TransVPR: Transformer-Based Place Recognition with Multi-Level Attention Aggregation Supplementary Material

Ruotong Wang* Yanqing Shen* Weiliang Zuo Sanping Zhou Nanning Zheng†
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
{wrt072@stu., qing1159364090@stu., weiliang.zuo@, spzhou@, nnzheng@mail.}xjtu.edu.cn

This supplementary material provides the following additional information: Sec. 1 presents some additional results and analysis. Several visualization examples are also illustrated in this section. Sec. 2 provides some experimental details, including training settings, ablation settings, dataset utilization, and implementation of compared methods.

1. Additional Results

1.1. Quantitative Results

Efficiency-accuracy trade-off. In Fig. 1, we show the computational time and storage requirements of TransVPR compared with other methods as well as their performances on MSLS validation set. TransVPR achieves the best trade-off between accuracy and efficiency in both terms of latency and memory.

Fine-grained results on RobotCar Seasons v2 dataset. Tab. 1 shows the fine-grained comparison results on RobotCar-S2 test set split by specific appearance changing conditions[1]. We observe that all methods, especially the methods (*i.e.*, NetVLAD, SFRS, Patch-NetVLAD, and TransVPR) trained on MSLS or Pitts30k datasets, do not work well under night conditions. This is because the distribution of samples under day and night conditions in training datasets is severely unbalanced. In future work, it would be useful to address this challenge by exploring some techniques to improve the model generalization ability based on limited and unbalanced data.

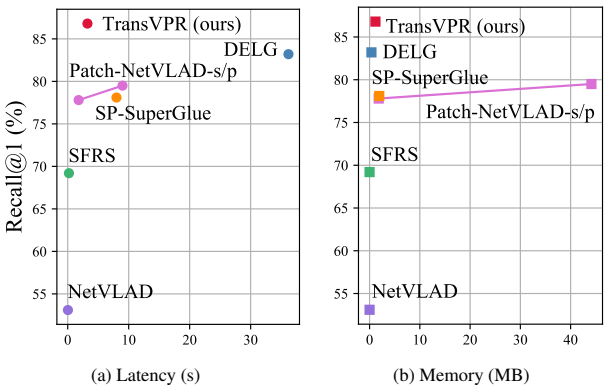


Figure 1. **Efficiency-accuracy trade-off.** The Recall@1 scores on MSLS validation set are shown on the y-axis, with (a) the accumulated time (*i.e.*, feature extraction and feature matching) and (b) the total memory cost to process one query image shown on the x-axis.

1.2. Qualitative Results

Effect of re-ranking. Examples of retrieved images before and after re-ranking by key-patch descriptors are illustrated in Fig. 8. In TransVPR model, the global image feature is a linear projection of the weighted summation of multi-level patch descriptors. Although it is robust to the changes in a scene, it carries no structural information of the scene and sometimes generates false positive results which roughly looks similar as the query image. Therefore, geometrical verification by key-patch descriptors is crucial to achieve high precision.

Visualization of multi-level attentions. Fig. 6 in the main paper illustrates some visualisation examples of multi-level attentions captured by TransVPR on MSLS validation set. Similarly, Fig. 9, Fig. 10, and Fig. 11 shows examples from Nordland, Pitts30k, and RobotCar-S2 datasets respectively. In Fig. 11 and under night conditions, we note that the mid-level attention masks also focus on some light reflection areas. This means that the model misidentifies

*Equal contribution.
†Corresponding author.
[1]We failed to reproduce the results reported in the original paper [6] of Patch-NetVLAD performance-focused configuration with the official implementation. We are trying to contact the author to solve this problem. In Tab. 2 of the main paper, we use the result given by the original paper. Here, we use the fine-grained results reproduced by our-selves since they are not provided by the author.

