Supplementary Material
R^2Former: Unified Retrieval and Reranking Transformer for Place Recognition

Sijie Zhu[1,2,†], Linjie Yang[1], Chen Chen[2], Mubarak Shah[2], Xiaohui Shen[1], Heng Wang[1]
[1] ByteDance Inc. [2]Center for Research in Computer Vision, University of Central Florida
{sijiezhu,linjie.yang,heng.wang,shenxiaohui.kevin}@bytedance.com, {chen.chen,shah}@crcv.ucf.edu

A. Overview

In this supplementary material, we provide the following content for a better understanding of the paper:

B. Limitations and Societal Impact.

C. Implementation Details.

D. Performance of R^2Former Trained on Pitts30k.

E. Performance of R^2Former with Different Number of Reranking Candidates.

F. Qualitative Results of Retrieval and Reranking.

G. Qualitative Results on Matched Pairs.

H. The Snapshot of MSLS Leaderboard.

I. Model Size.

J. Explanation of Learning Patch-level Correspondence.

K. Results on Nordland Dataset.

L. Different Training Settings.

M. Compare with MixVPR.

B. Limitations and Societal Impact

One limitation of our method is that our reranking module is not as explainable as RANSAC [3] and it does not guarantee a correct geometry correspondence between the two images. Based on the ablation study, most of the information used for reranking is still the correlation between local features, and the geometric information is not fully exploited. We might introduce direct homography estimation and verification into our model in future work. Another limitation is that our reranking module needs to be trained with the global hardest negative samples from the full dataset. A more elegant training strategy could be designed to achieve complete end-to-end training with a simple sampling strategy.

The proposed method could be used to improve localization or navigation systems of a wide range of real-world applications, and the authors do not foresee any negative societal impact.

C. Implementation Details

The global retrieval module is trained with Adam [5] optimizer with a learning rate of 0.00001. Each batch samples 16 triplet pairs, where each pair contains a query, a positive, and two negative reference images. The negative samples are generated using partial negative mining [2] on MSLS [8]. The module is trained until the recall@5 on the validation set is not improving. 50000 queries are sampled in each epoch for MSLS.

The partial negative mining [2] and positive sampling mining follow the default setting of the VG benchmark [2]. We did not use the full negative mining implementation of [2] to train the reranking module because it is very time-consuming. Instead, we freeze the reranking module first and precompute the global hardest samples for all the queries on GPU which is much faster in practice. It can fit in a single GPU because the dimension of our global retrieval module is only 256.

The 2D interpolation is conducted on the positional embedding (every channel is reshaped to width×height) so that its size is always the same as the input size w/p × h/p. We use the pre-trained best-performing NetVLAD models from VG benchmark [2] where PCA is not used by default.

D. Performance of R^2Former Trained on Pitts30k

In Table 1, we show the performance of our models and a typical standard method from [2] ("ResNet101+GeM") on Pitts30k [7] test set. The models are trained on either Pitts30k or MSLS [8] dataset. For global retrieval methods, *i.e.* "ResNet101+GeM", and "Ours w/o Reranking", training on Pitts30k achieves better performance than its

† This work was done during the first author's internship at ByteDance Inc.

	Trained on Pitts30k [7]			Trained on MSLS [8]		
	R@1	R@5	R@10	R@1	R@5	R@10
ResNet101+GeM	83.2	92.5	94.8	77.0	89.2	92.5
Ours w/o Reranking	77.7	90.5	93.5	73.1	88.7	92.5
Ours	**85.8**	**93.2**	**95.3**	**88.4**	**94.2**	**95.7**

Table 1. Performance on Pitts30k [7] test set for our models trained on different datasets.

counterpart trained on MSLS [8], because there is a generalization gap between MSLS and Pitts30k dataset. However, "Ours" trained on MSLS [8] performs significantly better than its counterpart trained on Pitts30k, indicating that Pitts30k is not suitable to train our reranking module. Different from RANSAC [3], our reranking module is data-driven, so it prefers a large-scale training dataset like MSLS. Besides, the reference/database images in Pitts30k are extracted from panorama, where 24 images are generated from each panorama with the same location. Although the positive sample mining [1, 2] is adopted, the positive samples are not guaranteed to visually overlap with the corresponding query images, which could lead to wrong supervision information for our reranking module. Therefore, we train our model on MSLS [8] dataset by default and finetune on Pitts30k if necessary.

