Attentional Pyramid Pooling of Salient Visual Residuals for Place Recognition (Supplementary Material)

Guohao Peng[1], Jun Zhang[1], Heshan Li[2], Danwei Wang[2]
Nanyang Technological University, Singapore
[1]{peng0086,jzhang061}@e.ntu.edu.sg, [2]{heshan.li,edwwang}@ntu.edu.sg

In this supplementary material, we provide some additional details regarding the proposed method.

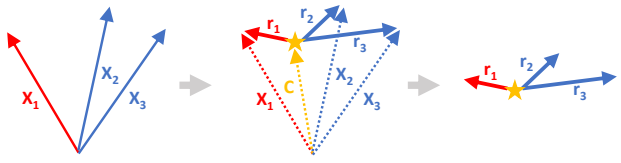
1. Underlying Intuitions of APPSVR

1.1. Characterize the Nuances of Local Details

Fig.1 gives two examples from the Tokyo24/7 dataset, which shows that different scenes in a city-scale environment can be composed of similar entities (*e.g.* buildings, roads, vegetation) with a consistent spatial layout. This requires the encoded descriptors should be able to characterize local details and distinguish subtle differences (*e.g.* architectural colors, textures and symbols).

Figure 1. With similar entities and spatial layout, two images from different scenes may appear to be the same place.

As illustrated in Fig.2, by introducing cluster centroids that represent the common characteristics of local features, the residuals ($\mathbf{r}_i$) between local features ($\mathbf{x}_i$) and their corresponding cluster centroid ($\mathbf{c}$) can better characterize the local distinctiveness.



(a) Local features (b) Introducing a centroid (c) Local residuals

Figure 2. Local residual is more discriminative than the original feature in terms of feature association.

Therefore, we follow the basic idea of embedding cluster-wise residuals instead of directly embedding local features. The superiority has been verified and discussed in Table.3 and Section.6.5 of the main text, where APPSVR and other VLAD-centric methods show overall advantages over the global pooling methods.

1.2. Distinguish the Significance of Visual Cues

Besides the changes in appearance due to illumination, weather, or seasons, another common problem encountered in city-scale VPR is dynamic occlusion caused by vehicles or pedestrians. Taking Fig.1 as an example, dynamic objects seem to help distinguish the two images to some extend. To avoid learning this harmful trick, a robust VPR model must be able to adaptively suppress dynamic visual cues while emphasizing long-term static structures.

Our APPSVR integrates triple attention into the encoding strategy, where local refinement can suppress misleading visual cues and global integration can distinguish the saliency of retained visual cues in the final image representation. These two internal driving forces enable APPSVR to learn the inference habit of highlighting the long-term static in a data-driven manner. Fig.3 below illustrates the process of APPSVR generating the overall attention of an image, and shows the subordinate relationship between triple attention and functional steps.

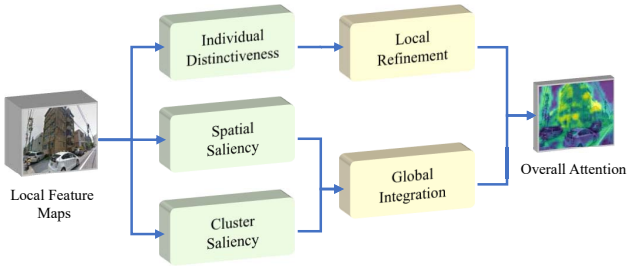


Figure 3. The flowchart of attention inferring by APPSVR.

Fig.4 shows the inferred attention of the two images in Fig.1 by APPSVR. As can be seen, APPSVR adaptively

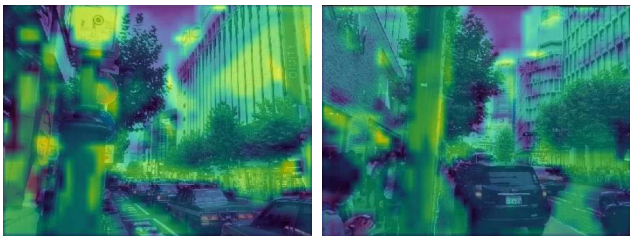


Figure 4. The inferred attention of the two images in Fig.1 by APPSVR. Architectural and traffic signs are most highlighted while vehicles and pedestrians are ignored. Those highlighted unique visual cues allow APPSVR to easily distinguish visually similar images from different places.

learns to focus on architectural and traffic signs, which are also discriminative in human cognition. Comparing the two saliency maps, one can easily judge that APPSVR is able to distinguish visually similar images by local details.

