A Survey on Deep Visual Place Recognition

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 18, 2021, accepted January 23, 2021, date of publication January 27, 2021, date of current version February 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3054937

A Survey on Deep Visual Place Recognition

CARLO MASONE[1], (Member, IEEE), AND BARBARA CAPUTO[1,2], (Member, IEEE)
[1]Visual and Multimodal Applied Learning Team, Istituto Italiano di Tecnologia, 10141 Torino, Italy
[2]Department of Computer Engineering and Automation, Politecnico di Torino, 10124 Torino, Italy

Corresponding author: Carlo Masone (carlo.masone@iit.it)

This work was supported in part by the H2020 European Research Council under Grant 637076 (project RoboExNovo).

ABSTRACT In recent years visual place recognition (VPR), i.e., the problem of recognizing the location of images, has received considerable attention from multiple research communities, spanning from computer vision to robotics and even machine learning. This interest is fueled on one hand by the relevance that visual place recognition holds for many applications and on the other hand by the unsolved challenge of making these methods perform reliably in different conditions and environments. This paper presents a survey of the state-of-the-art of research on visual place recognition, focusing on how it has been shaped by the recent advances in deep learning. We start discussing the image representations used in this task and how they have evolved from using hand-crafted to deep-learned features. We further review how metric learning techniques are used to get more discriminative representations, as well as techniques for dealing with occlusions, distractors, and shifts in the visual domain of the images. The survey also provides an overview of the specific solutions that have been proposed for applications in robotics and with aerial imagery. Finally the survey provides a summary of datasets that are used in visual place recognition, highlighting their different characteristics.

INDEX TERMS Visual place recognition, image representation learning, deep learning.

I. INTRODUCTION

''Where was this picture taken?'' – understanding the location of a generic photo is a problem that has interested researchers for nearly two decades, under the name of *visual place recognition* (VPR). The last decade in particular has seen a drastic acceleration of the research in this field, driven by three forces. Firstly, the mass diffusion of smartphones and the consequential demand for new services that can leverage their integrated cameras, such as consumer photography, vision based navigation and augmented reality. Secondly, the large availability of publicly shared pictures on social media and other platforms, which can be used to locate interesting venues, holiday sites, restaurants, etcetera. Thirdly, the rise of mobile robots operating in the open world, e.g., self-driving cars, and the inherent challenge of their long term autonomy. Pertaining to the last point, recognizing places by vision is regarded as a key component for localization and navigation, being used for loop-closure in SLAM algorithms in GPS denied environments as well as an input to learn navigation policies [1] under different conditions. Remarkably, the development of visual localization in robotics is also paving the way for new applications of VPR, such as assistive technologies for people with visual impairments [2].

Such a variety of use-cases and application domains translates to a rich research panorama where VPR is studied by different communities (computer vision, robotics, machine learning) and with different problem settings. For instance, in computer vision VPR is often studied as the task of recognizing the location of a single image. In robotics, VPR algorithms can typically leverage streams of heterogeneous data (e.g., videos, pointclouds, odometry, etc.) as well as some knowledge of the motion of the robot. Moreover, in robotics there is a stronger emphasis on computational efficiency and real-time execution. Even the definition of *place* may change depending on the task: a place could be denoted by the name of a landmark, a GPS coordinate or even a 6 DoF pose with respect to a frame of reference.

Given the breadth of research in VPR and its fragmentation across multiple scientific domains, it is arguably challenging for scientists to have a comprehensive view of the state of the field. This challenge is exacerbated by the profound evolution of VPR that was prompted recently by the adoption of deep learning techniques. This manuscript aims at describing the state of research in VPR, by collecting, analyzing and systematizing studies pertaining that are published within

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

the communities of computer vision, robotics and machine learning. We focus on the most recent literature to identify the current research trends, particularly from the perspective of deep learning. Yet, this manuscript is not intended to be an introduction to deep learning, and we assume that the reader is at least familiar with basic concepts regarding convolutional neural networks. Finally, we remark that the goal of this paper is not to provide an empirical validation or comparison of the numerous methods discussed. On one hand, it would be infeasible to experimentally assess all the many approaches discussed. On the other hand, we think that such an evaluation study is better left to a benchmarking report with a much narrower scope than a survey.

A. VISUAL PLACE RECOGNITION: CONCEPTS AND ORGANIZATION OF THE SURVEY

Visual place recognition is, broadly speaking, the task of recognizing the place depicted in an image (or a sequence of images). This task is commonly addressed as an image retrieval problem. In this formulation, the prior knowledge of the places of interest for the task is represented as a collection of images (*database*). Each image in the database is tagged with an identifier of its location, e.g., the name of a landmark or a GPS coordinate. When a new picture needs to be localized (*query*), the place recognition system searches through the database for images that are similar to it. If similar pictures are found, their tagged locations are used to infer the location of the query. This retrieval process is typically implemented as a three-stages pipeline (see Fig. 1):

1) an encoding procedure extracts from each image a vector representation of its content (*image representation*);
2) a similarity search performs a pairwise comparison between the representations of the query and of every image in the database according to a scoring function (e.g., Euclidean distance or cosine similarity), and returns the best matches;
3) a post-processing stage refines the results produced by the similarity search.

The first part of this survey (Secs. II to VII) expands on the three stages of the VPR retrieval pipeline. First, Sec. II gives an overview of the hand-crafted representations that were used for image retrieval and VPR in the pre-deep learning era. Then, Sec. III moves on to discuss how these representations have evolved with the advent of deep convolutional networks (CNN), highlighting similarities and differences with the engineered descriptors. In particular, Sec. III focuses on the architectural aspect of the CNN-based representations, discussing the methods used in VPR to extract a vector description of an image using CNNs. Section IV reasons about the dimensionality of the representations and how to reduce it, which is important for the scalability of VPR to large databases. Section V delves into the topic of how the CNN models for extracting the image representations used in VPR are trained. Finally, Secs. VI and VII focus on the last two stages of the VPR retrieval pipeline, with a brief discussion on the similarity search and then a detailed review of the post-processing methods used to refine the results of the search. In this regard, we observe that most methods in this last stage are still based on engineered approaches, however few learning-based solutions are recently emerging, e.g., using CNN-based local feature descriptors for geometric verification or graph convolutional networks (GCN) to revise the results of the search.

Although VPR is formulated as an image retrieval task, there are specific challenges and use-cases in the recognition of places that set it apart from other retrieval problems. The second part of the survey (Secs. VIII to X) elaborates on these unique challenges and research questions. These challenges are largely related to the complexity of the scenes and to the dynamic nature of the world. First of all, images of places hardly ever present a single identifying object in the foreground. On the contrary, they usually contain multiple visual elements. Many of these elements may carry no useful information regarding the place and they might even occlude more useful objects in the background. Moreover, the appearance of landscapes naturally changes over time, not only because of dynamic objects or physical modifications (e.g., a temporary construction site), but also due to variations in illumination, weather and seasons. Other challenges in VPR arise from i) the presence of recurring elements and architectural patterns that make different places look similar, and ii) the large variety of viewpoints from which a place can be observed. Section VIII discusses these challenges that are peculiar to place recognition and identifies the research trends that have emerged to address them. The survey then covers the development of VPR in two application domains with very specific characteristics. The first domain is that of aerial images taken either from a high altitude satellite/aircraft or from a low altitude micro aerial drone. Section IX discusses how the viewpoint and the lack of distinctive visual details in aerial imagery make VPR in this setting rather different than when performed with street-level images. Afterwards, Sec. X analyzes VPR in the context of robotics. In robotics, place recognition is a task that is performed continuously during the navigation of the robot and it can often leverage multiple streams of data. These reasons, combined with a strong emphasis on computational efficiency, have led to unique developments of VPR.

At the beginning of this introduction we have established that VPR is commonly formulated as an image retrieval task and we have described how the survey covers different aspects of this formulation, particularly from the perspective of deep learning. However, the influence of deep learning in VPR goes beyond its application to the image retrieval pipeline shown in Fig. 1. In fact, the remarkable results achieved by deep classification models have led some researchers to investigate VPR as a classification task. In this formulation, unlike in image retrieval, the database images are categorized in classes, with each class representing a place. A deep CNN is then trained for a classification task on these images. At inference time the database images are not needed: new queries are run through the classifier which

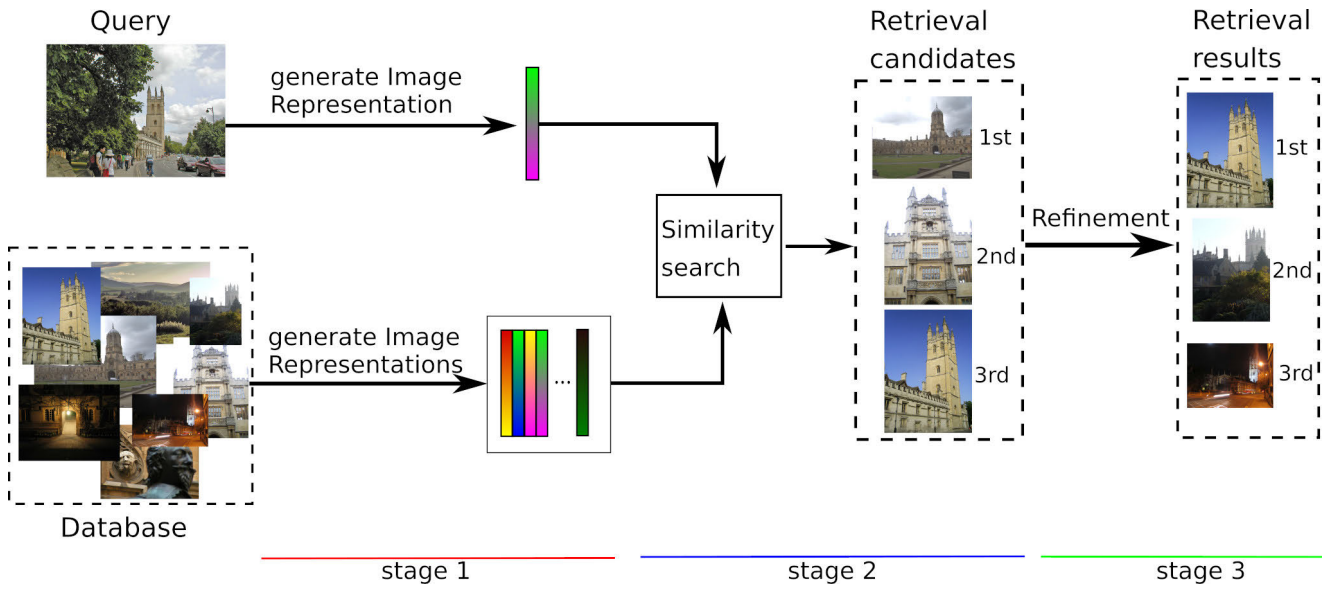

FIGURE 1. Visual place recognition is commonly formulated as an image retrieval problem. The known places are collected in a database and a new image to be localized is called query. The place retrieval is performed in three logical stages. 1) In the first stage, vector representations are generated for the query and the database images. From a practical perspective, the representation of the query is computed online, whereas the representations of the database images are computed offline. 2) the representation of the query is compared to those of the database images, to find the most similar ones (here only the top 3 are shown). 3) The best results of the comparison are further refined with post-processing techniques (here only the top 3 are shown).

predicts their corresponding classes, i.e., their places. Section XI discusses VPR as a classification task, looking in particular at the solutions that have been developed to partition large database of images, that are not necessarily a discrete set of landmarks, into classes.

Finally, the survey closes in Sec. XII with an overview of the evaluation metrics used in VPR, as well as with a comprehensive discussion on the publicly available datasets for VPR.

B. RELATED WORKS

Prior to this survey only few works have tried to provide an overview of the state of research in VPR. Lowry *et al.* [3] reviewed the evolution and state-of-the-art of visual place recognition, mostly from a robotics perspective. Their analysis focuses on the role of VPR in mapping and localization for mobile robots, discussing how the representation of the known world may include topological and metric information, besides appearance, and how all this information can be exploited. The paper also discusses image representations, albeit restricted mostly to classical hand-crafted or shallow learned descriptors. Strictly related to VPR, Piasco *et al.* [4] provide a survey on visual based localization. The main difference between visual place recognition and visual based localization is that the latter has the goal of precisely estimating the pose of the camera when it took the photo, whereas the former has the broader scope of recognizing the location. Therefore, a focal point of [4] is about methods that directly regress the pose of the camera. That survey also provides an insightful analysis on the role of heterogeneous data in VPR.

With respect to these prior works, this survey distinguishes itself not only for including more recent references, but most importantly for providing a review of visual place recognition from a different perspective. Although we comprehensively describe the broad landscape of VPR, we focus primarily on the advances that deep learning has introduced in this task. Among the other things we discuss the adoption of image representations based on convolutional neural networks and how these representations are trained using metric learning, we analyze multi-modal and multi-task architectures, we review the deep-learning based strategies that are emerging to cope with distractors and domain shifts. Given its different perspective, this survey can be considered complementary to [3], [4] and we encourage the reader who wants to get a broader view on the subject to consult also these other documents.

We also acknowledge that, during the time this manuscript was under review, another survey on VPR from the perspective of deep learning was published [5]. However, even with respect to [5], the present survey brings additional value because:

- the two surveys, which are the result of independent studies, organize and present the topics differently, thus providing readers with different points of view;
- this manuscript discusses aspects of VPR that are not touched upon in [5] (i.e., VPR as a classification task, in depth analyses on the application to robotics and aerial imagery) and delves with greater detail in some aspects of the problem that are discussed in both manuscripts (e.g., metric learning for place retrieval, techniques for refining the results of place retrieval);

- this manuscript cites around 25% more references than [5].

II. HAND-CRAFTED REPRESENTATIONS FOR PLACE RETRIEVAL

Visual place recognition is commonly framed as an image retrieval pipeline (see Fig. 1). This formulation relies on the ability to generate image representations that are discriminative w.r.t. places. This section briefly revisits the hand-crafted representations that were used for this task before the advent of CNNs. The following discussion is divided in two parts: representations generated from descriptors of local features and representations that describe an image as a whole. As it will be seen in Sec. III, the concepts and lessons learned from these representations provide a useful insight also for convolutional-based representations.

A. REPRESENTATIONS FROM LOCAL DESCRIPTORS

A local feature descriptor analyzes only a patch of the image, highlighting patterns that differ from its neighborhood [6]. These patches can be densely sampled [7], however in visual place recognition they are generally originated from a sparse detector that identifies points of interest (keypoints). Examples of detectors are the Hessian-Affine detector [8] or MSER [9]. A description is then extracted around each keypoint using methods such as SIFT [10], SURF [11], RootSIFT [12], BRIEF [13], DSP-SIFT [14] and kernel descriptors [15], [16].

Several studies, even recent ones, have showcased the use of hand-crafted detectors and local descriptors to represent images in VPR [17]–[30]. Two images can be compared by analyzing pairwise correspondences among their respective descriptors, however this approach is not effective and hardly scalable to a database-wide search. Not all detected features are discriminative for the task, so good features can be selected using shallow classifiers [27]. Far more effective and scalable is the idea that for searching the database the images should be compared by analyzing the statistics of their descriptors, rather than matching them on an individual basis. This idea was pioneered by Sivic and Zisserman [31] who adopted the Bag of Words (BoW) approach for image retrieval. In this method the descriptors are quantized in clusters, based on a codebook of visual words, and the image representation is then obtained as the histogram of the assignment of all image descriptors to visual words, weighted according to the ''term frequency – inverse document frequency'' (tf-idf). During retrieval, the images in the database are ranked based on the normalized inner product of their representation w.r.t. the query (i.e., the cosine similarity). Since the representation is sparse, retrieval can be implemented efficiently using an inverted file structure [32]. This was the first method that demonstrated efficient image retrieval, although on a small sized database. Following in its footsteps, other representations based on the quantization of local descriptors have been proposed, improving upon the BoW with a better ranking under the similarity measure [33]–[35], a reduction in the memory footprint [33] and a reduction in the number of visual words in the codebook [35]. Jégou and Zisserman [36] observe that methods that create a single vector representation from local feature descriptors can be regarded as two-steps approaches: i) an *embedding step* that individually maps each vector (feature) to a higher-dimensionality space, and ii) an *aggregation step* that generates a single representation from the mapped vectors. The rationale behind the embedding step is to improve the distinctiveness of the individual features and suppress false positives. For example VLAD embedding [33], [34] suppresses all matches between features that are adjacent to different centroids in the codebook. Notable examples of such methods are Fisher Vectors [35], VLAD [33], [34] and Triangular embeddings with democratic aggregation [36]. In [37] Tolias *et al.* proposed a general family of representations and similarity functions that, besides embedding and aggregation, includes a mechanism to select the contribution of each pair of descriptors per cluster. Using this formulation, the authors not only revisit methods such as BoW and VLAD, but also derive a novel *aggregated selective matching kernel* (ASMK) that achieves state-of-the-art performance in large scale place recognition. A regional version of ASMK was introduced in [38]. One property of these representations that makes them particularly good for image retrieval, as discussed in [39], is that they inherit to some extent the invariance properties (change in viewpoints, cropping, etc.,) of the local descriptors they are computed from.

B. REPRESENTATIONS FROM GLOBAL DESCRIPTORS

While the aggregation of local feature descriptors allows to obtain a single vector representation of an image, this can also be done directly using global feature descriptors, i.e., descriptors that encode holistic properties of the scene. Since they process the image as a whole, global descriptors do not require a detection phase, thus being less expensive to compute. Examples of whole-image descriptors are HOG [40] and Gist [41]. A low-dimensional binary coded representation of Gist was proposed in [42], which not only reduces the memory footprint but also allows for rapid recognition. Representations from global feature descriptors have been used in visual place recognition [26], [43]–[46]. Compared to the representation from local descriptors, global descriptors are less robust to viewpoint changes, clutter and occlusions. However, global descriptors like Gist are not dependent on illumination changes [47].

III. DEEP LEARNED REPRESENTATIONS FOR PLACE RETRIEVAL

Convolutional neural networks (CNNs) are a type of neural network that is specialized for processing data organized in a grid-like topology, e.g., images. CNNs have several remarkable properties, which have led them to become a powerful tool in different fields, including computer vision. In particular, since Krizhevsky *et al.* [48] demonstrated that deep CNNs can reach excellent performance in visual tasks, it has been recognized that these architectures can act

as powerful generators of image representations [49]–[51]. Moreover, it has been shown that CNNs can learn generic features that are, to some extent, transferable to other visual tasks [52]–[54]. These findings have also inspired the application of deep learned representations to image retrieval, where they have surpassed the performance achieved with handcrafted methods.

In the rest of this section we discuss how CNNs are used to generate image representations for VPR. Given the breadth of CNNs as a subject, we do not provide an introduction to it. That would be impractical to do in a limited space and it would unnecessarily dilute the survey. Rather we refer the reader to other sources for an introduction, e.g., [55], and hereinafter we assume that the reader is familiar with basic concepts of CNNs such as convolutional layers, pooling layers, fully connected layers, feature maps, etcetera.

