IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 January 2023, accepted 15 February 2023, date of publication 20 February 2023, date of current version 23 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3246803

RESEARCH ARTICLE

CoCALC: A Self-Supervised Visual Place Recognition Approach Combining Appearance and Geometric Information

KANGYU LI[1,2], XIFENG WANG[2], LEILEI SHI[1,2], AND NIUNIU GENG[1]
[1]Machinery Technology Development Company Ltd., Beijing 101407, China
[2]China Academy of Machinery Science and Technology, Beijing 100044, China

Corresponding author: Kangyu Li (liky@mtd.com.cn)

This work was supported by the Special Project of the National Key Research and Development Program of China under Grant 2020YFB1313304.

ABSTRACT Visual place recognition (VPR) is considered among the most complicated tasks in SLAM due to the multiple challenges of drastic variations in both appearance and viewpoint. To address this issue, this article presents a self-supervised and lightweight VPR approach (namely CoCALC) that fully utilizes the appearance and geometric information provided by images. The main thing that makes CoCALC ultra-lightweight (only 0.27 MB) is our use of Depthwise Separable Convolution (DSC), a simple but effective architecture that enables our model to generate a more robust image representation. The network trained specifically for VPR can efficiently extract deep convolutional features from salient image regions that have relatively higher entropy, thereby expanding its applications on resource-limited platforms without GPUs. To further eliminate the negative consequences of the high percent false matches, a novel band-matrix-based geometric check is employed to filter out the incorrect matching of image patches, and the impact of different bandwidths on the recall rate is discussed. Results on several benchmark datasets confirm that the proposed CoCALC can yield state-of-the-art performance and superior generalization with acceptable efficiency. All relevant codes are provided at https://github.com/LiKangyuLKY/CoCALC-VPR for further studies.

INDEX TERMS Convolutional neural network, robotic vision, visual place recognition, visual simultaneous localization and mapping.

I. INTRODUCTION

Over the past few decades, visual simultaneous localization and mapping (SLAM) [1] has been considerably advanced in robotics communities. Visual place recognition (VPR) denotes the task of ascertaining whether or not a place has already been visited using the visual information of images. As one of the essential components in the SLAM pipeline, it can help correct the accumulated drift and offer a more precise pose estimation by recognizing previously visited places [2]. Regarding the robots that operate autonomously for an extended period, ascertaining whether the current place is a revisited one is still considered a daunting task due to the challenges including appearance and viewpoint changes, occlusions, and perceptual aliasing [3].

VPR can be considered as the data association task [4] and is known as the appearance-based approach when the association is carried out in the image space [5], [6], [7]. This kind of approach is generically conducted within the framework of image retrieval and the performance is highly influenced by the image representation (so-called descriptors). In addition, the VPR systems are typically conducted with resource-constrained and internal-space-limited platforms, such as compact industrial PCs without GPUs. Therefore, the descriptors should be computationally efficient to attain the requirements of real-time running. This means that the proper trade-off between matching performance and computational efficiency is thus needed.

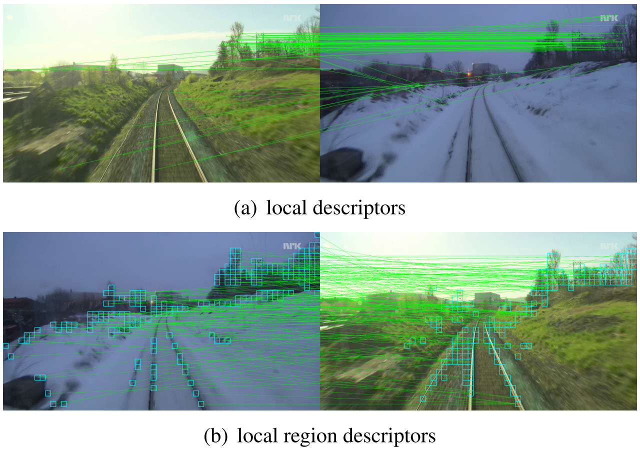
The associate editor coordinating the review of this manuscript and approving it for publication was Huiyu Zhou.

Early visual descriptors can be classified into two major divisions: local descriptors and global descriptors. Local descriptors describe the image by extracting the feature around each interest point, which exhibit robustness against viewpoint changes but suffer greatly from appearance changes. Conversely, global descriptors describe the image as a whole and generate one compact feature vector, thus they have advantages in appearance and illumination invariance but are not good at dealing with viewpoint changes. Consequently, some attempts [8], [9] were made to combine the complementary strengths of the local and global descriptors to arrive at more powerful hybrid approaches, namely local region descriptors. An impressive handcrafted feature-based work, called CoHOG [8], used the entropy map to extract regions of interest (ROI) and then used the Histogram of Oriented Gradients (HOG) descriptor to create the cooperative regional representations.

Lately, the focus on region-based VPR has turned to learning-based techniques, especially Convolutional Neural Networks (CNN), due to their great success in image retrieval and classification. In particular, the great potential of combining local region descriptors and CNN techniques is confirmed by the preliminary results in the VPR task [10], [11], [12], [13], [14]. However, CNN-based descriptors generally require a significant amount of computing resources, such as hardware-based acceleration using a GPU, which are not suitable for resource-limited devices. In this case, the development of lightweight networks offers potential for practical applications.

Although the above-mentioned approaches have achieved promising performance by extracting various features, they do not fully utilize the geometric information contained in images. Such additional geometric information has been shown to improve the precision of place recognition, especially for cases involving perceptual aliasing [3], [15]. In the existing works [16], [17], [18], geometric verification is performed with Random Sample Consensus (RANSAC) algorithm. However, RANSAC is generally applicable for local descriptors but not for local region descriptors. Specifically, it usually performs badly when the number of outliers is more than 50% [19], but it is difficult to guarantee that more than half of the matched image regions are correct (see Fig. 1). Fortunately, only a small quantity of correct matches is needed for successful place recognition, the key is therefore to develop an effective geometric check mechanism for filtering out the correct matches.

To bridge this gap, we further pursue the idea of combining the appearance and geometric information and propose a novel local region-based VPR approach called CoCALC, building on earlier HOG-based studies [8], [20]. The proposed method adopts the local-entropy-based regional proposal strategy, which can extract information-rich image patches. In particular, we construct a lightweight CNN architecture to reconstruct the HOG descriptor from image patches in a self-supervised way and combine them to generate a holistic image representation. Unlike the previous work [8] which uses max-pooling to find the best-matched (i.e., maximum similarity score) image region pairs, the proposed CoCALC here employs a geometry-associated band matrix to remove the semantically similar but spatially wrong-matched regions caused by perceptual aliasing (discussed in detail, in Section III-C). Evaluations on several benchmark datasets illustrate the state-of-the-art performance of the method in coping with diverse scenarios and extreme changes.


