Instance-level Image Retrieval using Reranking Transformers
Supplementary Material

Fuwen Tan
University of Virginia
fuwen.tan@virginia.edu

Jiangbo Yuan
eBay Computer Vision
jiayuan@ebay.com

Vicente Ordonez
Rice University
vicenteor@rice.edu

This document is organized as follows. In Sec. A, we discuss why we consider geometry verification, query expansion, and aggregated selective match kernel as the baseline methods. In Sec. B, we provide an ablation study on using different numbers of local descriptors in geometry verification (GV) [9] and RRT. In Sec. C, we perform experiments using SuperPoint [5] as the feature extractor for RRT, and compare with SuperGlue [11] on Stanford Online Products [12]. In Sec. D, we perform experiments using ResNet101 [6] as the CNN backbone. In Sec. E, we visualize the keypoint correspondences learned by RRT. In Sec. F, we discuss the limitation of the proposed method. Finally, in Sec. G, we present more qualitative examples.

The names of the training images sampled from GLDv2, as discussed in Section 4.1 of the main paper, are in a separate document.

A. Appropriate baselines

We consider geometry verification [9] and αQE [3] as the main baselines as they share the same spirit with our method: they make better use of the test-time information. When comparing the query and target images, geometry verification attends to *different* sub-regions of the query image when the target image is *different*, and vice versa, which is very similar to the proposed Reranking Transformers (RRTs). αQE also leverages test-time knowledge, but relies on analyzing the local affinity graph created during testing. We believe incorporating test-time knowledge is the key motivation of image reranking. It also distinguishes our method from most of the previous approaches that focus on feature learning. Note that we use pretrained and fixed feature representation in most of our experiments.

Fig. 1 provides an intuitive example of the partial-matching cases. In this example, the target images are some crops of the query. We believe the global descriptor + cosine similarity paradigm is not ideal for this case, as no matter how large is the global descriptor, it contains irrelevant information that hinders the cosine similarity measurement.

Aggregated Selective Match Kernel (ASMK) [13] was previously used as a global retrieval approach instead of an image reranking approach. Specifically, it proposes to create a set of new filters (i.e. visual codebook) by clustering. It then remaps/aggregates the local descriptors of each image into a global vector. We perform experiments on ASMK as it also relies on local descriptors.

Figure 1. An example where the target images are some crops of the query. In this case the global descriptor + cosine similarity retrieval paradigm may not be ideal.

B. Ablation on the number of local descriptors

In the DELG model, for each image, a maximum of 1000 local descriptors are extracted for geometric verification. In our experiment, we observe that for most of the images, the number of local descriptors is close to 1000. For example, on the sampled GLDv2 training set, the query and gallery sets of Revisited Oxford ($\mathcal{R}$Oxf) [10], DELG extracts 955/759/987 local descriptors per image on average.

We perform an ablation experiment by setting the maximum number of local descriptors used for each image to different values. The DELG model [2] used in this experiment is pretrained on the "v2-clean" split of Google Landmarks v2 (GLDv2) [15]. For purposes of comparison, we include the results of geometry verification (GV) and the proposed method (RRT). We report the mAP scores on Revisited Oxford ($\mathcal{R}$Oxf) in Table 1.

Both GV and RRT benefit from using more local descriptors in general. Nevertheless, the performance of RRT saturates at 500 local descriptors. As the local descriptors are

# Local Desc.	Medium		Hard	
	GV	RRT	GV	RRT
200	72.1	76.7	48.3	58.9
400	75.2	77.6	53.8	58.6
500	75.7	78.1	53.4	60.2
600	77.4	77.9	55.9	59.6
800	77.9	76.9	56.7	57.4
1000	78.3	78.1	57.9	60.4

Table 1. Ablation on the number of local descriptors used per image. We compare the proposed Reranking Transformer (RRT) model to geometric verification (GV) on Revisited Oxford [10]. The mAP scores on the Medium and Hard setups are reported.

