Visual Re-Ranking with Non-Visual Side Information

GUSTAV HANNING, GABRIELLE FLOOD, VIKTOR LARSSON

Lund University

Conclusion

Including **side information**, for example recorded radio signal strengths, database image poses or compass heading angles, **can significantly increase the accuracy** of image retrieval and downstream tasks such as visual localization.

Background

Image retrieval:

- Input: query image
- Output: top matching database images
- May include re-ranking step to re-order top DB images
- Only uses visual similarity (global image descriptors)

Our approach:

- Utilize side information to improve re-ranking
- Flexible framework: we combine visual similarity with other types, like similarity of recorded WiFi and BlueTooth signals
- Learning based: our network learns how to best weight the different modalities

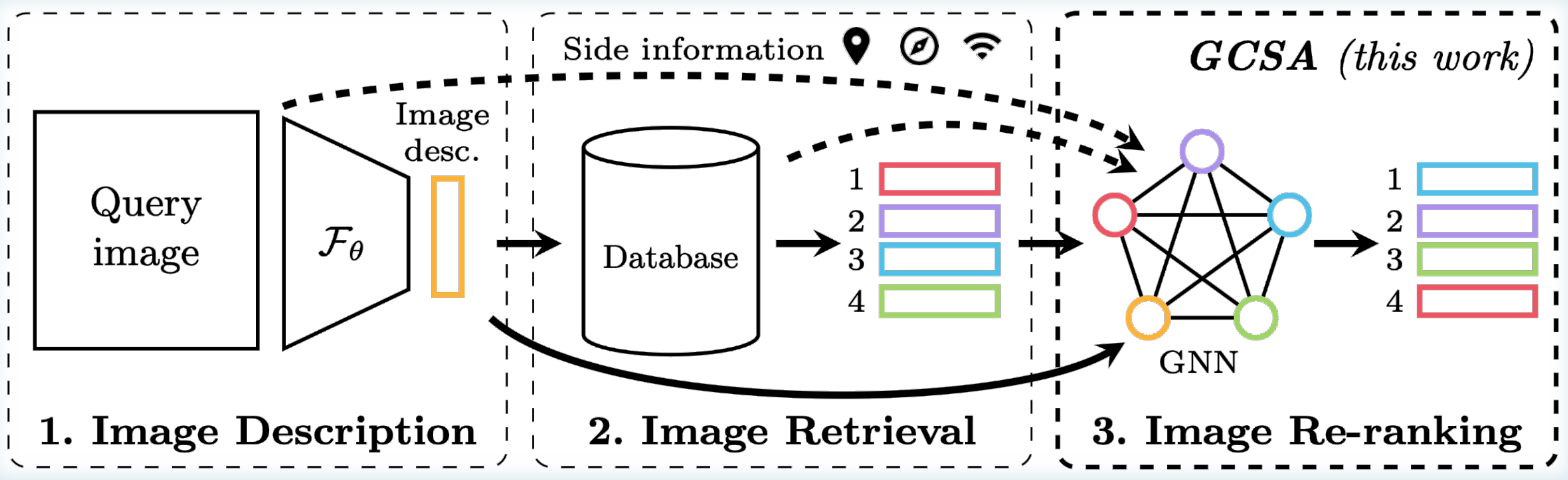

Figure 2. The image retrieval and re-ranking pipeline using GCSA. **(1)** For each image, a global image descriptor is computed. **(2)** An initial ordering is established by comparing descriptor similarity between the query and database descriptors. **(3)** Our proposed method (GCSA) takes the top-scoring descriptors, together with other side information, and re-ranks them to improve the accuracy of the retrieval.

Results

We train and evaluate GCSA on two large-scale datasets covering both outdoor and indoor scenarios.

I. GCSA achieves the highest precision and recall among re-ranking methods benchmarked on Mapillary Street-Level Sequences.