(a) local descriptors

(b) local region descriptors

FIGURE 1. The matched (a) local descriptors and (b) local region descriptors. The matches of (a) are mostly correct but (b) are not; hence RANSAC is only suitable for local descriptors to remove the outlier. Fortunately, only a certain number of matches are sufficient for place recognition, thus it is not necessary to pick out all correct matched region pairs.

To sum up, this work has three main contributions as follows:

- We present a local region-based descriptor for VPR and integrate it into a compact pipeline that effectively combines appearance and geometric information. The proposed approach is validated in five benchmark datasets, achieving state-of-the-art matching performance and remarkable generalization.
- We design a self-supervised and ultra-lightweight DSC-based network and any scene-centric large-scale datasets can be used for model training, which greatly enhances the practicability and ease of use.
- We propose a band-matrix-based geometric verification method specifically crafted for local region-based descriptors. This method can fully utilize the spatial geometric information and efficiently discard the erroneously matched image patches.

The rest of this article is organized as follows: We provide an overview of the related approaches that have influenced this work in the next section. In Section III we lay out the details of the proposed approach. In Section IV we test the proposed method with several benchmark datasets. A brief discussion in Section V, with a description of future research directions, concludes this work. We also provide all relevant codes at https://github.com/LiKangyuLKY/CoCALC-VPR for further studies.

II. RELATED WORK

The core need of an appearance-based VPR system is to describe places accurately and effectively, and this requirement has spawned many methods [2], [21], which can be generally categorized into local-, global-, and local-region-descriptor-based approaches.

A. LOCAL-DESCRIPTOR-BASED APPROACHES

Early handcrafted local descriptor methods for place recognition are generally based on gradient histograms such as Scale-Invariant Feature Transform (SIFT) [22] and Speeded Up Robust Features (SURF) [23]. These descriptors possess certain invariance to scale, rotation, illumination, and noise; hence they have been widely used for visual localizing [24], [25] and place recognition [5], [6], [15]. The work done in FAB-MAP [5] and FAB-MAP2.0 [15] are considered to be the classical SURF-based approaches. These methods, though simple and clear, are suffered from expensive computational costs and prohibitive memory consumption that are hard to run in real-time. To overcome these problems, one line of thought is to reduce the dimensionality of descriptors, a typical example is PCA-SIFT [26]. Another line of thought is that feature extraction and matching can be significantly sped up by encoding the descriptors in the binary format, methods include BRIEF [27], ORB [28], BRISK [29], and FREAK [30]. The most popular visual SLAM systems, namely ORB-SLAM2 [16] and its optimization version ORB-SLAM3 [17], utilized ORB as the descriptor and quantified them as visual words to represent the place.

More recently, numerous end-to-end learning-based approaches have also been applied to generate more robust local descriptors. Some previous works like [31], MatchNet [32], and LIFT [33] were demonstrated to outperform traditional descriptors in terms of matching quality. However, the descriptor encoding of them is very time-consuming even using GPUs. This adverse impact offsets their advantage of matching performance in VPR tasks. To address this problem, more attention has been paid to not only matching performance but also computational efficiency. Later learning-based local descriptors, such as PN-Net [34], LF-NET [35], and SuperPoint [36], were able to achieve superior matching performance and real-time inference with GPU acceleration.

B. GLOBAL-DESCRIPTOR-BASED APPROACHES

With these approaches, an image is represented as a compact feature vector. Global descriptors can be directly generated by extracting the global features of the images. Two commonly used global descriptors are Gist [37] and HOG [38]. Murillo and Kosecka [39] presented a panorama matching approach for recognizing the revisited places, promoting the application of Gist for VPR tasks. Shortly thereafter, Singh and Kosecka [40] conducted extensive experiments in a 13-mile urban area, demonstrating that the Gist descriptor is competent for large-scale place recognition. HOG descriptor can extract the structure information of the images by calculating the gradients and orientations of each pixel, which can yield good performance on VPR datasets with slightly changed viewpoints [41]. Alternatively, a global descriptor can be generated by aggregating the local descriptors via Bag of Visual Words (BoVW) [42] or Vector of Locally Aggregated Descriptors (VLAD) [43]. These approaches have also been confirmed as valid and efficient models for place recognition [3], [5], [17], [44], [45], particularly when used in conjunction with the inverted index.

After the seminal work of Chen et al. [46], research has increasingly focused on CNN-based methods and many authors have introduced the off-the-shelf CNN model (e.g., AlexNet [47], VGG [48], ResNet [49]) to the field of VPR [50], [51], [52]. A meaningful conclusion was drawn by Sünderhauf et al. [50] that the intermediate layers of the network are more robust against appearance change and this phenomenon has also been documented by Hou et al. [51]. Afterward, some authors customized the novel CNN architecture specifically for VPR tasks instead of using the off-the-shelf model. In particular, considering the clear physical meaning of handcrafted descriptors, some successful attempts such as NetVLAD [53], MobileNetVLAD [54], Convolutional Autoencoder for Loop Closure (CALC) [20], and E^2BoWs [55] have been carried out to redevelop handcrafted-based approaches through learning-based customized networks. Generally, the generalization performance of learning-based approaches is directly related to the training dataset, several relevant datasets are therefore presented in VPR tasks, such as Pittsburgh [56], Places365 [57], and Specific Places Dataset (SPED) [58]. Recently, the advents of FILD++ [59], HEAPUtil [60] and CosPlace [61] has facilitated rapid advancements in VPR domain.

C. LOCAL-REGION-DESCRIPTOR-BASED APPROACHES

This kind of approach aims to combine the advantages of the aforementioned local- and global-descriptor-based approaches. The crucial point in these studies is to find the most salient and distinct regions. In the context of VPR, Zaffar et al. [8] proposed a simpler but effective entropy-based mechanism to extract ROI. Gao and Zhang [62] select the detected keypoints with maximum feature response and then resize them into image patches. Sünderhauf et al. [13] utilized Edge Boxes [63] algorithm to extract the potential landmarks. Recent learning-based advances in object detection, such as RPN [64] and YOLOv3 [65], offer more adaptable solutions and have been introduced in VPR tasks. Notably, CNN-based region extraction approaches have good adaptability and generalization with regard to coping with severe appearance variations, nonetheless at the expense of considerable computing resources. Therefore, we propose a handcrafted-based region extractor to enhance the practicality of a resource-constrained mobile robot.