2. More Details on Datasets

As described in Section.6.1 of the main text, Pitts30k [1], Pitts250k [9], and Tokyo 24/7 [1, 8] are employed as the benchmark datasets for evaluation. These three datasets are all collected from Google Street View Time Machine and provide perspective images generated from street-level panoramics. The only ground-truth annotation is GPS-tag, which is used to identify the positive and negative correspondences for each query image. The statistics of the employed datasets can refer to Table 1.

Table 1. Description of Employed Datasets

Dataset	Split	Database	Query set
Pittsburgh	Pitts250k-train	91,464	7,824
	Pitts250k-val	78,648	7,608
	Pitts250k-test	83,952	8,280
	Pitts30k-train	10,000	7,416
	Pitts30k-val	10,000	7,608
	Pitts30k-test	10,000	6,816
TokyoTM	TokyoTM-train	49,104	7,277
	TokyoTM-val	49,056	7,186
	Tokyo24/7(test)	75,984	315

3. Unified Training And Evaluation Protocol

We notice that the recent SOTA methods [1, 3, 10, 12] for VPR use different learning frameworks (MatConvNet and Pytorch) and training protocols, which may cause deviations in the performance of a model. In our experiments, we require all comparisons to be done fairly so that the pure advantages of different architectures can be highlighted.

Therefore, we re-implement all comparative models in Pytroch, including NetVLAD [1], CRN [3], R-MAC [7], GeM [6], GhostVLAD [11], APANet [12], and SPENetVLAD [10]. Among different training and evaluation protocols, we follow the latest SOTA [10] to employ Pitts30k-train as the only training set, and test all trained models on Pitts250k-test, Pitts30k-test, and Tokyo24/7 respectively. Note that Tokyo24/7 is more challenging in terms of larger appearance changes and more dynamic objects. Training on Pitts30k but testing on Tokyo24/7 put forward higher requirements for the generalization ability of the trained models. We believe that such a rigorous evaluation protocol can better reflect the practicality of an architecture.

Table 2. Performance on Pittsburgh250k reported by our re-implemented models ($\star$) and original paper ($\circ$).

Method	Mode	Pitts250k-test		
		r@1	r@5	r@10
NetVLAD [1]	$\circ$	86.0	93.2	95.1
	$\star$	**86.5**	**93.8**	**95.5**
CRN [3]	$\circ$	85.5	93.5	95.5
	$\star$	**87.2**	**94.2**	**95.9**

As shown in Table.2 of the main text, despite the more rigorous protocol, APPSVR still achieves new SOTA results on both Pitts250k and Tokyo24/7 datasets. It demonstrates the excellent robustness and generalization ability of our proposed model. Besides, compared to benchmarks, the performance gain of APPSVR is larger on Tokyo24/7 than that on Pittsburgh, which shows our advantages are more pronounced in more challenging scenarios.

4. Additional Implementation Details

Centroid generation. As mentioned in Section.3.1 (Semantic constrained initialization) of the main text, clustering is performed to generate the centroids for model initialization. Specifically, 1k images are sampled from the training set, and forwarded to the backbone model for local feature extraction. From the activation maps of each image, 200 deep local features are randomly sampled and cached. Eventually, there are a total number of 1k*200=200k deep local features sampled for the subsequent clustering.

