Journal Pre-proof

Visual Place Recognition: A Survey from Deep Learning Perspective

Xiwu Zhang, Lei Wang, Yan Su

PII: S0031-3203(20)30563-X
DOI: https://doi.org/10.1016/j.patcog.2020.107760
Reference: PR 107760

To appear in: *Pattern Recognition*

Received date: 10 November 2019
Revised date: 4 September 2020
Accepted date: 25 November 2020

Please cite this article as: Xiwu Zhang, Lei Wang, Yan Su, Visual Place Recognition: A Survey from Deep Learning Perspective, *Pattern Recognition* (2020), doi: https://doi.org/10.1016/j.patcog.2020.107760

Highlights

- We provide a whole picture about deep learning-based visual place recognition.
- The differences and similarities between VPR and image retrieval are included.
- We review different kinds of CNN-based methods, novel CNN features and datasets for VPR.
- New tools such as GANs and multi-modality feature fusion are discussed for VPR.
- We discuss challenges, open issues and future directions of visual place recognition.

Visual Place Recognition: A Survey from Deep Learning Perspective

Xiwu Zhang[a], Lei Wang[b], Yan Su[a,*]

[a] *Nanjing University of Science and Technology, Nanjing, Jiangsu, 210094, P.R.China*
[b] *University of Wollongong, Wollongong, NSW 2522, Australia*

Abstract

Visual place recognition has attracted widespread research interest in multiple fields such as computer vision and robotics. Recently, researchers have employed advanced deep learning techniques to tackle this problem. While an increasing number of studies have proposed novel place recognition methods based on deep learning, few of them has provided a whole picture about how and to what extent deep learning has been utilized for this issue. In this paper, by delving into over 200 references, we present a comprehensive survey that covers various aspects of place recognition from deep learning perspective. We first present a brief introduction of deep learning and discuss its opportunities for recognizing places. After that, we focus on existing approaches built upon convolutional neural networks, including off-the-shelf and specifically designed models as well as novel image representations. We also discuss challenging problems in place recognition and present an extensive review of the corresponding datasets. To explore the future directions, we describe open issues and some new tools, for instance, generative adversarial networks, semantic scene understanding and multi-modality feature learning for this research topic. Finally, a conclusion is drawn for this paper.

Keywords: Visual place recognition, Deep learning, Visual SLAM, Survey

1. Introduction

Visual place recognition (VPR) aims to help a robot or a vision-based navigation system determine whether it locates in a previously visited place. It is one of the essential and challenging problems in the field of robotics and computer vision. These fields have witnessed a surge in the use of VPR for various applications in the last decade. For example, in a visual Simultaneous Localization and Mapping (SLAM) system [1], place recognition, which is also referred to as loop closure detection (LCD), is a key component. It can not only reduce the localization error induced by visual odometry (VO), but also avoid building an ambiguous map of the unknown environment [2].

While visual place recognition has received considerable attention and has been extensively studied in computer vision and robotics communities, there are still numerous open issues to deal with. The challenge for visual place recognition is twofold in general. First, false place recognition renders interference to a localization algorithm, which decreases the accuracy and even results in a catastrophic localization failure for a navigation system. Therefore, place

*Corresponding author
Email addresses: 314101002261@njust.edu.cn (Xiwu Zhang), leiw@uow.edu.au (Lei Wang), suyan@njust.edu.cn (Yan Su)

recognition algorithms must be able to achieve a very high recognition precision (100% precision is even needed in some cases). Second, most of the existing visual place recognition methods are appearance-based. However, the appearance of the same place may change drastically along with different illumination conditions, viewpoints, seasons, distance, occlusion and/or background clutter. As a result, it is challenging to correctly recognize the same place if it undergoes appearance changes. On the other hand, it is not uncommon that visual place recognition methods suffer perceptual aliasing issue [3], i.e. regarding different places as the same one. This is because i) the visual sensors (for example, cameras) obtain partial information of the environment, which limits the discrimination of the visual data, and ii) normally there are similar objects such as trees and buildings in the environment, which means that images from different places may have a similar appearance, resulting in difficulties for recognition.

The key part of a place recognition system is how to represent the place effectively. In this paper, we consider the concept of place in the field of localization or navigation. That is, a place indicates a spatial location as well as its nearby locations (within a small range), which are 2-dimensional or 3-dimensional, depending on its application scenarios. For visual place recognition, the sensory information corresponds to images collected by cameras at that location. In this case, the task of place representation is cast to an image representation problem. For a long period of time, VPR has been limited to approaches where images are represented by handcrafted features — either local features such as SIFT [4] and SURF [5] or global features such as HOG [6]. Some well-known VPR algorithms, for example, DBoW [7], FAB-MAP [8, 9] and the landmark-based relocalization approach [10], are all based on handcrafted local image features and have been widely used for visual SLAM or localization tasks.

Recently, deep learning (DL) has become one of the most attractive research topics and has achieved great success in many areas, including computer vision (CV) [11, 12] and robotics [13, 14]. In the last five years, researchers from computer vision and robotics communities have leveraged this advanced technology to address visual place recognition. Prevalent international conferences such as IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) and International Conference on Robotics and Automation (ICRA) also show their great interest by holding a series of workshops regarding DL-based visual place recognition. Numerous papers [15, 16, 13] have demonstrated that the performance of visual place recognition methods based on deep neural networks (DNNs), especially the convolutional neural networks (CNNs), is superior to that of traditional methods. Fig. 1 presents the milestones of visual place recognition over the past decade, including traditional handcrafted feature-based methods and the learned feature-based methods.

While a growing number of papers have developed DL-based VPR methods from various perspectives, there remains barriers for researchers (especially those who do not have sufficient experience on deep learning) to fully understand this research topic due to the lack of the literature that provides a whole picture about how and to what extent DL techniques have been employed in VPR. In light of this, this paper aims to give a comprehensive survey on VPR from DL perspective. This paper also tries to identify open issues and promising directions for future research. Note that this paper will not discuss traditional handcrafted feature-based approaches in detail. For a thorough historical review and a panorama of traditional approaches about visual place recognition, readers are referred to the survey

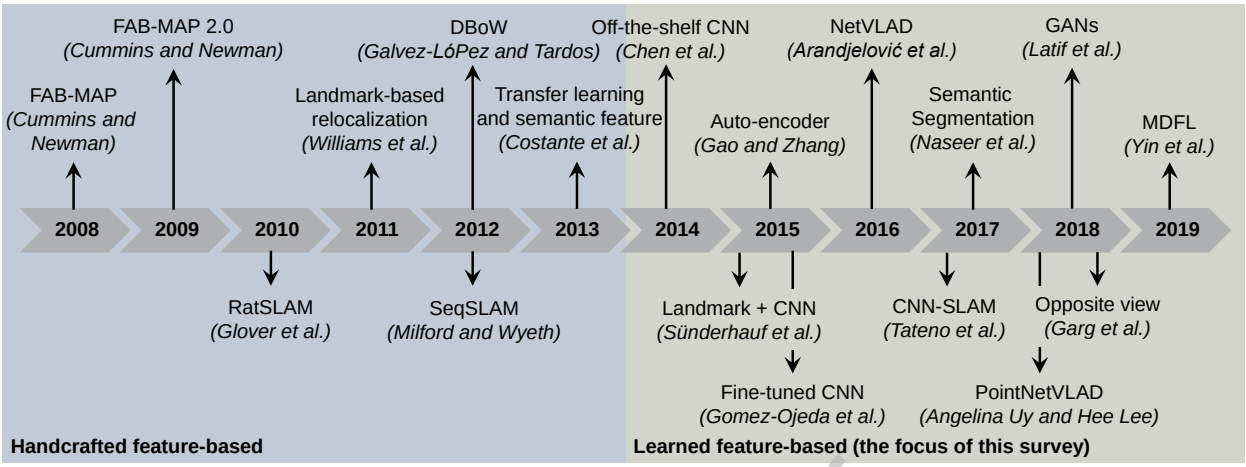

Figure 1: Milestones of visual place recognition methods over the past decade, from traditional handcrafted feature-based methods to more recent (since 2014) learned feature-based methods. This paper focuses on learned feature-based ones.

given by Lowry et al. [3].

This paper is organized as follows. An introduction of deep learning, especially convolutional neural networks, is presented in Section 2. The potential opportunities of deep learning in VPR are then discussed in Section 3. Section 4 presents CNN-based place recognition methods, including off-the-shelf and specifically designed architectures as well as the novel CNN feature representations. An exhaustive investigation of datasets for place recognition and the challenging problems such as recognizing places across seasons are described in Section 5. Open issues for this research topic and the new tools beyond CNN, for example, generative adversarial networks, semantic scene understanding and place recognition with heterogeneous data, are explored in Section 6. Finally, the conclusion of this paper and future directions regarding VPR problem are given in Section 7.

2. A Brief Introduction of Deep Learning

This section reviews the history of deep learning (DL) briefly. We first show some basic ideas behind this advanced technique in Section 2.1. After that, convolutional neural network (CNN) is focused in Section 2.2 since it plays the most important role in DL-based place recognition and will be mentioned frequently in the rest of this paper.

2.1. What is Deep Learning?

Deep learning [17] is a subfield of machine leaning (ML) and it aims to learn high-level representations from raw data — such as image [18, 12] and speech [19, 20]. Different from conventional machine learning algorithms, where engineers or domain experts design feature representation empirically for specific recognition tasks, deep learning is capable of discovering representations needed for pattern recognition automatically from raw data. This is done by using deep neural networks (DNNs), which consist of multiple trainable layers arranged hierarchically. The raw data is fed into DNNs as input. As the data flows through each layer, more and more abstract representations are

obtained. In this manner, the raw input data is eventually represented by a high-dimensional feature vector, which is very effective in discriminating different patterns.

Researchers have developed a series of deep neural networks, for example, Convolutional Neural Network (CNN) [18], Recurrent Neural Network (RNN) [21], auto-encoder [22], and so on for various application scenarios. Though the architectures of DNNs are diverse, most of them are composed of basic components, including convolutional layer, pooling layer, fully connected layer and non-linear operation such as the rectified linear unit (ReLU).

The key aspect of DNNs is that they contain a huge number (say, millions) of trainable parameters, which enable them to learn complex functions to map raw data to compact features for recognition tasks. Before DNNs are used for recognition, they must be trained carefully on training datasets. However, it is time-consuming to train a DNN with so many parameters. In addition, a large-scale dataset with labeled data is needed since most of the networks are supposed to be trained in a supervised manner. At the beginning, due to the limitation of labeled data, it is intractable to implement DNNs in practice. However, in the last decade, an explosion has been witnessed in the development of deep learning in various fields. The reason is twofold. First, the advanced computational technologies and excellent software and hardware enable researchers to program and train networks conveniently and efficiently. For example, according to [23], using graphics processing unit (GPU) makes it tens of orders of magnitude faster to train a network. Second, large-scale annotated datasets, for example, ImageNet [24], provide sufficient data for the training of DNNs.

2.2. *Convolutional Neural Network (CNN)*

One of the most successful application areas of deep learning is computer vision. This benefits immensely from convolutional neural network (CNN), which may be the most commonly used model in the field of deep learning. CNN is a type of deep neural network designed to deal with data in multiple arrays, for instance, a three-channel (RGB) image with a two-dimensional array in each channel.

As illustrated in Fig. 2(a), the typical architecture of a CNN is composed of convolutional layers, pooling layers and fully connected layers. Take an image as an example, as it flows through each convolutional layer, new feature maps are obtained by applying a set of functions. Specifically, we denote K filters and the corresponding biases in the l-th convolutional layer as $\mathcal{W} = \{\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_K\}$ and $\mathcal{B} = \{b_1, b_2, \ldots, b_K\}$, respectively. The k-th feature map $\mathbf{X}_k^l$ at the l-th layer is then obtained as $\mathbf{X}_k^l = f(\mathbf{W}_k * \mathbf{X}^{l-1} + b_k)$, where $\mathbf{X}^{l-1}$ is the output of previous layer, $*$ denotes convolution operation and $f(\cdot)$ is the activation function such as the rectified linear unit (ReLU), which is given by $f(x) = \max(x, 0)$. Pooling operation is applied to reduce the dimensions of the feature maps. With the layers going deeper, high-level image representations are learned. At the end, several fully connected layers are used for particular recognition tasks such as image classification [18].

Three significant points that make CNNs unique and effective are local connectivity, weight sharing and pooling. Local connectivity scheme is showed in Fig. 2(b). The neurons are connected to a sub-region, which is referred to as *receptive field*, of the previous layer. Local connectivity enables CNNs to learn strong responses to spatially local input patterns. Furthermore, with the convolutional layers stacked, neurons from deeper layers have a larger receptive

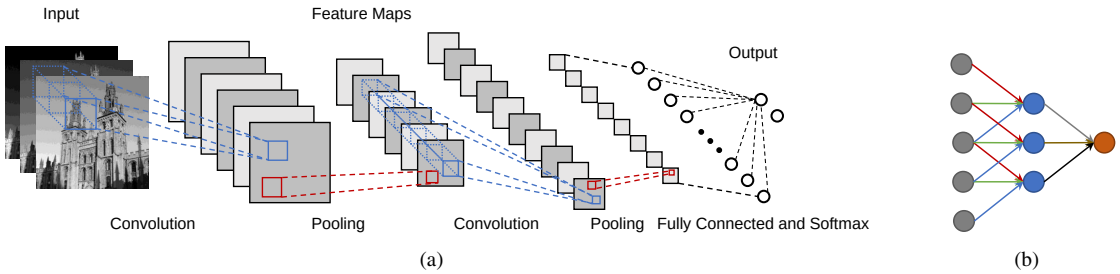

Figure 2: **(a)** Typical architecture of a CNN, including convolutional layers, pooling layers, fully connected layers and a Softmax layer. **(b)** Local connectivity and weight sharing scheme. To illustrate clearly, 1D data is showed as an example. A neuron is connected with a sub-region of its previous layer. The neurons in one layer could share the same weights, which are indicated using the lines with the same color.

field, which means that they encode more information of the input. This is why CNNs learn low-level features such as color or edges in images at the first several layers and learn high-level semantic features as the networks go deeper.

Another critical scheme of CNNs is weight sharing, which means that the entire visual field is convolved with the same filter (as illustrated in Fig. 2(b)). By exploiting shared weights, the number of parameters to be trained of a network no longer varies with the size of input images. Instead, it depends on the kernel size, which impressively reduces the total number of parameters to be trained. More importantly, weight sharing enables CNN models to be invariant to the translation of objects in images. That is, if an object appears in one part of an image, it could appear in any other locations. Accordingly, objects at different locations are supposed to be detected by the same filter.

Pooling is a type of operation to down-sample input feature maps. A typical pooling method used in CNNs is max-pooling, i.e. applying a max filter on a feature map to compute the maximum of sub-regions. The advantage of performing max-pooling is twofold. First, it is an effective way to reduce the dimensions of intermediate representations, and hence reduces the computational cost. Second, to some extent, max-pooling avoids the impact of shifting of patterns on the representation, which provides translation invariance property for image features learned by CNNs.

2.2.1. Popular CNN models

LeNet, introduced in 1998 by LeCun et al. [25], is considered to be an important early CNN model. It consists of two convolutional layers and two max-pooling layers, followed by fully connected layers. LeNet was successfully used for optical character recognition (OCR). After that, CNNs have not been widely used for a long period of time due to the complexity and difficulty for training them until 2012 when AlexNet [18] is introduced. AlexNet wins the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 and rekindles researchers' interest on DNNs, which brings a revolution of deep learning and computer vision. AlexNet consists of five convolutional layers and employs new techniques such as the ReLU, dropout [26] and GPUs for network training in CNNs for the first time.

In the following few years, many convolutional neural networks are developed with a preference of going far deeper. For instance, with a similar architecture to AlexNet [18], VGG16 [27] stacks 13 convolutional layers and achieves remarkable results on ImageNet dataset [24]. On the other hand, GoogLeNet [28] wins ILSVRC 2014 by proposing a new module referred to as Inception, which is able to reduce the number of parameters to be learned

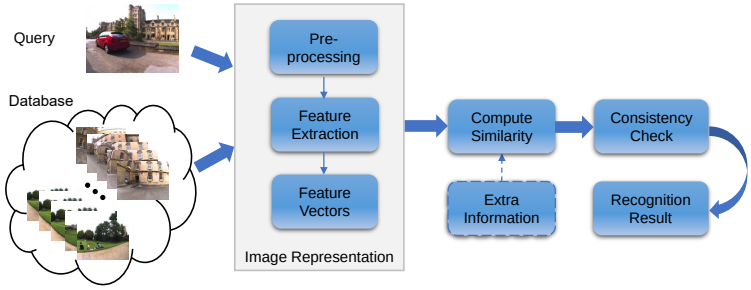

Figure 3: A basic pipeline of visual place recognition.

and process visual information at various scales. However, simply stacking more layers may render a degradation problem, which means that the accuracy of a suitably deep model does not increase if it is stacked more layers. The more recent model named ResNet proposed by He et al. [29] addresses this issue by using a residual block, where a residual mapping rather than the original mapping is learned. By doing so, the accuracy can be well improved even the depth of a CNN model increases and it is effective to train the model.