Another important aspect is to generate descriptors from the selecting salient region, a process similar to the global-descriptor-based approaches. As a consequence, more promising approaches such as CoHOG [8], Region-VLAD [66], and R-VLAD [67] have been developed. Of course, the most direct route for feature extraction is also to use a pre-trained off-the-shelf CNN model. Although their feasibility has been extensively validated by the work done in [11], [13], and [14], many improved or novel CNN architectures have been proposed, usually trained on specific VPR datasets, to improve the performance. An impressive approach is Patch-NetVLAD [10] which generates region-level features through NetVLAD [53], achieving superior VPR performance in challenging datasets. Interestingly, region extraction and descriptor generation can be merged into a compact process. Unlike the aforementioned approaches that rely on external region detectors, Chen et al. [14] discovered the salient regions directly from the CNN activations by finding the highest averaging activations energies, and computational cost can be reduced to a certain extent. In their later research [68], a multi-scale context-flexible attention model was presented to identify the salient regions. Xin et al. [69] designed and trained the landmark localization network (LLN) with image-level annotations to select the discriminative landmarks. However, the training of the above supervised or semi-supervised learning approaches is a multistep process that involves data annotating and complicated pretreatments, unlike the self-supervised mechanism used in this work.

III. METHODOLOGY

In this section, the proposed CoCALC approach is described in detail, and the entire workflow of the proposed approach is shown in Fig. 2. Our approach first uses the entropy map to extract the top-K most information-rich regions from an image; the image is resized and converted to grayscale in this process. We then compute and combine the descriptors for each image patch using a well-trained network, thus an image can be represented as an ensemble matrix holding K regional descriptors. Finally, the similarity between the query image and the candidate database image is measured using the cross-matching of the region-level descriptors, where the erroneous matching pairs of image patches are removed based on a band-matrix-based geometric check.

A. REGION EXTRACTION

Inspired by [8], we propose an local-entropy-based approach to extract the top-K most salient region of the input image. The raw image is resized to fixed-size ($L \times L$ pixels) with an aspect ratio of 1:1 and converted to grayscale to facilitate the detection of subtle changes in the gray level distribution and calculational simplicity. We can obtain the entropy map of an image (see Fig. 3(b)) where the entropy is calculated from the histogram of intensity using base a 2 logarithm

$$H = -\sum_{0}^{255} p_i \log_2 p_i \tag{1}$$

where p_i denotes gray value distribution of pixel i. The image is then divided into $n \times n$ small square patches, that is, the size of each patch is $(L/n) \times (L/n)$. Therefore, an image and its entropy map are represented as $n \times n$ matrix R and E, respectively

$$R = \begin{pmatrix} r_{11} & \dots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{pmatrix} \tag{2}$$

$$E = \begin{pmatrix} e_{11} & \dots & e_{1n} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nn} \end{pmatrix} \tag{3}$$

For image patch r_{ij}, its average entropy value e_{ij} is computed. Based on the sorted patches in the descending order of the average entropy value, K patches are selected for the subsequent processes. Formally, the selected image patches are represented as

$$R_K := \{r | r \in R \wedge e^r \in max_K(E)\} \tag{4}$$

where e^r is the average entropy value of the patch r, and $max_K(E)$ denotes the top-K maximum elements in matrix E.

Fig. 3(c, d, e) illustrates the top-$K = \{50, 100, 150\}$ regions/patches picked through the proposed entropy-based approach. It is clearly seen that the low-entropy regions such as large areas of sky and snowfield are discarded. This intuitively makes sense, since ground and sky (ceiling in an indoor case) usually cover large portions of the image and look similar in different places, which is the major trigger for perceptual aliasing [9]. Additionally, filtering out confusing regions can significantly reduce the computational cost.

B. SELF-SUPERVISED NETWORK

1) NETWORK ARCHITECTURE

The efficiency of several off-the-shelf CNN architectures, such as wider GoogLeNet [70] and deeper ResNet [49], has been confirmed by the computer vision community. However, it is not appropriate to directly apply them to VPR tasks, particularly in cases where we need to re-formulate the HOG descriptor for image representation. This is because (i) they do not take into account the specificity of VPR tasks, for example, higher layers in deeper networks are more semantically meaningful but more susceptible to perceptual aliasing [50]; (ii) due to the tremendous parameters and expatiatory sequential processing, their computational cost is too expensive to satisfy the real-time requirements for the robot; (iii) they contain more semantic information but less structural information, as well as unsuitable architecture, making it difficult to reconstruct a HOG descriptor.

Given that, a lightweight network is designed and adopted in the proposed CoCALC. As shown in Fig. 4, the network

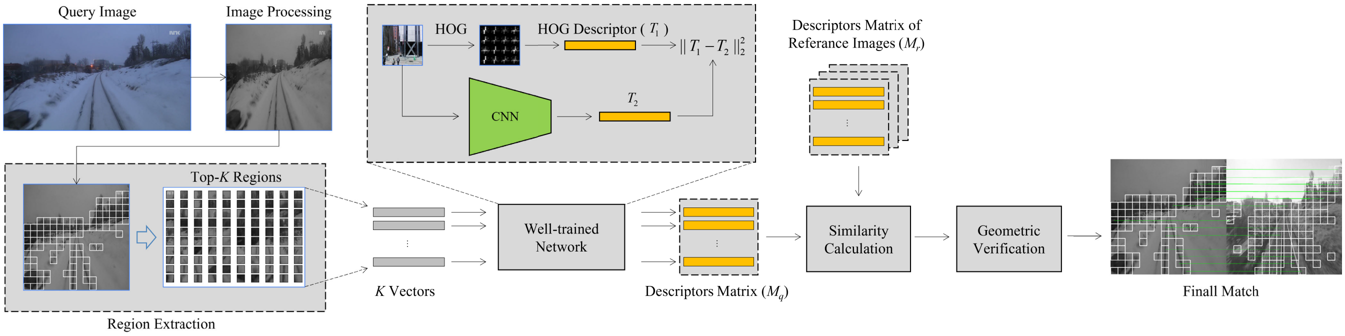


FIGURE 2. The entire workflow of the proposed CoCALC approach is shown here.

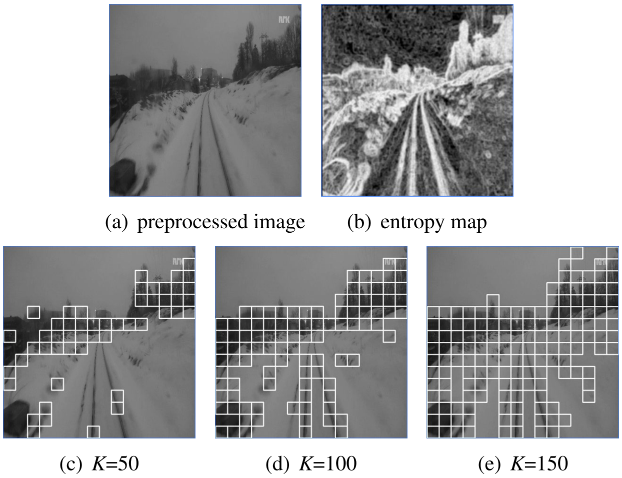


FIGURE 3. Entropy-based salient region extraction is shown here. The intensity of the entropy map ranges from light (white) to dark (black), and the darker the color, the lower the entropy.

