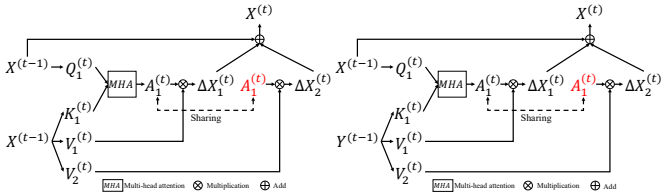
A.1. Training

We train our model on the MegaDepth dataset [7] from scratch. Following SuperGlue [9], we use 153 scenes with 130k images in total for training and 36 for validation. For each category, we first detect 4096 keypoints for all images and then build correspondences for image pairs with overlap ratio from 0.3-1.0. Matches with re-projection errors less than 5px are deemed as inliers, resulting in different number of inliers for different pairs. In the training process, for each epoch, we randomly choose 80 pairs for each scene. For each pair, we randomly choose 1024 keypoints with inliers ranging from 32 to 512 between two images, respectively. We observe that samples with high inlier ratios boost the convergence and those with low inlier ratios enhance the ability of models for finding matches for tough cases at test time.

Our matching loss is identical to the assignment loss of SGMNet [3], which is more stable than the original version in SuperGlue [9], as analyzed in the SGMNet paper. These modifications enable us to train the model on the MegaDepth dataset [7] from scratch without requiring any pretraining.

A.2. Architecture

As [3, 9, 11], we use self and cross attention to gather global information for each keypoint in two sets. Considering the similarities of attention matrices in two consecutive iterations, we adopt the shared attention mechanism [2] to speed up the message propagation process at low cost. A detailed architecture of self and cross attention is shown in Fig. 1. We adopt the identical position encoder as SuperGlue [9].



(a) Self attention (b) Cross attention

Figure 1. **Architecture.** Detailed architecture of self (a) and cross (b) attention with sharing attention matrix.

A.3. Inference

At test time, for relative pose estimation on YFCC100m [12] and Scannet [4] datasets, we use the identical testing pairs and number of keypoints for evaluation as SuperGlue [9] and SGMNet [3]. As for the metrics, we report the **exact** cumulative error curve (AUC) rather than the **approximate** one, because the former measures the error based on groundtruth poses while the latter does it based on predicted poses.

When evaluating our model on Aachen Day-Night v1.0 and v1.1 datasets [10, 13], we adopt HLoc [8] pipeline for mapping and localization, as previous methods [3, 9, 11]. We leverage NetVLAD [1] to provide 50 reference images for each query image in the localization process.

In our adaptive sampling process, in order to avoid losing too many potential inliers or informative keypoints, we set the minimum number of preserved keypoints for each image to 256. Once the number of keypoints is smaller or equal to 256, we don't perform any sampling. This strategy is also applied to the ratio-based sampling (R50).

B. Results

In this section, we first provide more ablation studies in Sec. B.1. Then, we show and discuss more qualitative results of SuperGlue* (official SuperGlue) [9], SGMNet [3] and our IMP and EIMP on Scannet [4], YFCC100m [12], and Aachen Day-Night datasets [10, 13] in Sec. B.2, Sec. B.3, and Sec. B.4, respectively.

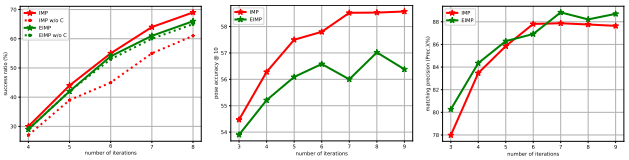
B.1. Ablation study

Pose-consistency loss (C). Pose-consistency loss forces our model to predict matches which are not only correct but also able to give a good pose by implicitly embedding geometric information into the matching module. Fig. 2a shows the influence of pose-consistency loss to the number of iterations required to find a good pose. With pose-consistency loss, as expected, when using the same number of iterations, IMP gives higher success ratios than IMP without this loss (IMP w/o C) because pose-consistency loss allows pose-aware matches pump out first to join the pose estimation, reducing the number of iterations. EIMP adopts the sampling strategy which filters some unreliable matches, therefore the pose-consistency has little influence on the success ratios of EIMP.

