Where Is Your Place, Visual Place Recognition?

Sourav Garg*, **Tobias Fischer*** and **Michael Milford**
QUT Centre for Robotics, Queensland University of Technology, Brisbane, Australia
{s.garg, tobias.fischer, michael.milford}@qut.edu.au

Abstract

Visual Place Recognition (VPR) is often characterized as being able to recognize the same place despite significant changes in appearance and viewpoint. VPR is a key component of Spatial Artificial Intelligence, enabling robotic platforms and intelligent augmentation platforms such as augmented reality devices to perceive and understand the physical world. In this paper, we observe that there are three "drivers" that impose requirements on spatially intelligent agents and thus VPR systems: 1) the particular agent including its sensors and computational resources, 2) the operating environment of this agent, and 3) the specific task that the artificial agent carries out. In this paper, we characterize and survey key works in the VPR area considering those drivers, including their place representation and place matching choices. We also provide a new definition of VPR based on the visual overlap – akin to spatial view cells in the brain – that enables us to find similarities and differences to other research areas in the robotics and computer vision fields. We identify several open challenges and suggest areas that require more in-depth attention in future works.

1 Introduction

Visual Place Recognition (VPR) is a rapidly growing topic: Google Scholar lists over 2300 papers matching this exact term, with 1600 of them published since the pivotal survey paper by Lowry *et al.* in 2016. While exhaustive surveys of works on VPR are given elsewhere [Lowry *et al.*, 2016; Zhang *et al.*, 2020; Masone and Caputo, 2021], our goal here is to lay a concrete understanding of VPR as a research problem. VPR capability is based on the fundamental ability to aptly represent incoming information and associate the incoming information with previously stored information. This is required for an embodied or augmented agent to intelligently understand the operating environment, to navigate within it and interact with it, thus evolving Simultaneous Localization And Mapping (SLAM) into Spatial AI [Davison, 2018][1]. We argue that research on VPR has increasingly become more dissociated: there is no standard definition of a 'place' and comparison of methods is challenging as benchmark datasets and metrics vary substantially.

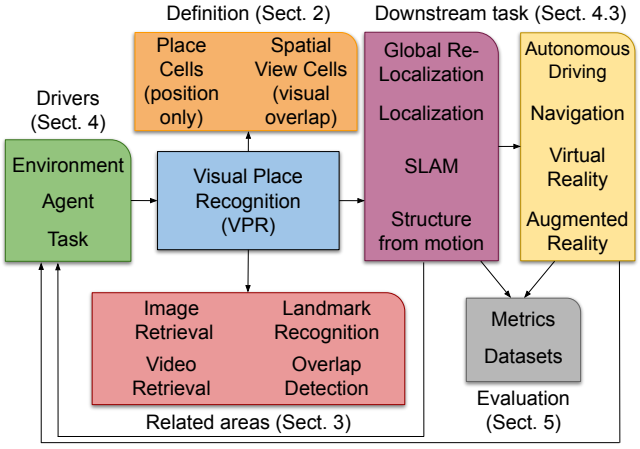

Figure 1: Visual Place Recognition (VPR) is the ability to recognize one's location based on reference and query observations perceived from overlapping field-of-views. This figure illustrates the main sections of this paper and how they interrelate.

In light of the dissociation and while retaining the accessibility of a compact treatment, we discuss VPR with regards to its definition (Section 2), how it closely relates to other areas of research (Section 3), what the key drivers of VPR research are (Section 4), how to evaluate VPR solutions (Section 5), and what key research problems still remain unsolved (Section 6). Figure 1 illustrates the outline of our paper and shows how the various sections are interrelated.

2 What is Visual Place Recognition?

Lowry *et al.* state that VPR addresses the question of "given an image of a place, can a human, animal, or robot decide whether or not this image is of a place it has already seen?" [Lowry *et al.*, 2016]. One can easily see that such a capability is of

*Sourav Garg and Tobias Fischer contributed equally to the paper.

[1]SLAM is the problem of incrementally estimating the environment's structure while simultaneously tracking the robot's pose within the environment. Traditionally SLAM's purpose was purely navigation, while the inclusion of semantics moved SLAM towards Spatial AI, whose aim is intelligent goal-driven interaction of the robot with the environment and other agents [Cadena *et al.*, 2016].

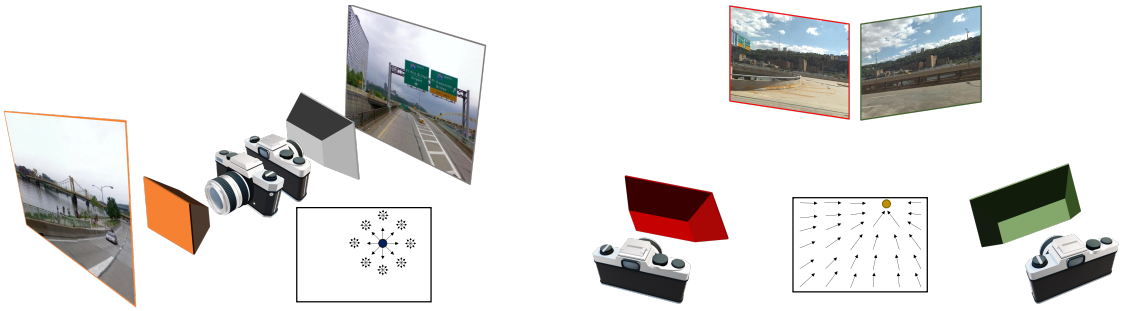

Figure 2: Two observations of the same place were previously considered matching if the physical distance between the observations was below a threshold. This meant that a VPR algorithm potentially needed to match two images that did not have any visual overlap because of a large orientation offset (left example). This definition was in line with place cells, which fire regardless of the orientation an animal is facing (left inset). We instead suggest that two observations should be matched based on their visual overlap – i.e. the left example would not be considered the same place, while the right example observations would be, despite the large physical distance between the locations (right). This is in line with spatial view cells, which fire whenever an animal gazes at a certain area (right inset).

crucial importance in tasks like localization and navigation, which in turn become ever more important with the advent of artificial intelligence (AI) in autonomous cars, mobile robots that interact with humans, and intelligent augmentation platforms such as the HoloLens 2. Inspiration for VPR is often drawn from the animal kingdom, given the remarkable localization and navigation capabilities of even "simple" animals, and the relatively well understood underlying mechanisms (even leading to 2014's Nobel Prize for the discovery of place cells and grid cells, as further detailed below).

Although it might seem inevitable to define a place first, we instead define VPR directly, remembering that it is a comparison of visual data, observed from same or different physical locations with same or different viewpoints. We argue that a reference and query observation can lead to successful recognition if there exists a certain degree of visual overlap[2] due to overlapping field-of-view of the underlying sensor, whereby the acceptable degree depends on the drivers introduced in Section 4. This implies that: 1) being at the same physical location is not sufficient, the orientation (i.e. viewpoint) needs to be somewhat similar as well, and 2) places can also be recognized when observed from distant physical locations (see Figure 2). In short, we define VPR as the ability to recognize one's location based on reference and query observations perceived from overlapping field-of-views. Note that our definition requires rethinking the typical notion of localization threshold (as used by almost all datasets and evaluation metrics, see Section 5), which is based on metric distances without considering orientation.