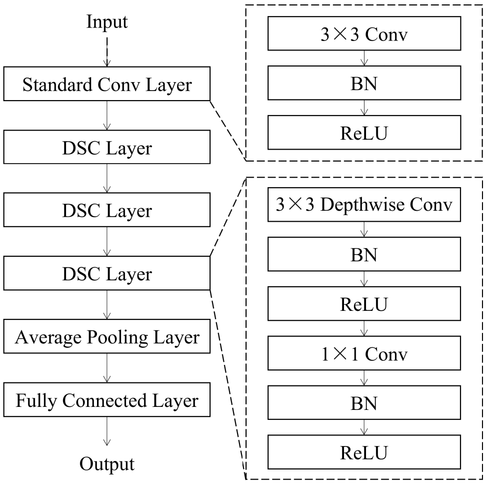


FIGURE 4. Backbone architecture of the proposed lightweight network.

architecture is a simplified version of MobileNet [71] and is built similarly on the Depthwise Separable Convolution (DSC). DSC has been introduced in detail in [71] and [72] so we will not further describe it here. The proposed network consists of a standard 3 × 3 Conv layer, three consecutive DSC layers, an average pooling layer, and a final fully connected layer. Note that all Conv layers are followed by batch-norm and ReLU. Our motivation for using DSC is that its smaller number of parameters serves to alleviate over-fitting and reduce computational costs. Consequently, the model size of our ultra-lightweight network is only 0.27 MB and the total number of parameters is 71,876.

2) NETWORK TRAINING AND INFERENCE

As mentioned previously, the proposed network is trained in a self-supervised way, which is implemented as follows. During the data loading process, the entire image is first resized to $W \times H \times C$, where W, H, and C denotes the width, height, and channel, respectively. Then, the processed image samples with batch size N are fed into two parallel pipelines (see Well-trained Network module in Fig. 2), where the first pipeline is a HOG descriptor extraction and the second pipeline is our lightweight network described in the previous subsection. In particular, the feature tensors generated from the first pipeline (denoted as T_1) and the second pipeline (denoted as T_2) have the same dimension of $W \times D$, where D denotes the dimension of a HOG descriptor. Finally, the model will iteratively learn to adjust the weights to minimize the distance between T_1 and T_2, where the distance is measured using the mean squared $L2$ norm

$$Loss = \frac{1}{N}\sum_{n=1}^{N}(T_1(n) - T_2(n))^2 \qquad (5)$$

In this way, any unannotated images can be fed into the proposed network because they will be automatically labeled without manual intervention.

Benefiting from the compact architecture and a small number of hyperparameters, the proposed network usually converges quickly and smoothly, obtaining the capability of extracting more robust features. In the inference phase, it should be noted that the top-K image patches were fed into the well-trained network in batches with a batch size equal to K. Benefiting from parallel inference, we can combine all selected patches and process them simultaneously, which

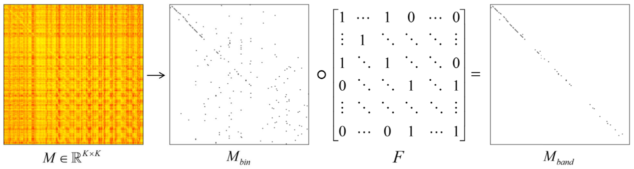


FIGURE 5. Visualization of matrix binarization and band-matrix-based filtering. '∘' denotes the Hadamard product.

results in higher throughput and higher efficiency. We have verified that when using parallel inference, the time consumption of processing one image patch and 200 image patches is almost the same. Finally, an input image containing K salient regions can be represented as a matrix $I \in \mathbb{R}^{K\times D}$, where each region is described by a $1 \times D$ descriptor.

C. GEOMETRIC-BASED SIMILARITY MEASUREMENT

For the proposed region-based approach, the similarity measurement can be divided into three steps: (i) calculating the cosine distance of each matched region pair; (ii) discarding the low-reliability matched region pairs; (iii) measuring the overall similarity of two images.

Given a query image containing K salient regions, it can be represented as a 2-dimensional matrix $I_q \in \mathbb{R}^{K\times D}$, as described above. Similarly, a database image can also be represented as $I_{db} \in \mathbb{R}^{K\times D}$. Two representation matrixes I_q and the transposed I_{db}^{T}, when multiplied, will produce a new matrix $M \in \mathbb{R}^{K\times K}$, where each row of matrix M indicates the cosine similarity between a region of a query image and all regions of a database image.

Obviously, not all matched region pairs are true-positive results, thus we present a band-matrix-based geometric check mechanism to remove wrong matching. Considering the most appropriate matched region pairs should have the largest similarity score, we first impose a binarization to matrix M so that the maximum value of each row is assigned to 1 and the rest are 0, as indicated in Fig. 5, thus we get a matrix M_{bin}. In particular, for the purpose of filtering out the outlier, we can then obtain a matrix M_{band} by calculating the Hadamard product for matrix M_{bin} with a band matrix F:

$$F = (e_{ij})_{K\times K}, \quad e_{ij} = 0 \ \ if \ |j - i| > d \ \ else \ \ e_{ij} = 1 \tag{6}$$

where the bandwidth of this diagonally band is determined by d.

The motivation for this is that two best matched regions should be geometrically adjacent to each other, despite the variation in viewpoint. More concretely, elements close to the diagonal of M_{bin} represent the more reliable matching. Finally, the overall similarity score between a query image and a database image can be calculated by taking the average of the non-zero value in matrix M_{band}.

IV. EXPERIMENTAL RESULTS

A. IMPLEMENTATION DETAILS

To achieve better VPR performance, we used grid search to exhaustively set different parameter combinations and test the effectiveness of these settings on the Gardens Point datasets. Finally, we obtained the appropriate configuration and employed it across all experiments. For the salient regions extraction module, the image size was set to $L = 512$ and $n = 16$, that is, a 512 by 512 pixels image is divided into 162 regions of 32×32 pixels each; the parameter of HOG was set to cell-size = 8×8, window-size = block-size = 16×16, stride = 16, bin-size = 9, thus a 32×32 image region can be represented as a 324-dimensional HOG descriptor.

As for the training module, our open-source approach is implemented in Pytorch using Adam optimizer with a learning rate of 1×10^{-4} and a decay factor of 0.5. The network is trained on the *Pittsburgh 250k* dataset, which contains diverse scene images downloaded from Google Street View. The epoch is set to 200 and the batch size is set to 512. To remain consistent with the extracted image region, the input image is also resized to 32×32 and converted to grayscale before entering the network. The training was performed on an Ubuntu 18.04 LTS operating system running on an Intel Xeon E5-2678 V3 CPU @ 2.5GHz and eight RTX 2080Ti GPUs.