3. Visual Place Recognition: from the Deep Learning Perspective

3.1. Visual Place Recognition Pipeline

A basic pipeline of a visual place recognition system is illustrated in Fig. 3. Given an image of a query place, the first step is to convert the raw image to a mathematical representation (normally a feature vector) that can be expressed quantitatively. The feature vector is then used to compute the similarities between the query and the database images. The similarity score of two images indicates a belief about whether or not they are from the same place. If the score is larger than a given threshold, the corresponding images will be considered to be matched. Note that a further verification (on the basis of some specific constraints such as temporal consistency or geometrical consistency [7]) for the matched images is normally needed before the final recognition result is reported.

The most significant part of a visual place recognition system is image representation, which is similar to the case in most visual recognition tasks such as image retrieval, image classification, object detection and so on. Traditionally, VPR systems utilize handcrafted feature-based models, for example, Bag-of-words (BoW) [30] model, for image representation. The well-known handcrafted local features such as SIFT [4] and SURF [5] are not robust enough with respect to the environmental changes such as lighting conditions, scales and viewpoints. On the other hand, as discussed in Section 2.2, features learned from CNNs are more robust and discriminative. Researchers [31, 32] demonstrate that learned features outperform handcrafted ones for visual recognition tasks. Inspired by these papers, researchers from VPR community start to utilize advanced deep learning techniques to tackle place recognition.

Chen et al. [15] propose a CNN-based place recognition method for the first time in 2014. They use a pre-trained CNN model named Overfeat [33] as a feature extractor. The output of each layer forms a feature vector to represent the input image. The feature vectors are then used to calculate the similarities between different images. Experimental

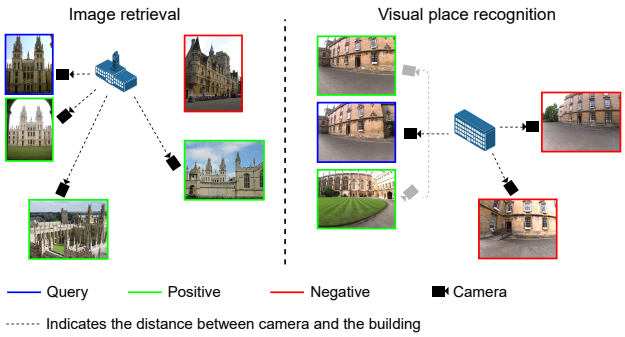

Figure 4: Illustration of image retrieval (left) and visual place recognition (right). Images related to the building are collected with a camera (in black). The position and direction of the camera in this figure indicate its position and perceptual viewpoint in the real world, respectively. Cameras (in gray) linked with the gray dashed lines indicate that they are from the same position (but with different viewpoints). For image retrieval, given a query (with a blue border), whether an image is a positive (with a green border) or a negative (with a red border) generally depends on the visual similarity, less considering the position information. However, for visual place recognition, position must be considered, which means that the image with drastic changes (induced by, for example, different viewpoints) in appearance could be a positive while the one with very similar appearance could be a negative.

results on a large-scale place recognition dataset show that CNN features perform better than handcrafted features. In the following few years, a series of interesting methods have been developed to explore the application of various deep learning approaches — from global CNN features [16, 34] to semantic representations [35, 36, 37] and from pre-trained generic CNN models [32, 38] to task-specific deep neural networks [13, 39] — to visual place recognition. Fig. 1 shows the timeline of representative methods where various techniques in the domain of DL are used to boost the performance of VPR. In the following sections, we will discuss each of these topics in detail.

3.2. Visual Place Recognition and Image Retrieval

Visual place recognition shares significant similarity with content-based image retrieval (CBIR) [40, 41], which is one of the important applications of deep learning in computer vision. The aim of CBIR is to search for the images that are similar to a query in appearance from a large database. To some extent, what lies at the heart of both CBIR and VPR is how to build image representations appropriately, based on which the system is able to obtain a measure (e.g., Euclidean distance of feature vectors) to decide whether or not the corresponding images are matched.

Although the key issues of the two tasks are similar, one shall not simply regard VPR as an image retrieval problem. The main differences between VPR and image retrieval are twofold. First, as illustrated in Fig. 4 (left), image retrieval systems find matched images generally based on the appearance similarity, and less consider the associated position information [3]. However, for VPR systems, it is emphasized that the recognition result should depend on whether or not the matched images are taken by cameras at the same place in the real world. For example, in Fig. 4 (right), the image (bottom-left with a green border) is deemed as positive with respect to the query (with a blue border) due to the fact that it is from the same or similar place (position), even though it has a different appearance from the query. By contrast, the image (right with a red border) that has a very similar appearance to the query could be regarded as negative because it is from the other facet of the same building. In a word, image retrieval mainly concerns the visual content of images, while place recognition concerns more the *place* or *position* where the images are taken. Second,

the database of an image retrieval system is relatively stable and a query image normally needs to be compared with all the images in database. However, for VPR, the database changes temporally in most cases. For instance, in a visual SLAM system [42], the number of images increases with the robot traveling through the environment. Fortunately, VPR systems are able to ensure a fast retrieving speed by comparing the query image only with those collected at the currently believed location and its adjacent neighbors inferred by using extra information or prior knowledge [43].

Visual place recognition methods greatly benefit from image retrieval due to the similarity of the two tasks. Various novel image representations and learning models developed in the domain of image retrieval have been used to tackle challenging VPR issues such as across season place recognition [14, 44, 45]. Another useful image retrieval technique that has been employed for visual place recognition is *query expansion* (QE) [46]. It is a kind of re-ranking method that is able to significantly improve the accuracy of retrieval. A commonly used variant is average query expansion (AQE) [47], which averages the high-ranked positives of the original query to issue a new query to conduct one more retrieval. It is an effective way to compensate for the adverse impact on retrieval caused by, for instance, viewpoint and illumination changes [48]. Recently, Zhang et al. [49] employ a more sophisticated query expansion method, which is referred to as *diffusion process* [50, 51], to tackle VPR for the first time. Experimental results on challenging datasets such as KITTI [52] show its promising performance. The detailed discussions on novel image representations and learning models will be conducted in Sections 4.2 and 4.3, respectively.

4. CNN-Based Place Recognition

In this section, we introduce the VPR methods based on CNNs. This category of methods starts from using pre-trained CNN models (Section 4.1) as feature extractor to construct image representation to measure image similarity. Various types of CNN features have been developed (Section 4.2). Soon after, instead of using pre-trained models, researchers either fine-tune a model on specific VPR datasets or design new architectures to improve the recognition performance (Section 4.3). Table 1 summarizes the typical work regarding VPR approaches based on CNNs. We also discuss another two critical steps in CNN-based VPR methods: similarity measure (Section 4.4) and evaluation criterion (Section 4.5). At the end, we compare the computational performance for some representative methods (Section 4.6).

4.1. Using Pre-trained CNN Models

4.1.1. Off-the-shelf Features

The most straightforward way to develop CNN-based place recognition methods is to regard pre-trained CNN models as feature extractor. Once a CNN is trained on a sufficiently large dataset (for instance, ImageNet [24]), the activations of its intermediate convolutional layers or fully connected layers can be used as global feature representation of an input image. For example, suppose that a convolutional layer consists of K feature maps with width W and height H. Then a vector in the length of $K \times W \times H$ can be obtained as the image representation by flattening those

Table 1: Representative work regarding CNN-based visual place recognition (VPR).

Reference	Architecture[a]	Training[b]	Highlights
[15] (2014)	Overfeat [33]/Caffe[53]	Pre-trained	The first work that employs off-the-shelf CNN features to tackle VPR problem
[34] (2015)	AlexNet [18]	Pre-trained	Evaluate the performance of CNN features compared with handcrafted features
[16] (2015)	AlexNet [18]	Pre-trained	Compare the performance of CNN features from different convolutional layers
[32] (2014)	Overfeat [33]	Pre-trained	Raw features are augmented with PCA and whitening to improve their performance
[14] (2015)	AlexNet [18]	Pre-trained	Landmark-based place recognition; Detect landmarks with EdgeBox
[54] (2016)	AlexNet [18]	Pre-trained	Place recognition with landmark distribution descriptor (LDD)
[55] (2016)	VGG [27]	Pre-trained	Landmark-based; Detect landmarks with superpixel segmentation (SP-Grid) technique
[56] (2018)	AlexNet [18]	Pre-trained	Use CNN descriptors to boost the performance of SeqSLAM (SeqCNNSLAM)
[57] (2017)	VGG16 [27]	Pre-trained	Detect salient landmarks and extract their features using one CNN
[49] (2019)	VGG16 [27]	Pre-trained	Use R-MAC feature; Employ diffusion technique to boost similar image association
[58] (2019)	HybridNet[59]/ AlexNet[18]	Pre-trained	Use feature map filter to remove distracting feature maps
[13] (2017)	AlexNet [18]	Fine-tuned	Fine-tune a pre-trained CNN for VPR for the first time; Features are created in an end-to-end fashion, without additional coding or pooling
[60] (2017)	Fast-Net [61]	Fine-tuned	Semantic-aware; Up-convolutional architectures
[62] (2019)	AlexNet [18]/ VGG16 [27]/ ResNet101 [29]	Fine-tuned	Employ structure-from-motion methods to select training data automatically; Siamese architecture for fine-tuning
[39] (2016)	AlexNet[18]/ VGG16 [27] + NetVLAD [39]	Hybrid	Design a specific architecture for place recognition; Achieve groundbreaking recognition result on large-scale datasets
[59] (2017)	AlexNet [18]	Hybrid	Present a massive dataset SPED for place recognition
[63] (2018)	VGG16 [27]	Hybrid	Develop an attention model to adaptively generate regions with flexible context
[64] (2018)	Task specific	From scratch	Convolutional autoencoder architecture; Trained in an unsupervised way
[65] (2017)	AlexNet [18]/ VGG16 [27] + CRN [65]	Hybrid	Introduce a context re-weighting network (CRN) to focus on substantial regions within an image
[66] (2019)	ResNet101 [29] + LLN[66]	Hybrid	Design a landmark localization network (LLN) to generate discriminative landmarks
[67] (2019)	Task specific	From scratch	Utilize distance metric learning methods to learn an enhanced similarity measure

[a] Specify the adopted underlying CNN architecture. **Task specific** means that the architecture is designed for place recognition task specifically.
[b] **Pre-trained** indicates that the network is initially trained on existing datasets (e.g., ImageNet [24]) for classification; **Fine-tuned** means that the network is initialized with pre-trained weights and then re-trained on particular datasets; **From scratch** means that the network is totally trained on particular datasets, without any prior knowledge. **Hybrid** means that two or three of the above choices are selected in the corresponding study.

feature maps. The very first work following this idea to tackle place recognition is presented by Chen et al. [15]. They utilize off-the-shelf CNN features extracted from a pre-trained model called Overfeat [33] to represent images.

The similar idea is adopted by Sünderhauf et al. [16] and Nasser et al. [45] where AlexNet [18] pre-trained on ImageNet [24] dataset is used. The performance of features from each individual layer is evaluated and compared with the methods [8, 68] based on handcrafted features. An interesting finding of [16] is that features extracted from intermediate layers (e.g., the third convolutional layer of AlexNet [18]) are more robust against different lighting and weather conditions of the environment, while the features from top layers (e.g., the last fully connected layer) gain more robustness towards viewpoint changes. By presenting a comparative study between CNN features extracted from a pre-trained model and four types of handcrafted features, i.e. GIST [69], VLAD [70], Bag-of-Visual-Words (BoVW) and Fisher Vector (FV) [30], Hou et al. [34] discover that CNN-based image representations significantly outperform handcrafted features when the dataset involves drastic illumination changes. Bai et al. [56] propose a method (called SeqCNNSLAM) that uses pre-trained AlexNet [18] to extract robust CNN feature to boost the performance of SeqSLAM [68], which is a prevalent place recognition algorithm based on handcrafted features.

The above studies demonstrate the potential of CNN-based visual place recognition methods and conclude that the learned features are more robust to the environmental variations than conventional handcrafted features. This provides new opportunities to tackle the challenging VPR problem.

4.1.2. Feature Post-processing

Compared with using the output of intermediate layers as image representation directly, a further *post-processing* or a *feature augmentation* procedure is of great benefit for recognition tasks [71, 72]. A simple normalization step is proved to be useful to improve the performance of CNN features. For instance, Hou et al. [34] perform ℓ_2-normalization on the raw CNN feature vector $\mathbf{f}$ before it is used to calculate similarity score, that is $\mathbf{f}' = \mathbf{f}/\|\mathbf{f}\|_2$, where $\|\cdot\|_2$ is the ℓ_2-norm of the vector. This makes $\mathbf{f}'$ to be a unit vector (whose inner product corresponds to the cosine similarity), which is helpful for comparing measurements of different images.

Another commonly used post-processing method is principal component analysis (PCA) [73]. The main advantage of performing PCA on CNN features is that it can reduce the dimensionality of feature vectors without significantly losing the accuracy of the result. This effectively decreases the memory footprint and makes the usage of CNN features particularly attractive. As demonstrated in [74] and [75], compressing the original feature vectors of high dimensionality (e.g., 4096) to a lower-dimensional one (e.g., 128 dimensions) does not significantly decrease their performance for image retrieval. Razavian et al. [32] adopt PCA compression to reduce the length of CNN feature vectors (from 4096 to 500) and exploit whitening [71] processing subsequently. Experimental results for recognition tasks on several datasets show that the accuracy could increase by a significant margin by adopting such a post-processing procedure. Zhang et al. [38] post-process the raw CNN features in their CNN-based visual place recognition algorithm in the following manner: the obtained features are ℓ_2-normalized at first, then PCA dimensionality reduction and whitening are performed and the whitened vectors are ℓ_2-normalized again at the end. Using the augmented features to perform place recognition is able to achieve a higher accuracy compared with using the raw features directly.

4.2. Powerful Image Representations

Employing the output of fully connected layers or flattening the activations of convolutional layers as feature vector [15, 16] is the simplest but not the optimal way for image representation. Disadvantages such as high-dimensionality and the lack of generalization capability make such representations less competitive in practice. Recent research has developed a variety of methods to obtain more powerful representations for image matching. Table 2 lists the typical image representations built upon CNN models. We divide them into two categories: **global descriptors** that describe the whole image by a single feature vector and **regional descriptors** that focus on the regions of interest (e.g., landmarks) of an image and describe them individually.

4.2.1. Global Descriptors

For global descriptors, the whole image is fed into a CNN model. Typical operations like *pooling* or *aggregating* are often used on the raw feature maps to create more compact and discriminative global CNN features. To review these representations, we use notations from [77] to denote corresponding variables in this paper. An image I is fed into a CNN. The activations of a convolutional layer, consisting of K feature maps (channels) with size of W (width) and H (height), will be used to construct the feature vector, which is denoted by $\mathbf{f}$. Mathematically, each feature map,

Table 2: Typical image representations built upon convolutional neural network (CNN). Abbreviations stand for: Dimensionality (Dim.), Principal Component Analysis (PCA), Regions of Interest (ROI), Region Proposal Network (RPN) and Gaussian Random Projection (GRP).