A. FULLY CONNECTED REPRESENTATIONS

The first attempts at using CNNs as representation generators for image and place retrieval date back to 2014-2015, when several studies [49], [56]–[60] demonstrated that the vector of activations of a fully connected (FC) layer of a classification network pre-trained on ImageNet [61] could be effectively used for retrieval. Soon afterwards, it was shown that the better retrieval results were achieved with FC representations when the model was trained specifically for the retrieval task using a triplet loss [60], [62].

From this early studies it became soon clear that the information extracted by a FC layer is akin to a global descriptor: it is not robust to the presence of distractors or occlusions and lacks invariance to translation and scale. In order to mitigate these issues, a few studies tried to extract multiple sub-patches from the input image, each with a FC representation, and use each patch for retrieval [56], [58], [62]. Although such a strategy was shown to close the gap with classical hand-crafted representations from local descriptors, especially when considering low-memory footprint, it is computationally expensive and it does not solve all the limitations of FC layers. In particular, FC representations are limited by the fixed input size and by requiring large numbers of parameters.

B. CONVOLUTIONAL REPRESENTATIONS

The limitations of FC representations have inspired researchers to investigate the generation of image representations directly from the output of convolutional layers. The first study in this direction was proposed by Babenko *et al.* [57]. In that work, the authors demonstrate that the feature maps produced by a convolutional layer of a CNN trained for classification can be used as representation for place retrieval. More specifically, the authors take the $H \times W \times C$ tensor produced by a convolutional layer of the network, where H is the height of the tensor, W is its width and C is the number of channels, and flatten it as a vector. This vector is then normalized and used as image representation. A similar approach was demonstrated also in [63]. Despite the interesting use of the convolutional feature maps, the results achieved with this simple method are not far off from those obtained with FC representations. Intuitively, simply flattening the feature maps of a convolutional layer does not take full advantage of the spatial information contained therein. This consideration has guided the development of the current state-of-the-art representations for place retrieval. These methods can be categorized into two families:

- aggregation of the convolutional features using methods derived from hand-crafted representations of local descriptors;
- pooling methods that summarize the convolutional features.

1) AGGREGATED REPRESENTATIONS

Rather than collapsing the $H \times W \times C$ features extracted from a convolutional layer to a vector, they can be considered as a $H \times W$ grid of C-dimensional feature descriptors, each one having a limited receptive field. Namely, the output of a convolutional layer can be assimilated to a set of densely extracted local descriptors. Following the lessons learned from classical non-learned approaches, these dense descriptors can be aggregated in a single vector representation and then compared using a similarity function (e.g., Euclidean distance or cosine similarity). Several studies demonstrated the applicability of classic encodings to these dense convolutional descriptors, e.g., VLAD [64], BoW [65], [66], ASMK [67]. Moving further, researchers have proposed aggregation modules that can be plugged on top of a CNN and allow end-to-end learning. In [68], the authors combine a fully convolutional network to a Fisher vector module. By computing the gradient of the contrastive loss w.r.t. the parameters of the Fisher Vector, this module can be trained together with the CNN. In [69] it is introduced NetVLAD, a layer that implements the VLAD embedding and aggregation with differentiable operations, thus allowing end-to-end training of the network. Moreover, NetVLAD presents more trainable parameters than VLAD, hence providing more flexibility.

2) POOLED REPRESENTATIONS

Researchers have shown that convolutional features from mid/late layers, unlike shallow non-learned features, can be successfully aggregated and compared without embedding. Babenko and Lemiptski [70] show that for shallow hand-crafted features like SIFT, the embedding step is fundamental to improve their discriminativity. However, they argue that raw convolutional features have a higher discriminative capability and therefore they can be pooled together with simpler schemes, thus providing not only a leaner pipeline and, in many cases, more compact representations, but also improving performance. Namely, an image representation can be generated by summarizing the statistics of the convolutional features. The simplest pooled representation is achieved by max-pooling the feature maps of a convolutional layer. Despite its simplicity, the low-dimensional

representation obtained with this scheme can outperform more complex hand-crafted representations with a similar memory footprint scheme [71]. Another popular representation is obtained by a parameterless sum-pooling of convolutional feature maps (SPoC) [70], which can be interpreted as an implementation of a simple match kernel. Sum-pooling has been shown to perform better than max-pooling, especially when the image representation is whitened [69], [70]. An intuition about these two pooling strategies is elaborated in [72]. The authors observe that max-pooling is more invariant to scale changes, whereas sum-pooling is less sensitive to distractors in the feature maps. To combine the advantages of both methods, they test an hybrid pooling that concatenates the sum and max-pooling descriptors. Inspired by the max-pooling described in [51], [71], Tolias *et al.* [73] design a new pooling procedure that encodes multiple regions. Rather than extracting multiple patches from the image and making a forward pass for each of them, they select regions directly on the feature maps using a uniform sampling scheme. A max-pooled vector is computed for each region and these regional descriptors are then summed and ℓ_2 normalized. This descriptor, called "Regional Maximum Activations of Convolutions" (R-MAC) can be implemented with integral images, which only requires specifying one parameter and yields a more efficient computation. This implementation reduces to sum-pooling for a specific choice of the parameter. Later, [74] introduces an implementation of the R-MAC descriptor using differentiable operations, thus yielding a module that can be plugged atop any CNN and that allows end-to-end training. R-MAC uses different pooling operations to capture multi-scale information from the regions. This approach is modified in [75] in two ways: i) the different pooling operations are applied to the whole image, and ii) the obtained feature maps are concatenated in a pyramid. Then, the multi-scale feature maps of the pyramid are fused using 1×1 convolutions. With this fusion operation, the network learns to combine the multi-scale context at each location. A generalization of sum-pooling is presented in [76] by using a cross-dimensional weighting scheme before the sum-pooling. The weights across both dimensions are engineered based on heuristic. The spatial weighting is based on the normalized total response across all channels and it tends to boost the response for locations in which multiple channels are active, which likely correspond to salient regions. The channel weighting is based on the sparsity of the feature maps. Conceptually, it is similar to an inverse document frequency and it boosts the contribution of rare features in the overall response.

Together with R-MAC, the current state-of-the-art pooling method is the generalized-mean aggregation layer (GeM) [77]. This layer implements a parametric generalized-mean. There is one parameter per feature map, however they can be shared reducing the parameters count to one. GeM generalizes both max and average pooling (SPoC), which can be considered as special cases corresponding to a proper selection of the parameters. Since the pooling operation is differentiable, the parameters can be learned as part of the back-propagation. Experiments show that GeM consistently outperforms max-pooling (MAC) and average pooling (SPoC), and even R-MAC (in the implementation with fixed region sampling).

IV. DIMENSIONALITY REDUCTION AND WHITENING

Another important aspect of image representations, besides the retrieval performance they yield, is their dimensionality. Intuitively, the number of dimensions of an image representation is directly connected to the size it occupies in memory. This has practical consequences for scalability, considering that for the similarity search in the retrieval pipeline (see Fig. 1) the representations of all the database images should be loaded in memory. Therefore, when developing a VPR system that needs to be deployed to large environments, i.e., with a large database of images, reducing the memory footprint of the representations becomes critical. Additionally, reducing the dimensionality of representations is also helpful to reduce the retrieval time. Scalability has indeed motivated many recent works to investigate short-code representations [57], [68], by adopting different dimensionality reduction approaches. With hand-crafted descriptors there is evidence that dimensionality reduction and whitening can, in some circumstances, slightly improve performance over the original embeddings [78], [79]. For example, Jégou and Chum [39] explain that PCA and whitening help with exploiting negative evidence and mitigating the problem of co-occurrences. Many recent studies adopt dimensionality reduction approaches with deep learned features [49], [56], [57], [67]–[70], [73], [77], [80]–[85], suggesting that the learned descriptors are better suited to compression. An explanation for this is that the network learns to discard much of the information that is irrelevant, thus allowing for a more aggressive dimensionality reduction [57]. However, it is also revealed that the effectiveness of the reduction may depend on the aggregation method. For example, training the PCA for high-dimensional engineered descriptors needs a lot of data and it is prone to overfitting [70]. Early works with FC descriptors report improvements when using PCA and whitening [49], [56], [57]. In [56], where a multi-patch approach is used, the dimensionality reduction is applied both to the FC descriptors and to the aggregated representation. Starting from a 4096 dimensional (4096-D) descriptor, the performance degradation is negligible up to 256-D and 128-D. Replacing the PCA with a learned projection matrix that optimizes distances of the projected features can further reduce the compression degradation.

Several works have confirmed that the effectiveness of dimensionality reduction and whitening on convolutional representations depends significantly on the method used to build the representation. PCA and whitening are shown to be more beneficial with sum-pooling than with max-pooling [68]–[70], [76]. One explanation of this phenomenon is that in sum-pooling whitening helps suppressing the contribution of features that are both common

across images and bursty. For max-pooling, burstiness of popular features is a minor issue and whitening actually causes a drop in performance [69]. Ong *et al.* [68] confirm that PCA-whitening works better for sum-pooling, whereas max-pooling shows better results when compressed with linear discriminant analysis (LDA). Several alternatives to PCA-whitening have also been used with convolutional descriptors. In [83] the authors use the linear discriminant projections originally proposed by Mikolajczyk and Matas for SIFT features [78], directly learning it on the training data. While this approach works better for high-dimensional representations, PCA is shown to be superior for very compact codes (64-D or less). The PCA projection is implemented in [80] with a shifting and a FC layer, which can be advantageously trained with the rest of the network. This implementation shows results that are comparable to classic whitening, albeit being sensitive to initialization (random orthogonal projection is reported to work best). This is explained by the fact that the FC layer introduces a huge number of parameters and it is prone to overfitting [69]. Another approach is to learn the projection matrix from the representations of semantic landmarks in the image, and then use it to reduce the dimensionality of a holistic descriptor [86], [87]. This solution effectively allows integrating both image-wide information and the information from the semantic landmarks in a unique representation.

While PCA whitening is an effective way to solve the problem of over-counting and co-occurrences [39], it may excessively penalize over-counting. Zhu *et al.* [84] observe that whitening balances the energy across the dimensions of the rotated descriptor, but they argue that it would be beneficial if the variance of the first few dimensions is preserved to some extent. For this reason they introduce a *PCA power whitening*, in which the variance scaling is modulated via a parameter. This parameter allows changing the tradeoff between reducing over-counting and preserving the energy distribution of the features. Experiments show that the power whitening improves upon the classic whitening, and it can even add a small gain to max-pooling (whereas the PCA whitening generally worsens results). A similar formulation is also proposed in [85] in the context of patches similarities. In [67] the authors use a convolutional autoencoder (AE) module to learn low-dimensionality local descriptors. This approach is appealing because the AE can be integrated in the network and trained with it by adding a reconstruction loss, without the need to perform post-processing learning steps as in the case of PCA. However, care must be taken to control the flow of the gradients from the autoencoder to the backbone. Experiments show that AE outperforms PCA and a simpler dimensionality reduction using a single FC layer.

V. LEARNING TO RETRIEVE PLACES

Section III introduced the topic of CNN-based representations used in VPR but only from an architectural perspective. However, besides the aggregation or pooling method used to build these representations, their effectiveness depends also on how they are learned from the data. This section discusses the approaches that are used to train the CNNs as representation generators for VPR, highlighting also new directions of research in this regard.

A. LEARNING FROM CLASSIFICATION

As mentioned in Sec. III, the first CNN-based representations used in VPR were actually generated from models trained for a classification task, not for retrieval. This is motivated by the fact that classification is the first visual task where CNN demonstrated extraordinary results, and also by the observations that CNNs can learn generic features that are, to some extent, transferable to other visual tasks [52]–[54].

The first attempts with CNN-based descriptors for image retrieval amounted to using an off-the-shelf classification network pre-trained on ImageNet as feature extractor [49]. Although the pre-trained network is shown to be reasonably capable of localizing queries, the retrieval performance was noticeably improved in [57] by fine-tuning on a landmarks dataset that is much closer to the target domain for urban place recognition. Despite these improvements, it is generally argued that learning for a classification task is sub-optimal because the extracted features are not necessarily suitable for the retrieval task [69], [80], [83]. One point to support this objection is that descriptors trained for classification learn to distinguish between semantic classes but are robust to intra-class variability, which is undesirable for instance retrieval [80].

Nevertheless, there a few recent studies that have gone back to investigating the use of classification models to generate global representations for retrieval (i.e., FC representation, see Sec. III-A). For instance, [67], [88] use the ArcFace loss [89] to train the global features with image level labels, achieving good retrieval results under the cosine similarity. There is a couple of reason that motivate this interest in revisiting classification models as generators of image representations for VPR. On one hand, the global descriptors they produce can be quite compact. On the other, they only need image level labels, without incurring in the cost of mining examples that is discussed in the next section for metric learning. Both these aspects are relevant for scalability to large databases.

B. LEARNING TO RANK

Image retrieval is akin to a "learning-to-rank" problem, therefore it naturally lends itself to metric learning, i.e., learning image descriptors that represent well the similarity under a distance function. Indeed, most studies in VPR train the CNN that generates the image representations with one of two ranking losses: the contrastive loss, using a siamese network setup, [68], [77], [83], or the triplet loss, using a triplet network setup [69], [74], [80], [90]–[92]. Albeit different, these two losses are based on a similar idea. For each training sample the model is provided also with positive and/or negative examples. The two losses enforce that the learned representation for the training sample is close to that of positive

examples and distant to that of negative examples, according to a metric. In the context of VPR, a positive example is an image of the same place as the training sample, whereas a negative example is an image of a different place. There are no explicit comparisons between these two losses in VPR, with only a few studies providing some indications: [93] reports superior performance using a triplet network, but only for a classification task; on the other hand, [77], [83] mention using a siamese architecture because it yielded faster convergence and better generalization, but no evidence is given.

Various modifications of the classic contrastive and triplet losses have been discussed in image retrieval. Mishchuk *et al.* [66] propose a triplet loss variant that maximizes the distance between the closest positive and closest negative example in the batch. In [94] a Quadratic Hinge Triplet loss which constraints first-order similarity is paired with a loss that regularizes second-order similarity for local descriptor learning. This idea is also applied to global descriptors learning in [95]. In [96], the triplet loss is modified by setting the distance to the hardest negative example to a constant value, so that the corresponding derivative of the loss is set to zero. The effect is to bring positive samples together and then distributing them in the space, satisfying the triplet margin criterion. A multi-scale version of the triplet loss is proposed for place recognition in [97], so as to create embeddings from features extracted at multiple layers.

As mentioned earlier, representation learning via ranking losses requires selecting positive and negative examples for each training image. This topic is discussed next.

1) EXAMPLES MINING

The selection of positive and negative examples is crucial when training a model with contrastive and triplet losses. If they are too easy, the network will not learn to properly discriminate the images. On the other hand, forcing the network to learn extremely hard examples could lead to overfitting [83] and bad local minima [98]. The pairwise similarity learning process is also not very tolerant to outliers, therefore it needs clean training data [80]. Another challenge in mining hard examples is efficiency, especially for scaling to large databases. Various solutions have been proposed to mine positive and negative examples. In [99] the authors use stochastic sampling to select a set of samples, propagate them through the network, and retain only those with the largest losses. A similar approach is taken in [80] and the authors report that, to reduce the computational effort, they mine hard triplets every 16 network updates.

Learning through examples allows for weakly-supervised [69], [91], [100] or unsupervised training [77], by exploiting additional information to guide the mining process. Radenovic *et al.* [77] exploit 3D models (clusters) constructed via structure-from-motion (SfM) to inform the choice of the examples. Hard negatives are mined from clusters different from the one the query belongs to. Positive examples are instead selected at random from the images that co-observe enough 3D points with the query, but with a threshold on the scale. Since the positive examples do not depend on the current state of the CNN, but only on the images and 3D models, they can be mined once and then kept fixed during training. GPS information is exploited in [69], [91], [100]. In [91], the set of positive examples is selected as those images that are within 50 meters from the GPS tag of the training query. Considering that images from the same GPS coordinates could be taken pointing the camera at different directions, the positive candidates are refined via geometric verification. For the negative examples, the authors mimic the image geo-localization process within the training batch and for each iteration they pick as a negative candidate the top retrieved image at least 225 meters away from the GPS location of the training query. Since a learning based only on the hardest negatives can lead to a bad local minima [98], some negatives are also randomly selected from the batch. The main difference in [69], [100] is that the loss considers all the negative examples for each training sample. A naïve computation of all negatives is infeasible because it would require for each query to perform a forward pass on all database images. Moreover, many negative examples would have a negligible contribution to the loss, so considering them would be a waste of computations. The authors propose three strategies to make the mining process more efficient:

- Sampling: the loss is computed only for a set of negative samples and each step inherits the hardest negatives from the previous epoch.
- Caching: the representations are cached and recomputed after a certain number of training queries. This number can be chosen depending on the learning rate.
- Clustering: the queries are clustered according to their GPS location and the queries in a cluster share the same negative examples.

C. LISTWISE RANKING

Although the contrastive and triplet loss are the most popular methods used in VPR for learning image representations, they both come with two limitations. The first limitation is practical: the procedures for mining examples can add a significant overhead to the training and might even lead to poor results if the examples are not chosen properly. The second limitation is theoretical: these losses have been shown to be only upper bounds on the mean average precision (mAP) [101]. Hence, optimizing these ranking losses is not guaranteed to also optimize the mAP. A few recent works have instead proposed to use a different loss that can address both limitations. The idea is to directly optimize the mAP by leveraging a *listwise loss* formulation (Fig. 2) [82], [102], [103]. This formulation approximates the non-smooth and non-differentiable AP using the method of histogram-binning with a differentiable soft-assignment. This allows to compute a quantized and differentiable mAP and to use backpropagation. From an implementation perspective, backpropagation with the listwise-loss needs large batch sizes, which is generally intractable. This problem can be bypassed using a multi-stage backpropagation [82] or heuristics to split the

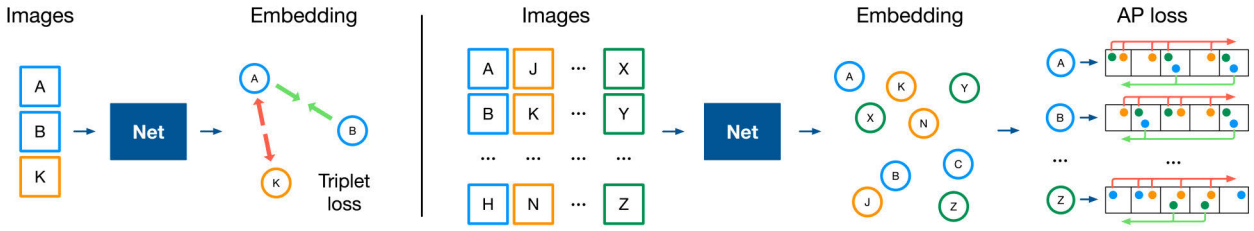

FIGURE 2. The triplet loss (left) only considers few examples at a time. The listwise loss (right) considers all the images simultaneously and directly optimizes the AP. Image from [82], Copyright ©2019, IEEE.

batch in mini-batches [102]. Experiments in image retrieval demonstrate that learning with the listwise loss consistently surpasses the results achieved with contrastive or triplet loss, even compared to methods that perform multi scale analysis. Moreover, this is achieved by using a smaller number of forward/backward passes, a smaller number of iterations, with a training process that is several times faster and without the need to mine hard positive and negative examples [82].