	day conditions							night conditions	
	dawn	dusk	OC-summer	OC-winter	rain	snow	sun	night	night-rain
m	.25 / .50 / 5.0	.25 / .50 / 5.0	.25 / .50 / 5.0	.25 / .50 / 5.0	.25 / .50 / 5.0	.25 / .50 / 5.0	.25 / .50 / 5.0	.25 / .50 / 5.0	.25 / .50 / 5.0
deg	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10
NetVLAD	11.5 / 30.0 / 78.0	4.6 / 24.4 / 92.4	7.1 / 28.4 / 87.7	1.2 / 22.0 / 97.6	11.7 / 42.9 / 100.0	11.6 / 32.1 / 95.3	4.9 / 15.2 / 72.8	0.4 / 0.9 / 2.7	1.0 / 2.0 / 12.8
SFRS	12.3 / 33.5 / 89.0	6.6 / 32.0 / 98.5	10.0 / 37.0 / 92.9	2.4 / 26.2 / 97.6	12.7 / 43.4 / 99.5	14.0 / 38.6 / 96.3	9.4 / 24.6 / 87.9	0.9 / 3.1 / 31.4	2.0 / 8.4 / 36.9
SP-SuperGlue	15.0 / 45.8 / 97.4	9.1 / 39.1 / 99.0	9.5 / 45.0 / 96.7	**3.0** / 31.7 / **100.0**	15.1 / **52.7** / **100.0**	14.0 / 46.5 / 97.7	12.1 / 33.5 / 94.2	3.1 / 8.8 / 32.7	3.0 / 15.8 / 56.7
DELG	4.0 / 10.6 / **97.8**	0.0 / 1.5 / 57.4	1.4 / 3.8 / 72.5	0.0 / 2.4 / 67.8	0.0 / 2.4 / 67.8	1.4 / 3.7 / 89.8	2.2 / 10.3 / **94.6**	**4.4 / 13.7 / 52.2**	**5.9 / 24.6 / 70.9**
Patch-NetVLAD-s	4.8 / 10.1 / 43.6	2.5 / 8.1 / 39.6	3.8 / 10.9 / 35.5	0.0 / 6.1 / 37.8	5.9 / 21.0 / 43.9	3.7 / 12.1 / 40.0	2.7 / 12.1 / 46.0	0.0 / 0.0 / 9.3	0.0 / 0.0 / 9.9
Patch-NetVLAD-p	5.3 / 12.8 / 44.1	3.0 / 9.6 / 39.1	4.7 / 10.4 / 32.7	0.0 / 6.7 / 37.2	4.9 / 19.0 / 43.9	4.7 / 13.5 / 39.5	3.6 / 12.9 / 46.0	0.4 / 3.1 / 16.8	1.0 / 1.5 / 10.3
TransVPR (ours)	**18.5** / **52.0** / 95.6	**10.7** / **44.7** / **100.0**	**12.3** / **45.5** / **99.1**	1.2 / **36.6** / 99.4	**15.1** / 50.7 / 99.5	**14.0** / **42.8** / **99.1**	**13.4** / **34.4** / 91.1	0.9 / 4.9 / 30.5	0.0 / 1.0 / 10.3

Table 1. Fine-grained performance comparison on RobotCar Seasons v2 dataset.

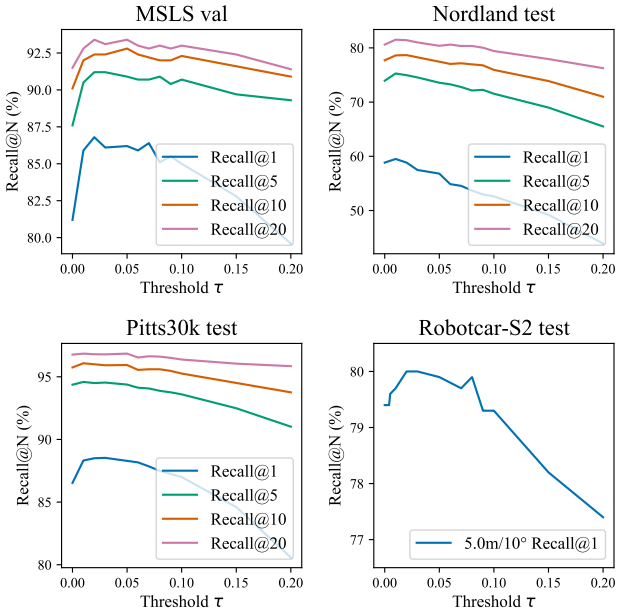


Figure 2. Sensitivity of TransVPR to key-patch filtering threshold.

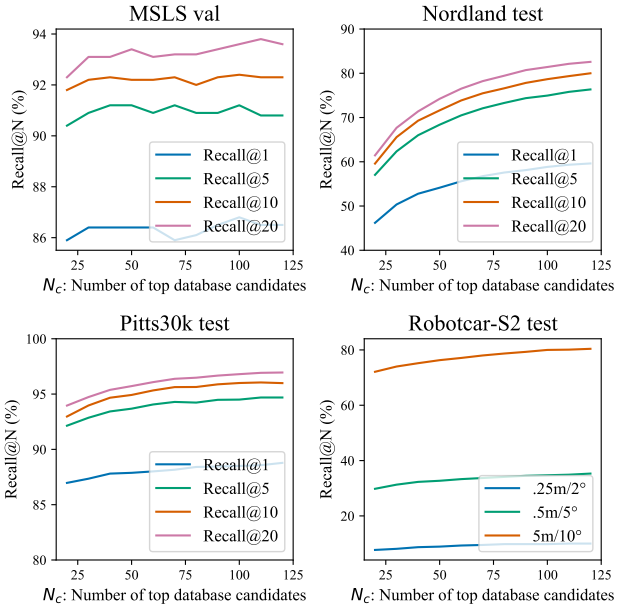


Figure 3. Sensitivity of TransVPR to the number of candidates to be re-ranked.

these areas as landmark objects and thus leads to poor performance at night. How can we identify and filter out these ares is a problem that needs to be solved in future work.