Image representation	Dim.[a]	Commentary
Global Descriptor		
Raw CNN feature [15, 16]	Varies w.r.t. input image[b]	Simply stack activations of feature maps; High-dimensional
Augmented CNN feature [32, 38]	500 (depending on PCA)	Augmented with PCA, whitening, normalization, etc.
MAC[76, 77]	512 (fixed)	Perform max-pooling on each feature map; Compact
SPoC [75]	256 (fixed)	Perform sum-pooling on each feature map; Compact
CroW [78]	128/256/512 (fixed)	Use cross-dimensional weighting to weight the feature maps before sum-pooling; Dimensionality reduction with PCA
GeM [62]	256/512/2048 (fixed)	Propose a generalized pooling method; Dimensionality reduction with PCA
R-MAC [77]	256/512 (fixed)	Perform max-pooling on sub-regions of feature maps at different scales; Be sensitive to ROI of the original image
Deep Image Representation [79]	2048 (fixed)	Modify R-MAC by using a RPN and multi-resolution scheme
NetVLAD [39]	16 to 4096	Specifically designed for VPR; Dimensionality reduction with PCA
Regional Descriptor		
Landmark-Based [14, 80, 81, 57]	-	Identify ROI of the original image and describe them individually; More robust against changing environment
LDD [54]	1024	Encode spatial distribution of landmarks; Dimensionality reduction with GRP

[a] Fixed means that the dimensionality depends on the architecture of adopted CNN and is independent of an input image.
[b] This indicates the case that the raw feature are extracted from convolutional layers. When raw features are extracted from fully connected layers, the dimensionality will be fixed, depending on the CNN architecture.

represented by $\mathcal{X}_i, i = 1, 2, ..., K$, is a 2D tensor. We use Ω and $\mathcal{X}_i(p)$ to indicate the set of spatial locations of each individual 2D tensor and the activation at position p (where $p \in \Omega$) of this tensor, respectively.

MAC. Azizpour et al. [76] perform spatial max-pooling for each feature map when building the representation. The feature vector $\mathbf{f}$ can be described as $\mathbf{f} = [\mathrm{f}_1, ..., \mathrm{f}_i, ..., \mathrm{f}_K]^\top$, where $\mathrm{f}_i = \max_{p \in \Omega} \mathcal{X}_i(p)$. Experimental results on several source-target transfer learning datasets show that such a representation, which is named MAC (Maximum activation of convolutions) by Tolias et al. [77], is of great transferability for practical use.

SPoC. Babenko and Lempitsky [75] differentiate their image feature descriptor called SPoC (sum-pooled convolutional features) by using sum-pooling rather than max-pooling. In this case, the feature vector is given by $\mathbf{f} = [\mathrm{f}_1, ..., \mathrm{f}_i, ..., \mathrm{f}_K]^\top$, where $\mathrm{f}_i = \sum_{p \in \Omega} \mathcal{X}_i(p)$. According to [75], sum-pooling strategy leads to superior results compared with max-pooling, especially when post-processing such as whitening is performed on the feature vectors.

CroW. Instead of sum-pooling the activations directly, Kalantidis et al. [78] present a cross-dimensional weighting mechanism to produce weighted feature maps before sum-pooling is performed. That is, the activation $\mathcal{X}_i(p)$ in each feature map in SPoC is recalculated according to $\mathcal{X}'_i(p) = \alpha_i \beta_i \mathcal{X}_i(p)$, where α_i and β_i are the channel-wise (inter-channel) weight and location-wise (intra-channel) weight, respectively. The proposed representation, which is denoted by CroW (cross-dimensional weighting), achieves excellent performance on popular image search benchmarks [82].

GeM. Radenović et al. [62] propose a generalized pooling method, called generalized-mean (GeM) pooling to modify max-pooling and sum-pooling. It results in $\mathbf{f} = [\mathrm{f}_1, ..., \mathrm{f}_i, ..., \mathrm{f}_K]^\top$, where $\mathrm{f}_i = \left(\frac{1}{|\mathcal{X}_i|} \sum_{p \in \Omega} \mathcal{X}_i(p)^{m_i}\right)^{\frac{1}{m_i}}$ and m_i is the pooling parameter that can be set empirically or learned from training data. GeM is the generalized form of MAC [76] and SPoC [75]. When $m_i \to \infty$ and $m_i = 1$, it reduces to max-pooling and average pooling, respectively.

R-MAC. The representations described above are built upon the entire feature map, which means that they encode the whole image into a single vector. Tolias et al. [77] propose a method that focuses on sub-regions of the feature

map at different scales. Suppose that $\mathcal{R}$ is a sub-region of the feature map, i.e. $\mathcal{R} \subseteq \Omega$. Then the feature vector of this particular region is obtained in a similar way as MAC [76], i.e. $\mathbf{f}_{\mathcal{R}} = [\mathrm{f}_{\mathcal{R},1}, ..., \mathrm{f}_{\mathcal{R},i}, ..., \mathrm{f}_{\mathcal{R},K}]^\top$, where $\mathrm{f}_{\mathcal{R},i} = \max_{p \in \mathcal{R}} \mathcal{X}_i(p)$. Multiple feature vectors for corresponding sub-regions at different scales can be obtained in the same manner. Then the image representation is created as follows. First, each regional feature vector is ℓ_2-normalized, PCA-whitened and ℓ_2-normalized again. Then all post-processed region-level feature vectors belonging to one image are aggregated by summing them together and ℓ_2-normalized once again, producing a new compact feature vector. The proposed representation is referred to as R-MAC (regional maximum activation of convolutions). It has been used for VPR in [49], which demonstrates the excellent image representation ability on challenging datasets. Radenović et al. [83] also show that R-MAC outperforms MAC by a significant margin for image retrieval.

A crucial characteristic of aforementioned [62, 76, 77, 75, 78] representations is as follows. According to the definitions of these feature vectors, their dimensionality is equal to the number of feature maps (i.e. K) of the convolutional layer from which the representations are derived. Due to the fact that K is constant once the CNN model and the corresponding layer are determined, one can generally obtain a fixed-length feature vector for a reasonably sized image without the need of resizing or cropping it to a pre-defined size.

Deep image representation. The selection of regions for R-MAC is on the basis of a rigid grid of the CNN feature maps. One problem is that it treats all regions of a feature map with the same importance (Note that here we use "importance" to indicate that to what extent a region contributes to the similarity score of a pair of images). Gordo et al. [79] modify R-MAC by employing a region proposal network (RPN) [84] that is capable of focusing on more important (relevant) regions. In addition, they introduce a multi-resolution scheme. That is, an image is resized into multiple resolutions (three different resolutions are used in their work) and R-MAC features are extracted from each resized image independently. Then the obtained features are summed into a single vector to produce the final representation. These modifications lead to a big boost on the performance of the R-MAC representation.

NetVLAD. VLAD (Vector of Locally Aggregated Descriptors) aggregation scheme [70] is a prevalent descriptor aggregation method for hand-engineered features. Inspired by VLAD, Arandjelović et al. [39] design a novel architecture to mimic it. It regards the output (i.e. $K \times W \times H$ map) of a convolutional layer as $W \times H$ local descriptors with length of K and aggregates them with a specifically designed pooling layer to obtain the final representation. Experimental results in [39] show its superior performance on challenging VPR dataset such as Tokyo 24/7 [85].

4.2.2. Regional Descriptors

Note that, in this paper, we use regional descriptors to refer to those methods that feed relevant patches of an image (instead of the whole image as previously mentioned methods do) into a CNN and represent them individually, resulting in multiple feature vectors to represent an image. They focus on the regions of interest (e.g., *landmarks*) of an image. This is of great significance for image matching when the scene undergoes severe changes due to, for example, different viewpoints, changing weather or season and occlusions, which are not uncommon in practice for place recognition [44]. As the key step of the construction of regional descriptors, the region proposal (or object

proposal) methods can be divided into two categories according to [86]: **window scoring** methods such as Edge Boxes [87], YOLOv2 [88] and BING [89] and **grouping-based** methods such as Geodesic [90] and MCG [91].

Sünderhauf et al. [14] present a landmark-based visual place recognition method by combining object proposal techniques and CNN features. They use Edge Box [87] to detect potential landmarks within an image and then extract CNN features using AlexNet [18] for each detected landmark. They demonstrate that describing a scene based on regional landmarks rather than the whole image enables their method to gain more robustness against environmental changes caused by occlusions and viewpoint changes. Hou et al. [81] conduct a comprehensive evaluation of commonly used object proposal methods, including the region proposal network (RPN) [84] and the state-of-the-art CNN features for landmark-based place recognition. It is worth noting that Neubert et al. [80] present an evaluation criterion that measures the repeatability of region detectors to guide the selection of a proper region proposal method. In their later research [55], a novel region detector, coined SP-Grid (superpixel grid), is presented. SP-Grid generates landmarks based on a superpixel segmentation method [92] rather than the aforementioned object proposal methods and has been demonstrated to have a better performance, which is in agreement with Xin et al. [93].

Interestingly, Chen et al. [57] propose a method to detect salient landmarks and extract their features using one CNN directly. In particular, they use a high-level convolutional layer of a CNN (VGG16 [27] is employed in their study) to identify distinctive regions of landmarks, and then use another convolutional layer at a lower level of the same model to describe these regions. The main idea behind their method is that activations from the latter convolutional layers contain rich semantic information, which is helpful for identifying meaningful regions in an image. Similarly, the very recent work proposed by Maldonado-Ramírez and Torres-Mendez [94] adopts a visual attention model [95] to identify salient landmarks for place recognition in underwater environments.

Panphattarasap et al. [54] extend the method of [14] by further considering the spatial distribution of the landmarks within an image. On this basis, they propose the landmark distribution descriptor (LDD) to represent a place. Compared with global CNN features and the primal landmark-based method [14], LDD obtains higher precision on the datasets whose images change drastically in appearance due to different viewpoints. Xin et al. [93] propose to utilize superpixel segmentation techniques [92] to generate multi-scale landmarks. They also consider spatial distributions as well as scale distributions of landmarks when describing an image, which increases the robustness of their method. Yang et al. [96] present a multi-scale sliding window (MSW) scheme for landmark generation. In contrast to object-based methods, MSW performs better when the images undergo severe illumination and viewpoint variations.

4.3. Using Fine-tuned CNN Models or New Architectures

Although the methods that use pre-trained generic CNN models as described in Section 4.1 have achieved remarkable performance, researchers continue to develop novel methods by either fine-tuning pre-trained CNN models on specific place recognition datasets (or training from scratch if sufficient images are available) or designing and training new architectures to improve the performance of VPR methods. In these cases, the network is usually rearranged or designed particularly to ensure that the image representations can be generated in an end-to-end way, without the need

of further pooling or encoding procedures [97, 98].

As pointed out in [99], initializing a CNN model with pre-trained weights and re-training (which is also referred to as fine-tuning) it on specific training datasets leads to significant improvement of its adaptation ability. Gomez-Ojeda et al. [13] propose, for the first time, to fine-tune a particular CNN for place recognition with the target of learning appearance-invariant representations. They construct their CNN architecture by only keeping the first four convolutional layers of AlexNet [18] and replacing the following layers with a fully connected (FC) layer. The output (of 128 dimensions) of the last FC layer is regarded as image representation to recognize revisited places without additional post-processing. Chen et al. [59] develop a large-scale (containing 2.5 million images) dataset, which is coined as SPED, for place recognition particularly. On this basis, they fine-tune a network, named HybridNet (initialized with pre-trained AlexNet [18]) and train a network, named AMOSNet, from scratch respectively for recognizing places involving drastic viewpoint and condition changes. Radenović et al. [62] fine-tune CNNs in an automated manner. In particular, they utilize the structure-from-motion (SFM) [100] technique to help them to select high-quality training data, without the need of manual annotation.

In addition to resorting to off-the-shelf CNN architectures, developing specific models and training them for VPR also becomes active research topics. As mentioned in Section 4.2, Arandjelović et al. [39] design a trainable layer, called NetVLAD, to replace the last fully connected layer of the primal CNN model, resulting in an architecture that is able to be trained end-to-end for large-scale place recognition. Chen et al. [63] train an attention model that uses VGG16 [27] as the underlying model for VPR. The proposed multi-scale attention scheme is capable of generating regions of interest adaptively according to the image features, which improves the robustness of their method. With the similar goal, that is, focusing on the regions that contribute positively to the similarity of images, authors of [65] and [66] also introduce their specifically designed architectures for VPR. In particular, Kim et al. [65] introduce a novel network referred to as Contextual Re-weighting Network (CRN). The CRN is added after the conv5 layer of AlexNet [18] or VGG16 [27] to re-weight the original feature maps, producing the representation of an input image. By training it carefully, the presented network is able to produce spatially varying weights, which helps to focus on important regions. Xin et al. [66] design a landmark localization network (LLN), which predicts the discrimination of local features for the corresponding regions within an image, to help generate discriminative landmarks. Pre-trained ResNet101 [29] is utilized as the underlying network to produce feature maps, which is taken as the input of LLN. The LLN is then trained on Retrieval-SFM dataset [83] in an end-to-end manner. Most of the aforementioned methods belong to supervised learning, relying on labeled data. Merrill and Huang [64] construct a novel autoencoder-based CNN architecture with the objective of recognizing revisited places in changing environment. The specific architecture enables them to train the network in an unsupervised way.

In [101], authors formulate place recognition as a *distance metric learning* [102] problem. Metric learning concerns about learning a reasonable distance between two exemplars. In [101], off-the-shelf features extracted from Overfeat [33] are used to train a classifier to guide the recognition of same places. On the other hand, Zhao et al. [67] harness distance metric learning to obtain discriminative representations for VPR. They use SPED dataset [59]

to train their network, obtaining a self-adaptively enhanced similarity metric for the measurement of pairwise images.

It has been demonstrated that typical architectures such as *Siamese* [103] or *Triplet* [104] networks are more suitable for the training of metric learning models. Specifically, Siamese or Triplet networks employ the idea that the distance of a matching pair is smaller than that of a non-matching pair and build contrastive loss [105] or triplet loss [106, 107] to train the networks. Given a query image, Lopez-Antequera et al. [13] regard the images from the same location (but with different appearance) as matchings (positives) while those from different locations as non-matchings (negatives), obtaining a triplet dataset to train their network. In this fashion, they can learn appearance-invariant features for place recognition. Triplet architecture is also used in [58, 66, 67] to train networks for VPR purpose. On the other hand, Radenović et al. [62] adopt a Siamese architecture to train the networks for image matching.

4.4. Similarity Measure

After the image representations are obtained by one of the aforementioned approaches (see Sections 4.1 and 4.2), how to utilize them to measure image similarity for accurate place recognition becomes the main concern. A *similarity score*, which indicates how likely two images are from the same place, is necessary to enable the place recognition system to report the result. Usually, once the similarity score is larger than a pre-set threshold, the corresponding image will be reported as a candidate (a further verification is normally performed for the final decision as discussed in Section 3.1). To this end, a large number of similarity measures have been developed. We group them into two categories according to the type of CNN features based on which the similarity is computed: **global similarity** and **region- or landmark-based similarity**.

4.4.1. Global Similarity

Global similarity indicates that images are represented as single vectors as discussed in Section 4.2.1. In this case, similarity scores are calculated based on a type of distance metric. The most commonly used distance metric is Euclidean distance. Suppose a pair of images (I_a, I_b) are represented by global vectors $(\mathbf{f}_a, \mathbf{f}_b) \in \mathbb{R}^n$, the Euclidean distance is then given by $d_{ab} = \|\mathbf{f}_a - \mathbf{f}_b\|_2$. In [15, 34, 13, 39], authors use Euclidean distance directly to measure the dissimilarity of images. By contrast, Gomez-Ojeda et al. [13] propose to use a normalized Euclidean distance to construct a confusion matrix, whose rows and columns correspond to the database and query images, respectively. That is, $d^*_{ab} = d_{ab}/\max\{D(i, j)\}$, where $D(i, j)$ denotes the (i, j)th element of the confusion matrix. On this basis, Zhang et al. [38] define the similarity score as $S_{ab} = 1 - d^*_{ab}$ to ensure the values to lie in $[0, 1]$.

As discussed previously, most of the CNN representations are ℓ_2-normalized, making $\mathbf{f}$ to be a unit feature vector. This makes it convenient to compute the cosine similarity of two vectors, given by $s_{ab} = \mathbf{f}_a \cdot \mathbf{f}_b$. Cosine similarity has been widely used for image matching in VPR [45, 108]. In particular, to obtain more distinctive score values, Naseer et al. [108] perform normalization for the original cosine similarities over each column of the confusion matrix. A practical issue is that the computation of cosine similarity between many high-dimensional feature vectors (for instance, off-the-shelf features from intermediate convolutional layers) is expensive. To tackle this issue and ensure

the efficiency of the proposed method, Sünderhauf et al. [16] employ locality sensitive hash (LSH) function [109] to hash the original high-dimensional vectors to obtain relatively short-bit vectors, which will be used to compute Hamming distance to approximate the cosine similarity of original vectors. By doing so, they significantly speed up the computation of similarities (compressing vectors by 99.6%) without losing much performance (95% is retained), which is in agreement with Lowry and Andreasson [110]. Wu et al. [111] propose a method named similar hierarchy deep supervised hashing, which is able to achieve real-time speed on CPU, to deal with place recognition under the circumstance of severe illumination or viewpoint changes.

