Robust place recognition based on salient landmarks screening and convolutional neural network features

Jie Niu¹ and Kun Qian²

Abstract
In this work, we propose a robust place recognition measurement in natural environments based on salient landmark screening and convolutional neural network (CNN) features. First, the salient objects in the image are segmented as candidate landmarks. Then, a category screening network is designed to remove specific object types that are not suitable for environmental modeling. Finally, a three-layer CNN is used to get highly representative features of the salient landmarks. In the similarity measurement, a Siamese network is chosen to calculate the similarity between images. Experiments were conducted on three challenging benchmark place recognition datasets and superior performance was achieved compared to other state-of-the-art methods, including FABMAP, SeqSLAM, SeqCNNSLAM, and PlaceCNN. Our method obtains the best results on the precision–recall curves, and the average precision reaches 78.43%, which is the best of the comparison methods. This demonstrates that the CNN features on the screened salient landmarks can be against a strong viewpoint and condition variations.

Keywords
Saliency map, visual place recognition, CNN features, Siamese network, changing environment

Date received: 31 December 2019; accepted: 26 September 2020

Introduction
Visual place recognition consists of identifying locations from a database that corresponds to the same place. From the perspective of image processing, place recognition can be considered to be an image retrieval problem. Many researchers have tried to find invariant features of scene images. Typical features include Haar-like features,¹ scale-invariant feature transform (SIFT),² histogram of oriented gradient (HOG),³ local binary patterns (LBP),⁴ and so on. These handcrafted features can be part of a bag-of-words model, which is widely used in some computer vision applications.⁵,⁶ However, in real applications, such as navigating vehicles through an unstructured environment,⁷ these handcrafted features have their inherent limitations, and they cannot work well in some situations, such as dynamic place recognition. Recently, with the development of deep learning, deep features trained on large-scale annotated datasets have shown great power in representing images

¹ School of Electronic Engineering, Changzhou College of Information Technology, Changzhou, Jiangsu, China
² School of Automation, Southeast University, Nanjing, Jiangsu, China

Corresponding author:
Jie Niu, School of Electronic Engineering, Changzhou College of Information Technology, 22 Mingxin Road, Wujing, Jiangsu, Changzhou 213164, China.
Email: niujie@czcit.edu.cn

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
that are robust to such variations.\textsuperscript{8,9} On the other hand, deep learning methods also have shortcomings in obtaining high-quality data annotations and the lack of various filtering mechanisms. For instance, when we need to recognize a place, using features extracted from dynamic objects, such as pets or bicycles, may lead to recognition failure or decrease recognition performance. In summary, except for some specific simple cases, the global features of images extracted by convolutional neural networks (CNNs) contain redundant information, which affect the accuracy of retrieval.

This article proposes a robust place recognition measurement in natural environments based on salient landmark screening and CNN features. We use the salient object detection algorithm to extract salient objects in the image. Then, we use the category-screening network to obtain candidate landmark sets by removing specific objects that are not suitable for environmental modeling. Finally, a Siamese network is designed to complete the scene recognition task. The overall flowchart of the proposed method is shown in Figure 1. The main contributions of our work can be summarized as follows.

We propose a novel salient landmark detection technique that can extract the salient candidate targets and then automatically remove the specific object regions according to the predefined categories, which improve the quality and effectiveness of image retrieval in the later stage.

We use CNNs to learn the representations of image patches and train a Siamese network architecture to compare pairs of patches corresponding and noncorresponding for better matching.

The remainder of this article is organized as follows: the second section gives a brief overview of related work. Our method is introduced in detail in the third section. In the fourth section, three datasets used in this article and their characteristics are firstly described. Then, experimental results compared with other methods are also provided. Finally, the article is concluded in the fifth section.