Our definition is complementary to that of Lowry *et al.*, but has a different underlying motivation. Lowry *et al.*'s definition is in line with the notion of place cells, which fire when an animal is in a particular place in the environment, irrespective of the animal's orientation. Instead, our definition is in line with spatial view cells, which fire when a specific area of the environment is gazed at by the animal, irrespective of the particular location [Georges-François *et al.*, 1999] (see Figure 2).

We note that in the context of robotics, VPR often involves sequential imagery (i.e. using multiple images that were captured over tens or hundreds of meters) rather than single images, as this can significantly improve place recognition performance, especially so for challenging environments [Lowry *et al.*, 2016]. For such sequence-based methods, the equivalent of visual overlap is the overlap of the volume spanned by the sequence.

3 Related Areas

In this section, we highlight similarities and differences of VPR with a handful of related areas of research. This section does not include downstream tasks such as visual SLAM, which are instead covered in Section 4.3 (see also Figure 1). While the relation to image retrieval has been discussed in other works [Lowry *et al.*, 2016; Zhang *et al.*, 2020; Masone and Caputo, 2021], it is for the first time that VPR's relation to video retrieval, visual landmark recognition, and overlap detection is systematically presented. We argue that for each of those areas, there is a potential for mutual benefits: research into VPR can offer insights for these areas and vice versa.

Image Retrieval: Image retrieval refers to the general problem of retrieving relevant images from a large database [Arandjelović *et al.*, 2017]. VPR is commonly cast as an image retrieval problem that involves a nearest neighbor search of compact global descriptors [Arandjelović *et al.*, 2017] or cross-matching of local descriptors [Liu *et al.*, 2021; Hausler *et al.*, 2021; Tourani *et al.*, 2021]. With regards to solving the nearest neighbor search problem, VPR and image-retrieval systems face similar challenges. However, the underlying goals differ between the two areas. For image retrieval, similarity criteria could be based on semantic categories such as 'clothes' as a product category or nighttime image as an environmental condition category. However, with the additional context of being a 'place' (see Section 2), VPR deviates from the process of merely retrieving a "similar" image, which instead is one of the challenges of VPR and referred to as *perceptual aliasing* (see Section 4). The notion of similarity in VPR is constrained

[2]Here visual overlap is defined as the common visibility of 3D points/regions in 2D image views. We do not consider VPR based on prior knowledge, e.g. scene completion [Hays and Efros, 2007; Song *et al.*, 2017], where visual overlap may not be needed.

to matching spatial information, where images captured from the *same* place would be considered a true match even if environmental conditions are dissimilar (e.g. day vs night).

Video Retrieval: The video retrieval problem is analogous to image retrieval, except that relevant videos are retrieved from the database leveraging sequential information. In VPR literature, sequence-based VPR is typically implemented as a decoupled approach, where single image-based retrieval is followed by sequence score aggregation [Ho and Newman, 2007; Milford and Wyeth, 2012]. The recent introduction of explicit sequence-based place representations (where the representation itself describes the sequence), posing VPR as a video retrieval problem, opens up new opportunities to obtain solutions robust to extreme appearance variations [Garg and Milford, 2021; Garg *et al.*, 2020a; Arroyo *et al.*, 2015].

Visual Landmark Recognition and Retrieval: Visual landmark recognition is the classification problem of given an image and a set of images belonging to a large variety of landmarks, deciding to which landmark this image belongs. Recently, the Google-Landmarks dataset [Weyand *et al.*, 2020] presented a new large-scale *instance-level* recognition and retrieval challenge, with the number of landmarks[3] increased from 30,000 to 200,000 in its second version. This large-scale recognition is an *extreme classification* problem [Choromanska *et al.*, 2013], where existing recognition solutions have relied on retrieval (nearest neighbor search) [Teichmann *et al.*, 2019]. Google-Landmarks comprises *specific* places (with the semantics of unique proper names) as opposed to general place categories (with the semantics of common names) [Zhou *et al.*, 2017; Wu *et al.*, 2009].

In contrast, VPR refers to the ability of distinctively recognizing *any ordinary* place or a region in the 3D world, thus posing an 'extremer' classification problem. It remains to be seen how methods developed for landmark recognition and retrieval can be leveraged in the context of VPR – recent advances include learning to aggregate 'relevant' landmarks [Teichmann *et al.*, 2019], as well as jointly training local and global descriptors [Cao *et al.*, 2020; Sarlin *et al.*, 2019].

Visual Overlap Detection: As discussed in Section 2, our definition of VPR is based on an overlapping field of view between the two places that should be matched; thus VPR and the area of visual overlap detection become more closely linked. The contrast between defining VPR using visual overlap as opposed to "positional offset" impacts the choice of ground truth for both training and evaluation procedures. This contrast has recently been shown to lead to noticeable changes in absolute performance when benchmarking localization algorithms [Pion *et al.*, 2020].

As the ground truth visual overlap might not be available for all datasets, overlap detection measures could be used as a supervision signal [Rau *et al.*, 2020; Chen *et al.*, 2020] to develop better VPR techniques. A noteworthy recent proposal on overlap detection [Rau *et al.*, 2020] introduced the 'normalized surface overlap' to measure the number of pixels of image A visible in image B (and vice versa). This leads to an asymmetric, but interpretable, measure that can also estimate the relative scale between pairs of images.

[3]In the context of mobile robotics, the term 'landmark' is typically used to indicate *any* specific visual entity in the scene relevant for localization [Luo *et al.*, 1992; Xin *et al.*, 2019].

4 What *Drives* VPR Research?

This section outlines the three key drivers of spatially intelligent systems, including intelligent autonomous systems operating in industry and household domains. As drivers, we refer to components that typically impose requirements on the system with regards to a) how the problem should be defined, b) how the solution (in the context of VPR: place representation and matching) should be designed, and c) how these solutions should be evaluated, both in terms of datasets and metrics. The three drivers are the *Environment* where an agent operates (Section 4.1), the *Agent* on which the spatially intelligent system is deployed (Section 4.2), and the *Downstream Task* that is performed (Section 4.3). In practice, different aspects of each of these drivers are simultaneously at play. We detail why it is crucial to understand the influence of these drivers to design better spatially intelligent systems, in particular in the VPR domain.

4.1 Environment

The first driver of VPR research is the operating environment, where research often branches out, as methods that work in certain environment types might cease to do so in other environment types. Differing branches include indoor vs outdoor, suburban vs highway, structured vs open, and human-made vs natural. The operating environment is often tightly coupled with the robotic agent choice (Section 4.2) – for example, driverless cars do not operate in office environments, or at least should not.

While the general aim of VPR systems is often stated to be invariance to changes in viewpoint *as well as* changes in appearance (including structural, seasonal, and illumination changes) [Lowry *et al.*, 2016; Arandjelović *et al.*, 2017; Garg *et al.*, 2018b; Zhang *et al.*, 2020], we argue that 1) not all environments/agents require invariance to both viewpoint and appearance (as detailed below), and 2) that there is a trade-off between viewpoint and appearance invariance achievable by current systems (as detailed in Section 6.1). Therefore, knowing the operating environment can provide crucial advantages when deciding how to represent and match places.