D. LEARNING FROM EXPERT KNOWLEDGE

Learning to rank allows training a network directly for the retrieval task, but the process can be long and costly. If one such network has already been trained, knowledge distillation [104] can be used to train a student network. This is the approach followed in [105], [106], where the student network is based on a lightweight MobileNet architecture. Following a similar concept, the feature extractor for the retrieval task can also be trained from an expert system based on hand-crafted features. For example, [107] uses an autoencoder architecture that, instead of reconstructing the input image, is tasked to decode a holistic feature vector that tries to reconstruct a handcrafted HOG descriptor. The HOG descriptor provides the geometric prior knowledge needed to train the network, while also guaranteeing some invariance to illumination conditions. This approach is shown to be capable of producing a lightweight feature extractor and it can be trained without supervision. The limitation of these methods is that they depend on the availability of an expert system.

VI. SIMILARITY SEARCH

The second step of the retrieval procedure is a k-nearest neighbour search (kNN), i.e., finding the k database instances that are closest to the query. Albeit simple, this task is quite expensive. Even though there are efficient algorithms that implement exact nearest neighbour search for low-dimensional cases, in high-dimensional problems they can even be outperformed by a naïve linear search due to complex effects (curse of dimensionality [108]). The search can be drastically sped up using approximate nearest neighbour methods (ANN) that perform a non-exhaustive search implemented using different indexing structures, encoding and stopping criteria [18], [59], [78], [109]–[113]. Efficient implementations of several approximate nearest neighbour algorithms are available in the FLANN library [109] and in the more recent FAISS library [113] which also supports GPU operations. The similarity search for visual place recognition can also be implemented by matching multiple features per image. This approach has been demonstrated using a nearest neighbour (NN) for each individual local feature in the query and resorting to techniques such as the Generalized Minimum Clique Problem [114] or the Dominant Set Clustering [30]. These techniques are combined with NN pruning strategies and with filtering based on global features to shortlist the results from the matched images.

For image retrieval, it is also important to consider the memory requirements of the indexing structure of the similarity search method because large image representations can lead to unsustainable memory footprints for big databases. In the literature of image retrieval and visual place recognition, several techniques have been used to make the similarity search more efficient and scalable. Indexing structures based on an inverted file [115] have been adopted to implement a non-exhaustive search that is particularly effective with sparse vector representations [24], [31], [37], [65], [81], [116]. In [24] retrieval time is reduced by first grouping similar database images and then performing the matching by cluster. Quantization techniques such as k-means [116], [117], binarization [35], [37], [67], [118], [119] and product quantization [74], [81], [120], [121] have been used to reduce the memory requirements for storing the data, in some cases by more than one order of magnitude. They have also been combined with asymmetric distance computation [120] and multiple assignments [37], [118], [120], [122]–[124] to mitigate the quantization errors. The inverted index has also been generalized to work with product quantization [58], [120], [125], further improving the speed and accuracy of the search, at a small memory cost. For an in-depth review on the topic of approximate and efficient methods for nearest neighbors search that is out of the scope of this document we refer the reader to [111].

VII. RETRIEVAL REFINEMENT

The shortlist of database images retrieved by the similarity search provides a set of hypotheses of the place corresponding to the query. Due to the complexity of representing the similarity, noise in the data and approximations, the hypotheses can contain false positives or they can miss relevant instances from the database. This section reviews several methods that

can be applied to improve precision and recall by re-ranking and even expanding the shortlist of candidates.

A. SPATIAL VERIFICATION

Spatial (geometric) verification is a popular technique for boosting the performance of image retrieval and particularly visual place recognition [24], [66], [67], [96], [106], [116], [117], [119], [124], [126]–[131]. The gist of this method is to first detect feature-to-feature correspondences among a pair of images and then verify their reliability by analyzing the consistency of spatial transformations between them. The result of this analysis is then used to re-rank the shortlisted results. Although spatial verification is generally used at the refining stage, the same principle can also be used as a procedure clean the database from labeling noise [80] before setting up the place retrieval pipeline.

Spatial verification is typically implemented by using model-based methods, such as RANSAC [116], [132] or PROSAC [133], to generate transformation hypotheses based on feature-to-feature correspondences, which are typically pruned by imposing different kinds of constraints, such as geometric [24] or semantic [128]. Each hypothesis is evaluated based on the number of ''inliers'' among all features under that hypothesis, which can then be used as score for re-ranking. Alternatively to model-based methods, some works use model-free methods for the verification step [124], [128].

Spatial verification methods use sparse local descriptors to detect correspondences and check the consistency between images. In pre-deep learning VPR, the same hand-crafted local descriptors that were aggregated to build the representation of an image for the retrieval could also be used for spatial verification [117]. However, the transition to CNN-based image representations raises the question of how these sparse local descriptors for spatial verification can be extracted, given that the methods discussed in Sec. III to generate holistic image representations do not make sparse local descriptors readily available. Aside from the naïve solution of additionally computing hand-crafted local feature descriptors especially for the spatial verification [119], new approaches that leverage CNNs have been proposed. These methods can be categorized in three families:

- use a single CNN specialized to generate image representations and then extract sparse local descriptors from the same network using some heuristics;
- use two CNNs, one to generate image representations and the other to extract sparse local descriptors;
- use a hybrid CNN trained to both generate holistic image representations and to extract sparse local descriptors.

1) HEURISTIC EXTRACTION OF SPARSE LOCAL DESCRIPTORS

These methods aim at simplicity and efficiency, trying to extract spatial local descriptors without the need of retraining the model for this task nor a second model. One strategy proposed by Taira *et al.* [119] is based on the observation that the feature maps from a convolutional layer of a CNN can be interpreted as a dense grid of local feature descriptors, but these descriptors can be matched in a coarse-to-fine-manner to sparsify them. For this purpose, the authors first find broader matches among the features of the fifth convolutional layer (*conv5*) and then look for matches in the finer features from the third convolutional layer (*conv3*) restricted by the already found *conv5* correspondences. Another solution is to extract sparse local descriptors from the CNN used for the first stage by selecting high activations of the convolutional feature maps [128], [131]. This method is based on the observation that the output of a convolutional layer can be interpreted as a collection of 2D response maps of pattern detectors. Hence, the selection of the high activations can be seen as choosing the local features with the most confident detections.

2) SEPARATE MODEL FOR EXTRACTING SPARSE LOCAL DESCRIPTORS

This family of methods aims at using a separate model specialized to extract the sparse local descriptors. These methods are based on the observation that the feature maps from a convolutional layer of a CNN can be interpreted as a dense grid of local feature descriptors because they lack the detection step that is instead used for hand-crafted local descriptors. Therefore, they propose to combine the detection and description steps in the model that is specialized for extracting the sparse local descriptors (Fig. 3).

An architecture for this purpose that is tailored for VPR is DELF [81]. In DELF the detection step is implemented as an attention module that sits on top of the convolutional layer and weights its activations. Effectively, this module works as a keypoint detection, albeit the detection happens after the description step. Revaud *et al.* [103] argue that salient regions are not necessarily discriminative and that the model must learn to detect keypoints that are both repeatable and reliable for matching. For this purpose, they filter local descriptors extracted at each pixel position based on a repeatability map, learned in a self-supervised way, and a reliability map, trained using a modified listwise loss. A non-learned detection approach is used in [134], where detections are obtained by performing a non-local-maximum suppression on convolutional feature maps followed by a non-maximum suppression across each descriptor.

3) HYBRID MODELS TO EXTRACT BOTH IMAGE REPRESENTATIONS AND SPARSE LOCAL DESCRIPTORS

The use of two specialized models, one to generate image representations and the other to extract local descriptors, clashes with the limited resources and need for efficiency that arise in many applications. For this reason several researchers have investigated hybrid solutions that combine the computation of both global descriptors (for similarity comparison) and local descriptors (for spatial verification) into a single CNN with multiple heads. This approach is used in DELG [67], where local and global features are extracted from a common backbone with two heads: i) a GeM pooling that produces

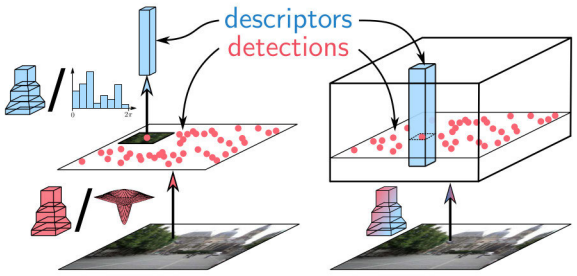

FIGURE 3. Different approaches for detecting keypoints and describing local descriptors from CNN features. a) Classical pipeline. b) Pipeline that combines the two steps. Figure from [134], Copyright ©2019, IEEE.

the global representation, and ii) an attention module inspired by DELF [81] to produce the local descriptors. In order to train the two tasks simultaneously, the authors leverage the concept of hierarchical representations in CNNs [53]: global features are associated with the deep layers of the network that encode high-level cues, while local features are associated to the mid-levels that encode more localized information. Therefore, at training time, only the gradient of the similarity loss of global descriptors is propagated back to the backbone whereas the gradients of the losses concerning the local descriptors are stopped before. This is due to the observation that a naïve optimization of the three losses would disturb the hierarchical feature representation and produce weak models. Another two-headed architecture is proposed in [106], but in this case using distillation to learn the tasks directly from off-the-shelf teacher networks. In particular, the authors use a NetVLAD [69] based network for the image representations and SuperPoint [135], a generic detector-descriptor architecture, to extract the local descriptors.

B. NON-GEOMETRIC RE-RANKING

Even though spatial verification is the most popular method for re-ranking, other methods that do not rely on geometric correspondences are also used. In [73] the re-ranking stage is performed by computing the matching scores between the MAC representation of the query and all the individual R-MAC regions for the database image. The shortlisted images are re-ranked based on the maximum similarity between their regions and the query. In [88], [136] the authors use a discriminative ranking method based on the similarity of labels assigned to the images by a kNN search with soft voting. Namely, the search results are re-ranked by first moving up all the shortlisted images that have the same label as the query, and then by adding the images from the database with the same label as the query and that were not retrieved by the search. Another manually engineered re-ranking method is presented in [137]. This solution is a brute force algorithm based on matching descriptors from mid-level convolutional layers while accounting for their spatial location.

C. QUERY EXPANSION

One of the most successful and widely used techniques to improve the retrieval result is query expansion (QE) [12], [23], [37], [65], [66], [70], [73], [74], [76], [77], [80], [82], [88], [92], [96], [123], [126], [127], [129], [130], [138]. The idea of query expansion is to use the shortlisted images as a feedback to produce an enriched representation that is re-submitted for a new search through the database. This solution can significantly increase the recall by retrieving relevant images that were not selected with the first search. However, it requires the initial candidates to be reliable and accurate enough, hence it benefits from a prior verification step. Moreover, queries with few relevant images might see a degradation in performance after query expansion [129].

There are several versions of QE that are commonly used. Average Query Expansion (AQE) [126] creates the enriched representation as the average of the high ranked results. The Discriminative Query Expansion (DQE) [12] instead uses the top and bottom ranked results as positive and negative examples to train a linear SVM. The SVM learns a weight vector that is then used to re-rank all the candidates. The Hamming Query Expansion revisits query expansion making it compatible with Hamming Embedding [123]. The α-weighted query expansion (αQE) [77] is a generalization of AQE that uses a weighted average. Namely, each of the top retrieved results is weighed by its similarity score raised to the power of a tunable scalar parameter.

FIGURE 4. Example of diffusion. Left) retrieval from two queries using a nearest neighbour search. Right) the diffusion process allows bettering capturing the underlying data manifold. Image from [139], Copyright ©2013, IEEE.

D. DIFFUSION

One of the limitations of retrieval by similarity search is that the pairwise formulation ignores the structure of the data manifold. Instead, similarities could be estimated more accurately along the geodesic path on the data manifold. Even query expansion, which has been shown to boost the retrieval performance, only uses the closest neighbours selected according to the pairwise similarity values to issue new queries. In contrast to these methods, diffusion is a technique that exploits the context similarities between all elements of the database to unveil the data manifold and it uses this information to perform a search in a principled way (Fig. 4). The manifold here is interpreted as a weighted graph, where each instance is represented by a node and the weight on an edge is a pairwise similarity measure between the connected nodes. The diffusion process then follows the concept of a random walk that propagates a ranking score through the whole graph.

The diffusion technique has recently gained popularity in retrieval tasks because it has been shown to significantly increase the performance [139], but this comes with

some caveats. Firstly, this is a very expensive procedure that can dominate the computational time of all other phases in the retrieval. Secondly, the diffusion process assumes that the query is part of the graph (i.e., the database), which is not the case in visual place recognition. A major step towards overcoming these issues was the Regional Diffusion algorithm by Iscen *et al.* [140], which included several strategies to make the diffusion refinement more efficient. Firstly, new queries are handled without augmenting the original graph but rather by expressing the vector that selects the initial graph nodes in terms of the top ranked nearest neighbors retrieved with the similarity search. The scores are then propagated with a conjugate gradient solver. To further speed up the online computation, diffusion is applied only on the truncated graph corresponding to the nearest neighbors, with a trade-off between precision and size of the truncation. This technique is also generalized to handle multiple query vector representations, typical of regional methods. The Fast Spectral Ranking algorithm (FSR) [141] improves upon [140], by moving more computations offline, effectively reducing the online stage to a sequence of sparse matrix-vector multiplications. The Regularized Diffusion Process (RDP) [142] is an algorithm that introduces a smoothness criterion that simultaneously regularizes four vertices in the affinity graph. This regularization is used to guide the iterative diffusion process on the tensor product graph. An even more significant speed boost is achieved by Yang *et al.* [143]. In comparison to Iscen's method their algorithm completely moves offline the computation of the graph Laplacian that is used for the diffusion, thus reducing the online process to a linear combination of precomputed vectors. With this solution, the cost of the diffusion process becomes almost negligible w.r.t. the nearest neighbour search. Additionally, the authors use a late truncation (truncation of the Laplacian) and demonstrate that, contrary to Iscen's early truncation (truncation of the affinity matrix) [140], this does not reduce performance. The motivation for this is that the subgraph obtained with an early truncation contains incomplete manifolds and the later normalization raises the probabilities to reach nodes on such incomplete manifolds. The improved truncation allows implementing more sizeable graph reductions with a benefit in terms of the memory footprint. Regarding the offline computation, building the affinity matrix of the graph requires an exhaustive pairwise similarity check among all the images of the database. Even though the operation is not as critical as the online stage, this procedure cannot scale to large databases. Generally an approximate NN search is used [140]. Magliani *et al.* [112] propose an ANN algorithm based on local sensitivity hashing that is specifically tailored for the diffusion task. Inspired by diffusion, [130] introduces a graph traversal approach called Explore-Exploit Graph Traversal (EGT) to be applied to the kNN graph from the similarity search. The main idea is to combine the strength of QE (exploiting the neighbours) and diffusion (exploring the descriptor space) by alternating exploitation and exploration steps. Additionally, a variant of this traversal algorithm includes spatial verification to adjust the weights of the edges. The spatial verification helps mitigating the problem of topic drift – the exploration drifting away from the original query – and improves results on a number of benchmarks.

The essential idea of classic diffusion methods is to unveil the manifold structure to better guide the similarity search. This idea has been revisited with the use of graph convolutional networks (GCN). GCNs can be used to encode the information from the kNN graph directly into the image descriptors used for the similarity search. This generates new descriptors that encode the high-order neighbour information. In [144] this idea is explored with a GCN that is trained without supervision by using a loss function inspired by the concept of clustering: similar descriptors should move closer, while dissimilar descriptors should be pushed apart. After the model is learned, the image descriptors can be forwarded through it to get the updated representation. At inference time, computing the updated representation for the query requires first to update the adjacency matrix of the graph, which can be done in an approximated manner to limit the time cost. This method depends on the quality of the adjacency matrix, so the authors suggest using spatial verification in the offline construction of the graph of the database images. A similar idea is explored in [145], where three different implementations of the transition equation for the graph are demonstrated but not for the specific task of retrieval. Another GCN-based method is presented in [146], with a model that is trained without supervision using two loss functions that directly depict the diffusion process without any labeled information: a local loss that enforces smoothness (if two nodes are topologically close in the graph the similarity of their features should be high), and a global loss that enforces global order (similarities measured by different nodes from two neighborhoods should remain consistent). The learned feature space can be applied to unseen queries without a second nearest neighbour search, by a resorting to a QE-like averaging.

VIII. CHALLENGING CONDITIONS IN VISUAL PLACE RECOGNITION AND HOW TO TACKLE THEM

Although VPR is mostly treated as an image retrieval task, there are numerous challenges specific to the recognition of places that set it apart from other retrieval problems. One peculiar problem is that two places might present common elements that make them difficult to be distinguished. For example, man-made structures in urban environments are rich of recurring patterns such as building facades or fences [117]. These recurring patterns cause the phenomenon of ''visual burstiness'' [147], i.e., the presence of visual elements that are more frequent than predicted by a statistically independent model. Another problem is that the same scene can appear significantly different if viewed from different viewpoints [121], [148] or there can be little overlap between query and database images [149], making the retrieval task harder. Moreover, a scene can experience structural modifications over time, e.g., when there is a temporary

FIGURE 5. Examples of challenging conditions. Images taken from: CMU Seasons dataset [151], [152], Aachen Day/Night dataset [151], [153], [154], Nordland dataset [155].

construction site. Unlike other instance retrieval tasks, e.g., catalogue search, in visual place recognition usually there is not a single object of interest centered and well visible in the picture. Scenes can be cluttered with non informative elements, such as people or vehicles, that might distract or even occlude distinctive elements of the environment see (Fig. 5c). There are also challenges that are specific to certain environments. For instance, indoor scenes, such as campuses or hospitals, may have similarly shaped corridors and textureless areas [119]. Another major problem is the fact that the same scene can appear drastically different due to changes in environmental conditions such as illumination (day/night/shadows), weather or season (see Fig. 5a-b and Fig. 5d). The rest of this section discusses the various techniques that have been proposed to tackle these specific challenges of VPR, categorizing them in order to highlight their different goals and properties. Once again, we stress the fact that the following discussion is not a comparison of multiple methods, as they are too many and pursue different goals. For a not exhaustive comparison of few of these methods we refer the reader to [150].

A. SELECTING WHERE TO LOOK

The problem of coping with visual clutter and distractors has inspired different solutions for guiding the visual inspection pipeline to focus on the most informative parts of the images and avoiding those elements that may induce confusion. These methods not only can extract more informative and discriminative features for the localization task, but they can also make the system more efficient. The idea of selecting relevant visual information for image geo-localization is not new, and it has been investigated also with not CNN-based descriptors and/or handcrafted schemes [18], [23], [27], [117], [156], [157]. Some of the lessons learned by these studies have translated to CNN-based architectures.