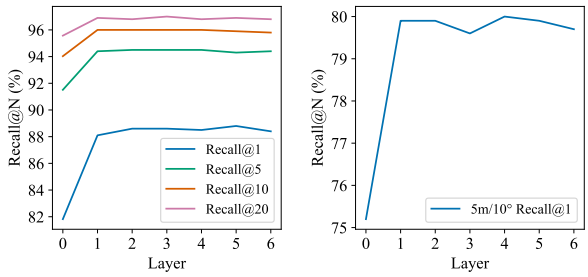
1.3. Further Ablations

Sensitivity to hyper-parameters. We further evaluate the sensitivity of our model to changes in two main hyper-parameters: the number of candidates to be re-ranked N_c and the key-patch filtering threshold τ. The results are shown in Fig. 2 and Fig. 3.

For all values of N_c, ranging from 20 to 120, TransVPR maintains state-of-the-art recall performance on MSLS validation set compared to all other VPR methods, and achieves competitive performance on other three datasets. This demonstrates the effectiveness of the TransVPR global representation which achieves a high recall rate before re-ranking. Note that the matching time of re-ranking top-20 candidates is only 1.02 seconds per query, 3.1 times faster than that of re-ranking top-100, with only 0.9% and 1.6% recall@1 degradation on MSLS and Pitts30k datasets. TransVPR provides an efficient solution in real-time VPR applications.

While changing the key-patch filtering threshold gradually, the performance of TransVPR reaches a peak at $\tau = 0.02$ ($\tau = 0.005$ for RobotCar-S2 dataset) and remains high until about $\tau = 0.08$. Larger values of τ result in too few key-patches reserved for matching. Note that using all local descriptors without filtering ($\tau = 0$) yields a degradation of performance, indicating the necessity of our proposed key-patch detection module.

Ablations on other datasets. In Section 5.4 of the main paper, we conduct ablations on MSLS validation set and Nordland test set to study the effect of the choice of patch



(a) Pitts30k test (b) RobotCar-S2 test

Figure 4. **Ablations of local descriptor set selection.** On Pitts30k test set and RobotCar-S test set, descriptors from any Transformer layer have similar performances and significantly outperform raw patch descriptors (*i.e.*, layer 0).

Attention mask	Pitts30k test			Robotcar-S2 test		
	R@1	R@5	R@10	.25m/2°	.5m/5°	5.0m/10°
None	86.5	94.4	95.8	9.9	35.0	79.4
$\mathbf{a_L}$	87.9	94.2	96.0	9.3	34.2	**80.1**
$\mathbf{a_M}$	87.2	93.8	95.3	9.3	33.2	79.3
$\mathbf{a_H}$	84.4	92.6	94.9	8.9	29.9	77.2
$\mathbf{a_L}$ & $\mathbf{a_M}$	88.4	94.4	96.0	9.6	32.7	80.0
A	**88.5**	**94.5**	**96.0**	**10.1**	**35.0**	79.6

Table 2. Performance of TransVPR on Pitts30k test set and RobotCar-S2 test set when using several combinations of attention masks from different Transformer levels to select key-patch descriptors. For RobotCar-S2 dataset, we set the key-patch filtering threshold τ to 0.005 which is the most optimized value.

descriptor sets and attention masks used for key-patch detection. Here, we show results of same experiments on Pitts30k test set and RobotCar-S2 test set in Fig. 4 and Tab. 2 respectively. We have the same observations as in the main paper: Patch descriptors with global context significantly outperform raw patch descriptors which have only local perception fields. Among all attention mask combinations, the fused multi-level attention mask **A** achieves the best performance.

2. Experimental Details

2.1. Training Details

TransVPR is optimized by AdamW optimizer [18] with 0.03 weight decay using cosine learning rate decay schedule.