Structured Environments and Viewpoint Variations: In well-structured environments such as road infrastructure, the extent of 6-DoF viewpoint variations is generally confined, e.g. for driverless cars on roads, viewpoint varies mostly in the yaw direction [Maddern *et al.*, 2017]. Similar effects in viewpoint variations can be observed for other platforms, too. For example, as soon as aerial vehicles reach a certain height, one can assume a planar homography ("flat world"), simplifying template matching [Saurer *et al.*, 2016; Mount *et al.*, 2019]. Planar homographies are also present when mounting the camera at a fixed distance from the surface and pointing towards the surface. In structured indoor environments such as warehouses and offices, aisles and corridors enable Manhattan world assumptions and often simplify

Simultaneous Localization And Mapping (SLAM) [Li *et al.*, 2018].

Environment-Dependent Appearance Variations: Appearance invariance is similarly often constrained when assuming a certain operating environment. However, this kind of invariance is harder to quantify as changes in appearance can originate from a wide range of factors: Examples include changes in the time of the day, seasonal changes, structural changes, and weather changes. Therefore, while viewpoint change could be quantified by the metric shift in translation and rotation, there is no linear scale in the difficulty of appearance invariance [Zaffar *et al.*, 2021]. There are even some counter-intuitive examples, where a reference image captured outdoor in the morning might be easier to match to a well-lit nighttime image than to an image captured at noon which has shadows cast on a large area of the image [Corke *et al.*, 2013].

For different platforms and environments, the requirement of representing and matching places in an appearance-invariant or viewpoint-invariant manner can differ significantly. For example, driverless cars typically traverse a well-defined route and could trade-off viewpoint-invariance with appearance-invariance which can be relatively more challenging due to variations in the time of day, season, structural changes including roadworks and differing traffic conditions [Warburg *et al.*, 2020]. On the other hand, when an autonomous agent is deployed indoors or when considering an unmanned aerial vehicle, its route or maneuvers may not always be constrained, thus requiring viewpoint-invariance more than appearance-invariance.

Perceptual Aliasing: Another consideration that is tied to the operating environment is the extent of perceptual aliasing. Perceptual aliasing is the problem that two distinct places can look strikingly similar, often more similar than the same place observed under different conditions [Lowry *et al.*, 2016]. For example, indoor environments often contain corridors and hallways that are hard to distinguish. In outdoor settings, different places along a highway or a natural vegetative environment tend to be more perceptually aliased than different places within the man-made urban or suburban dwellings.

Dynamic Environments: Problems related to the operating environment that – to our knowledge – have not yet been addressed in VPR research are sensor dust, reflections (in glass or puddles) and undesired objects close to the camera (e.g. windscreen wipers). Such conditions are expected in challenging environments like mines and forests, which have come into focus in recent years [Nardari *et al.*, 2020; Garforth and Webb, 2020]. It would be interesting to model the impact of such 'noise' explicitly or measure the impact of sensor noise in existing VPR systems.

4.2 Agent

VPR has widespread applications and is thus deployed on a large variety of robotic platforms, including unmanned ground vehicles and autonomous cars [Doan *et al.*, 2019], unmanned aerial vehicles [Zaffar *et al.*, 2019] and unmanned underwater vehicles [Li *et al.*, 2015b]. Other platforms where VPR is applied are those tightly coupled to human users such as virtual/augmented reality devices [Sattler *et al.*, 2016] and mobile phones [Torii *et al.*, 2018].

Computational Resources: A robotic agent typically runs a large number of processes, many of them interacting with each other through tools like the Robot Operating System (ROS) [Quigley *et al.*, 2009; Fischer *et al.*, 2021]. These processes share limited onboard resources and often require cognitive architectures [Fischer *et al.*, 2018] to interact efficiently. Thus the resources dedicated to the VPR system might be relatively small, and a GPU (that significantly boosts inference times of deep networks) might not be available. Similarly, storage limitations could mean that the reference map of the operating environment (in the form of images, global/local descriptors, point clouds) has to be of reasonable size. Section 6 discusses some of the open problems in VPR in this context, for example, compact global description, efficient indexing and quantization, and hierarchical place matching pipelines.

Suitable Sensor Suite: Depending on the agent and the operating environment, robust VPR solutions can be developed by using additional suitable sensors. For example, event cameras perform exceptionally well when a high dynamic range is required, such as when exiting a dark tunnel and moving into bright sunlight [Fischer and Milford, 2020]. LIDAR-based systems can perceive the scene's geometry even in the most challenging nighttime conditions, although those systems lack appearance information [Guo *et al.*, 2019]. Using omnidirectional cameras or multi-camera rigs increase the field-of-view and thus the visual overlap, which results in reduced complexities in image matching.

Correct sensor type choice can also aid in tackling specific challenges such as nighttime conditions. Crucially, the sensor capabilities should drive the research regarding what characteristics are required in our learned descriptors. We believe that using novel sensor types such as 3D ultrasonic sensors (e.g. the Toposens TS3) and sensor fusion [Jacobson *et al.*, 2015] could further improve the robustness of VPR systems.

While some sensors can be a replacement for RGB cameras, another area worthy of more thorough investigation is the use of additional information in the form of prior position or egomotion. For example, one can assume that autonomous cars are equipped with a GPS sensor. Still, despite the popularity of datasets such as Oxford RobotCar [Maddern *et al.*, 2017] that contain GPS information, it is rarely used for VPR [Vysotska *et al.*, 2015] – using GPS information in environments where available could refocus research on GPS-denied environments like tunnels or underground mines that have distinct challenges. While there are many examples of GPS-denied environments, almost all mobile robots have some odometry information, but it has only been used in a limited manner for VPR [Pepperell *et al.*, 2014].

4.3 Downstream Task

Here, we consider the different tasks that an agent (robotic platform or intelligent augmentation) might perform. In this paper, we consider the VPR problem in the context of localization, although VPR does not strictly imply localization. For example, VPR can also be used to solve the problem of scene change detection where localization based image pairing is assumed [Park *et al.*, 2021], especially in the context of updating the map of the environment [Alcantarilla *et al.*, 2018; Tanaka, 2018].

It is also worth noting that localization does not strictly require VPR. For example, satellite-based (e.g. GPS) coarse position estimation can be fused with additional sensor information to enable 6-DoF visual SLAM in outdoor environments [Schneider *et al.*, 2016]. However, limitations of satellite-based pose estimation such as unavailability indoors, sensitivity to atmospheric changes, and signal obstruction in cluttered environments, deem vision-based alternatives such as VPR necessary.

Localization, SLAM and Kidnapped Robot Problem: A purely *topological* visual SLAM system[4] can be directly defined through VPR, which is highly relevant for large-scale mapping [Cummins and Newman, 2008; Doan *et al.*, 2019]. Such a topological SLAM system requires determining whether the currently observed place is a revisited one or is a new 'unseen' place, thus posing unique design requirements on VPR.

