Supplementary Material for SeqMatchNet: Contrastive Learning with Sequence Matching for Place Recognition & Relocalization

Sourav Garg
QUT, Australia
s.garg@qut.edu.au

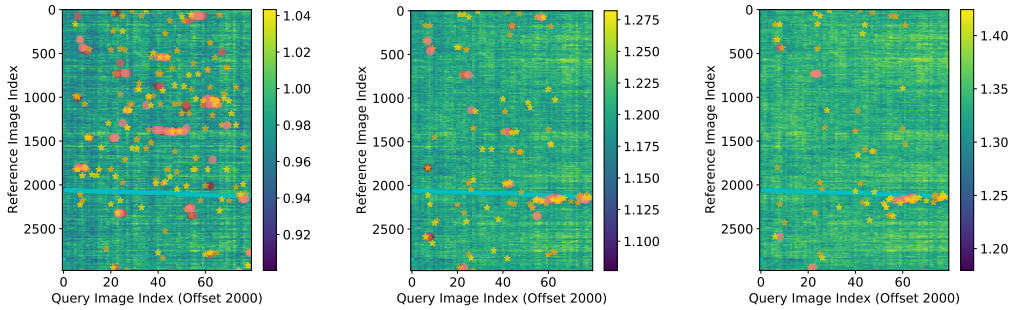
Madhu Vankadari
University of Oxford, UK

Michael Milford
QUT, Australia

In this supplementary, we provide a visualization of negatives during training, insights into the latent space distance margins post training, a comparison of qualitative matches, details of the datasets and training settings, and additional experiments and analysis considering deeper network configurations and cross-testing of single/sequence matching with/without sequence-based training.

1 Negatives Distribution during Training

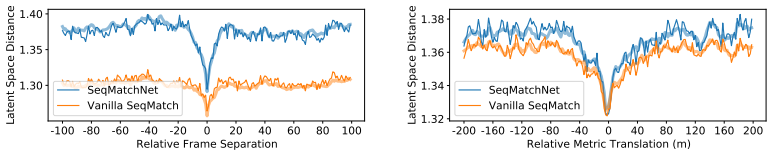
Figure 1 shows the distance matrix with top-3 (dark color to light color) hard negatives per query for the training split of the Oxford dataset [1], computed using single image matching (yellow stars) and sequence-based matching (orange circles). The cyan colored line near the bottom of each matrix shows the ground truth matches (positives). In the leftmost matrix (after Epoch 1), it can be observed that the sequence-based negatives (circles) are far fewer than the single image based negatives (stars). This occurs because most of the queries are *easier* for sequence matching than for single image matching. This effect is particularly visible for the columns of query region around index 30, where stars outnumber circles. Now consider a few query images on the immediate left of the query index 60 – several stars can be found scattered across the rows whereas the circles are mostly clustered at the top. This shows how harder negatives from a particular region of the environment can be tapped by sequence matching while single image matching randomly points to different regions of the environment. As training progresses, sequence-matching based negatives are prioritized and thus by the end of 50 epochs, most of the circles that remain are the ones close to the positives (in cyan), which are typically much harder to resolve due to visual overlap between some of the positive and negatives since images beyond 20 meters are considered as negatives both for training and evaluation.



Epoch 1 Epoch 10 Epoch 50

Figure 1: Hard negatives plotted over distance matrix for the Oxford dataset computed using single image (yellow stars) and sequence-based (orange circles) matching.

5th Conference on Robot Learning (CoRL 2021), London, UK.



(a) Nordland (b) Oxford

Figure 2: Latent space distance vs relative frame separation between a query and its nearby reference images. Distance margin, that is, the difference between the trough and the peak values of a curve is relatively higher when using SeqMatchNet-based NetVLAD transformation (blue) than using Vanilla SeqMatch+NetVLAD (orange).