Number of iterations. A our model is able to predict matches at each iteration. We test its performance on relative pose estimation by progressively increasing the number of iterations from 3 to 9. As the number of iterations increases, keypoints become more discriminative with more geometric information embedded, so both IMP and EIMP achieve more precise poses, as shown in Fig. 2b. Additionally, their matching precision (Prec.) also gets higher gradually, indicating that they find more inliers, as shown in Fig. 2c. Fig. 2b and Fig. 2c show that both IMP and EIMP report almost the best performance when using 7 or 8 iterations, which means IMP and EIMP could even run faster by reducing the maximum number iterations from 9 to 7 or with marginal performance loss. Note that both IMP and EIMP are trained to predict matches at each iteration, so we don't need to retrain or fine-tune IMP and EIMP to achieve this.

B.2. Qualitative results on Scannet

In Fig. 3, we visualize predicted matches and relative poses of SuperGlue* [9], SGMNet [3], our IMP and EIMP. We observe that for simple cases (Fig. 3 (1)), all matchers give similar numbers of inliers. However, both our IMP and EIMP obtain smaller rotation and translation errors than SuperGlue* and SGMNet because the embedded geometric information in our matching module. In the iteration process, rather than finding inliers from a cluster, both IMP and EIMP expand the areas with inliers, which allows our models to find more potential inliers and make the pose estimation more stable. When testing images become more difficult (Fig. 3 (2) and (3)), the behavior of our models in expanding inliers over the whole meaningful regions of images can be observed more clearly. Especially for Fig. 3 (3) where all keypoints are extracted from regions with repetitive textures, matching methods based on pure descriptors can hardly discriminate these keypoints, so SuperGlue* and SGMNet fail to give enough inliers. At this time, geometric constraints play an indispensable role at finding correct matches, which explains the success of IMP and EIMP.



(a) Number of iterations (b) Pose accuracy (c) Precision of matches

Figure 2. We report the number of iterations for IMP and EIMP with and without pose-consistency (w/o C) loss (a), relative pose accuracy @ 10° (b), and precision of matches (c) of IMP and EIMP using different number of iterations on YFCC100m dataset [12].

B.3. Qualitative results on YFCC100m

Fig. 4 shows the predicted matches and relative poses reported by SuperGlue* [9], SGMNet [3], our IMP and EIMP. For simple cases (Fig. 4 (1)), all methods give a large number of inliers while IMP and EIMP report smaller pose errors in the iteration process even using fewer inliers. Note that instead of keeping all keypoints, our EIMP with adaptive sampling effectively reduces the number of keypoints from 2,0000 to 865 and 255 in the iteration process, which significantly decreases the time complexity for self and cross attention computation 1. We also observe that for tough cases (Fig. 4 (2) and (3)), due to large viewpoint and illumination changes, SuperGlue* and SGMNet fail to report comparable number of inliers to our models, resulting in higher rotation and translation errors. In contrast, our models still progressively increase the number of inliers from different regions in the iteration process. These well distributed inliers lead to smaller pose errors. By comparing the results of EIMP in Fig. 4 (2) and Fig. 4 (3), we see that the number of preserved keypoints are based on the number of potential inliers in the image pair: more potential inliers result in more retrained keypoints. That is because our sampling strategy is fully adaptive.

B.4. Qualitative results on Aachen

In Fig. 5, we show the inliers between query and reference images in the large-scale localization task on Aachen v1.1 dataset [10, 13]. Different with image pairs in Scannet [4] and YFCC100m [12] datasets, query and reference images in Aachen dataset are captured under totally different conditions usually with extremely large viewpoint (Fig. 5 (1)-(4)) and illumination (Fig. 5 (5)-(8)) changes, making finding matches difficult. As the groundtruth poses of query images are not available, we use HLoc [8] framework to visualize the inliers between query and reference images after the PnP [6] + RANSAC [5] for all methods.

Fig. 5 ((1)-(4)) shows that when image pairs have large viewpoint changes, both SuperGlue* and SGMNet fail to

find enough correct matches. That is because geometric constraints are more useful for finding matches in two images with large viewpoint changes and both SuperGlue* and SGMNet ignore this information. However, we embed the geometric information into the matching module, so our IMP and EIMP work much better, guaranteeing the localization success.