To generate the K cluster centroids $\{\mathbf{c}_k\}$, k-means clustering is first performed on the filtered features with the static semantics. The informative centroid of each cluster is initialized to be the same as the cluster centroid. Then for each dynamic or task-irrelevant semantic, features with this label are clustered to generate M candidate centroids, which are added to the shadow candidate list. During initialization, the L shadow centroids $\{\mathbf{c}^s_{kl}\}$ of the k^{th} cluster are initialized by the top L candidates with the shortest distances from the cluster centroid $\mathbf{c}_k$.

Traning tuple mining. Since GPS-tag is the only ground-truth annotation provided by the training set, we use it and the distance thresholds to identify the positive and negative samples for each query image.

To create the training tuple $(X_q, X_r^{p*}, \{X_r^n\})$ for a query X_q, we first mine its potential positives (images within 10 meters) and negatives (images further away than 225 meters) based on the GPS tags. Among the potential positives $\{X_r^p\}$, the one that has the smallest representation distance to the query X_q is chosen as the positive X_r^{p*} of the tuple:

$$X_r^{p*} = \underset{X_r \in \{X_r^p\}}{\arg\min} \|\tilde{f}(X_q) - \tilde{f}(X_r)\|. \quad (1)$$

$\tilde{f}(X)$ is the cached representation of an image X, which is recomputed every 1000 training queries.

To choose the N negatives $\{X_r^n\}$ of the tuple, we adopt both random and hard mining strategies. The half of $\{X_r^n\}$ is selected as the hardest negatives (with the most similar representation as X_q) from a pool of 1000 randomly sampled negatives, and the other half is randomly selected from the cached hard negative list of the query. The cached hard negative list contains the hard negatives from previous epochs that violated the triplet criterion in Eq.(2).

$$\|\tilde{f}(X_q), \tilde{f}(X_r^{p*})\| - \|\tilde{f}(X_q), \tilde{f}(X_r^n)\| + m \leq 0 \quad (2)$$

5. More Experimental Results

To assess the generalization ability of our method, we also conduct experiments on datasets for different tasks. San Francisco landmark dataset [2] is used for landmark identification. It contains a database of 1.2M architectural images, and a difficult query set of 803 images taken with different camera phones. The ground-truth annotations for correct matches are given in the benchmark. $\mathcal{R}$Oxford5k [5] and $\mathcal{R}$Paris6k [5] datasets are used for image retrieval. They both contain a query set of 70 images, and their reference set consists of 5k and 6k images respectively.

Landmark identification. We first compare the different methods on San Francisco for landmark identification. All evaluated models are pre-trained on Pitts30k and generate image representations of two sizes (original dimension and 4096D). As presented in Table 3, our model steadily outperforms benchmark methods in both representations. An improvement of about 4% has been achieved in Recall@1 index compared with NetVLAD. Combining the results in the main text, APPSVR shows good generalization ability on both TokyoTM and San Francisco. It demonstrates that the incorporated triple attention can greatly improve the adaptability of our image representation to different data domains. This endows APPSVR with strong practicability, especially in practical applications where a large amount of data for fine-tuning is not always available.

Image retrieval. Further comparisons are performed on the image retrieval benchmarks $\mathcal{R}$Oxford5k [5] and $\mathcal{R}$Paris6k [5]. Same as in [5], all compared models are fine-tuned on street-view images (Pitts30k) and evaluated by mean average precision (mAP). Fig.5 shows the performance of different models in the evaluation of three difficulty levels. As can be seen, our representation achieves the best mAP on $\mathcal{R}$Oxford5k and $\mathcal{R}$Paris6k in all cases. Compared with NetVLAD, APPSVR achieves an improvement of about 4% on both benchmark datasets.

Table 3. Evaluate the generalization ability of different models on San Francisco landmark dataset.