2 Visualizing Distance Margins

In Figure 2, we show how the latent space distance margins are affected by the proposed transformation of NetVLAD descriptors via SeqMatchNet (in orange) as compared to when no transformation is performed and vanilla NetVLAD descriptors are used (Vanilla SeqMatch in blue). Both the single image distances (thin opaque lines) and sequence matching distances (semi-transparent thick lines) are plotted against the physical distance (frame-wise or metric) between a query and reference images, selected relative to the ground truth reference image in either direction of the trajectory. The distance values are averaged over 20 uniformly sampled query images. The latent space distance margin can be observed as the difference between the trough and the peak values of a curve. For both the datasets, distance margins are relatively higher when using SeqMatchNet as compared to the vanilla system. Here, we use the Brisbane-trained model on the Oxford test set and the Nordland's train set model on the Nordland's test set. The increase in margin is apparent for both the single image distances and the sequence matching distances.

3 Qualitative Matches

Figure 3 and 4 show qualitative matches for the Brisbane and the Oxford dataset respectively. Vanilla SeqMatch is based on sequence matching applied directly to NetVLAD [2] whereas the proposed SeqMatchNet is based on the NetVLAD descriptors transformed using sequence-based loss and negative mining. For both the datasets, the first row represents a success case for both the methods, the second and third row show the cases where SeqMatchNet found the correct match but the vanilla method failed, and the last row shows a failure case for both the methods.

Query Ground Truth Vanilla SeqMatch SeqMatchNet

Figure 3: Qualitative matches for the *Brisbane dataset* comparing the proposed method and vanilla sequence matching applied on top of NetVLAD.

Figure 4: Qualitative matches for the *Oxford dataset* comparing the proposed method and vanilla sequence matching applied on top of NetVLAD.

4 Dataset and Training Details

4.1 Datasets

Nordland [3]: This dataset is comprised of 728 km rail journey repeated under four seasons. We use its Summer traverse as the reference database and Winter traverse as the query, and for both of them train/val/test splits are defined with geographical separation. We remove the image frames where the vehicle was stationary or passing through tunnels [4, 5]. Since this dataset is vastly different from the road/street imagery datasets described next, we only use it for an ablation study of loss and negative mining method types (see Section 5.1).

Oxford Robotcar [1]: This dataset comprises several 10 km traverses of Oxford captured over an year under substantially different appearance conditions. We use the left stereo images from the traverse ids: 2015-03-17-11-08-44 and 2014-12-16-18-44-24 which respectively form our daytime reference database and nighttime query.

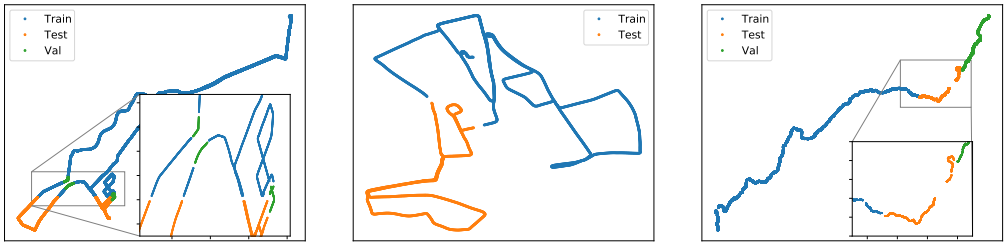
Brisbane City Loop [6]: This dataset comprises two 38 km traverses of Brisbane: City Loop Day and City Loop Night, which are used as reference and query respectively. We use Oxford Robotcar and Brisbane City Loop datasets for cross-testing generalization capabilities of models trained on each city, following the protocol of [7] from where the datasets were sourced. Since the Oxford traverse is much smaller than the Brisbane one, a common validation set is defined for both the methods using the Brisbane dataset, and train/test splits are defined for each city individually.

Mapillary Street Level Sequences (MSLS) [8]: This dataset comprises short image sequences of street-view imagery captured from a variety of camera configurations, several different cities and under a range of appearance conditions. For all the cities, reference and query databases are pre-defined [8]. We used Melbourne for training, Austin for validation and Amman for testing.

