Viewpoint Invariant Dense Matching for Visual Geolocalization Supplementary material

Gabriele Berton[1,2], Carlo Masone[2], Valerio Paolicelli[2] and Barbara Caputo[1,2]
[1]Politecnico di Torino [2]Italian Institute of Technology
[gabriele.berton, barbara.caputo]@polito.it [carlo.masone, valerio.paolicelli]@iit.it

1. Qualitative results

Figure 1 displays qualitative results that extend the quantitative discussion of the experiments presented in the main paper. The figure showcases the results achieved on a few queries from the R-Tokyo [5] dataset. The selected queries are a good example of challenging conditions for urban visual geolocalization, as they are images taken at night/evening (whereas the gallery is built from images captured during the day) and with many dynamic and static objects occluding the scene in the background. In particular, we show the top-1 retrieved result from a shortlist of 5 predictions. We note that the best baseline (ResNet-50 [3] + NetVLAD [1]) fails in all the examples. Spatial verification using the method from DELG [2] manages to correctly localize only a few of the queries. Lastly, building our proposed dense matching method but using the warping module from [4], which is not tailored for the geolocalization task, only manages to retrieve a correct prediction for one of the queries, particularly when the query and the prediction have only small perspective difference and a large overlap. On the other hand, GeoWarp manages to correctly localize all but one query. However, this failure case is due to the fact that the top-5 predictions retrieved by the global search on the gallery do not contain any positive match for the query, therefore re-ranking them cannot help improving the localization. The last two rows from Fig. 1 illustrate the warped query and the first prediction generated with GeoWarp.

2. Effects of k on the warping operation

The main paper includes a quantitative ablation study on the impact of $k \in [0, 1]$ on the recall@1 metric. Here, we give further intuition about the influence of k on the self-supervised generation of the training quadruplets $\{I_a, I_b, \boldsymbol{t}_a, \boldsymbol{t}_b\}$. In Fig. 2 we see that for small values of k the training quadruplets are generated sampling quadrilaterals whose corners are close to the corners of the image I. This means that the two images I_a and I_b have large overlaps and little perspective difference. On the other hand, high values of k lead to the images I_a and I_b to have have little overlap, simulating very different views of the same scene.

This effect on the generation of the training quadruplets $\{I_a, I_b, \boldsymbol{t}_a, \boldsymbol{t}_b\}$ translates to the fact that the network trained with higher values of k learns to perform stronger warping. We can see this qualitatively in Fig. 3, which showcases a few examples of the warped images generated by various query-prediction pairs, with our homography regression network trained using different values of k. We can observe in Figs. 3a and 3d, that a network trained with small values of k is not capable to compensate for strong viewpoint shifts and images with small overlap. On the other hand, the model trained with high values of k may become able to take two images of the same scene but at far away locations from each other and transform them to have similar appearance (see Fig. 3b). This explains why higher values of k are better suited for the case of a rougher geolocalization. Finally, Fig. 3c shows the result produced when the model is given a query and a false prediction. We can see that, regardless of the value of k, the warping operation has little effect. This demonstrates that our dense matching is rather robust to the case when the predictions to be re-ranked contain false positives.

3. Comparison with other warping methods

In the main paper we have discussed how our warping regression module is inspired by the work of Rocco et al. [4], detailing the conceptual differences and corroborating these considerations with quantitative and qualitative results. In Fig. 4 we provide further qualitative evidence to give a better intuition of the differences between the two methods. Firstly, since [4] only transforms one image while keeping the other one unchanged, it gives rather different results depending on whether the transformed image is the query or the prediction (see Fig. 4a). On the other hand, our method transforms both images achieving a higher robustness. Moreover, the single image transformation from [4] can lead to the creation of artifacts, as pixels outside

of the source image boundaries are filled with grey color, which might further complicate the retrieval task (see Figs. 4a and 4c). Finally, Fig. 4b shows a difficult example with little overlap, in which GeoWarp clearly manages to output similar representations whereas [4] has almost no effect.