\textbf{Related Work}

The visual place recognition task can be considered as an image retrieval problem, and many researchers have explored this area. Many researchers have tried to operate directly on raw pixels\textsuperscript{10} or find the condition-invariant feature of places. Ali et al.\textsuperscript{11} used SIFT and SURF to describe the image. Gálvez-López and Tardos\textsuperscript{12} experimented with using a BRIEF features-based bag-of-words model in several datasets and demonstrated its effectiveness. However, in a real-world scenario, the characteristics of a place change all the time, and it is challenging to find correspondences between handcraft features.

Considering the above challenges, deep neural networks have become the mainstream image representation framework.\textsuperscript{13} Both viewpoint and appearance robustness have been demonstrated when CNNs are used for a wide variety of vision jobs like object recognition,\textsuperscript{14} object detection,\textsuperscript{15} and scene recognition.\textsuperscript{16} Inspired by these works, many researchers have tried to utilize intermediate representations from CNNs and have achieved competitive results. Wang et al.\textsuperscript{17} proposed a novel omnidirectional CNN to handle severe camera pose variation for place recognition. Sun et al.\textsuperscript{18} proposed a novel point-cloud-based place recognition method using a pretrained CNN. Park et al.\textsuperscript{19} proposed a lightweight CNN place recognition method for embedded systems, and the experimental results show that the method is significantly...
better than traditional methods in accuracy and calculation time. By combining the deep learning features with spatio-temporal filtering, Chen et al.\textsuperscript{20} completed the place recognition task based on CNN models. Sünderhauf et al.\textsuperscript{21} provided an in-depth analysis of the viewpoint-invariant properties of deep learned features for place recognition; however, there is at least one problem with the above approaches. The training process of deep learning network needs high-performance hardware devices and is time consuming. The deep model trained on the object recognition dataset (e.g. ImageNet) is not fully adapted to the application task of place recognition.\textsuperscript{22} In the real scene, the image of a place will change over time due to illumination, dynamic objects, and so on. Therefore, it is also inappropriate to extract features on the entire image, and a whole-image approach such as this is sensitive to viewpoint variations.

Some researchers have attempted to intelligently select useful information within an image for recognizing places. Sünderhauf et al.\textsuperscript{21} developed a landmark extraction algorithm and computed the neural responses to each landmark region in a scene. Hou et al.\textsuperscript{23} designed an objectness measure method to detect the landmarks and then expressed these regions as the features computed by the CNNs for matching in the next step. Additionally, saliency detection has attracted extensive research efforts in recent years. Salient detection is widely used in many applications, such as image classification,\textsuperscript{24} image retrieval,\textsuperscript{25} and visual place recognition.\textsuperscript{26} Visual attention can guide computing resources to fixed important areas. Compared with image processing on the entire image, it can effectively improve the recognition performance, that is because not all contents in the image play a positive role in representing a place.\textsuperscript{27–29} Among these methods, feature extraction on the entire image for place recognition introduces redundant information and reduces efficiency. Considering that CNN leads to the large space of image representation, Chen et al.\textsuperscript{30} proposed a new method of reserving salient feature maps and make adaptive binarization on it, and the experimental results proved the effectiveness. Since the highly representative landmark features are robust to appearance changes, Xin et al.\textsuperscript{31} proposed an effective method based on CNNs and content-based multiscale landmarks to complete the task of place recognition. Chen et al.\textsuperscript{32} proposed a method for place recognition based on CNN features, which achieved the aim of recognizing the place. However, their landmark detectors are based on all the salient objects in the image. In contrast, our method is able to remove salient objects that are not suitable for place recognition, and furthermore, a Siamese network is used to calculate the similarity, which can improve the matching efficiency.

Proposed system

This article proposes a robust place recognition measurement in natural environments based on salient landmark screening and CNN features. The main contribution is that our category-screening network removes the salient objects that do not suit the recognition in the current environment. Moreover, a Siamese network is used to calculate the similarity for efficiency optimization.