When query images have large illumination changes with reference images, corresponding keypoints from two images are less discriminative, so SuperGlue* and SGMNet only give slightly more inliers than for images with large viewpoint, as shown in Fig. 5((5)-(8)). As our model additionally leverages geometric constraints to find matches, both IMP and EIMP successfully obtain a large number of inliers. Note that compared to SuperGlue* and SGMNet, both IMP and EIMP find inliers from the almost the whole overlap regions of the two images as opposed to some clusters (Fig. 5 (2)-(5) and (8)).

References

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

[2] Boyu Chen, Peixia Li, Baopu Li, Chuming Li, Lei Bai, Chen Lin, Ming Sun, Junjie Yan, and Wanli Ouyang. PSViT: Better vision transformer via token pooling and attention sharing. *arXiv preprint arXiv:2108.03428*, 2021. 1

[3] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to match features with seeded graph matching network. In *CVPR*, 2021. 1, 2, 4, 5, 6

[4] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM ToG*, 2017. 1, 2, 4

[5] 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*, 1981. 2, 6

[6] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An Accurate O(n) Solution to the PnP Problem. *IJCV*, 81:155–166, 2009. 2, 6

[7] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 1

[8] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *CVPR*, 2019. 1, 2, 6

[9] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 1, 2, 4, 5, 6

[10] 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, 2, 6

[11] Yan Shi, Jun-Xiong Cai, Yoli Shavit, Tai-Jiang Mu, Wensen Feng, and Kai Zhang. ClusterGNN: Cluster-based Coarse-to-Fine Graph Neural Network for Efficient Feature Matching. In *CVPR*, 2022. 1

[12] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 2016. 1, 2, 5

[13] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference pose generation for long-term visual localization via learned features and view synthesis. *IJCV*, 2021. 1, 2, 6

