Hindawi
Journal of Robotics
Volume 2023, Article ID 9150357, 17 pages
https://doi.org/10.1155/2023/9150357

Review Article

Evaluation of Global Descriptor Methods for Appearance-Based Visual Place Recognition

Kangyu Li ,[1,2] Yuhan Ma,[1] Xifeng Wang,[2] Lijuan Ji,[3] and Niuniu Geng[1]

[1] *Machinery Technology Development Co., Ltd., Beijing 100037, China*
[2] *China Academy of Machinery Science and Technology, Beijing 100044, China*
[3] *China Productivity Center for Machinery, Beijing 101407, China*

Correspondence should be addressed to Kangyu Li; liky@mtd.com.cn

Received 26 September 2022; Revised 20 March 2023; Accepted 8 April 2023; Published 22 April 2023

Academic Editor: Xianfeng Yuan

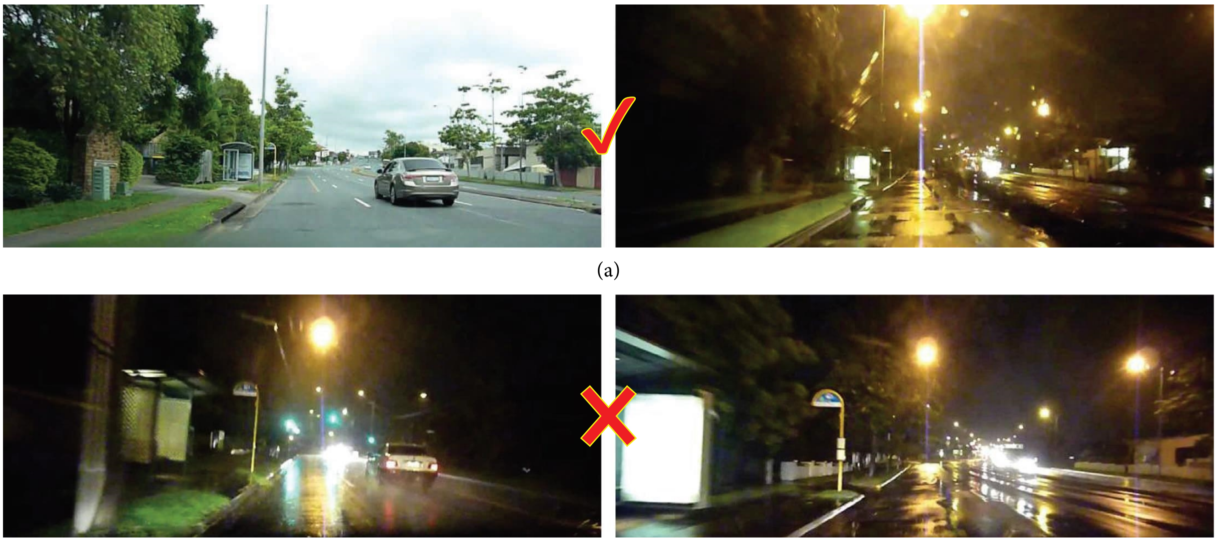
Visual place recognition (VPR) is considered among the most challenging problems due to the extreme variations in appearance and viewpoint. Essentially, appearance-based VPR can be considered as an image retrieval task, thus the key is to accurately and efficiently describe the images. Recently, global descriptor methods have attracted substantial attention from the VPR community, which has contributed to numerous important outcomes. Despite the growing number of global descriptors presented, little attention has been paid to the comparison and evaluation of these methods and so it remains difficult for researchers to disentangle the factors that led to better performance. This study provided comprehensive insight into global descriptors from a practical application perspective. We present a systematic evaluation that integrates 15 commonly used global descriptors, 6 benchmark datasets, and 5 evaluation metrics, and subsequently extended this evaluation to discuss the key factors impacting the matching performance and computational efficiency. We also report practical suggestions for constructing promising CNN descriptors, based on the experimental conclusions. Our analysis reveals both advantages and limitations of three different types of global descriptors, including handcrafted features-based ones, off-the-shelf CNN-based ones, and customized CNN-based ones. Finally, we evaluate the practicality of reported global descriptors to mediate the trade-offs between matching performance and computational efficiency.

1. Introduction

Over the past few decades, visual simultaneous localization and mapping (SLAM) [1] has been considerably advanced in robotics research communities. As one of the essential components in the visual SLAM, visual place recognition (VPR) denotes the task of ascertaining whether or not the current place has already been visited [2, 3]. In this manner, the system can impose additional constraints for map building and trajectory optimization, subsequently eliminating the incremental drift [4–6]. As regards the robots that require autonomous operation for an extended period, the appearance of the surrounding environment may change drastically over time. Severe changes in appearance, as well as the adverse impacts caused by the occlusions, dynamic scenes, and perceptual aliasing [4] (see Figure 1), make VPR still considered a daunting task.

VPR tends to be essentially viewed as the data association task and is known as the appearance-based method when the data type is an image. In most cases, the appearance-based methods are conducted within the framework of image retrieval. Specifically, the comparison is drawn between the query image of the current place and images of previously visited places stored in the historical database, and their similarity acts as a key factor in determining whether they are the same place. Therefore, it is very important for appearance-based methods to generate appropriate and accurate image descriptions [5].

Feature descriptor provides the compact and efficient representation of distinctive characteristics in an image [7], and ideally, the descriptor would yield good invariance under image transformation. The correspondences can be established between the query image and database images by


(a)

(b)

Figure 1: The challenge of VPR is that (a) the same places can look totally different while (b) the diverse locations have a similar sensory appearance.

a similarity measurement of the descriptors; thus, it is helpful to distinguish one place (image) from another. The VPR approaches tend to be classified into three major divisions, namely, local descriptor-based, global descriptor-based, and local region descriptor-based. Since global descriptor methods include the aggregation of local descriptors and local region descriptors refer to describing the image patches with a global descriptor, we focus on global descriptor methods to limit the scope of this study.

Global descriptors simply describe the image with only one compact feature vector. On the one hand, global descriptors can be directly constructed by extracting the global features of the image, for instance, histogram of oriented gradients (HOG) [8] and Gist [9]. Conversely, they could equally be aggregated from multiple local descriptors. A typical example is bag of visual words (BoVW) [10], which clusters local features (e.g., SIFT [11], ORB [12]) and then generates the vector of frequency histogram for global description. In recent times, many deep learning-based methods are introduced into VPR to extract global features [13–18], both off-the-shelf and customized convolutional neural network (CNN) models. While the global descriptor methods in VPR have attracted extensive attention, few papers have contrasted them from a comprehensive perspective. It remains complicated for researchers, especially inexperienced novices to thoroughly comprehend this research topic. Therefore, one of the main focuses of our work is to figure out the mechanism that contributes to improved performance. In particular, considering the great potential of deep learning techniques, we also integrated recent advances into our VPR evaluation framework.

The drastically changing environment will allow for false place recognition, which will disrupt the global consistency of the map and lead to a wrong localization [19, 20]. Therefore, high hopes have been pinned for VPR methods to achieve higher or even 100% recognition precision [21]. Although false-positive results could be filtered out by temporal [4, 22–24] or geometric consistency check [25–27], it is more important to develop novel descriptors that have better performance being more robust and recognizable. Appearance and viewpoint invariance is very important for descriptors but it is not the only property to be considered in the VPR task. The descriptors should be computationally efficient to attain the requirements of real-time running. Therefore, another key contribution of this research is to assess the global descriptors from a practical application perspective.

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

(i) We present a comprehensive assessment of the global descriptor methods commonly used in VPR tasks, thereby figuring out the motivational factors for improved performance. Our work covers 15 global descriptor methods, 6 benchmark datasets, and 5 metrics.

(ii) We give practical advice for the design of better global descriptors for VPR tasks, based upon quantitative and qualitative analysis. The specific analysis of hierarchal features and backbone networks was implemented for this purpose.

(iii) We provide valuable information regarding VPR performance from the point of view of practical applications. This investigation offers the trade-offs between matching performance and computational efficiency.