For metric or topo-metric SLAM systems[4], the final goal is typically to estimate the agent's 6-DoF pose, thus requirements of VPR systems vary in this case. For example, when VPR is used to provide a coarse estimate of the pose within a 6-DoF localization-only [Toft *et al.*, 2020] or SLAM algorithm [Mur-Artal *et al.*, 2015], the error bounds need to be very tight and the visual overlap between the two places relatively large with sufficient parallax. This is opposed to a scenario where loose error bounds are sufficient – a rough location estimate might sufficiently narrow down the search space for a subsequent laser-based pose estimation for global re-localization of a mobile robot ("kidnapped robot problem") [Jacobson *et al.*, 2021].

The requirements with regards to precision and recall are also varying for different scenarios. When using VPR for generating loop closures for SLAM (i.e. recognizing that a location has been visited previously, so that a globally consistent map can be built), incorrect matches can lead to catastrophic failures, thus requiring high precision VPR [Cadena *et al.*, 2016]. On the other hand, one could use VPR to select top k matches which are then passed to computationally more intensive stages for further processing; in this case, higher recall is more important than the precision. Thus, the downstream task is a key determining factor for formulating and evaluating VPR, as further discussed in Section 5.

Higher-level Tasks: The requirements of some downstream tasks like SLAM and Structure from Motion (SfM) are relatively well understood; yet, these requirements are very distinct and probably need a suitably tailored treatment. For example, SfM-based large-scale 3D reconstruction is typically performed offline [Schönberger and Frahm, 2016] and needs sub-pixel accurate alignment of images. The computational requirements of a VPR system then play a much lesser role than in real-time deployments on a mobile platform mapping an unknown environment using visual SLAM.

[4]Topological SLAM captures the connectivity of the environment rather than building a geometrically accurate map [Brooks, 1985]. This is opposed to metric SLAM, where geometrically accurate maps are built. Topo-metric SLAM systems build local maps that are geometrically accurate and then topologically connected to form the overall map [Cadena *et al.*, 2016].

The requirements of other "higher-level" tasks such as those of augmented reality platforms and navigation are not yet well established. This is in part due to the complex hierarchical nature of typical spatially intelligent systems, for example an augmented reality platform would involve many interrelated components such as image retrieval, sequential localization, local feature matching, visual odometry, and pose refinement [Stenborg *et al.*, 2020]. Furthermore, the utility of VPR and mapping for navigation purposes [Milford and Wyeth, 2007; Dall'Osto *et al.*, 2020] is a vastly unexplored area, and a deeper understanding of task requirements is needed.

5 How to Evaluate Visual Place Recognition?

This section discusses the evaluation datasets and metrics, in the context of the *drivers*.

5.1 Evaluation Datasets

There are numerous place recognition datasets, each covering different aspects of VPR (for recent overviews see [Warburg *et al.*, 2020; Masone and Caputo, 2021]). Thus some datasets are better suited to investigate specific configurations of proposed drivers (i.e. environments, agents and downstream tasks, see Section 4), while other datasets better represent different scenarios. This highlights the importance of clearly stating the application scenario targeted by a particular VPR system – it may be sufficient that the VPR system excels in datasets that are close to the actual use-case (but not in others).

Recent progress has enabled easier comparison of different methods. VPR-Bench [Zaffar *et al.*, 2021] provides a mechanism for the comparison of a new method on an extensive range of datasets. In the light of highly successful standard benchmark datasets in other research areas like visual object tracking [Kristan *et al.*, 2018], we believe that such benchmarking will accelerate VPR research. Mapillary Street Level Sequences (MSLS) [Warburg *et al.*, 2020] is a single dataset that tries to capture all variations of appearance/viewpoint change at once. MSLS notably also introduces 'sub-tasks' that can be separately investigated, including sub-tasks like summer to winter, day to night, and old to new. An additional benefit of MSLS is that it provides a hold-out test set that can be used for challenges.

If the aim is to design a VPR system applicable in all different scenarios, an open challenge is to design systems that are equally applicable indoors and outdoors. Few studies evaluate systems both indoors and outdoors, one of them being the above mentioned VPR-Bench [Zaffar *et al.*, 2021]. VPR-Bench has shown that performance trends can vary noticeably across environment types, e.g. indoor vs outdoor. However, care should be taken to not make generic assumptions about an architecture when the trained descriptors heavily depend on the training data – the training data should be representative of the data encountered at deployment time. Most recently, [Warburg *et al.*, 2020] have shown that training on more diverse data drastically improves performance on unseen data. This is distinct from the approach where different network configurations are explicitly trained for different scenarios, e.g. one for indoors and another for outdoors [Sarlin *et al.*, 2020].

5.2 Evaluation Metrics

The previous examples show that the downstream task and the relevant evaluation metrics are tightly coupled. However, we note that many VPR papers do not state why a particular evaluation metric was chosen. Notable exceptions include system papers where VPR is one of many components, and a specific downstream task is considered, such as [Cummins and Newman, 2008]. The computer vision community typically uses the Recall@k measure, which indicates that in this context the VPR system is benchmarked based on its ability to retrieve a correct match within the top-k retrievals regardless of the false matches. On the other hand, the mean average precision (mAP) metric [Philbin *et al.*, 2007], used in the image retrieval community, explicitly penalizes selection of false matches. The mAP metric could be adopted to measure VPR performance for SLAM-like downstream tasks (Section 4.3) where precision is more important, complementing measures like Recall at 100% Precision.

The area under the precision-recall curve and the F-score are sometimes used as summary statistics [Molloy *et al.*, 2021]. However, their practical use is unclear, as these summary statistics imply that recall and precision are of similar importance, which is unlikely the case for most downstream tasks. Moreover, these measures are based on the distribution of match scores which may only be relevant for topological SLAM-like scenarios where VPR needs to be highly precise and no subsequent outlier rejection method is employed.

Most of the VPR datasets in robotics are in the form of trajectories with inherent sequential information (Section 2). Thus, evaluation metrics such as 'maximum open-loop distance traveled' (that is, the extent of visual odometry or dead reckoning based robot motion without loop closures) have also been considered in the literature [Clement *et al.*, 2020]. We believe it would be beneficial to investigate metrics that tightly couple single-image and sequence-based VPR.

6 What Are Open Research Problems?

This section aims to highlight open research problems, considering the drivers discussed in Section 4. For space reasons and to avoid duplication, we do not cover the open research problems discussed in recent surveys on deep learning methods for VPR [Zhang *et al.*, 2020; Masone and Caputo, 2021], which include using autoencoders as an alternative to Convolutional Neural Networks (CNNs), use of generative methods including Generative Adversarial Networks (GANs), using semantic information, making use of heterogeneous data including multi-sensory fusion, and the choice of loss function.

Here, we broadly classify the open research problems into: 1) representation, discussing the need for better global descriptors and enriched/synthesized reference maps, and 2) matching, discussing the need for better hierarchical matching frameworks, relevant distance metrics and 'learning to match'.

