Supplementary Materials
Revisiting Self-Similarity: Structural Embedding for Image Retrieval

Seongwon Lee Suhyeon Lee Hongje Seong Euntai Kim*
School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
{won4113, hyeon93, hjseong, etkim}@yonsei.ac.kr

S1. Additional Ablation Studies

In this section, we introduce the additional ablation studies that could not be included due to the space limitations of the main paper.

S1.1. Comparison with Re-ranking Solutions

Since this paper aims to extract improved global embeddings, we did not consider re-ranking solutions in the main paper. In this subsection, we compare our proposed solution with existing re-ranking solutions and further show that our proposed solution is more effective when combined with re-ranking solutions. The comparison results are shown in Tab. S1. Our proposed method outperforms most of the previous two-stage (global retrieval and re-ranking) solutions (*e.g.* GeM+DSM [13], DELG+GV [1], DELG+RRT [14], DELG+SuperGlue [14]) and shows the advantage of handling structural information on the global stage. Recently, a powerful re-ranking solution called CVNet-Rerank [7] has emerged. We applied this CVNet-Rerank to our proposed Network. Our proposed network, SENet, shows quite high performance with global embedding alone, and when combined with CVNet-Rerank, it surpasses the original CVNet-Global + CVNet-Rerank method, further showing the powerful effect of robust global embedding.

S1.2. Re-ranking with Query Expansion

In Sec. S1.1, only methods for re-ranking through precise matching between image pairs (*e.g.* GV, RRT, and CVNet-Rerank) were presented, and methods for traversing the entire database (*e.g.* query expansion [3–5, 16] and diffusion [2, 6]) were not reported. In this subsection, we additionally apply alpha query expansion (αQE) [4], which is a representative method among query expansion methodologies, to our method, and show that our method can be harmoniously connected with various re-ranking methods. we tune the hyper-parameters of αQE, the number of the query expansion candidates n and power parameter α, on $\mathcal{R}$Oxf / $\mathcal{R}$Par benchmarks and fixed on their 1M-add experiments following the previous studies [7, 14]. Finally, we choose $n = 5, \alpha = 2$ for $\mathcal{R}$Oxf and $n = 20, \alpha = 1$ for $\mathcal{R}$Par experiments. Tab. S2 shows the results when the αQE methods is applied to our proposed SENet. Due to the characteristic of query expansion, which shows better performance as the global retrieval result is more accurate, our proposed global embedding network shows a huge performance improvement when combined with query expansion.

S1.3. Model Design Consideration

Channel-wise similarity (Tab. S3). Self-similarity can be calculated per channel or directly using all channels. We conduct additional ablation studies on two methods: channel-wise self-similarity and direct self-similarity. The results are shown in Tab. S3. The method using channel-by-channel self-similarity shows superior performance. We believe that channel-wise self-similarity is a way to fully exploit the valuable semantic information of each channel, and our experimental results support this belief.

Similarity type (Tab. S4). Self-similarity can be measured through several similarity metrics. We conduct additional ablation studies on two similarities: cosine similarity and dot product. The results are shown in Tab. S4. The model using the dot product shows quite good performance in the base $\mathcal{R}$Oxf and $\mathcal{R}$Par experiments, but shows relatively weak performance in the 1M-add experiments than the model using cosine similarity. Since cosine similarity helps to measure the absolute similarity without being affected by the scale of the features, it showed higher performance than the model using the dot product.

Fusion method (Tab. S5). Feature fusion can also proceed with several fusion methods. We conduct additional ablation studies on two fusion methods, sum and concatenate. The results are shown in Tab. S5. While both experiment results with each fusion method show better performance than the baseline for all measures with the help of self-similarity, the model using sum fusion shows slightly better performance while intuitively learning the consensus of the visual features and structural features.

*Corresponding author.