The remainder of this paper is structured as follows. In Section 2, we provide a review of the typical global descriptor methods used in VPR tasks. Then, Section 3 describes the implementation details of the evaluation experiments. The experimental results, comparison, and analysis are carried out in Section 4. Finally, Section 5 gives a conclusion to this study.

2. Literature Review

The similarity between two global descriptors is readily measured by cosine similarity or Euclidean distance, thus it is easy to implement and maintain. Global descriptor methods describe the image as a whole and extract the handcrafted or learning-based advanced features through specified approaches [7]. In this case, the existing methods of global descriptors can be divided into two: (1) handcrafted-based global descriptor methods; (2) deep learning-based global descriptor methods. To facilitate the understanding of deep learning-based global descriptor methods, we also give a brief introduction to recent advances in deep learning techniques, especially CNN models.

2.1. Handcrafted-Based Global Descriptor Methods. Such methods focus on the information present in the image itself, such as the changes in pixel intensity. Histogram of Oriented Gradients (HOG) [8] is one of the most commonly used handcrafted global descriptors. It can extract the structure information of the images by calculating the gradients and orientations of each pixel. McManus et al. [28] extracted HOG features from image patches containing unique visual elements, improving robustness to extreme appearance changes. An impressive work is CoHOG [29], which uses HOG to represent salient image regions for convolutional matching. Another prominent invariant global descriptor is Gist [9]. Murillo and Kosecka [30] presented a Gist-based panorama matching approach for recognizing the revisited places, promoting the application of Gist in VPR tasks. Shortly thereafter, Singh and Kosecka [31] conducted extensive experiments in a 13-mile urban area, demonstrating that the Gabor–Gist descriptor is competent for large-scale scenes.

Instead of using one global feature for the entire images, another class of global descriptors is based on the aggregated local descriptors. BoVW was initially used for image retrieval but has already been demonstrated to be an effective model for place recognition [3, 4, 22–24, 32, 33]. Importantly, BoVW also confers scalability of the map, which is very important for place recognition in large-scale environments and long-term autonomy. Benefiting from the tree-structure vocabulary [10] and inverted index, the previously visited places (images) can be stored and retrieved in a highly efficient manner. An early application case of BoVW in VPR was performed by Schindler et al. [34], who stored more than 100 million SIFT features using a vocabulary tree and successfully achieved place recognition on 20 km of urban roads. Gálvez-López and Tardos [4] proposed an enhanced BoVW method and open-sourced their C++ library named DBoW for converting images into a bag-of-word representation and constructing the visual dictionary [35]. Owing to its convenience and efficiency, the improved version DBoW2/3 has been utilized in many excellent SLAM systems [23, 24]. Similar approaches were successfully presented with Fisher vectors (FV) [36] or vector of locally aggregated descriptors (VLAD) [37].

2.2. Deep Learning-Based Global Descriptor Methods. Given the rapid development of deep learning, the novel properties of learning-based techniques have inspired researchers to leverage them to remedy the shortcoming of handcrafted descriptors. After the seminal work of Chen et al. [38], research has increasingly focused on learning-based descriptors which are mainly built on CNN features [13, 39–41]. Sünderhauf et al. [13] used pretrained AlexNet [42] as a descriptor extractor and concluded that mid-level features have better robustness to appearance variations. Hou et al. [39] have reported a similar finding, where a CNN model was pretrained on the scene-centric database called Places365 [43]. Zhang et al. [40] constructed a graph-based VPR method through the integration of visual features extracted from VGG16 [44] and temporal information from image sequences. Wang et al. [41] used pretrained ResNet [45] as the image descriptor to realize place recognition in a dynamic environment. Furthermore, researchers have focused on developing specialized neural network architectures for VPR tasks. Reconstructing the traditional approaches via deep learning-based techniques motivated the emergence of novel VPR methods, such as CALC [46], NetVLAD [14], and E2BoWs [47]. These methods combine the complementary strengths of handcrafted and learning-based descriptors to arrive at a remarkable performance. Additionally, autoencoder and its modified versions have also been introduced into the VPR domain [16]. The advantage of these unsupervised learning methods is that they require less manual data preparation.

In theory, the performance of CNN-based descriptors, especially supervised learning ones, depends on high-quality, large-scale training datasets. Driven by the booming VPR research communities, more relevant datasets in this field have been constructed and the development of global descriptors has been further promoted. A typical example is the specific place dataset (SPED), developed by Merrill and Huang [46] in 2017. This study highlighted the differences between a network trained on SPED versus ImageNet and indicated that the CNN-based descriptors trained on tailored datasets tend to have enhanced performance gain. Additionally, it was also found that adaptability and generalization can be improved by fine-tuning the targeted dataset [47–49].

2.3. Popular Deep Learning Models. Deep learning is particularly adept at extracting high-level abstract features from raw images. CNN is one of the most popular deep learning networks. The CNN model was first proposed by LeCun et al. [50] for recognizing handwritten digits. Extending work AlexNet [42] has surged a wave in the computer vision community. Several important backbone networks were proposed in subsequent years, such as VGG [44], GoogLeNet [51], ResNet [45], Xception [52], and DenseNet [53]. With the reorganization of processing units and the emergence of new modules, a wide variety of CNN architectures are constantly presented in response to different applications. Over, the trend has been towards deeper and more complicated architectures to derive better performance. However, going deeper means increasing sequential processing and latency. For robots with computationally limited platforms, it is critical to develop lightweight and low-latency models.

To expand the applications in mobile and embedded devices, some attempts were made to reduce the parameter quantity of the CNN model. SqueezeNet [54] employs 1 × 1 convolutions to compress the model to less than 0.5 MB. MobileNet [55, 56] builds on depthwise separable convolution and efficiently balances latency and accuracy. The core component of ShuffleNet [57] is pointwise group convolution and channel shuffle, which significantly reduce computation costs. Benefiting from neural architecture search, EfficientNet [58] offers adjustable depth-width-resolutiontrade-offs and leads to better accuracy and efficiency.

3. Implementation Details

Based on the significant work reviewed above, this paper chooses 15 typical global feature descriptors for evaluation (see Table 1), while the motivation for the choices and implementation details are briefly presented. Then, the datasets and evaluation metrics used in our experiments are presented.

3.1. Global Descriptor Methods

3.1.1. Handcrafted Feature-Based Descriptors

(1) HOG. HOG is a simple but effective descriptor and has good invariance to illumination changes. We use a window size of 16 × 32, a block size of 16 × 16, and a cell size of 8 × 8. The number of bins is set equal to 9. Resultantly, an input image with 160 × 120 size is represented as a 16416-dimensional HOG descriptor.

(2) Gist. The essential idea behind Gist is that an image can be described by the responses of Gabor filters at diverse scales and orientations. In our work, input images are convolved with 20 Gabor filters at 3 scales (8, 8, and 4 orientations). Each feature map is divided into 16 regions by a 4 × 4 grid, thus the output descriptor size is 960 dimensions.

(3) DBoW3-ORB. As mentioned previously, DBow3 is one of the most widely used methods in VPR. We utilize the vocabulary file *ORBvoc* provided by ORB-SLAM2, which is trained on a large-scale dataset and has good adaptability and generalization.

3.1.2. Off-the-ShelfCNN-Based Descriptors. In this paper, we select six off-the-shelfCNN-based descriptor methods, including *AlexNet*, *VGG16*, *ResNet50*, *MobileNet v3*, *ShuffleNet v2*, and *EfficientNet B0*. They have been introduced in detail in their paper so we will not repeat their description. Here, we provide the implementation details and motivation.

To a certain extent, the performance of the CNN models depends on both the scale and richness of training datasets. For all experiments, six off-the-shelf CNN models pretrained on ImageNet are used to generate global feature descriptors. For AlexNet, VGG16, and ResNet50, since the performance of hierarchal features has attracted considerable attention, the features extracted from different layers are used as global descriptor vectors, respectively. Analysis results are discussed in Section 4. To more comprehensively survey the global descriptors, we also selected three very recent advances to generate holistic features for describing images, including MobileNet v3, ShuffleNet v2, and EfficientNet B0. The choice of them is motivated by their lightweight architectures and practicality: larger models are less suitable for resource-constrained robots or mobile devices.