4.2 Training Data Splits

For each of the datasets, training, validation and test sets are defined without any geographical overlap, following [7]. Furthermore, the end points of these splits are ensured to be 40 meters apart to avoid visual overlap between the end point images of data splits. Figure 5 shows the data splits for the Brisbane, Oxford and Nordland dataset. Mapillary dataset is not displayed here as the images are sourced from *different cities* for training, validation and testing. For day-night place recognition via Brisbane and Oxford datasets, a common validation set is used from the Brisbane traverse. All the results in the main paper are presented on the test sets of the respective datasets using models

trained on the train split of the same or different cities. Table 1 shows the number of images within the reference and query databases of each of the splits of the four datasets.



(a) Brisbane (b) Oxford (c) Nordland

Figure 5: Data splits for Brisbane, Oxford and Nordland datasets.

Table 1: Data Splits: Reference (R) and Query (Q) Database Size

Split\Dataset	Oxford		Brisbane		Nordland		MSLS	
	R	Q	R	Q	R	Q	R	Q
Train	2981	2931	11994	13186	15000	15000	4973	4474
Val	-	-	477	582	3000	3000	4927	1732
Test	1574	1576	2858	2770	3000	3000	953	835

4.3 Training Parameters Settings

For the triplet loss, we used margin $m = 0.3$ and optimized using SGD with weight decay rate of 0.001 and momentum 0.9. We used PyTorch [9] for our experiments. Training is run for 50 epochs for all the datasets with an initial learning rate of 0.001 which is then reduced by a factor of 0.5 every 5 epochs. Following the original NetVLAD [2] training protocol, we use epoch-wise negative caching and optimize with the easiest positive and 10 hard negatives per anchor within a batch[1]. The distance matrix used for computing negatives is refreshed after processing every 1000 anchors (queries) during an epoch. For training, positives are defined within a localization radius of 5 meters and negatives are defined outside a radius of 20 meters for all the datasets except Nordland for which these values are set to be 10 and 40 frames respectively. Images near the boundaries of a trajectory, where L-length sequence cannot be centered, are not considered for testing.

5 Additional Experiments, Results and Visualizations

5.1 Deeper Network Configurations

We experimented with alternative network configurations including both an end-to-end training with image encoder and that based on pre-computed descriptors.

5.1.1 End-to-end training:

We used AlexNet as the image encoder with global max pooling to obtain a 256-dimensional single image descriptor. We found that with and without the proposed sequence-matching based loss and negative mining, Recall@1 dropped from 49.3 to 47.6 respectively on the Oxford Robotcar dataset. This demonstrates that the relative performance gains of the proposed method are still achievable even when using an end-to-end training with a deep architecture. However, the absolute performance is affected by the choice of architecture, pooling and the training dataset type.

[1]Note that hard positive mining [10], soft margins [11, 12] and other recent enhancements [13, 14, 15, 16, 17] proposed for single image based contrastive learning still remain complementary to sequence matching based contrastive learning.

Table 2: Deeper Network Configurations using Pre-computed Descriptors

FC 4K-4K	FC-FC 4K-2K-4K	FC-ReLU-FC 4K-2K-4K	FC-FC 4K-4K-4K	FC-ReLU-FC 4K-4K-4K	FC-FC-FC 4K-4K-4K-4K
0.78	0.69	0.46	0.74	0.66	0.55

Table 3: Cross-testing Single/Sequence Matching with/without Sequence-based Training

Oxford	Test Single	Test Sequence
No Training	0.47	0.67
Train Single	0.46	0.71
Train Sequence	0.46	0.71

Brisbane	Test Single	Test Sequence
No Training	0.20	0.21
Train Single	0.23	0.28
Train Sequence	0.24	0.29

5.1.2 Pre-computed Descriptors:

Table 2 shows Recall@1 on the Oxford Robotcar dataset when using different network configurations for pre-computed single image descriptors - in this case, PCA-transformed 4096-dimensional NetVLAD [2] vectors. FC refers to a fully-connected layer with bias and ReLU refers to the use of Rectified Linear Unit as non-linear activations as an intermittent layer. The size of the descriptors at the input and output of FC layers is shown in the second row. The first column represents the same configuration as that described in Equation 1 in the main paper. It can be observed that the use of additional layers with or without the use of non-linearity does not further improve performance. This can be attributed to two specific aspects of the input NetVLAD descriptors: i) PCA-transformation, due to which learning linear combination of features can potentially be sufficient, and ii) high dimensionality (4096), due to which using several layers on top significantly increases the network size (number of learnable parameters) affecting generalization capability since the training dataset size remains the same.