Category-screening network for salient landmarks

The category-screening network includes two parts of functions: salient object detection and object category recognition. Its core innovation lies in trying to exclude certain types of objects that are not suitable for environmental modeling. To achieve this, we firstly needed to select a state-of-art saliency object detection method. Since deep learning-based processing methods have gained great advantages in various fields of computer vision, our focus is on a salient object detection scheme based on CNN-based saliency approaches. Hou et al.\textsuperscript{33} found that in most CNN-based salient models, the outputs of the deeper side and the shallower side have their respective contributions to the correct results. They found that making short connections between the shallower and deeper side output layers improved the performance. After considering the performance and execution efficiency, in this article, we choose the DSS [33] method as the salient object detection algorithm.

After obtaining a large number of salient object images, pictures must be identified as to whether they belong to some specific categories. The above requirements can be considered to be a typical image classification problem, and good results have been achieved by the conventional CNN-based approaches. One of the characteristics of these methods is the large number of labeled training samples and powerful GPU resources required. For those who lack these conditions, parameter transfer learning can be used to improve performance by fine-tuning on the target dataset.

For this study, we built an AlexNet based on Caffe and trained it on the ILSVRC2012 dataset. Then, the pretraining model is used to estimate the training samples in the target domain. The purpose of optimization is to purify the samples on the training set. According to the study results of Koh and Liang,\textsuperscript{34} even a single sample in the training set can affect the final testing loss. Wang et al. proposed an optimal scheme for the training data.\textsuperscript{35} In our work, their implementation was modified, which is detailed in Algorithm 1. A pretrained model using training data from the ImageNet is chosen as the initial model, and then, the influence of each sample in the dataset based on the pretrained model is calculated. Finally, after removing the samples that reduce the loss of the verification set, the final model can be trained again with the new training set. In the scheme, \( x_i \) denotes a single training sample and \( f_s(x_i) \) represents a CNN model with input \( x \). Let \( L(f_s(x_i)) \) be the loss. After the removal of sample \( x_i \), the influence on the validation \( x_j \) is defined as \( I_{\text{loss}}(x_i, x_j) \). According to Koh and Liang,\textsuperscript{34} the \( I_{\text{loss}}(x_i, x_j) \) can be approximated as follows:

\[
I_{\text{loss}}(x_i, x_j) = -\nabla_{\theta} L(f_s(x_i))^T H^{-1}_\theta \nabla_{\phi} L(f_s(x_i))
\]  

(1)
where $L$ is usually a loss function like cross entropy, and so on. $H^{-1} = \frac{1}{n} \sum_{i=1}^{n} \nabla^2 f(f(x_i))$ is the Hessian and positive definite.

After following the above-stated steps, a newly optimized training set is obtained. Next, the weights of all convolutional layers of pretrained AlexNet are frozen, the weights of the last three fully connected layers are finely tuned, and the last layer of the network is truncated. Since the pretraining model we selected is generated based on the ImageNet dataset, and the softmax layer output of the model is 1000. Because the number of categories we need to recognize is far less than the number, a redesign softmax layer is required. For example, if 10 object categories that are not suitable as landmarks in the environment, the new softmax layer of the network should be 11 (including the type other than the target categories). After the model is trained, the salient images are sent into the model for classification, and the images classified into specific target categories, which are unsuitable for environmental modeling, will be removed. The whole process is shown in Figure 2.

**Algorithm 1**

**Input:**
- $f_\theta$: pre-trained deep learning model;
- $\chi$: training data;
- $v$: validation set;
- $i/j$: ratio of training to verification set;
- $\text{len}()$: sample size of the dataset.

**Output:**
- $\chi'$: Optimized training data;

**Begin**

for $i \leftarrow 1$ to $\text{len}(\chi)$ do
  for $j \leftarrow 1$ to $\text{len}(v)$ do
    use $f_\theta$ to get $I_{\text{loss}}(\chi(i), v(j));$
    if $\sum_I I_{\text{loss}}(\chi(i), v(j)) > 0$ then
      delete $\chi(i)$ from $\chi$
  return $\chi'$

**Similarity computing**

Traditional classification methods require an accurate label for each sample. When the number of categories to be classified is large and the number of samples in each category is small, the traditional method is not suitable. On the other hand, the Siamese network can learn a similarity measure and can be applied to a classification problem, where the number of categories is large or the training sample cannot be trained with the previous method. Therefore, after the landmark proposals have been extracted, we use a Siamese network to enforce descriptor similarity and dissimilarity for better matching. The Siamese network architecture that employs two CNNs with identical parameters to learn the representation of the image patches compares pairs of patches, and we minimize a loss that enforces the L2 norm of their difference to be small for corresponding patches and large otherwise. Figure 3 shows the overall idea of the similarity computing method.