1) REGION SELECTION

An approach for dealing with clutter and visual distractors is to extract regions of interest from the image, i.e., regions that contain only the elements that are most relevant for the recognition task. This idea can be naïvely implemented as a multi-scale search on the input image. Namely, patches are cropped from the image and for each of them a representation is extracted. The regional representations can then be compared for the retrieval. Even though such an implementation has shown to be more robust against scale and viewpoint variations [71], [149], [158], [159], extracting patches directly from the input image is inefficient because it requires multiple forward passes through the network. Inspired by the advances in object detection, most recent works extract the regions directly on the convolutional feature maps. Another consideration, which applies also to regions extracted on the feature maps, is that using a fixed-grid of proposals (e.g., [71], [73]) is sub-optimal. Since the fixed grid is not informed by the content of the image, the proposed regions may fail to fully contain relevant elements. Moreover, many of the regions may only cover clutter, hence they can negatively affect performance [80]. This problem, cannot be solved by simply using a finer grid because increasing the number of regions would not only improve coverage of the informative areas but also of the irrelevant ones. Additionally, this would not be a scalable solution, because increasing the number of regions would also increment the retrieval latency and the required memory [129]. Following these considerations, [80] modifies the R-MAC descriptor [73] by replacing the fixed-grid sampling with a region proposal network, akin to the one introduced in Faster R-CNN [160], and trained on the Landmarks dataset [57]. For a similar number of regions, the region proposal network yields better performance than the fixed sampling. A limitation of this proposal method is that it is trained in a supervised fashion, from a dataset with labeled regions. A similar modification was proposed for the ASMK aggregation [37] by Teichmann *et al.* [38], using a MobileNet-V2 [161] based SSD detector [162] for selecting the regions and modifying the ASMK kernel to apply regional aggregation. Another strategy for computing the regions of interest, without requiring annotated region proposals, is to directly mine them on the convolutional feature maps [163], [164]. Features at late layers tend to be sparse and representative of semantically meaningful elements such as a shape or an object [53]. Therefore, saliency regions can be extracted from these layers by clustering the activations and selecting those with highest energy. The main difference between [163] and [164], besides the aggregation method from the local features, is in the definition of the cluster: a set of non-zero spatially proximal 8-connected activations in [163], a set of neighboring activations with similar values in [164]. When multiple regions are selected for each image, an approximate linear bidirectional similarity search across the database can become prohibitive. To mitigate this problem, in [165] the k-NN search from regional descriptors is replaced by a Locality-Sensitive Hashing based method, which not only provides a significant speedup but also improves matching.

2) ATTENTION MODULES AND WEIGHTING MASKS

Attention modules are an approach to select the more relevant information from the images (see Fig. 6) that can significantly improve the performance of place retrieval [166].

FIGURE 6. Examples of images from urban environment (top) and the attention scores (bottom) generated by the AdAGeo architecture [168] which implements attention via class specific activation maps. The attention is focused on building, disregarding dynamic objects, uninformative areas (sky) and objects that are not stable over time (trees).

Differently from region proposal methods, which effectively extract portions of the image that are deemed interesting, with attention modules the image is processed as a whole but the individual features are weighted according to a relevance criterion. The weighting scheme, particularly in early works, can be established according to some heuristic. In [70] the weighting heuristic is based on the assumption that objects of interest tend to be closer to the center of the image. The CroW descriptor [76] combines spatial and per-channel weighting. The spatial weighting is based on the normalized total response across all channels, effectively boosting the response for locations in which multiple channels are active. The per-channel weighting is based on the sparsity of feature maps, effectively penalizing low-sparsity filters that are very recurrent and not discriminative. Heuristics based weighting masks can improve specific aspects of the retrieval process, but they lack flexibility. In the spirit of deep learning, these weighting masks are better learned end-to-end from the data. A learned attention module is used in [81], where it serves as a keypoint detection unit. The module learns a non-negative scoring function for each feature with a weakly supervised training, i.e., requiring only image-level labels. The same attention module, both in additive and multiplicative form, is also used in [138] to improve the localization from aerial images (remote sensing). A learned contextual reweighting network (CRN) is described in [91]. The CRN is implemented as a concatenation of multi-scale context filters followed by 1×1 convolutions, with downsampling/upsampling layers used for dimensionality consistency. Being implemented with conventional differentiable layers, the CRN can sit on top of a fully convolutional network and be trained in an unsupervised way, meaning that the training does not need explicitly annotated boxes. The effect of this contextual modulation is to produce a weighting mask based on semi-global context. Qualitatively, the mask gives positive weights to relevant structures while it penalizes repetitive lattices or not meaningful content. This contextual reweighting surpasses the predefined weighting mask used in Crow [76]. A method similar to the CRN is presented in [167]. In [168] an attention mechanism is implemented through class specific activation maps [169], which are used as score maps to weigh the features (see Fig. 6).

Selection of salient regions can be guided by context at multiple scales as demonstrated by MSCAN [92] with a two layer LSTM network. The first layer of the LSTM network generates an initial multi-scale context memory that is then fed to the second LSTM layer to produce a multi-scale aware attention. MSCAN produces very focused responses on the relevant portion of the images (e.g., buildings) but not on occlusions such as people in the foreground. Another model that uses an attention map based on multi-scale context is proposed in [170]. There, a global latent context is created for each location of the feature maps by adaptively pooling all local descriptors. The attention mask is created by fusing these context maps, thus combining the local information and the global context at multiple scales. Global image information is also used in [84] to guide the attention module with a cascaded scheme: a first attention block of 1×1 convolutions produces a global attention descriptor; a second attention block uses the global attention descriptor from the first one as content prior. In [75], a multi-scale attention map is implemented using two 1×1 convolutional layers with sigmoid activation on a multi-scale feature map. However, the results show that this multi-scale attention map may give inconsistent results in presence of lighting variations.

Another solution to produce more informative attention maps is to leverage second-order spatial information, as done in [95] using a non-local block [171]. Second order spatial information allows to generate a feature map in which local features reflect the correlations between all spatial locations, in contrast to first order features where each local feature has a limited receptive field. This method allows to learn dense local descriptors that account for the contribution of each local feature in relation to the others.

B. VIRTUAL VIEWS AND WARPING

View synthesis is an approach that has been adopted in visual place recognition to address the problem of viewpoint variations. Namely, the query or database images are replaced/augmented by artificial views that show the same scene but from a different viewpoint. In [121], view synthesis is used as a way to augment the database in order to help recognition of night-time queries against day-time database images. This use of synthesized views is motivated by the fact that the dense local descriptors used for the matching, while more robust than sparse local descriptors to illumination changes [172], suffer from limited invariance to geometric transformations (scale and viewpoint). In this case, the views are synthesized from panoramic Google Street View images and their associated coarse depth-maps. Despite containing significant visual artifacts, these artificial views yield better localization than the non-augmented database. View synthesis is used in [119] for pose verification after the initial retrieval stage. Namely, after a set of candidate poses are estimated for the query, a view for each estimate is synthesized to show how the scene would look from that pose. The synthesis

leverages a database of RGBD images that provides a dense and accurate 3D structure of the environment. The re-ranking is finally based on the count of matching and non-matching pixels between query and synthesized views.

Warping is used in the case of cross-view localization in which the query is a street-level image and the database is made of aerial images, or viceversa. Such an extreme viewpoint difference makes it impossible to directly match query and database images. One workaround is to project ground and aerial images to a common artificial view. In [173] the authors first project the 2D street view images to 3D world coordinates and then re-project them onto the aerial view image plane through the street-view depth estimates and using some simplifying assumptions. A similar use of rectified views is displayed in [174], where a street-level query is geo-localized against a geographic information system (i.e., a map from satellite imagery). In [28] virtual views are created from Google Street View images to simulate different camera tilt angles and improve matching with aerial queries.

C. SEMANTIC INFORMATION

Semantic information can be leveraged to guide the extraction of the most informative and distinctive visual elements for the retrieval task, providing more robustness to distractors and changing conditions (Fig. 7). This approach leverages a prior knowledge about the environment where the localization is considered. For example, knowing that people and vehicles carry no relevant information for the localization, one could discard or penalize the visual cues corresponding to these semantic labels. Even in the pre-CNN setting, semantic information has been used to improve image retrieval, e.g., to select correspondences between images only on areas recognized as man-made structures because they are more distinctive and stable over time [175], to remove the sky since it carries no localization information [176], or to filter out patch descriptors based on the semantic content of the patches [177]. Place categorization is instead used in [178] to inform the recognition of the location. Namely, a classification network trained on Places365 [179] is used to extract the most likely semantic attributes from a scene, thus categorizing it. The semantic category of the query is then used in the place recognition module to bias matches within the same semantic category. The method proposed in [178] actually uses sequences of images and the segmentation is used to create subsequences with coherent category both in the query and in the database. Nevertheless, the principle is applicable to the single image scenario. A similar concept is explored in [180], where the semantic classes of the objects in the query are used to filter the database images, thus reducing the search space for the image retrieval.

Dense, pixel level semantic information extracted from a semantic segmentation network is used in [128], [181] to tackle the specific challenge of localizing a vehicle along a road that was previously traversed only in the opposite direction, meaning that there is a 180° viewpoint difference between queries and database images. The visual semantic

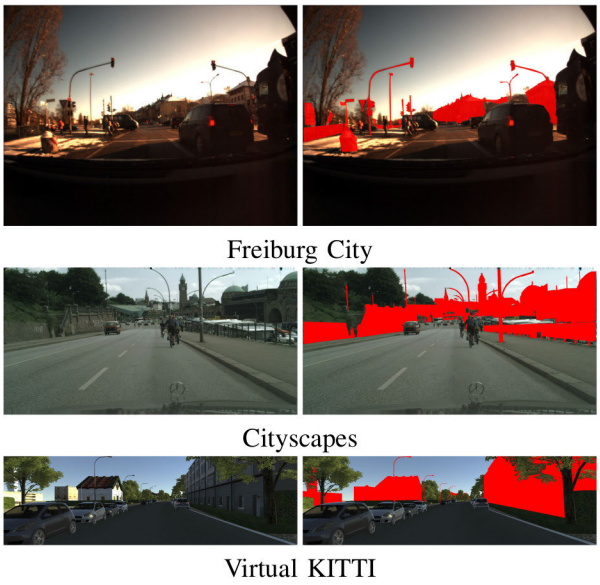

FIGURE 7. Pixel-wise semantic information can be leveraged to select only portion of the image corresponding to content that is stable across seasons, e.g., man-made structures. Image from [183], Copyright ©2017, IEEE.

information is used in two ways. First, the semantic label probability and convolutional maps of the semantic segmentation network are used to construct a descriptor called Local Semantic Tensor (LoST), which concatenates descriptors for each semantic class. Then, the maximally-activated location of the same feature maps are exploited to select semantically labeled keypoints for verifying and re-ranking the candidate matches. This use of the semantic information not only mitigates the problem of viewpoint changes, but it also boosts the recognition performance in varying weather/illumination/season condition. A similar problem and solution are considered in [182]. Here, temporal sequences of pixel-wise semantic masks are used to build graphs where the vertices are semantic blobs extracted from the masks and the edges are built based on proximity constraints. Given a query graph and a database graph (map), the place is then retrieved by matching descriptors built using a random walk approach.

Pixel-wise semantic labels are used in [183] to produce an image representation that is more robust to changes over time. For this purpose, the authors use a semantic segmentation network trained over a dataset with scenes in varying ambient conditions to extract a binary semantic mask. Pixels corresponding to stable elements (buildings, signals) are marked as discriminative and preserved, whereas dynamic objects (pedestrian, cars), uninformative content (road, sky) and objects with unstable appearance over time (trees) are marked as not discriminative and removed. Both the original and segmented image are then fed through a feature extractor and the corresponding convolutional features are aggregated to form the image representation. Hou *et al.* [184] employ a similar semantic binary mask, however they use it

not to compute an holistic image representation but rather to filter out the non-discriminative regions proposed by a landmark detector. Similarly to [183], also [185] combines appearance-based and semantic features to produce a more robust representation. The main difference is that [185] also uses semantic and appearance information to estimate a multi-modal attention module. The multi-modal attention module, informed by both appearance and semantics, helps the network to selectively focus on visual elements that are more discriminative and stable. Pixel-wise semantic labels and depth labels are combined in [97] to train a multi-task architecture with a shared feature extractor. With this strategy, the feature extractor implicitly learns to fuse geometric and semantic information, thus producing more discriminative embeddings for the retrieval task. While semantic masks can be highly informative in urban environments, the lack of content in bucolic scenes might make them ineffective. For this reason [186] proposes to use descriptors created from semantic edges, rather than the full semantic masks. However this solution is extremely vulnerable to noise in the semantic mask.

Despite the improvements in the localization performance achieved by several methods that exploit pixel-wise semantic information, the semantic classes that are used are often few and chosen by experience or inherited by other tasks. Larsson *et al.* [187] argue that such a choice of few classes is not optimal for visual place recognition because it limits the discriminative power of the learned representations. To overcome this problem they propose a fine-grained segmentation network with a high number of classes ($\approx 10^2 - 10^3$) that are learned in a self-supervised way by clustering the features extracted. The clustering, and thus the selection of classes, is updated at a fixed number of training steps. Additionally, a correspondence loss is applied to different views of the same scene to encourage the model to learn semantic representations that are robust to viewpoint and ambient changes.

Semantic information has also been used to address the place recognition problem in the case of extremely different views. Castaldo *et al.* [174] consider the case in which the query is a street-level image and the database is made of semantically annotated top-view satellite images. In this case, the matching is performed by searching on the map for a tile with a spatial layout of semantic segments (e.g., road, building, grass) which is consistent with the one in the query image. Finally, even though not strictly relevant for the visual place recognition task using only 2D images, it is noteworthy that semantic information also been shown to yield more robust results in the case of 3D-based place recognition [29]. In this case, the semantic information is exploited to generate semantically complete 3D models, from which robust 3D descriptors are extracted.

D. DEPTH INFORMATION

Depth maps are an auxiliary source of information that can be combined with appearance-based processing to leverage scene geometry in the place recognition task, hence providing robustness to visual changes. Depth information can be used to guide the process of extracting a global image representation for the image retrieval. This strategy is used in [97], where the authors propose an architecture where the encoder used to extract the image representation for the recognition task is shared with two auxiliary tasks: depth map reconstruction and segmentation mask reconstruction. Thus, the encoder learns to use the geometry in the scene (as well as the semantic content) to extract the appearance-based representation. Piasco *et al.* [188] also use depth reconstruction as an auxiliary task to help the place recognition task. Unlike [97], they use separate encoders for the two tasks (plus a decoder to generate the depth map) and combine representations obtained from these two encoders in a unique descriptor. In both [188] and [97] the training is supervised, requiring a depth map for each training image, but at inference time only the RGB image is needed.

E. ADAPTING TO DIFFERENT ENVIRONMENTAL CONDITIONS

The challenges posed to vision place recognition by changing conditions such as illumination, weather and seasons are widely acknowledged [151] and represent an open problem. A significant body of literature has investigated the shortcomings of descriptors and deep features in such difficult conditions [172], [189], [190] and the insight gained by these analyses has provided some guidance to find more robust descriptors. In this sense, it has been shown that sparse local features are not robust to appearance variations such as drastic illumination changes (day/night) [172]. An explanation for this poor performance is that keypoint detectors only consider small image regions and use low-level information that is highly affected by pixel intensities. This can lead to unstable detections under strong appearance changes [134]. These observations have led researchers to use dense local descriptors without a detection phase to address place recognition across day and night cycles [121]. An alternative is to preprocess the images with a learned photometric normalization to cope with significant illumination changes [191]. Notably, deep learned descriptors have shown better performance than hand-crafted ones in benchmarks with day/night conditions [69]. For similar reasons, dense local descriptors have also found use in indoor localization where keypoint detection is hindered not only by changes in lighting (artificial/natural) but also but also by the lack of textures [119]. Small performance gains have been recorded by using methods that introduce a selective focus on parts of the image, e.g., with regions of interest [158], [163], [164], attention modules [84], [170] or semantic guidance [128], [183]–[185]. These gains can however be attributed to the fact that these techniques help focusing on elements in the scene, such as buildings, that have a more stable appearance over time. It is also worth noting that methods that use sequences of images have shown a greater robustness to changing conditions [192]–[195].

A different solution to the problem of changing conditions has been proposed for robotics and autonomous driving, and it consists of continuously growing the database with new images captured in different conditions. Disregarding the collection costs, this idea is difficult to implement as the dimension of the database, and consequently the computational cost of the place recognition algorithm, can quickly become unmanageable. To solve this scalability problem, Doan *et al.* [196] propose a solution that consists of three elements. Firstly, a VPR algorithm based on a Hidden Markov Model that is efficient both in terms of training time and testing time. Secondly, a strategy for growing the database with the query sequences that only adds images with significant new content. Thirdly, a compression step that merges connected portions of the map, thus decreasing its size.

More recently, a few studies have explicitly targeted the cross-domain problem where the target domain (query) differs from the source domain (database) due to changes in illumination, weather or season. A method that is investigated in [197], [198] is to replace the query with a synthetic image that depicts the same scene but with the appearance of the source domain. In [197], the query image is translated to a synthetic image in the source domain by using a CycleGAN [199] that has been tailored for SURF matching by adding a loss term on the feature detector and descriptor. In particular, the authors use a generator based on UResNet [200] and train it with a two-stages procedure. In the first, unsupervised, stage the image generators are trained using a small set of unpaired images from the two domains. In the second, supervised, stage the generators are fine-tuned using pairs of pixel-aligned images from the two domains, learning certain feature transformations that might not have been captured in the first stage. A similar idea is developed in [198], albeit using a ComboGAN [201] that allows for n-domain translations. The architecture in [198] is tailored for the retrieval task by using a triple discriminator: one focused on texture, one on color, and the third one on horizontal/vertical gradients. The main difference with [197] is that there the cycle-consistency loss is used to enforce that the feature detectors/descriptors are translated properly, whereas in [198] the third discriminator emulates the process of extracting SIFT descriptors, thus inducing the creation of matching-relevant features in the translated version. A ComboGAN is used also in [202], together with a feature consistency loss, to learn domain invariant latent features for retrieval-based place recognition. GANs with cycle consistency are also used in [203] to tackle the cross-domain problem, however there the authors directly utilize the feature extracted by the first fully connected layer in the discriminator as image representation to be used for the similarity search. Rather than aligning the features of different domains, another strategy is to learn multi domain features and then separate condition dependent features from the condition invariant ones using a separation module [204].

One specific case of cross-domain place recognition is considered in [205]: the database is composed of present day RGB images and the queries are historic images from the same area. In this case the domain shift is caused not only by possible changes in the scene but also by the different technology used to take the photos. The architecture proposed to tackle this specific setting is based on a CNN feature extractor with VLAD aggregation [33] and two key elements: i) an attention module that weighs the features and residuals in the aggregation within the VLAD module, and ii) a multi-kernel maximum mean discrepancy (MK-MMD) domain adaptation loss that guides the CNN to learn a latent space where the two domains are not distinctive. Experiments show that the attention module brings only a modest improvement, whereas the boost due to the domain adaptation loss is significant. A different application of domain adaptation is proposed in [97]. There, the authors propose to use a virtual dataset to train their model which requires both depth and semantic labels. An adversarial training with an adaptation loss is then used to ensure that the latent features extracted from the virtual and real domain have similar distributions. Generative and domain adaptation approaches are combined in [168] to tackle the problem of changing conditions in place recognition. The authors show that generative and domain adaptation techniques bring orthogonal improvements to the recognition results, and their combination further boosts performance. Additionally, among the solutions discussed here, [168] is the only one that addresses domain adaptation in VPR in a few-shot setting.