5.2 Cross-testing Single/Sequence Matching with/without Sequence-based Training

In Table 3, we investigate the effect of training with and without sequences on both single image and sequence-based matching, and compare it against the vanilla NetVLAD descriptors. In column 'Test Single', Recall@1 is reported for single descriptor matching and in column 'Test Sequence', sequence matching based results are presented. The row 'No Training' refers to the vanilla NetVLAD descriptors, and 'Train Single' and 'Train Sequence' respectively refer to training with and without the use of sequence-based metrics for computing loss and mining negatives. For both the datasets (tested using models trained on the other city), it can be observed that sequence matching based Recall@1 improvement is much higher for the trained systems as compared to the vanilla NetVLAD descriptors. Furthermore, using sequential information both during training and testing consistently leads to superior performance, which is closely followed by training on single and testing on sequence. Finally, for the Oxford dataset, it is interesting to note that recall of single descriptor testing reduces slightly due to training (0.46) as compared to the vanilla descriptors (0.47), even though using sequence matching on top of vanilla descriptors is significantly inferior to its trained counterparts.

References

[1] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 year, 1000 km: The oxford robotcar dataset. *IJ Robotics Res.*, 36(1):3–15, 2017.

[2] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] N. Sünderhauf, P. Neubert, and P. 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.

[4] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford. On the performance of convnet features for place recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 4297–4304. IEEE, 2015.

[5] S. Hausler, S. Garg, M. Xu, M. Milford, and T. 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.

[6] M. Milford, S. Garg, and J. Mount. P1-007: How automated vehicles will interact with road infrastructure now and in the future. Technical report, iMove, QUT and Queensland Government, January 2020. URL https://imoveaustralia.com/wp-content/uploads/2020/02/P1-007-Milestone-6-Final-Report-Second-Revision.pdf.

[7] S. Garg and M. J. Milford. Seqnet: Learning descriptors for sequence-based hierarchical place recognition. *IEEE Robotics and Automation Letters*, 2021.

[8] F. Warburg, S. Hauberg, M. López-Antequera, P. Gargallo, Y. Kuang, and J. 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.

[9] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[10] J. Thoma, D. P. Paudel, A. Chhatkuli, and L. Van Gool. Learning condition invariant features for retrieval-based localization from 1m images. *arXiv preprint arXiv:2008.12165*, 2020.

[11] J. Thoma, D. P. Paudel, and L. Van Gool. Soft contrastive learning for visual localization. *Advances in Neural Information Processing Systems 33*, 2020.

[12] M. Leyva-Vallina, N. Strisciuglio, and N. Petkov. Generalized contrastive optimization of siamese networks for place recognition. *arXiv preprint arXiv:2103.06638*, 2021.

[13] Q. Hu, X. Wang, W. Hu, and G.-J. Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1074–1083, June 2021.

[14] F. Wang and H. Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504, June 2021.

[15] J. Revaud, J. Almazán, R. S. Rezende, and C. R. d. Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5107–5116, 2019.

[16] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao. Spatial pyramid-enhanced netvlad with weighted triplet loss for place recognition. *IEEE Trans. Neural Netw. Learn. Syst.*, 31(2): 661–674, 2019.

[17] J. Thoma, D. P. Paudel, A. Chhatkuli, and L. Van Gool. Geometrically mappable image features. *IEEE Robotics and Automation Letters*, 5(2):2062–2069, 2020.