E. Performance of R^2Former with Different Number of Reranking Candidates

In Table 2, we show the performance of our method (R^2Former) and "Ours+RANSAC" (RANSAC with our backbone) with different numbers of reranking candidates. There is no observable improvement when the number of candidates is increased from 100 to 200 for both methods, indicating that 100 candidates are enough for reranking-based methods. We also validate that using only 20 candidates does not cause much performance drop ($\sim$ 1%), which could be a good trade-off to reduce the computational cost by $5\times$.

	Number of Candidates	MSLS R@1	MSLS R@5	MSLS R@10
Ours + RANSAC	20	85.1	93.2	94.7
	50	85.0	92.8	94.2
	100	84.9	93.0	94.5
	200	84.2	92.3	93.6
Ours	20	88.9	93.4	94.9
	50	89.1	94.2	95.1
	100	**89.7**	95.0	**96.2**
	200	89.5	**95.1**	95.8

Table 2. Performance of our methods on MSLS [8] dataset with different numbers of reranking candidates.

F. Qualitative Results of Retrieval and Reranking

In Figs. 1, 2, and 3, we show qualitative results of our methods on challenging scenarios of MSLS [8] dataset, *i.e.* dramatic lighting change, seasonal change, and viewpoint variation. All the methods are based on the same backbone, "No Reranking" only adopts our global retrieval module. "Ours" and "RANSAC" adopt our reranking module and RANSAC [3] respectively on our global retrieval module. Both "Ours" and "RANSAC" perform better than "No Reranking". "Ours" shows strong robustness in these challenging scenarios.

G. Qualitative Results on Matched Pairs

In Fig. 4, we show the qualitative comparison between RANSAC and our reranking module in terms of local pairs. Although our reranking module does not guarantee geometric correspondence, most of the highlighted local pairs of our reranking module are correct local matches.

H. The Snapshot of MSLS Leaderboard

In Fig. 5, we show the snapshot of the MSLS [8] leaderboard at the time of submission, and the proposed method (named "Anonymous006" for double-blind policy) is ranked 1st among all methods. Its recall@5 is much better than the other methods.

I. Model Size

Table 3 shows ViT-Small and ResNet backbones have very similar model sizes, *e.g.* # of parameters, and GFLOPs.

Method	# of para.	GFLOPs
ResNet+GeM	23.5 M	25.2
ViT-Small	22.2 M	25.9

Table 3. Model sizes of different backbones.

J. Explanation of Learning Patch-level Correspondence

Although only image-level supervision is used, similar patches are still likely to have similar features/tokens in

Figure 1. Qualitative results of our methods with different reranking configurations on dramatic **lighting change**. Top-5 matching results are presented with green/red boxes for correct/wrong predictions.

Figure 2. Qualitative results of our methods with different reranking configurations on **seasonal change**. Top-5 matching results are presented with green/red boxes for correct/wrong predictions.

the embedding space. The raw similarity/correlation information might not be accurate, but our reranking module learns to determine whether the two images are correct matches according to the inaccurate correlation matrix. The image-level supervision helps the model put more attention on candidate patch pairs with high confidence to be correct local matches. It does not estimate any homography transformation, and thus is not as explainable as RANSAC. However, it could take advantage of additional information to get better performance than RANSAC.

K. Results on Nordland Dataset

We follow Patch-NetVALD [4] GitHub to download Nordland [6] dataset. Table 4 shows that our method significantly outperforms Patch-NetVLAD.

Method	R@1	R@5	R@10
Patch-NetVLAD	44.9	50.2	52.2
Ours	**60.6**	**66.8**	**68.7**

Table 4. Comparison with PatchNetVLAD on Nordland dataset.