The backbone network (the four-layer CNN and the six-layer Transformer) is pretrained on Places365 dataset [19] for 100 epochs with an initial learning rate of 0.0001. Images are scaled to 224×224 for faster training. The attention layers (W_i^a) and the dimensionality reduction layer (W_g) are initialized by training for 5 epochs on MSLS training set with 0.0003 initial learning rate, 384×384 input image size, and the margin m of 0.1. The pre-training time for

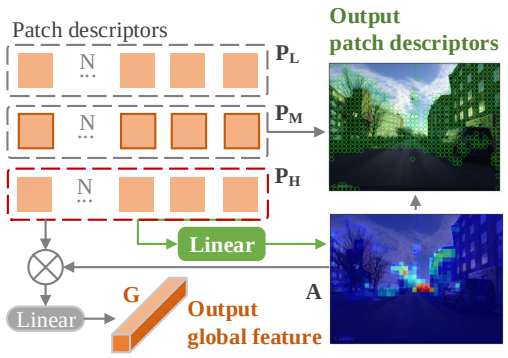


Figure 5. Single level & single attention map (sL-sATT).

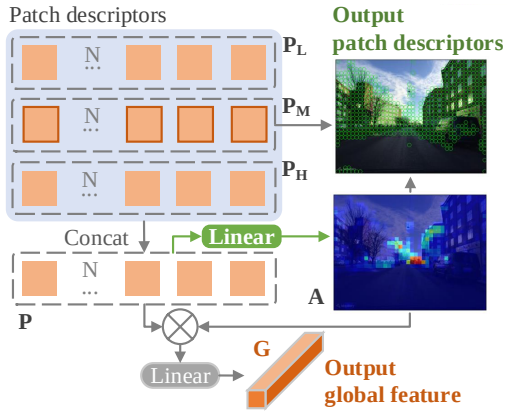


Figure 6. Multi-level & single attention map (mL-sATT).

TransVPR is about 4.5 days on 4 NVIDIA GeForce RTX 2080 Ti GPUs.

When fine-tuning on VPR datasets, we use initial learning rate 0.00001 and train for 15 epochs. Input image size is set to 384×384 for MSLS training set and 480×640 for pitts30k training set. It takes approximately 2.5 days on 4 GPUs to pre-train on MSLS dataset and 0.5 days on 2 on Pitts30k dataset.

2.2. Attention Aggregation Strategy

In section 5.4 of the main paper, we implement three degenerate models based on different attention aggregation strategies (*i.e.*, sL-sATT, mL-sATT, and mL-mATT-plain). Their architectures are illustrated in Fig. 5, Fig. 6, and Fig. 7 respectively.

2.3. Dataset Description and Utilization

Mapillary Street Level Sequences (MSLS) [17]. MSLS is a large-scale place recognition dataset which contains more than 1.6 million urban and suburban images from 30 cities across six continents. It covers various types of environmental changes, including weather, season, day and night, viewpoint, dynamic objects, and structural modifications. GPS coordinates and compass angles are provided

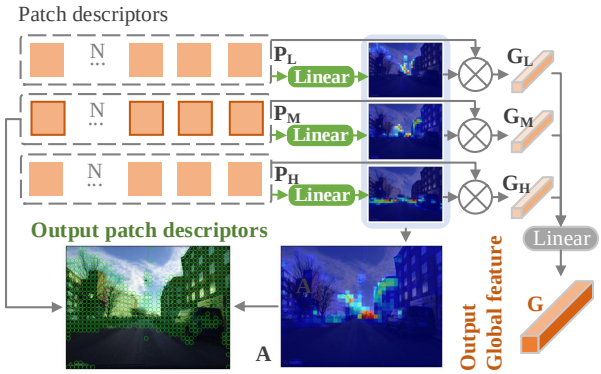


Figure 7. Multi-level & multiple attention maps & plain connection (mL-mATT-plain).

for each image. Reference images which locate within 25m and 40° from the query are considered as ground truths. The dataset is divided into a training set, a public validation set, and a withheld test set (MSLS challenge). When training, we define a mixed distance d_{ij} to represent the field of view overlap between two images i and j:

$$d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||_2/40 + (\theta_i - \theta_j)/25, \quad (1)$$

where $\mathbf{x}$ is GPS coordinate vector and θ is angle value. Positive samples are selected from image pairs with $d_{ij} < 1$ to ensure that there are overlapping regions between them.

Pittsburgh [16]. Pittsburgh dataset contains 250k street-level images derived from Google Street View panoramas. It has significant viewpoint variations and dynamic objects. Only GPS labels are available in Pittsburgh, and ground truths are defined as reference images within 25m from the query. Its subset Pitts30k, proposed by [1], containing 6k queries images and 10k database images in each of the training, validation and test sets is used in our experiments.

