A2-GNN: Angle-Annular GNN for Visual Descriptor-free Camera Relocalization

Supplementary Material

1. Additional Details and Results

Data Preparation. Following the experimental settings in GoMatch [7] and DGC-GNN [6], we use the MegaDepth dataset [2] for training. MegaDepth is a large-scale outdoor dataset comprising 196 scenes from various landscapes around the world. We utilize 99 scenes for training, 16 scenes for validation, and 53 scenes for testing. The ground truth sparse 3D point clouds are reconstructed using COLMAP [4]. During data preprocessing, a maximum of 500 query images are selected per scene. For each query, we gather its k co-visible views, ensuring at least 35% visual overlap. Queries lacking sufficient co-visible views are excluded from the training set. Visual overlap is computed as the ratio of co-observed 3D points to the total number of 3D points in the query image. The training set comprises 25,624 queries from 99 scenes, the validation set includes 3,146 queries from 16 scenes, and the test set consists of 12,399 samples from 49 scenes. For the training dataset, we control the number of keypoints per image to range from 100 to 1,024. During inference, this range is adjusted to 10 to 1,024 keypoints per image.

Representation Ablation Study. We present results with different 3D representations, as shown in Table 1. The bearing vector as the representation in 3D side plays a crucial role in enhancing the results. The insight behind this improvement is that it integrates the pose of database images into feature learning, bringing one step further towards middle representation from 3D to 2D.

Generalizability. Our model is trained on the MegaDepth dataset [2] using the SIFT [3] detector. To demonstrate the generalizability of our model, we conducted evaluations on the 7Scenes dataset using two keypoint detectors: SIFT and SuperPoint [1]. The results are presented in Table 2. The similar results in translation and rotation errors between the two detectors further demonstrate the robustness and generalizability of our model.

Hyperparameters Selection. Ablation studies on various hyperparameters in the self-attention layer are presented in Table 3. The outlier rejection threshold of $t = 0.7$ yields the best results, achieving higher AUC and lower rotation and translation errors. We select $t = 0.5$ in the main paper to make a fair comparison with other methods. The choice of the nearest neighbors parameter k has minimal impact on performance. However, when fewer nearest neighbors are processed, it becomes more challenging to accurately capture the local geometric structures.

Timing and Model Size. The inference time per query image for A2-GNN is ∼34 ms, comprising four main components: the feature encoding (∼2 ms), the attention layers (∼14 ms), optimal transport (∼9 ms), and the outlier rejection process (∼6 ms). Our model contains 2.7 million parameters, with a total size of ∼10.6 MB. All experiments were conducted on a 32GB NVIDIA Tesla V100 GPU, using a maximum of 1,024 keypoints.

3D representation	Reproj. AUC (%)	Rotation (°)	Translation (m)
	@1 / 5 / 10px (↑)	Quantile @25 / 50 / 75% (↓)	
Coordinate	7.69 / 27.96 / 32.82	0.28 / 12.6 / 59.64	0.02 / 1.32 / 5.34
Bearing vector	**12.72 / 41.84 / 48.02**	**0.12 / 0.79 / 26.37**	**0.01 / 0.08 / 2.80**

Table 1. Ablation results with different 3D representations on MegaDepth on top-1 image retrieval.

7Scenes [5]	SIFT [3]	SuperPoint [1]
Chess	3 / 1.37	3 / 1.41
Fire	5 / 1.78	6 / 1.99
Heads	4 / 2.70	2 / 3.12
Office	6 / 1.56	6 / 1.48
Pumpkin	7 / 1.86	9 / 2.28
Redkitchen	7 / 2.00	8 / 2.08
Stairs	72 /17.05	66 / 16.02

Table 2. Comparison on sift and superpoint as detector on 7Scenes dataset. Median translation and rotation errors (cm,°) are reported.

Acknowledgement. This work was supported by the Academy of Finland (grants No). We acknowledge the computational resources provided by the CSC-IT Center for Science, Finland.

Methods	Neighbors	Groups	OR Threshold	Reproj. AUC (%) @1 / 5 / 10px (↑)	Rotation (°) Quantile@25 / 50 / 75% (↓)	Translation
A2-GNN	9	3	0.5	17.29 / 54.41 / 62.24	**0.06** / 0.19 / 4.6	**0.01** / 0.02 / 0.48
HyperParam.	9	3	0.7	**18.59 / 58.84 / 66.48**	**0.06 / 0.16 / 2.16**	**0.01 / 0.01 / 0.22**
	9	3	0.3	14.82 / 48.57 / 56.7	0.07 / 0.34 / 8.42	**0.01** / 0.03 / 0.9
	12	3	0.5	17.21 / 54.36 / 62.18	**0.06** / 0.2 / 4.45	**0.01** / 0.02 / 0.46
	9	no groups	0.5	15.35 / 49.81 / 57.64	0.08 / 0.28 / 5.48	**0.01** / 0.03 / 0.58
	6	3	0.5	16.41/ 51.84 / 59.52	0.06 / 0.22 / 7.49	**0.01** / 0.02 / 0.81

Table 3. Ablations on Hyperparameters. The results are evaluated on MegaDepth with top-10 retrieval images. The best results are bold.

References

[1] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 1

[2] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 1

[3] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 1

[4] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[5] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013. 1

[6] Shuzhe Wang, Juho Kannala, and Daniel Barath. Dgc-gnn: Leveraging geometry and color cues for visual descriptor-free 2d-3d matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20881–20891, 2024. 1

[7] Qunjie Zhou, Sérgio Agostinho, Aljoša Ošep, and Laura Leal-Taixé. Is geometry enough for matching in visual localization? In *European Conference on Computer Vision*, pages 407–425. Springer, 2022. 1