Taking into account, factors, such as the size of the dataset, a mature three-layer CNN architecture is used to make the matching process more efficient, as shown in Figure 3. Considering the pixel size of salient objects, the sizes of the salient landmarks were adjusted to $64 \times 64$, while only CNN3’s output is used as the descriptors, and then, their L2 norm is computed, which is commonly used for image similarity measure. This is an instance of semi-supervised embedding. The margin-based loss function, which has been shown to outperform other methods in such applications, was used in this study

$$
\text{Loss}(f_i, f_j, w_{ij}) = \begin{cases} 
  \frac{1}{2} \| f_i - f_j \|^2 & \text{if } w_{ij} = 1 \\
  \max(0, m - \| f_i - f_j \|^2) & \text{if } w_{ij} = 0
\end{cases}
$$

where $f_i$ is the CNN3’s output of salient landmark $x_i$ in proposal landmark dataset, $f_j$ is the CNN3’s output of new input salient image patch $x_j$; $w_{ij}$ is a label, and if $x_i$ and $x_j$ should be similar, then $w_{ij} = 1$, or otherwise $w_{ij} = 0$; and $m$ is Euclidean distance that dissimilar patches should be away from each other.

![Figure 2. The workflow of the transfer learning method.](image-url)
Experiments

To verify the effectiveness of the method, we designed a series of experiments on place recognition problems.

Datasets

We selected three datasets that are widely used in the field of place recognition for verification. The details of the datasets are summarized in Table 1.

The Gardens Point dataset contains three image sequences, which were recorded on the Gardens Point Campus of QUT and along the Brisbane River by a pedestrian. The St Lucia dataset is a dataset gathered from a car with a forward-facing web camera driven in a 9.5-km circuit around the University. The synthesized Nordland dataset contains video clips from four different seasons of train driving perspectives. A special feature of the dataset is that the viewpoint and the field-of-view are in all four videos.

Evaluation methodology

We compared our work with three advanced place recognition methods, the FABMAP, SeqSLAM, SeqCNN-SLAM, and Place-CNN. FABMAP is a method, which makes use of the handcrafted features for place recognition and mapping. SeqSLAM is a sequence-based approach, which uses a global HOG descriptor and has been considered to be one of the most successful algorithms for place recognition. SeqCNN-SLAM is an improvement of SeqSLAM, SeqCNN-SLAM uses state-of-the-art deep learning techniques to generate robust feature representations of images. Place-CNN is a deep learning-based recognition model trained on the Places205 dataset. In view of
the good performance of the conv3 layer in the deep network, when the appearance changes, we choose conv3 layer to extract the features. The experimental results adopt the conventional precision–recall (PR) curves, which is often used in visual place recognition.\textsuperscript{42–44}

**Results and discussion**

As shown in Figures 4 to 6, the PR curves are computed on all three test datasets for our proposed approach against the other four algorithms. In all three results, our method, which combines the advantages of salient object detection, category-screening network, and CNN features, is always superior to other comparison methods.

Figure 4 shows the PR curves generated by these methods on the Gardens Point dataset. Our model achieves slightly better performance than SeqSLAM and SeqCNN-SLAM. The average accuracy (AP) of FABMAP is only 10.4\% in three datasets, which is the worst performance. However, there is still unfairness when the FABMAP method is directly compared with other methods in the article. One of the reasons is that FABMAP is focusing on place recognition to be used in mapping and localization applications, and the handcrafted features used by FABMAP have no advantage over the deep features in dealing with environmental changes. On the other hand, the SeqSLAM method gets better performance than the place-CNN method. This is mainly because the SeqSLAM method adopts sequence matching rather than single-image matching for place recognition, which fits well with the image features of the test dataset.