Nordland [15] [12]. Nordland dataset contains four aligned image sequences recorded during a 728 km long train journey in four seasons. It has seasonal changes but very few viewpoint changes. We used the dataset partition[2] presented by [12], which has a test set containing 3450 images per sequence. Ground truth tolerance is set to ± 2 frames away from the query. Following [2, 6, 7, 9], winter sequence is used as query set, while summer sequence is used as reference set.

RobotCar Seasons v2 [10, 14] RobotCar Seasons v2 is a subset of RobotCar dataset which captures 20k images in Oxford by cameras mounted on a car. It typically includes various weather and seasonal conditions during day and night with minor viewpoint changes. It is divided into a public training set and a withheld test set. Following the setting in [6], results in the main paper are computed by summarizing the results of different conditions weighted by the number of query images of each condition.

[2] https://webdiis.unizar.es/ jmfacil/pr-nordland/

2.4. Implementation Details of Baselines

NetVLAD [1]. A learnable VLAD layer is proposed in this method to aggregate local descriptors from CNN feature maps with learnable cluster centers. We use the pytorch implementation[3] and its released model trained on Pitts30k training set with VGG-16 backbone.

SFRS [5]. This work is based on NetVLAD and improves its performance by training under self-enhanced and fine-grained supervision. We use the official implementation[4] and the released model state trained on Pitts30k training set.

SP-SuperGlue [4, 13]. This patch-level feature matching approach is used to re-rank candidates retrieved by NetVLAD. We use the official implementation[5] and the released outdoor model trained on MegaDepth dataset [8].

DELG [3]. This is a unified CNN model which extracts both global and patch-level image features. We use the official implementation[6] and the released model trained on Google Landmarks Dataset v2 dataset [11] with ResNet-50 backbone.

Patch-NetVLAD [6]. This method extracts patch-level features based on NetVLAD residuals and uses them to re-rank NetVLAD retrieved candidates. We use the official implementation[7]. Both speed-focused and performance-focused configurations are evaluated. Following the original paper, the released model trained on Pitts30k training set is evaluated on Pitts30k test set, while the one trained on MSLS training set is evaluated on all other datasets.

[3] https://github.com/Nanne/pytorch-NetVlad
[4] https://github.com/yxgeee/OpenIBL
[5] https://github.com/magicleap/SuperGluePretrainedNetwork
[6] https://github.com/tensorflow/models/tree/master/research/delf
[7] https://github.com/QVPR/Patch-NetVLAD

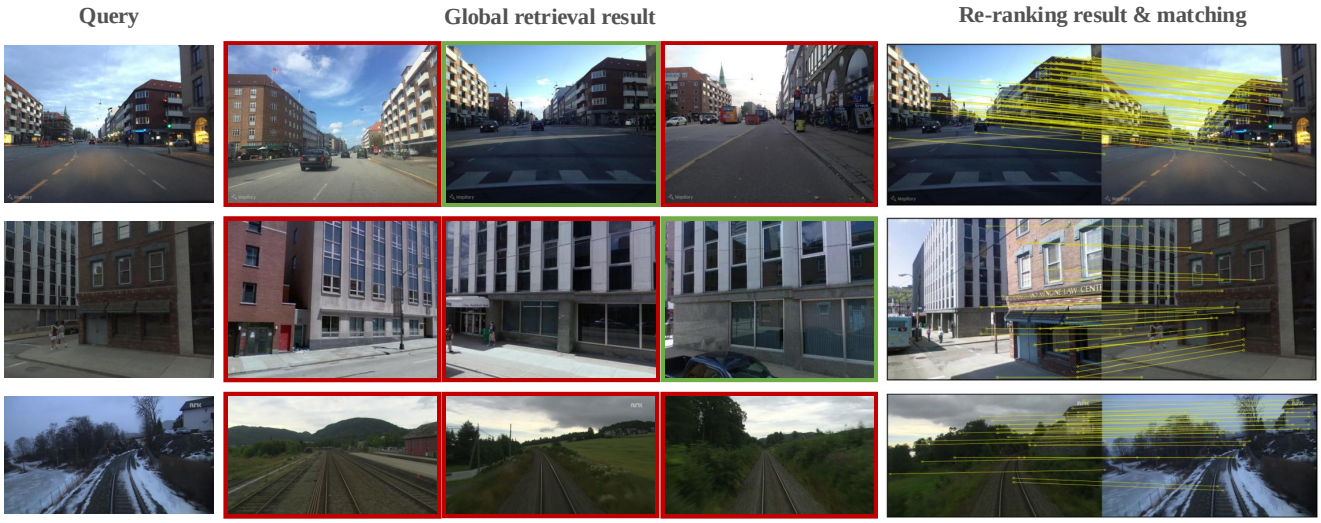


Figure 8. Examples of global retrieval and re-ranking results by TransVPR on MSLS val (top), Nordland (middle) and Pitts30k (bottom) datasets. In these cases, re-ranking results are all correct, and global retrieval results framed in green are correct while those in red are incorrect.