Method	R@1	R@10	R@100
Global-only	32.8	45.4	60.5
SuperGlue [11]	45.5	54.6	60.5
RRT (w pos, frozen)	47.3	56.5	60.5
RRT (w/o pos, frozen)	50.2	57.9	60.5
RRT (w/o pos, finetuned)	**51.9**	**59.0**	60.5

Table 2. Comparison to the pretrained SuperGlue model [11] on Stanford Online Products [12], using SuperPoint [5] as the CNN backbone. The SuperGlue model is pretrained on ScanNet [4]. The R@K (K =1, 10, 100) scores on the SOP [12] test set are reported. Note that as only the top-100 neighbors are reranked, the R@100 scores remain unchanged for all the models.

extracted from seven image scales, we conjecture that in each image there are descriptors extracted from the same location, thus providing duplicate information. To verify this, we compute the number of *distinct* local descriptors extracted from different grid locations. In particular, we assign each local descriptor $\mathbf{x}_{l,i}$ to a grid location (gu, gv) by $(gu, gv) = (\lfloor u/16 \rfloor, \lfloor v/16 \rfloor)$. Here (u, v) is the coordinate of $\mathbf{x}_{l,i}$ provided by the DELG model, 16 is the stride of the convolutional feature map where $\mathbf{x}_{l,i}$ is extracted from. We then group the descriptors sharing the same grid location as a *distinct* descriptor. We observe that, the number of *distinct* local descriptors is significantly smaller than the number of all local descriptors per image. For example, on the sampled GLDv2 training set, the query and gallery sets of Revisited Oxford ($\mathcal{R}$Oxf), the numbers of *distinct* local descriptors per image are 585/465/655 on average.

When using the same number of local descriptors, RRT outperforms GV in four of the six experiments on the Medium setup, and consistently outperforms GV on the Hard setup.

C. SuperPoint as the CNN backbone.

In the main paper, we compare Reranking Transformer (RRT) with SuperGlue [11] on Revisited Oxford/Paris, but the feature extractors used for the two models are different: ResNet50 for RRT, SuperPoint [5] for SuperGlue. In this experiment, we use SuperPoint [5] as the feature extractor for RRT, so that it has the same backbone architecture as SuperGlue. We compare the new model with SuperGlue on Stanford Online Products [12]. We also explore finetuning the SuperPoint backbone (we tried finetuning SuperPoint on Google Landmarks v2-clean [15] but found it requires much more computing resources than we can afford). The SuperGlue model in this experiment is pretrained on ScanNet [4]. ScanNet is a large-scale dataset that contains 2.5 million images of 1513 indoor scenes. Both SuperGlue and our method take a 320x320 grayscale image as input. We extract the global descriptor by averaging all the local responses, and sample the top-500 local descriptors for all the models. We also investigate the benefit of using the position embedding for this task. The training and evaluation settings remain the same as in the SOP experiment presented in the main paper. We do not finetune SuperGlue on SOP as SOP does not include pixel-level annotations.

As shown in Table 2, reranking by either SuperGlue or RRT can significantly improve the retrieval performance. RRT outperforms SuperGlue with a frozen SuperPoint backbone. Interestingly, RRT does not benefit from the position embedding in this task, as is also the case in the DELG experiment. On the other hand, we observe that the position embedding is helpful in the SOP experiment of the main paper, where the descriptors of all the grid positions are used. We conjecture that the keypoints sampling may result in imbalanced sampled positions that potentially hinder the training. Finally, finetuning the SuperPoint backbone leads to the best performance.

D. ResNet101 as the CNN backbone.

Following [2], we perform experiments using ResNet101 as the CNN backbone. We train the Reranking Transformer on two extra sets of image descriptors: the DELG R101 descriptors pretrained on Google Landmarks (GLD) v1 [8] and v2-clean [15]. The training and evaluation settings remain the same as the main experiment on ResNet50, except that we also clip the gradient with a maximal norm of 0.1, and find that it can stabilize the training and lead to better performance. Here we compare our model with geometry verification [9] and SuperGlue [11] (pretrained on MegaDepth [7]) on Revisited Oxford/Paris [10], as shown in Table 3.

When evaluated on the "v1" descriptors, our method performs favorably to both geometry verification and SuperGlue on all the settings. When evaluated on the "v2-clean"

Method	Desc. version	# local desc.	Desc. dim.	Medium $\mathcal{R}$Oxf	Medium $\mathcal{R}$Par	Hard $\mathcal{R}$Oxf	Hard $\mathcal{R}$Par
DELG global	R101-v1	0	-	73.2	82.4	51.2	64.7
GV	R101-v1	1000	128	78.5	82.9	59.3	65.5
SuperGlue	SuperPoint	500	256	74.6	82.5	51.7	62.5
SuperGlue	SuperPoint	1024	256	76.9	82.9	57.2	64.7
RRT (ours)	R101-v1	500	128	**78.8**	**83.2**	**62.5**	**68.4**
DELG global	R101-v2-clean	0	-	76.3	86.6	55.6	72.4
GV	R101-v2-clean	1000	128	**81.2**	87.2	64.0	72.8
SuperGlue	SuperPoint	500	256	77.1	86.8	55.5	69.3
SuperGlue	SuperPoint	1024	256	79.7	87.1	62.1	71.5
RRT (ours)	R101-v2-clean	500	128	79.9	**87.6**	**64.1**	**76.1**

Table 3. Comparison to geometric verification [9] and SuperGlue [11] on Revisited Oxford/Paris [10] using ResNet101 [6] as the backbone. The SuperGlue model is pretrained on MegaDepth [7] with SuperPoint [5] as the backbone. The mAP scores on the Medium and Hard setups are reported.

descriptors, our method is inferior to geometry verification on $\mathcal{R}$Oxf-Medium but performs favorably to geometry verification and SuperGlue on the rest settings.

