Rethinking Visual Geo-localization for Large Scale Applications: Supplementary Material

Gabriele Berton
Politecnico di Torino
gabriele.berton@polito.it

Carlo Masone
CINI

Barbara Caputo
Politecnico di Torino

1. Further information on SF-XL

General information. In Fig. 1 we show the density of the training set (*i.e.* number of panoramas within each cell), in Fig. 2 is its temporal distribution, and in Fig. 3 we show the temporal variability of SF-XL test v1's queries. The StreetView images composing the train set, val set and test database, are 512×512 images cropped from 360° panoramas.

SF-XL test v1. While the database from SF-XL test v1 is very homogeneous, given that StreetView images are all taken at daytime with the same camera and good weather, the queries present large degrees of domain changes: there are night images, grayscale, with heavy changes in viewpoint and occlusions. Coming from the crowd-sourced platform Flickr, these queries are collected by a large number of users, also ensuring variety in the typologies of cameras. We resized these images so that their shorter side is 480 pixels. In Fig. 4, we show more examples of queries, besides the ones shown in Fig. 2 of the main paper.

SF-XL test v2. While the database of SF-XL test v2 is the same as SF-XL test v1, the two sets use different queries. With the advantage of having 6 DoF labels, SF-XL test v2 can also be used for pose estimation. The downside of this set is the homogeneity among its images, given that almost all are taken during sunny days, with clear views and without heavy occlusions. Some examples of the queries are shown in Fig. 4.

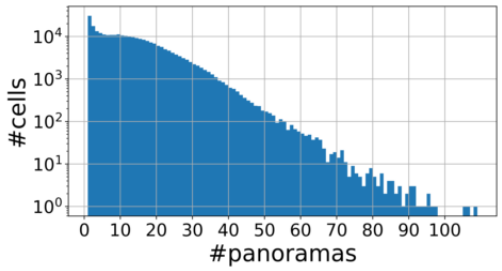

Figure 1. **Histogram showing how many cells contain a given number of panorama.** We can see that cells with only one panorama (which are discarded at train time as explained in Sec. 2.2) are the most common. Note that the y axis is in logarithmic scale. The side of each cell (i.e. the hyperparameter M) is $M = 10$ meters, as in our final experiments.

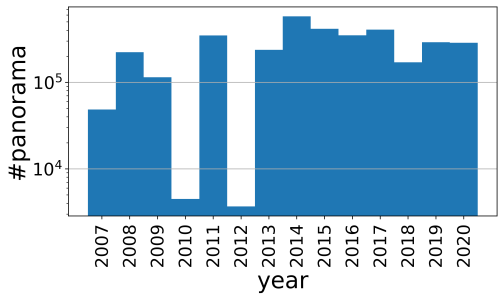

Figure 2. **Number of 360°panorama of SF-XL for each given year.**

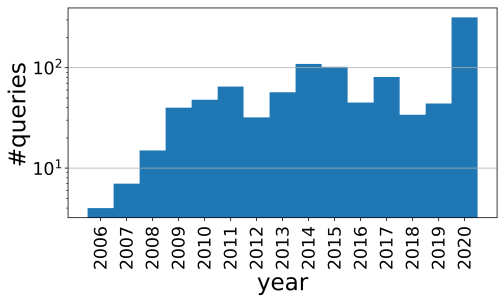

Figure 3. **Number of queries from SF-XL test v1 for each given year.**

2. Experiments

2.1. Further ablations

In this section, we provide further results obtained by changing the hyperparameters of CosPlace to better understand their correlations to the final results.

In Fig. 5 we report an extensive ablation obtained by changing the parameters used to split the dataset into groups and classes, namely M, α, N and L, as well as using a different number of groups for training. Among other results,