6.1 Place Representation

Global Descriptors – Appearance & Viewpoint Invariance: Section 4.1 discussed the requirements on viewpoint and appearance invariance depending on the operating environment. Here we note that there is a trade-off when learning a descriptor of a fixed size/type: increasing viewpoint-invariance will inevitably reduce some degree of appearance invariance (assuming the same amount of training data) [Arandjelović *et al.*, 2017; Chen *et al.*, 2017; Garg *et al.*, 2019]. This is evident from significant differences observed in place recognition performance when considering a cross-combination of datasets such as Nordland (same view, varying appearance) [Sünderhauf *et al.*, 2013] and Pittsburgh (similar appearance, varying view) [Arandjelović *et al.*, 2017] with feature learning/aggregation methods such as HybridNet (viewpoint-assumed) [Chen *et al.*, 2017] and NetVLAD (viewpoint-agnostic) [Arandjelović *et al.*, 2017].

There is a vast research gap, and a need for global description mechanisms that go beyond the binary nature of encoding viewpoint, that is, viewpoint-assumed vs viewpoint-invariant. This might be achieved by learning novel ways to incorporate local geometric information in the global descriptor formation, such as using vertical blocks (Stixels) [Hernandez-Juarez *et al.*, 2019], semantic blobs [Gawel *et al.*, 2018], objects [Qin *et al.*, 2021] or superpixels [Neubert *et al.*, 2015], where learning could be based on attention mechanisms such as that employed in Transformers [Vaswani *et al.*, 2017] and Graph Neural Networks [Veličković *et al.*, 2018].

Global Descriptors – Efficiency: Most of the state-of-the-art global image descriptors are high-dimensional (with dimensions varying from 512 [Sarlin *et al.*, 2019] to 70,000 [Sünderhauf *et al.*, 2015a]). Increasing the descriptor's dimensionality directly leads to increased computational requirements. To improve efficiency, researchers commonly employ dimension-reduction methods such as Principal Component Analysis (PCA), and have also explored quantization [Jegou *et al.*, 2010; Brandt, 2010; Ge *et al.*, 2013], binarization [Lowry and Andreasson, 2018; Arroyo *et al.*, 2015; Jegou *et al.*, 2008], hashing [Vysotska and Stachniss, 2017; Gionis *et al.*, 1999; Andoni and Indyk, 2006] and efficient indexing [Cao *et al.*, 2020] techniques.

However, there have not been any attempts to learn these efficiency-inducing processes for VPR, particularly considering that retrieving places can include additional information in the form of sequences or odometry. This could be achieved by learning to reduce dimensions [McInnes *et al.*, 2018; Amid and Warmuth, 2019] or to hash [Wang *et al.*, 2017], while maintaining the overall structure of the appearance space learnt through the existing global descriptor methods. Existing efficient VPR techniques consider sequential or odometry information in a decoupled manner [Vysotska and Stachniss, 2017; Garg and Milford, 2020] but could benefit from jointly considering additional information when optimizing for efficiency.

Enriched Reference Maps: With the rapid increase in data gathering, more so in the field of autonomous driving, it is high time to consider the use of an enriched reference map, which could be in the form of multiple reference images per location [Churchill and Newman, 2012] or semantically-annotated 3D maps [Garg *et al.*, 2020b]. In the simplest case, choosing the best reference set can lead to vast performance improvements. More sophisticated approaches fuse multiple reference sets to achieve even better performance [Churchill and Newman, 2012; Linegar *et al.*, 2015;

Molloy *et al.*, 2021]. Multiple reference sets are often used when long-term autonomy is required, as structural changes can be detected and incorporated over time. While significant progress using multiple reference maps has been made in the past, open questions remain around the increased storage and computational requirements when using multiple reference sets. There is clearly a trade-off, and preliminary efforts in that direction [Doan *et al.*, 2019] need further attention.

View Synthesis: To deal with significant viewpoint variations, researchers have also explored matching through multiple synthesized views of the observed places [Torii *et al.*, 2018; Milford *et al.*, 2015], although the process requires additional computation. However, this can be mitigated by performing view synthesis offline during the mapping traverse and using compact global descriptors with an efficient nearest neighbor search for retrieval. The current bottleneck of automatically generating accurate and relevant views can potentially be addressed by recent advances in volumetric rendering [Mildenhall *et al.*, 2020], performed offline to generate novel views under novel lighting conditions.

6.2 Place Matching

Mutually-informed Hierarchical Systems: Different downstream tasks can impose very different requirements on the VPR system (see Section 4.3). For example, a visual SLAM system can be built *with* [Cummins and Newman, 2011] or *without* [Angeli *et al.*, 2009; Cummins and Newman, 2008] using a geometric verification step based on local feature matching. Cummins and Newman found this verification to be particularly essential for large datasets. However, it remains an open question how an effective hierarchical system should be designed, where the variables are: the number of complementary VPR techniques fused [Hausler *et al.*, 2019], number of stages, and types of unique methods involved, e.g., query expansion [Chum *et al.*, 2011] and keypoint filtering [Garg *et al.*, 2018b].

The complementarity and information transfer within different stages in the hierarchy requires an in-depth investigation. This can reveal answers to several overarching questions: should these stages always operate independently or could earlier stages better *inform* subsequent stages (beyond simply providing candidate images); could one selectively apply a subset of the techniques to save computational resources; how such behavior of hierarchical retrieval can be learnt; and in doing so do some of the stages become redundant.

Choice of Distance Metric: When comparing the global descriptors of two images, one has to choose a suitable distance metric or a similarity measure. Some of the most commonly employed measures include Euclidean [Arandjelović *et al.*, 2017; Sünderhauf *et al.*, 2015b], cosine [Sünderhauf *et al.*, 2015a; Garg *et al.*, 2018a; Garg *et al.*, 2018b], and Hamming [Lowry and Andreasson, 2018; Neubert *et al.*, 2019] distance. While some descriptors are better matched using one distance than the other, the range of distances distribution is typically relatively narrow, even in non-matching images of a completely different appearance. Therefore a more systematic investigation considering both a theoretical viewpoint and practical performance implications is needed. In particular, important consideration factors include suitability to loss functions (e.g. max-margin triplet loss) [Arandjelović *et al.*, 2017; Revaud *et al.*, 2019; Garg and Milford, 2021], descriptor normalization [Arandjelovic and Zisserman, 2013; Garg *et al.*, 2018a; Schubert *et al.*, 2020], whitening [Jégou and Chum, 2012; Arandjelovic and Zisserman, 2013], feature scaling [Beatty and Manjunath, 1997; Li *et al.*, 2015a], and quantization/binarization [Jegou *et al.*, 2010; Lowry and Andreasson, 2018].

Learning to Match: While learning to 'describe' (i.e. local or global descriptors) has been widely explored, there have been limited attempts to learn to 'match'. Such matchers can either be learnt through siamese networks [Altwaijry *et al.*, 2016] or cross-attention based on graph neural networks [Sarlin *et al.*, 2020]. The outcome of such matcher could either be a matched reference index or relative 3D pose [Gridseth and Barfoot, 2020]. Learning-to-match frameworks for VPR and localization could potentially eradicate the need for sophisticated matching pipelines.