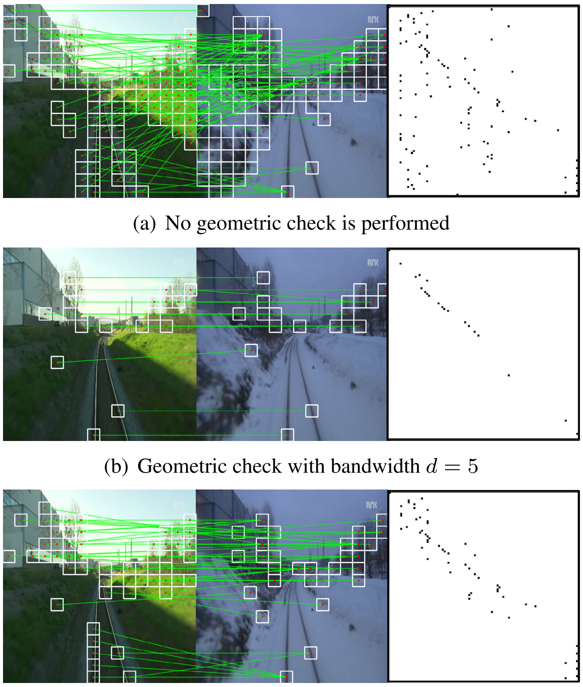
TABLE 1. The ground-truth tolerance used in our experiments.

Nordland	SPEDTest	Gardens Point	Campus Loop	Cross-Seasons
±1 frames	frame-to-frame	±2 frames	frame-to-frame	±5 frames

B. DATASETS AND EVALUATION METRICS

We carried out evaluation studies to validate the proposed CoCALC approach on 5 benchmark datasets, covering diversity scenarios and complicated conditions. Here we provide a brief introduction to these datasets to facilitate analyzing the performance of CoCALC against diversity challenges. For challenges brought by appearance change, we use the **Nordland** dataset [73], which collects both natural and urban landscapes in four seasons using a camera mounted in front of the train, but no viewpoint changes are involved due to the fixed track route. Another dataset we used is **SPEDTest** [58], which captured in diverse seasons and illumination conditions but again no viewpoint changes. Then, we use the **Gardens Point** dataset [74], which contains two day-time traversals with images recorded on the left and right sides of the walking path, resulting in strong variations in viewpoints and illumination. For comprehensive challenges, we use **Campus Loop** [20] and **Cross-Seasons** dataset [75], which were both captured under extreme viewpoint and appearance changes caused by diverse illumination, weather, or seasonal conditions. Furthermore, the judgment tolerance of the same place for those datasets is different because of the divergence in running speed and shooting frequency, thus the ground-truth tolerance we used in our experiments was presented in Table 1.

An ideal VPR method aims to achieve 100% precision and 100% recall, nevertheless, negative correlations were



(a) No geometric check is performed

(b) Geometric check with bandwidth $d = 5$

(c) Geometric check with bandwidth $d = 20$

FIGURE 6. Visualization of the band-matrix-based geometric check. Drawn in the color images for better visualization.

observed between precision and recall. Thus, Precision-Recall (PR) curves and Area-under-the-PR curves (AUC) are usually used to assess the comprehensive VPR performance [3], [5], [7], [20]. The other two metrics used in our study are recall rate at 100% precision (R_{P100}, for short) and *Recall@N*. As for the computational performance, we take the feature encoding time and descriptor matching time into consideration.

C. VISUALIZATION OF BAND-MATRIX-BASED GEOMETRIC CHECK

We visualize the band-matrix-based geometric check for a more intuitive illustration. From left to right are the visualizations of query image I_q, database image I_{db} and their binarized similarity matrix M_{bin}. Fig. 6(a) demonstrates that no geometrical verification is performed on it. We can find that not only most of the matches are incorrect, but multiple different regions in I_q are also erroneously matched to one region in I_{db}.

Two examples of using band-matrix-based geometric check are illustrated in Fig. 6(b, c). Clearly, most wrong matches are successfully filtered out when bandwidth is set to 5, and as the bandwidth grows more correct matches are retained but fugitives also increase. These results demonstrates the effectiveness of the proposed methods for retrieving true-positives from matches containing a larges of outliers. It is important to point out that even a small number of correct matched region pairs are sufficient for robust place recognition, as shown later in sub-section IV-E.

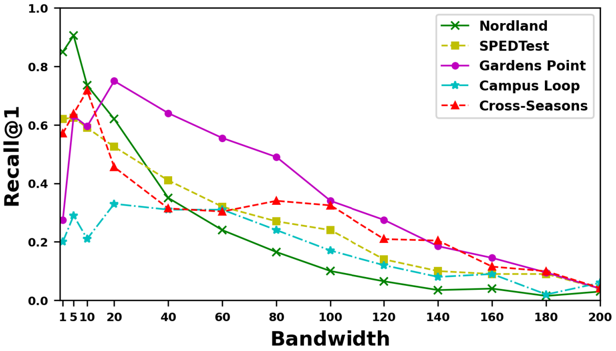


FIGURE 7. The *Recall@1* at different bandwidths evaluated on five datasets are presented here.

D. EXPERIMENT ON DIFFERENT BANDWIDTH

To further illustrate the impact of the bandwidth d used for the band-matrix-based geometric check, we evaluate the *Recall@N* (*N* is set to 1) performance at different bandwidths. The motivation behind using *Recall@1* is that this metric reflects the percentage of correctly recognized query images. The value of d is set to $\{1, 5, 10, 20, 40, \ldots, 200\}$, where 1 denotes only the diagonal elements of M_{bin} are taken and 200 means that no geometric verification is performed. As shown in Fig. 7, we find that all curves have similar trends, that is, as the bandwidth d grows, the curves of *Recall@1* quickly reach the maximum values and then smoothly decrease. The maximum values of *Recall@1* occur in the range of $5 \leqslant d \leqslant 20$. These phenomena show the duality of bandwidth size, too small d may lead to the loss of potentially correct matches whereas too large d will aggravate the interference of wrong matches. Overall, the proposed band-matrix-based approach with an appropriate bandwidth can bring considerable improvement in place recognition performance.