Method	SanFrancisco (Orgdim)			SanFrancisco (4096-D)		
	r@1	r@5	r@10	r@1	r@5	r@10
NetVLAD [1]	68.7	75.6	78.6	79.2	85.1	86.8
GhostVLAD [11]	67.6	76.7	79.2	80.3	86.6	87.5
CRN [3]	69.1	78.6	81.0	81.8	86.9	89.0
SRALNet [4]	70.4	78.8	81.8	79.0	85.6	87.9
APPSVR-SC-PN-L3	**72.7**	**80.2**	**83.0**	**82.3**	**87.1**	**89.1**

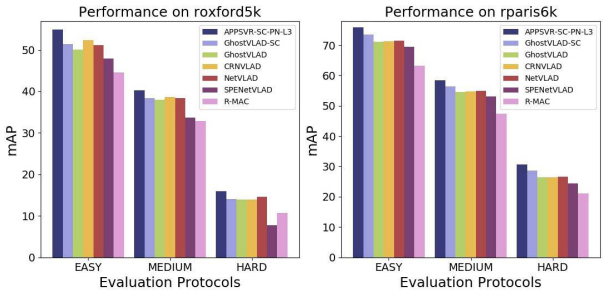


Figure 5. The performance comparisons of image retrieval on $\mathcal{R}$Oxford5k and $\mathcal{R}$Paris6k datasets. On both datasets, our best model (APPSVR-SC-PN-L3) outperforms the baseline methods at all three difficulty levels.

6. Additional Visualization Results

Fig.6 presents some challenging queries in Tokyo24/7 dataset and the top retrieved images using different models. As can be seen, when other benchmark models fail, APPSVR can still retrieve the queries correctly. Besides, the inferred attention on query and reference images shows that APPSVR can steadily focus on long-term static structures without being affected by appearance or viewpoint changes. This demonstrates the strong robustness of our proposed model.

In addition, Fig.7 presents the top five retrieved images by APPSVR on other groups of Tokyo24/7 queries. Each group contains three queries from daytime, dusk and night. Our model can smoothly deal with the illumination changes among them. Besides, the first two groups show that APPSVR can handle the billboard changes over time, which can be attributed to its greater emphasis on structural information during feature embedding. The remaining examples also demonstrate the robustness of APPSVR to dynamic objects and viewpoint variations.

Overall, APPSVR is able to cope with the practical challenges that may be encountered in city-scale visual place recognition.

Figure 6. Example retrieval results on Tokyo24/7 dataset. From left to right: query image, our inferred query attention, the top retrieved image using our method (APPSVR), our inferred reference attention, the top retrieved image using NetVLAD [1], the top retrieved image using CRN [3], and the top retrieved image using SPENetVLAD [10]. Green and red borders indicate correct and incorrect retrieved results respectively. As can be seen, when other benchmark methods fail, our APPSVR can still retrieve the queries correctly. The inferred attention shows that APPSVR can steadily focus on long-term static structures without being affected by appearance and viewpoint changes.

References

[1] Relja Arandjelovic, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, 2016.

[2] David M. Chen, Georges Baatz, Kevin Köser, Sam S. Tsai, Ramakrishna Vedantham, Timo Pylvänäinen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, Bernd Girod, and Radek Grzeszczuk. City-scale landmark identification on mobile devices. *CVPR 2011*, pages 737–744, 2011.

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

[4] G Peng, Y Yue, J Zhang, Z Wu, X Tang, and D Wang. Semantic reinforced attention learning for visual place recognition. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[5] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Revisiting oxford and paris: Large-scale image retrieval