7 Conclusions

This survey defines VPR based on the visual overlap of two observations, in line with spatial view cells found in primates. This new definition enabled us to discuss how VPR closely relates to other research areas. This paper also identified and detailed three key drivers of Spatial AI: the environment, agent and downstream task. Considering these drivers, we then discussed numerous open research problems that we think are worth addressing in future VPR research.

To date, VPR research has addressed the problems of representing, associating (matching), and searching of *spatial data*, and is a key enabler of Spatial AI. Further advances in VPR research will require unifying the efforts of the artificial intelligence, computer vision, robotics, and machine learning communities, particularly taking into account embodied agents. To achieve this, an in-depth understanding of the problem, research goals and evaluation protocols is necessary, and this paper takes a step in that direction.

Acknowledgments

This work received funding from the Australian Government, via grant AUSMURIB000001 associated with ONR MURI grant N00014-19-1-2571. The authors acknowledge continued support from the Queensland University of Technology (QUT) through the Centre for Robotics.

References

[Alcantarilla *et al.*, 2018] Pablo F Alcantarilla, et al. Street-view change detection with deconvolutional networks. *Autonomous Robots*, 42(7):1301–1322, 2018.

[Altwaijry *et al.*, 2016] Hani Altwaijry, et al. Learning to match aerial images with deep attentive architectures. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3539–3547, 2016.

[Amid and Warmuth, 2019] Ehsan Amid and Manfred K Warmuth. TriMap: Large-scale dimensionality reduction using triplets. *arXiv preprint arXiv:1910.00204*, 2019.

[Andoni and Indyk, 2006] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *IEEE Symposium Foundations Computer Science*, pages 459–468, 2006.

[Angeli *et al.*, 2009] Adrien Angeli, et al. Visual topological SLAM and global localization. In *IEEE Int. Conf. Robot. Autom.*, pages 4300–4305, 2009.

[Arandjelovic and Zisserman, 2013] Relja Arandjelovic and Andrew Zisserman. All about VLAD. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1578–1585, 2013.

[Arandjelović *et al.*, 2017] Relja Arandjelović, et al. NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1437–1451, 2017.

[Arroyo *et al.*, 2015] Roberto Arroyo, et al. Towards life-long visual localization using an efficient matching of binary sequences from images. In *IEEE Int. Conf. Robot. Autom.*, pages 6328–6335, 2015.

[Beatty and Manjunath, 1997] Morris Beatty and BS Manjunath. Dimensionality reduction using multi-dimensional scaling for content-based retrieval. In *IEEE Int. Conf. Image Process.*, volume 2, pages 835–838, 1997.

[Brandt, 2010] Jonathan Brandt. Transform coding for fast approximate nearest neighbor search in high dimensions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1815–1822, 2010.

[Brooks, 1985] Rodney Brooks. Visual map making for a mobile robot. In *IEEE Int. Conf. Robot. Autom.*, volume 2, pages 824–829, 1985.

[Cadena *et al.*, 2016] Cesar Cadena, et al. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Trans. Robot.*, 32(6):1309–1332, 2016.

[Cao *et al.*, 2020] Bingyi Cao, et al. Unifying deep local and global features for efficient image search. In *Eur. Conf. Comput. Vis.*, pages 726–743, 2020.

[Chen *et al.*, 2017] Zetao Chen, et al. Deep learning features at scale for visual place recognition. In *IEEE Int. Conf. Robot. Autom.*, pages 3223–3230, 2017.

[Chen *et al.*, 2020] Xieyuanli Chen, et al. OverlapNet: Loop closing for LiDAR-based SLAM. In *Robotics: Science and Systems*, 2020.

[Choromanska *et al.*, 2013] Anna Choromanska, et al. Extreme multi class classification. In *NeurIPS Workshop: eXtreme Classification*, 2013.

[Chum *et al.*, 2011] Ondřej Chum, et al. Total recall II: Query expansion revisited. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 889–896, 2011.

[Churchill and Newman, 2012] Winston Churchill and Paul Newman. Practice makes perfect? Managing and leveraging visual experiences for lifelong navigation. In *IEEE Int. Conf. Robot. Autom.*, pages 4525–4532, 2012.

[Clement *et al.*, 2020] Lee Clement, et al. Learning matchable image transformations for long-term metric visual localization. *IEEE Robot. Autom. Lett.*, 5(2):1492–1499, 2020.

[Corke *et al.*, 2013] Peter Corke, et al. Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, pages 2085–2092, 2013.

[Cummins and Newman, 2008] Mark Cummins and Paul Newman. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *Int. J. Robot. Res.*, 27(6):647–665, 2008.

[Cummins and Newman, 2011] Mark Cummins and Paul Newman. Appearance-only SLAM at large scale with FAB-MAP 2.0. *Int. J. Robot. Res.*, 30(9):1100–1123, 2011.

[Dall'Osto *et al.*, 2020] Dominic Dall'Osto, et al. Fast and robust bio-inspired teach and repeat navigation. *arXiv preprint arXiv:2010.11326*, 2020.

[Davison, 2018] Andrew J Davison. FutureMapping: The Computational Structure of Spatial AI Systems. *arXiv preprint arXiv:1803.11288*, 2018.

[Doan *et al.*, 2019] Anh-Dzung Doan, et al. Scalable place recognition under appearance change for autonomous driving. In *Int. Conf. Comput. Vis.*, pages 9319–9328, 2019.

[Fischer and Milford, 2020] Tobias Fischer and Michael Milford. Event-based visual place recognition with ensembles of temporal windows. *IEEE Robot. Autom. Lett.*, 5(4):6924–6931, 2020.

[Fischer *et al.*, 2018] Tobias Fischer, et al. iCub-HRI: A software framework for complex human-robot interaction scenarios on the iCub humanoid robot. *Front. Robot. AI*, 5(22):1–9, 2018.

[Fischer *et al.*, 2021] Tobias Fischer, et al. RoboStack: Using the Robot Operating System alongside the Conda and Jupyter Data Science Ecosystems. *arXiv preprint arXiv:2104.12910*, 2021.

[Garforth and Webb, 2020] James Garforth and Barbara Webb. Lost in the woods? Place recognition for navigation in difficult forest environments. *Front. Robot. AI*, 7, 2020.

[Garg and Milford, 2020] Sourav Garg and Michael Milford. Fast, compact and highly scalable visual place recognition through sequence-based matching of overloaded representations. In *IEEE Int. Conf. Robot. Autom.*, pages 3341–3348, 2020.

[Garg and Milford, 2021] Sourav Garg and Michael Milford. SeqNet: Learning descriptors for sequence-based hierarchical place recognition. *IEEE Robot. Autom. Lett.*, 6(3):4305–4312, 2021.

[Garg *et al.*, 2018a] Sourav Garg, et al. Don't look back: Robustifying place categorization for viewpoint- and condition-invariant place recognition. In *IEEE Int. Conf. Robot. Autom.*, pages 3645–3652, 2018.

[Garg *et al.*, 2018b] Sourav Garg, et al. LoST? appearance-invariant place recognition for opposite viewpoints using visual semantics. In *Robotics: Science and Systems*, 2018.