3.1.3. Customized CNN-Based Descriptors

(1) CALC. CALC is a lightweight convolutional autoencoder model proposed by Merrill et al. to address the shortcoming of HOG which is not robust to viewpoint changes. Our implementation of CALC retains the last three fully connected layers but the original backbone network was substituted with a pretrained ResNet18. Noting that the last pooling layer along with the fully connected layer of ResNet18 is eliminated. For training, we fine-tuned the modified model on the Places365 dataset to better focus on the VPR task and followed the training setup of CALC's open-sourced work, so the corresponding output descriptors are also 3648-dimensional.

(2) NetVLAD. The reformulating of VLAD through CNN-based techniques contributed to this significant outcome. It provides a differentiable pooling mechanism with trainable parameters. The proposed NetVLAD layer serves as a plug-and-play module and presents a rich yet compact image representation. We implemented the NetVLAD in Pytorch and also used Pittsburgh dataset for training.

(3) MobileNetVLAD. MobileNetVLAD was initially proposed for 6-DoF pose estimation. We integrated it into our evaluation work as a reference for comparison. Interestingly, MobileNetVLAD is trained in a self-supervised manner. Under the supervision of a well-trained NetVLAD model (the teacher), this network (the student) uses knowledge distillation to transfer the knowledge, thus being able to extract NetVLAD descriptors with a more lightweight network.

(4) DBoW3-SuperPoint. SuperPoint is a CNN-based interest point detector and descriptor. With respect to handcrafted local descriptors, it achieves superior homography estimation results in the premise of high real-time performance. To fit the VPR task, we further incorporated this local descriptor into the BoVW model and led to a novel global descriptor. Our implementation is also carried out on the C++ DBoW3 library.

Table 1: 15 typical global feature descriptors evaluated in our work.

Category	Method	Use in VPR
Handcrafted-feature-based	HOG [8]	[28, 29, 46]
	Gist [9]	[30, 31]
	DBoW3-ORB [4]	[4, 23, 24, 59]
Off-the-shelf CNN-based	AlexNet [42]	[2, 13, 39, 49, 60, 61]
	VGG [44]	[14, 17, 40]
	ResNet [45]	[45, 62]
	ShuffleNet [57, 63]	[64]
	MobileNet [55, 56]	[65–68]
	EfficientNet [58]	[69, 70]
Customized CNN-based	CALC [46]	[46]
	NetVLAD [14]	[14, 70–75]
	MobileNetVLAD [76]	[76, 77]
	DBoW3-SuperPoint [78]	[79]
	Autoencoder [80]	[16, 46, 81]
	Variational autoencoder [82]	[83]

(5) Autoencoder (AE). This type of neural network can learn efficient image representations in an unsupervised manner, and thus is well suited for VPR tasks that lack high-quality labeled data. It is an appropriate case for a customized CNN-based descriptor, based on its outstanding performance as shown by Gao and Zhang [16] and Park et al. [81]. In our work, the encoder of our AE is similar to the implementation of [46] and the decoder is composed of deconvolution and unpooling layers. The network was trained on the Places365 dataset, and the output of the well-trained encoder was considered as the global descriptor.

(6) Variational Autoencoder (VAE). VAE was proposed by Kingma et al. where explicit regularization is introduced to ensure the good properties of its latent space. An interesting VPR method built upon a VAE was proposed by Merrill and Huang [83]. We designed a network structure similar to the autoencoder described above and also trained it on the Places365 dataset. The original idea that we used VAE is for dimensionality reduction rather than image generation, thus the decoder was removed during inference.

3.2. Evaluation Metrics. A descriptor that has state-of-the-art place matching performance, but an unacceptable longer place retrieval time, will fail to meet the rigid demand for real-time localization systems. For practical reasons, we integrated multiple evaluation metrics in this work to comprehensively evaluate these global descriptors in terms of matching performance and computational efficiency. Details of each metric are presented as follows.

3.2.1. Matching Performance. For VPR tasks, true positives (TP) denote the correct image/place matching results, false positives (FP) refer to the situations where the actually incorrectly matched images are judged to be the same place, while false negatives (FN) represent the situations where the true matching cases are not screened out. For most VPR datasets, it should be pointed out that every query image has a ground-truth match in the database images, thus there are usually no true negatives [13]. Precision and recall are computed as follows:

$$\begin{aligned}\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}},\\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}.\end{aligned} \tag{1}$$

We present the evaluation metrics used in this paper as follows:

(1) AUC. Ideally, a VPR method should achieve 100% precision and 100% recall, and indeed, a negative correlation was found between precision and recall, that is, increasing precision frequently leads to a reduction in the recall. Therefore, many works [3, 4, 32, 46, 83–85] have focused on the area-under-the-precision-recall (AUC) curves for a comprehensive evaluation and we also introduced it into our assessment work. The precision-recall (PR) curves reflect the changing trend of precision with rising recall, thus it can help in making an informed decision when facing the precision/recall dilemmas. AUC summarizes precision and recall in a single visualization, only a single correct match result is taken into account when calculating AUC.

(2) Recall at 100% Precision. Furthermore, another metric applied in this study is recall at 100% precision (R_{P100}, for short). It is also a commonly used metric to evaluate the VPR methods [16, 40, 46, 68]. The motivation of using this metric is the favoritism of precision in the VPR system. In general, 100% precision is very important for the VPR system because false positives are extremely disruptive and unacceptable.

(3) Recall@1. The requirement of Recall@1 is that the best-matched database image for a query image must be a true positive. Although Recall@N has been widely used in image retrieval tasks, that is, the correct retrieval only needs to be among the Top-N candidates, the allowable range for VPR or loop closure detection is more stringent from the viewpoint of practical application. In addition, the motivation behind

using Recall@1 is that this metric actually reflects the percentage of correctly matched query images.

3.2.2. Computational Efficiency. Despite limited computing resources, the place recognition module in a mobile robot must perform in real-time to maintain good localization accuracy. In this case, the computational efficiency of the descriptors is another major consideration, including feature encoding time and descriptor matching time.

(1) Feature Encoding Time. The feature encoding process of most global descriptors is relatively time-efficient because no keypoints detection process is involved. In such a case, feature encoding time denotes the time spent in extracting the global features. BoVW-based descriptors (i.e., DBoW3-ORB and DBoW3-SuperPoint in this article) are exceptions, their time-consuming process includes the time spent in detecting keypoint, describing, and mapping into bag-of-words space. For statistical validation, the encoding time corresponds to the average of over 200 runs.

(2) Descriptor Matching Time. The total time consumption of descriptor matching is proportional to the scale of the map (database images). For a fair comparison, descriptor matching time here refers to the time required to match two global descriptors, which is also a statistical mean. The manner of similarity measurement also exerts influence in matching time while cosine similarity was used for all global descriptors evaluated in this study.

3.3. Evaluation Datasets. We integrated 6 benchmark datasets to evaluate the performance of the above 15 global descriptors. These datasets feature images from diverse scenarios, including indoors, urban roads, suburbs, and natural scenery. Each dataset has two separate folders, one for organizing query images and one for database images. The ground-truth information is provided by the filename, that is, the images with the same filename indicate the same place. For each query image, there is a database image that was taken at the same place but has undergone changes in appearance and/or viewpoint. Table 2 provides a summary of the datasets and their major challenges; Figure 2 gives the sample images.

Due to the differences in shooting frequency or traveling speed, the image sequences in some datasets may be consecutive and the adjacent images have overlapping visual content. Therefore, the setup of ground-truth tolerance is commonly accepted in VPR tasks but generally stricter than that of computer vision tasks. The ground-truth tolerance used in our work is presented in Table 2.

Here, we provide a brief introduction to these datasets to facilitate analyzing the performance of each descriptor method. The download links of all datasets are available in the footnote.

(1) Nordland Dataset (https://nrkbeta.no/2013/01/15/). The Nordland dataset is extracted from video footage recording four 729 km journeys on the same route. It collects both natural and urban landscapes in four seasons. Severe cross-seasonal changes lead to strong appearance changes but no viewpoint changes are involved due to the fixed route. We choose Spring versus Winter image sequences for experimental analysis.