F. USING 3D MODELS

Appearance based recognition of a place can be supplemented with the information of a 3D model. This information can be exploited in various ways. Firstly and most importantly, the 3D model can be used to accurately regress the pose of the camera that captured the query image with respect to a given coordinate system. Additionally, the 3D information can be used to improve some aspects of the retrieval pipeline, such as the construction of the database. The 3D model can be built directly from the database images using structure-from-motion (SfM) [151], [206], [207], as long as the images present enough overlaps and provide different viewpoints of the same scene. Since the 3D points reconstructed from SfM are usually sparse, multiview stereo algorithms [207] or densification [208] can be used to recover more dense and accurate models. Mining techniques such as zoom-in and zoom-out or sideways crawl can also be included in the SfM pipeline [208] to better capture fine details. 3D models can also be created by other sensor information, e.g., using depth information [209] or lidar measurements [29].

In visual place recognition 3D models are exploited in different ways. In [210] it is used to learn a codebook to be used for classic aggregation methods such as BoW or VLAD. The 3D model can also be used for re-ranking with geometric verification [208], or as an instrument to guide the mining of positive examples for metric learning [77], [83]. A predominant use of 3D models is to estimate the camera pose with respect to the map [28], [105], [106], [119]. First, the image

is matched using descriptors, then a precise pose is computed from 2D-3D matches using a PnP [211] solver within a RANSAC loop [212]. Under strong viewpoint changes, which are typical for instance in place recognition for aerial robots, it can be difficult to establish 2D-3D correspondences, so the 3D map can be densified using depth completion [213]. One problem with using 3D models is that they are expensive to store and maintain, and do not scale well to large environments. To overcome this problem [214] proposes to not use a large 3D model created offline, but rather to create a small local model online (SfM-on-the-fly), using the top retrieved images from the database. This solution is shown to achieve good results, although its effectiveness deteriorates in the case of weakly textured scenes. A 3D model is also used in [119] to synthesize views from the predicted camera pose, which can then be used for verification and re-ranking. It is worth noting that there are deep camera pose regression methods based on a single CNN [215]–[217] without the need of storing the 3D model for online inference, however these methods have not yet achieved an accuracy comparable to those using explicitly 3D structure. A 3D model derived from SfM is used in [134] to generate training data for D2, a network that learns to detect keypoints for local descriptors. Structure-from-motion is also used as a way to generate ground-truth sparse patch correspondences between pairs of images depicting the same scene. These correspondences can then be used to train a model that extracts sparse local descriptors [218] for spatial verification.

While all the aforementioned studies use 3D information to support visual-based recognition of places, recent studies have proposed to do the inverse. Favoured by the increasing availability of 3D sensors, several researchers are investigating to directly use 3D-3D matching for place recognition. In this case the images are used as an auxiliary information, e.g., to extract semantic information to be fused with the 3D pointcloud [29].

IX. VISUAL PLACE RECOGNITION WITH AERIAL IMAGES

The task of visual place recognition is predominantly studied in the context of images captured from the street level, however the availability of satellite imagery and the diffusion of camera-equipped aerial robots has led to new and specific developments. On one hand, aerial images allow to obtain a bigger variety of viewpoints as well as wider views of an area. On the other, they introduce new challenges, such as drastic viewpoint variations and a lack of distinctive visual details. The rest of this section discusses few scenarios of VPR with aerial images.

A. REMOTE SENSING

In remote sensing image retrieval, like in classical VPR, the task is to identify a query image location by retrieving similar images from a database. However, the images are taken from a downfacing camera onboard an aircraft flying at high altitude or from a satellite. Therefore, the images depict large geographic areas, with objects that may have significantly different scales. Moreover, elements that are very distinctive from the street level, such as buildings, may be less informative when looked from a great distance above. On the other hand, visual elements that are not particularly informative from the street level, such as roads, are important for remote sensing.

Despite these differences, methods that are used in VPR have also been applied to remote sensing place recognition. In [219] the authors use an approach based on the bag-of-words framework. First, the images are divided in patches using different schemes, namely uniform grid and superpixel. Then, the image representation is constructed by stacking together the latent features extracted by feeding each patch through the encoding part of a deep convolutional autoencoder. Finally, the bag-of-words is generated using these representations. Rather than using patches, which is computationally expensive, [138] proposes to use the DELF architecture [81] to extract attentive local features that are then combined via the VLAD aggregator. Additionally, since geometric verification is difficult to apply to remote sensing imagery, the authors use a query expansion based on memory vectors [220] to improve the retrieval results.

B. CROSS-VIEW GEO-LOCALIZATION

Another use-case for aerial images in place recognition is cross-view geo-localization. In this setting the query is taken from the street level whereas the images in the database are aerial views (or vice versa). Lin *et al.* [173] consider the specific case in which the aerial images in the database are taken at approximately 45° angle and the database images are taken from Google Street View together with a coarse depth map. This information, together with the assumption of an orthographic camera model, allows to reproject the street level images onto the aerial plane and establish ground-aerial matches. These matches are then used as examples to train a CNN-based feature extractor with a contrastive loss. The idea of cross-view training is also explored in [221]. There the authors propose to train a CNN to extract the FC representation of aerial images by using an ℓ_2 loss function that aligns these representations with those extracted from a pre-trained model for the corresponding ground images. The cross-view scenario is specialized in [174] for the case of geographic information system images endowed with a semantic map. Similarly to [173], a reprojection is used to rectify the street level image, however the projection is applied to a semantically segmented copy of the query. The cross-view matching is then cast as a search for consistent spatial layouts of the semantically labeled regions.

C. MICRO AERIAL ROBOTS

When the aerial images are captured from a front facing camera onboard a multicolor micro aerial vehicle flying at low height, the geographic area covered is not as wide as in the case of remote sensing imagery. However, the large roll-pitch rotations that are typical in the motion of these robots cause drastic viewpoint changes (Fig. 8). This prob-

FIGURE 8. Example of viewpoint variations due to the drone pitch/roll angles during flight. Images from the camera onboard a drone flying in Zurich. Dataset presented in [28].

lem is particularly pronounced in the cross-view setting in which the database images are taken at the ground-level. Majdik *et al.* [28] suggest using virtual views that simulate different tilt angles as well as a verification step. The problem of drastic viewpoint variations is also present in the aerial-to-aerial setting and it is demonstrated in [148], [213], where the authors propose a new dataset specifically for VPR for aerial robots. Along the same lines, Zaffar *et al.* [222] show that methods that perform very well on ground-level localization show a significant degradation in performance when applied to the aerial case with 6 DoF viewpoint changes. The viewpoint variations encountered in this setting can be partially mitigated by the use of geometric verification [148].

X. VISUAL PLACE RECOGNITION IN ROBOTICS APPLICATIONS

Visual place recognition is a fundamental component in the navigation stack of robotic systems, being used for example for loop closure detection in GPS denied environments. Although the techniques from computer vision research have carried over to this scenario, the unique characteristic implicit in the robotic application have led to specific developments. The most significant peculiarity of this problem setting is that place recognition is intended as a continuous task that processes streams of observations and that can leverage a knowledge of the motion of the robot available through egomotion estimation and motion models. Additionally, robots are often equipped with different sensor technologies other than vision that can be used for VPR, such as 3D lidars [223], [224] or range sensors [225]. Lastly, the place recognition problem is often combined with visual based localization, where the goal is to regress the 6 DoF pose of the robot w.r.t. a known map.

A. MAPS

The task of recognizing a location from visual information requires prior knowledge of all the places of interest. Since in robotic navigation the observations are collected continuously, consecutive images from the video stream of the camera are connected by spatio-temporal constraints. In this setting, the prior knowledge of the world is thus naturally organized as a topological map [3], where nodes represent places with associated observations of the world (images) and edges indicate transitions between places. These transitions allow to naturally describe motion constraints that are posed by the structure of the environment and by the robot itself. For example, the navigation of autonomous cars in a urban environment is constrained not only by their mechanical structure but also by the roads and the traffic rules. In this context, the transition between two spatially near places might not be allowed because of a "no entry" traffic sign. Topological maps, combined with the continuous localization of the robot, can effectively speed up matching because the location prior can be used to limit the search to a sub-graph with an adaptive window approach [194]. A similar idea is proposed in [226], where the database images for an indoor environment are partitioned in subspaces based on spatial rules. This partitioning stems from the observation that there is a relation between image similarity and distance. Moreover, the partitioning rules used by the authors also leverage the natural subdivision in spaces that is present in indoor environments (rooms, corridors). The partitioning of the database is used to improve the computational efficiency of the VPR system by limiting, if possible, its search space to the last visited subspace.

Topological maps might also be augmented with metric information assigned to nodes or edges [3]. Such metric information can be exploited to guide the place recognition process in conjunction with motion models and odometry measurements [182], [194]. The addition of odometry information has been demonstrated to improve place recognition performance [176]. Moreover, the information stored in the nodes is not necessarily limited to the appearance of the corresponding places, i.e., images, but can also contain semantic information. For example, places might be described using scene text such as billboards, shop names, road signs, etcetera [194], [227]. Hong *et al.* [194] use a topological-metric map where each node represents an image with an associated a descriptor of the scene text present therein. The similarity matching is then implemented as a combination of semantic information (Levenshtein distance between text strings) and appearance information (IoU between the text bounding boxes). The semantic information that can be stored in the map is not limited to textual descriptors. Gawel *et al.* [182] build graphs using blobs of connected regions extracted from semantically segmented images, i.e., regions with the same pixel-wise class label.

One question that arises in this setting is how to efficiently construct and expand the map as the robot navigates the world. This question is critical for long-term autonomy and scalability to wide geographical regions. In particular, when a new sequence of images is acquired it is necessary to add only new and relevant information to the map and to establish connections with previously visited places. One approach used to selectively grow the map is to link the acquisitions of new observations to localization failures [25]. This strategy follows the idea that the localization failure is an indication that the prior knowledge available is not sufficient for recognizing the current place. Churchill and Newman [25] apply this solution and introduce four different methods to connect different acquisition with new edges. New nodes can also be added to the map based on metric information, i.e., when the

camera moves a certain distance from an existing node [194]. Metric information can also be used to combine newly created graphs with previous acquisition, by merging close vertices into a single location [182]. Doan *et al.* [196] have recently proposed a strategy to efficiently expand a map. This strategy is based on two different optimizations: i) adding images that provide only new information based on the localization belief (culling); ii) merging nodes that refer to the same place but were visited in different occasions (compression).

B. CONTINUOUS PLACE RECOGNITION AND LOCALIZATION

In robotics, VPR is intended as a continuous task in which the algorithms can access streams of video frames, rather than single images, as well as odometry measurements and hypotheses about the motion model. The availability of such information can be leveraged by ad-hoc solutions to achieve a more robust and accurate place recognition.

1) STOCHASTIC MODELS

Stochastic models are used to generate a belief distribution about the location of the current observation (image) captured by the robot with respect to the known map. This generation process is performed recursively, exploiting previous estimates, egomotion measurements and motion models to provide the prior localization belief. An important milestone in this line of research is FAB-MAP [19], [21], an appearance-only algorithm for place recognition that extends the bag-of-words framework with a recursive Bayes estimation. FAB-MAP implements Bayes recursion by approximating the likelihood that the observation was originated from an unseen place. This approximation can cause perceptual aliasing, however this problem is mitigated by the introduction of a smoothing operator. Although FAB-MAP can work as a pure image retrieval process by always assuming a uniform location prior, using even a simple motion model to leverage the prior estimate improves performance. Hidden Markov Models (HMM) are also used extensively for visual place recognition in robotics to exploit the temporal order of the captured images and the high correlation between time and place due to motion constraints [193], [196]. In [193], a HMM is used to combine image representations produced by four methods: two hand-crafted representations and two types of deep learned representations. For this purpose, the authors introduce a multi-process fusion algorithm that compares the matching performance of each representation and ranks them by voting. Then, in the application of the Viterbi algorithm the emission matrix is re-computed at each element of the sequence using only the top voted representation for that particular image.

2) SIMILARITY MATRIX

Another family of methods for VPR in robotics is based on the use of similarity matrices. Given a sequence of frames (query), a similarity matrix is built by comparing individually each frame of the query (rows) with each image in the database sequences (columns) (see Fig. 9). The similarity matrix is then used to estimate the most likely trajectory followed with respect to the known map. This kind of approach was popularized by SeqSLAM [45], which used image differences with local contrast enhancement to create the similarity matrix. SeqSLAM estimates the current location by searching on the matrix for the best fitting trajectory along the map, i.e., the trajectory that provides the best matching score, possibly given motion constraints. One weakness of SeqSLAM is the susceptibility to perceptual aliasing which can be counteracted by using long sequences of images. The approach of SeqSLAM has been expanded by subsequent works, e.g., using a Hidden Markov Model to allow searching for non-linear trajectories [228], filtering out the sky from the images [176], querying frames by their relative distance to allow for traversals at different speeds [176], using learned features instead of handcrafted ones [229], using semantic-based descriptors to create the similarity matrix [181]. The similarity matrix can also be used in conjunction with more complex trajectory search methods to achieve more robust and accurate localization results. Naseer *et al.* [192] use the similarity matrix to build a flow network that associates the frames in the query and in the database. The nodes in the network represent matches between images on the edges are associated a cost based on the similarity score. The trajectory search is thus interpreted as finding the minimum cost flow through the network. Not only this method allows to find trajectories traversed at different speeds and with stops, but using special nodes it can also manage trajectories with non-matching frames.

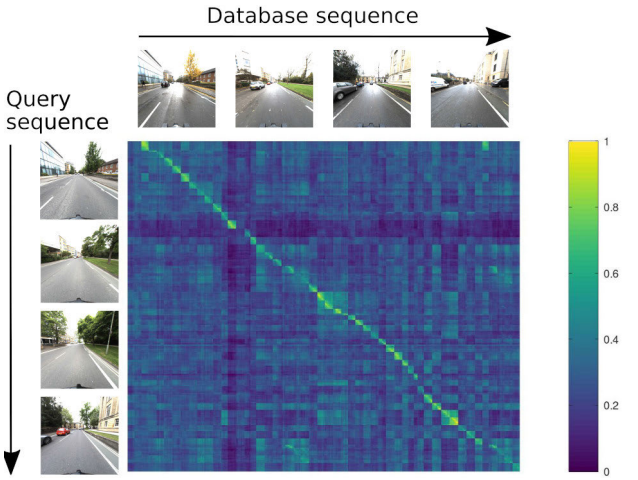

FIGURE 9. Example of similarity matrix computed from sequences of images. The rows refer to the query frames, whereas the columns refer to the database frames. The sequences are taken from Oxford RoboCar and the image representations are computed using a pretrained ResNet-101 with a GeM pooling.

3) SEQUENCE REPRESENTATIONS

The images in the query sequence can also be processed all at once to create a combined representation that implicitly contains the temporal information among frames and that can

be matched to other sequences [182], [203], [230]. In [230] the authors test three different methods to combine deep learned representations, i.e., concatenation, fusion via a FC layer and recurrent representations built via LSTMs. Interestingly, naïve grouping works better on standard sequences, perhaps because it is equivalent to imposing a coherence check. Although such a grouping can become more distinctive as the length of the sequence grows [203], the representation size also increases with the number of combined frames thus imposing a trade-off. Fusion and recurrent representations work better when the sequences are altered (different speed, reverse traversal) as they are able to learn complex relations among the frames. Sequence representations are created in [182] using a two-stage strategy. First, a semantic graph is created from a sequence of images by selecting connected semantic components. Then, for every node of the graph a random walk is repeated and the sequences of visited semantic nodes are stored as a matrix. In this case, the similarity between pairs of sequences is scored by the number of matching random walks in their representations.

4) BIOLOGICALLY INSPIRED METHODS

Biologically inspired methods mimic the cognitive processes of animals with relatively small brains, such as insects and rodents, in order to create efficient and resource limited algorithms. RatSLAM [231] is a notable example of such biologically inspired methods. RatSLAM is inspired by computational models of the hippocampus of rodents and it represents the pose of the robot by the activity in a continuous attractor network (CAN) that integrates odometry with visual landmark sensing. In [232] the authors present a place recognition approach that is inspired by a recently discovered type of spatial encoding cell, called grid cell, that is found within the mammalian brain and whose firing structure reveals the characteristics of multiple discrete and overlapping scales. The proposed approach mimics the discrete multi-scale encoding patterns of grid cells by utilizing multiple place recognition channels, each of which adaptively selects spatial scales based on environmental similarity. NeuroSLAM [233] employs multi-dimensional CANs to represent a multilayered head direction cell model and a 3D grid cell model. These models are used to perform 3D SLAM with a robot whose state is given by a 3D position and heading (yaw angle). Chancan *et al.* [195] propose a hybrid system that concatenates a compact and sparse two layer neural network inspired by the brain structure of fruit flies with a one-dimensional CAN that encodes the places. The first part of this architecture, the FlyNet network, imitates how the fruit fly brain assigns similar activity patterns to similar odors. The 1D CAN filters temporal information, with units being inhibited/excited by movement. Despite its extremely compact format, this hybrid solution manages to achieve competitive results in several place recognition scenarios and it is even shown to surpass more complex algorithmic methods in the day/night domain shift case.

C. MULTI-TASK ARCHITECTURES

Visual place recognition is just one of several tasks that a robot needs to perform while navigating. In this context, the performance of place recognition can be improved by leveraging the information extracted by other related tasks, or viceversa. For example, VLocNet++ [234] processes incoming images with three streams: one for visual odometry, one for global pose regression and the last one for semantic segmentation. The visual odometry and pose regression streams employ hybrid hard parameter sharing up to the third residual block. This influences the pose regression network to integrate motion specific features. Semantic features are also fused at the fourth residual block of the pose regression network via an adaptive fusion module. This allows the pose regression to leverage also semantic information. While VLocNet++ is an architecture with parallel tasks, DeLS-3D [235] serializes them. First, the pose is estimated using a regression network that is fed the image stream as well as the semantic mask obtained from a semantic 3D map and the coarse pose an inertial navigation system. A multi-layer RNN follows the pose regression to include temporal information. Finally, the estimated pose is used to render a precise semantic mask from the 3D map which is then fed into a segmentation CNN together with the RGB image. A different take on multi-task architectures is to have a single primary task, e.g., place recognition, assisted by other auxiliary tasks. This strategy is used in [97] with an architecture that extracts multi-scale features for the place retrieval task using an encoder that is shared with other two tasks: semantic segmentation and depth estimation, each one using a specialized map generator. This solution implicitly fuses geometric and semantic information in the features extracted for place recognition. Another example of multi-task architecture, in the domain of assistive technologies, is demonstrated in [2], which introduces a model with a single backbone and two heads, one for VPR and the other for scene recognition.