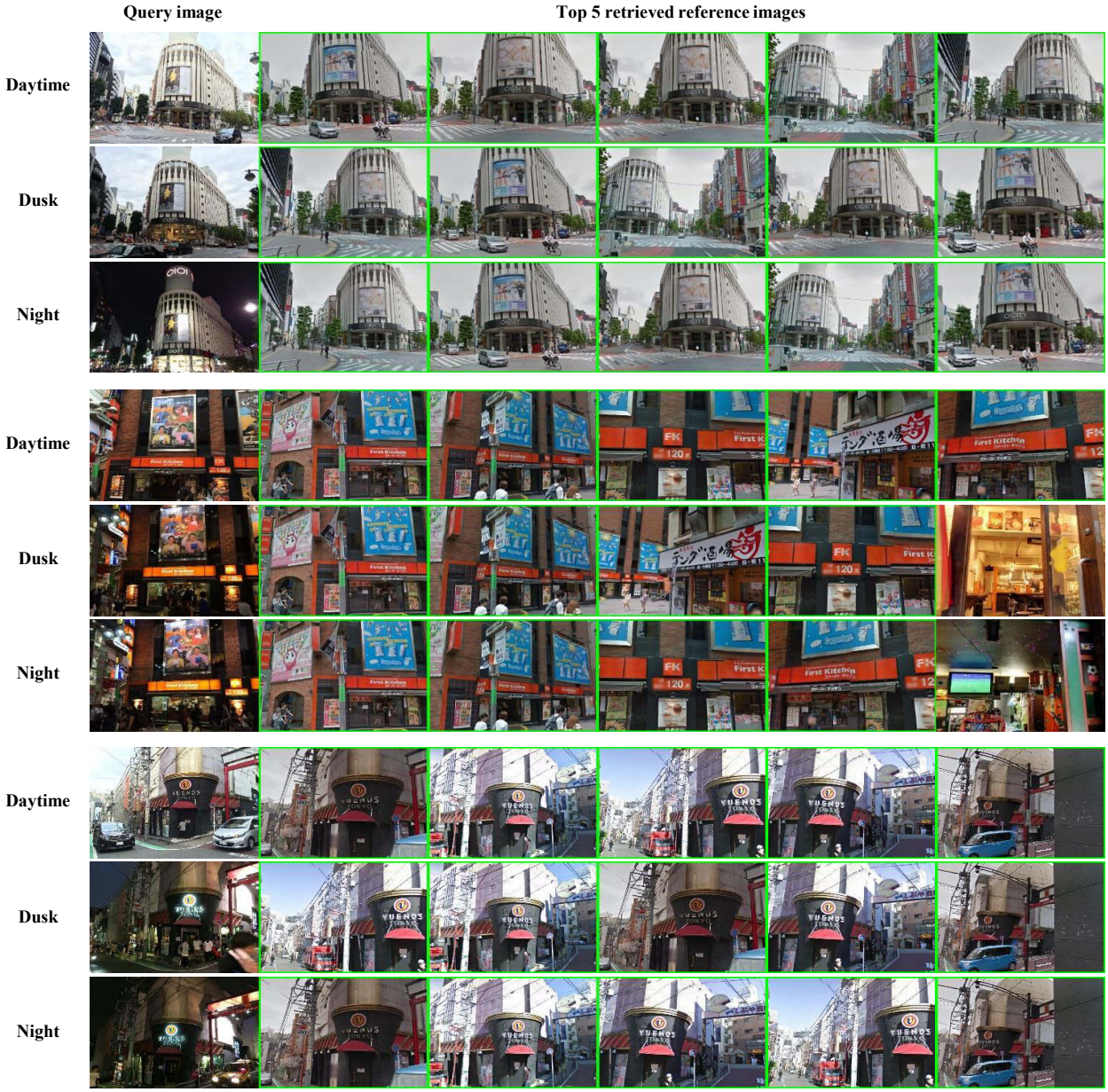


Figure 7. Top 5 retrieved images for a few groups of Tokyo24/7 queries by APPSVR. Each group contains three queries taken from three periods of the day, to validate the robustness of the model against illumination changes. Besides, the first two groups show that APPSVR can handle billboard changes due to time spans. These examples also demonstrate that our model is robust to dynamic objects, occlusions, and viewpoint variations.

benchmarking. In *CVPR*, 2018.

[6] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2019.

[7] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *CoRR*, abs/1511.05879, 2015.

[8] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, 2015.

[9] Akihiko Torii, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. Visual place recognition with repetitive structures. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 883–890, 2013.

[10] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao. Spatial pyramid-enhanced netvlad with weighted triplet loss for place recognition. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2019.

[11] Yujie Zhong, Relja Arandjelovic, and Andrew Zisserman.

Ghostvlad for set-based face recognition. In *ACCV*, 2018.

[12] Yingying Zhu, Jiong Wang, Lingxi Xie, and Liang Zheng. Attention-based pyramid aggregation network for visual place recognition. In *ACM Multimedia*, 2018.