II. We use GCSA to localize the query images in the LaMAR dataset, showing improved accuracy compared to CSA.

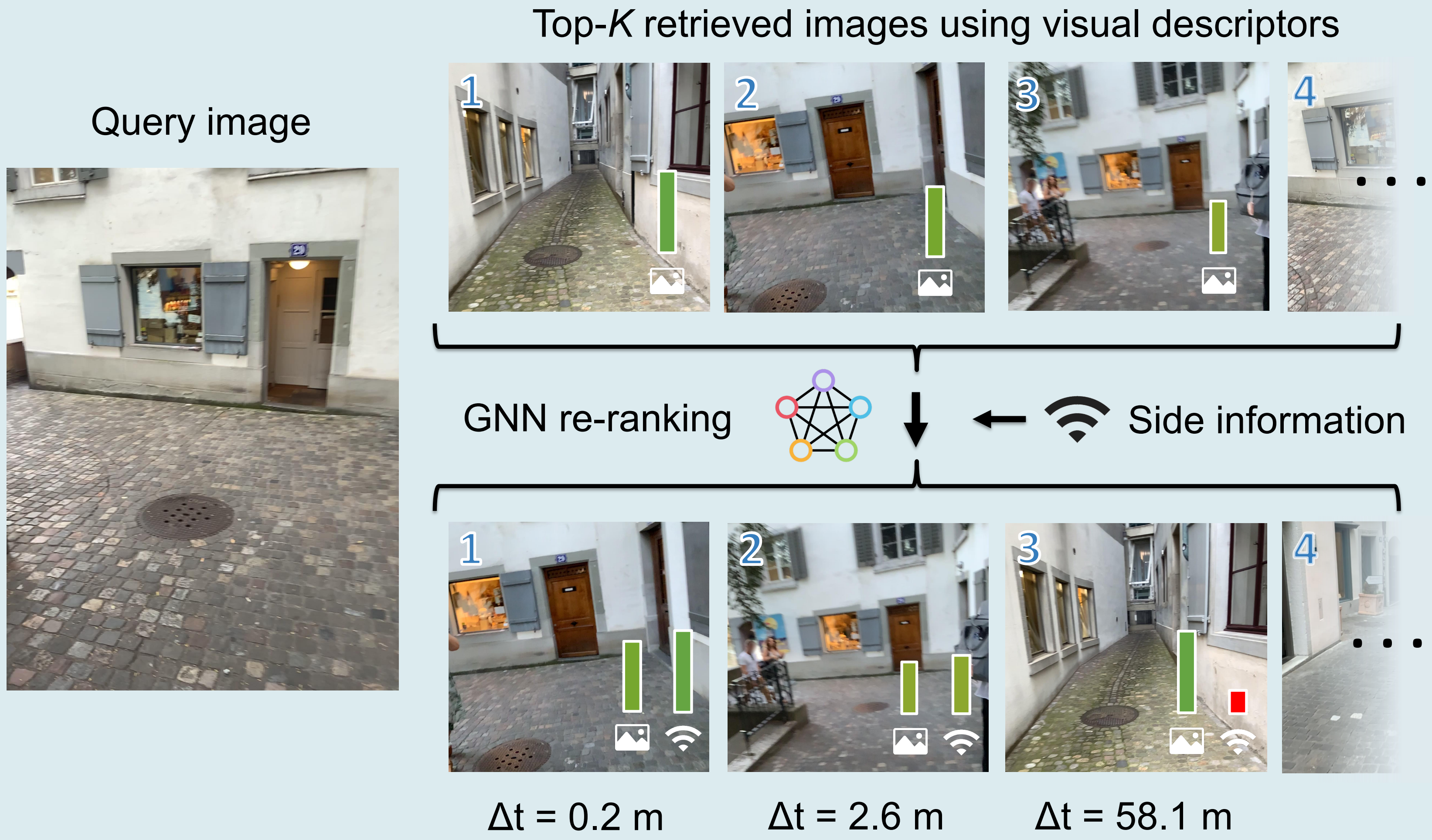

Figure 1. Our image retrieval re-ranking method GCSA utilizes both the visual (🖼) similarity between the query and database images as well as non-visual side information such as the similarity of recorded radio signal strengths (📶). A database image that is visually similar to the query (top left) can be downranked if the radio signals do not match. Δt denotes the ground truth distance to the query camera.

Method

Contextual Similarity Aggregation (CSA) [1]:

- Encode visual similarity with top L images in affinity vector
  $a_i^{vis} = [s(d_i,\, d_0),\ s(d_i,\, d_1),\ \ldots,\ s(d_i,\, d_L)],\ 0 \le i \le K$
- Refine affinity vectors in a GNN with self-attention
- Re-rank based on distance between refined affinity vectors

Generalized CSA (ours):

- Extend the affinity-based representation to other modalities
  $a_i^x = [s_x(I_i,\, I_0),\ s_x(I_i,\, I_1),\ \ldots,\ s_x(I_i,\, I_L)],\ 0 \le i \le K$
- Simplify network architecture and improve training process

[1] Ouyang, J., Wu, H., Wang, M., Zhou, W., Li, H.: Contextual Similarity Aggregation with Self-attention for Visual Re-ranking. NeurIPS (2021)

		mAP				Recall		
	Method	@1	@5	@10	@20	@5	@10	@20
NetVLAD	No re-ranking	34.5	22.4	19.7	18.6	45.4	50.7	55.8
	AQE	34.5	27.5	24.4	23.1	39.0	41.4	44.5
	αQE	34.5	25.3	22.6	21.6	41.8	45.0	48.3
	SuperGlobal	33.5	26.0	23.8	22.6	40.7	42.3	45.4
	CSA	34.6	27.8	26.0	24.9	51.0	57.6	63.4
	GCSA (ours)	**53.7**	**42.9**	**39.7**	**38.1**	**69.4**	**75.1**	**79.1**
SALAD	No re-ranking	75.5	63.1	60.1	59.0	89.2	91.5	93.3
	AQE	75.5	66.2	64.0	63.1	87.5	89.1	90.4
	αQE	75.3	65.3	63.5	62.5	87.9	89.8	91.3
	SuperGlobal	74.4	65.9	64.1	63.4	87.9	89.6	91.0
	CSA	76.2	67.6	65.6	64.9	88.8	91.9	93.0
	GCSA (ours)	**77.1**	**71.2**	**70.2**	**69.9**	**91.3**	**93.4**	**94.7**

Table 1. Re-ranking results on the Mapillary Street-Level Sequences test set using NetVLAD and DINOv2 SALAD descriptors. Our method includes visual, heading and positional affinity.

	HoloLens – Top 1 / 10		Phone – Top 1 / 10	
Method	(1°, 10 cm)	(5°, 1 m)	(1°, 10 cm)	(5°, 1 m)
No re-ranking	23.9 / 34.8	35.3 / 48.1	25.9 / 36.5	37.6 / 49.1
CSA	23.0 / 34.2	35.5 / 46.9	25.8 / 38.0	38.6 / 50.6
GCSA (ours) 🖼	25.8 / 36.8	39.3 / 50.6	27.8 / 39.3	41.6 / 52.5
GCSA (ours) 🖼 📍 📶	**26.1** / **39.2**	**41.9** / **54.5**	**30.6** / **43.7**	**47.0** / **59.7**

Table 2. Localization results for the LaMAR test set with NetVLAD descriptors. We compare our method, with and without side information, to CSA and the baseline of no re-ranking and report the recall at one fine and one coarse threshold.

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM