DIGGING HIERARCHICAL INFORMATION FOR VISUAL PLACE RECOGNITION WITH WEIGHTING SIMILARITY METRIC

Hong Liu, Qian Zhang, Guoliang Hua, Chenyang Zhao

Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School

ABSTRACT

Visual place recognition is a challenging task due to the appearance of a place varying with the change of illumination, seasonal variations, and diverse viewpoints. Although significant progress has been made recently, how to dig sufficient hierarchical information in the real scenario for visual place recognition remains a problem. To this end, a hierarchical feature extraction module (HFM), as well as a weighting similarity metric module (WSM), is proposed in this paper. Specifically, the context-aware feature extraction block in HFM is designed to exploit multi-scale features containing contextual information. The additional complementary features extracted from the shallow layer are refined by a recalibration block for reserving detailed information. Furthermore, the WSM, which consists of a part-based similarity metric layer and a weighting layer, can make the best of the hierarchical information to calculate the similarity scores. Experiments conducted on three typical benchmarks show that our method achieves state-of-the-art performance on visual place recognition.

Index Terms— Visual Place Recognition, Hierarchical Feature Extraction, Similarity Metric, Feature Embedding

1. INTRODUCTION

Visual place recognition (VPR), which aims to match a query place with previously visited places, has received considerable attention for its applications in autonomous navigation [1, 2], mobile robotics [3, 4], and augmented reality [5, 6]. Traditional methods usually rely on hand-crafted descriptors [7, 8] to represent the image and achieve the place recognition through feature matching [9, 10]. With success in deep learning [11, 12], recent researchers focus on designing Convolutional Neural Networks (CNNs) to release the difficulty of extracting high-level features for place recognition. Despite significant progress, there remains two inherent limitations of extracting features by standard CNNs to represent the scene image. First, convolutional filters only integrate the information of their receptive fields and fail to capture spatial correspondence. Second, with deeper networks are designed, the size of feature maps becomes smaller, resulting in loss of detailed information. As a result, existing CNN-based methods lack full use of the visual cues in the scenario, thus causing a bottleneck to further improve the performance of VPR.

Relation to prior work: Gomez-Ojeda et al. [13] trained a CNN to extract features for representing the image and then achieved place recognition by comparing the features of different images. To select features of effective regions, Kim et al. [14] proposed a Contextual Reweighting Network to generate a weighted mask. Meanwhile, Chen et al. [15] designed a context-flexible attention learning model to generate attention maps for adaptively selecting relevant features. Feature pooling operations were performed on different scales to aggregate regional features [16], which can suppress the local confusions. Moreover, Liu et al. [17] proposed a two-branch siamese network to reduce the redundant features and enhance feature representation capability.

In contrast to the methods mentioned above, we engage in addressing the challenges in VPR by emphasizing the significance of contextual information and detailed information. The contextual information helps to leverage spatial correspondence, and additional detailed information improves the integrality of feature embedding. Both two types of information can provide powerful visual cues for VPR. Considering this nature, a hierarchical feature extraction module (HFM), associated with a weighting similarity metric module (WSM), is proposed in this paper. Specifically, in HFM, a context-aware feature extraction block is designed to extract multi-scale features containing contextual information, and a recalibration block is introduced to refine the features from the shallow layer, which reserve more detailed information. Meanwhile, to make the best of hierarchical information, WSM is introduced to measure the similarity score between features in the same hierarchy and efficiently fuse the similarity scores to the final similarity decision by learning the fusion weights.

The main contributions of this paper are as follows: (1) A hierarchical feature extraction module is proposed to extract both high-level contextual features and refined detailed features. (2) A weighting similarity metric module is introduced to aggregate hierarchical information into a final similarity decision.

This work is supported by National Natural Science Foundation of China (U1613209, 61673030), National Natural Science Foundation of Shenzhen (JCYJ20190808182209321).

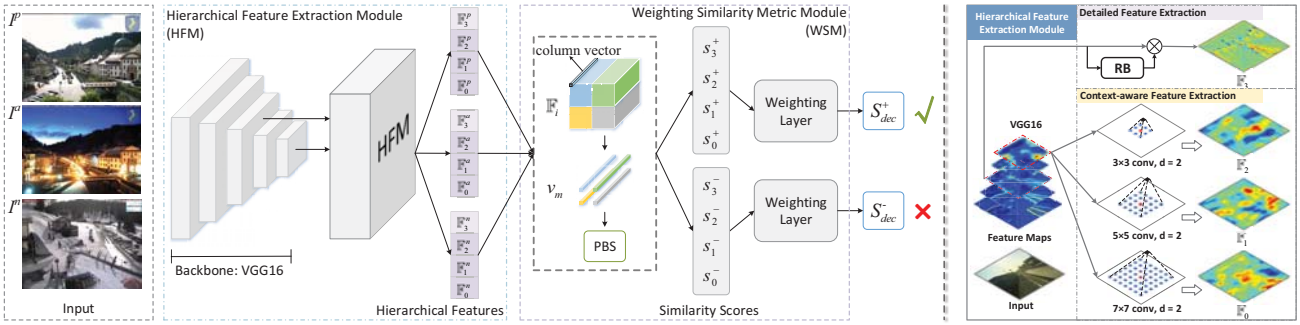

Fig. 1. **(a)** Overview of the proposed framework. VGG16 [18] is adopted as the backbone network. The HFM takes outputs of *Conv 3* and *Conv 5* as inputs to extract hierarchical features. Then, the WSM calculates the similarity scores of corresponding features and fuses them into a final similarity decision. PBS denotes the part-based similarity metric. **(b)** Details of HFM. RB represents the recalibration block.

2. PROPOSED METHOD

In this section, the overall framework is given at first. Then, the proposed hierarchical feature extraction module (HFM) and weighting similarity metric module (WSM) are introduced respectively. Finally, the loss function is presented.

2.1. Framework

The overall framework is shown in Fig. 1(a). The input of our framework is an image tuple $\langle I^a, I^p, I^n \rangle$, where I^a is an anchor image, I^p represents a positive sample image from the same place, and I^n denotes a negative sample image from a different place. VGG16 [18] is adopted as the backbone network, except that the last max pooling layer is removed. Following the backbone network, a hierarchical feature extraction module (Sec. 2.2) is introduced to generate hierarchical feature embedding $\{\mathbb{F}_i | i \in [0,3]\}$, where $\mathbb{F}_i$ means features in the hierarchy i. In particular, when $i = 3$, the features come from the detailed feature extraction process. Otherwise, they come from the context-aware feature extraction process, as shown in Fig. 1(b). After that, these features are sent to the weighting similarity metric module to calculate the similarities between image pairs (Sec. 2.3). The whole network is optimized by triplet loss in an end-to-end manner (Sec. 2.4).

2.2. Hierarchical Feature Extraction

The hierarchical feature extraction module, which contains two parts: context-aware feature extraction and detailed feature extraction, is introduced to generate hierarchical feature embedding.

Context-aware feature extraction. For the reason that the operation of the standard single convolutional kernel is performed on the local receptive field, the spatial correspondence among landmarks in the scenario is neglected. In our context-aware feature extraction block, dilated convolutional layers with different kernel sizes are adopted to aggregate the multi-size regions on the CNN feature maps. Given the output features $\mathbf{f}$ of *Conv 5* block, the operation to capture the contextual features $\mathbb{F}_{i,c}$ at scale i is defined as:

$$\mathbb{F}_{i,c} = \mathcal{F}\left(\boldsymbol{W}_{k_i,d}, \mathbf{f}\right), \quad \mathbb{F}_{i,c} \in \mathbb{R}^{D_1 \times H_1 \times W_1}, \tag{1}$$

where $\mathcal{F}$ represents dilated convolutional operation, $\boldsymbol{W}_{k_i,d}$ is the convolutional filters with the kernel size k_i and dilation rate d. In our implementation, three kernel sizes, i.e., 3, 5 and 7, are utilized. The dilation rate is set to 2. Then, the extracted multi-scale contextual features can be summarized as $\{\mathbb{F}_{0,c}, \mathbb{F}_{1,c}, \mathbb{F}_{2,c}\}$. In this way, spatial correspondence among landmarks is sufficiently aggregated, which is effective for encoding the local cues for complex scenes in VPR.

Detailed feature extraction. In contrast to semantic feature maps captured from high layers, feature maps extracted from shallow layers remain higher spatial resolution and reserve more detailed information, such as edges and corners. It is argued that detailed information helps to have a better representation of scene structure and can increase the integrality of feature embedding. Therefore, features from shallow layers are added into the final feature embedding. In practice, features extracted from *Conv 3* block are utilized for a balance between reserving detailed information and encoding sufficiently. Furthermore, in order to filter out the redundant and useless information of the shallow layer features, a recalibration block is designed. Specifically, a scaling mask is generated by the recalibration block to refine the detailed information through a Hadamard product of $\mathbf{b}$ and $\mathcal{H}(\mathbf{b})$:

$$\mathbb{F}_{3,l} = \mathbf{b} \odot \mathcal{H}(\mathbf{b}), \quad \mathbb{F}_{3,l} \in \mathbb{R}^{D_2 \times H_2 \times W_2}, \tag{2}$$

where $\mathbf{b}$ is the output features of *Conv 3*, $\mathcal{H}(\cdot)$ is the learned function of the recalibration block. To limit the model complexity, the recalibration block is made up of a 3×3 convolutional layer and a sigmoid function, which is simple yet efficient.

By combining high-level contextual features with refined detailed features, the final hierarchical feature embedding X generated by HFM is summarized as:

$$X = \{\mathbb{F}_{0,c}, \mathbb{F}_{1,c}, \mathbb{F}_{2,c}\} \cup \{\mathbb{F}_{3,l}\} = \{\mathbb{F}_i | i \in [0,3]\}. \tag{3}$$

2.3. Weighting Similarity Metric

After extracting hierarchical features, a weighting similarity metric module, which includes a part-based similarity metric layer and a weighting layer, is designed to calculate the similarities. Specifically, the part-based similarity metric layer

(a) (b) (c)

Fig. 2. Sample images in (a) Nordland dataset (b) St Lucia dataset, and (c) Gardens Point dataset. Each column corresponds to an image pair that depicts the same place under different conditions.

measures the similarity scores between paired features in the same hierarchy. The weighting layer fuses the similarity scores into a final similarity decision by learning a set of parameters that evaluate the importance of each score.

Before calculating the similarity, we first partition the features $\mathbb{F}_i$ into four parts and then aggregate column vectors in each part into a part-level column vector $\boldsymbol{v}_m$ by a global average pooling operation $\overline{(\cdot)}$, as shown in Fig. 1(a). By this way, an enhanced feature representation $\overline{\mathbb{F}_i}$ with partial information is obtained:

$$\overline{\mathbb{F}_i} = \{\boldsymbol{v}_m | m \in [0,3]\}, \quad i \in [0,3]. \tag{4}$$

Then, the similarity score between $\mathbb{F}_i^q$ (from query image feature embedding X^q) and $\mathbb{F}_i^r$ (from reference image feature embedding X^r) is calculated by:

$$s_i = PBS\left(\overline{\mathbb{F}_i^q}, \overline{\mathbb{F}_i^r}\right), \quad i \in [0,3], \tag{5}$$

where $PBS(\cdot)$ is the part-based similarity metric [19]. The calculation process of $PBS(\cdot)$ is as follows:

1) By calculating the similarity of each pair of part-level column vector, a 4×4 similarity matrix $\boldsymbol{M}$ is obtained. In detail, the element $M_{m,n}$ denotes the cosine similarity between $\boldsymbol{v}_m^q \in \overline{\mathbb{F}_i^q}$ and $\boldsymbol{v}_n^r \in \overline{\mathbb{F}_i^r}$:

$$M_{m,n} = cos < \boldsymbol{v}_m^q, \boldsymbol{v}_n^r >, \quad m,n \in [0,3]. \tag{6}$$

2) For two images that come from the same place, the diagonal similarity scores derived from the same part-level should be higher, while others are lower. Diagonal similarity scores are then averaged and weighted to the final similarity score s_i:

$$\begin{gathered} s_i = \mathcal{G}(d_2 - d_1) \cdot d_1, \\ d_1 = \frac{1}{4}\sum_{m=0}^{3} M_{m,m}, \quad d_2 = \frac{1}{6}\sum_{m=0}^{3}\sum_{n>m} M_{m,n}, \end{gathered} \tag{7}$$

where $\mathcal{G}(\cdot)$ is a nonlinear function utilized to self-adaptively reweight d_1 to a more discriminative similarity score, according to the difference between average diagonal similarity scores and average off-diagonal similarity scores. In practice, sigmoid is exploited as the nonlinear function.

At last, similarity scores $\{s_i | i \in [0,3]\}$ between X^q and X^r in different hierarchical correspondence are fused to be the final similarity decision S_{dec} through the weighting layer:

$$S_{dec} = \sum_{i=0}^{3} w_i \cdot s_i, \quad s.t. \sum_{i=0}^{3} w_i = 1, \tag{8}$$

where w_i is the weight parameter. Instead of tuning the weights by grid search or random search, the designed weighting layer learns these weights automatically through a standard backpropagation algorithm. In the testing stage, the weights are fixed to the best-learned values, and S_{dec} is used to determine whether the query image and reference image belongs to the same place.

Table 1. Summarization of Testing Datasets.

Datasets	No. of frames		Variation	
	Reference	Test	Appearance	Viewpoint
Nordland	2000	2000	Severe	Moderate
St Lucia	900	3600	Moderate	Moderate
Gardens Point	200	400	Severe	Severe

2.4. Loss Function

In the training stage, given an image triplet $\langle I^a, I^p, I^n \rangle$ and their feature embedding $\langle X^a, X^p, X^n \rangle$, the similarity scores $\left\{s_i^+ | i \in [0,3]\right\}$ between X^a and X^p, $\left\{s_i^- | i \in [0,3]\right\}$ between X^a and X^n are calculated by the $PBS(\cdot)$. The loss function is based on triplet loss and summarized as:

$$L = \sum_{i=0}^{3} w_i \cdot max\left(s_i^- - s_i^+ + \delta, 0\right), \tag{9}$$

where w_i is the weight from the weighting layer, δ is a constant parameter giving the margin.

3. EXPERIMENTS AND DISCUSSIONS

This section first describes the details of datasets and evaluation protocols. Then the implementation details are given. Finally, experimental results and analyses on three datasets are presented in detail.

3.1. Datasets and Evaluation Protocols

Training dataset. The whole network is trained on the SPED_900 [20] dataset, which contains 57600 images from 900 locations in the world.

Testing datasets. Three typical benchmarks, Nordland [21] dataset, St Lucia [22] dataset, and Gardens Point [23] dataset are used to evaluate the effectiveness of the proposed method. **Nordland** dataset was recorded on a train with severe appearance changes (across seasons). *Spring* set is used as a reference run, and *winter* set is served as a revisit. **St Lucia** dataset was recorded in suburb at five different times (*8:45, 10:00, 12:10, 14:10, 15:45*) in a day, and on different days over a time of two weeks with temporal variations. *08:45* sequence is adopted as a reference sequence, and others are served as test sequences. **Gardens Point** dataset was captured on a campus which contained three traverses (*day_left, day_right, night_right*) with severe appearance and viewpoint

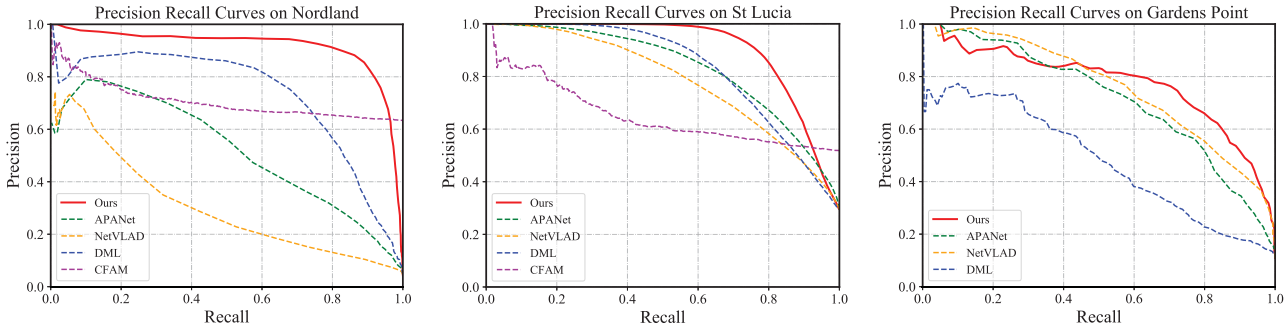

Fig. 3. PR curve comparison. **(a)** PR curves on Nordland dataset. **(b)** PR curves on St Lucia dataset. **(c)** PR curves on Gardens Point dataset.

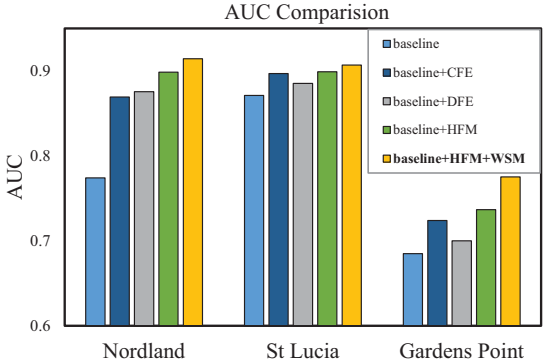

Fig. 4. AUC comparison on three datasets. CFE represents context-aware feature extraction, DFE means detailed feature extraction, HFM is a combination of CFE and DFE.

changes. *Day_left* set is served as a reference run, and others are applied as test sequences. The GPS annotations provided with St Lucia dataset are adopted as ground truth for coarse correspondence. As for Nordland dataset and Gardens Point dataset, frame-level correspondence is adopted as ground truth. Details of testing datasets are summarized in Table 1, and some example images are shown in Fig. 2.

Comparison methods. The performance of our method is compared with several state-of-the-art methods, including APANet [16], NetVLAD [24], CFAM [15] and DML [19].

Evaluation metrics. Following the recent visual place recognition methods, Precision Recall (PR) Curve and Area under Curve (AUC) are adopted to evaluate the performance.

3.2. Implementation Details

Our backbone network is initialized with the VGG16 [18] model pre-trained on ImageNet [25]. Parameters of the whole network are then fine-tuned during training. We resize the input images to 320×240. $[D_1, H_1, W_1]$, $[D_2, H_2, W_2]$ are the corresponding output size of *Conv 3* and *Conv 5* block, i.e., $[512, 20, 15]$, $[256, 80, 60]$. The margin δ in Eq. (9) is set to 0.5. The network is optimized by SGD with fixed momentum 0.9, and the batch size is 48. The learning rate is initially set to 0.01 and then reduced by a factor of 10 every 5000 iterations.

3.3. Experimental Results

Comparison with the state-of-the-art methods. Fig. 3(a) and Fig. 3(b) show the PR curves on Nordland dataset and St Lucia dataset, respectively. It is evident that our method outperforms other methods by a large margin. The results reveal the superiority of our method in handling seasonal and temporal variations. Fig. 3(c) shows the PR curves on Gardens Point dataset. Because CFAM [15] did not conduct experiments on this dataset, its experimental result is not shown in the figure. Although NetVLAD [24] and APANet [16] have a better performance at the low recall rate on this dataset, our method gives a higher precision as the recall rate increasing. Results on these three benchmarks demonstrate that by digging hierarchical information and exploiting an efficient metric mechanism, our method outperforms other methods.

Ablation study. Fig. 4 shows the experimental results of ablation study. Baseline denotes our framework without HFM and WSM. Note that when WSM is not used, we use the cosine similarity metric and average fusion mechanism. Either embedding CFE or DFE into the baseline (**baseline+CFE/baseline+DFE**), it shows consistent improvements over the baseline method on all the three datasets, indicating the effectiveness of exploiting contextual information and refined detailed information in visual place recognition. More importantly, by cooperating DFE with CFE, **baseline+HFM** outperforms **baseline+CFE** and **baseline+DFE**, showing complementary property of DFE and CFE. At last, combining WSM with HFM (**baseline+HFM+WSM**) brings about promising improvement, which verifies that the proposed WSM can make full use of hierarchical information and fuse them efficiently.

4. CONCLUSION

This paper presents a hierarchical feature extraction module and a weighting similarity metric module to exploit hierarchical information for visual place recognition. Compared with previous methods, our method makes full use of visual cues in the scenario by extracting multi-scale contextual features and refined detailed features. What's more, for efficiently exploiting the hierarchical information, the weighting similarity metric module is designed to measure the similarity score between features in the same hierarchy and learn the decision fusion weights automatically. Experimental results show that our method achieves state-of-the-art performance on three typical benchmarks. Besides, ablation study validates the effectiveness of each component of our method.

5. REFERENCES

[1] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2012, pp. 1643–1649.

[2] A.-D. Doan, Y. Latif, T.-J. Chin, Y. Liu, T.-T. Do, and I. Reid, "Scalable place recognition under appearance change for autonomous driving," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[3] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics (TR)*, vol. 32, no. 1, pp. 1–19, 2015.

[4] P. Yin, L. Xu, X. Li, C. Yin, Y. Li, R. A. Srivatsan, L. Li, J. Ji, and Y. He, "A multi-domain feature learning method for visual place recognition," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 319–324.

[5] D. M. Chen, S. S. Tsai, R. Vedantham, R. Grzeszczuk, and B. Girod, "Streaming mobile augmented reality on mobile phones," in *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2009, pp. 181–182.

[6] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt, "Scalable 6-dof localization on mobile devices," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 268–283.

[7] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics (TR)*, vol. 28, no. 5, pp. 1188–1197, 2012.

[8] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research (IJRR)*, vol. 27, no. 6, pp. 647–665, 2008.

[9] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision (ICCV)*, 2003, p. 1470.

[10] Z. Li and J. Tang, "Unsupervised feature selection via nonnegative spectral analysis and redundancy control," *IEEE Transactions on Image Processing (TIP)*, vol. 24, no. 12, pp. 5343–5355, 2015.

[11] H. Liu, W. Shi, W. Huang, and Q. Guan, "A discriminatively learned feature embedding based on multi-loss fusion for person search," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1668–1672.

[12] W. Shi, H. Liu, F. Meng, and W. Huang, "Instance enhancing loss: Deep identity-sensitive feature embedding for person search," in *IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 4108–4112.

[13] R. Gomezojeda, M. Lopezantequera, N. Petkov, and J. Gonzalezjimenez, "Training a convolutional neural network for appearance-invariant place recognition," *Computer Science*, 2015.

[14] H. J. Kim, E. Dunn, and J.-M. Frahm, "Learned contextual feature reweighting for image geo-localization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3251–3260.

[15] Z. Chen, L. Liu, I. Sa, Z. Ge, and M. Chli, "Learning context flexible attention model for long-term visual place recognition," *IEEE Robotics and Automation Letters (RA-L)*, vol. 3, no. 4, pp. 4015–4022, 2018.

[16] Y. Zhu, J. Wang, L. Xie, and L. Zheng, "Attention-based pyramid aggregation network for visual place recognition," in *ACM Multimedia Conference on Multimedia Conference (ACMMM)*, 2018, pp. 99–107.

[17] H. Liu, C. Zhao, W. Huang, and W. Shi, "An end-to-end siamese convolutional neural network for loop closure detection in visual SLAM system," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 3121–3125.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015, pp. 1–14.

[19] C. Zhao, R. Ding, and H. Liu, "End-to-end visual place recognition based on deep metric learning and self-adaptively enhanced similarity metric," in *IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 275–279.

[20] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 3223–3230.

[21] P. Neubert, N. Sünderhauf, and P. Protzel, "Superpixel-based appearance change prediction for long-term navigation across seasons," *Robotics and Autonomous Systems (RAS)*, vol. 69, pp. 15–27, 2015.

[22] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, "FAB-MAP+ RatSLAM: Appearance-based SLAM for multiple times of day," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2010, pp. 3507–3512.

[23] "Gardens Point dataset," https://wiki.qut.edu.au/display/cyphy/Day+and+Night+with+Lateral+Pose+Change+Datasets, accessed February 7, 2020.

[24] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5297–5307.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.