4.4.2. Region- or landmark-based Similarity

Compared with the case of global similarity, the computation for region- or landmark-based similarity is a bit more complicated. This class of methods need to firstly recognize matched landmarks from different images. Note that *cross-check* is normally applied to guarantee that only mutually matched landmarks are selected. For example, for nearest neighbor search, if landmark l_i^a from image I_a is the nearest neighbor of landmark l_j^b from image I_b and l_j^b is the nearest neighbor of l_i^a as well, then l_i^a and l_j^b can be regarded as a pair of matched landmarks. Then the calculation of similarity between two images can be roughly divided into two steps: (i) computing the similarity of each pair of matched landmarks and (ii) calculating the overall similarity of two images.

Sünderhauf et al. [14] develop a unified framework, which is referred to as *QUT framework* by Hou et al. [81] later. It measures the similarity of different images for landmark-based place recognition. Specifically, suppose l_i^a from image I_a and l_j^b from image I_b are a pair of matched landmarks, they first define the shape similarity as

$$l_{ij} = \exp\left(\frac{1}{2}\left(\frac{\left|w_i - w_j\right|}{\max\left(w_i, w_j\right)} + \frac{\left|h_i - h_j\right|}{\max\left(h_i, h_j\right)}\right)\right), \tag{1}$$

where (w_i, h_i) and (w_j, h_j) are the size of the bounding boxes of matched landmarks. Then the overall similarity score between images I_a and I_b is calculated as follows,

$$S_{ab} = \frac{1}{\sqrt{n_a \cdot n_b}} \sum_{i,j} 1 - \left(l_{ij} \cdot d_{ij}\right), \tag{2}$$

where n_a and n_b are the number of landmarks extracted from each image and $d_{ij} = 1 - \mathbf{f}_i^a \cdot \mathbf{f}_j^b$ is the cosine distance between l_i^a and l_j^b. By considering the shape similarity, they penalize mismatched landmarks (false positives) that have similar feature vectors but are significantly different in size, which is proved to be helpful to improve the performance of VPR [14]. *QUT framework* as well as its variants have been widely used for landmark-based VPR [54, 81, 93].

Instead of using the shape similarity as Eq. (1), Xin et al. [66] propose to weight similarities using a weighting scheme based on the spatial distribution of landmarks. In particular, they first analyze coordinate (of the center of a landmark) differences of all matched landmark pairs to obtain the coordinate differences with the highest frequency, denoted as d_x and d_y for x and y directions, respectively. On this basis, the weight is calculated as $w_{ij} = \exp(-\frac{1}{2}(((x_i^a -$

Table 3: Evaluation criteria for visual place recognition (VPR) methods. Abbreviations refer to: precision (*pre.*), recall (*rec.*), True Positive (TP), False Positive (FP), False Negative (FN), Average Precision (AP) and mean Average Precision (mAP).

Criterion	Definition	Used in VPR[a]
precision-recall curve	$pre. = \frac{TP}{TP+FP}$; $rec. = \frac{TP}{TP+FN}$	[15, 34, 14]
rec.@100%pre.	-	[49, 112]
rec.@N	-	[39, 65, 85]
AP or mAP	$AP = \sum_{k=1}^{n} pre.^k(rec.^k - rec.^{k-1})$	[49, 81, 113]
F_1 score	$F_1 = \frac{2 \cdot pre. \cdot rec.}{pre. + rec.}$	[16, 58, 45]

[a] Only some representative references are listed.

$x_j^b) - d_x)^2 + ((y_i^a - y_j^b) - d_y)^2))$, where (x_i^a, y_i^a) and (x_j^b, y_j^b) are the coordinates of the center of the landmarks l_i^a and l_j^b, respectively. The weighted similarities of all matched landmark pairs are then summed to obtain the final similarity of two images, given by $S_{ab} = \sum_{i,j} w_{ij} s_{ij}$, where s_{ij} is the cosine similarity between l_i^a and l_j^b.

Chen et al. [57] train a vocabulary by clustering $K \times M$ feature vectors of landmarks into N different visual words, where K is the number of images in a training dataset and M is the number of regions (landmarks) of each image. A weight is assigned to each word as $w_c = \log_{10}(K/n_c), c = 1, ..., N$, where n_c is the number of images that contain word c. The similarity of I_a and I_b is then computed according to $S_{ab} = \frac{1}{M} \sum_{i,j} w_i w_j s_{ij}, \{i, j\} = 1, ..., M$, where w_i and w_j are the weights of visual words that landmarks l_i^a and l_j^b belong to, respectively, and s_{ij} is the cosine similarity of them.

4.5. Evaluation criteria

Historically, *precision-recall curve* is commonly used to evaluate place recognition methods [3]. Precision, denoted as *pre.* in this paper, is defined as the ratio of the number of true positives to the number of reported positives, while recall, denoted as *rec.*, refers to the ratio of the number of true positives to the total number of ground truth in a database. By varying the threshold, based on which the recognition candidates are selected, multiple precision-recall pairs can be obtained to plot a precision-recall curve. Since a VPR system is expected to report recognition results without false positives, Garcia-Fidalgo and Ortiz [112] and Zhang et al. [49] use recall at 100% of precision, denoted as *rec.@100%pre.*, as an indicator of the performance. Another prevalent criterion, which is referred to as *rec.@N* [39, 65], computes the recall rate when top N retrieved database images are given. By varying the value of N, one can plot a curve similar to the precision-recall curve.

Precision-recall curve illustrates the recognition results with respect to different thresholds intuitively, while other criteria are able to measure the overall place recognition performance using a scalar solely. For example, *average precision* (AP) computes the mean of precision over all recall rates. Intuitively, AP can be considered as the area under the precision-recall curve (AUC). Therefore, a larger AP indicates a better performance. Considering that there are usually multiple query images for a database, *mean average precision* (mAP) can be calculated by averaging all individual APs. In [49, 81], AP or mAP is used to evaluate the overall performance of a VPR method. By contrast,

Table 4: Comparison of runtime performance for representative VPR methods. "Para." stands for the number of parameters of the network. "DB Scale" stands for the number of images of the database on which the method is evaluated. "Avg. Time" stands for average processing time for a single image. "Fea. Ext." and "Img Mat." indicate the time for feature extraction and image matching, respectively. Runtime is given in Second except those marked with "ms" (milliseconds).

Method	Architecture	Para.	Feature	DB Scale	GPU	Avg. Time (in Second)		
						Fea. Ext.	Img Mat.	Total
[9] (2011)	-	-	handcrafted	20,862	✗ (Core i7-4790K CPU)	-	0.039	-
[68] (2012)	-	-	handcrafted	20,862	✗ (Core i7-4790K CPU)	-	0.251	-
[15] (2014)	LeNet [53]	60K	raw CNN feature	4,789	✗ (standard CPU)	0.1	0.3	0.40
[16] (2015)	AlexNet [18]	60M	raw CNN feature	100,000	NVIDIA Quadro K4000	0.015	35.7	35.715
			raw CNN feature+hashing			0.015	0.189	0.204
[39] (2016)	VGG16 [27]	138M	learned feature	20,862	GeForce GTX TITAN X	-	0.137	-
[57] (2017)[a]	VGG16 [27]	138M	regional landmark-based	1,000	NVIDIA Titan X Pascal	0.4084	0.007	0.4154
[81] (2018)[a]	AlexNet [18]	60M	regional landmark-based	-	GeForce GTX TITAN X	0.171	-	-
[64] (2018)	Task Specific	9.8M	learned feature	4,541	GeForce GTX 960M	0.862 ms	1.47 ms	2.332 ms
[114] (2020)[a]	AlexNet365	60M	Region-VLAD	1,125	NVIDIA P100	0.41	0.12 ms	0.41
[115] (2020)	FlyNet+CANN	72K	learned feature	1,000	-	0.035	0.025	0.06

[a] For the methods where features are built based on regional landmark, the *Fea. Ext.* includes the time used for region proposal, feature extraction and encoding. The timing results of methods of FAB-MAP [9], SeqSLAM [68] and NetVLAD [39] are quoted from [116].

Sünderhauf et al. [16] use F_1 *score* to compare the performance of their proposed method with other state-of-the-art. Table 3 lists the commonly used evaluation criteria for VPR methods.

4.6. Runtime Performance

Runtime of VPR algorithms is one of the main considerations, especially for mobile robots which normally need to meet real-time requirement. Compared with handcrafted feature-based methods, DL-based ones are more time-consuming and typically deployed on dedicated GPUs [115]. This can be found in Table 4, which summarizes the runtime performance of typical methods. Table 4 also lists the number of parameters of the employed embedding networks to provide an intuitive comparison regarding the memory complexity when implementing these methods. Increasing the number of parameters normally means the cost of more memory. Typical CNN architectures such as VGG16 [27] have a relatively large number (138 million) of parameters because they contain more layers. Chancán et al. [115] develop a lightweight and compact architecture, which has only three layers and far few parameters (72K), to guarantee a smaller computational footprint for visual place recognition.

Normally, the runtime of a CNN-based VPR algorithm mainly consists of two parts: feature extraction time and image matching time. For feature extraction time, it is highly depends on the complexity of CNN models and the hardware where the algorithm is deployed. Lightweight and compact neural networks [64, 114] can be used for feature extraction to achieve low computational cost. For example, Merrill and Huang [64] propose a lightweight neural network built upon autoencoder for feature embedding for VPR. Their model achieves a faster speed (0.862 ms) for feature extraction of an image while this costs AlexNet [18] 16.6 ms on the same GPU. For image matching time, it is proportional to the total number of images stored in a database. The length of a CNN feature vector has its

impact as well, especially when the database is large. Carefully designed algorithms can improve the image matching efficiency significantly. As mentioned in Section 4.4.1, Sünderhauf et al. [16] propose to employ hashing approach [109] to avoid directly computing the cosine similarity of high-dimensional feature vectors. By doing so, the average time for image matching on a database with 100K images is reduced from 35.7 seconds to 0.189 second as shown in Table 4. This enables their method to perform real-time place recognition. In general, for DL-based VPR methods, researchers pay more attention to the improvement of recognition accuracy. Comparatively, computational complexity may be overlooked more or less. However, it is of great significance for practical applications. It is still an open issue and more explorations are expected for the development of VPR approaches with low computational cost.

5. Place Recognition Datasets and Challenging Issues

In this section, we introduce the commonly used datasets and discuss the challenging issues in visual place recognition. The performance of typical methods on these datasets is also described. High-quality datasets are important for the development of VPR because they not only provide benchmark for the evaluation of novel methods, but also make it possible to train deep learning models. Researchers have developed a large number of datasets, including the generic ones used for multiple purposes and the specific ones used for challenging topics. In Table 5, we provide an exhaustive investigation of existing datasets that are publicly available and can be used for VPR.

5.1. Generic and Long-term Datasets

Generic. The New College and City Centre dataset, which is initially proposed by Cummins and Newman [8] to evaluate their appearance-based place recognition method (FAB-MAP) initially, is one of the most famous and commonly used VPR datasets. The images in this dataset undergo viewpoint changes and slight illumination changes. In addition, it provides manually labeled ground truth that indicates the true loop closures for each image as well as GPS (global positioning system) coordinates. Therefore, it has become a benchmark, as used in [8, 34, 49, 112], to measure the robustness of VPR methods against appearance changes. Another dataset derived by Smith et al. [117] later is also referred to as New College. In addition to containing more images, it also records laser and IMU (inertial measurement unit) data as well as omni-directional imagery for multiple usages. Galvez-LóPez and Tardos [7] evaluate their handcrafted feature-based method, coined DBoW2, with this dataset, obtaining 55.92% recall rate at 100% of precision. By contrast, Naseer et al. [108] obtain a better performance (79% rec.@100%pre.) by using CNN features and a novel data association approach on the same dataset.

A number of datasets have been created thanks to the development of autonomous mobile robot. Rawseeds dataset built by Bonarini et al. [120] contains image sequences gathered from both indoor and outdoor environments. It is initially developed for localization and mapping problems of autonomous robot and includes high-quality data sets in multiple types, such as LiDAR data and omni-directional images. Similarly, Ford Campus dataset [121] and Malaga dataset [123] are developed for the study of SLAM techniques at the beginning. All these datasets contain loop

Table 5: Datasets used in visual place recognition (VPR). Abbreviations refer to: Localization or pose ground truth (GT), Information (Info.), color images (RGB), color and depth images (RGB-D), grayscale images (Gray.), Global Positioning System (GPS), Inertial Measurement Unit (IMU) and Light Detection and Ranging (LiDAR). The terms of "Illumination", "Viewpoint" and "Labels" marked with ✓mean that the images in a dataset are with illumination changes, viewpoint changes and human-annotated labels, respectively.

Topic	Name	Image type	Environment	Illumination	Viewpoint	GT	Labels	Extra Info.	Used in VPR
Generic	New College and City Centre[1] [8]	RGB	Outdoor	slight	✓	✓	✓	GPS	[8, 34, 49, 112]
	New College Vision and Laser[2] [117]	Gray.	Outdoor	slight	✓	✓		GPS, IMU, LiDAR	[108, 7, 118, 119]
	Rawseeds[3] [120]	RGB	Indoor/ Outdoor		✓	✓		GPS, LiDAR	[7, 119]
	Ford Campus[4] [121]	RGB	Urban	slight		✓		GPS, IMU, LiDAR	[7, 122]
	Malaga Parking 6L[5] [123]	RGB	Outdoor			✓		GPS, IMU, LiDAR	[35, 7]
	KITTI Odometry[6] [52]	Gray./ RGB	Urban	slight		✓		GPS, IMU, LiDAR	[112, 49, 13, 124]
Long-term	St. Lucia[7] [125]	RGB	Urban	✓	slight			GPS	[16, 112, 81]
	COLD[8] [126]	RGB	Indoor	✓	✓	✓	✓	LiDAR	[127, 128, 129]
	Oxford RobotCar[9] [130]	RGB	Urban	✓		✓		GPS, IMU, LiDAR	[131, 37, 113, 132]
	Gardens Point Walking[10] [133]	RGB	Indoor/ Outdoor	✓	✓			-	[16, 57, 14, 93]
	MSLS[31] [134]	RGB	Urban	✓	✓	✓		GPS	-
Across seasons	Nurburgring and Alderley[11] [68]	RGB	Urban	✓		✓	✓	-	[68, 13]
	Nordland[12] [135]	RGB	Outdoor	✓		✓		GPS	[108, 16, 136, 57]
	CMU[13] [137]	RGB	Urban	✓	✓	✓		GPS	[93, 113, 138]
	Freiburg (FAS)[14] [44]	RGB	Urban	✓		✓	✓	GPS	[108, 44, 136]
	VPRiCE[15] [139]	RGB	Outdoor	✓	✓			-	[45, 136, 113]
RGB-D	TUM RGB-D[16] [140]	RGB-D	Indoor		✓	✓		IMU	[141, 142, 143]
	Microsoft 7-Scenes[17] [144]	RGB-D	Indoor		✓	✓	✓	-	[145, 146]
	ICL-NUIM[18] [147]	RGB-D	Indoor		✓	✓		-	[148, 142]
Semantic	KITTI Semantic[19] [149]	RGB	Urban			✓	✓	GPS, IMU, LiDAR	[150]
	Cityscapes[20] [151]	RGB	Urban			✓	✓	GPS	[60, 138]
	CSC[21] [152]	RGB	Outdoor	✓		✓		LiDAR	-
Train networks	Cambridge Landmarks[22] [145]	RGB	Outdoor	✓	✓	✓	✓	-	[145]
	Pittsburgh250k[23] [153]	RGB	Urban	✓	✓	✓	✓	GPS	[153, 39, 154]
	Tokyo 24/7[24] [85]	RGB	Urban	✓	✓	✓		GPS	[85, 39]
	SPED[25] [59]	RGB	Outdoor	✓	✓			-	[59, 63, 155]
Omni-directional	New College Vision and Laser[2] [117]	Gray.	Outdoor	slight	✓	✓		GPS, IMU, LiDAR	[108, 7, 118, 119]
	MOLP[26] [156]	Gray./D	Outdoor	✓		✓		GPS	[156]
	NCLT[27] [157]	RGB	Outdoor	✓	✓	✓		GPS, LiDAR	[158, 159]
Aerial/UAV	Shopping Street 1/2[28] [160]	Gray.	Urban	slight	✓	✓		-	[160, 161]
	EuRoC[29] [162]	Gray.	Indoor		✓	✓		IMU	[160]
Underwater	UWSim[30] [163]	RGB	Underwater			✓		GPS	[94]
Range sensors	MulRan[32] [164]	3D Point clouds	Urban	✓		✓		LiDAR, RADAR	-

[1] http://www.robots.ox.ac.uk/~mobile/IJRR_2008_Dataset/data.html [2] https://ori.ox.ac.uk/older-projects/new-college-dataset/
[3] http://www.rawseeds.org/home/ [4] http://robots.engin.umich.edu/SoftwareData/Ford
[5] https://www.mrpt.org/malaga_dataset_2009 [6] http://www.cvlibs.net/datasets/kitti/eval_odometry.php
[7] https://wiki.qut.edu.au/display/cyphy/St+Lucia+Multiple+Times+of+Day [8] https://www.nada.kth.se/cas/COLD/
[9] https://robotcar-dataset.robots.ox.ac.uk/ [10] https://goo.gl/tqmWyq
[11] https://wiki.qut.edu.au/display/cyphy/Michael+Milford+Datasets+and+Downloads [12] https://nrkbeta.no/2013/01/15/
[13] http://3dvis.ri.cmu.edu/data-sets/localization/ [14] http://aisdatasets.informatik.uni-freiburg.de/freiburg_across_seasons/
[15] https://goo.gl/R0QYU2 [16] https://vision.in.tum.de/data/datasets/rgbd-dataset
[17] https://www.microsoft.com/en-us/research/project/rgb-d-dataset-7-scenes/
[18] https://www.doc.ic.ac.uk/~ahanda/VaFRIC/iclnuim.html [19] http://www.cvlibs.net/datasets/kitti/eval_semantics.php
[20] https://www.cityscapes-dataset.com/ [21] https://www.visuallocalization.net/datasets/
[22] http://mi.eng.cam.ac.uk/projects/relocalisation/#dataset [23] http://www.ok.ctrl.titech.ac.jp/~torii/project/reptile/
[24] http://www.ok.ctrl.titech.ac.jp/~torii/project/247/ [25] https://goo.gl/OXeL2X [26] http://hcr.mines.edu/code/MOLP.html
[27] http://robots.engin.umich.edu/nclt/ [28] http://www.v4rl.ethz.ch/research/datasets-code.htmls
[29] https://projects.asl.ethz.ch/datasets/doku.php?id=kmavvisualinertialdataset [30] https://goo.gl/GtMQkv
[31] https://www.mapillary.com/dataset/places [32] https://sites.google.com/view/mulran-pr

closures within a sequence more or less, which makes them descent choices for the evaluation of place recognition algorithms, either conventional handcrafted feature-based [7] or CNN feature-based [35].