Figure 5 shows the PR curves generated by these methods on the St Lucia dataset. The images in the Lucia dataset change dramatically in light and appearance, both the FABMAP method and the SeqSLAM method based on handcrafted features do not perform well. When the SeqSLAM method replaces the features with CNN features, we can observe that SeqCNN-SLAM method performs far better than the original one. In addition, our method performs best among all the comparison methods using CNN features, leading by about 20\% on the average accuracy, indicating the benefits of using salient object detection and category-screening network.

Figure 6 shows the PR curves generated by these methods on the Nordland dataset. Our method is still ahead of other comparison methods. Compared with the SeqSLAM method, the SeqCNN-SLAM method has improved the average recognition accuracy by about 20.29\%. In addition, it is noteworthy that the PlaceCNN method performed poorly on the Nordland dataset, which is mainly because the appearance of the sequence in the dataset changes dramatically and was recorded only once per season. Since the PlaceCNN method extracts features from the whole image, there were a lot of redundant features. The performance of the model decreases seriously when the environment appearance changes dramatically.
Conclusion and future work

Inspired by the success of saliency attention models in other computer vision tasks and the vigorous development of deep learning techniques, in this article, we propose a place recognition method that takes advantage of the salient landmark screening and CNN features, which achieved good recognition results. Compared with the method of extracting features on the entire image, our method can automatically select informative regions for recognizing places. To get more accurate landmarks, we also designed a category-screening network, which can effectively remove specific object categories, and it makes the subsequent CNN features more stable. Evaluation against the state-of-the-art on several datasets reveals that our method obtains the best results on the PR curves, and the AP reaches 78.43%, which is nearly 15% points ahead of the second.

In the future, we will try to optimize the network structure in an end-to-end fashion, such as retaining the CNN features for the Siamese network while detecting the object class, to improve the place recognition performance.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Changzhou Applied Basic Research Planned Project [CJ20180010], “QingLan” Project of Jiangsu Province, CCIT Key Laboratory of Industrial IoT [KYPT201803Z], Natural Science Foundation of CCIT [XJ2020000601], and Changzhou Key Laboratory of High Technology [CM20183004].