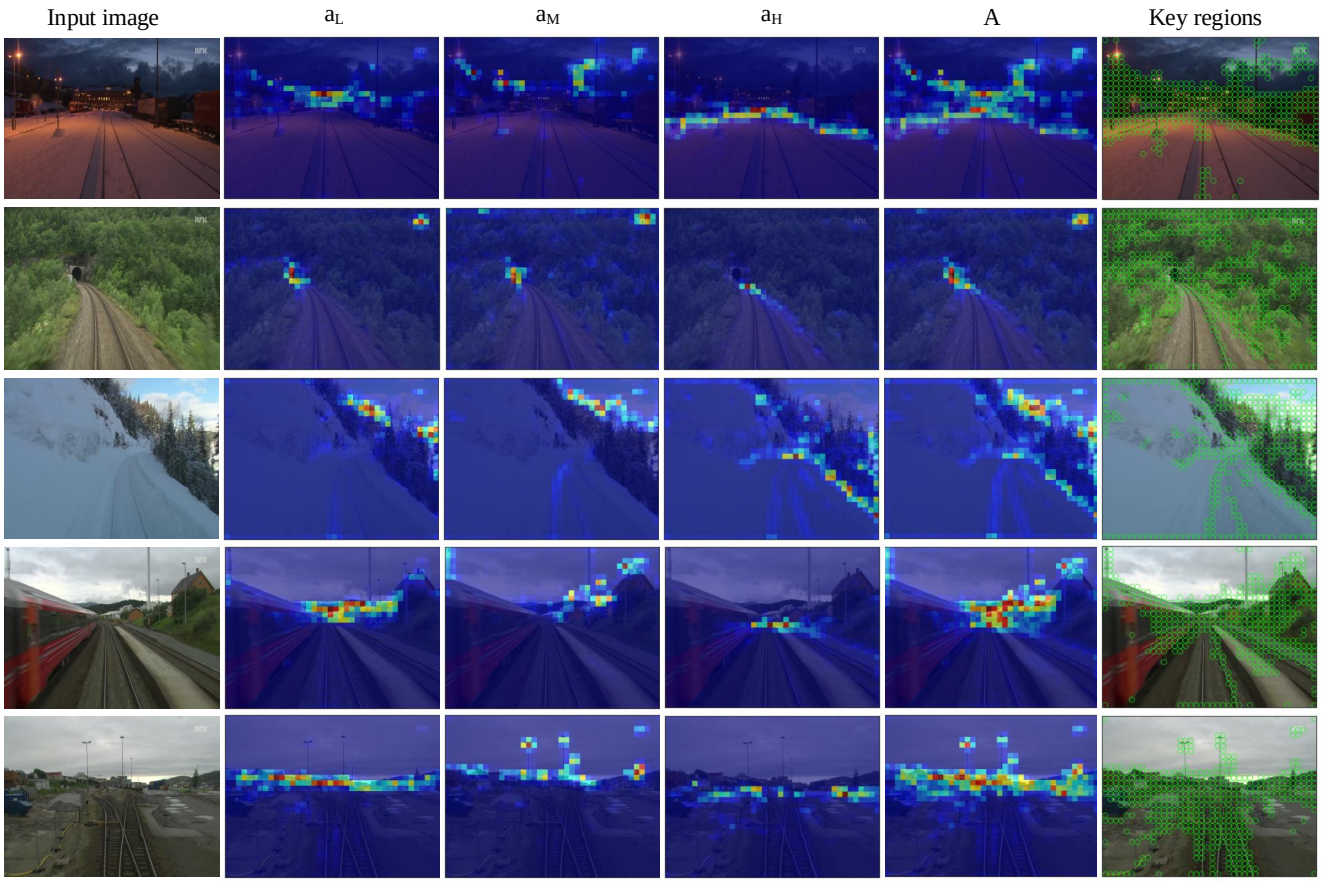


Figure 9. Attention visualization examples on Nordland dataset.