E. COMPARISON WITH STATE-OF-THE-ART APPROACHES

1) EXPERIMENTAL SETUP

In this study, we compare the performance of the proposed CoCALC against several VPR approaches, including HOG [38], two earlier HOG-based approaches CALC [20] and CoHOG [8], and three recent works NetVLAD [53], MobileNetVLAD [54], and Patch-NetVLAD [10]. These approaches were implemented with the standard configurations recommended by their authors. Notably, CoCALC, CALC and NetVLAD are all trained on the *Pittsburgh 250k* dataset. MobileNetVLAD is initially proposed for 6-DoF pose estimation, which uses knowledge distillation to transfer the knowledge of NetVLAD to a more lightweight network. Here, we integrated it into our work as a reference for comparison. Patch-NetVLAD is one of the most representative local region descriptors and achieves state-of-the-art VPR results validated on several challenging datasets. We used the official code and pre-trained models provided by Patch-NetVLAD's authors.

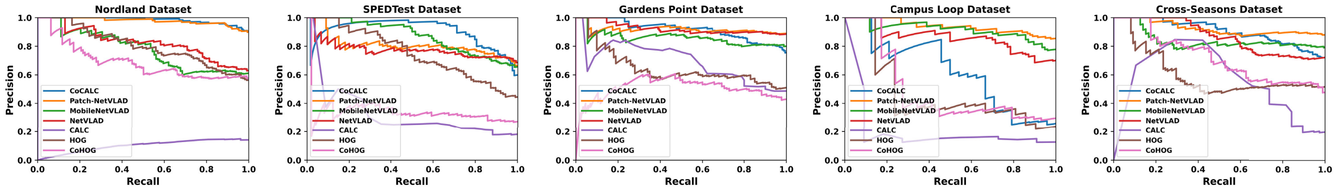


FIGURE 8. Here are the comparison results of PR curves generated on five benchmark datasets.

TABLE 2. The values of AUC and recall rate at 100% precision (R_{P100}) are listed here.

Approaches	Norland		SPEDTest		Gardens Point		Campus Loop		Cross-Seasons	
	AUC	R_{P100}	AUC	R_{P100}	AUC	R_{P100}	AUC	R_{P100}	AUC	R_{P100}
HOG	0.80	0.16	0.75	0.15	0.65	0.06	0.46	0.14	0.60	0.09
CoHOG	0.69	0.06	0.34	0.02	0.51	0	0.49	0.17	0.72	0.18
CALC	0.10	0	0.33	0.03	0.69	0.03	0.20	0	0.56	0
NetVLAD	0.83	0.13	0.78	0.04	0.91	0.15	0.86	0.18	0.87	0.34
MobileNetVLAD	0.78	0.11	0.87	0.21	0.86	0.06	0.91	0.12	0.84	0.13
Patch-NetVLAD	0.97	0.31	0.81	0.09	0.92	0.05	0.94	0.27	0.92	0.17
CoCALC	0.98	0.51	0.90	0.02	0.92	0.23	0.63	0.13	0.91	0.23

The PR curves for each approach are presented in Fig. 8, and the AUC and R_{P100} are listed in Table 2. The value of bandwidth d was set depend on the optimal results in Fig. 7 (for example, $d = 5$ for Nordland and SPEDTest dataset), while Top-K was set to 200 and used across all experiments. It should be pointed out that it may not be the optimal setting for each test dataset, that is, other values of Top-K may yield better VPR results.

2) EXPERIMENTAL RESULTS

a: NORDLAND DATASET

As showcased in Fig. 8 and Table 2, the PR curve and AUC of the proposed CoCALC outperform all other approaches yet the improvement of performance are not remarkable. Unexpectedly, there is a large gap in the both AUC and PR curves between the CoCALC and other approaches, with the exception of Patch-NetVLAD, which achieved nearly the same performance. This implies that the performance boost of our approach originates from the cooperation between salient region features. As for the R_{P100} value, it should be emphasized that we set a very strict ground-truth tolerance, so it is difficult for these tested approaches to obtain a high R_{P100} value. Furthermore, images collected from similar scenes introduce strong perceptual aliasing into Nordland dataset. Even though, the proposed CoCALC still achieves a satisfying result with an R_{P100} of 0.51, which is remarkably higher than the other approaches. The improvement of the performance mainly benefits from the band-matrix-based geometric check mechanism.

b: SPEDTEST DATASET

Similar to the Nordland dataset, the SPEDTest dataset exhibits extreme appearance variations but no viewpoint variations and is therefore helpful for evaluating the properties in response to the single appearance change factor. Fig. 8 and Table 2 show that the proposed CoCALC still achieves the best results in terms of PR curve and AUC, demonstrating its robustness against complex conditions including illumination, weather or season changes. However, it is an extremely challenging task for all approaches to find true-positive results, especially under the premise of 100% precision and tight frame-to-frame tolerance. It can be observed that these lead to bad R_{P100} values for almost all approaches. HOG produces a relatively good result for both AUC and R_{P100}, indicating that it is good at handling datasets without viewpoint changes.

c: GARDENS POINT DATASET

Fig. 8 shows that in this case, the performance of CoCALC is significantly better than the other three HOG-based approaches (i.e., HOG, CALC, and CoHOG), while NetVLAD and Patch-NetVLAD performs on par with our proposed CoCALC. The outperformance of CoCALC is also visible in Table 2, where we achieve the highest AUC (0.92) and R_{P100} (0.23). Further, we also noted that the PR curves of CoCALC, NetVLAD, and Patch-NetVLAD decrease gently after the R_{P100} value, whereas the PR curves of the other methods decline rapidly. This confirms that CoCALC can still maintain satisfying precision under gradually higher recall performance. We also noticed that MobileNetVLAD can achieve (and sometimes even surpass) NetVLAD-level VPR results. Similarity to MobileNetVLAD, the proposed CoCALC is also built upon the DSC-based architecture, demonstrating the potential of lightweight CNN in VPR tasks.

d: CAMPUS LOOP DATASET

Two sequences in this dataset were separately captured on a sunshiny day and a cloudy snowy day, which provides strong variations in viewpoint and appearance as well as

TABLE 3. Time (in milliseconds) for feature encoding (t_e) as well as descriptor matching (t_m), using a CPU-based platform (Intel i5-10500TE CPU @ 2.30GHz, 16GB RAM).