References

[1] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1437–1451, 2018.

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

[3] 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.

[4] I. Rocco, R. Arandjelović, and J. Sivic. Convolutional neural network architecture for geometric matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 2553–2567, 2018.

[5] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020.

Figure 1. Qualitative results, each column corresponds to one test case. The first row represents the queries, the next four rows show the first prediction with various methods, and the last two rows show the warped queries and first prediction with GeoWarp (**Ours**).

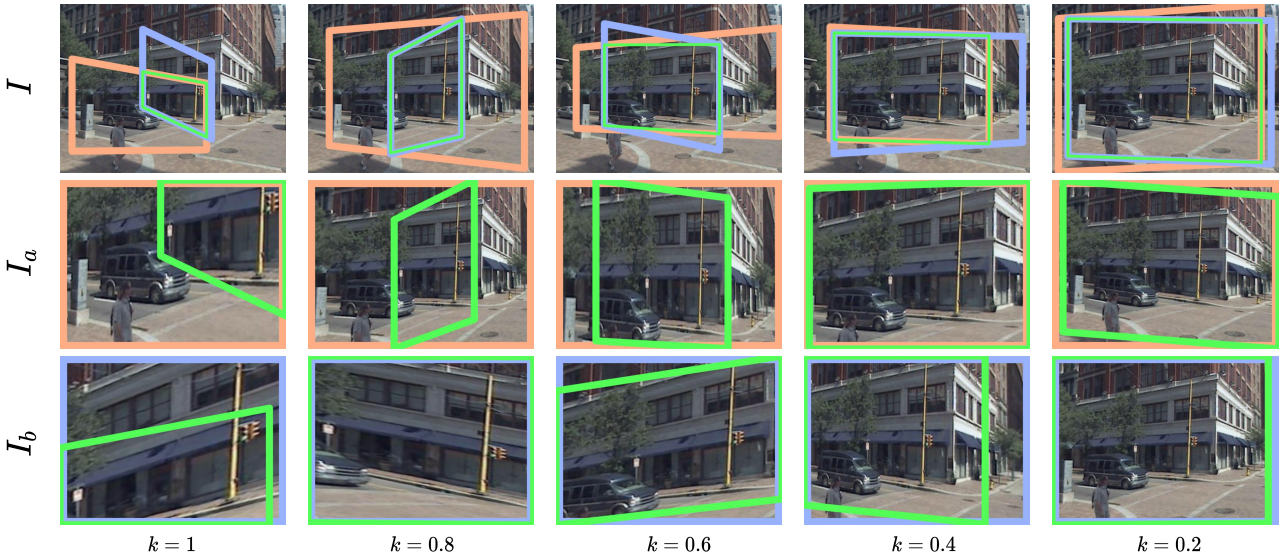

Figure 2. Examples of training quadruplets generated in a self-supervised fashion, using different values of k (1, 0.8, 0.6, 0.4, 0.2). For each column, the top image represents the source image I, the two other images represent I_a and I_b. The green quadrilaterals in the second and third rows indicate $\boldsymbol{t}_a$ and $\boldsymbol{t}_b$, respectively.