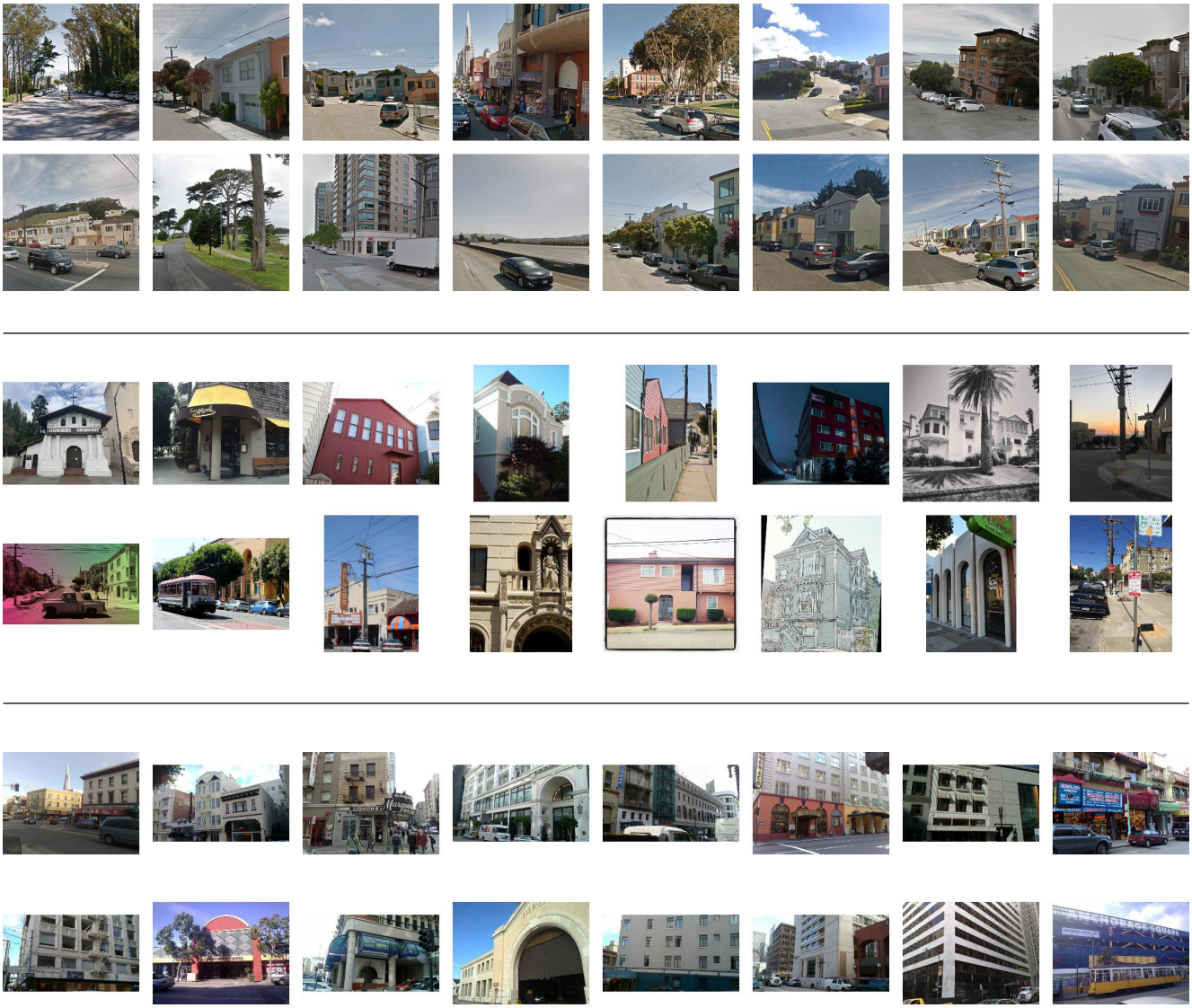

Figure 4. **Examples from SF-XL**. The first two rows of images are from the train set, the next two from the queries of SF-XL test v1, and the last two rows from the queries of SF-XL test v2.

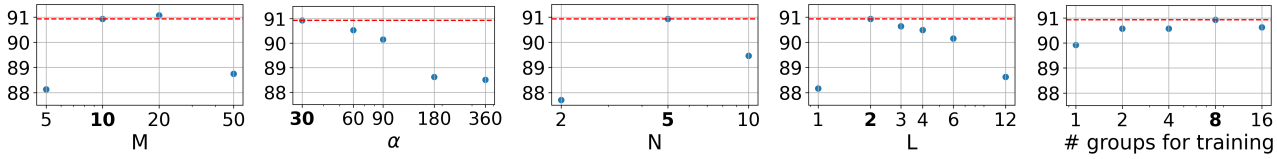

Figure 5. **Full ablation on each hyperparameter.** On the x axis are values for the hyperparameters, and on the y axis their respective recall@1 on the SF-XL val set, computed with a ResNet-18. Values in bold are the chosen ones for all experiments besides ablations, and the red line represents their recall@1.

we see in the rightmost plot that using just a single group for training the model leads to a drop in recall@1 of just 1%, and that the optimal results are achieved using 8 of the 50 groups.