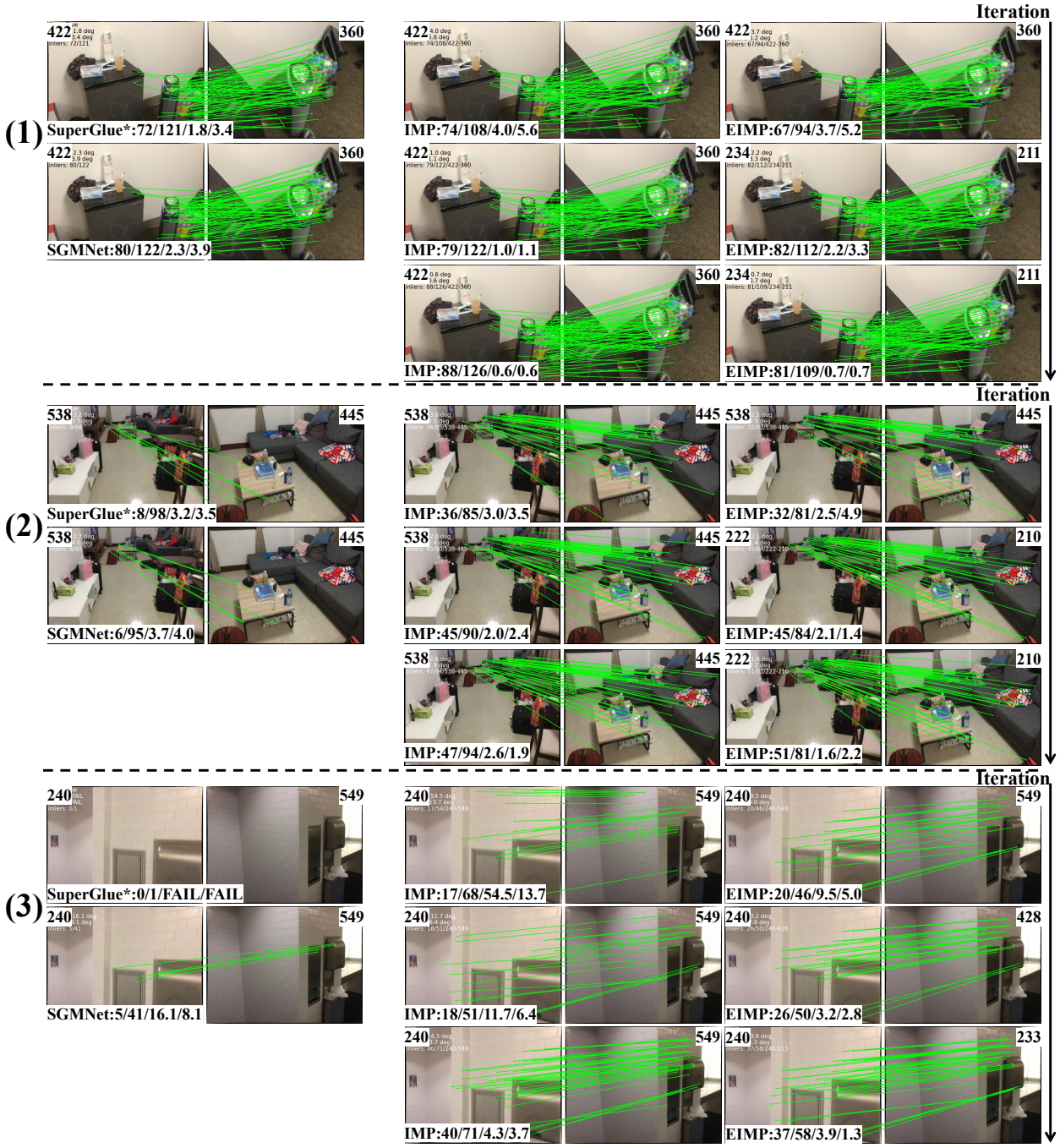


Figure 3. **Qualitative results on Scannet dataset [4].** At the left-bottom of each image pair, we report the number of inliers/matches and rotation/translation errors of SuperGlue* [9] (official SuperGlue), SGMNet [3], our IMP and EIMP. Besides, the number of keypoints in each image are shown at the top of each pair. For simple case (1), all methods give similar numbers of inliers, but IMP and EIMP obtain smaller rotation and translation errors than SuperGlue* and SGMNet because of the embedded geometric information. In the iteration process, rather than finding inliers from a cluster, both IMP and EIMP expand the areas with inliers, which allows our models to find more potential inliers and make the pose estimation more stable. When testing images become more difficult ((2), (3)), the behavior of our models in expanding inliers over the whole meaningful regions of images can be observed more clearly. Especially for (3) where all keypoints are extracted from regions with repetitive textures, matching methods based on descriptors can hardly discriminate these keypoints, so SuperGlue* and SGMNet fail to give enough inliers. At this time, geometric constraints play an indispensable role at finding correct matches, which explains the success of IMP and EIMP.