The KITTI odometry benchmark suite proposed by Geiger et al. [52] is one of the most well-known datasets due to its multiple application scenarios (e.g., visual odometry and monocular/stereo SLAM). KITTI odometry is composed of 22 sequences totally and 12 of them contain loop closures. The images are gathered in urban dynamic environment, including moving cars and pedestrians (See Fig. 5 for examples). To evaluate the proposed hierarchical place recognition method named HTMap, Garcia-Fidalgo and Ortiz [112] conduct experiments on KITTI dataset, obtaining 90.24% recall at 100% of precision on KITTI 00 sequence. By comparison, Zhang et al. [49] use R-MAC feature and employ diffusion process [51] to boost data association for place recognition, obtaining 95.37% recall at 100% of precision on the same sequence.

Long-term. For long-term place recognition in outdoor environment, illumination is a crucial factor because the appearance of the same place may change drastically under different illumination conditions. A dataset called St. Lucia Multiple Times of Day, created by Glover et al. [125] has been widely used by researchers [16, 155] to evaluate the robustness of their methods to illumination changes. This dataset consists of 10 sequences, which are collected along the same route but at different time during sunny days — from early morning to late afternoon. The images from different sequences are apparently different even though they are from the same place. Pronobis and Caputo [126] present a dataset, named COLD, that contains sequences collected at different time of a day (e.g., day and night) or under different weather (e.g., sunny and cloudy day). This dataset is collected in indoor environment such as laboratories and corridors and each image is annotated with a semantic label, which is very convenient for researchers to use to study semantic place recognition [128, 129].

Recently, Maddern et al. [130] present a new dataset called Oxford RobotCar. This dataset comprises over 100 sequences by traversing through the same route of Oxford city repetitively with a robot car platform over a period of over a year. As a result, it contains extensive images (collected at the same place) that are apparently different due to various factors such as different weathers, illuminations, seasons as well as traffic/road conditions. Apart from this, it provides 3D LiDAR data, GPS and IMU ground truth. Therefore, it is extensively used for multiple research topics in VPR, for example, PointNetVLAD [131] based on LiDAR data, semantic place recognition [37, 36] and the very recent work regarding biologically-inspired place recognition[113].

5.2. Specific Datasets and Challenging Issues

5.2.1. Across Seasons

Recognizing places across seasons is attracting widespread interest in recent years. It is an essential and challenging topic for long-term place recognition and has boosted the development of VPR a lot by producing many appearance-invariant recognition methods [13, 37, 59, 110, 165]. Remarkable datasets related to this topic play an important role in this progress. A key aspect of this class of datasets is that there are no loops within an individual

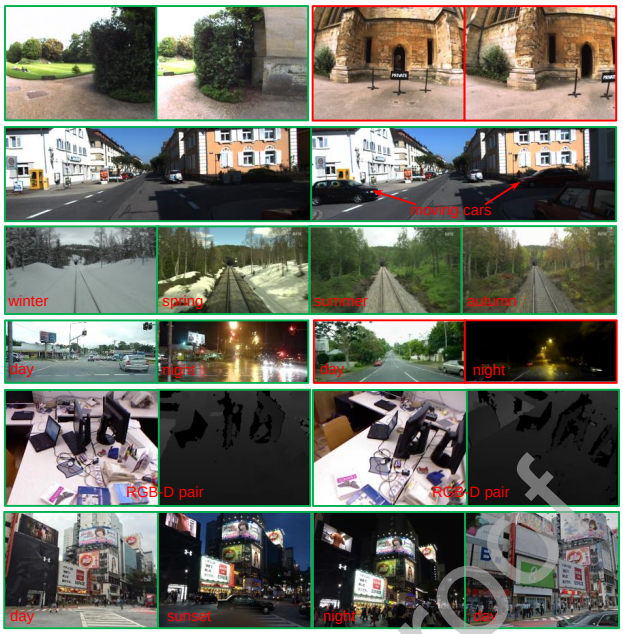

Figure 5: Image examples in VPR datasets. For each row, the images bordered with the same color (green or red) are collected from the same place. From top to bottom, the examples are from: New College [8], KITTI odometry (dynamic environments, moving cars or pedestrians) [52], Nordland (across seasons) [135], Alderley (sunny day-rainy night) [68], TUM RGB-D [140] and Tokyo 24/7 (day-night) [85], respectively.

image sequence in most cases. Instead, they are normally composed of multiple sequences collected along the same route but at different time or season.

Nurburgring and Alderley datasets [68] are two of the earliest datasets presented for image matching under severely changing environment. They are created by Milford and Wyeth [68] originally to evaluate the performance of SeqSLAM, which is a visual navigation approach based on sequence matching. Nurburgring dataset is gathered at two different times of a year while Alderley is collected during a sunny morning and a heavy rain night, respectively, leading to severe appearance changes for images (See Fig. 5 for examples). Researchers [68, 13] use Alderley dataset to evaluate the robustness of visual place recognition approaches with respect to condition changes. The Nordland dataset proposed by Sünderhauf [135] comprises four sequences recorded in winter, spring, fall and summer, respectively, along a 728-km long train journey, presenting massive images undergoing drastic appearance changes, as illustrated in Fig. 5. The dataset is manually synchronized to ensure that images (in different sequences) with the same indices are from the same place, which is convenient for researchers to use. An early dataset dubbed CMU (Carnegie Mellon University) [137], focusing on long-term vision-based localization in outdoor environment, also consists of images collected along the same route but during different seasons.

In the following years, a series of publicly available datasets such as Freiburg Across Seasons (FAS) [44], VPRiCE [139] as well as the aforementioned Oxford RobotCar [130] are created. All these datasets, in one way another, provide video frames with large variability in appearance. The VPRiCE (Visual Place Recognition in Changing Environments) challenge 2015 dataset proposed by Sünderhauf [139] aims to encourage researchers from various fields like robotics,

computer vision and visual neuroscience to tackle challenging VPR problems.

5.2.2. RGB-D Datasets

Depth images provided by RGB-D cameras are not vulnerable to illumination changes, which is significant for retrieving images in changing environment. RGB-D cameras such as Kinect become popular for robot localization [146, 166] and pattern recognition [167] recently. Here we focus on the RGB-D datasets that can be employed for place recognition. Detailed discussions about RGB-D image-based VPR methods will be presented in Section 6.3.2.

TUM RGB-D [140] is a prevalent dataset that contains color-depth image pairs and has been used to evaluate various place recognition approaches [141, 142, 168]. It consists of numerous sequences with loop closures, captured with Kinect camera in indoor environment. For some of the sequences, the camera is hand-held, so the images undergo strong viewpoint changes. The Microsoft 7-Scenes [144] and ICL-NUIM [147] are another two RGB-D datasets created for camera relocalization task, which shares significant similarity with place recognition [169]. Microsoft 7-Scenes dataset is designed by Shotton et al. [144] for indoor relocalization with RGB-D images. It comprises 17,000 query images and 26,000 database images collected from seven different scenes. It is employed in [145, 146] as a benchmark to evaluate the methods for determining poses of a visual acquisition system. The synthetic dataset called ICL-NUIM [147] introduced by Handa et al. aims at benchmarking SLAM and place recognition algorithms. Sizikova et al. [148] conduct experiments on ICL-NUIM to show the feasibility of their method that jointly leverages color and depth images to tackle place recognition.

5.2.3. For Network Training

The performance of deep neural networks (either fine-tuned or trained from scratch) crucially depends on a high-quality dataset that contains sufficient images for training. Arandjelović et al. [39] employ two large datasets namely Pittsburgh250k [153] and Tokyo 24/7 [85] to train and test the specifically designed model NetVLAD for place recognition. Pittsburgh250k [153] contains 250k database images and 24k query images with geotagged labels generated from Google Street View. It is a challenging place recognition dataset because the images are collected under different illumination conditions (in different seasons in some cases) with multiple view directions. Tokyo 24/7 [85] dataset is composed of 76k database images and 1,125 query images gathered with smartphones from 125 urban places. For each place, the images are captured at three different times (day, sunset and night) during a day and from three different viewpoints. As a consequence, the appearance of images varies significantly (See Fig. 5 for examples).

A relevant dataset called Cambridge Landmarks is created by Kendall et al. [145] to train their PoseNet model, designed for camera localization. The dataset contains 12,000 images obtained at five outdoor scenes and each image is labeled with its 6-DOF camera pose. In view of the absence of large-scale datasets for the training of deep neural networks for place recognition, Chen et al. [59] develop a rather large dataset, named SPED (Specific PlacEs Dataset). It contains 2.5 million images gathered from over one thousand of places with hundreds of images for each place. The images encounter changing conditions, including weather, seasons and day-night, which is very helpful to learn

condition-invariant representations. SPED has been increasingly used in recent years, as employed in [59, 63, 155] for training large models specifically designed for place recognition.

Finally, miscellaneous datasets as listed in Table 5 also play an important role in enriching the research of visual place recognition. These datasets are created for all kinds of specific purposes. For example, MOLP [156] is presented for omni-directional image-based place recognition. The Shopping Street dataset [160] and EuRoC dataset [162] can be used for UAV (Unmanned Aerial Vehicle) related studies [161], while UWSim [163] is proposed for recognizing places in underwater environment as used in [94]. Others like KITTI Semantic [149], COLD [126], Cityscapes [151] and CSC [152] datasets can be used for semantic place recognition, which will be discussed in Section 6.2.

6. New Tools and Open Issues

In this section, we describe the new tools in the domain of deep learning that have been employed for visual place recognition, for example, generative adversarial networks (GANs), semantic scene understanding, multi-modality feature learning, domain adaptation/transfer learning, and so forth. Although some of these techniques may not be in the research focus of the community of VPR, relevant studies have demonstrated their effectiveness and advantages for tackling the VPR problem.

6.1. Beyond Convolutional Neural Networks

6.1.1. AutoEncoder (AE)

CNNs have been extensively studied for place recognition as discussed in Section 4. One handicap is that a huge number of labeled data is needed when training the CNN models. Taking advantages of autoencoder (AE), one can train models in an unsupervised manner. The AE adopts an encoder-decoder paradigm, that is, it first encodes the input to a compressed representation and then maps it to a reconstruction of the original input. During the encoder-decoder stage, the intrinsic structure of input data can be discovered in order to learn representations (output of encoder).

Gao and Zhang [141] propose an autoencoder-based loop closure detection method. In particular, they leverage a modified stacked denoising autoencoder (SDA), which is a variant of AE, to learn representations from corrupted image data. The proposed network can be trained in an unsupervised way. Experimental results on the New College and City Centre dataset [8] and the TUM RGB-D dataset [140] show that their method achieves comparable performance to FAB-MAP 2.0 [9]. However, their network cannot learn condition-invariant features. Taking this into consideration, Merrill and Huang [64] propose to utilize a generalized viewpoint alteration approach to learn condition-invariant features when training the networks. In addition, instead of using a fully connected AE for the encoder and decoder stage [141], they adopt deconvolution and unpooling layers at the decoder stage. However, the network used in [64] cannot capture 2D structures and hierarchical features of images due to the fully connected encoder. By contrast, a further study presented by Maldonado-Ramírez and Torres-Mendez [94] utilizes a convolutional autoencoder (CAE) [170] to learn features of salient landmarks extracted with a visual attention system to solve place recognition. The

encoder part of a CAE is same as a standard CNN, which enables it to deal with 2D images properly, while the decode part is reversed where up-sampling layers, instead of max-pooling layers, are adopted. Once the network is trained, the encoder can be considered as a feature extractor. Similarly, Mukherjee et al. [124] propose a locally connected autoencoder (LCA) architecture to learn representations for loop closure detection in SLAM. More recently, Wang et al. [171] propose a network, named graph-regularization stacked denoising auto-encoder (G-SDAE), jointly uses autoencoder and manifold learning [172]. It not only learns discriminative image representations, but can also capture the local geometry structure of the embedded feature space, which helps them achieve better accuracy when performing place recognition.

6.1.2. Generative Adversarial Networks (GANs).

In the last five years, generative adversarial networks (GANs) [173] have been successfully used for a multitude of tasks, especially for *domain adaptation* or *transfer learning* [174]. A GAN consists of two main parts: a generator and a discriminator. By simultaneously training them carefully, the generator is able to learn specific object representations for major components of an image (e.g., windows and beds for indoor scenes) while the discriminator learns features of the domain that are useful for recognition tasks [175].

Latif et al. [176] propose a GAN-based method for across season place recognition. Based on the observation that images of the same place collected at different seasons own different distributions, which can be considered as from different domains, they formulate VPR as a domain translation problem [177]. In their study, for example, the images collected in summer are regarded as source domain $\mathcal{S}$, while the ones collected in winter can be regarded as target domain $\mathcal{W}$. Following this idea, they train two coupled GANs to translate images from one domain (e.g., summer) to the other (e.g., winter) and vice versa. Then they compare the generated images (translated from domain $\mathcal{W}$) in domain $\mathcal{S}$ instead of the original images in domain $\mathcal{W}$ with the images in domain $\mathcal{S}$ to perform place recognition. By doing so, they obtain an improved recognition performance on challenging datasets such as Nordland [135].

With the similar idea, Porav et al. [132] train a cyclic GAN inspired by [178] to transfer image appearance from source domain (e.g., night) to target domain (e.g., day) for image matching under adverse conditions. Different the work in [176], they use convolutional architectures, instead of fully connected layers, for generator and discriminator. Experimental results on Oxford RobotCar [130] show that the localization error calculated on *real day* to *synthesized day* image matching is indeed smaller than that based on *real day* to *real night* image matching. In addition, Yin et al. [179] apply a GAN to improve the generalization ability of their networks for feature-to-data transformation for VPR.

Open Issues. Despite GANs have shown their advantages for tackling place recognition under severe appearance changes (for instance, across seasons and day-night), there are open issues to be studied. As indicated in [176], it is relatively easy to obtain the mapping between domains with regular changes (e.g., summer-winter and/or day-night). However, images may undergo various changes in practice. A method that is able to efficiently translate images with these changes from one domain to another domain is necessary for better performance and wider applications.