E. Visualizing the correspondences

Following the previous instance recognition [14] and image matching [11] works, we visualize the correspondences learned by RRT in Fig. 2. We extract the attention scores from the last transformer layer (i.e. $\mathbf{Z}_C$) of RRT. Correspondences are computed by solving a linear sum assignment problem [1] using the attentions as the affinity. The examples show that RRT is not good at learning the pixel-wise correspondences of keypoints. It also indicates that rather than estimating the local correspondences, RRT learns distinct knowledge to compute the similarity of images.

F. Limitation

Interpretability. Compared to the homography that explicitly models the alignment of the image-pair, the similarity score predicted by our model is less interpretable. In the future, we'd like to extend the work to learning more visual relation concepts, e.g. homography, dense matching, optical flow, which may lead to more interpretable results.

Domain shift. In the DELG [2] experiment, our method is trained on Google Landmarks v2 [15] and tested on Revisited Oxford/Paris [10]. In the SOP [12] experiment, the training and test sets have no overlapping instance categories. Both experiments demonstrate that the proposed Reranking Transformer can transfer the knowledge across different instance categories to a certain extent. On the other hand, similar to all learning-based approaches, our method might have difficulty in handling large domain shifts. It is also a major challenge for most of the recent approaches as another key component of the image retrieval pipeline, the feature extractor, may also suffer from domain shift. Learning transferable feature representation/matching could be an interesting topic for future research.

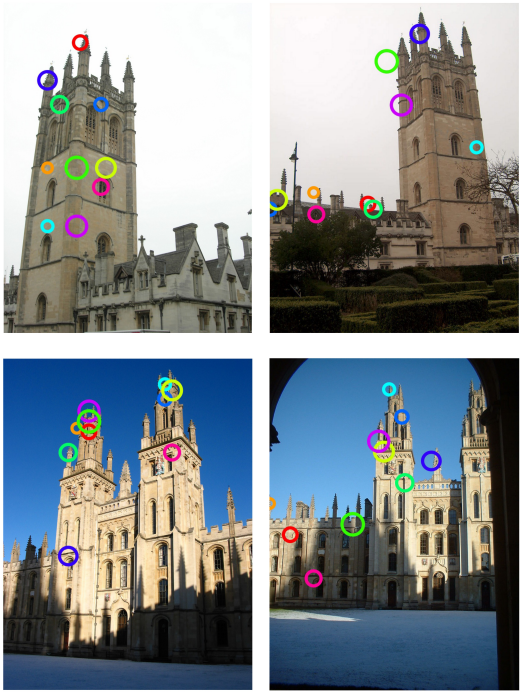

Figure 2. Visualization of the correspondences estimated by a trained RRT model (RRT-R50-v2-clean). Each row shows a pair of matching images. Two keypoints with the same color and scale are considered as a correspondence.

G. More qualitative examples

In Fig. 3, we provide qualitative examples on Stanford Online Products [12]. Here, we compare the results from the global-only model (*CO*) and the proposed model (*CO + RRT (finetuned)*). In particular, we showcase the examples of rigid objects (e.g. coffee maker, kettle) and deformable objects (e.g. stapler, lamp). The proposed method outperforms the global-only retrieval on challenging cases such as partial-matching (example (A)(C)(D)), articulated objects (example (E)(F)), and irrelevant context (example (B)).

In Fig. 4, we provide reranking examples produced by geometry verification and the proposed Reranking Transformer on Revisited Oxford/Paris [10]. It is shown that, compared to geometry verification, the proposed method performs favorably when large viewpoint variations are present. For example, the queries in example (A) and (B) represent the same landmark but exhibit a large viewpoint change. While geometry verification predicts two different sets of top neighbors, our model predicts the same set of top ranked images for the two queries. Example (E) and (F) show failure cases of our model.