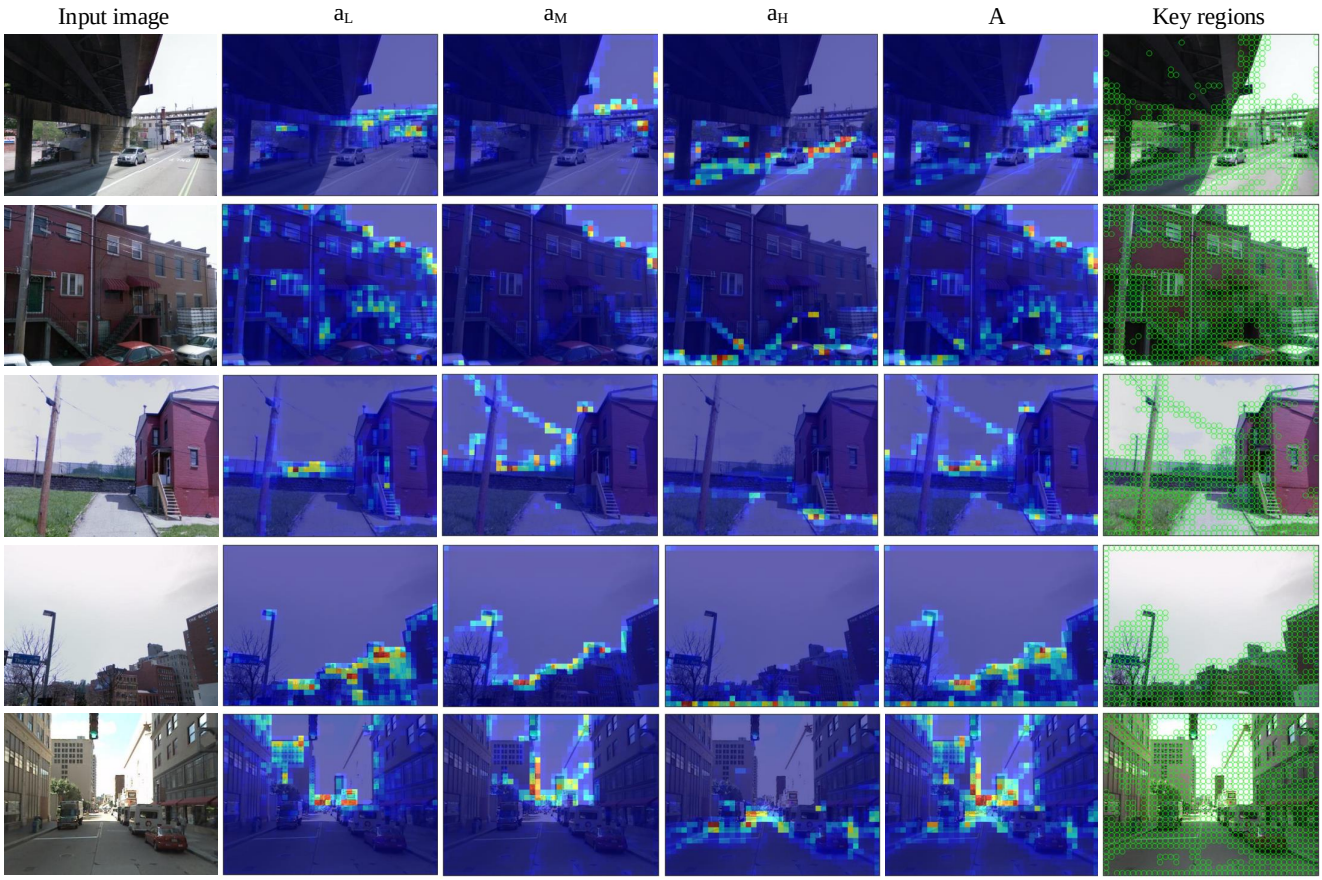


Figure 10. Attention visualization examples on Pitts30k dataset.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, pages 5297–5307, 2016. 4

[2] Luis G Camara and Libor Přeučil. Visual place recognition by spatial matching of high-level cnn features. *Robot. Autom. Syst.*, 133:103625, 2020. 4

[3] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *ECCV*, pages 726–743, 2020. 4

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR*, pages 224–236, 2018. 4

[5] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *ECCV*, pages 369–386, 2020. 4

[6] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *CVPR*, pages 14141–14152, 2021. 1, 4

[7] Stephen Hausler and Michael Milford. Hierarchical multi-process fusion for visual place recognition. In *IEEE Int. Conf. Robot. Autom.*, pages 3327–3333, 2020. 4

[8] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018. 4

[9] Feng Lu, Baifan Chen, Xiang-Dong Zhou, and Dezhen Song. Sta-vpr: Spatio-temporal alignment for visual place recognition. *IEEE Robot. Autom. lett.*, 6(3):4297–4304, 2021. 4

[10] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *Int. J. Rob. Res.*, 36(1):3–15, 2017. 4

[11] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, pages 3456–3465, 2017. 4

[12] Daniel Olid, José M. Fácil, and Javier Civera. Single-view place recognition under seasonal changes. In *PPNIV Workshop at IROS 2018*, 2018. 4