(2) SPEDTest Dataset (https://goo.gl/OXeL2X). The SPEDTest dataset is a subset picked from the original SPED dataset. It is captured with the outdoor cameras that are used to collect the long-term scenarios changes and hence contains extreme appearance variations in changeable seasons and illumination conditions. Due to the limitation of the camera's fixed view angle, this dataset exhibits no viewpoint changes.

(3) Campus Loop Dataset (https://github.com/rpng/calc/tree/master/TrainAndTest/test_data). This dataset consists of two image sequences, captured in indoor and outdoor environments. For the purposes of covering multiple challenges, the first image sequence was taken on a cloudy snowy day with buildings and roads covered in snow, whereas the second image sequence was taken on a sunny day.

(4) Gardens Point Dataset (https://zenodo.org/record/4561862). This dataset was collected at the campus scenes with a handheld mobile phone. In this study, we used two daytime image sequences recorded under different illumination conditions and left/right walking paths. These factors render this dataset containing strong lateral viewpoint changes and modest appearance changes.

(5) Cross-Seasons Dataset (https://www.visuallocalization.net/datasets/). The cross-seasons dataset used in our work contains two image sequences taken in different illumination, seasons, or weather conditions. The interference from the dynamic object and perceptual aliasing further made it challenging to perform place recognition.

(6) Alderley Day/Night Dataset (https://www.dropbox.com/s/ejmnz9vfp4n7o7s/alderley.zip?dl=0). This dataset was created by Milford et al. where two image sequences were captured on a bright sunny day and an extremely heavy rainy night, respectively. Furthermore, night storms cause extreme appearance changes and blurring, making it complicated even for humans to achieve successful place recognition.

4. Results and Discussion

In the following, we will present the experimental evaluation of the 15 global descriptor methods and discuss the driving forces behind these results. The analysis was generally carried out from two aspects, including matching performance and computational efficiency, to facilitate more consideration of the practicability.

Table 2: Information on benchmark datasets used in our work.

Dataset	Environment	Ground-truth tolerance	Major challenges
Nordland [86]	Natural scenery	±3 frames	Extreme changes due to shifts in season, blurry appearance
SPEDTest [48]	Urban and natural scenery	Frame-to-frame	Strong appearance variations in changeable seasons, illumination conditions
Campus loop [46]	Campus outdoor/indoor	±1 frames	Extreme brightness, viewpoint and seasonal variations, many dynamic objects
Gardens Point [87]	Campus outdoor/indoor	±2 frames	Strong viewpoint variations, dynamic scenes
Cross-seasons [88]	Urban road	±3 frames	Weather change, dynamic scenes
Alderley [89]	Urban roads	±15 frames	Extreme appearance changes due to day-night and weather variations

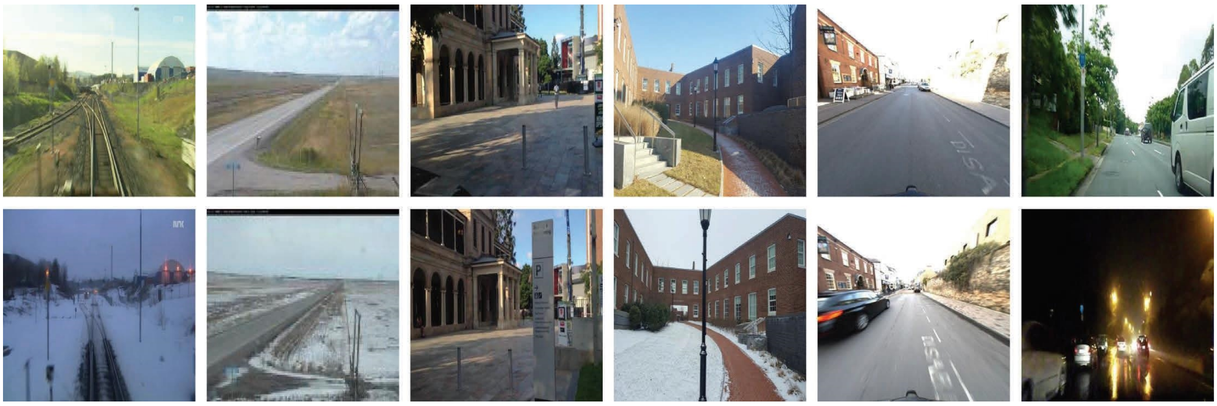


Figure 2: Sample images from 6 benchmark datasets.

4.1. Matching Performance Analysis. The PR curves for all 15 global descriptors are presented in Figure 3, and the values of AUC and R_{P100} for each descriptor are presented in Tables 3–5.

For three handcrafted feature-based descriptors, Figure 3 shows that their place matching precision either retains at a relatively low level or degrades substantially with the increasing recall when encountering drastic visual changes. HOG can achieve good performance on SPEDTest and Nordland datasets that do not involve any viewpoint changes. However, its performance is significantly degraded when meeting the dual challenge of viewpoint and appearance variations. DBoW3-ORB and Gist can only yield acceptable matching performance on the less-challenging datasets, such as Gardens Point and cross-seasons. Although aggregated as a global descriptor through DBoW3, as a local descriptor, ORB is not robust to appearance changes, which leads to its bottom-ranked AUC on the Norland and Alderley datasets.

Despite the choice of datasets having a moderate influence on assessment results, we observe that the performance of CNN-based descriptors has an overall better performance over non-CNN-based descriptors, particularly for datasets that have extreme variations in viewpoint and appearance. For instance, while all descriptors suffer from multiple challenges, the PR curve of CNN-based descriptors decreases relatively gently, whereas that of handcrafted descriptors declines rapidly. The quantitative comparisons are presented in Table 3–5. The results of the AUC indicator show that CNN-based descriptors almost always outperform handcrafted feature-based descriptors, a similar picture is seen on the R_{P100} metric. One exception is DBoW3-SuperPoint, which cannot reach the same level as other CNN-based methods. Like DBoW3-ORB, DBoW3-SuperPoint only performs well on the Gardens Point dataset with slight appearance changes. This further shows that the aggregation of local descriptors cannot hold for the strong appearance changes to the same level as global descriptors that represent the image as a whole. In most cases, we observed that CALC achieves better PR performance than HOG, demonstrating that better matching results can be delivered by integrating the advantages of CNN and traditional methods. Despite their lightweight networks, the matching performance of ShuffleNet, MobileNet and EfficientNet is marginally better than that of the other three off-the-shelf CNN-based descriptors.

In addition, the customized CNN-based descriptors generally have better robustness compared to off-the-shelf CNN-based ones, thereby maintaining good adaptability and generalization in common challenges. In terms of PR curves, the decay of the former's precision is relatively slow. This means that customized CNN-based descriptors generally achieve higher precision under the same recall. An impressive method is NetVLAD that nearly in most cases attains state-of-the-art performance. MobileNetVLAD can achieve (and sometimes even surpass) NetVLAD-level place matching performance, demonstrating the potential of lightweight CNN in VPR tasks.

4.2. Computational Efficiency Analysis. We now discuss the computational efficiency of the 15 global descriptor methods. In this experiment, we use the Gardens Point dataset with an image resolution of 960 × 540, and the values of feature encoding time and descriptor matching time are listed in Table 6. Note that a unified CPU-only platform was used for both conventional and CNN-based descriptors, whereas CNN-based ones generally require more computational resources. This experiment was performed on an Ubuntu 18.04 LTS operating system running on an Intel Xeon E5-2678 V3 CPU @ 2.5 GHz and RTX 2080Ti GPU.

It can be seen that the matching time and dimension are positively associated when using the same similarity measure. For instance, the HOG descriptor achieves the fastest feature encoding of only 1.46 ms, but this descriptor matching time is significantly higher because of its larger dimension. Similarly, 6 off-the-shelf descriptors are of dimension 1000, therefore matching time for them is nearly identical.