ORCID iD

Jie Niu https://orcid.org/0000-0001-7085-5862

References

1. Viola P and Jones M. Rapid object detection using a boosted cascade of simple features. In: IEEE computer society conference on computer vision and pattern recognition, Kauai, HI, USA, 8–14 December 2001, pp. 511–518. Los Alamitos: IEEE.
2. Cao J, Chen T, and Fan J. Landmark recognition with compact BoW histogram and ensemble ELM. Multimed Tools Appl 2016; 75: 2839–2857.
3. Dalal N and Triggs B. Histograms of oriented gradients for human detection. In: International conference on computer vision and pattern recognition, San Diego, CA, USA, 20–25 June 2005, pp. 886–893. Los Alamitos: IEEE.
4. Ameur B, Masmoudi S, Derbel AG, et al. Fusing gabor and LBP feature sets for KNN and SRC-based face recognition. In: 2016 2nd international conference on advanced technologies for signal and image processing, Monastir, Tunisia, 21–23 March 2016, pp. 453–458. Piscataway: IEEE.
5. Wang C and Huang K. How to use bag-of-words model better for image classification. Image Vis Comput 2015; 38: 65–74.
6. Ali NM, Jun SW, Karis MS, et al. Object classification and recognition using bag-of-words (BoW) model. In: 12th international colloquium on signal processing and its applications, Malacca City, Malaysia, 4–6 March 2016, pp. 216–220. Piscataway: IEEE.
7. Abad EC, Alonso JM, Garcia MJG, et al. Methodology for the navigation optimization of a terrain-adaptive unmanned ground vehicle. Int J Adv Robot Syst 2018; 15(1): 1–11.
8. Arandjovic R, Gronat P, Torii A, et al. NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, USA, 26 June–1 July, 2016, pp. 5297–5307. Los Alamitos: IEEE.
9. Liang M and Hu X. Recurrent convolutional neural network for object recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, Massachusetts, USA, 7–12 June 2015, pp. 3367–3375. Los Alamitos: IEEE.
10. Milford MJ and Wyeth GF. SeqSLAM: visual route-based navigation for sunny summer days and stormy winter nights. In: International conference on robotics and automation, Saint Paul, MN, USA, 14–18 May 2012, pp. 1643–1649. Piscataway: IEEE.
11. Ali N, Bajwa KB, Sablatnig R, et al. A novel image retrieval based on visual words integration of SIFT and SURF. PLoS One 2016; 11: 1–20.
12. Gálvez-López D and Tardos JD. Bags of binary words for fast place recognition in image sequences. IEEE Trans Robot 2012; 28: 1188–1197.
13. Lowry S, Sünderhauf N, Newman P, et al. Visual place recognition: a survey. IEEE Trans Robot 2016; 32: 1–19.
14. Agrawal P, Girshick R, and Malik J. Analyzing the performance of multilayer neural networks for object recognition. In: European conference on computer vision, Zurich, Switzerland, 6–12 September 2014, pp. 329–344. Cham: Springer.
15. Vu T-H, Osokin A, and Laptev I. Context-aware CNNs for person head detection. In: Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 13–16 December 2015, pp. 2893–2901. Los Alamitos: IEEE.
16. Zhou B, Lapedriza A, Xiao J, et al. Learning deep features for scene recognition using places database. In: Advances in neural information processing systems, Montreal, Quebec, Canada, 8–13 December 2014, pp. 487–495. Red Hook: NIPS.
17. Wang T-H, Huang H-J, Lin J-T, et al. (eds). Omnidirectional CNN for visual place recognition and navigation. In: International conference on robotics and automation, Brisbane, QLD, Australia, 21–25 May 2018, pp. 2341–2348. Piscataway: IEEE.
18. Sun T, Liu M, Ye H, et al. Point-cloud-based place recognition using CNN feature extraction. IEEE Sensors J 2019; 19(24): 12175–12186.
19. Park C, Jang J, Zhang L, et al. (eds). Light-weight visual place recognition using convolutional neural network for mobile robots. In: International conference on consumer electronics, Las Vegas, NV, USA, 12–14 January 2018, pp. 1–4. Piscataway: IEEE.

20. Chen Z, Lam O, Jacobson A, et al. Convolutional neural network-based place recognition. In: Australasian conference on robotics and automation, Melbourne, Australia, 2–4 December 2014, pp. 1–8. Red Hook: Curran Associates, Inc.

21. Sünderhauf N, Shirazi S, Jacobson A, et al. Place recognition with convnet landmarks: viewpoint-robust, condition-robust, training-free. In: Proceedings of robotics: science and systems, Rome, Italy, 13–17 July 2015, pp. 1–10. Berlin: Springer.

22. Sünderhauf N, Shirazi S, Dayoub F, et al. On the performance of convnet features for place recognition. In: International conference on intelligent robots and systems, Hamburg, Germany, 28 September–2 October 2015, pp. 4297–4304. Danvers: IEEE.

23. Hou Y, Zhang H, and Zhou S. Evaluation of object proposals and convnet features for landmark-based visual place recognition. J Intell Robot Syst 2018; 92: 505–520.

24. Lei B, Tan E-L, Chen S, et al. Saliency-driven image classification method based on histogram mining and image score. Pattern Recognit 2015; 48: 2567–2580.

25. Wei S, Liao L, Li J, et al. Saliency inside: learning attentive CNNS for content-based image retrieval. IEEE Trans Image Process 2019; 28(9): 4580–4593.