[Garg *et al.*, 2019] Sourav Garg, et al. Semantic–geometric visual place recognition: a new perspective for reconciling opposing views. *Int. J. Robot. Res.*, page 0278364919839761, 2019.

[Garg *et al.*, 2020a] Sourav Garg, et al. Delta descriptors: Change-based place representation for robust visual localization. *IEEE Robot. Autom. Lett.*, 5(4):5120–5127, 2020.

[Garg *et al.*, 2020b] Sourav Garg, et al. Semantics for robotic mapping, perception and interaction: A survey. *Found. Trends Robot.*, 8(1–2):1–224, 2020.

[Gawel *et al.*, 2018] Abel Gawel, et al. X-view: Graph-based semantic multi-view localization. *IEEE Robot. Autom. Lett.*, 3(3):1687–1694, 2018.

[Ge *et al.*, 2013] Tiezheng Ge, et al. Optimized product quantization for approximate nearest neighbor search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2946–2953, 2013.

[Georges-François *et al.*, 1999] Pierre Georges-François, et al. Spatial View Cells in the Primate Hippocampus: Allocentric View not Head Direction or Eye Position or Place. *Cerebral Cortex*, 9(3):197–212, 1999.

[Gionis *et al.*, 1999] Aristides Gionis, et al. Similarity search in high dimensions via hashing. In *Int. Conf. Very Large Data Bases*, pages 518–529, 1999.

[Gridseth and Barfoot, 2020] Mona Gridseth and Timothy D Barfoot. DeepMEL: Compiling visual multi-experience localization into a deep neural network. In *IEEE Int. Conf. Robot. Autom.*, pages 1674–1681, 2020.

[Guo *et al.*, 2019] Jiadong Guo, et al. Local descriptor for robust place recognition using LiDAR intensity. *IEEE Robot. Autom. Lett.*, 4(2):1470–1477, 2019.

[Hausler *et al.*, 2019] Stephen Hausler, et al. Multi-process fusion: Visual place recognition using multiple image processing methods. *IEEE Robot. Autom. Lett.*, 4(2):1924–1931, 2019.

[Hausler *et al.*, 2021] Stephen Hausler, et al. Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[Hays and Efros, 2007] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Trans. Graph.*, 26(3):4–es, 2007.

[Hernandez-Juarez *et al.*, 2019] Daniel Hernandez-Juarez, et al. Slanted stixels: A way to represent steep streets. *Int. J. Comput. Vis.*, 127(11-12):1643–1658, 2019.

[Ho and Newman, 2007] Kin Leong Ho and Paul Newman. Detecting loop closure with scene sequences. *Int. J. Comput. Vis.*, 74(3):261–286, 2007.

[Jacobson *et al.*, 2015] Adam Jacobson, et al. Autonomous multisensor calibration and closed-loop fusion for SLAM. *J. Field. Robot.*, 32(1):85–122, 2015.

[Jacobson *et al.*, 2021] Adam Jacobson, et al. What localizes beneath: A metric multisensor localization and mapping system for autonomous underground mining vehicles. *J. Field. Robot.*, 38(1):5–27, 2021.

[Jégou and Chum, 2012] Hervé Jégou and Ondřej Chum. Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening. In *Eur. Conf. Comput. Vis.*, pages 774–787, 2012.

[Jegou *et al.*, 2008] Herve Jegou, et al. Hamming embedding and weak geometric consistency for large scale image search. In *Eur. Conf. Comput. Vis.*, pages 304–317, 2008.

[Jegou *et al.*, 2010] Herve Jegou, et al. Product quantization for nearest neighbor search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1):117–128, 2010.

[Kristan *et al.*, 2018] Matej Kristan, et al. The sixth visual object tracking VOT2018 challenge results. In *Eur. Conf. Comput. Vis. Worksh.*, pages 3–53, 2018.

[Li *et al.*, 2015a] Dong Li, et al. A feature-scaling-based k-nearest neighbor algorithm for indoor positioning systems. *IoT J.*, 3(4):590–597, 2015.

[Li *et al.*, 2015b] Jie Li, et al. High-level visual features for underwater place recognition. In *IEEE Int. Conf. Robot. Autom.*, pages 3652–3659, 2015.

[Li *et al.*, 2018] Haoang Li, et al. A monocular SLAM system leveraging structural regularity in Manhattan world. In *IEEE Int. Conf. Robot. Autom.*, pages 2518–2525, 2018.

[Linegar *et al.*, 2015] Chris Linegar, et al. Work smart, not hard: Recalling relevant experiences for vast-scale but time-constrained localisation. In *IEEE Int. Conf. Robot. Autom.*, pages 90–97, 2015.

[Liu *et al.*, 2021] Dongfang Liu, et al. DenserNet: Weakly supervised visual localization using multi-scale feature aggregation. In *AAAI*, 2021.

[Lowry and Andreasson, 2018] Stephanie Lowry and Henrik Andreasson. Lightweight, viewpoint-invariant visual place recognition in changing environments. *IEEE Robot. Autom. Lett.*, 3(2):957–964, 2018.

[Lowry *et al.*, 2016] Stephanie Lowry, et al. Visual place recognition: A survey. *IEEE Trans. Robot.*, 32(1):1–19, 2016.

[Luo *et al.*, 1992] Ren C Luo, et al. Neural network based landmark recognition for robot navigation. In *Int. Conf. Ind. Electron. Control Instrum. Autom.*, pages 1084–1088, 1992.

[Maddern *et al.*, 2017] Will Maddern, et al. 1 Year, 1000km: The Oxford RobotCar dataset. *Int. J. Robot. Res.*, 36(1):3–15, 2017.

[Masone and Caputo, 2021] Carlo Masone and Barbara Caputo. A survey on deep visual place recognition. *IEEE Access*, 9:19516–19547, 2021.

[McInnes *et al.*, 2018] Leland McInnes, et al. UMAP: Uniform manifold approximation and projection. *J. Open Source Softw.*, 3(29), 2018.

[Mildenhall *et al.*, 2020] Ben Mildenhall, et al. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Eur. Conf. Comput. Vis.*, pages 405–421, 2020.

[Milford and Wyeth, 2007] Michael Milford and Gordon Wyeth. Spatial mapping and map exploitation: a bio-inspired engineering perspective. In *Int. Conf. Spatial Inf. Theory*, pages 203–221, 2007.

[Milford and Wyeth, 2012] Michael J. Milford and Gordon F. Wyeth. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *IEEE Int. Conf. Robot. Autom.*, pages 1643–1649, 2012.

[Milford *et al.*, 2015] Michael Milford, et al. Sequence searching with deep-learnt depth for condition-and viewpoint-invariant route-based place recognition. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 18–25, 2015.

[Molloy *et al.*, 2021] Timothy L. Molloy, et al. Intelligent reference curation for visual place recognition via Bayesian selective fusion. *IEEE Robot. Autom. Lett.*, 6(2):588–595, 2021.

[Mount *et al.*, 2019] James Mount, et al. Automatic coverage selection for surface-based visual localization. *IEEE Robot. Autom. Lett.*, 4(4):3900–3907, 2019.