XI. VISUAL GEO-LOCALIZATION AS CLASSIFICATION

Although visual place recognition from a single image has been predominantly formulated as an instance retrieval task, few recent works have proposed to cast this problem as a classification task. Note that this is different from what discussed in Sec. V-A, where the classification task is only used to train a model to generate representations for the retrieval pipeline (see Fig. 1). In the alternative formulation discussed here, it is the classification task itself that predicts the place of an image, without any retrieval. This idea stems not only from the remarkable results that deep classifiers have achieved on large-scale tasks, but also from the observation that humans can estimate the location of a photograph without having to perform instance level or landmark recognition. This is particularly interesting when trying to achieve a global scale localization, in which case category level information can help [26].

The pioneering study from Weyand *et al.* [236] is the first one to explicitly formulate visual geolocalization on a global scale as a classification problem. In this setting, the surface of the earth (or just the area of interest) is divided in non overlapping cells. Each cell, corresponds to a class for the classification problem. A CNN with a Softmax output layer (PlaNet) computes a discrete probability distribution assigning a confidence to all the cells. The partition in cells is performed using an adaptive subdivision, so that each cell is recursively divided if it contains more than a certain number of images. Finally, cells with too few samples are discarded. Such a subdivision allows to have balanced classes, while assigning more classes (or equivalently, more of the network's parameter space) to areas with higher density of images. A shortcoming of this solution is that the size of the regions determines the maximum accuracy achievable. It would be tempting to use a finer subdivision in cells to make the classification more accurate, however this does not work too well because: i) if the cells are very fine the number of training examples per class will reduce; ii) with the increase in the number of classes the number of parameters of the classifier grows substantially, thus leading to problems with scalability and generalization. Indeed, while in some instances increasing the number of cells may increase accuracy, in others it can worsen the results [237] (see Fig. 10). To overcome this problem, [238] proposes a combinatorial partitioning in cells. Namely, different coarse partitions of earth's surface are generated and then overlapped to create finer regions. With this strategy it is possible to use a single backbone classifier with different fully connected layers per coarse partition, resulting in an architecture that has far fewer parameters than a single classifier network trained for many more classes. At inference time, the subregions overlapped by multiple class sets are given cumulative scores from multiple classifiers, with a normalization that accounts for the number of cells per class. Experiments with this architecture show that the classifier is able to correctly locate images from a wide variety of environment types, not just urban pictures. However, the performance drops critically as the accuracy required increases. This can be explained not only by the limitation of the training dataset, but also because the discretization used to create the classes is naturally lossy. The performance of the classification-based place recognition can be boosted by combining it with scene recognition [239]. In this approach, a first network classifies the category of the scene and, based on the label, forwards the image to a classifier specialized for that category. The main improvement comes from the fact that the specialized classifiers can learn more specific features for their respective domain, however this solution is not easily scalable.

So far, only one study has attempted to compare the performance of geolocalization using image retrieval and classification approaches [237]. Even though this comparison is not extensive and it does not include the latest architectures, it does offer some insight. The authors observe that image retrieval performs generally better at finer scales than the classification method, however both have some shortcomings. Image retrieval requires a database that provides images with significant overlap with the query. Moreover, it may be difficult to create a retrieval solution that generalizes to different environment types (e.g., urban cities and naturalistic scenes). On the other hand, the formulation as a classification task can provide a more general localization solution, but it suffers from the lossy partitioning in classes.

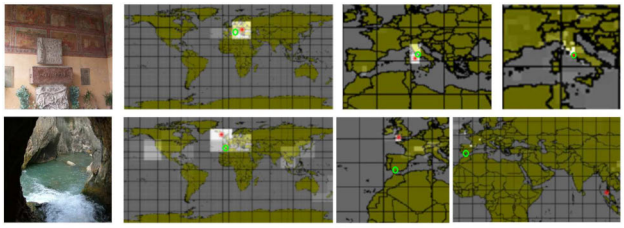

FIGURE 10. Example that shows the effect of different partitioning schemes on the localization result with a classification formulation. The two rows show two different examples, where the picture on the left is the query and the charts on the right show the predicted places for increasing number of cells. The red point indicates the prediction, the green point indicates the ground truth. Image from [237], Copyright ©2017, IEEE.

XII. DATASETS AND EVALUATION

A. DATASETS

The datasets that are openly available and used for visual place recognition focus on different use cases or problems, and therefore have substantial differences from one another. A first broad categorization of these datasets comes from the distinction between robotics and non-robotics datasets. Robotics datasets [21], [45], [151], [152], [214], [240], [241], [247]–[249], [249] are typically created by recording videos from cameras mounted on a vehicle (e.g., car) or a smaller robot. The data is thus available in sequences, with the temporal coherence among successive frames that can be leveraged to formulate motion hypotheses. Moreover, the point of view is consistently at the street level, without big changes in the vertical orientation. Non robotics datasets typically are not collected as sequences of frames [22], [23], [38], [69], [80], [81], [116]–[118], [122], [125], [190], [215], [242], [244], [246], [251]. In many cases, they are created by collections of online images, with variable viewpoints and resolutions. Since such collections are quite noisy and may include mislabeled images, it has been reported that an automatic cleaning stage can be critical to improve retrieval results [80], [88], [136]. A source of images that is closer to the robotics domain is Google StreetView. This is a collection of panoramic images taken from a vehicle (street level) but not organized as time-coherent sequences. Several datasets use images from StreetView, divided in perspective images, to create maps of cities [30], [69], [114], [117], [121], [243]. Another distinction among these datasets is the way they encode places. Some datasets focus on recognizing famous landmarks or discretely sampled locations, so they encode places by labels [22], [116], [118], [122], [190].

TABLE 1. Summary of commonly used datasets in VPR. Among the changing conditions, D/N stands for Day/Night, W stands for Weather, and S stands for Season. The column denoted as 3D indicates if the dataset includes 3D models.

Dataset	Date	Scene	Scale	# Images	Changing Conditions			3D	Place
					D/N	W	S		
Oxford [116]	2007	Urban	City	∼5k					Label
Paris [122]	2008	Urban	City	∼6k					Label
Holidays [118]	2008	Outdoor	World	∼2k					Label
Eynsham [21]	2009	Urban	City	∼70k					GPS
St. Lucia [240], [241]	2010	Urban	City	∼66k					GPS
European Cities 50k [22]	2010	Urban	Continent	∼50k					Label
Geotagged StreetView [23]	2010	Urban	City	∼17k					GPS
Rome 16k [242]	2010	Urban	City	∼16k				✓	Pose
Dubrovnik 6k [242]	2010	Urban	City	∼6.8k				✓	Pose
San Francisco [243]	2011	Urban	City	∼1.06M					GPS
Alderley [45]	2012	Urban	City	∼31k	✓	✓			GPS
7 Scenes [244]	2013	Indoor	Building	∼43k				✓	Pose
Nordland [155]	2013	Outdoor	Region	∼143k			✓		GPS
Google StreetView 62k [114]	2014	Urban	City	∼62k					GPS
Freiburg Across Seasons [192], [245]	2014	Urban	City	∼43k			✓		GPS
Cambridge Landmarks [215]	2015	Urban	City	∼10.8k				✓	Pose
Paris500k [246]	2015	Urban	City	∼504k					Label
Pittsburgh [117]	2015	Urban	City	∼278k					GPS
Landmarks-full [80], [125]	2016	Urban	World	∼192k					Label
NCLT [247]	2016	Outdoor + Indoor	Campus	∼3.8M		✓	✓	✓	Pose
Oxford Robotcar [248]	2017	Urban	City	∼20M	✓	✓	✓		GPS
SPED [190]	2017	Outdoor	World	∼1.3M	✓	✓	✓		Label
Google-Landmarks [38], [81]	2017	Outdoor	World	∼1.2M					GPS
$\mathcal{R}$Oxford [129]	2018	Urban	City	∼5k					Label
$\mathcal{R}$Paris [129]	2018	Urban	City	∼6k					Label
Tokyo 24/7 [121]	2018	Urban	City	∼ 2.8M	✓				GPS
Aachen Day/Night [151], [153], [154]	2018	Urban	City	∼7.6k	✓			✓	Pose
RobotCar Seasons [151]	2018	Urban	City	∼31k	✓	✓	✓	✓	Pose
CMU Seasons [151], [152]	2018	Urban	City	∼116k	✓	✓	✓	✓	Pose
TokyoTM [69]	2018	Urban	City	∼190k	✓				GPS
InLoc Dataset [119], [209]	2018	Indoor	Building	∼10k				✓	Pose
TB Places v2 [249], [250]	2019	Garden	City	∼59k					Label
San Francisco Revisited [214]	2019	Urban	City	∼790k				✓	Pose
WorldCities [30]	2019	Urban	City	∼300k					GPS
Google-Landmarks v2 [251]	2020	Outdoor + Indoor	World	∼4.2M					GPS
Mapillary SLS [252]	2020	Urban	World	∼1.68M	✓	✓	✓		GPS

Other datasets use the GPS information tagged on the images as the place information [69], [121]. This is particularly useful when the database densely covers an area rather than being a sparse list of landmarks. Finally, few datasets encode places with 6 DoF camera poses [151], [153], [154]. An additional aspect that sets apart the datasets is the kind of environment they consider. The majority of databases are focused on urban environments, being this the most relevant use case for many applications. However, a few datasets consider indoor environments [119], [251] and non urban areas [81], [118], [190], [249]–[251].

The available datasets have also evolved over time, in order to better represent the current problems to be solved in VPR and provide more challenging benchmarks. For example, the Oxford [116] and Paris [122] datasets that served as the main VPR benchmarks for several years have been recently revised, not only to correct inaccuracies but also to introduce new protocols of increasing difficulty [129]. The Google Landmark dataset [81] has also been expanded with a second version [251], not only to increase its size but also to ensure that the images are stable and are not be removed from their sources. Some datasets have also been revisited to add extra information, such as manually annotated boxes (Google Landmark v1 [38]) or 3D models created using structure from motion (CMU Seasons [151]). Most recent datasets also provide images from different ambient conditions (day/night cycles, different weather/seasons), because dealing with such variations is an open problem in VPR [69], [121], [151], [151]–[155], [155], [190], [248], [252]. Table 1 summarizes the datasets that are commonly used in visual place recognition.

The significant efforts recently poured into autonomous driving research has led to the release of several datasets that, although not directly aimed at visual place recognition, could be potentially adapted and used for that purpose [254]–[260]. In fact, these datasets offer long sequences of data recordings that generally include not only multiple camera streams, but also egomotion measurements, lidar scans and sometimes depth masks and semantic segmentation. However, to be used for visual place recognition these recordings need to be pre-processed, filtering the video streams and associating ''place label'' to the frames.

TABLE 2. Summary of datasets used for VPR with aerial robots. Among the changing conditions, D/N stands for Day/Night, W stands for Weather, and S stands for Season. The column denoted as 3D indicates if the dataset includes 3D models.

Dataset	Date	Scene	Scale	# Images	Changing Conditions			3D	Place
					D/N	W	S		
Old City [253]	2017	Urban	City	∼12k					GPS
Clausius Street [148]	2018	Urban	City	∼17k					GPS
L'Agout and Corvin [213]	2019	Outdoor	Single Landmark	∼5k					Pose

This could be done by leveraging the data collected from the vehicle inertial navigation system, although this typically operates at a different frequency than the cameras. Additionally, there are several virtual datasets aimed at autonomous driving that manage to generate pseudo-realistic images from a variety of scenarios, while at the same time providing exact annotations for semantic segmentation and depth masks [261]–[265]. Most of these virtual datasets do not associate a pose or coordinate to the images yet, but if geotagged images were to be added they could become a valuable instrument to study VPR. Lastly, there are also datasets specialized for visual place recognition with aerial robots [28], [148], [213], [253]. These datasets emphasize the problem of viewpoint changes, which is extremely common for aerial robots for two reasons: i) multi-rotor drones tilt significantly during flight, and ii) they can fly at different heights. Table 2 summarizes the datasets that are available to study VPR with aerial robots.

B. EVALUATION OF RESULTS

Evaluating the performance of a visual place recognition system requires first to define when a query is correctly localized. This definition changes depending on how a place is identified. For localization of landmarks identified by a label, the place is considered recognized if the label retrieved for the query matches the ground truth [117]. In the case of places identified by GPS coordinates, a query image is deemed correctly localized if the retrieved image is within a certain distance from the ground truth position [69], [121]. Finally, if the place is identified by a pose, the correctness of the retrieval is based on a maximum error on position and orientation with respect to the ground truth [119]. The latter two definitions give the flexibility to set the error threshold, adapting it to the use-case. For example, the GPS error may be set differently for recognition on street level or city level [238].

Using these definitions, various metrics are used to assess the performance of the recognition system. The most common metric applied to both retrieval-based and classification-based methods is the fraction of correctly recognized queries. This metric is indicated with different names, such as accuracy [238] or recall (in this context with a slight different meaning than in pure image retrieval) [69], [121]. Another quantity of interest in the metric is the number of hypotheses that are considered to verify a query, i.e., the number of top ranked retrieved images or most likely classes. The parameter is indicated in the metric with the notation @N, i.e., recall@N. Methods can be compared for specific values of N or by considering full curves over a range of N.

For datasets with a constant number of positive database images for each query, such as Oxford [116] and Paris [122], performance is assessed using the mean average precision (mAP) [83]. In the classification formulation the retrieval is evaluated by the mean average precision (mAP@N), i.e., sorting the top N retrieved images in order of relevance and averaging the AP of the individual queries. A modified version of the mAP that is similar to the μAP [266] is considered in [81] to account for distractors among the set of queries. Full precision-recall curves and the corresponding "area under curve" (AUC) are also used for evaluation [81], [150].

XIII. DISCUSSION AND FUTURE DIRECTIONS OF RESEARCH

The main proposition of this document is to present a comprehensive overview on visual place recognition, breaking down this topic in its multiple facets. This overview is built upon over 250 research items published by different scientific communities, i.e., computer vision, machine learning and robotics. By categorizing and analyzing all these documents, we tried to identify the research trends in VPR and to emphasize them. In this final section we summarize the contents of the survey with a short discussion. Once again, we stress that a quantitative comparison would be impossible given the breadth and diversity of methods and aspects covered by the survey. Afterwards, we elaborate on possible future directions of research for this field.

A. DISCUSSION

1) IMAGE REPRESENTATIONS FOR PLACE RETRIEVAL

In Secs. II to V we touched upon various aspects concerning image representations for place retrieval. This has been a central focus of research in recent years, with convolutional based representations becoming the state-of-the-art. From an architectural perspective, excellent results have been achieved both with aggregation and pooling schemes, but we observe that most recent studies are leaning towards pooling schemes. Apart from their simplicity, these solutions are shown to produce more effective compact representations than aggregation methods. Another important aspect to be considered is the method used to train the representation generator. We observe that researchers in this field are mostly using contrastive or triplet metric learning losses. Nevertheless, the mining step involved with these methods can become

a weakness moving forward to large scale databases, due to the complexity in selecting hard examples and the computational overhead it adds to the training. The recent application of listwise loss, which requires no mining, appears promising but it should be tested more extensively, particularly on large databases. Novel ideas, such as including second order appearance information or geometric/semantic information in the metric learning process, show that this is an area where there can be further improvements.

2) RETRIEVAL POST-PROCESSING

Although most modern research in VPR is centered around learning better representations for the similarity search stage, the post-processing stage is equally important. In Sec. VII we have mentioned various refinement methods that can boost the performance of a place recognition system. For each of these methods we have discussed strengths and limitations, with regards to the computational effort required, how effectively the information in the database is used and applicability to unseen queries. The different families of methods discussed are not mutually exclusive, but the selection of the most appropriate one depends heavily on the specific problem setting and requirements. We observe that most of the refinement methods used nowadays are based on well-established principles. Nevertheless, there are some notable advances to be acknowledged. Firstly, the promising application of diffusion techniques. Secondly, there are some attempts to implement deep learning solutions also at this stage, e.g., for the computation of local feature descriptors for geometric matching and for the implementation of diffusion processes via graph convolutional networks.

3) VISUAL PLACE RECOGNITION IS NOT JUST ANOTHER RETRIEVAL PROBLEM

In the survey we observed that visual place recognition, albeit viewed as a retrieval task, is very different from other image retrieval problems. Section VIII paints a picture of the numerous challenges that make visual place recognition a unique and very difficult problem, and it identifies several research trends that have emerged as a consequence of these challenges. From these trends we can draw some observations:

- The selection of salient areas of the image is particularly important in place recognition, helping not only with occlusions and distractors but also with the identification of the most stable elements across seasons. Among these methods, attention maps stand out not only because they do not need a separate supervised training, but also because they allow to modulate how much focus should be given to different elements.
- The recognition of a place based on appearance can be greatly improved by exploiting also semantic and geometric information.
- The variation of viewpoints can greatly affect the recognition of places, but the integration of view synthesis or warping techniques in the retrieval pipeline can mitigate this problem.

4) VISUAL PLACE RECOGNITION IN ROBOTICS

Robotics is a major domain of application for visual place recognition. The specific challenges and requirements in robotics have led to several key developments in visual place recognition (see Sec. X). Many of these developments revolve around the continuous nature of robotics navigation. In particular, the availability of temporal and topological information in robotics applications can be leveraged in the retrieval process. Most of the solutions developed with this goal combine a multi-frame retrieval process with trajectory exploration strategies. However, there are few studies that try to embed the temporal information directly in the place recognition process, e.g., using sequence descriptors or biologically inspired methods. Another emerging theme in this context is the integration of VPR in multi-task architectures. These architectures may not only be aimed at realizing more efficient perception stacks, but they can also combine the information extracted from multiple sensor input to perform better in different tasks.

5) GEO-LOCALIZATION BY CLASSIFICATION

In Sec. XI we discussed an alternative formulation of VPR as a classification task. This formulation has interesting characteristics and it can provide a more effective solution for the coarse recognition of a place given a large and sparse database of images. The different strengths and limitations of image retrieval and classification methods for place recognition also raise some interest in hybrid solutions that could combine the accuracy of retrieval based methods with the ability to generalize and the resilience to viewpoint changes from classification methods.

B. FUTURE DIRECTIONS

Based on the analysis provided in this survey we can formulate some considerations about the future directions of research in VPR. One observation is that scalability is a fundamental problem to be solved to make VPR viable in real world applications. As discussed in the survey, there are several aspects of the retrieval task that have repercussions on scalability and where researchers can investigate new solutions. Firstly, there is a need to investigate compact representation that can be discriminative for large scale problems. Moreover, using classic ranking losses may be infeasible for massive databases, given the problems involved with mining examples. From this point of view, it could be interesting to revisit representations generated from classification tasks. Indeed, in the survey we observe that these representations were quickly abandoned in favour of generators based on a contrastive or a triplet loss. However, that trend was motivated by the results achieved on small scale problems. For large scale problems, using a classification model may be beneficial to obtain compact representations without the need to mine examples. Additionally, even similarity search on a massive database can become prohibitive. In this sense, methods to reduce the search space in the database may

be necessary. Lastly, we advocate the necessity to investigate scalability both with sparse databases (images collected on a large area, without overlaps) and dense databases (where a possibly limited area is covered by many images with overlaps), because these two scenarios present different problems.