We now turn to the discussion of feature encoding time. As illustrated in Table 6, CNN-based global descriptors are computationally intensive, thereby

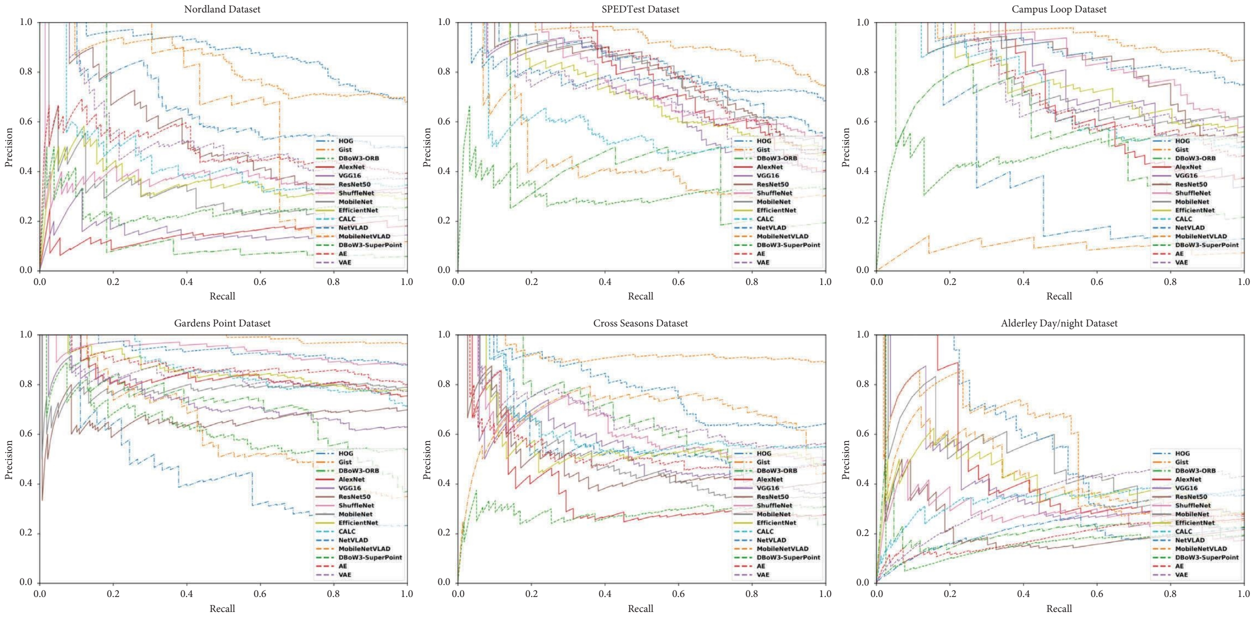


Figure 3: The precision-recall curves for all 15 global descriptors generated on the 6 benchmark datasets.

Table 3: The values of AUC and R_{P100} for 3 handcrafted feature-based descriptors.

Dataset	HOG		Gist		DBoW3-ORB	
	AUC	R_{P100}	AUC	R_{P100}	AUC	R_{P100}
Norland	65.71	10.10	61.68	30.43	24.47	18.18
SEPDTest	81.58	11.22	44.18	6.90	44.20	14.29
Campus loop	38.78	18.18	9.46	0	55.41	5.26
Gardens Point	46.06	11.11	62.54	12.86	71.53	8.11
Cross-seasons	55.22	9.78	67.38	0	63.47	17.78
Alderley	18.64	0	44.10	2.44	21.15	0

Table 4: The values of AUC and R_{P100} for 6 off-the-shelf CNN-based descriptors.

Dataset	AlexNet		VGG16		ResNet50		ShuffleNet		MobileNet		EfficientNet	
	AUC	R_{P100}	AUC	R_{P100}	AUC	R_{P100}	AUC	R_{P100}	AUC	R_{P100}	AUC	R_{P100}
Norland	14.85	0	15.21	0	53.25	8.06	36.87	1.49	30.78	2.50	35.00	0
SEPDTest	77.18	36.71	71.42	7.79	78.48	10.00	75.89	22.86	79.55	16.49	70.92	14.13
Campus loop	70.49	35.13	76.68	39.40	83.16	14.00	86.61	31.00	76.71	33.33	76.67	21.43
Gardens Point	83.38	11.26	73.09	2.38	66.26	0.72	93.52	4.55	78.02	0.63	85.72	7.69
Cross-seasons	37.46	3.92	53.25	5.80	44.22	2.63	61.91	5.38	49.92	6.06	56.38	7.69
Alderley	46.71	16.67	40.18	3.85	22.50	2.33	27.26	0	47.37	3.23	40.29	0

commonly spending more time in feature encoding with a few exceptions. The lightweight CNN-based descriptors are able to reach the same encoding efficiency as non-CNN-based descriptors on a CPU-only platform. Such lightweight networks are important for meeting the need of the real time. A typical example is MobileNetVLAD (a lightweight version of NetVLAD), which achieves a significant speed boost. The encoding time speed roughly doubled after using MobileNet instead of the original VGG16 backbone.

We also report the encoding times of 12 CNN-based descriptors accelerated with GPU in Table 7. As can be seen, all the CNN-based global descriptors are able to achieve real-time performance under the GPU acceleration. For three lightweight networks, including ShuffleNet, MobileNet, and EfficientNet, they are outstanding regarding the number of parameters (#Params) and floating-point operations per second (FLOPs), but their inference speeds are less prominent than in the CPU platform. This is because of their specific design for mobile and embedded devices. In addition, to these three descriptors, it is apparent that the major factor impacting the descriptor encoding time is the required floating-point operations.

Table 5: The values of AUC and R_{P100} for 6 customized CNN-based descriptors.

Dataset	CALC		NetVLAD		MobileNetVLAD		DBoW3-SuperPoint		AE		VAE	
	AUC	R_{P100}	AUC	R_{P100}	AUC	R_{P100}	AUC	R_{P100}	AUC	R_{P100}	AUC	R_{P100}
Norland	44.27	7.25	88.02	12.69	82.64	9.56	24.84	0	49.11	0	52.81	8.11
SEPDTest	56.65	7.29	77.53	3.70	92.55	21.09	32.24	0	80.77	30.68	68.48	8.14
Campus loop	71.02	12.24	88.24	15.98	94.11	16.67	47.65	0	72.47	26.67	74.06	20.37
Gardens Point	86.37	25.87	93.79	16.03	98.83	50.78	62.63	1.85	87.71	8.75	83.69	8.44
Cross-seasons	61.95	8.65	78.59	10.66	92.21	12.94	28.63	0	52.29	3.30	67.20	5.66
Alderley	33.33	0	58.47	21.13	52.76	1.96	17.59	2.56	18.82	0	34.47	0

Table 6: Times (in milliseconds) for feature encoding t_e (ms) as well as descriptor matching t_m (ms), using a CPU-based platform.

Methods	Input size	t_e	t_m	Dimensions
HOG	160 × 120	1.46	0.0074	16416 × 1
Gist	960 × 540	40.25	0.0024	960 × 1
DBoW3-ORB		18.13	0.0071	32 × 500
AlexNet	224 × 224	12.48	0.0027	1000 × 1
VGG16		54.42	0.0029	
ResNet50		48.27	0.0029	
ShuffleNet		16.84	0.0026	
MobileNet		18.25	0.0028	
EfficientNet		51.71	0.0029	
CALC	160 × 120	24.71	0.0032	3648 × 1
NetVLAD	224 × 224	71.35	0.0136	32768 × 1
MobileNetVLAD		34.53	0.0043	7680 × 1
DBoW3-SuperPoint	960 × 540	32.81	0.0001	284 × 1
AE	224 × 224	52.45	0.0063	12544 × 1
VAE		61.40	0.0061	12544 × 1

Table 7: Times (in milliseconds) for feature encoding t_e (ms) of 12 CNN-based descriptors, using a GPU-based platform.

Methods	#Params (M)	FLOPs (G)	t_e
AlexNet	61.10	0.72	2.13
VGG16	138.36	15.48	5.07
ResNet50	25.56	4.12	4.94
ShuffleNet	2.28	0.15	9.07
MobileNet	2.54	0.06	8.03
EfficientNet	5.28	0.40	12.85
CALC	21.30	0.72	5.71
NetVLAD	14.74	15.36	15.48
MobileNetVLAD	0.95	0.06	22.82
DBoW3-SuperPoint	—	—	17.38
AE	1.14	3.21	3.31
VAE	1.64	3.25	4.26

Table 8: Comparison of different feature levels of CNN-based descriptors (with a GPU platform).