To better understand the importance that the GeM pooling [12] has within the architecture used for CosPlace, we provide a set of experiments by replacing it with the average or max pooling in Tab. 1. From the table, we can see

Pooling	Pitts250k	Pitts30k	Tokyo 24/7	MSLS	St Lucia
Average	88.5	87.6	73.7	78.5	98.7
Max	**90.8**	89.3	78.1	80.5	98.7
GeM	90.4	**89.5**	**81.6**	**81.8**	**98.8**

Table 1. **Ablation over different pooling layers.** This table shows results obtained by replacing the GeM layer with a max or average pooling. Results refer to the recall@1 obtained with a ResNet-18.

that CosPlace would outperform the previous state-of-the-art even with a standard architecture used for classification, made of a CNN backbone, a max pooling, and a fully connected layer.

2.2. Further implementation details

Regarding CosPlace training, to ensure that each class is well represented, only cells with at least 10 panoramas are considered for training, effectively discarding about 15% of the images. The hyperparameters of $M = 10$, $\alpha = 30$, $N = 5$, and $L = 2$ lead to the creation of 50 groups, where each group ends up with roughly 35k classes, and each class contains on average 19.8 images. As explained in the main paper, we only train on 8 (out of 50) groups, which together contain roughly 5.6M images. Note that the total size of the SF-XL training set is 41.2M (*i.e.*, we only use 13.6% of the images), meaning that train-time scalability is a factor that can still be vastly improved in future works.

We use the Adam optimizer [7] with a learning rate of 0.00001, and a batch size of 32 images. We use color jittering as in [3]. For results to be fair with [3], which uses a smart region cropping method, we also employ random cropping. Finally, the margin of the cosFace loss is set to 0.40.

Number of hyperparameters. Although CosPlace introduces a considerable amount of hyperparameters, we also note that there is no more need for many other ones used in previous state-of-the-art methods [1, 3, 9], such as the number of negatives per query (usually set to 10), refresh rate of the cache (1000), pool size of randomly sampled negatives (1000), threshold distance for train-time potential positives (10 meters) and the number of cluster in NetVLAD layer (64). Moreover, the intuitive meaning of the hyperparameters in CosPlace in comparison to the less obvious mining hyperparameters makes it easier to set them using common sense: for example, it is clear that a small M (or α) leads to little intra-class spatial variations, while a large M (or α) may cause two images of the same class to be too different; similarly, using a small value for N leads to a higher similarity between inter-class (but same group) images, while using a very high N leads to classes being very geographically spread out, which can be a problem with smaller datasets (because groups would have few classes).

2.3. Exploratory experiments

Further results on backbones and descriptors dimensionality. Given that previous methods (as recent as 2021) in Visual geo-localization rely on relatively old VGG-16 [13] or AlexNet [8] backbones [1, 3, 6, 9–11, 14], we believe that this is widening an already large gap between research and real-world applications, where one would want to obtain the best possible results with the lowest computational complexity. To narrow such a gap, we investigate how the use of more recent backbones can enhance CosPlace and lead to better results, smaller descriptors, and faster computation. To this end, we train CosPlace using a number of backbones, namely VGG-16 [13], ResNet-18, ResNet-50, ResNet-101 [5], ViT [2], CCT224 and CCT384 [4]. All CNN backbones (*i.e.*, VGG-16 and ResNets) are followed by a GeM pooling [12] and a fully connected layer, Moreover, we experiment with various powers of 2 (from 32-D to 2048-D) as output dimension. Regarding transformers-based neural networks, we use the 384-D SeqPool output of CCT as descriptors, and for ViT, we obtain the 768-D output by feeding the CLS token to a multi-layer perceptron with tanh, following its original implementation [2]. Preliminary results showed that directly using ViT's CLS token led to lower recalls.