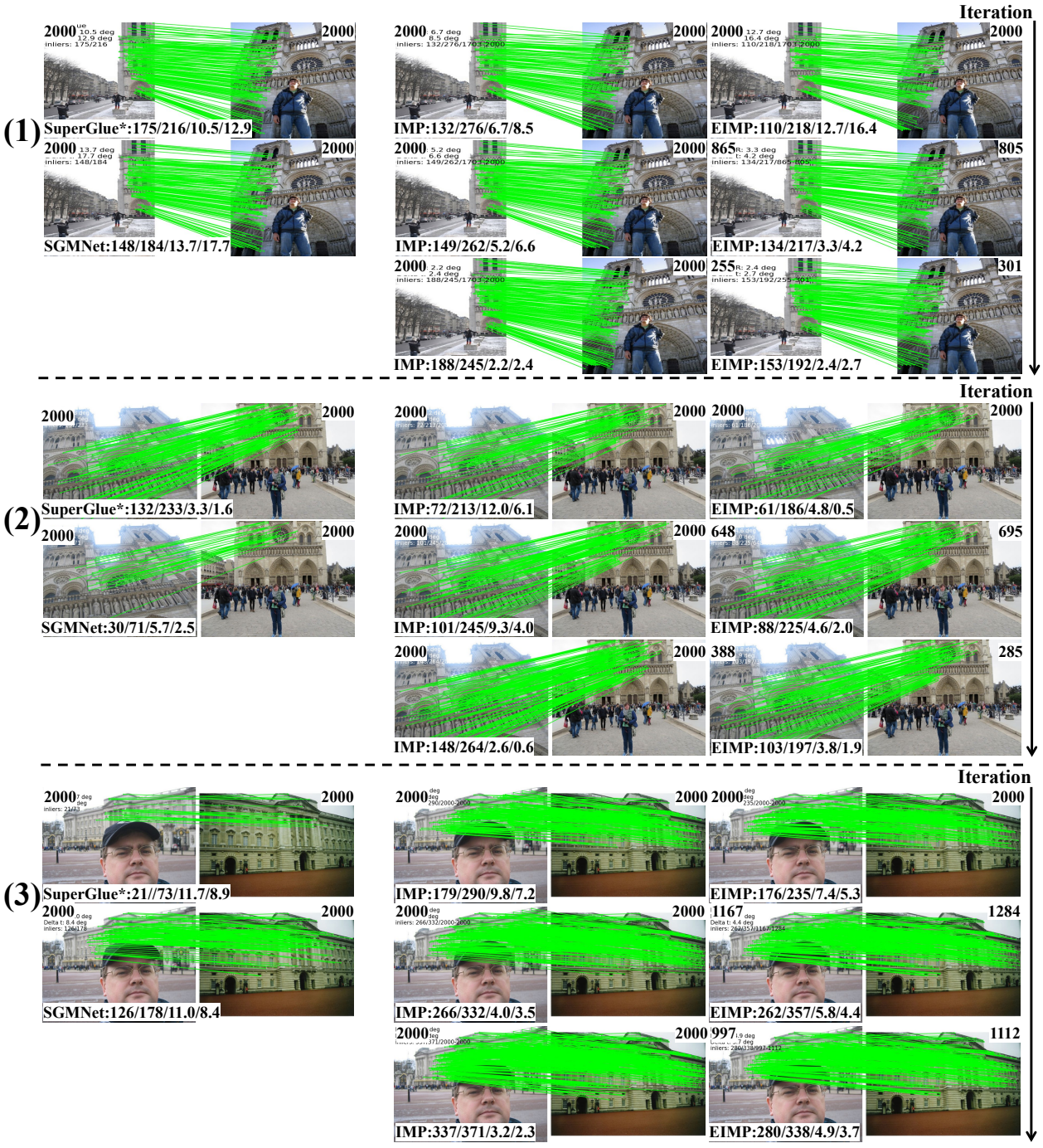


Figure 4. **Qualitative results on YFCC100m dataset [12]** At the left-bottom of each image pair, we report the number of inliers/matches and rotation/translation errors of SuperGlue* [9] (official SuperGlue), SGMNet [3], our IMP and EIMP. Besides, the number of keypoints in each image are shown at the top of each pair. For simple case (1), all methods give a large number of inliers, but IMP and EIMP report smaller pose errors even when using fewer inliers. Note that instead of keeping all keypoints, EIMP effectively reduces the number of keypoints from 2,0000 to 865 and 255 in the iteration process, significantly decreasing the time complexity for self and cross attention computation 1. For tough cases ((2), (3)), due to large viewpoint and illumination changes, SuperGlue* and SGMNet fail to report comparable number of inliers to our models, resulting in higher rotation and translation errors. In contrast, our models still progressively increase the number of inliers from different regions in the iteration process. These well distributed inliers lead to smaller pose errors. By comparing the results of EIMP in (2) and (3), we see that the number of preserved keypoints are based on the number of potential inliers in the image pair: more potential inliers result in more retrained keypoints. That is because our sampling strategy is fully adaptive.