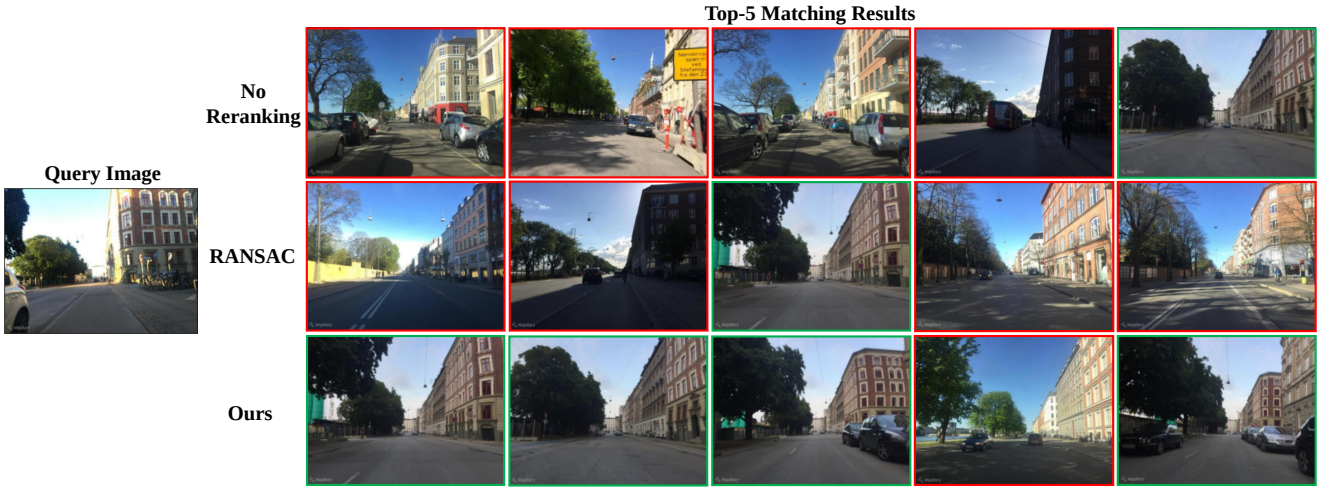


Figure 3. Qualitative results of our methods with different reranking configurations on **viewpoint variation**. Top-5 matching results are presented with green/red boxes for correct/wrong predictions.

L. Different Training Settings

Our results with the model trained only on MSLS are included in Table 2. We also evaluate the Pitts30k fine-tuned model on MSLS Val in Table 5 and it still outperforms TransVPR.

Method	R@1	R@5	R@10
TransVPR	86.8	91.2	92.4
Ours-Pitts30k Finetuned	**88.4**	**94.5**	**95.4**

Table 5. Evaluation of Pitts30k-finetuned model on MSLS Val.

M. Compare with MixVPR

Since MixVPR is published in 2023 (after the CVPR submission deadline) and the code is not released, we are not able to reproduce it during the rebuttal. Based on their reported results, our method performs much better (9% ↑) on the hold-out MSLS Challenge set in Table 6.

Method	R@1	R@5	R@10
MixVPR	64.0	75.9	80.6
Ours	**73.0**	**85.9**	**88.8**

Table 6. Comparison with MixVPR on MSLS Challenge set.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 2

[2] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5407, 2022. 1, 2

[3] 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*, 24(6):381–395, 1981. 1, 2

[4] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14141–14152, 2021. 3

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[6] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. In *Proc. of workshop on long-term autonomy, IEEE international conference on robotics and automation (ICRA)*, page 2013, 2013. 3