Results from Fig. 6 clearly show that CosPlace presents encouraging results regardless of the depth of the backbone, and we argue that future works should focus on more modern architectures, such as the ResNets, which are generally faster, lighter and achieve comparable or better results than the commonly used VGG-16. We also see that CosPlace is able to reach remarkable recalls and robustness with very low dimensions; for example, we see that any 128-D architecture trained with CosPlace outperforms 4096-D NetVLAD (which is trained on Pitts30k) on any test dataset.

While transformers achieve lower results, we want to point out that we used the same hyperparameters for all experiments (*e.g.*, same learning rate and optimizer), and we believe that performing a proper hyperparameter tuning independently for each backbone can increase the results shown in Fig. 6, at the cost of a large number of experiments. Moreover, while we used a resolution of 512×512 for CNNs, transformers require a smaller size, respectively 224×224 for ViT and CCT224, and 384×384 for CCT384, and this can provide a further explanation of the lower results with transformers.

Comparison with other methods using same descriptors dimensionality. Given that CosPlace uses much lower dimensionality of descriptors, in Tab. 2 and Tab. 3 we report the equivalent experiments of Tab. 3 and Tab. 4 of the main paper, but using the same (512) dimensionality for all methods. We can see that in this scenario, the advantages of

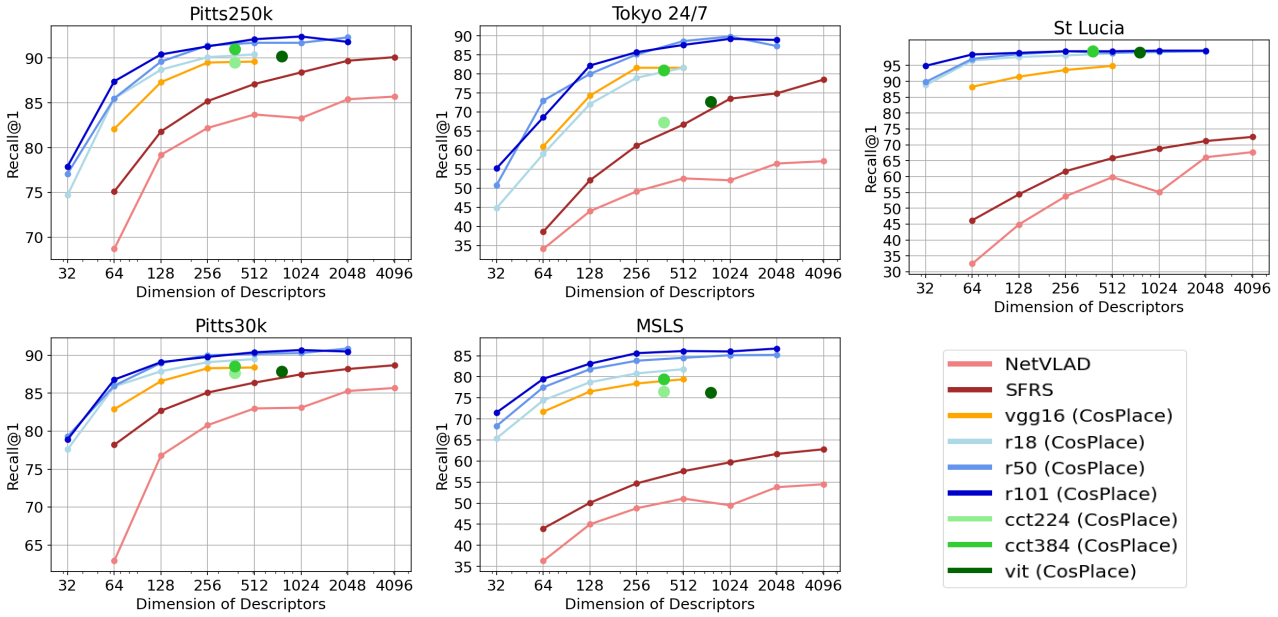

Figure 6. **Further results on backbones and descriptors dimensionality.** Results on a number of datasets of CosPlace using different backbones and dimensionalities, compared with SFRS and NetVLAD trained on Pitts30k.