26. Kim HJ, Dunn E, and Frahm J-M. Learned contextual feature reweighting for image geo-localization. In: IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017, pp. 3251–3260. Piscataway: IEEE.

27. Chen L-C, Yang Y, Wang J, et al. Attention to scale: scale-aware semantic image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, LAS VEGAS, USA, 26 June–1 July 2016, pp. 3640–3649. IEEE.

28. Chen L, Zhang H, Xiao J, et al. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, Hawaii, USA, 21–26 July 2017, pp. 5659–5667. Piscataway: IEEE.

29. Jie N, Xiongzhu B, and Kun Q. A method of extracting natural landmarks for mobile robot navigation. J Harbin Eng Univ 2019; 40: 844–850.

30. Chen Y, Gan W, Jiao S, et al. Salient feature selection for CNN-based visual place recognition. IEICE Trans Inform Syst 2018; 101(12): 3102–3107.

31. Xin Z, Cui X, Zhang J, et al. Real-time visual place recognition based on analyzing distribution of multi-scale CNN landmarks. J Intell Robot Syst 2019; 94(3–4): 777–792.

32. Chen Z, Maffra F, Sa I, et al. Only look once, mining distinctive landmarks from convnet for visual place recognition. In: IEEE/RSJ international conference on intelligent robots and systems, Vancouver, BC, Canada, 24–28 September 2017, pp. 9–16. IEEE.

33. Hou Q, Cheng M, Hu X, et al. Deeply Supervised salient object detection with short connections. IEEE Trans Pattern Anal Mach Intell 2019; 41: 815–828.

34. Koh PW and Liang P. Understanding black-box predictions via influence functions. In: Proceedings of the 34th international conference on machine learning, Sydney, Australia, 6–11 August 2017, pp. 1885–1894. Stroudsburg: International Machine Learning Society (IMLS).

35. Wang T, Huan J, and Li B. Data dropout: optimizing training data for convolutional neural networks. In: 30th international conference on tools with artificial intelligence, Volos, Greece, 5–7 November 2018, pp. 39–46. Los Alamitos: IEEE.

36. Simo-Serra E, Trulls E, Ferraz L, et al. Discriminative learning of deep convolutional feature point descriptors. In: International conference on computer vision, Santiago, Chile, 13–16 December 2015, pp. 118–126. Piscataway: IEEE.

37. Hadsell R, Chopra S, and LeCun Y. Dimensionality reduction by learning an invariant mapping. In: IEEE computer society conference on computer vision and pattern recognition, New York, NY, USA, 17–22 June 2006, pp. 1735–1742. Los Alamitos: IEEE.

38. Chen Z, Jacobson A, Sünderhauf N, et al. Deep learning features at scale for visual place recognition. In: IEEE international conference on robotics and automation, Singapore, 29 May–3 June 2017, pp. 3223–3230. Piscataway: IEEE.

39. Lowry S and Andreasson H. Lightweight, viewpoint-invariant visual place recognition in changing environments. IEEE Robot Autom Lett 2018; 3: 957–964.

40. Cummins M and Newman P. Highly scalable appearance-only SLAM - FAB-MAP 2.0. In: Robotics: science and systems V, Seattle, USA, June 28–July 1 2009, pp. 1–8. Cambridge: MIT Press.

41. Bai D, Wang C, Zhang B, et al. CNN feature boosted SeqSLAM for real-time loop closure detection. Chin J Electron 2018; 27: 488–499.

42. Wang Z, Peng Z, Guan Y, et al. Manifold regularization graph structure auto-encoder to detect loop closure for visual SLAM. IEEE Access 2019; 7: 59524–59538.

43. Chen Z, Liu L, Sa I, et al. Learning context flexible attention model for long-term visual place recognition. IEEE Robot Autom Lett 2018; 3: 4015–4022.

44. Fan C, Chen Z, Jacobson A, et al. Biologically-inspired visual place recognition with adaptive multiple scales. Robot Autonom Syst 2017; 96: 224–237.