Descriptors		Dimensions	Time (ms)		Metric	
			t_e	t_m	AUC	$\mathrm{R}_{\mathrm{P100}}$
AlexNet	pool1	(64, 27, 27)	**0.55**	0.0086	0.69	0.07
	pool2	(192, 13, 13)	0.77	0.0068	0.83	0.05
	conv3	(384, 13, 13)	1.02	0.0116	0.90	0.07
	conv4	(256, 13, 13)	1.18	0.0072	0.89	0.21
	pool5	(256, 6, 6)	1.25	0.0048	**0.91**	**0.31**
	fc6	(1, 4096)	1.76	0.0030	0.89	0.22
	fc7	(1, 4096)	1.76	0.0027	0.89	0.06
	ALL	(1, 1000)	2.13	**0.0024**	0.83	0.11
VGG16	pool1	(64, 112, 112)	**1.03**	0.1091	0.61	0.04
	pool2	(128, 56, 56)	1.65	0.0567	0.60	0.16
	pool3	(256, 28, 28)	2.47	0.0330	0.81	0.18
	pool4	(512, 14, 14)	3.90	0.0174	0.81	0.11
	pool5	(512, 7, 7)	4.66	0.0060	**0.89**	0.11
	fc6	(1, 4096)	4.97	0.0040	0.88	**0.22**
	fc7	(1, 4096)	4.96	0.0040	0.82	0.08
	ALL	(1, 1000)	5.07	**0.0029**	0.73	0.02

Table 8: Continued.

Descriptors		Dimensions	Time (ms)		Metric	
			t_e	t_m	AUC	R_{P100}
ResNet18	pool1	(64, 56, 56)	**0.69**	0.0580	0.58	0.00
	stage2	(64, 56, 56)	1.83	0.0362	0.61	0.15
	stage3	(128, 28, 28)	2.80	0.0230	0.73	**0.16**
	stage4	(256, 14, 14)	3.64	0.0085	0.81	0.15
	stage5	(512, 7, 7)	4.79	0.0060	**0.90**	0.12
	ALL	(1, 1000)	4.94	**0.0029**	0.74	0.02
ResNet50	pool1	(64, 56, 56)	**0.76**	0.4271	0.60	0.00
	stage2	(256, 56, 56)	3.22	0.0860	0.63	0.06
	stage3	(512, 28, 28)	6.73	0.0622	0.76	0.11
	stage4	(1024, 14, 14)	11.49	0.0378	0.81	**0.14**
	stage5	(2048, 7, 7)	13.24	0.0029	**0.86**	0.11
	ALL	(1, 1000)	12.64	**0.0024**	0.66	0.01

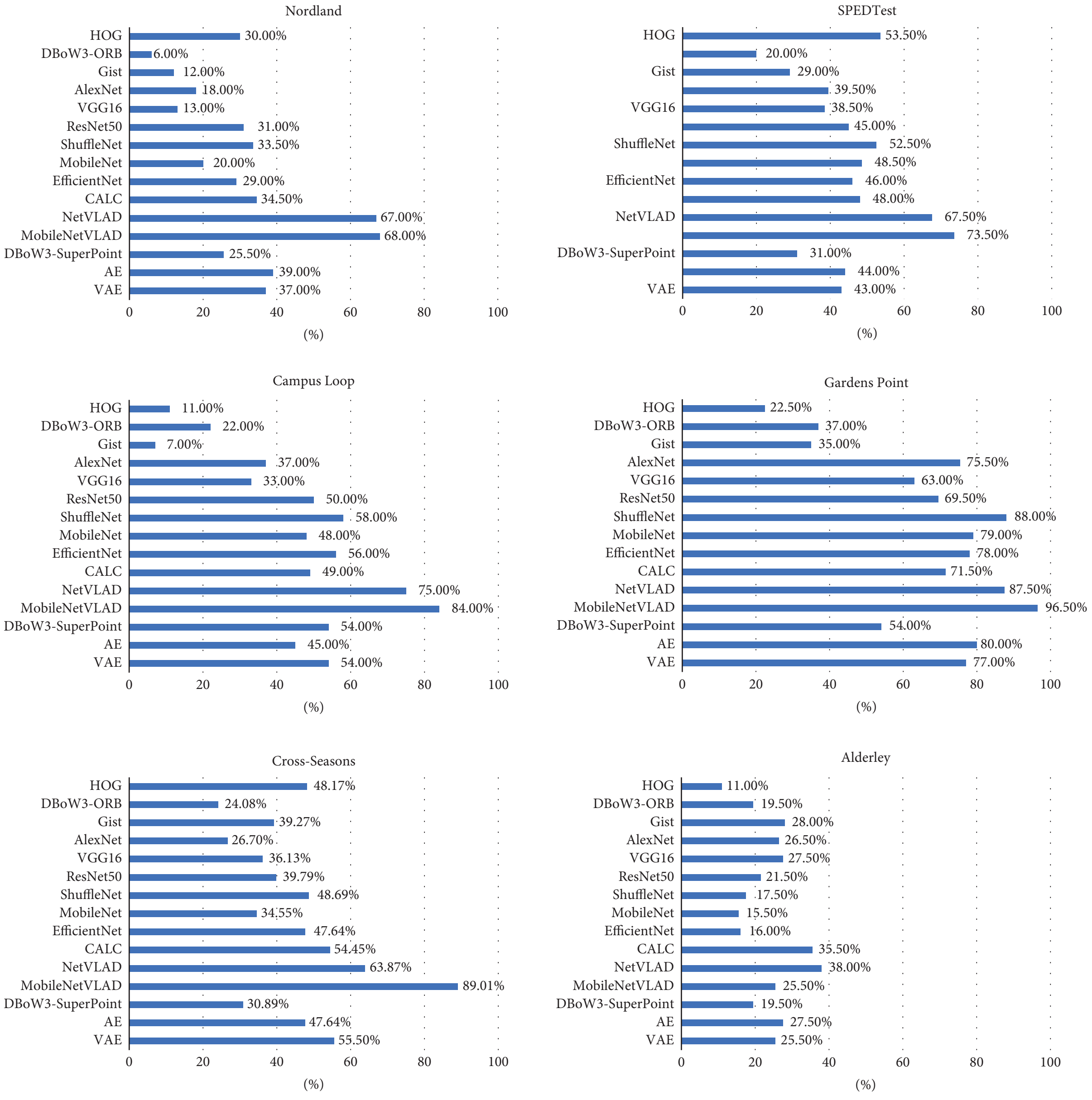


Figure 4: The values of Recall@1 for 15 global descriptors.

4.3. The Inquiry in Constructing Better CNN Descriptors. Based on the abovementioned results and discussion, we attempted to figure out which aspects a descriptor can benefit from. As discussed earlier, the contradiction lies in that VPR methods generally run on a computationally limited platform but are asked to meet real-time requirements. Therefore, a better global descriptor should be sufficiently lightweight and efficient to compute. Apparently, compressing the size of the CNN-based descriptors does not lead to the loss in matching performance. Comparing MobileNetVLAD and NetVLAD, it will be seen that the former can achieve a better comprehensive performance, even under complex challenges.

Sünderhauf et al. [13] reported that the mid-level layer of AlexNet can yield outstanding performance even under strong appearance changes. Being broader in scope than this work, our work utilizes more models to verify the correlation between matching performance and feature levels. The results are shown in Table 8, and the optimal value for each indicator is indicated in bold in the table. The results of this paper agree with the conclusion drawn in [13], and a similar phenomenon can also be observed in VGG16 and ResNet18/50. Features from shallower layers, particularly the outputs from the first pooling layer cannot achieve matching performance to the same level as features from other higher layers. It is demonstrated that lower-level features (i.e., edges, lines as well as blobs) exhibit sensitivity to the variations in viewpoint and appearance. Contrary to popular belief in the computer vision community, the performance degradation happens in the features from high-level layers. One possible reason is that high-level features are more semantically meaningful but thus suffer significantly on perceptual aliasing. Given the analysis above, we considered that the depth of a CNN descriptor should be moderate, and a fully connected layer should be avoided. It is also readily observed that extracting features from higher layers will introduce more sequential processing, thus decreasing the number of CNN layers will be beneficial in the computational efficiency as well.