Method	Desc. dim.	Train set	Pitts250k		Pitts30k		Tokyo 24/7		MSLS		St Lucia	
			R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
GeM [12]	512	Pitts30k	75.3± 0.2	88.4± 0.3	77.9± 0.4	90.5± 0.3	46.4± 0.9	65.3± 0.7	51.8± 0.9	64.4± 0.9	59.9± 1.6	76.3± 2.0
GeM [12]	512	MSLS	65.3± 1.2	81.0± 1.6	71.6± 2.1	85.1± 1.9	44.9± 1.7	62.6± 1.2	66.7± 0.7	78.9± 0.5	84.6± 1.1	93.3± 0.7
GeM [12]	512	SF-XL*	64.7± 0.8	81.4± 0.8	67.8± 0.6	83.6± 0.7	37.9± 2.3	51.0± 2.1	46.8± 2.1	58.1± 1.2	68.5± 2.4	82.7± 1.8
NetVLAD [1]	512	Pitts30k	83.7± 0.3	92.8± 0.1	83.0± 0.2	92.6± 0.3	52.6± 1.1	70.9± 1.2	51.1± 1.0	63.5± 0.8	59.8± 0.5	74.5± 1.2
NetVLAD [1]	512	MSLS	74.6± 1.3	86.8± 1.2	77.0± 0.9	88.6± 1.2	50.5± 2.2	65.1± 1.7	72.6± 0.5	83.0± 0.3	92.6± 0.6	97.1± 0.4
NetVLAD [1]	512	SF-XL*	77.5± 0.5	88.5± 0.2	79.7± 0.3	90.0± 0.4	53.0± 0.9	70.2± 0.5	53.1± 3.2	64.2± 2.2	78.7± 1.6	88.1± 1.8
CRN [6]	512	Pitts30k	84.6± 0.6	93.6± 0.3	84.3± 0.2	92.7± 0.2	53.4± 0.5	70.6± 0.8	54.1± 0.6	66.1± 0.6	56.6± 2.7	75.5± 2.9
APANet [14] †	512	Pitts30k	83.7	92.6	-	-	67.0	81.0	-	-	-	-
SARE [9]	512	Pitts30k	84.3± 0.7	92.6± 0.4	84.7± 0.7	92.6± 0.5	62.0± 0.6	74.9± 0.4	55.8± 3.3	67.8± 3.3	63.4± 3.0	79.0± 1.9
SFRS [3]	512	Pitts30k	87.1± 0.4	94.6± 0.2	86.4± 0.5	93.8± 0.2	66.7± 1.0	79.6± 0.9	57.6± 1.1	68.9± 1.0	65.8± 3.1	80.1± 2.3
SRALNet [10] †	512	Pitts30k	84.8	93.5	-	-	60.6	76.5	-	-	-	-
APPSVR [11] †	512	Pitts30k	85.3	94.0	-	-	62.0	76.5	-	-	-	-
CosPlace (Ours)	512	SF-XL	89.3± 0.2	**96.2± 0.3**	88.5± 0.1	**94.5± 0.2**	**82.2± 0.5**	**88.9± 0.9**	**79.6± 0.5**	**87.2± 0.4**	**94.1± 0.8**	97.4± 0.1

Table 2. **Comparisons of various methods on popular datasets with 512-D descriptors.** This table is the equivalent of Tab. 3 in the main paper, but here all descriptors have the same dimensionality.

Method	Desc. dim.	Train set	SF-XL test v1			SF-XL test v2		
			R@1	R@5	R@10	R@1	R@5	R@10
GeM	512	Pitts30k	21.7	30.3	34.4	43.1	63.7	69.2
GeM	512	MSLS	8.1	15.6	20.2	29.3	46.3	53.8
GeM	512	SF-XL*	9.8	17.6	21.2	34.8	55.5	63.0
NetVLAD	512	Pitts30k	27.4	38.1	43.6	66.7	79.3	82.9
NetVLAD	512	MSLS	14.5	21.0	28.9	40.5	59.7	64.4
NetVLAD	512	SF-XL*	25.4	32.9	40.5	66.9	78.6	82.8
CRN	512	Pitts30k	31.4	43.0	49.7	68.2	81.3	83.3
SARE	512	Pitts30k	30.8	42.1	46.5	69.2	81.1	83.1
SFRS	512	Pitts30k	35.6	49.7	54.8	78.1	88.5	91.3
CosPlace (Ours)	512	SF-XL	**65.1**	**73.6**	**77.6**	**83.4**	**92.1**	**94.8**

Table 3. **Comparisons of various methods on SF-XL test v1 and SF-XL test v2 with 512-D descriptors.** This table is the equivalent of Tab. 4 in the main paper.