[13] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, pages 4938–4947, 2020. 4

[14] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi

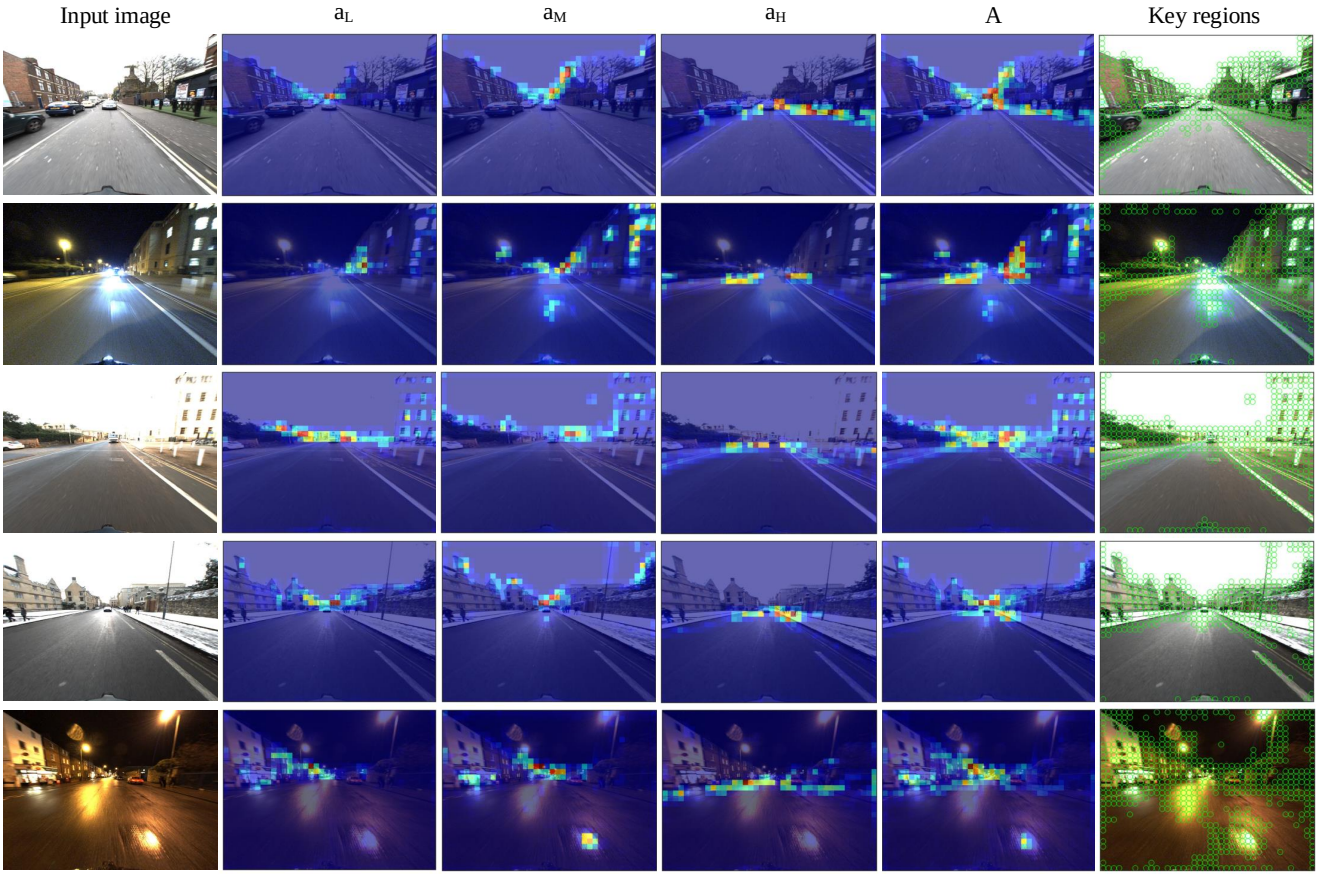


Figure 11. Attention visualization examples on RobotCar-S2 dataset.

Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *CVPR*, 2018. 4

[15] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. In *IEEE Int. Conf. Robot. Autom. Worksh.*, 2013. 4

[16] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *CVPR*, pages 883–890, 2013. 4

[17] 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 *CVPR*, pages 2626–2635, 2020. 3

[18] Hui Zhong, Zaiyi Chen, Chuan Qin, Zai Huang, Vincent W Zheng, Tong Xu, and Enhong Chen. Adam revisited: A weighted past gradients perspective. *Frontiers of Computer Science*, 14(5):1–16, 2020. 3

[19] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 2017. 3