6.2. Semantic Information

Humans are capable of recognizing a previously visited place effortlessly even if they view it from an opposite direction. However, for computer, most of the state-of-the-art VPR systems are feature-based, which means that they determine revisited places largely based on the distance of corresponding features in a feature vector space. This is conducted without sufficiently utilizing semantic information, which has been proved useful for many tasks such as pattern recognition [180] and visual localization [142, 143]. Intuitively, semantic cues of a scene are not as vulnerable as pixel intensity to condition variations. As illustrated in [138], the appearance of color images of the same place changes a lot as the weather, illumination or season varies. By contrast, the corresponding semantic labels for each pixel obtained with semantic segmentation techniques show their overall invariance to the above condition changes.

Semantic cues have been utilized for place recognition in early years by Costante et al. [128], although they develop their method based on hand-engineered features. In particular, they consider to transfer knowledge learned from source domain to target domain. Semantic categories of the environment guide them to determine what and how much knowledge should be transfered. Cascianelli et al. [35] utilize high-level semantic features extracted from CNN to specify patches in an image to construct a covisibility graph in their landmark-based VPR method. However, they do not associate each patch to a specific object label. Therefore, they regard their approach as semi-semantic-based.

Recently, successes in the areas of semantic segmentation [152, 181] and semantic scene understanding [151, 182] rekindle research interest regarding semantic place recognition. Stenborg et al. [138] propose a robust place recognition approach for long-term visual localization. Naseer et al. [60] segment images based on Fast-Net [61] to obtain appearance-invariant objects such as buildings and employ features of these salient regions to form a robust descriptor. Garg et al. [37, 36] look deeper into semantic information for VPR in their serial work. In [37], they focus on a very challenging problem, i.e. recognizing a revisited place from the opposite view. For this purpose, they propose a novel descriptor coined Local Semantic Tensor (LoST) built on feature maps of RefineNet [183], which is a high-resolution semantic segmentation network designed to match images semantically. In [36], they improve this work by presenting a pipeline that simultaneously uses semantic information at three levels: database level for environment segmentation, image level for place matching and pixel level for final spatial-consistency check. Schönberger et al. [184] propose to learn an embedding by jointly utilizing high-level semantic information and 3D geometric information for visual localization under severe illumination, season and viewpoint changes. Finally, Hong et al. [185] propose, for the first time, to leverage *textual information* contained in the (normally urban) environment such as street signs and shop signage for VPR. They show the superiority of semantic information for place recognition when images undergo serious condition/viewpoint changes and occlusions.

6.3. Heterogeneous Data

Instead of resorting to one specific modality of visual information solely, an increasing number of studies in recent years have focused on utilizing heterogeneous imaging data [169] or combining multiple modalities of visual information together for pattern recognition [41, 167, 186, 187]. The core idea is that features derived from different

visual modalities or imaging data can provide complementary information, which can significantly improve recognition accuracy. This category of methods is known as *multi-modality feature fusion (or learning)*. Note that the term of *multi-modality* includes two cases: (i) multi-modality features for one type of image, for example, local binary pattern (LBP) [188] and histograms of oriented gradients (HOG) [6] of gray-scale images; and (ii) multiple types of imaging data such as RGB image, depth image and LiDAR data. Both of them are discussed below.

6.3.1. Multi-modality Feature Fusion

The significance of combining multiple modalities of visual information has been observed by researchers in visual place recognition. Costante et al. [128] combine two types of descriptors — SPACT [189] and SPMK [190] — for their transfer leaning-based semantic VPR approach. Experimental results in indoor environment show that, by leveraging feature combination, the recognition accuracy increases by 20.63% and 4.57% compared with using the SPACT or the SPMK solely, respectively. Qiao and Zhang [191] propose a multi-modality feature fusion approach, where the HOG [6] and LBP [188] features of gray-scale images and the LBP feature of the disparity map of stereo images are fused for recognizing places.

Seymour et al. [155] present a framework named SAANE (Semantically-Aware Attentive Neural Embeddings) for robust visual localization under severe appearance changes. In particular, they train an attention model that is able to focus on discriminative regions in an image by jointly using appearance and semantic representations. The modality fusion module and spatial pooling module of the framework are used to fuse representations from two modalities and generate the final representations of an input image, respectively. Yin et al. [179] introduce a Multi-Domain Feature Learning (MDFL) approach to the selection of condition-invariant features of an image. Specifically, they first employ CapsuleNet [192] to extract multi-domain features, including condition-related and condition-invariant features. After that, a feature separation module, including a decoder module, a discriminator module and two reconstruction modules, is applied to guide the selection of condition-invariant features for robust place recognition.

6.3.2. Depth Information

Aforementioned studies utilize multi-modality features of one type of image, whereas others employ auxiliary imaging data when recognizing places. Depth information, collected with RGB-D cameras, is one of the commonly used information for robot localization or navigation [146, 166, 193]. Researchers find that depth images can provide structural information of environment, which benefits various pattern recognition tasks such as object recognition [194, 195] and scene understanding [167, 196]. Wang et al. [194] design deep CNN layers for color modality and depth modality, respectively, and then connect them with proposed multi-modal layers to fuse color and depth information. To fully exploit RGB-D information, Li et al. [167] present a unified framework called MAPNet to capture the intra-modality and cross-modality patterns for scene classification.

Depth information is invariant to illumination changes, which is an appealing property for place recognition. Sizikova et al. [148] propose to leverage depth information to compensate for the ambiguous color information

induced by illumination changes when performing place recognition. Specifically, they adopt two CNN models to process RGB image and depth image separately. Two-modality features are combined to obtain the final descriptor. It is worth noting that they train the depth Siamese CNN using synthetic depth images generated with synthetic 3D models [197] and then test it on an unseen dataset from the real world such as TUM RGB-D [140]. In terms of color images, they employ AlexNet [39] pre-trained on Places Dataset [198] for feature extraction. Li et al. [146] show the advantage of depth information for the relocalization task when color images are not available due to, for example, night time or dimly indoor environment.

Open Issues. While the studies described in Sections 6.3.1 and 6.3.2 fuse complementary visual features for robust and efficient place recognition, few of them [179] concerns the *feature selection* issues. Feature selection aims to select relevant features of a dataset. It is of great importance for pattern recognition, especially when the feature vectors are high-dimensional and tend to contain redundant and irrelevant information. This situation is not uncommon when conducting multi-modality information fusion. One can be referred to [199, 200] for more discussions about feature selection.

Although depth information shows its unique advantage for VPR, most commonly used RGB-D cameras such as Kinect have a limited working range, which restricts RGB-D-based methods largely to indoor setting. In addition, the lack of sufficiently large RGB-D datasets with labeled data, especially for outdoor environment, also makes RGB-D cameras hard to be implemented in practice. As mentioned above, Sizikova et al. [148] use synthesized depth images to train their network. However, their method is still restricted to indoor environment. In the past several years, the research area of *depth prediction* [201] has received more attention with the ability of obtaining depth maps for both indoor and outdoor scenes. Depth prediction concerns the generation of depth images based on the corresponding monocular color images by means of deep neural networks. Normally the networks are in the form of encoder-decoder architecture and can be trained end-to-end in a supervised [202] or unsupervised [203] manner. As indicated in [142], detecting loop closures with depth prediction is expected to be a representative approach in the future in the field of visual SLAM.

6.3.3. Omni-directional Image

Compared with the single-view [68, 135] or side-view [8] images used by most of the previously mentioned place recognition methods, omni-directional images [127] are able to perceive the entire surrounding environment and hence more robust against viewpoint changes. Kumar et al. [159] propose the first deep learning-based VPR method using omni-directional images. The images are from NCLT dataset [157], collected with a visual acquisition module composed of five cameras, providing almost 360° planar view. A pre-trained CNN model is used to extract features of images for each place. Experimental results show that, benefiting from omni-directional images, the recognition performance is boosted a lot compared with using single-view images, especially when the viewpoint changes. Iscen et al. [154] introduce a simple yet effective way for recognizing locations with illumination variations by means of panorama to panorama image matching.

In [156], the authors propose a method referred to as Fusion of Omni-directional Multisensory Perception (FOMP). It not only estimates the importance of different viewing angles of a panoramic imaging, but is also able to fuse multi-modality information obtained with heterogeneous sensors (e.g., intensity and depth) to build discriminative features. Wang et al. [204] design a novel CNN architecture coined O-CNN, which takes omni-directional image as input and outputs a feature map used for constructing the representation later. They propose two mechanisms to deal with omni-directional images specifically. That is, *circular padding*, which enables them to leverage all information in the panoramic image adequately, and *roll branching*, which makes the learned features invariant to horizontal rotations of the camera. Their VPR method outperforms the state-of-the-art ones such as NetVLAD [39].

6.3.4. LiDAR Data

LiDAR (Light Detection and Ranging)-based place recognition has become a compelling research topic due to the irreplaceable advantages of LiDAR data. It contains informative 3D structural information of the environment and is more robust against illumination and seasonal variations. Furthermore, laser sensors have a longer working range than RGB-D cameras, which makes them more competitive for robot perception in outdoor scenes. Dubé et al. [205] propose a VPR method named *SegMatch* to use segments in LiDAR point clouds for place matching. SegMatch can recognize places at object-level even though there is no intact object. It takes advantage of descriptive shape information provided by 3D point clouds, making it more reliable under unconstructed scenes.

Unlike the case of traditional images, learning representations (either local or global) from raw 3D point clouds for pattern recognition is still an open issue [206]. Yin et al. [158] design an architecture to learn representations from LiDAR data in an end-to-end fashion. Considering that it is intractable to feed raw LiDAR data into networks directly, Yin et al. [158] first propose an octree structure to map the 3D points into a 2D space. The obtained bird-view images are then used as the input of an adversarial feature learning network to learn low-dimensional features for VPR. They show that LiDAR data gain more robustness to viewpoint changes compared with traditional appearance-based methods like SeqSLAM [68]. Angelina Uy and Hee Lee [131] integrate the novel PointNet [207] and NetVLAD [39], obtaining an architecture referred to as PointNetVLAD, to tackle place recognition. PointNet is capable of taking 3D point clouds as input directly while NetVLAD is proved to be suitable for laser scans due to its permutation-invariant property. The authors formulate place recognition as a metric learning (See Section 4.3 for more description) problem and present a *lazy triplet and quadruplet* loss function to train the proposed network end-to-end. As a consequence, PointNetVLAD is able to extract generalizable and discriminative global representations from raw LiDAR data. With the similar purpose, Liu et al. [208] present a network coined LPD-Net (Large-scale Place Description Network) to learn global descriptors from 3D point clouds in their recent study. Compared with PointNetVLAD [131], LPD-Net considers the spatial distribution of similar local structures, which is capable of improving the recognition performance and gaining more robustness with respect to weather or illumination changes. At last, to incorporate the strength of LiDAR data and intensity image, Guo et al. [209] construct a novel descriptor named ISHOT (Intensity Signature of Histograms of OrienTations) using calibrated intensity returns of LiDAR. As pointed out in [209], when compared

with the descriptors built upon 3D geometric data solely, for example, SHOT [210] and NBLD [211], ISHOT obtains significantly improved place recognition performance.

7. Conclusion and Research Directions

Deep learning (DL) has been becoming ubiquitous in the field of pattern recognition during the last decade. The visual place recognition (VPR) community has recently witnessed a surge on the use of this advanced technique. To the best of our knowledge, this paper presents, for the first time, a comprehensive survey on DL-based VPR methods. We have provided a whole picture about this research topic — from classic convolutional neural networks (CNNs) to more recent generative adversarial networks (GANs), from off-the-shelf CNN features to specifically learned semantic representations and from traditional color images to heterogeneous data. Generally speaking, the main concern of VPR is *how to construct discriminative and robust image representations*. From the DL perspective, we have answered this question by introducing numerous off-the-shelf and specifically designed deep neural networks for representation learning as well as a series of novel descriptors. In addition, we have provided the comparison of runtime performance for representative methods and an exhaustive summarization on the publicly available VPR datasets. At last, we have attempted to identify some (but certainly not complete) open issues and new tools in DL for the VPR research.

Overall, according to the surveyed papers, we find that most of the proposed methods focus on improving the precision and robustness of VPR algorithms, which are expected to recognize revisited places unfailingly under various conditions. Although significant progress has been made in this domain, challenging problems and open issues regarding DL-based VPR methods for real-world application render the following promising research directions: (i) Develop more discriminative and robust image representations (*feature learning*); (ii) Train models that can be self-adaptive with different working scenarios or easily transferred between different domains (*transfer learning/domain adaptation*); (iii) Explore new methods that focus on regions of interest of images to improve the robustness of VPR methods (*visual attention model*); (iv) Propose new methods that utilize high-level *semantic* or *textual information* of the environment; (v) Leverage *heterogeneous imaging data* (e.g., depth image, panoramic image and 3D point clouds) for complementary information; (vi) Develop *multi-modality feature fusion* and/or *feature selection* methods to maximally exploit relevant visual information; (vii) Construct high-quality large-scale (annotated) *datasets* for challenging issues and VPR-specific deep network modeling; and (viii) Develop strategies to better balance between the complexity and efficiency of VPR systems (*runtime consideration*). In the future, this survey can be extended to cover the new exciting developments and the wider application of VPR to practical tasks.

References

[1] H. F. Durrant-Whyte, T. Bailey, Simultaneous Localisation and Mapping (SLAM): Part I, IEEE Robotics & Automation Magazine 13 (2) (2006) 99–110. doi:10.1109/MRA.2006.1638022.

[2] P. Newman, Kin Ho, Slam-loop closing with visually salient features, in: Proceedings of the 2005 IEEE International Conference on Robotics and Automation, 2005, pp. 635–642. doi:10.1109/ROBOT.2005.1570189.

[3] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, M. J. Milford, Visual place recognition: A survey, IEEE Transactions on Robotics 32 (1) (2016) 1–19. doi:10.1109/TRO.2015.2496823.

[4] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[5] H. Bay, A. Ess, T. Tuytelaars, L. V. Gool], Speeded-up robust features (surf), Computer Vision and Image Understanding 110 (3) (2008) 346 – 359, similarity Matching in Computer Vision and Multimedia. doi:https://doi.org/10.1016/j.cviu.2007.09.014.

[6] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1, 2005, pp. 886–893 vol. 1.

[7] D. Galvez-LóPez, J. D. Tardos, Bags of binary words for fast place recognition in image sequences, IEEE Transactions on Robotics 28 (5) (2012) 1188–1197. doi:10.1109/TRO.2012.2197158.

[8] M. Cummins, P. Newman, FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance, The International Journal of Robotics Research 27 (6) (2008) 647–665. doi:10.1177/0278364908090961.

[9] M. Cummins, P. Newman, Appearance-only slam at large scale with fab-map 2.0, The International Journal of Robotics Research 30 (9) (2011) 1100–1123.

[10] B. Williams, G. Klein, I. Reid, Automatic relocalization and loop closing for real-time monocular slam, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (9) (2011) 1699–1712. doi:10.1109/TPAMI.2011.41.

[11] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, T. Chen, Recent advances in convolutional neural networks, Pattern Recognition 77 (2018) 354 – 377. doi:https://doi.org/10.1016/j.patcog.2017.10.013.

[12] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013) 1915–1929.

[13] M. Lopez-Antequera, R. Gomez-Ojeda, N. Petkov, J. Gonzalez-Jimenez, Appearance-invariant place recognition by discriminatively training a convolutional neural network, Pattern Recognition Letters 92 (2017) 89–95.

[14] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, M. Milford, Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free, in: Proceedings of Robotics: Science and Systems XII, 2015, pp. 1–10.

[15] Z. Chen, O. Lam, A. Jacobson, M. Milford, Convolutional neural network-based place recognition, in: 2014 Australasian Conference on Robotics and Automation (ACRA 2014), 2014, pp. 1–8.

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

[17] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[18] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: International Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.

[19] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Processing Magazine 29 (6) (2012) 82–97.

[20] K.-Y. Huang, C.-H. Wu, M.-H. Su, Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses, Pattern Recognition 88 (2019) 668 – 678. doi:https://doi.org/10.1016/j.patcog.2018.12.016.

[21] H. Sak, A. Senior, F. Beaufays, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, in: Fifteenth Annual Conference of the International Speech Communication Association, 2014, pp. 338–342.

[22] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, Journal of Machine Learning Research 11 (Dec) (2010) 3371–3408.

[23] R. Raina, A. Madhavan, A. Y. Ng, Large-scale deep unsupervised learning using graphics processors, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 873–880.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.

[25] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al., Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The Journal of Machine Learning Research 15 (1) (2014) 1929–1958.

[27] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[30] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, C. Schmid, Aggregating local image descriptors into compact codes, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (9) (2012) 1704–1716. doi:10.1109/TPAMI.2011.235.

[31] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, in: International Conference on Machine Learning, 2014, pp. 647–655.