[Mur-Artal *et al.*, 2015] Raul Mur-Artal, et al. Orb-slam: a versatile and accurate monocular slam system. *IEEE Trans. Robot.*, 31(5):1147–1163, 2015.

[Nardari *et al.*, 2020] Guilherme V Nardari, et al. Place recognition in forests with urquhart tessellations. *IEEE Robot. Autom. Lett.*, 6(2):279–286, 2020.

[Neubert *et al.*, 2015] Peer Neubert, et al. Superpixel-based appearance change prediction for long-term navigation across seasons. *Robotics and Autonomous Systems*, 69:15–27, 2015.

[Neubert *et al.*, 2019] Peer Neubert, et al. A neurologically inspired sequence processing model for mobile robot place recognition. *IEEE Robot. Autom. Lett.*, 4(4):3200–3207, 2019.

[Park *et al.*, 2021] Jin-Man Park, et al. Changesim: Towards end-to-end online scene change detection in industrial indoor environments. *arXiv preprint arXiv:2103.05368*, 2021.

[Pepperell *et al.*, 2014] Edward Pepperell, et al. All-environment visual place recognition with smart. In *IEEE Int. Conf. Robot. Autom.*, pages 1612–1618, 2014.

[Philbin *et al.*, 2007] James Philbin, et al. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1–8, 2007.

[Pion *et al.*, 2020] Noé Pion, et al. Benchmarking image retrieval for visual localization. In *3DV*, 2020.

[Qin *et al.*, 2021] Cao Qin, et al. Semantic loop closure detection based on graph matching in multi-objects scenes. *J. Vis. Comm. Image Rep.*, page 103072, 2021.

[Quigley *et al.*, 2009] Morgan Quigley, et al. ROS: an open-source Robot Operating System. In *IEEE Int. Conf. Robot. Autom. Worksh.*, 2009.

[Rau *et al.*, 2020] Anita Rau, et al. Predicting visual overlap of images through interpretable non-metric box embeddings. In *Eur. Conf. Comput. Vis.*, pages 629–646, 2020.

[Revaud *et al.*, 2019] Jerome Revaud, et al. Learning with average precision: Training image retrieval with a listwise loss. In *Int. Conf. Comput. Vis.*, pages 5107–5116, 2019.

[Sarlin *et al.*, 2019] Paul-Edouard Sarlin, et al. From coarse to fine: Robust hierarchical localization at large scale. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12716–12725, 2019.

[Sarlin *et al.*, 2020] Paul-Edouard Sarlin, et al. SuperGlue: Learning feature matching with graph neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4938–4947, 2020.

[Sattler *et al.*, 2016] Torsten Sattler, et al. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(9):1744–1756, 2016.

[Saurer *et al.*, 2016] Olivier Saurer, et al. Homography based egomotion estimation with a common direction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(2):327–341, 2016.

[Schneider *et al.*, 2016] Johannes Schneider, et al. Fast and effective online pose estimation and mapping for uavs. In *IEEE Int. Conf. Robot. Autom.*, pages 4784–4791, 2016.

[Schönberger and Frahm, 2016] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4104–4113, 2016.

[Schubert *et al.*, 2020] Stefan Schubert, et al. Unsupervised learning methods for visual place recognition in discretely and continuously changing environments. In *IEEE Int. Conf. Robot. Autom.*, pages 4372–4378, 2020.

[Song *et al.*, 2017] Shuran Song, et al. Semantic scene completion from a single depth image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1746–1754, 2017.

[Stenborg *et al.*, 2020] Erik Stenborg, et al. Using image sequences for long-term visual localization. In *3DV*, pages 938–948, 2020.

[Sünderhauf *et al.*, 2013] Niko Sünderhauf, et al. Are we there yet? challenging SeqSLAM on a 3000 km journey across all four seasons. In *IEEE Int. Conf. Robot. Autom. Worksh.*, 2013.

[Sünderhauf *et al.*, 2015a] Niko Sünderhauf, et al. On the performance of ConvNet features for place recognition. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, pages 4297–4304, 2015.

[Sünderhauf *et al.*, 2015b] Niko Sünderhauf, et al. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. In *Robotics: Science and Systems*, 2015.

[Tanaka, 2018] Kanji Tanaka. Simultaneous localization and change detection for long-term map learning: A scalable scene retrieval approach. In *IEEE Int. Conf. Soft Computing Intell. Syst. and Int. Symposium Adv. Intell. Syst.*, pages 1039–1045, 2018.

[Teichmann *et al.*, 2019] Marvin Teichmann, et al. Detect-to-retrieve: Efficient regional aggregation for image search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5109–5118, 2019.

[Toft *et al.*, 2020] Carl Toft, et al. Long-term visual localization revisited. *IEEE Trans. Pattern Anal. Mach. Intell.*, to appear, 2020.

[Torii *et al.*, 2018] Akihiko Torii, et al. 24/7 place recognition by view synthesis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(2):257–271, 2018.

[Tourani *et al.*, 2021] Satyajit Tourani, et al. Early bird: Loop closures from opposing viewpoints for perceptually-aliased indoor environments. In *Int. Conf. Comput. Vis. Imaging Comput. Graph. Theory Applications*, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, et al. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, pages 5998–6008, 2017.

[Veličković *et al.*, 2018] Petar Veličković, et al. Graph attention networks. In *Int. Conf. Learn. Represent.*, 2018.

[Vysotska and Stachniss, 2017] Olga Vysotska and Cyrill Stachniss. Relocalization under substantial appearance changes using hashing. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. Worksh.*, 2017.

[Vysotska *et al.*, 2015] Olga Vysotska, et al. Efficient and effective matching of image sequences under substantial appearance changes exploiting gps priors. In *IEEE Int. Conf. Robot. Autom.*, pages 2774–2779, 2015.

[Wang *et al.*, 2017] Jingdong Wang, et al. A survey on learning to hash. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):769–790, 2017.

[Warburg *et al.*, 2020] Frederik Warburg, et al. Mapillary street-level sequences: A dataset for lifelong place recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2626–2635, 2020.

[Weyand *et al.*, 2020] Tobias Weyand, et al. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2575–2584, 2020.

[Wu *et al.*, 2009] Jianxin Wu, et al. Visual place categorization: Problem, dataset, and algorithm. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, pages 4763–4770, 2009.

[Xin *et al.*, 2019] Zhe Xin, et al. Localizing discriminative visual landmarks for place recognition. In *IEEE Int. Conf. Robot. Autom.*, pages 5979–5985, 2019.

[Zaffar *et al.*, 2019] Mubariz Zaffar, et al. Are state-of-the-art visual place recognition techniques any good for aerial robotics? In *IEEE Int. Conf. Robot. Autom. Worksh.*, 2019.

[Zaffar *et al.*, 2021] Mubariz Zaffar, et al. Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. *Int. J. Comput. Vis.*, pages 1–39, 2021.

[Zhang *et al.*, 2020] Xiwu Zhang, et al. Visual place recognition: A survey from deep learning perspective. *Pattern Recognition*, 2020.

[Zhou *et al.*, 2017] Bolei Zhou, et al. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2017.