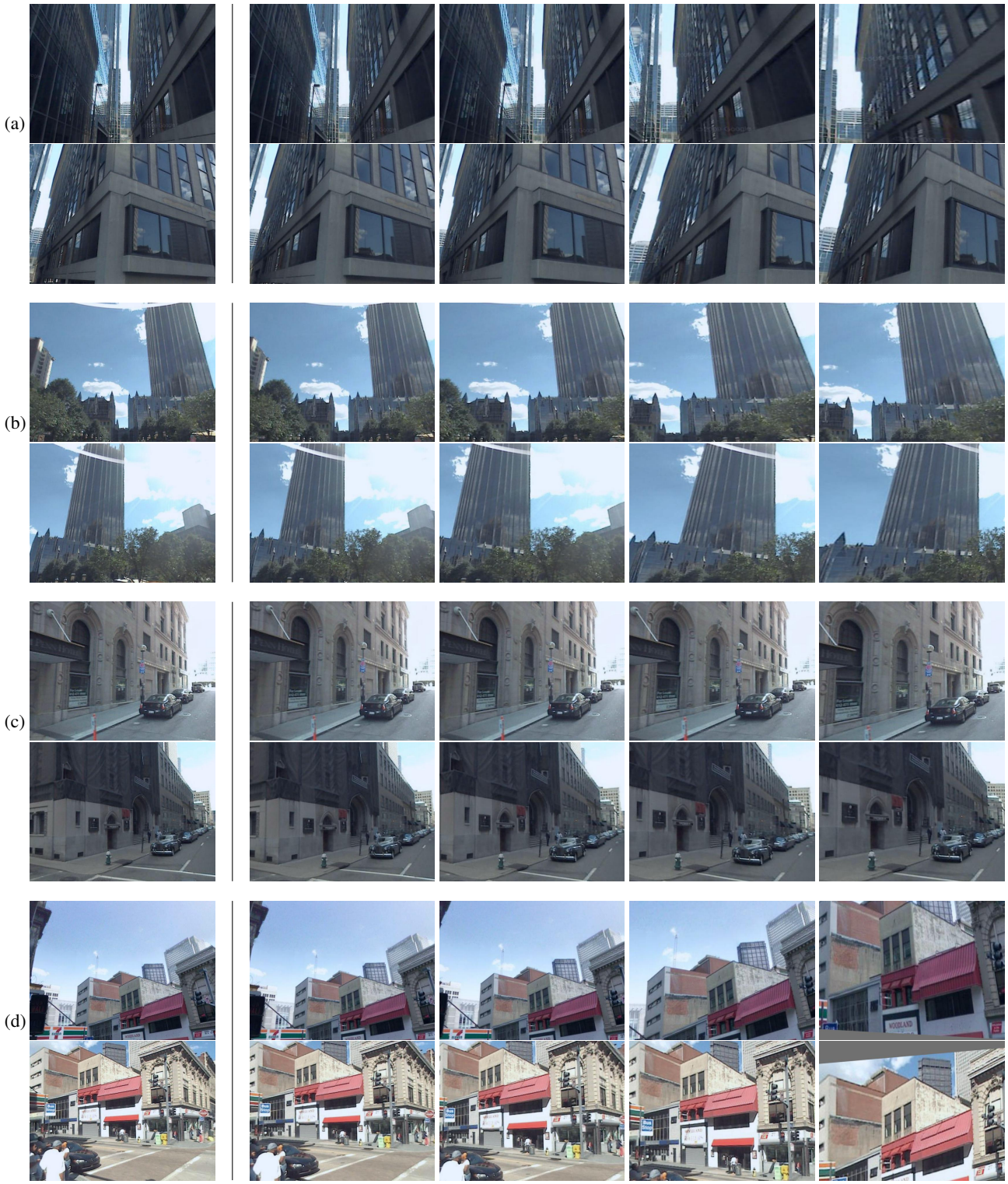

Figure 3. Qualitative results: each pair of rows corresponds to one test case. The first column represents a query-prediction pair, the other columns show our pairwise warping's results using models trained with different values of k (0.3, 0.5, 0.7, 0.9): a) the two images are 1 meter away, in this case a heavier warping is helpful; b) the two images are 39 meters away, and a heavy warping might be useful, depending on the chosen threshold for positive images; c) the prediction is wrongly retrieved by the global features, and warping has little to no effect on the warped pair, regardless of k's value; d) a query-prediction pair with little visual overlap.

Figure 4. Qualitative results of homography, each pair of rows corresponds to one test case. The first column represents a query-prediction pair, the second column shows warping on the prediction using [4], the third shows warping on the query using [4], and the rightmost column is the output of our pairwise warping, using our best network (ResNet-50 [3] backbone trained with $k = 0.6$.)