Another finding from the PR curves is that place matching performance for CNN-based descriptors is affected by the relevance of the training dataset. Despite the simple structure, AE and VAE yield promising results because they use scene-centric Places365 datasets. In comparison with the off-the-shelf models which trained on ImageNet, they can learn more relevant features for place representation.

4.4. Practicality Analysis. From the perspective of practicality, the system primarily focuses on the proportion of true positives successfully retrieved under acceptable computational efficiency. Taking closed-loop detection in SLAM as an example, the more correct closed loops detected, the more likely the reliable localization accuracy will be maintained. Therefore, we also evaluated the performance of 15 reported descriptors under the Recall@1 metric, as shown in Figure 4. This metric actually reflects the success rate of the place recognition. Figure 4 demonstrates visually the percentage of correct matching for each descriptor.

Although our results are preliminary due to the limited scope of the survey, we recommend the following:

(1) In the presence of none or slight viewpoint changes (e.g., a robot whose routes barely change), the HOG descriptor is a very suitable candidate because of its computational efficiency and effectiveness. DBoW3 is a cost-effective and convenient alternative for less-challenging scenes.

(2) For more complex and changeable environments, CNN-based global descriptors can retrieve more correct matches, nonetheless at the expense of considerable computing resources. Therefore, most CNN-based descriptors are not suited for a CPU-only platform unless their architectures are lightweight enough to acquire low computations.

(3) For an effective but computing-heavy CNN-based descriptor, compressing it into a smaller network is a worthwhile attempt. We have verified that replacing the backbone network with a lightweight one and knowledge distillation are feasible solutions.

5. Conclusion

This article presents a comprehensive evaluation of global descriptor methods for appearance-based visual place recognition. The experiments were conducted on six benchmark datasets, covering diverse scenarios, and utilized five commonly used metrics to assess the matching performance and computational efficiency of 15 global descriptor methods.

Our analysis revealed that each type of descriptor has its own strengths and weaknesses, and we provided valuable insights regarding practicality. Specifically, CNN descriptors generally exhibit significant matching performance, albeit at a higher computational cost, indicating the potential of lightweight CNN descriptors. On the other hand, descriptor methods based on traditional features also have their own utility, with non-CNN-based descriptors being particularly useful in scenarios with less-challenging conditions, owing to their training-free nature and computational efficiency. In addition, our evaluation extended to identify the motivational factors that contribute to improved performance in VPR. The investigation centered around hierarchical features, backbone network designs, model compression, and the choice of training datasets. Our experiential results suggest that utilizing an overly deep network architecture may not be necessary for achieving optimal performance in VPR, given that mid-level features demonstrate more robust performance. Additionally, the network structure should not be too cumbersome to be deployed on a resource-constrained robot platform. Model compression techniques, such as knowledge distillation, may provide feasible solutions to this issue. It is also critical to emphasize the importance of using relevant datasets to train the CNN model.

We anticipate that this study could assist researchers in gaining a more comprehensive understanding of appearance-based VPR and global descriptor methods, especially for novice learners. As a future direction, we will focus on novel methods used in VPR tasks, such as generative adversarial networks and deep multimodal learning. Consequently, this assessment could be extended to integrate additional descriptors, datasets, and metrics, thereby enhancing our understanding of this field.

Data Availability

The visual place recognition data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Special Project of the National Key Research and Development Program of China under grant no. 2020YFB1313304.

References

[1] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *IEEE Robotics and Automation Magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[2] N. Sünderhauf, S. Shirazi, A. Jacobson et al., "Place recognition with convnet landmarks: viewpoint-robust, condition-robust, training-free," in *Proceedings of Robotics: Science and Systems XII*, pp. 1–10, 2015.

[3] M. Cummins and P. Newman, "Fab-map: probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

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

[5] S. Lowry, N. Sünderhauf, P. Newman et al., "Visual place recognition: a survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[6] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," in *Proceedings of the 2000 ICRA. Millennium Conference IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, vol. 2, pp. 1023–1029, San Francisco, CA, USA, April 2000.

[7] R. Feng, H. Shen, J. Bai, and X. Li, "Advances and Opportunities in Remote Sensing Image Geometric Registration: a systematic review of state-of-the-art approaches and future research directions," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 4, pp. 120–142, 2021.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pp. 886–893, California, CA, USA, June 2005.

[9] A. Oliva and A. Torralba, "Building the gist of a scene: the role of global image features in recognition," *Progress in Brain Research*, vol. 155, pp. 23–36, 2006.

[10] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 2161–2168, New York, NY, USA, June 2006.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: an efficient alternative to sift or surf," in *Proceedings of the 2011 International conference on computer vision*, pp. 2564–2571, Barcelona, Spain, November 2011.

[13] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *Proceedings of the 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 4297–4304, Hamburg, Germany, September 2015.

[14] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5297–5307, Las Vegas, NV, USA, June 2016.

[15] D. Bai, C. Wang, B. Zhang, X. Yi, and X. Yang, "Sequence searching with cnn features for robust and fast visual place recognition," *Computers & Graphics*, vol. 70, pp. 270–280, 2018.

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

[17] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9–16, Vancouver, Canada, September 2017.

[18] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, and K. McDonald-Maier, "Levelling the playing field: a comprehensive comparison of visual place recognition approaches under changing conditions," 2019, http://www.arxiv.org/abs/0704.4001.

[19] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 55–81, 2015.

[20] M. Magnusson, H. Andreasson, A. Nuchter, and A. J. Lilienthal, "Appearance-based loop detection from 3d laser data using the normal distributions transform," in *Proceedings of the 2009 IEEE International Conference on Robotics and Automation*, pp. 23–28, Kobe, Japan, May 2009.

[21] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: a survey from deep learning perspective," *Pattern Recognition*, vol. 113, Article ID 107760, 2021.

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

[23] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[24] R. Mur-Artal and J. D. Tardós, "Orb-slam2: an open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[25] K. A. Tsintotas, L. Bampis, and A. Gasteratos, "Assigning visual words to places for loop closure detection," in *Proceedings of the 2018 IEEE International Conference on Robotics*

and Automation (ICRA), pp. 5979–5985, Brisbane, Australia, May 2018.

[26] E. Johns and G.-Z. Yang, "Feature co-occurrence maps: appearance-based localisation throughout the day," in *Proceedings of the 2013 IEEE International Conference on Robotics and Automation*, pp. 3212–3218, Karlsruhe, Germany, May 2013.

[27] M. Milford, W. Scheirer, E. Vig et al., "Condition-invariant, top-down visual place recognition," in *Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5571–5577, Hong Kong, China, May 2014.

[28] C. McManus, B. Upcroft, and P. Newman, "Scene signatures: localised and point-less features for localisation," in *Proceedings of Robotics: Science and Systems X*, pp. 1–9, 2014.

[29] 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 Robotics and Automation Letters*, vol. 5, no. 2, pp. 1835–1842, 2020.

[30] A. C. Murillo and J. Kosecka, "Experiments in place recognition using gist panoramas," in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 2196–2203, Kyoto, Japan, September 2009.

[31] G. Singh and J. Kosecka, "Visual loop closing using gist descriptors in manhattan world," in *Proceedings of the ICRA omnidirectional vision workshop*, pp. 4042–4047, Alaska, AK, USA, April 2010.

[32] T. Nicosevici and R. Garcia, "Automatic visual bag-of-words for online robot navigation and mapping," *IEEE Transactions on Robotics*, vol. 28, no. 4, pp. 886–898, 2012.

[33] S. Khan and D. Wollherr, "Ibuild: incremental bag of binary words for appearance based loop closure detection," in *Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5441–5447, Washington, DC, USA, May 2015.

[34] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7, Minneapolis, Minnesota, June 2007.

[35] D. Arthur and S. Vassilvitskii, *k-means++: The Advantages of Careful Seeding*, , 2006.

[36] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, June 2007.

[37] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3304–3311, IEEE, San Francisco, CA, USA, June 2010.