Metric	HOG	CoHOG	CALC	NetVLAD	MobileNetVLAD	Patch-NetVLAD	CoCALC	CoCALC*
Input size	32×32	512×512	160×120	224×224	224×224	224×224	512×512	512×512
t_e	0.12	26.80	10.34	205.71	36.86	880.48	86.27	1.02
t_m	0.003	-	0.004	0.019	0.019	0.004	0.274	0.274
Dimensions	1×324	-	1×3648	1×32768	1×7680	1×4096	200×324	200×324

Note that the t_e of CoCALC was computed throughout the entire feature encoding procedure, while that of CoCALC does not include the process of region extraction.

many dynamic objects. Therefore, all approaches suffer on these synthetic challenges, which result in degraded performance. In comparison to other approaches, Patch-NetVLAD achieves the best performance in both AUC and R_{P100}, closely followed by MobileNetVLAD and NetVLAD. In this case, our proposed CoCALC is slightly inferior to three NetVLAD-based approaches but significantly outperforms other early HOG-based approaches. This indicates that the proposed lightweight network can extract robust representation descriptors even though it is only 0.27 MB.

e: CROSS-SEASONS DATASET

This dataset is captured by a car-mounted camera under long-term changing conditions and therefore consists of many different combinations of viewpoints, illumination, weather and dynamic objects. The AUC and R_{P100} value in Table 2 show that CoCALC, NetVLAD, and Patch-NetVLAD each have strong points. Patch-NetVLAD achieves the best performance in AUC (0.92) but NetVLAD achieves better R_{P100}(0.34), while CoCALC produces a more balanced results. In terms of PR curves, it can also be observed that the precision at 100% recall of CoCALC and NetVLAD is at the same level, but the curve of CoCALC declines relatively slowly. CALC cannot meet these combined challenges to the same level as CoCALC, despite they are both trained on the Pittsburgh 250k dataset.

F. COMPUTATIONAL PERFORMANCE

We now discuss the computational performance of the proposed CoCALC. The experiment was performed on the Gardens Point dataset which contains 200 query images and 200 database images, and the image resolution is 960 × 540 pixels. The feature encoding time (denoted as t_e) and descriptor matching time (denoted as t_m) was computed by averaging processing time over the entire dataset, and the results are listed in Table 3. Note that a unified CPU-only platform was used for both conventional and CNN-based approaches, whereas CNN-based ones generally require more computational resources.

From Table 3, we can find that HOG descriptor achieves the fastest feature encoding of only 0.12 ms. Benefiting from the lightweight architecture, the inference speed of CoCALC is almost 7 times faster than CALC. Admittedly, region extraction is the most time-consuming part of CoCALC, but it is still faster than NetVLAD and Patch-NetVLAD on the CPU platform. This shows the practical value of CoCALC on resource-constrained robotic systems without GPUs when considering its state-of-the-art matching performance and acceptable efficiency. For the descriptor matching time, the best and second results are achieved by HOG and CALC. CoHOG was not included in the test due to its variable descriptor dimension. Descriptor matching for CoCALC spends 0.274 ms on average due to the introduction of geometric check. Even though, it is highly desirable driven by the considerable improvement in matching performance.

TABLE 4. Time (in milliseconds) for feature encoding t_e, using a GPU-based platform (Nvidia 2080Ti GPU 11GB RAM).

Approaches	# Params(M)	FLOPs(G)	t_e
CALC	9.80	0.21	0.88
NetVLAD	14.74	15.36	14.30
MobileNetVLAD	0.95	0.06	22.82
Patch-NetVLAD	14.72	15.35	53.45
CoCALC	0.07	0.003	89.63
CoCALC*	0.07	0.003	0.84

Although this work was developed for the resource-limited devices without GPU acceleration, we also report the encoding time of CNN-based approaches using a GPU platform in Table 4. Accelerated with the GPU, feature encoding for CALC and NetVLAD achieves significant speed boosts. Since the most time-consuming region extraction cannot benefit from GPU acceleration, no obvious improvement is observed in the encoding time of CoCALC.

V. CONCLUSION

To address the dual challenge of extreme variation in viewpoint and appearance, research has increasingly focused on local-region-descriptor-based approaches. We took a step in this direction and presented a self-supervised approach CoCALC that incorporates appearance and geometric information. Since the annotation information is automatically generated using the HOG, less manual data preparation is required during the training. The proposed lightweight network is constructed on the alternating DSC layer and used to extract features from salient image patches that have relatively higher entropy. The proposed CoCALC improves the VPR performance by a simple but effective band-matrix-based geometric check, and the impact of the bandwidth is discussed. Assessment of CoCALC on several benchmark datasets revealed that it yields state-of-the-art performance and satisfying generalization.

Future research should consider the potential effects of inserting another descriptor as the automated annotation scheme. In addition, combining semantic information into the proposed approach may achieve a further performance boost. We are currently working on integrating the proposed CoCALC into a visual SLAM system of automated forklifts.

REFERENCES

[1] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," *IEEE Robot. Autom. Mag.*, vol. 13, no. 2, pp. 99–110, Jun. 2006.

[2] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2015.

[3] D. Galvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1188–1197, Oct. 2012.

[4] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. Tardós, "A comparison of loop closing techniques in monocular SLAM," *Robot. Auton. Syst.*, vol. 57, no. 12, pp. 1188–1197, Dec. 2009.

[5] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *Int. J. Robot. Res.*, vol. 27, no. 6, pp. 647–665, 2008.

[6] M. Labbé and F. Michaud, "Appearance-based loop closure detection for online large-scale and long-term operation," *IEEE Trans. Robot.*, vol. 29, no. 3, pp. 734–745, Jun. 2013.

[7] E. Garciafidalgo and A. Ortiz, "iBoW-LCD: An appearance-based loop-closure detection approach using incremental bags of binary words," in *Proc. Int. Conf. Robot. Autom.*, May 2018, vol. 3, no. 4, pp. 3051–3057.

[8] M. Zaffar, S. Ehsan, M. Milford, and K. McDonald-Maier, "CoHOG: A light-weight, compute-efficient, and training-free visual place recognition technique for changing environments," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1835–1842, Apr. 2020.

[9] C. Cheng, D. L. Page, and M. A. Abidi, "Object-based place recognition and loop closing with jigsaw puzzle image segmentation algorithm," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2008, pp. 557–562.

[10] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14141–14152.

[11] P. Neubert and P. Protzel, "Beyond holistic descriptors, keypoints, and fixed patches: Multiscale superpixel grids for place recognition in changing environments," *IEEE Robot. Autom. Lett.*, vol. 1, no. 1, pp. 484–491, Jan. 2016.

[12] Z. Xin, X. Cui, J. Zhang, Y. Yang, and Y. Wang, "Real-time visual place recognition based on analyzing distribution of multi-scale CNN landmarks," *J. Intell. Robotic Syst.*, vol. 94, nos. 3–4, pp. 777–792, Jun. 2019.

[13] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," in *Proc. Robotics, Sci. Syst. XI*, 2015, pp. 1–10.

[14] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from ConvNet for visual place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 9–16.

[15] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *Int. J. Robot. Res.*, vol. 30, no. 9, pp. 1100–1123, Jun. 2011.

[16] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[17] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source slam system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[18] N. Merrill and G. Huang, "CALC2.0: Combining appearance, semantic and geometric information for robust and efficient visual loop closure," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 4554–4561.

[19] A. Hast, J. Nysjö, and A. Marchetti, "Optimal RANSAC—Towards a repeatable algorithm for finding the optimal set," *J. WSCG*, vol. 21, no. 1, 2013, pp. 1–10.