	Medium				Hard			
	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M
(a) Existing Global Retrieval Solutions + Re-ranking								
DELF-D2R-R-ASMK* (GLDv1) [15]	73.3	61.0	80.7	60.2	47.6	33.6	61.3	29.9
+ Spatial Verification (SP) [15]	76.0	64.0	80.2	59.7	52.4	38.1	58.6	29.4
R101-GeM (SfM-120k) [11, 13]	65.3	46.1	77.3	52.6	39.6	22.2	56.6	24.8
+ Deep Spatial Matching (DSM) [13]	65.3	47.6	77.4	52.8	39.2	23.2	56.2	25.0
R50-DELG (GLDv2-clean) [1]	73.6	60.6	85.7	68.6	51.0	32.7	71.5	44.4
+ Geometric Verification (GV) [1]	78.3	67.2	85.7	69.6	57.9	43.6	71.0	45.7
+ Reranking Transformer (RRT) [14]	78.1	67.0	86.7	69.8	60.2	44.1	75.1	49.4
R101-DELG (GLDv2-clean) [1]	76.3	63.7	86.6	70.6	55.6	37.5	72.4	46.9
+ Geometric Verification (GV) [1]	81.2	69.1	87.2	71.5	64.0	47.5	72.8	48.7
+ Reranking Transformer (RRT) [14]	79.9	-	87.6	-	64.1	-	76.1	-
+ SuperGlue [12, 14]	79.7	-	87.1	-	62.1	-	71.5	-
R50-CVNet-Global (GLDv2-clean) [7]	**81.0**	72.6	88.8	79.0	62.1	50.2	76.5	60.2
+ R50-CVNet-Rerank [7]	**86.1**	77.6	89.4	79.9	72.8	61.1	78.6	63.9
R101-CVNet-Global (GLDv2-clean) [7]	80.2	**74.0**	**90.3**	**80.6**	**63.1**	**53.7**	**79.1**	**62.2**
+ R101-CVNet-Rerank [7]	85.6	**79.6**	**90.6**	**81.5**	**72.9**	**64.5**	**80.4**	**66.2**
(b) Ours + Re-ranking								
R50-SENet-$\mathcal{L}_{cls}$ & $\mathcal{L}_{con}$ (GLDv2-clean)	81.9	74.2	90.0	79.1	63.0	52.0	78.1	59.9
+ R50-CVNet-Rerank‡ [7]	85.8	78.7	90.8	80.1	72.4	62.7	81.0	63.7
R101-SENet-$\mathcal{L}_{cls}$ & $\mathcal{L}_{con}$ (GLDv2-clean)	**82.8**	**76.1**	**91.7**	**83.6**	**66.0**	**55.7**	**82.8**	**67.8**
+ R101-CVNet-Rerank‡ [7]	**86.5**	**80.0**	**92.0**	**84.3**	**74.4**	**65.4**	**83.5**	**70.7**

Table S1. **Comparison with state-of-the-art re-ranking models.** All re-rankings were applied to the top 100 candidates among the global retrieval results for each query. The best scores for each group are **boldfaced**. ‡ denotes extract re-ranking scores with the official models.

model	Loss		Medium				Hard			
	$\mathcal{L}_{cls}$	$\mathcal{L}_{con}$	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M
R50-SENet	✓		81.4	72.9	90.5	79.0	62.3	48.7	80.3	59.9
+ αQE			84.8	78.6	93.1	86.6	66.9	57.8	85.3	73.0
R50-SENet	✓		81.9	74.2	90.0	79.1	63.0	52.0	78.1	59.9
+ αQE			84.0	79.6	92.6	86.4	67.1	60.8	83.8	72.7
R101-SENet	✓	✓	80.0	72.5	**91.6**	82.1	61.7	49.2	**82.2**	64.6
+ αQE			83.2	78.4	**93.7**	88.1	64.4	56.8	**86.2**	75.4
R101-SENet	✓	✓	82.8	76.1	91.7	83.6	66.0	55.7	82.8	67.8
+ αQE			**85.0**	**81.2**	93.2	**88.2**	**69.3**	**63.0**	85.7	**76.5**

Table S2. **Effect of the Alpha Query Expansion (αQE).**

model	Medium				Hard			
(R50, $\mathcal{L}_{cls}$)	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M
baseline	78.6	70.7	89.5	77.4	58.8	44.8	77.9	57.7
directly	79.8	71.4	90.1	77.8	60.4	46.3	78.5	58.1
channel-wise	**81.4**	**72.9**	**90.5**	**79.0**	**62.3**	**48.7**	**80.3**	**59.9**

Table S3. **Ablation experiments on channel-wise self-similarity.**

model	Medium				Hard			
(R50, $\mathcal{L}_{cls}$)	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M
baseline	78.6	70.7	89.5	77.4	58.8	44.8	77.9	57.7
Dot Product	80.3	71.7	90.1	77.3	61.5	47.0	78.9	56.7
Cosine Similarity	**81.4**	**72.9**	**90.5**	**79.0**	**62.3**	**48.7**	**80.3**	**59.9**