[38] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," 2014, https://arxiv.org/abs/1411.1509.

[39] Y. Hou, H. Zhang, and S. Zhou, "Convolutional neural network-based image representation for visual loop closure detection," in *Proceedings of the 2015 IEEE international conference on information and automation*, Lijiang, China, August 2015.

[40] X. Zhang, L. Wang, Y. Zhao, and Y. Su, "Graph-based place recognition in image sequences with cnn features," *Journal of Intelligent and Robotic Systems*, vol. 95, no. 2, pp. 389–403, 2019.

[41] S. Wang, X. Lv, X. Liu, and D. Ye, "Compressed holistic convnet representations for detecting loop closures in dynamic environments," *IEEE Access*, vol. 8, pp. 60552–60574, 2020.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[43] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: a 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, NV, USA, July 2016.

[46] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," 2018, https://arxiv.org/abs/1805.07703.

[47] 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, 2020.

[48] Z. Chen, A. Jacobson, N. Sünderhauf et al., "Deep learning features at scale for visual place recognition," in *Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3223–3230, Singapore, May 2017.

[49] M. Lopez-Antequera, R. Gomez-Ojeda, N. Petkov, and J. Gonzalez-Jimenez, "Appearance-invariant place recognition by discriminatively training a convolutional neural network," *Pattern Recognition Letters*, vol. 92, pp. 89–95, 2017.

[50] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[51] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, Boston, MA, USA, June 2015.

[52] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, Hawaii, HI, USA, July 2017.

[53] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, Hawaii, HI, USA, July 2017.

[54] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size," 2016, https://arxiv.org/abs/1602.07360.

[55] A. G. Howard, M. Zhu, B. Chen et al., "Mobilenets: efficient convolutional neural networks for mobile vision applications," 2017, https://arxiv.org/abs/1704.04861.

[56] A. Howard, M. Sandler, G. Chu et al., "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, Montreal, Canada, October 2019.

[57] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: an extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, Salt Lake City, UT, USA, June 2018.

[58] M. Tan and Q. Le, “Efficientnet: rethinking model scaling for convolutional neural networks,” in *Proceedings of the International conference on machine learning*, pp. 6105–6114, Zhuhai China, February 2019.

[59] C. Campos, R. Elvira, J. J. G. Rodriguez, J. M. M Montiel, and J. D Tardos, “Orb-slam3: an accurate open-source library for visual, visual–inertial, and multimap slam,” *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[60] S. Hausler, A. Jacobson, and M. Milford, “Filter early, match late: improving network-based visual place recognition,” in *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3268–3275, Macau, China, November 2019.

[61] 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 Transactions on Robotics*, vol. 36, no. 2, pp. 561–569, 2020.

[62] Z. Xin, Y. Cai, T. Lu et al., “Localizing discriminative visual landmarks for place recognition,” in *Proceedings of the 2019 International Conference on Robotics and Automation (ICRA)*, pp. 5979–5985, Quebec, Canada, May 2019.

[63] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, “Shufflenet v2: practical guidelines for efficient cnn architecture design,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, Munich, Germany, September 2018.

[64] M. Zhu and L. Huang, “Fast and robust visual loop closure detection with convolutional neural network,” in *Proceedings of the 2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC)*, pp. 595–598, Greenville, South Carolina, November 2021.

[65] Z. Zhu, X. Xu, X. Liu, and Y. Jiang, “LFM: a lightweight lcd algorithm based on feature matching between similar key frames,” *Sensors*, vol. 21, no. 13, p. 4499, 2021.

[66] H. Baumgartl and R. Buettner, “Development of a highly precise place recognition module for effective human-robot interactions in changing lighting and viewpoint conditions,” in *Proceedings of the HICSS-53 Proceedings: 53nd Hawaii International Conference on System Sciences (HICSS-53)*, Hawaii, HI, USA, January 2020.

[67] Y. Chen, Y. Zhong, W. Wang, and H. Peng, “Fast and robust loop-closure detection using deep neural networks and matrix transformation for a visual SLAM system,” *Journal of Electronic Imaging*, vol. 31, no. 06, Article ID 061816, 2022.

[68] S. An, G. Che, F. Zhou, X. Liu, X. Ma, and Y. Chen, “Fast and incremental loop closure detection using proximity graphs,” in *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 378–385, Macau, China, November 2019.

[69] S. Sathiamoorthy, “Seagull optimization with deep learning driven condition invariant visual place recognition model,” *Indian Journal of Computer Science and Engineering*, vol. 13, no. 5, pp. 1497–1508, 2022.

[70] Y. Xu, J. Huang, J. Wang, Y. Wang, H. Qin, and K. Nan, “Esa-vlad: a lightweight network based on second-order attention and netvlad for loop closure detection,” *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6545–6552, 2021.

[71] Q. Gong, Y. Liu, L. Zhang, and R. Liu, “Ghost-dil-NetVLAD: a lightweight neural network for visual place recognition,” 2021, https://arxiv.org/abs/2112.11679.

[72] K. Zhang, J. Ma, and J. Jiang, “Loop closure detection with reweighting NetVLAD and local motion and structure consensus,” *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 6, pp. 1087–1090, 2022.

[73] Z. Liu, C. Suo, S. Zhou et al., “Seqlpd: sequence matching enhanced loop-closure detection based on large-scale point cloud description for self-driving vehicles,” in *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1218–1223, Macau, China, November 2019.

[74] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, “Patch-netvlad: multi-scale fusion of locally-global descriptors for place recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14141–14152, Nashville, TN, USA, June 2021.

[75] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, “Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 2, pp. 661–674, 2020.

[76] P.-E. Sarlin, F. Debraine, M. Dymczyk, R. Siegwart, and C. Cadena, “Leveraging deep visual descriptors for hierarchical efficient localization,” in *Proceedings of the Conference on Robot Learning*, pp. 456–465, Zürich, Switzerland, October 2018.

[77] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, “From coarse to fine: robust hierarchical localization at large scale,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12716–12725, Long Beach, CA, USA, June 2019.

[78] D. DeTone, T. Malisiewicz, and A. Rabinovich, “Superpoint: self-supervised interest point detection and description,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 224–236, Salt Lake City, UT, USA, June 2018.

[79] H. Yue, J. Miao, Y. Yu, W. Chen, and C. Wen, “Robust loop closure detection based on bag of SuperPoints and graph verification,” in *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3787–3793, Macau, China, November 2019.

[80] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[81] C. Park, H.-W. Chae, and J.-B. Song, “Robust place recognition using illumination-compensated image-based deep convolutional autoencoder features,” *International Journal of Control, Automation and Systems*, vol. 18, no. 10, pp. 2699–2707, 2020.

[82] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” 2013, https://arxiv.org/abs/1312.6114.

[83] N. Merrill and G. Huang, “CALC2. 0: combining appearance, semantic and geometric information for robust and efficient visual loop closure,” in *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4554–4561, Macau, China, November 2019.

[84] E. Garcia-Fidalgo and A. Ortiz, “ibow-lcd: an appearance-based loop-closure detection approach using incremental bags of binary words,” *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3051–3057, 2018.

[85] S. M. Siam and H. Zhang, “Fast-seqslam: a fast appearance based place recognition algorithm,” in *Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5702–5708, Singapore, June 2017.

[86] N. Sünderhauf, P. Neubert, and P. Protzel, “Are we there yet? challenging seqslam on a 3000 km journey across all four seasons,” in *Proceedings of the workshop on long-term autonomy IEEE international conference on robotics and automation (ICRA)*, Karlsruhe, Germany, May 2013.

[87] A. Glover, *Day and Night with Lateral Pose Change Datasets*, 2014.

[88] M. Larsson, E. Stenborg, L. Hammarstrand, M. Pollefeys, T. Sattler, and F. Kahl, "A cross-season correspondence dataset for robust semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9532–9542, Long Beach, CA, USA, June 2019.

[89] M. J. Milford and G. F. Wyeth, "Seqslam: visual route-based navigation for sunny summer days and stormy winter nights," in *Proceedings of the 2012 IEEE international conference on robotics and automation*, pp. 1643–1649, Minnesota, MN, USA, May 2012.