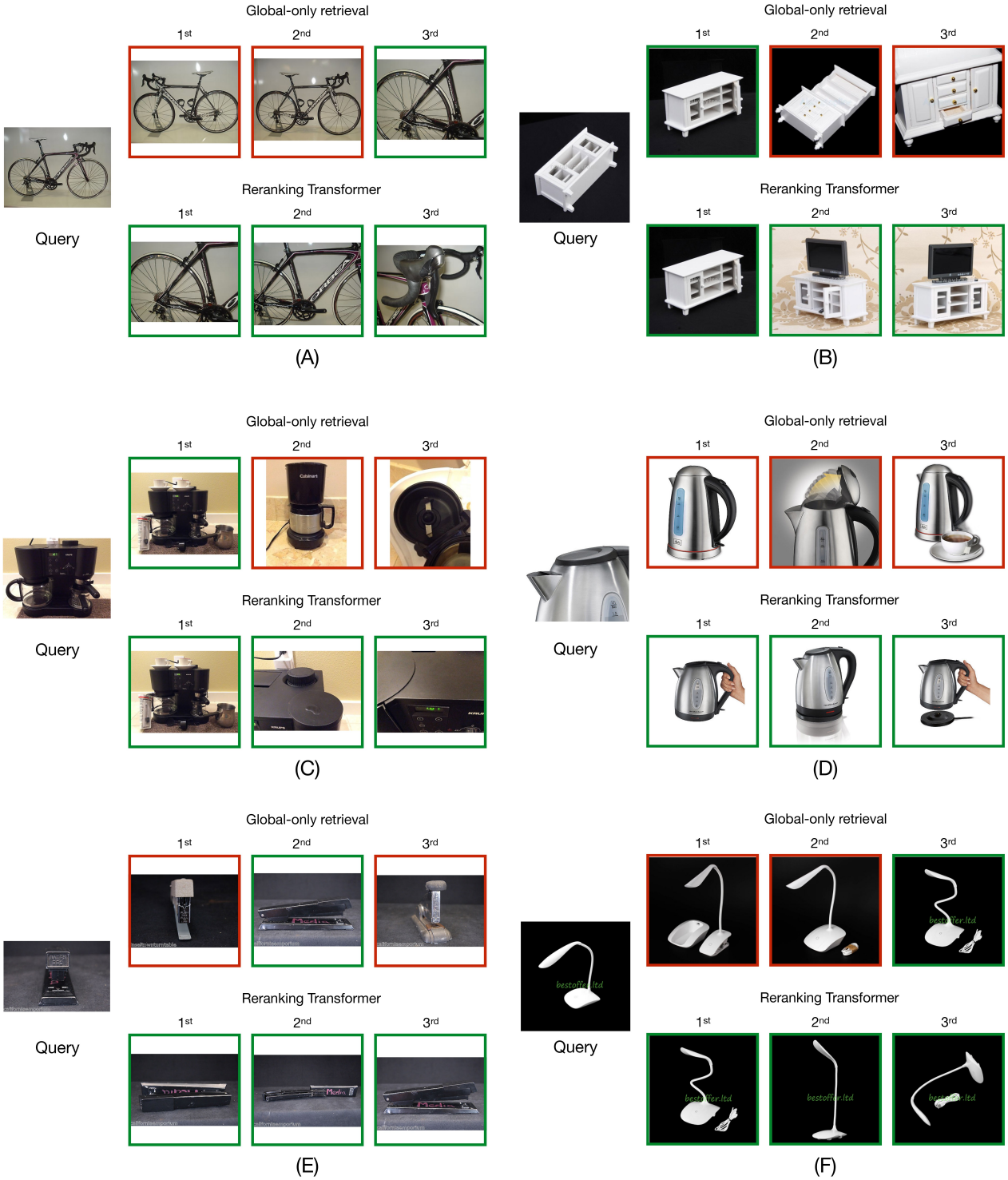

Figure 3. Qualitative examples from Stanford Online Products [12]. For each query, the top-3 neighbors predicted by the global-only retrieval and the proposed Reranking Transformer are presented. Correct/incorrect neighbors are marked with green/red borders.

