Supplementary material: Adaptive-Attentive Geolocalization from few queries: a hybrid approach

Gabriele Moreno Berton*, Valerio Paolicelli*, Carlo Masone and Barbara Caputo
Italian Institute of Technology
Turin, Italy
[gabriele.berton, valerio.paolicelli, carlo.masone]@iit.it barbara.caputo@polito.it

1. Additional dataset details

The main idea behind SVOX is to build a dataset that contains the RobotCar dataset [4], in order to test the accuracy of cross-domain visual geolocalization methods. To create the dataset we downloaded images from Google Street View, which provides 360° equirectangular panoramas at various resolutions. From each panorama we then cropped two rectangles at opposite sides, corresponding to the front and rear view of the car.

The original resolution of the images from the RobotCar dataset [4] is 1280x960, and we resized them to 512x384, keeping the original ratio of 4:3. We resized the images cropped from Google Street View panoramas to the same size, again keeping the same ratio.

Thanks to the Google Street View Time Machine we are able to download panoramas taken in the same location in different years. We chose to use images from the years of 2012 and 2014 as gallery and queries respectively, as these are the years with most panoramas in the Oxford area. Moreover, using gallery and queries taken in different years helps to ensure that methods that achieve accuracy must focus on long-term elements, instead of short-term or changing elements such as vegetation or scaffolding. The RobotCar dataset [4] was collected between 2014 and 2015, ensuring that the queries from RobotCar [4] are at least two years apart from the SVOX gallery. Some examples are shown in Fig. 1.

To build SVOX we chose a geographical area that would enclose the whole urban part of the city of Oxford. We then removed by hand images taken in the countryside, given the lack of buildings that are crucial to the geolocalization process. Moreover, we removed queries (from both SVOX and RobotCar [4]) which do not have a positive image within gallery, i.e. and image within 25 meters of distance. Finally we split SVOX in train, validation and test sets. As shown in Fig. 1 of the main paper, the RobotCar dataset [4] is included only in the train and test set, as it is intended to be used only as an unlabeled target dataset for domain adaptation, therefore not requiring a validation set.

*The authors equally contributed

2. Qualitative results

In Figs. 2, 3 and 4 we show for each target scenario of RobotCar [4], some visualizations of top1 images retrieved by our method (AdAGeo) versus the best baseline (NetVLAD [1] + GRL [2]), which are trained and tested with the ResNet18 [3] as encoder.

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 (CVPR)*, June 2016.

[2] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 2015. PMLR.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[4] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 2017.

(a) (b) (c) (d)

Figure 1: Examples of Oxford places at different times by means of Google Time Machine API. On the top row there are the images from 2012 used as gallery set, while on the bottom row there are the images from 2014 used as query set.

Figure 2: Comparison between our method and the best baseline, showing the top1 images retrieved for the target scenario Snow. The images with green border correspond with the ground truth, while the ones with a red border are wrong predictions.

(a)

(b)

Figure 3: Comparison between our method and the best baseline, showing the top1 images retrieved for the target scenarios Rain (a) and Sun (b). The images with green border correspond with the ground truth, while the ones with a red border are wrong predictions.

(a)

(b)

Figure 4: Comparison between our method and the best baseline, showing the top1 images retrieved for the target scenarios Night (a) and Overcast (b). The images with green border correspond with the ground truth, while the ones with a red border are wrong predictions.