[20] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," 2018, *arXiv:1805.07703*.

[21] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognit.*, vol. 113, May 2021, Art. no. 107760.

[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[23] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.* Graz, AUSTRIA: Springer, 2006, pp. 404–417.

[24] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *Int. J. Robot. Res.*, vol. 21, no. 8, pp. 735–758, 2002.

[25] A. C. Murillo, J. J. Guerrero, and C. Sagues, "SURF features for efficient robot localization with omnidirectional images," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, pp. 3901–3907.

[26] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2004, p. 2.

[27] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 778–792.

[28] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

[29] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2548–2555.

[30] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 510–517.

[31] K. Simonyan, A. Vedaldi, and A. Zisserman, "Learning local feature descriptors using convex optimisation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1573–1585, Aug. 2014.

[32] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3279–3286.

[33] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned invariant feature transform," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 467–483.

[34] V. Balntas, E. Johns, L. Tang, and K. Mikolajczyk, "PN-Net: Conjoined triple deep network for learning local image descriptors," 2016, *arXiv:1601.05030*.

[35] Y. Ono, E. Trulls, P. Fua, and K. M. Yi, "LF-Net: Learning local features from images," in *Proc. Adv. neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[36] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 224–236.

[37] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Prog. Brain Res.*, vol. 155, pp. 23–36, Oct. 2006.

[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.

[39] A. C. Murillo and J. Kosecka, "Experiments in place recognition using gist panoramas," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Sep. 2009, pp. 2196–2203.

[40] G. Singh and J. Kosecka, "Visual loop closing using gist descriptors in Manhattan world," in *Proc. ICRA Omnidirectional Vis. Workshop*, 2010, pp. 4042–4047.

[41] M. Zaffar, S. Garg, M. Milford, J. Kooij, D. Flynn, K. McDonald-Maier, and S. Ehsan, "VPR-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *Int. J. Comput. Vis.*, vol. 129, no. 7, pp. 2136–2174, Jul. 2021.

[42] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 2161–2168.

[43] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.

[44] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

[45] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "Fast and incremental method for loop-closure detection using bags of visual words," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 1027–1037, Oct. 2008.

[46] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," 2014, *arXiv:1411.1509*.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[50] N. Sunderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of ConvNet features for place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 4297–4304.

[51] Y. Hou, H. Zhang, and S. Zhou, "Convolutional neural network-based image representation for visual loop closure detection," in *Proc. IEEE Int. Conf. Inf. Autom.*, Aug. 2015, pp. 2238–2245.

[52] X. Zhang, L. Wang, Y. Zhao, and Y. Su, "Graph-based place recognition in image sequences with CNN features," *J. Intell. Robotic Syst.*, vol. 95, no. 2, pp. 389–403, 2018.

[53] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5297–5307.

[54] P.-E. Sarlin, F. Debraine, M. Dymczyk, R. Siegwart, and C. Cadena, "Leveraging deep visual descriptors for hierarchical efficient localization," in *Proc. Conf. Robot. Learn.*, 2018, pp. 456–465.

[55] X. Liu, S. Zhang, T. Huang, and Q. Tian, "E2BoWs: An end-to-end bag-of-words model via deep convolutional neural network for image retrieval," *Neurocomputing*, vol. 395, pp. 188–198, Jun. 2020.

[56] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 883–890.

[57] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[58] Z. Chen, A. Jacobson, N. Sunderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 3223–3230.

[59] S. An, H. Zhu, D. Wei, K. A. Tsintotas, and A. Gasteratos, "Fast and incremental loop closure detection with deep features and proximity graphs," *J. Field Robot.*, vol. 39, no. 4, pp. 473–493, Jun. 2022.

[60] N. V. Keetha, M. Milford, and S. Garg, "A hierarchical dual model of environment- and place-specific utility for visual place recognition," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 6969–6976, Oct. 2021.

[61] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4878–4888.

[62] X. Gao and T. Zhang, "Unsupervised learning to detect loops using deep neural networks for visual SLAM system," *Auton. Robots*, vol. 41, no. 1, pp. 1–18, 2017.

[63] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 391–405.

[64] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[65] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[66] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight CNNs for significant ViewPoint and appearance changes," *IEEE Trans. Robot.*, vol. 36, no. 2, pp. 561–569, Apr. 2019.

[67] M. Teichmann, A. Araujo, M. Zhu, and J. Sim, "Detect-to-retrieve: Efficient regional aggregation for image search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5109–5118.

[68] Z. Chen, L. Liu, I. Sa, Z. Ge, and M. Chli, "Learning context flexible attention model for long-term visual place recognition," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4015–4022, Oct. 2018.

[69] Z. Xin, Y. Cai, T. Lu, X. Xing, S. Cai, J. Zhang, Y. Yang, and Y. Wang, "Localizing discriminative visual landmarks for place recognition," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 5979–5985.

[70] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[71] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[72] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[73] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging SeqSLAM on a 3000 km journey across all four seasons," in *Proc. Workshop Long-Term Autonomy Int. Conf. Robot. Autom. (ICRA)*, 2013, pp. 1–3.

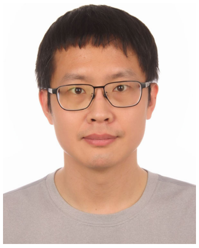
[74] A. Glover, "Day and night, left and right," Mar. 2014.

[75] M. Mans Larsson, E. Stenborg, L. Hammarstrand, M. Pollefeys, T. Sattler, and F. Kahl, "A cross-season correspondence dataset for robust semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9532–9542.

KANGYU LI received the B.S. degree in mechanical design manufacture and automation from the China University of Petroleum (East China), Qingdao, China, in 2013, and the M.S. degree in mechanical design manufacture and automation from the China Academy of Machinery Science and Technology (CAM), Beijing, China, in 2017, where he is currently pursuing the Ph.D. degree. His research interests include visual place recognition, computer vision, and deep learning.

XIFENG WANG received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 1986. He is currently a professor-level Senior Engineer and a Ph.D. Supervisor with the China Academy of Machinery Science and Technology. His research interests include fault diagnosis, machine vision, and condition monitoring.

LEILEI SHI received the B.S. degree in mechanical design manufacture and automation from Wuhan University, Wuhan, China, in 2021. He is currently pursuing the M.S. degree with the China Academy of Machinery Science and Technology (CAM), Beijing, China. His research interests include visual slam and GPU acceleration.



NIUNIU GENG received the B.S. degree from the Harbin Institute of Technology, China, in 2008, and the M.S. degree from the China Academy of Machinery Science and Technology, Beijing, China, in 2011. His research interests include autonomous mobile robots and simultaneous localization and mapping.