Table S4. **Ablation experiments on self-similarity type.**

model	Medium				Hard			
(R50, $\mathcal{L}_{cls}$)	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M
baseline	78.6	70.7	89.5	77.4	58.8	44.8	77.9	57.7
Concatenate	80.1	71.7	89.9	78.3	61.4	47.3	78.6	58.5
Sum	**81.4**	**72.9**	**90.5**	**79.0**	**62.3**	**48.7**	**80.3**	**59.9**

Table S5. **Ablation experiments on feature fusion method.**

S1.4. Additional Feature Visualization

We additionally visualize the intermediate features of our proposed network in Fig. S1 to see the effect of the proposed modules. In this figure, original features $\mathbf{F}$ and self-similarity descriptor $\mathbf{D}$ are fused to structural feature $\mathbf{F}^{\mathrm{s}}$ while raising the similarities where both visual and structural cues form a consensus and diminishing the similarities that do not.

S1.5. Additional Qualitative Results

Additional qualitative results on $\mathcal{R}$Oxford5k [8, 10] and $\mathcal{R}$Paris6k [9, 10] benchmark are shown in Fig. S2 and Fig. S3, respectively. All results are reported from experiments with the addition of a 1M distractor on "hard" difficulty ($\mathcal{R}$Oxf-Hard+1M and $\mathcal{R}$Par-Hard+1M). These results show that the proposed structural embedding finds the correct answer more accurately, even when the baseline solution often retrieves incorrect answers due to similar visual properties.

Figure S1. Additional visualization of the intermediate feature similarity between query-positive and query-hard negative images. Our network enhances the similarity where the visual and structural cues form a consensus and diminishes other parts. $S_c(\cdot, \cdot)$ denotes cosine similarity between two inputs. All features are extracted using R50-SENet-$\mathcal{L}_{cls}$ model.

Figure S2. **Additional qualitative results with R50-DELG† and R50-SENet-$\mathcal{L}_{cls}$ models on $\mathcal{R}$Oxford5k-Hard+1M benchmark**. The upper line is the result of R50-DELG†, and the lower line is the result of R50-SENet-$\mathcal{L}_{cls}$. Correct and incorrect answers are marked with green / red borders around the image, respectively. yellow dotted line indicates the area of the positive image that overlaps the query. All query images are cropped following the evaluation protocol of [10]. Our purpose is to visualize the difference between the baseline and our proposed methods so we skip the correct results that both models correct.