Another open problem in VPR is long term reliability (or long term autonomy), which is crucial to effectively deploy these systems in the wild. We recognize that there are two aspects of this problem that have barely received any attention. Firstly, these systems must be able to generalize or adapt to different domains. As discussed in the survey, there are many solutions that have been shown to help in this sense (i.e., saliency selection, use of multi-modal information) however these methods are either not specifically developed for this purpose or based on heuristics. There are only few studies that explicitly tackle domain adaptation in VPR and only one that does it in a few-shot setting. Secondly, these systems must be able to acquire new knowledge after they are deployed. To the best of our knowledge, the question of incremental learning has not been addressed in this context.

At the beginning of this manuscript we mentioned that the increasing interest in VPR largely derives from the many potential use-cases, from apps on a smartphone to an autonomous driving car. So far VPR has only been studied in the scenario of a single system, however both smartphones and autonomous cars are naturally networks of distributed systems. From this perspective, it becomes interesting to frame VPR as a task across multiple devices, where the experience of one could help increase the knowledge the others. There are several interesting challenges in such a scenario, from the sensitivity of data, to the limited onboard resources available on such devices. Therefore, we think that VPR makes for a fascinating problem to be studied from the perspective of edge computing and federated learning.

Currently, one prominent direction of research in VPR concerns the development of multi-modal solutions that also leverage semantic and geometric information to help in the recognition of places. However, we find that the lack of datasets built for VPR and that include also these other input modalities is a limiting factor. Therefore, we believe that to further advance the research in this direction it is urgently required to create new multi-modal datasets for VPR or expand existing ones in this direction. Along this line, we also think that autonomous driving platforms could offer the opportunity to create VPR datasets with heterogeneous sensor inputs (e.g., video streams and lidar pointclouds), which could be used to further expand the visual place recognition problem.

REFERENCES

[1] M. Chancan and M. Milford, "CityLearn: Diverse real-world environments for sample-efficient navigation policy learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 1697–1704.

[2] R. Cheng, K. Wang, J. Bai, and Z. Xu, "Unifying visual localization and scene recognition for people with visual impairment," *IEEE Access*, vol. 8, pp. 64284–64296, 2020.

[3] S. Lowry, N. Sunderhauf, 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. 2016.

[4] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet, "A survey on visual-based localization: On the benefit of heterogeneous data," *Pattern Recognit.*, vol. 74, pp. 90–109, Feb. 2018.

[5] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognit.*, vol. 28, Nov. 2020, Art. no. 107760.

[6] T. Tuytelaars and K. Mikolajczyk, "A survey on local invariant features," *Found. Trends Comput. Graph. Vis.*, vol. 3, no. 3, pp. 176–280, 2008.

[7] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2005, pp. 524–531.

[8] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proc. ECCV*, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Berlin, Germany: Springer, 2002, pp. 128–142.

[9] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, Sep. 2004.

[10] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1150–1157.

[11] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[12] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2911–2918.

[13] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "BRIEF: Computing a local binary descriptor very fast," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1281–1298, Jul. 2012.

[14] J. Dong and S. Soatto, "Domain-size pooling in local descriptors: DSP-SIFT," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5097–5106.

[15] A. Bursuc, G. Tolias, and H. Jégou, "Kernel local descriptors with implicit rotation matching," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, p. 595.

[16] A. Mukundan, G. Tolias, and O. Chum, "Explicit spatial encoding for deep local descriptors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9394–9403.

[17] W. Zhang and J. Kosecka, "Image based localization in urban environments," in *Proc. 3rd Int. Symp. 3D Data Process., Visualizat., Transmiss.*, 2006, pp. 33–40.

[18] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–7.

[19] 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, Jun. 2008.

[20] J. Hays and A. A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[21] M. Cummins and P. Newman, "Highly scalable appearance-only SLAM-FAB-MAP 2.0," in *Proc. Robot., Sci. Syst.*, Seattle, WA, USA, Jun. 2009, pp. 1–8.

[22] Y. Avrithis, G. Tolias, and Y. Kalantidis, "Feature map hashing: Sub-linear indexing of appearance and global geometry," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 231–240.

[23] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *Proc. ECCV*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Germany: Springer, 2010, pp. 748–761.

[24] E. Johns and G.-Z. Yang, "From images to scenes: Compressing an image cluster into a single scene model for place recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 874–881.

[25] W. Churchill and P. Newman, "Experience-based navigation for long-term localisation," *Int. J. Robot. Res.*, vol. 32, no. 14, pp. 1645–1661, Dec. 2013.

[26] J. Hays and A. A. Efros, *Large-Scale Image Geolocalization*. Cham, Switzerland: Springer, 2015, pp. 41–62.

[27] H. Jin Kim, E. Dunn, and J.-M. Frahm, "Predicting good features for image geo-localization using per-bundle VLAD," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1170–1178.

[28] A. L. Majdik, D. Verda, Y. Albers-Schoenberg, and D. Scaramuzza, "Air-ground matching: Appearance-based GPS-denied urban localization of micro aerial vehicles," *J. Field Robot.*, vol. 32, no. 7, pp. 1015–1039, 2015.
[29] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic visual localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6896–6906.
[30] E. Zemene, Y. T. Tesfaye, H. Idrees, A. Prati, M. Pelillo, and M. Shah, "Large-scale image geo-localization using dominant sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 148–161, Jan. 2019.
[31] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 1470–1477.
[32] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes*, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann, 1999.
[33] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.
[34] R. Arandjelovic and A. Zisserman, "All about VLAD," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 1578–1585.
[35] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-scale image retrieval with compressed Fisher vectors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3384–3391.
[36] H. Jegou and A. Zisserman, "Triangulation embedding and democratic aggregation for image search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3310–3317.
[37] G. Tolias, Y. Avrithis, and H. Jégou, "Image search with selective match kernels: Aggregation across single and multiple images," *Int. J. Comput. Vis.*, vol. 116, no. 3, pp. 247–261, Feb. 2016.
[38] 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. 5104–5113.
[39] H. Jãgou and O. Chum, "Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening," in *Proc. ECCV*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Germany: Springer, 2012, pp. 774–787.
[40] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.
[41] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," in *Vision Perception* (Progress in Brain Research), vol. 155, S. Martinez-Conde, S. Macknik, L. Martinez, J.-M. Alonso, and P. Tse, Eds. Amsterdam, The Netherlands: Elsevier, 2006, pp. 23–36.
[42] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large image databases for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[43] 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.
[44] N. Sãnderhauf and P. Protzel, "Brief-gist-closing the loop by simple means," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Sep. 2011, pp. 1234–1241.
[45] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1643–1649.
[46] Y. Liu and H. Zhang, "Visual loop closure detection with a compact image descriptor," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 1051–1056.
[47] C. Azzi, D. Asmar, A. Fakih, and J. Zelek, "Filtering 3D keypoints using GIST for accurate image-based localization," in *Proc. Brit. Mach. Vis. Conf.*, 2016, p. 127.
[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105.
[49] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features Off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 512–519.
[50] J. Donahue, Y. Jia, and O. Vinyals, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.
[51] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 36–45.
[52] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.
[53] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 818–833.
[54] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[55] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
[56] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 392–407.
[57] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Proc. ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 584–599.
[58] Y. Liu, Y. Guo, S. Wu, and M. S. Lew, "DeepIndex for accurate and efficient image retrieval," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, p. 43.
[59] L. Xie, R. Hong, B. Zhang, and Q. Tian, "Image classification and retrieval are ONE," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 3–10.
[60] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *Proc. 22nd ACM Int. Conf. Multimedia*, New York, NY, USA, Nov. 2014, p. 157.
[61] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[62] R. Gomez-Ojeda, M. Lopez-Antequera, N. Petkov, and J. Gonzalez-Jimenez, "Training a convolutional neural network for appearance-invariant place recognition," 2015, *arXiv:1505.07428*. [Online]. Available: http://arxiv.org/abs/1505.07428
[63] 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.
[64] M. Paulin, M. Douze, Z. Harchaoui, J. Mairal, F. Perronin, and C. Schmid, "Local convolutional features with unsupervised training for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 91–99.
[65] E. Mohedano, K. McGuinness, N. E. O'Connor, A. Salvador, F. Marques, and X. Giro-i-Nieto, "Bags of local convolutional features for scalable instance search," in *Proc. ACM Int. Conf. Multimedia Retr.*, New York, NY, USA, 2016, pp. 327–331.
[66] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, "Working hard to know your neighbor margins: Local descriptor learning loss," in *Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 4826–4837.
[67] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 726–743.
[68] E.-J. Ong, S. Husain, and M. Bober, "Siamese network of deep Fisher-vector descriptors for image retrieval," 2017, *arXiv:1702.00338*. [Online]. Available: http://arxiv.org/abs/1702.00338
[69] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1437–1451, Jun. 2018.
[70] A. Babenko and V. Lempitsky, "Aggregating deep convolutional features for image retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2015, pp. 1269–1277.
[71] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "[Paper] visual instance retrieval with deep convolutional networks," *ITE Trans. Media Technol. Appl.*, vol. 4, no. 3, pp. 251–258, Dec. 2016.

[72] A. Mousavian and J. Kosecka, "Deep convolutional features for image based retrieval and scene categorization," 2015, *arXiv:1509.06033*. [Online]. Available: http://arxiv.org/abs/1509.06033

[73] G. Tolias, R. Sicre, and H. Jãgou, "Particular object retrieval with integral max-pooling of CNN activations," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2016, pp. 1–5.

[74] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *Int. J. Comput. Vis.*, vol. 124, no. 2, pp. 237–254, 2017.

[75] J. Mao, X. Hu, X. He, L. Zhang, L. Wu, and M. J. Milford, "Learning to fuse multiscale features for visual place recognition," *IEEE Access*, vol. 7, pp. 5723–5735, 2019.

[76] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV) Workshops*, G. Hua and H. Jégou, Eds. Cham, Switzerland: Springer, 2016, pp. 685–701.

[77] F. Radenovic, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, Jul. 2019.

[78] K. Mikolajczyk and J. Matas, "Improving descriptors for fast tree matching by optimal linear projection," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[79] A. Gordoa, J. A. Rodriguez-Serrano, F. Perronnin, and E. Valveny, "Leveraging category-level labels for instance-level image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3045–3052.

[80] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *Proc. ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 241–257.

[81] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3476–3485.

[82] J. Revaud, J. Almazan, R. Rezende, and C. De Souza, "Learning with average precision: Training image retrieval with a listwise loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Nov. 2019, pp. 5106–5115.

[83] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," in *Proc. ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 3–20.

[84] Y. Zhu, J. Wang, L. Xie, and L. Zheng, "Attention-based pyramid aggregation network for visual place recognition," in *Proc. 26th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2018, p. 99.

[85] A. Mukundan, G. Tolias, A. Bursuc, H. Jégou, and O. Chum, "Understanding and improving kernel local descriptors," *Int. J. Comput. Vis.*, vol. 127, nos. 11–12, pp. 1723–1737, Dec. 2019.

[86] K. Liu, H. Wang, F. Han, and H. Zhang, "Visual place recognition via robust ℓ_2-norm distance based holism and landmark integration," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8034–8041.

[87] F. Han and H. Wang, "Learning integrated holism-landmark representations for long-term loop closure detection," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Jan. 2018, pp. 6501–6508.

[88] S. Yokoo, K. Ozaki, E. Simo-Serra, and S. Iizuka, "Two-stage discriminative re-ranking for large-scale landmark retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1012–1013.

[89] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4690–4699.

[90] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1386–1393.

[91] H. Jin Kim, E. Dunn, and J.-M. Frahm, "Learned contextual feature reweighting for image geo-localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3251–3260.

[92] Y. Lou, Y. Bai, S. Wang, and L.-Y. Duan, "Multi-scale context attention network for image retrieval," in *Proc. 26th ACM Int. Conf. Multimedia*, New York, NY, USA, 2018, p. 1128–1136.

[93] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognition*, A. Feragen, M. Pelillo, and M. Loog, Eds. Cham, Switzerland: Springer, 2015, pp. 84–92.

[94] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas, "SOSNet: Second order similarity regularization for local descriptor learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, p. 11 016–11 025.

[95] T. Ng, B. Vassileios, T. Yurun, and M. Krystian, "SOLAR: Second-order loss and attention for image retrieval," in *Proc. ECCV*, 2020, pp. 253–270.

[96] D. Mishkin, F. Radenovic, and J. Matas, "Repeatability is not enough: Learning affine regions via discriminability," in *Proc. ECCV*, Sep. 2018, pp. 284–300.

[97] H. Hu, Z. Qiao, M. Cheng, Z. Liu, and H. Wang, "DASGIL: Domain adaptation for semantic and geometric-aware image-based localization," *IEEE Trans. Image Process.*, vol. 30, pp. 1342–1353, 2021.

[98] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[99] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, and F. Moreno-Noguer, "Fracking deep convolutional image descriptors," 2014, *arXiv:1412.6537*. [Online]. Available: http://arxiv.org/abs/1412.6537

[100] 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.

[101] T.-Y. Liu, *Learning to Rank for Information Retrieval*. Cham, Switzerland: Springer, 2011.

[102] F. Cakir, K. He, X. Xia, B. Kulis, and S. Sclaroff, "Deep metric learning to rank," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1861–1870.

[103] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, "R2d2: Reliable and repeatable detector and descriptor," in *Advances in Neural Inf. Processing System*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, p. 12 405–12 415.

[104] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPS Deep Learn. Represent. Learn. Workshop*, 2015, pp. 1–5.

[105] P.-E. Sarlin, F. Debraine, M. Dymczyk, R. Siegwart, and C. Cadena, "Leveraging deep visual descriptors for hierarchical efficient localization," in *Proc. The 2nd Conf. Robot Learn.*, vol. 87, A. Billard, A. Dragan, J. Peters, and J. Morimoto, Eds., 2018, pp. 456–465.

[106] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, p. 12716.

[107] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," in *Proc. Robot., Sci. Syst.*, Pittsburgh, PA, USA, Jun. 2018, pp. 1–25.

[108] C. Böhm, S. Berchtold, and D. A. Keim, "Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases," *ACM Comput. Surv.*, vol. 33, no. 3, p. 322–373, Sep. 2001.

[109] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, vol. 2, Feb. 2009, pp. 331–340.

[110] M. Muja and D. G. Lowe, "Fast matching of binary features," in *Proc. 9th Conf. Comput. Robot Vis.*, May 2012, pp. 404–410.

[111] J. Wang, T. Zhang, j. song, N. Sebe, and H. Tao Shen, "A survey on learning to hash," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 769–790, Apr. 2018.

[112] F. Magliani, K. McGuinness, E. Mohedano, and A. Prati, "An efficient approximate kNN graph method for diffusion on image retrieval," in *Image Analysis Processing*, E. Ricci, S. R. Bulò, C. Snoek, O. Lanz, S. Messelodi, and N. Sebe, Eds. Cham, Switzerland: Springer, 2019, pp. 537–548.

[113] J. Johnson, M. Douze, and H. Jegou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, early access, Jun. 7, 2019, doi: 10.1109/TBDATA.2019.2921572.

[114] A. R. Zamir and M. Shah, "Image geo-localization based on MultipleNearest neighbor feature matching UsingGeneralized graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1546–1558, Aug. 2014.

[115] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0306457388900210

[116] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[117] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla, "Visual place recognition with repetitive structures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2346–2359, Nov. 2015.

[118] H. Jăgou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. ECCV*, D. Forsyth, P. Torr, and A. Zisserman, Eds. Berlin, Germany: Springer, 2008, pp. 304–317.

[119] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "InLoc: Indoor visual localization with dense matching and view synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7199–7209.

[120] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, Jan. 2011.

[121] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 257–271, Feb. 2018.

[122] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[123] G. Tolias and H. Jégou, "Local visual Query expansion: Exploiting an image collection to refine local descriptors," INRIA, Chile, France, Res. Rep. RR-8325, Jul. 2013.

[124] X. Li, M. Larson, and A. Hanjalic, "Pairwise geometric matching for large-scale object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5153–5161.

[125] A. Babenko and V. Lempitsky, "The inverted multi-index," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1247–1260, Jun. 2015.

[126] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic Query expansion with a generative feature model for object retrieval," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[127] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall II: Query expansion revisited," in *Proc. CVPR*, Jun. 2011, pp. 889–896.

[128] S. Garg, N. Suenderhauf, and M. Milford, "LoST appearance-invariant place recognition for opposite viewpoints using visual semantics," in *Proc. Robot., Sci. Syst.*, Pittsburgh, PA, Jun. 2018, pp. 1–8.

[129] F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Revisiting oxford and paris: Large-scale image retrieval benchmarking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5706–5715.

[130] C. Chang, G. Yu, C. Liu, and M. Volkovs, "Explore-exploit graph traversal for image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9423–9431.

[131] O. Simeoni, Y. Avrithis, and O. Chum, "Local features and visual words emerge in activations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11651–11660.

[132] O. Chum, J. Matas, and S. Obdrzalek, "Enhancing RANSAC by generalized model optimization," in *Proc. ACCV*, vol. 2, Jan. 2004, pp. 812–817.

[133] O. Chum and J. Matas, "Matching with PROSAC-progressive sample consensus," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 220–226.

[134] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable CNN for joint description and detection of local features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8092–8101.

[135] 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. 337–349.

[136] K. Ozaki and S. Yokoo, "Large-scale landmark retrieval/recognition under a noisy and diverse dataset," 2019, *arXiv:1906.04087*. [Online]. Available: http://arxiv.org/abs/1906.04087

[137] L. G. Camara and L. Preucil, "Spatio-semantic ConvNet-based visual place recognition," in *Proc. Eur. Conf. Mobile Robots (ECMR)*, Sep. 2019, pp. 1–8.

[138] R. Imbriaco, C. Sebastian, E. Bondarev, and P. H. N. de With, "Aggregated deep local features for remote sensing image retrieval," *Remote Sens.*, vol. 11, no. 5, p. 493, Feb. 2019.

[139] M. Donoser and H. Bischof, "Diffusion processes for retrieval revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1320–1327.

[140] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum, "Efficient diffusion on region manifolds: Recovering small objects with compact CNN representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2077–2086.

[141] A. Iscen, Y. Avrithis, G. Tolias, T. Furon, and O. Chum, "Fast spectral ranking for similarity search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7632–7641.

[142] S. Bai, X. Bai, Q. Tian, and L. J. Latecki, "Regularized diffusion process on bidirectional context for object retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1213–1226, May 2019.

[143] F. Yang, R. Hinami, Y. Matsui, S. Ly, and S. Satoh, "Efficient image retrieval via decoupling diffusion into online and offline processing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 9087–9094.

[144] C. Liu, G. Yu, M. Volkovs, C. Chang, H. Rai, J. Ma, and S. K. Gorti, "Guided similarity separation for image retrieval," in *Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 1556–1566.

[145] B. Jiang, D. Lin, J. Tang, and B. Luo, "Data representation and learning with graph diffusion-embedding networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, p. 10406.

[146] Z. Dou, H. Cui, L. Zhang, and B. Wang, "Learning global and local consistent representations for unsupervised image retrieval via deep graph diffusion networks," 2020, *arXiv:2001.01284*. [Online]. Available: http://arxiv.org/abs/2001.01284

[147] H. Jegou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1169–1176.