[7] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 883–890, 2013. 1, 2

[8] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2626–2635, 2020. 1, 2, 6

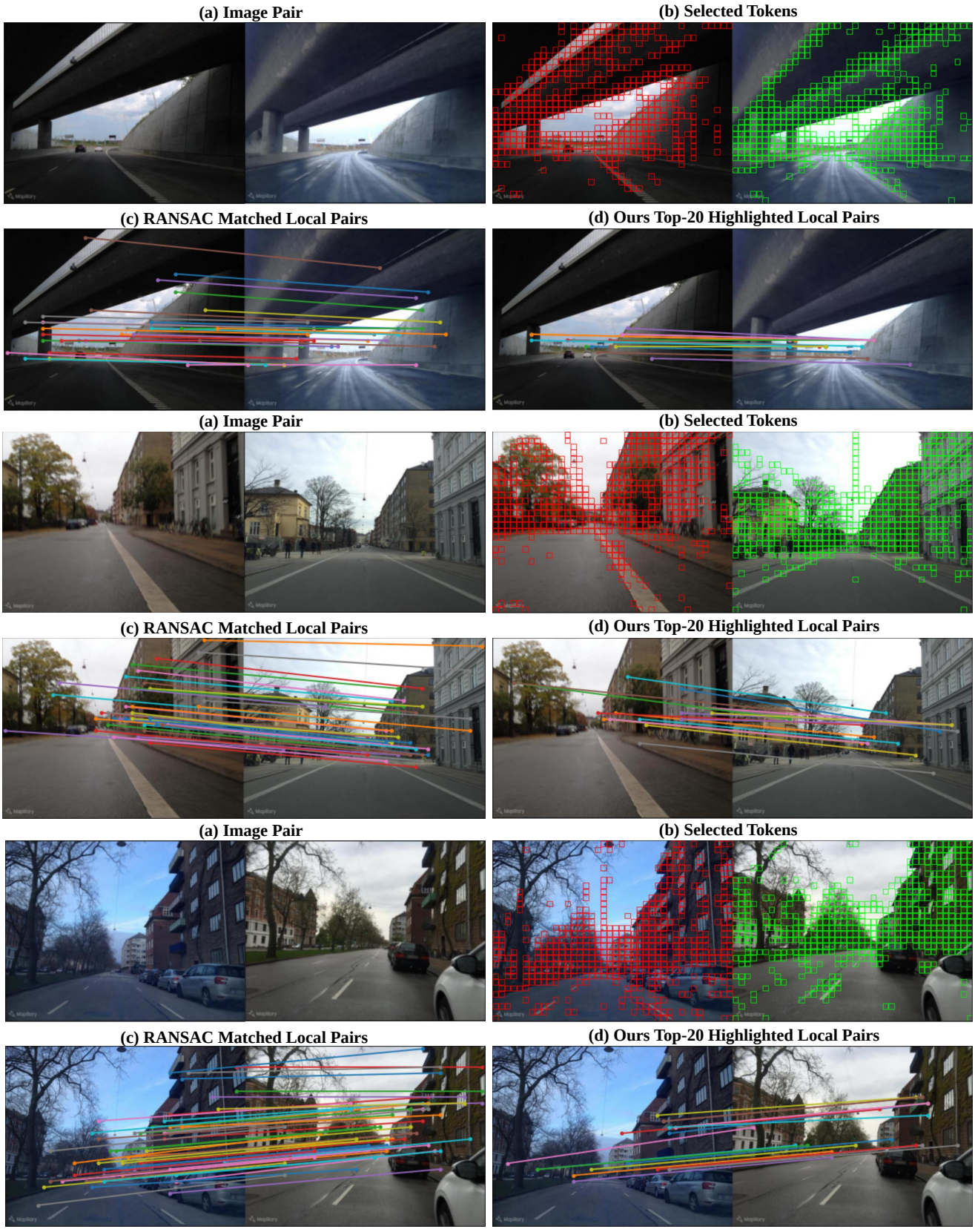
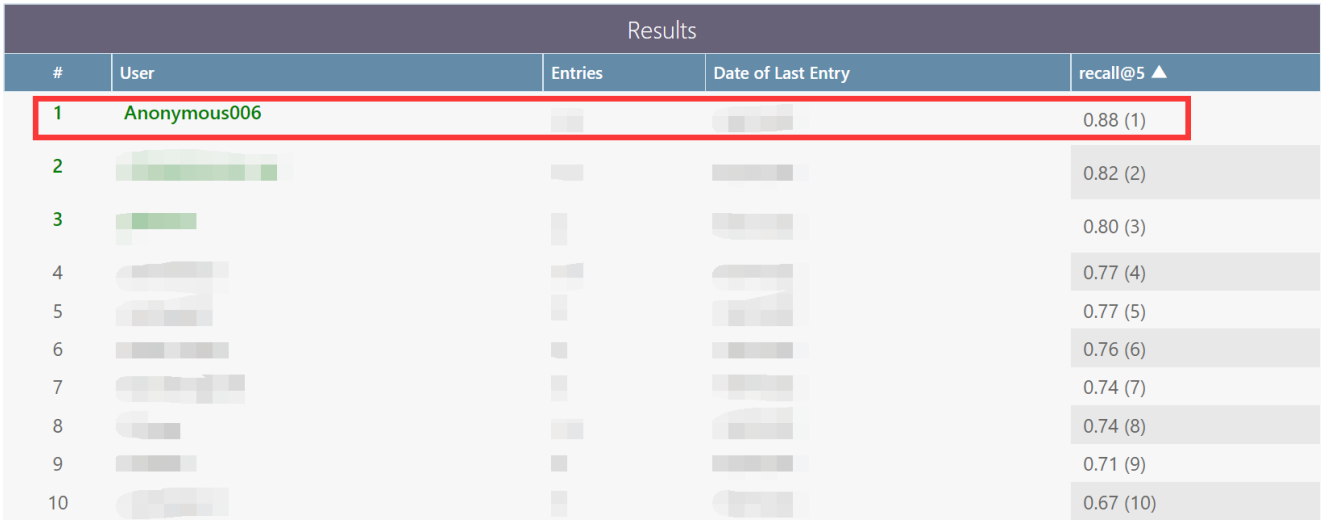


Figure 4. Qualitative results on selected local pairs of RANSAC and our reranking module.



Results

#	User	Entries	Date of Last Entry	recall@5 ▲
1	Anonymous006			0.88 (1)
2				0.82 (2)
3				0.80 (3)
4				0.77 (4)
5				0.77 (5)
6				0.76 (6)
7				0.74 (7)
8				0.74 (8)
9				0.71 (9)
10				0.67 (10)

Figure 5. The snapshot on MSLS [8] leaderboard. The proposed method named "Anonymous006" (for double-blind policy) is ranked 1st at the time of submission.