[32] A. S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, Cnn features off-the-shelf: An astounding baseline for recognition, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 512–519.

[33] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, arXiv preprint arXiv:1312.6229.

[34] Y. Hou, H. Zhang, S. Zhou, Convolutional neural network-based image representation for visual loop closure detection, in: IEEE International Conference on Information and Automation, 2015, pp. 2238–2245.

[35] S. Cascianelli, G. Costante, E. Bellocchio, P. Valigi, M. L. Fravolini, T. A. Ciarfuglia, Robust visual semi-semantic loop closure detection by a covisibility graph and cnn features, Robotics and Autonomous Systems 92 (2017) 53–65.

[36] S. Garg, N. Suenderhauf, M. Milford, Semantic–geometric visual place recognition: a new perspective for reconciling opposing views, The International Journal of Robotics Research (2019) 0278364919839761.

[37] S. Garg, N. Suenderhauf, M. Milford, Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics, Robotics: Science and Systems (RSS)doi:10.15607/RSS.2018.XIV.022.

[38] X. Zhang, Y. Su, X. Zhu, Loop closure detection for visual slam systems using convolutional neural network, in: Automation and Computing (ICAC), 2017 23rd International Conference on, IEEE, 2017, pp. 1–6.

[39] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, J. Sivic, Netvlad: Cnn architecture for weakly supervised place recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5297–5307.

[40] A. Gordo, J. Almazán, J. Revaud, D. Larlus, Deep image retrieval: Learning global representations for image search, in: European Conference on Computer Vision (ECCV), Springer, 2016, pp. 241–257.

[41] H. Wang, Z. Li, Y. Li, B. Gupta, C. Choi, Visual saliency guided complex image retrieval, Pattern Recognition Letters 130 (2020) 64 – 72, image/Video Understanding and Analysis (IUVA). doi:https://doi.org/10.1016/j.patrec.2018.08.010.

[42] R. Mur-Artal, J. M. M. Montiel, J. D. Tardos, Orb-slam: a versatile and accurate monocular slam system, IEEE Transactions on Robotics 31 (5) (2015) 1147–1163. doi:10.1109/TRO.2015.2463671.

[43] I. Ulrich, I. Nourbakhsh, Appearance-based place recognition for topological localization, in: IEEE International Conference on Robotics and Automation (ICRA), Vol. 2, Ieee, 2000, pp. 1023–1029.

[44] T. Naseer, L. Spinello, W. Burgard, C. Stachniss, Robust visual robot localization across seasons using network flows, in: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI14, AAAI Press, 2014, pp. 2564–2570.

[45] T. Naseer, M. Ruhnke, C. Stachniss, L. Spinello, W. Burgard, Robust visual slam across seasons, in: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2015, pp. 2529–2535.

[46] G. Tolias, H. Jégou, Visual query expansion with or without geometry: refining local descriptors by feature aggregation, Pattern recognition

47 (10) (2014) 3466–3476. doi:https://doi.org/10.1016/j.patcog.2014.04.007.

[47] R. Arandjelović, A. Zisserman, Three things everyone should know to improve object retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 2911–2918.

[48] X. Wu, G. Irie, K. Hiramatsu, K. Kashino, Query expansion with diffusion on mutual rank graphs, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 1653–1657. doi:10.1109/ICASSP.2018.8461360.

[49] X. Zhang, L. Wang, Y. Zhao, Y. Su, Graph-based place recognition in image sequences with cnn features, Journal of Intelligent & Robotic Systems 95 (2) (2019) 389–403. doi:10.1007/s10846-018-0917-2.

[50] M. Donoser, H. Bischof, Diffusion processes for retrieval revisited, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1320–1327.

[51] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, O. Chum, Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 926–935.

[52] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 3354–3361. doi:10.1109/CVPR.2012.6248074.

[53] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, 2014, pp. 675–678.

[54] P. Panphattarasap, A. Calway, Visual place recognition using landmark distribution descriptors, in: Asian Conference on Computer Vision, Springer, 2016, pp. 487–502.

[55] P. Neubert, P. Protzel, Beyond holistic descriptors, keypoints, and fixed patches: Multiscale superpixel grids for place recognition in changing environments, IEEE Robotics and Automation Letters 1 (1) (2016) 484–491.

[56] B. Dongdong, W. Chaoqun, B. Zhang, Y. Xiaodong, Y. Xuejun, et al., Cnn feature boosted seqslam for real-time loop closure detection, Chinese Journal of Electronics 27 (3) (2018) 488–499.

[57] Z. Chen, F. Maffra, I. Sa, M. Chli, Only look once, mining distinctive landmarks from convnet for visual place recognition, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 9–16.

[58] S. Hausler, A. Jacobson, M. Milford, Filter early, match late: Improving network-based visual place recognition, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019, pp. 3268 – 3275. doi:10.1109/IROS40897.2019.8967783.

[59] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, M. Milford, Deep learning features at scale for visual place recognition, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, Singapore, 2017, pp. 3223–3230. doi:10.1109/ICRA.2017.7989366.

[60] T. Naseer, G. L. Oliveira, T. Brox, W. Burgard, Semantics-aware visual localization under challenging perceptual conditions, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 2614–2620.

[61] G. L. Oliveira, W. Burgard, T. Brox, Efficient deep models for monocular road segmentation, in: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2016, pp. 4885–4891.

[62] F. Radenović, G. Tolias, O. Chum, Fine-tuning cnn image retrieval with no human annotation, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (7) (2019) 1655–1668. doi:10.1109/TPAMI.2018.2846566.

[63] Z. Chen, L. Liu, I. Sa, Z. Ge, M. Chli, Learning context flexible attention model for long-term visual place recognition, IEEE Robotics and Automation Letters 3 (4) (2018) 4015–4022. doi:10.1109/LRA.2018.2859916.

[64] N. Merrill, G. Huang, Lightweight unsupervised deep loop closure, in: Robotics: Science and Systems, 2018, pp. 1–10.

[65] H. J. Kim, E. Dunn, J. Frahm, Learned contextual feature reweighting for image geo-localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3251–3260. doi:10.1109/CVPR.2017.346.

[66] Z. Xin, Y. Cai, T. Lu, X. Xing, S. Cai, J. Zhang, Y. Yang, Y. Wang, Localizing discriminative visual landmarks for place recognition, 2019 IEEE International Conference on Robotics and Automation (ICRA) (2019) 5979–5985.

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

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

[69] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, International Journal of Computer Vision 42 (3) (2001) 145–175. doi:https://doi.org/10.1023/A:1011139631724.

[70] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2010, pp. 3304–3311.

[71] H. Jégou, O. Chum, Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening, in: European Conference on Computer Vision (ECCV), Springer, 2012, pp. 774–787.

[72] Y. Gong, L. Wang, R. Guo, S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features, in: European Conference on Computer Vision (ECCV), Springer, 2014, pp. 392–407.

[73] H. Abdi, L. J. Williams, Principal component analysis, Wiley interdisciplinary reviews: computational statistics 2 (4) (2010) 433–459.

[74] A. Babenko, A. Slesarev, A. Chigorin, V. Lempitsky, Neural codes for image retrieval, in: European Conference on Computer Vision (ECCV), Springer, 2014, pp. 584–599.

[75] A. Babenko, V. Lempitsky, Aggregating local deep features for image retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1269–1277.

[76] H. Azizpour, A. Sharif Razavian, J. Sullivan, A. Maki, S. Carlsson, From generic to specific deep representations for visual recognition, in: Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition Workshops, 2015, pp. 36–45.

[77] G. Tolias, R. Sicre, H. Jégou, Particular object retrieval with integral max-pooling of cnn activations, International Conference on Learning Representations (ICLR).

[78] Y. Kalantidis, C. Mellina, S. Osindero, Cross-dimensional weighting for aggregated deep convolutional features, in: European Conference on Computer Vision (ECCV), Springer, 2016, pp. 685–701.

[79] A. Gordo, J. Almazan, J. Revaud, D. Larlus, End-to-end learning of deep visual representations for image retrieval, International Journal of Computer Vision 124 (2) (2017) 237–254. doi:https://doi.org/10.1007/s11263-017-1016-8.

[80] P. Neubert, P. Protzel, Local region detector+ cnn based landmarks for practical place recognition in changing environments, in: 2015 European Conference on Mobile Robots (ECMR), IEEE, 2015, pp. 1–6.

[81] Y. Hou, H. Zhang, S. Zhou, Evaluation of object proposals and convnet features for landmark-based visual place recognition, Journal of Intelligent & Robotic Systems 92 (2018) 505 – 520. doi:https://doi.org/10.1007/s10846-017-0735-y.

[82] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Lost in quantization: Improving particular object retrieval in large scale image databases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2008, pp. 1–8.

[83] F. Radenović, G. Tolias, O. Chum, Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples, in: European Conference on Computer Vision (ECCV), Springer, 2016, pp. 3–20.

[84] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[85] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, T. Pajdla, 24/7 place recognition by view synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1808–1817.

[86] J. Hosang, R. Benenson, P. Dollár, B. Schiele, What makes for effective detection proposals?, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (4) (2016) 814–830. doi:10.1109/TPAMI.2015.2465908.

[87] C. L. Zitnick, P. Dollár, Edge boxes: Locating object proposals from edges, in: European Conference on Computer Vision (ECCV), Springer, 2014, pp. 391–405.

[88] J. Redmon, A. Farhadi, Yolo9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7263–7271.

[89] M.-M. Cheng, Z. Zhang, W.-Y. Lin, P. Torr, Bing: Binarized normed gradients for objectness estimation at 300fps, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3286–3293.

[90] P. Krähenbühl, V. Koltun, Geodesic object proposals, in: European Conference on Computer Vision (ECCV), Springer, 2014, pp. 725–739.

[91] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, J. Malik, Multiscale combinatorial grouping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 328–335.

[92] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, Slic superpixels compared to state-of-the-art superpixel methods, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (11) (2012) 2274–2282.

[93] Z. Xin, X. Cui, J. Zhang, Y. Yang, Y. Wang, Real-time visual place recognition based on analyzing distribution of multi-scale cnn landmarks, Journal of Intelligent & Robotic Systems 94 (3-4) (2019) 777–792.

[94] A. Maldonado-Ramírez, L. A. Torres-Mendez, Learning ad-hoc compact representations from salient landmarks for visual place recognition in underwater environments, in: 2019 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 5739–5745.

[95] A. Maldonado-Ramírez, L. A. Torres-Méndez, Robotic visual tracking of relevant cues in underwater environments with poor visibility conditions, Journal of Sensors 2016.

[96] B. Yang, X. Xu, J. Li, H. Zhang, Landmark generation in visual place recognition using multi-scale sliding window for robotics, Applied Sciences 9 (15) (2019) 3146. doi:10.3390/app9153146.

[97] L. Zheng, Y. Yang, Q. Tian, Sift meets cnn: A decade survey of instance retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (5) (2017) 1224–1244. doi:10.1109/TPAMI.2017.2709749.

[98] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, M. Pietikäinen, From bow to cnn: Two decades of texture representation for texture classification, International Journal of Computer Vision 127 (1) (2019) 74–109. doi:10.1007/s11263-018-1125-z.

[99] N. Zhang, J. Donahue, R. Girshick, T. Darrell, Part-based r-cnns for fine-grained category detection, in: European Conference on Computer Vision (ECCV), Springer, 2014, pp. 834–849.

[100] J. L. Schonberger, F. Radenovic, O. Chum, J.-M. Frahm, From single image query to detailed 3d reconstruction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5126–5134.

[101] Z. Chen, S. Lowry, A. Jacobson, Z. Ge, M. Milford, Distance metric learning for feature-agnostic place recognition, in: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2015, pp. 2556–2563.

[102] X. Zhe, S. Chen, H. Yan, Directional statistics-based deep metric learning for image classification and retrieval, Pattern Recognition 93 (2019) 113–123. doi:https://doi.org/10.1016/j.patcog.2019.04.005.

[103] Y. Xiao, J. Li, B. Du, J. Wu, J. Chang, W. Zhang, Memu: Metric correlation siamese network and multi-class negative sampling for visual tracking, Pattern Recognition 100 (2020) 107170. doi:https://doi.org/10.1016/j.patcog.2019.107170.

[104] E. Hoffer, N. Ailon, Deep metric learning using triplet network, in: A. Feragen, M. Pelillo, M. Loog (Eds.), Similarity-Based Pattern Recognition, Springer International Publishing, Cham, 2015, pp. 84–92.

[105] S. Chopra, R. Hadsell, Y. LeCun, et al., Learning a similarity metric discriminatively, with application to face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 539–546. doi:10.1109/CVPR.2005.202.

[106] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based cnn with improved triplet loss function, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1335–1344.

[107] W. Ge, Deep metric learning with hierarchical triplet loss, in: European Conference on Computer Vision (ECCV), 2018, pp. 269–285.

[108] T. Naseer, W. Burgard, C. Stachniss, Robust visual localization across seasons, IEEE Transactions on Robotics 34 (2) (2018) 289–302. doi:10.1109/TRO.2017.2788045.

[109] D. Ravichandran, P. Pantel, E. Hovy, Randomized algorithms and nlp: Using locality sensitive hash functions for high speed noun clustering, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05), 2005, pp. 622–629.

[110] S. Lowry, H. Andreasson, Lightweight, viewpoint-invariant visual place recognition in changing environments, IEEE Robotics and Automation Letters 3 (2) (2018) 957–964. doi:10.1109/LRA.2018.2793308.

[111] L. Wu, Y. Wu, Deep supervised hashing with similar hierarchy for place recognition, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019, pp. 3781–3786.

[112] E. Garcia-Fidalgo, A. Ortiz, Hierarchical place recognition for topological mapping, IEEE Transactions on Robotics 33 (5) (2017) 1061–

1074. doi:10.1109/TRO.2017.2704598.

[113] P. Neubert, S. Schubert, P. Protzel, A neurologically inspired sequence processing model for mobile robot place recognition, IEEE Robotics and Automation Letters 4 (4) (2019) 3200–3207. doi:10.1109/LRA.2019.2927096.

[114] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, K. McDonald-Maier, A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes, IEEE Transactions on Robotics 36 (2) (2020) 561–569.

[115] M. Chancn, L. Hernandez-Nunez, A. Narendra, A. B. Barron, M. Milford, A hybrid compact neural architecture for visual place recognition, IEEE Robotics and Automation Letters 5 (2) (2020) 993–1000. doi:10.1109/LRA.2020.2967324.

[116] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, T. Pajdla, Benchmarking 6dof outdoor visual localization in changing conditions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8601–8610.

[117] M. Smith, I. Baldwin, W. Churchill, R. Paul, P. Newman, The new college vision and laser data set, The International Journal of Robotics Research 28 (5) (2009) 595–599. doi:http://dx.doi.org/10.1177/0278364909103911.

[118] S. Lynen, M. Bosse, R. Siegwart, Trajectory-based place-recognition for efficient large scale localization, International Journal of Computer Vision 124 (1) (2017) 49–64. doi:https://doi.org/10.1007/s11263-016-0947-9.

[119] Y. Latif, C. Cadena, J. Neira, Robust loop closing over time for pose graph slam, The International Journal of Robotics Research 32 (14) (2013) 1611–1626. doi:https://doi.org/10.1177/0278364913498910.

[120] A. Bonarini, W. Burgard, G. Fontana, M. Matteucci, D. G. Sorrenti, J. D. Tardos, Rawseeds: Robotics advancement through web-publishing of sensorial and elaborated extensive data sets, in: International Conference on Intelligent Robots and Systems (IROS), Vol. 6, 2006, p. 93.

[121] G. Pandey, J. R. McBride, R. M. Eustice, Ford campus vision and lidar data set, The International Journal of Robotics Research 30 (13) (2011) 1543–1552. doi:https://doi.org/10.1177/0278364911400640.

[122] G. Zhang, M. J. Lilly, P. A. Vela, Learning binary features online from motion dynamics for incremental loop-closure detection and place recognition, in: 2016 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2016, pp. 765–772.

[123] J.-L. Blanco, F.-A. Moreno, J. González, A collection of outdoor robotic datasets with centimeter-accuracy ground truth, Autonomous Robots 27 (4) (2009) 327–351. doi:10.1007/s10514-009-9138-7.

[124] A. Mukherjee, S. Chakraborty, S. K. Saha, Learning deep representation for place recognition in slam, in: International Conference on Pattern Recognition and Machine Intelligence, Springer, 2017, pp. 557–564.