[148] F. Maffra, Z. Chen, and M. Chli, "Viewpoint-tolerant place recognition combining 2D and 3D information for UAV navigation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 2542–2549.

[149] T. Kanji, "Self-localization from images with small overlap," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 4497–4504.

[150] 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, *arXiv:1903.09107*. [Online]. Available: http://arxiv.org/abs/1903.09107

[151] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla, "Benchmarking 6DOF outdoor visual localization in changing conditions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8601–8610.

[152] H. Badino, D. Huber, and T. Kanade, "Visual topometric localization," in *Proc. IEEE Intell. Vehicles Sympo. (IV)*, Oct. 2011, pp. 794–799.

[153] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image retrieval for image-based localization revisited," in *Proc. Brit. Mach. Vis. Conf.*, 2012, p. 76.

[154] Z. Zhang, T. Sattler, and D. Scaramuzza, "Reference pose generation for long-term visual localization via learned features and view synthesis," *Int. J. Comput. Vis.*, vol. 4, pp. 1–24, Dec. 2020.

[155] 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. Automat. (ICRA)*, 2013, p. 2013.

[156] R. Arandjeloviä and A. Zisserman, "DisLocation: Scalable descriptor distinctiveness for location recognition," in *Proc. ACCV*, vol. 4, D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eds. Cham, Switzerland: Springer, 2015, pp. 188–204.

[157] C. McManus, B. Upcroft, and P. Newman, "Scene signatures: Localised and point-less features for localisation," in *Proc. Robot., Sci. Syst.*, D. Fox, L. E. Kavraki, and H. Kurniawati, Eds., 2014, pp. 1–9.

[158] 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. Robot., Sci. Syst.*, 2015, pp. 1–10.

[159] P. Panphattarasap and A. Calway, "Visual place recognition using landmark distribution descriptors," in *Proc. ACCV*, S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds. Cham, Switzerland: Springer, 2017, pp. 487–502.

[160] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Neural Information Processing Systems*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2015, pp. 91–99.

[161] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[162] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 21–37.

[163] 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.

[164] 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. 2020.

[165] Y. Kong, W. Liu, and Z. Chen, "Robust convnet landmark-based visual place recognition by optimizing landmark matching," *IEEE Access*, vol. 7, pp. 30754–30767, 2019.

[166] E. Mohedano, K. McGuinness, X. Giro-i-Nieto, and N. E. O'Connor, "Saliency weighted convolutional features for instance search," in *Proc. Int. Conf. Content-Based Multimedia Indexing (CBMI)*, Sep. 2018, pp. 1–6.

[167] 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.

[168] G. Moreno Berton, V. Paolicelli, C. Masone, and B. Caputo, "Adaptive-attentive geolocalization from few queries: A hybrid approach," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2918–2927.

[169] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[170] 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.

[171] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[172] H. Zhou, T. Sattler, and D. W. Jacobs, "Evaluating local features for day-night matching," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV) Workshops*, G. Hua and H. Jégou, Eds. Cham, Switzerland: Springer, 2016, pp. 724–736.

[173] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5007–5015.

[174] F. Castaldo, A. Zamir, R. Angst, F. Palmieri, and S. Savarese, "Semantic cross-view matching," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 1044–1052.

[175] A. Mousavian, J. Kosecka, and J.-M. Lien, "Semantically guided location recognition for outdoors scenes," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 4882–4889.

[176] E. Pepperell, P. I. Corke, and M. J. Milford, "All-environment visual place recognition with SMART," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 1612–1618.

[177] R. Arandjelović and A. Zisserman, "Visual vocabulary with a semantic twist," in *Proc. ACCV*, D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eds., vol. 1. Cham, Switzerland: Springer, 2015, pp. 178–195.

[178] S. Garg, A. Jacobson, S. Kumar, and M. Milford, "Improving condition- and environment-invariant place recognition with semantic place categorization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, Sep. 2017, pp. 6863–6870.

[179] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2014, pp. 487–495.

[180] W. Zhang, G. Liu, and G. Tian, "A coarse to fine indoor visual localization method using environmental semantic information," *IEEE Access*, vol. 7, pp. 21963–21970, 2019.

[181] S. Garg, N. Suenderhauf, and M. Milford, "Semantic–geometric visual place recognition: A new perspective for reconciling opposing views," *Int. J. Robot. Res.*, vol. 2019, Apr. 2019, Art. no. 027836491983976.

[182] A. Gawel, C. D. Don, R. Siegwart, J. Nieto, and C. Cadena, "X-view: Graph-based semantic multi-view localization," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 1687–1694, Jul. 2018.

[183] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, "Semantics-aware visual localization under challenging perceptual conditions," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 2614–2620.

[184] Y. Hou, H. Zhang, S. Zhou, and H. Zou, "Use of roadway scene semantic information and geometry-preserving landmark pairs to improve visual place recognition in changing environments," *IEEE Access*, vol. 5, pp. 7702–7713, 2017.

[185] Z. Seymour, K. Sikka, H. Chiu, S. Samarasekera, and R. Kumar, "Semantically-aware attentive neural embeddings for 2D long-term visual localization," in *Brit. Mach. Vis. Conf. (BMVC)*, 2019, pp. 1–15.

[186] A. Benbihi, S. Arravechia, M. Geist, and C. Pradalier, "Image-based place recognition on bucolic environment across seasons from semantic edge description," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 3032–3038.

[187] M. Larsson, E. Stenborg, C. Toft, L. Hammarstrand, T. Sattler, and F. Kahl, "Fine-grained segmentation networks: Self-supervised segmentation for improved long-term visual localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 31–41.

[188] N. Piasco, D. Sidibe, V. Gouet-Brunet, and C. Demonceaux, "Learning scene geometry for visual localization in challenging conditions," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 9094–9100.

[189] 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.

[190] 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.

[191] T. Jenicek and O. Chum, "No fear of the dark: Image retrieval under varying illumination conditions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9696–9704.

[192] T. Naseer, W. Burgard, and C. Stachniss, "Robust visual localization across seasons," *IEEE Trans. Robot.*, vol. 34, no. 2, pp. 289–302, Apr. 2018.

[193] S. Hausler, A. Jacobson, and M. Milford, "Multi-process fusion: Visual place recognition using multiple image processing methods," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1924–1931, Apr. 2019.

[194] Z. Hong, Y. Petillot, D. Lane, Y. Miao, and S. Wang, "TextPlace: Visual place recognition and topological localization through reading scene texts," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2861–2870.

[195] M. Chancan, L. Hernandez-Nunez, A. Narendra, A. B. Barron, and M. Milford, "A hybrid compact neural architecture for visual place recognition," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 993–1000, Apr. 2020.

[196] D. Doan, Y. Latif, T.-J. Chin, Y. Liu, T.-T. Do, and I. Reid, "Scalable place recognition under appearance change for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9319–9328.

[197] H. Porav, W. Maddern, and P. Newman, "Adversarial training for adverse conditions: Robust metric localisation using appearance transfer," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1011–1018.

[198] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. V. Gool, "Night-to-Day image translation for retrieval-based localization," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 5958–5964.

[199] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[200] R. Guerrero, C. Qin, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, M. C. Valdés-Hernández, D. A. Dickie, J. Wardlaw, and D. Rueckert, "White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks," *NeuroImage*, vol. 17, pp. 918–934, Oct. 2018.

[201] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool, "ComboGAN: Unrestrained scalability for image domain translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 896–903.

[202] H. Hu, H. Wang, Z. Liu, C. Yang, W. Chen, and L. Xie, "Retrieval-based localization based on domain-invariant feature learning under changing environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 3684–3689.

[203] Y. Latif, R. Garg, M. Milford, and I. Reid, "Addressing challenging place recognition tasks using generative adversarial networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 2349–2355.

[204] P. Yin, L. Xu, X. Li, C. Yin, Y. Li, R. A. Srivatsan, L. Li, J. Ji, and Y. He, "A multi-domain feature learning method for visual place recognition," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 319–324.

[205] Z. Wang, J. Li, S. Khademi, and J. van Gemert, "Attention-aware age-agnostic visual place recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1437–1446.

[206] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from Internet photo collections," *Int. J. Comput. Vis.*, vol. 80, no. 2, pp. 189–210, Nov. 2008.

[207] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building Rome in a day," *Commun. ACM*, vol. 54, no. 10, p. 105–112, Oct. 2011.

[208] J. L. Schonberger, F. Radenovic, O. Chum, and J.-M. Frahm, "From single image Query to detailed 3D reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5126–5134.

[209] E. Wijmans and Y. Furukawa, "Exploiting 2D floorplan for building-scale panorama RGBD alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1427–1435.

[210] A. Bergamo, S. N. Sinha, and L. Torresani, "Leveraging structure from motion to learn discriminative codebooks for scalable landmark classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 763–770.

[211] L. Kneip, D. Scaramuzza, and R. Siegwart, "A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation," in *Proc. CVPR*, Jun. 2011, pp. 2969–2976.

[212] T. Sattler, C. Sweeney, and M. Pollefeys, "On sampling focal length values to solve the absolute pose problem," in *Proc. ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 828–843.

[213] F. Maffra, L. Teixeira, Z. Chen, and M. Chli, "Real-time wide-baseline place recognition using depth completion," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1525–1532, Apr. 2019.

[214] A. Torii, H. Taira, J. Sivic, M. Pollefeys, M. Okutomi, T. Pajdla, and T. Sattler, "Are large-scale 3D models really necessary for accurate visual localization?" *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 16, 2019, doi: 10.1109/TPAMI.2019.2941876.

[215] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2938–2946.

[216] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5974–5983.

[217] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using LSTMs for structured feature correlation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 627–637.

[218] Y. Ono, E. Trulls, P. Fua, and K. M. Yi, "LF-net: Learning local features from images," in *Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2018, pp. 6234–6244. [Online]. Available: http://papers.nips.cc/paper/7861-lf-net-learning-local-features-from-images.pdf

[219] X. Tang, X. Zhang, F. Liu, and L. Jiao, "Unsupervised deep feature learning for remote sensing image retrieval," *Remote Sens.*, vol. 10, no. 8, p. 1243, 2018.

[220] A. Iscen, T. Furon, V. Gripon, M. Rabbat, and H. Jegou, "Memory vectors for similarity search in high-dimensional spaces," *IEEE Trans. Big Data*, vol. 4, no. 1, pp. 65–77, Mar. 2018.

[221] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3961–3969.

[222] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, K. Alexis, and K. D. McDonald-Maier, "Are state-of-the-art visual place recognition techniques any good for aerial robotics?" 2019, *arXiv:1904.07967*. [Online]. Available: https://arxiv.org/abs/1904.07967

[223] M. A. Uy and G. H. Lee, "PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4470–4479.

[224] Z. Liu, S. Zhou, C. Suo, P. Yin, W. Chen, H. Wang, H. Li, and Y. Liu, "LPD-net: 3D point cloud learning for large-scale place recognition and environment analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2831–2840.

[225] Y. S. Park, J. Jeong, Y. Shin, and A. Kim, "Radar dataset for robust localization and mapping in urban environment," in *Proc. ICRA*, Montreal, QC, Canada, May 2019, p. 152.

[226] X. Zhang, J. Lin, Q. Li, T. Liu, and Z. Fang, "Continuous indoor visual localization using a spatial model and constraint," *IEEE Access*, vol. 8, pp. 69800–69815, 2020.

[227] N. Radwan, G. D. Tipaldi, L. Spinello, and W. Burgard, "Do you see the bakery? Leveraging geo-referenced texts for global localization in public maps," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 4837–4842.

[228] P. Hansen and B. Browning, "Visual place recognition using HMM sequence matching," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2014, pp. 4549–4555.

[229] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," in *Proc. 16th Australas. Conf. Robot. Autom.*, 2014, pp. 1–8.

[230] J. M. Facil, D. Olid, L. Montesano, and J. Civera, "Condition-invariant multi-view place recognition," 2019, *arXiv:1902.09516*. [Online]. Available: http://arxiv.org/abs/1902.09516

[231] M. J. Milford, G. F. Wyeth, and D. Prasser, "RatSLAM: A hippocampal model for simultaneous localization and mapping," in *Proc. IEEE Int. Conf. Robot. Autom.*, Oct. 2004, pp. 403–408.

[232] C. Fan, Z. Chen, A. Jacobson, X. Hu, and M. Milford, "Biologically-inspired visual place recognition with adaptive multiple scales," *Robot. Auton. Syst.*, vol. 96, pp. 224–237, Aug. 2017.

[233] F. Yu, J. Shang, Y. Hu, and M. Milford, "NeuroSLAM: A brain-inspired SLAM system for 3D environments," *Biol. Cybern.*, vol. 113, nos. 5–6, pp. 515–545, Dec. 2019.

[234] N. Radwan, A. Valada, and W. Burgard, "VLocNet++: Deep multitask learning for semantic visual localization and odometry," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4407–4414, Oct. 2018.

[235] P. Wang, R. Yang, B. Cao, W. Xu, and Y. Lin, "DeLS-3D: Deep localization and segmentation with a 3D semantic map," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5860–5869.

[236] T. Weyand, I. Kostrikov, and J. Philbin, "Planet-photo geolocation with convolutional neural networks," in *Proc. ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 37–55.

[237] N. Vo, N. Jacobs, and J. Hays, "Revisiting IM2GPS in the deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2640–2649.

[238] P. H. Seo, T. Weyand, J. Sim, and B. Han, "CPlaNet: Enhancing image geolocalization by combinatorial partitioning of maps," in *Proc. ECCV*, Sep. 2018, pp. 536–551.

[239] E. Muller-Budack, K. Pustu-Iren, and R. Ewerth, "Geolocation estimation of photos using a hierarchical model and scene classification," in *Proc. ECCV*, Sep. 2018, pp. 563–579.

[240] M. Warren, D. McKinnon, H. He, and B. Upcroft, "Unaided stereo vision based pose estimation," in *Proc. Australas. Conf. Robot. Autom.*, G. Wyeth and B. Upcroft, Eds. Brisbane, Australia: Australian Robotics and Automation Association, 2010, pp. 1–8

[241] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, "FAB-MAP+RatSLAM: Appearance-based SLAM for multiple times of day," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 3507–3512.

[242] Y. Li, N. Snavely, and D. P. Huttenlocher, "Location recognition using prioritized feature matching," in *Proc. ECCV*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Germany: Springer, 2010, pp. 791–804.

[243] D. M. Chen, G. Baatz, K. Koser, S. S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk, "City-scale landmark identification on mobile devices," in *Proc. CVPR*, Jun. 2011, pp. 737–744.

[244] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in RGB-D images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2930–2937.

[245] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 2564–2570.
[246] T. Weyand and B. Leibe, "Visual landmark recognition from Internet photo collections: A large-scale evaluation," *Comput. Vis. Image Understand.*, vol. 135, pp. 1–15, Jun. 2015.
[247] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of michigan north campus long-term vision and lidar dataset," *Int. J. Robot. Res.*, vol. 35, no. 9, pp. 1023–1035, Aug. 2016.
[248] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford robotcar dataset," *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, Jan. 2017.
[249] M. Leyva-Vallina, N. Strisciuglio, and N. Petkov, "Place recognition in gardens by learning visual representations: Data set and benchmark analysis," in *Computer Analysis of Images and Patterns*, M. Vento and G. Percannella, Eds. Cham, Switzerland: Springer, 2019, pp. 324–335.
[250] M. Leyva-Vallina, N. Strisciuglio, M. L. Antequera, R. Tylecek, M. Blaich, and N. and Petkov, "TB-Places: A data set for visual place recognition in garden environments," *IEEE Access*, vol. 7, pp. 52277–52287, 2019.
[251] T. Weyand, A. Araujo, B. Cao, and J. Sim, "Google landmarks dataset V2—A large-scale benchmark for instance-level recognition and retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2572–2581.
[252] F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2626–2635.
[253] F. Maffra, L. Teixeira, Z. Chen, and M. Chli, "Loop-closure detection in urban scenes for autonomous robot navigation," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 356–364.
[254] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. (2019). *Lyft Level 5 Perception Dataset*. [Online]. Available: https://level5.lyft.com/dataset/
[255] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The ApolloScape open dataset for autonomous driving and its application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2702–2719, Oct. 2020.
[256] P. Sun, H. Kretzschmar, and X. Dotiwalla, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2443–2451.
[257] M.-F. Chang, D. Ramanan, J. Hays, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, and S. Lucey, "Argoverse: 3D tracking and forecasting with rich maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8740–8749.
[258] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2633–2642.
[259] J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A. S. Chung, L. Hauswald, V. Hoang Pham, M. Mählegg, S. Dorn, T. Fernandez, M. Jänicke, S. Mirashi, C. Savani, M. Sturm, O. Vorobiov, M. Oelker, S. Garreis, and P. Schuberth, "A2D2: Audi autonomous driving dataset," 2020, *arXiv:2004.06320*. [Online]. Available: http://arxiv.org/abs/2004.06320
[260] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, p. 11621.
[261] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3234–3243.
[262] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proc. ECCV*, vol. 9906, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 102–118.
[263] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 4340–4349.
[264] Y. Cabon, N. Murray, and M. Humenberger, "Virtual KITTI 2," 2020, *arXiv:2001.10773*. [Online]. Available: http://arxiv.org/abs/2001.10773
[265] E. Alberti, A. Tavera, C. Masone, and B. Caputo, "IDDA: A large-scale multi-domain dataset for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5526–5533, Oct. 2020.
[266] F. Perronnin, Y. Liu, and J.-M. Renders, "A family of contextual measures of similarity between distributions with application to image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2358–2365.

CARLO MASONE (Member, IEEE) received the B.S. and M.S. degrees in control engineering from Sapienza University of Rome, Rome, Italy, in 2006 and 2010, respectively, and the Ph.D. degree in control engineering from the University of Stuttgart in collaboration with the Max Planck Institute for Biological Cybernetics (MPI-Kyb), Stuttgart, Germany, in 2014. From 2014 to 2017, he was a Postdoctoral Researcher with MPI-kyb, within the Autonomous Robotics and Human-Machine Systems Group. From 2017 to 2020, he worked with industry on the development of self-driving cars. Since 2020, he has been a Senior Researcher with the Visual and Multimodal Applied Learning Team, Italian Institute of Technology (IIT).

BARBARA CAPUTO (Member, IEEE) received the Ph.D. degree in computer science from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2005. From 2007 to 2013, she was a Senior Researcher with Idiap-EPFL (CH). Then, she moved to Sapienza Rome University thanks to a MUR professorship, and joined the Politecnico di Torino, in 2018. Since 2017, she has been a double affiliation with the Italian Institute of Technology (IIT). She is currently a Full Professor with the Politecnico of Torino, where she leads the Hub AI@PoliTo. She is one of the 30 experts who contributed to write the Italian Strategy on AI, and coordinator of the Italian National Ph.D. on AI and Industry 4.0, sponsored by MUR. She is also an ERC Laureate and an ELLIS Fellow. Since 2019, she serves on the ELLIS Board.