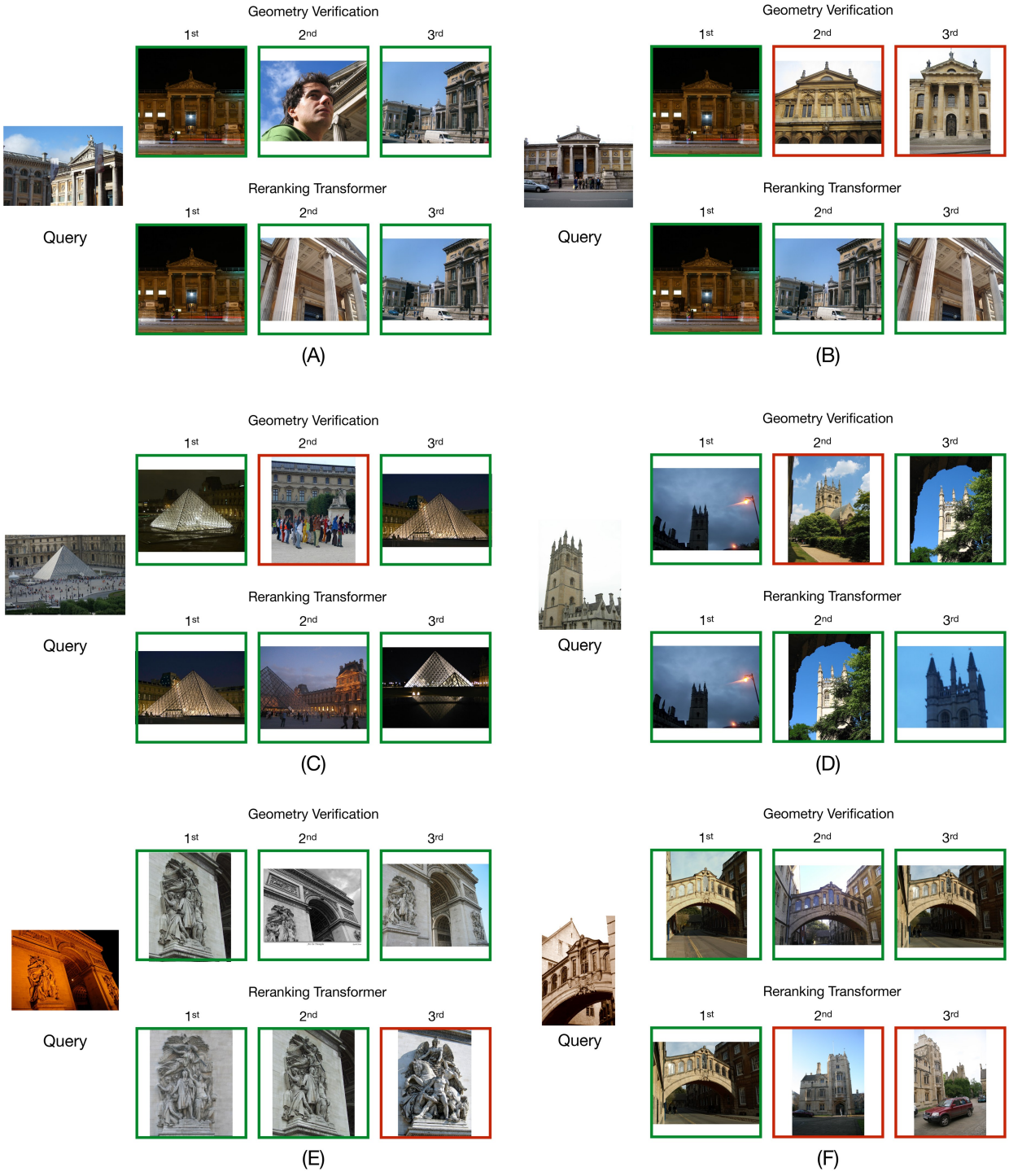

Figure 4. Qualitative examples from Revisited Oxford/Paris [10]. For each query, the top-3 neighbors predicted by geometry verification and the proposed Reranking Transformer are presented. Correct/incorrect neighbors are marked with green/red borders.

References

[1] Assignment problem. https://en.wikipedia.org/wiki/Assignment_problem. 3

[2] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *Eur. Conf. Comput. Vis.*, 2020. 1, 2, 3

[3] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Int. Conf. Comput. Vis.*, 2007. 1

[4] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2018. 1, 2, 3

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 1, 3

[7] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2, 3

[8] Hyeonwoo Noh, Andre Araujo, Jack Sim, and Bohyung Han. Image retrieval with deep local features and attention-based keypoints. In *Int. Conf. Comput. Vis.*, 2017. 2

[9] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1–8, 2007. 1, 2, 3

[10] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1, 2, 3, 5

[11] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 2, 3

[12] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 1, 2, 3, 4

[13] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. Image search with selective match kernels: Aggregation across single and multiple images. *Int. J. Comput. Vis.*, 116(3):247–261, 2016. 1

[14] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *Eur. Conf. Comput. Vis.*, 2020. 3

[15] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 – a large-scale benchmark for instance-level recognition and retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 2, 3