[125] A. J. Glover, W. P. Maddern, M. J. Milford, G. F. Wyeth, Fab-map + ratslam: Appearance-based slam for multiple times of day, in: 2010 IEEE International Conference on Robotics and Automation (ICRA), 2010, pp. 3507–3512. doi:10.1109/ROBOT.2010.5509547.

[126] A. Pronobis, B. Caputo, Cold: The cosy localization database, The International Journal of Robotics Research 28 (5) (2009) 588–594. doi:https://doi.org/10.1177/0278364909103912.

[127] H. Lu, Z. Zheng, Two novel real-time local visual features for omnidirectional vision, Pattern Recognition 43 (12) (2010) 3938 – 3949. doi:https://doi.org/10.1016/j.patcog.2010.06.020.

[128] G. Costante, T. A. Ciarfuglia, P. Valigi, E. Ricci, A transfer learning approach for multi-cue semantic place recognition, in: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2013, pp. 2122–2129.

[129] M. Mancini, S. R. Bulò, E. Ricci, B. Caputo, Learning deep nbnn representations for robust place categorization, IEEE Robotics and Automation Letters 2 (3) (2017) 1794–1801. doi:10.1109/LRA.2017.2705282.

[130] W. Maddern, G. Pascoe, C. Linegar, P. Newman, 1 Year, 1000km: The Oxford RobotCar Dataset, The International Journal of Robotics Research (IJRR) 36 (1) (2017) 3–15. doi:10.1177/0278364916679498.

[131] M. Angelina Uy, G. Hee Lee, Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4470–4479.

[132] H. Porav, W. Maddern, P. Newman, Adversarial training for adverse conditions: Robust metric localisation using appearance transfer, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 1011–1018.

[133] A. Glover, Gardens point walking (2014).

URL https://wiki.qut.edu.au/display/cyphy/Day+and+Night+with+Lateral+Pose+Change+Datasets

[134] F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, J. Civera, Mapillary street-level sequences: A dataset for lifelong place recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1–10.

[135] N. Sünderhauf, P. Neubert, P. Protzel, Are we there yet? challenging seqslam on a 3000 km journey across all four seasons, in: Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA), 2013, p. 1.

[136] O. Vysotska, C. Stachniss, Lazy data association for image sequences matching under substantial appearance changes, IEEE Robotics and Automation Letters 1 (1) (2016) 213–220.

[137] H. Badino, D. Huber, T. Kanade, Visual topometric localization, in: 2011 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2011, pp. 794–799.

[138] E. Stenborg, C. Toft, L. Hammarstrand, Long-term visual localization using semantically segmented images, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 6484–6490.

[139] N. Sünderhauf, The vprice challenge 2015 visual place recognition in changing environments (2015).
URL https://goo.gl/ROQYU2

[140] J. Sturm, N. Engelhard, F. Endres, W. Burgard, D. Cremers, A benchmark for the evaluation of rgb-d slam systems, in: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2012, pp. 573–580. doi:10.1109/IROS.2012.6385773.

[141] X. Gao, T. Zhang, Unsupervised learning to detect loops using deep neural networks for visual slam system, Autonomous Robots 41 (1) (2017) 1–18.

[142] K. Tateno, F. Tombari, I. Laina, N. Navab, Cnn-slam: Real-time dense monocular slam with learned depth prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6243–6252.

[143] O. Mendez, S. Hadfield, N. Pugeault, R. Bowden, Sedar: Reading floorplans like a humanusing deep learning to enable human-inspired localisation, International Journal of Computer Vision 128 (5) (2020) 1286 – 1310. doi:10.1007/s11263-019-01239-4.

[144] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, A. Fitzgibbon, Scene coordinate regression forests for camera relocalization in rgb-d images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2930–2937.

[145] A. Kendall, M. Grimes, R. Cipolla, Posenet: A convolutional network for real-time 6-dof camera relocalization, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2938–2946.

[146] R. Li, Q. Liu, J. Gui, D. Gu, H. Hu, Indoor relocalization in challenging environments with dual-stream convolutional neural networks, IEEE Transactions on Automation Science & Engineering PP (99) (2017) 1–12.

[147] A. Handa, T. Whelan, J. McDonald, A. J. Davison, A benchmark for rgb-d visual odometry, 3d reconstruction and slam, in: 2014 IEEE international conference on Robotics and automation (ICRA), IEEE, 2014, pp. 1524–1531.

[148] E. Sizikova, V. K. Singh, B. Georgescu, M. Halber, K. Ma, T. Chen, Enhancing place recognition using joint intensity-depth analysis and synthetic data, in: European Conference on Computer Vision (ECCV), Springer, 2016, pp. 901–908.

[149] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, J. Gall, Semantickitti: A dataset for semantic scene understanding of lidar sequences, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9296–9306. doi:10.1109/ICCV.2019.00939.

[150] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, D. Anguelov, Scalability in perception for autonomous driving: Waymo open dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2446–2454.

[151] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3213–3223.

[152] M. Larsson, E. Stenborg, L. Hammarstrand, M. Pollefeys, T. Sattler, F. Kahl, A cross-season correspondence dataset for robust semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9532–9542.

[153] A. Torii, J. Sivic, T. Pajdla, M. Okutomi, Visual place recognition with repetitive structures, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 883–890.

[154] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, O. Chum, Panorama to panorama matching for location recognition, in: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ACM, 2017, pp. 392–396.

[155] Z. Seymour, K. Sikka, H.-P. Chiu, S. Samarasekera, R. Kumar, Semantically-aware attentive neural embeddings for long-term 2d visual localization, in: 2019 30th British Machine Vision Conference (BMVC), 2019, pp. 1–15.

[156] S. Siva, H. Zhang, Omnidirectional multisensory perception fusion for long-term place recognition, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 5175–5181. doi:10.1109/ICRA.2018.8461042.

[157] N. Carlevaris-Bianco, A. K. Ushani, R. M. Eustice, University of michigan north campus long-term vision and lidar dataset, The International Journal of Robotics Research 35 (9) (2016) 1023–1035.

[158] P. Yin, L. Xu, Z. Liu, L. Li, H. Salman, Y. He, W. Xu, H. Wang, H. Choset, Stabilize an unsupervised feature learning for lidar-based place recognition, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 1162–1167.

[159] D. Kumar, H. Neher, A. Das, D. A. Clausi, S. L. Waslander, Condition and viewpoint invariant omni-directional place recognition using cnn, in: 2017 14th Conference on Computer and Robot Vision (CRV), IEEE, 2017, pp. 32–39.

[160] F. Maffra, Z. Chen, M. Chli, Viewpoint-tolerant place recognition combining 2d and 3d information for uav navigation, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 2542–2549. doi:10.1109/ICRA.2018.8460786.

[161] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, K. Alexis, K. McDonald-Maier, Are state-of-the-art visual place recognition techniques any good for aerial robotics?, IEEE ICRA 2019 Workshop on Aerial Robotics.

[162] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, R. Siegwart, The euroc micro aerial vehicle datasets, The International Journal of Robotics Research 35 (10) (2016) 1157–1163.

[163] A. C. Duarte, G. B. Zaffari, R. T. S. da Rosa, L. M. Longaray, P. Drews, S. S. Botelho, Towards comparison of underwater slam methods: An open dataset collection, in: OCEANS 2016 MTS/IEEE Monterey, IEEE, 2016, pp. 1–5.

[164] G. Kim, Y. S. Park, Y. Cho, J. Jeong, A. Kim, Mulran: Multimodal range dataset for urban place recognition, in: IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 1–8.

[165] M. Milford, W. Scheirer, E. Vig, A. Glover, O. Baumann, J. Mattingley, D. Cox, Condition-invariant, top-down visual place recognition, in: 2014 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2014, pp. 5571–5577.

[166] F. Endres, J. Hess, J. Sturm, D. Cremers, W. Burgard, 3-d mapping with an rgb-d camera, IEEE Transactions on Robotics 30 (1) (2014) 177–187. doi:10.1109/TRO.2013.2279412.

[167] Y. Li, Z. Zhang, Y. Cheng, L. Wang, T. Tan, Mapnet: Multi-modal attentive pooling network for rgb-d indoor scene classification, Pattern Recognition 90 (2019) 436 – 449. doi:https://doi.org/10.1016/j.patcog.2019.02.005.

[168] O. Guclu, A. B. Can, Integrating global and local image features for enhanced loop closure detection in rgb-d slam systems, The Visual Computer (2019) 1–20.

[169] N. Piasco, D. Sidibé, C. Demonceaux, V. Gouet-Brunet, A survey on visual-based localization: On the benefit of heterogeneous data, Pattern Recognition 74 (2018) 90–109. doi:https://doi.org/10.1016/j.patcog.2017.09.013.

[170] J. Masci, U. Meier, D. Cireşan, J. Schmidhuber, Stacked convolutional auto-encoders for hierarchical feature extraction, in: International Conference on Artificial Neural Networks, Springer, 2011, pp. 52–59.

[171] Z. Wang, Z. Peng, Y. Guan, L. Wu, Manifold regularization graph structure auto-encoder to detect loop closure for visual slam, IEEE Access 7 (2019) 59524–59538.

[172] L. Zhang, Q. Zhang, L. Zhang, D. Tao, X. Huang, B. Du, Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding, Pattern Recognition 48 (10) (2015) 3102–3112.

[173] M. D. Yongqiang Zhang, Yancheng Bai, B. Ghanem, Multi-task generative adversarial network for detecting small objects in the wild, International Journal of Computer Vision 128 (2020) 1810 – 1828.

[174] A. Atapour-Abarghouei, S. Akcay, G. P. de La Garanderie, T. P. Breckon, Generative adversarial framework for depth filling via wasserstein

metric, cosine transform and domain transfer, Pattern Recognition 91 (2019) 232 – 244. doi:https://doi.org/10.1016/j.patcog.2019.02.010.

[175] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D. N. Metaxas, Stackgan++: Realistic image synthesis with stacked generative adversarial networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (8) (2019) 1947–1962.

[176] Y. Latif, R. Garg, M. Milford, I. Reid, Addressing challenging place recognition tasks using generative adversarial networks, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 2349–2355.

[177] C. Wang, W. Niu, Y. Jiang, H. Zheng, Z. Yu, Z. Gu, B. Zheng, Discriminative region proposal adversarial network for high-quality image-to-image translation, International Journal of Computer Visiondoi:https://doi.org/10.1007/s11263-019-01273-2.

[178] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

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

[180] R. Li, W. Cao, Q. Jiao, S. Wu, H.-S. Wong, Simplified unsupervised image translation for semantic segmentation adaptation, Pattern Recognition 105 (2020) 107343. doi:https://doi.org/10.1016/j.patcog.2020.107343.

[181] R. M. Abhinav Valada, W. Burgard, Self-supervised model adaptation for multimodal semantic segmentation, International Journal of Computer Vision 128 (2020) 1239 – 1285. doi:https://doi.org/10.1007/s11263-019-01188-y.

[182] A. Lpez-Cifuentes, M. Escudero-Violo, J. Bescs, lvaro Garca-Martn, Semantic-aware scene recognition, Pattern Recognition 102 (2020) 107256. doi:https://doi.org/10.1016/j.patcog.2020.107256.

[183] G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1925–1934.

[184] J. L. Schönberger, M. Pollefeys, A. Geiger, T. Sattler, Semantic visual localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6896–6906.

[185] Z. Hong, Y. Petillot, D. Lane, Y. Miao, S. Wang, Textplace: Visual place recognition and topological localization through reading scene texts, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2861–2870.

[186] T. BaltruÅaitis, C. Ahuja, L. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2) (2019) 423–443. doi:10.1109/TPAMI.2018.2798607.

[187] X. Ma, T. Zhang, C. Xu, Deep multi-modality adversarial networks for unsupervised domain adaptation, IEEE Transactions on Multimedia 21 (9) (2019) 2419–2431. doi:10.1109/TMM.2019.2902100.

[188] T. Ojala, M. Pietikäinen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern recognition 29 (1) (1996) 51–59. doi:https://doi.org/10.1016/0031-3203(95)00067-4.

[189] J. Wu, J. M. Rehg, Centrist: A visual descriptor for scene categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (8) (2010) 1489–1501. doi:10.1109/TPAMI.2010.224.

[190] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2, IEEE, 2006, pp. 2169–2178.

[191] Y. Qiao, Z. Zhang, Visual localization by place recognition based on multifeature (d-ƛlbp, Journal of Sensors 2017 (2017) 1–18.

[192] S. Sabour, N. Frosst, G. E. Hinton, Dynamic routing between capsules, in: Advances in Neural Information Processing Systems, 2017, pp. 3856–3866.

[193] N. Yang, R. Wang, J. Stückler, D. Cremers, Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry, in: European Conference on Computer Vision (ECCV), Springer, 2018, pp. 835–852.

[194] A. Wang, J. Cai, J. Lu, T.-J. Cham, Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1125–1133.

[195] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, Q. Yang, Rgbd salient object detection via deep fusion, IEEE Transactions on Image Processing 26 (5) (2017) 2274–2285. doi:10.1109/TIP.2017.2682981.

[196] C. Zou, R. Guo, Z. Li, D. Hoiem, Complete 3d scene parsing from an rgbd image, International Journal of Computer Vision 127 (2) (2019) 143162. doi:10.1007/s11263-018-1133-z.

[197] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, R. Cipolla, Understanding real world indoor scenes with synthetic data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4077–4085.

[198] B. Zhou, A. L. Garcia, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, Advances in Neural Information Processing Systems 1 (2015) 487–495.

[199] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, Neurocomputing 300 (2018) 70–79.

[200] S. Sharmin, M. Shoyaib, A. A. Ali, M. A. H. Khan, O. Chae, Simultaneous feature selection and discretization based on mutual information, Pattern Recognition 91 (2019) 162 – 174. doi:https://doi.org/10.1016/j.patcog.2019.02.016.

[201] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, A. L. Yuille, Towards unified depth and semantic prediction from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2800–2809.

[202] B. Li, Y. Dai, M. He, Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference, Pattern Recognition 83 (2018) 328 – 339. doi:https://doi.org/10.1016/j.patcog.2018.05.029.

[203] A. Tonioni, M. Poggi, S. Mattoccia, L. Di Stefano, Unsupervised domain adaptation for depth prediction from images, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 1–1doi:10.1109/TPAMI.2019.2940948.

[204] T.-H. Wang, H.-J. Huang, J.-T. Lin, C.-W. Hu, K.-H. Zeng, M. Sun, Omnidirectional cnn for visual place recognition and navigation, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 2341–2348.

[205] R. Dubé, D. Dugas, E. Stumm, J. Nieto, R. Siegwart, C. Cadena, Segmatch: Segment based place recognition in 3d point clouds, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 5266–5272.

[206] C. Zou, B. He, M. Zhu, L. Zhang, J. Zhang, Learning motion field of lidar point cloud with convolutional networks, Pattern Recognition Letters 125 (2019) 514 – 520. doi:https://doi.org/10.1016/j.patrec.2019.06.009.

[207] C. R. Qi, H. Su, K. Mo, L. J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 652–660.

[208] Z. Liu, S. Zhou, C. Suo, P. Yin, W. Chen, H. Wang, H. Li, Y. Liu, Lpd-net: 3d point cloud learning for large-scale place recognition and environment analysis, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2831–2840.

[209] J. Guo, P. V. Borges, C. Park, A. Gawel, Local descriptor for robust place recognition using lidar intensity, IEEE Robotics and Automation Letters 4 (2) (2019) 1470–1477.

[210] F. Tombari, S. Salti, L. Di Stefano, Unique signatures of histograms for local surface description, in: European Conference on Computer Vision (ECCV), Springer, 2010, pp. 356–369.

[211] T. Cieslewski, E. Stumm, A. Gawel, M. Bosse, S. Lynen, R. Siegwart, Point cloud descriptors for place recognition using sparse visual information, in: 2016 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2016, pp. 4830–4836.

Xiwu Zhang received his B.Eng. degree from Nanjing University of Science and Technology (NJUST), China, in 2014. He is currently pursuing his Ph.D. degree at NJUST, China. He was visiting University of Wollongong, Australia, from 2017 to 2018. He focuses on visual place recognition, deep learning, computer vision and visual SLAM.

Lei Wang received his Ph.D. degree from Nanyang Technological University, Singapore in 2004. Now he works as associate professor at School of Computing and Information Technology of University of Wollongong, Australia. His research interests include pattern recognition, machine/deep learning, computer vision and image retrieval.

Yan Su received his M.Eng. and Ph.D. degrees from Southeast University, China in 1996 and 2001, respectively. He serves as a professor at School of Mechanical Engineering, Nanjing University of Science and Technology since 2005. The main research interests of Yan Su are in navigation systems, autonomous robots/vehicles and sensors.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.