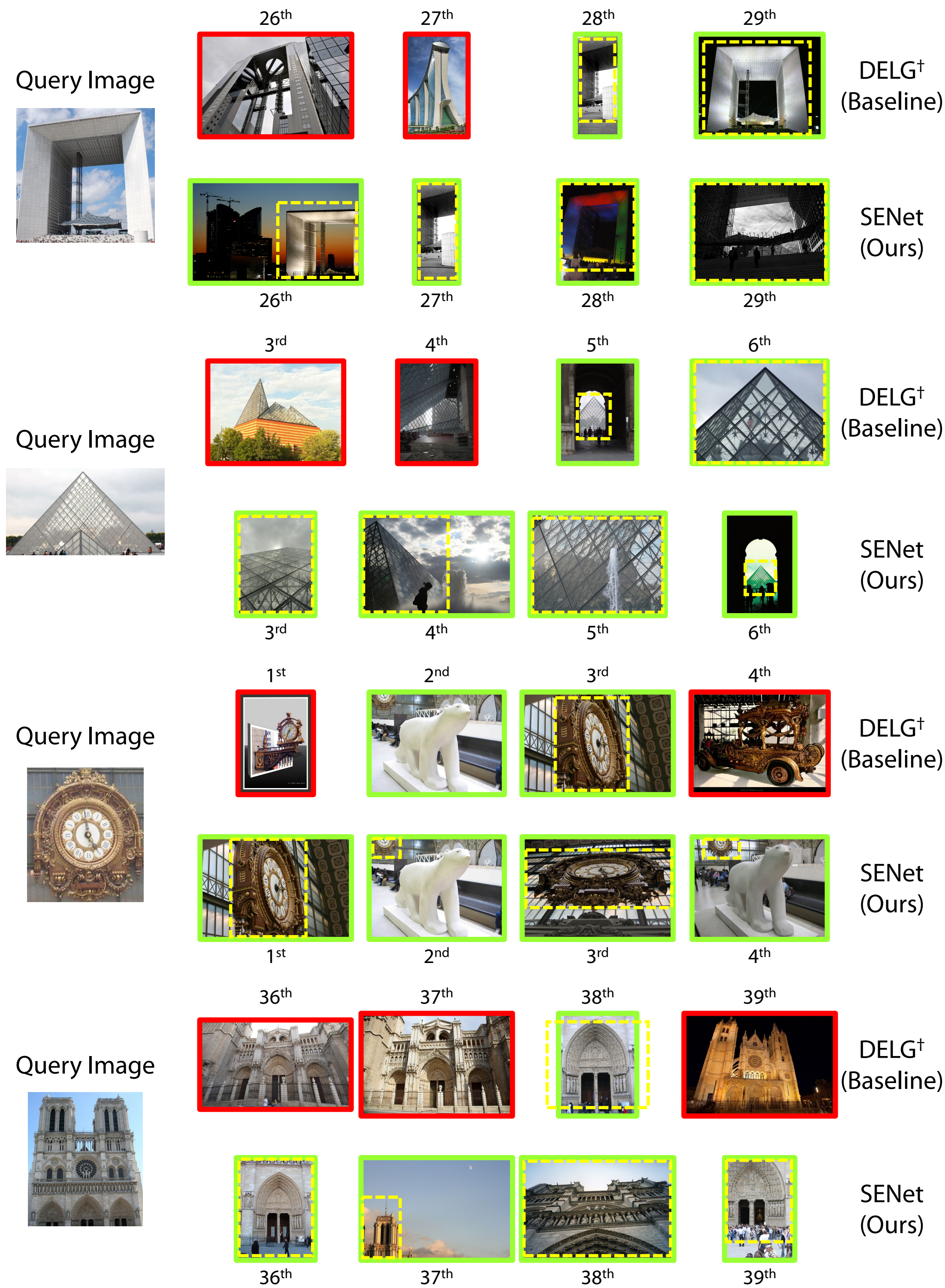

Figure S3. **Additional qualitative results with R50-DELG† and R50-SENet-$\mathcal{L}_{cls}$ models on $\mathcal{R}$Paris6k-Hard+1M benchmark**. The upper line is the result of R50-DELG†, and the lower line is the result of R50-SENet-$\mathcal{L}_{cls}$. Correct and incorrect answers are marked with green / red borders around the image, respectively. yellow dotted line indicates the area of the positive image that overlaps the query. All query images are cropped following the evaluation protocol of [10]. Our purpose is to visualize the difference between the baseline and our proposed methods so we skip the correct results that both models correct.

References

[1] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 726–743. Springer, 2020. 1, 2

[2] Cheng Chang, Guangwei Yu, Chundi Liu, and Maksims Volkovs. Explore-exploit graph traversal for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9423–9431, 2019. 1

[3] Ondřej Chum, Andrej Mikulik, Michal Perdoch, and Jiří Matas. Total recall ii: Query expansion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 889–896. IEEE, 2011. 1

[4] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007. 1

[5] Albert Gordo, Filip Radenovic, and Tamara Berg. Attention-based query expansion learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–188. Springer, 2020. 1

[6] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Teddy Furon, and Ondrej Chum. Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2077–2086, 2017. 1

[7] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5374–5384, June 2022. 1, 2

[8] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 2

[9] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008. 2

[10] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5706–5715, 2018. 2, 4, 5

[11] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(7):1655–1668, 2018. 2

[12] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4938–4947, 2020. 2

[13] Oriane Siméoni, Yannis Avrithis, and Ondrej Chum. Local features and visual words emerge in activations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11651–11660, 2019. 1, 2

[14] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[15] Marvin Teichmann, Andre Araujo, Menglong Zhu, and Jack Sim. Detect-to-retrieve: Efficient regional aggregation for image search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5109–5118, 2019. 2

[16] Giorgos Tolias and Hervé Jégou. Visual query expansion with or without geometry: refining local descriptors by feature aggregation. *Pattern recognition*, 47(10):3466–3476, 2014. 1