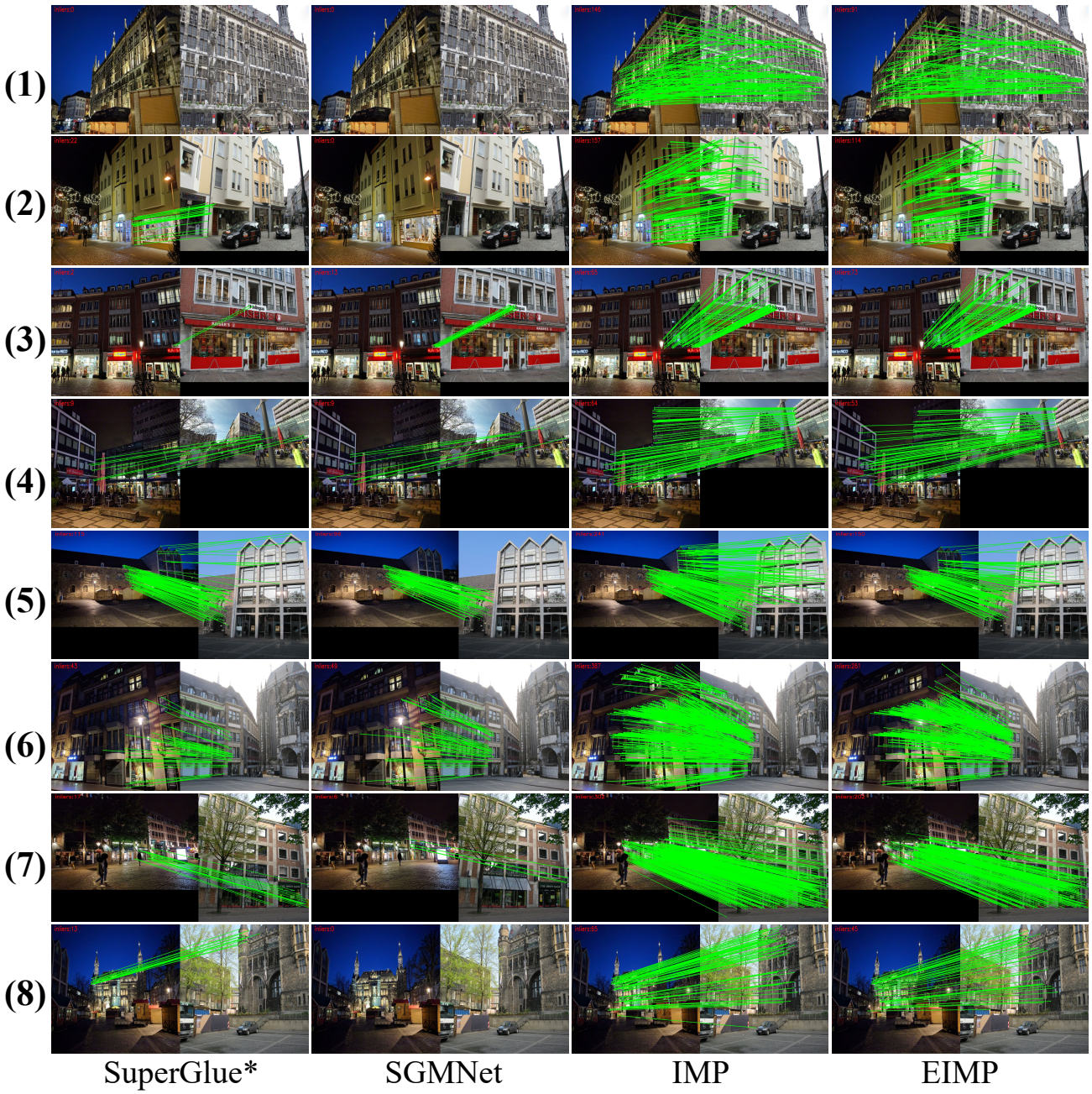


Figure 5. **Qualitative results on Aachen v1.1 dataset [10, 13]**. We visualize inliers of query (left) and reference (right) images under larger viewpoint ((1)-(4)) and illumination ((5)-(8)) changes of SuperGlue* [9] (official SuperGlue), SGMNet [3], our IMP and EIMP. As the groundtruth poses of each query images are not available, we utilize HLoc [8] framework to visualize inliers given by PnP [6] + RANSAC [5]. Testing pairs (1)-(4) show that when image pairs have large viewpoint changes, both SuperGlue* and SGMNet fail to find enough correct matches. That is because geometric constraints are more useful for finding matches in two images with large viewpoint changes and both SuperGlue* and SGMNet ignore it. However, we embed the geometric information into the matching module, so our IMP and EIMP work much better, guaranteeing the localization success. For cases (5)-(8), when query images have large illumination changes with reference images, corresponding keypoints from two images are less discriminative, so SuperGlue* and SGMNet only give slightly more inliers than for images with large viewpoint. As our model additionally leverages geometric constraints to find matches, both IMP and EIMP successfully obtain a large number of inliers. Note that compared to SuperGlue* and SGMNet, both IMP and EIMP find inliers from the almost the whole overlap regions of the two images as opposed to some clusters ((2)-(5), (8)).