CosPlace w.r.t. previous works are even more noticeable.

Comparison with models trained on Google Landmark. In Tab. 4, we compare models trained using CosPlace on SF-XL with models trained on two popular landmark retrieval (LR) datasets, namely the Google Landmark (GLD) and SfM120k [12]. Models trained on GLD and SfM120k are downloaded from the official repository of [12] [1], which relied on a triplet loss for training. CosPlace can't be used on such landmark retrieval datasets, as they lack GPS coordinates and heading labels.

Note that these experiments are not aimed at providing a rigorous comparison of CosPlace vs triplet losses or SF-XL vs standard retrieval datasets, given that the underlying tasks (*i.e.* VG and LR) present many differences; we just want to provide an intuition on how popular models trained for LR fare on VG datasets.

[1] https://github.com/filipradenovic/cnnimageretrieval-pytorch

Training Dataset	Backbone	Pitts250k	Pitts30k	Tokyo 24/7	MSLS	St Lucia
SfM120k	ResNet-50	84.5	83.4	75.2	64.5	73.9
GLD	ResNet-50	85.8	84.1	77.8	69.5	77.3
SF-XL	ResNet-50	**92.3**	**90.9**	**87.3**	**85.2**	**99.5**
SfM120k	ResNet-101	85.0	83.9	77.5	64.7	76.3
GLD	ResNet-101	86.9	85.1	77.8	72.4	83.4
SF-XL	ResNet-101	**91.8**	**90.5**	**88.9**	**86.7**	**99.7**

Table 4. **Comparison with models trained on large landmark retrieval datasets.** The models trained on SF-XL is trained with CosPlace, while models trained on GLD and SfM120k rely on a triplet loss. All models are equivalent (*i.e.* ResNets followed by a GeM pooling and a fully connected layer with output dimensionality 512).

Comparison with other methods: qualitative results. Figure 7 shows some qualitative results of retrieved images with CosPlace compared to previous SOTA methods such as NetVLAD [1], CRN [6], SARE [9] and SFRS [3].

References

[1] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1437–1451, 2018. 3, 4, 5

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929, 2021. 3

[3] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 369–386, Cham, 2020. Springer International Publishing. 3, 4, 5

[4] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the Big Data Paradigm with Compact Transformers. *ArXiv*, abs/2104.05704, 2021. 3

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[6] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. Learned contextual feature reweighting for image geo-localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3251–3260, 2017. 3, 4, 5

[7] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 3

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 3

[9] Liu Liu, Hongdong Li, and Yuchao Dai. Stochastic Attraction-Repulsion Embedding for Large Scale Image Localization. In *IEEE International Conference on Computer Vision*, 2019. 3, 4, 5

[10] Guohao Peng, Yufeng Yue, Jun Zhang, Zhenyu Wu, Xiaoyu Tang, and Danwei Wang. Semantic reinforced attention learning for visual place recognition. In *IEEE International Conference on Robotics and Automation*, pages 13415–13422. IEEE, 2021. 3, 4

[11] Guohao Peng, Jun Zhang, Heshan Li, and Danwei Wang. Attentional pyramid pooling of salient visual residuals for place recognition. In *IEEE International Conference on Computer Vision*, pages 885–894, October 2021. 3, 4

[12] F. Radenović, G. Tolias, and O. Chum. Fine-tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 2, 3, 4

[13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 3

[14] Yingying Zhu, Jiong Wang, Lingxi Xie, and Liang Zheng. Attention-based pyramid aggregation network for visual place recognition. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 99–107. ACM, 2018. 3, 4

Figure 7. **Qualitative comparisons of retrieved images for a number of methods.**