Noname manuscript No.
(will be inserted by the editor)

VPR-Bench

An Open-Source Visual Place Recognition Evaluation Framework with Quantifiable Viewpoint and Appearance Change.

Mubariz Zaffar · Shoaib Ehsan · Michael Milford · David Flynn · Klaus McDonald-Maier

Please note that this is a pre-print version and the paper is currently under-review. Open-source code and further improvements to follow in a few months based on the community and reviewers feedback.

Keywords Visual Place Recognition · SLAM · Autonomous Robotics · Robotic Vision

Abstract Visual Place Recognition (VPR) is the process of recognising a previously visited place using visual information, often under varying appearance conditions and viewpoint changes and with computational constraints. VPR is related to concepts of localisation, loop closure and is a critical component of many autonomous navigation systems ranging from autonomous vehicles to drones. While the concept of place recognition has been around for many years, visual place recognition research has grown rapidly as a field over the past decade due to both improving camera hardware technologies and its suitability for application of deep learning-based techniques. With this growth however has come field fragmentation and a lack of standardisation especially with respect to evaluation, and a disconnect between current performance metrics and the actual utility of a VPR technique when deployed in applications. In this paper we address these key challenges through a new comprehensive open-source framework for assessing the performance of VPR techniques, dubbed VPR-Bench. VPR-Bench introduces two much-needed capabilities for researchers: firstly, a framework for quantifying viewpoint and illumination variation, replacing what has largely been assessed qualitatively in the past, and secondly, new metrics Extended precision (EP), Performance-Per-Compute-Unit (PCU) and Number of Prospective Place Matching Candidates (NPPMC). These new metrics complement the limitations of traditional Precision-Recall curves and AUC measures, by providing measures that are more informative to the wide range of potential VPR applications that vary in requirements with respect to required precision or recall levels and that relate performance to computational requirements. Mechanistically, we develop new unified templates that facilitate the implementation, deployment and evaluation of a wide range of VPR techniques and datasets. We incorporate the most comprehensive combination of state-of-

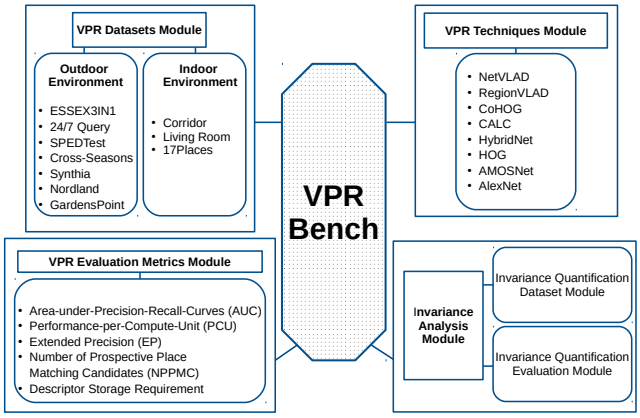

Fig. 1 A block-diagram overview of the developed VPR-Bench framework is shown here. All modules can be interlinked within the framework and can also be independently modified for graceful updates in the future.

Mubariz Zaffar, Shoaib Ehsan and Klaus McDonald-Maier
School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ, United Kingdom
E-mail: mubariz.zaffar,sehsan,kdm@essex.ac.uk

Michael Milford
School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4000, Australia
E-mail: michael.milford@qut.edu.au

David Flynn
School of Engineering and Physical Sciences, Smart Systems Group, Heriot-Watt University, Edinburgh, Currie EH14 4AS, United Kingdom
E-mail: D.Flynn@hw.ac.uk

the-art VPR techniques and datasets to date into VPR-Bench and demonstrate how it provides a rich range of previously inaccessible insights both with respect to techniques as well as benchmark datasets, such as the nuanced relationship between viewpoint invariance, different techniques and different types of VPR datasets.

1 Introduction

Visual Place Recognition (VPR) is a challenging and a widely investigated problem within the computer vision community (Lowry et al. (2015)). It identifies the ability of a system to match a previously visited place using onboard computer vision prowess, with resilience to perceptual aliasing, seasonal-, illumination- and viewpoint-variations. This ability to correctly and efficiently recall previously seen places using only visual input has several important applications. A key application lies in loop-closure to correct error drifts in a SLAM (Simultaneous Localisation and Mapping) pipeline (Cadena et al. (2016)). The applications of VPR systems extend to several other domains that utilise computer vision modules, e.g., image search based on visual content (Tolias et al. (2016a)), location-refinement given human-machine interfaces (Robertson and Cipolla (2004)), query-expansion (Johns and Yang (2011)), improved representations (Tolias et al. (2013)), vehicular navigation (Fraundorfer et al. (2007)), asset-management using aerial imagery (Odo et al. (2020)) and 3D-model creation (Agarwal et al. (2011)).

Consequently, researchers working within VPR come from various backgrounds, and some of the top robotics and computer vision groups across the world have dedicated their resources to investigating this problem. Several workshops have been organised in top-tier conferences, including but not limited to, 'Long-Term Visual Localisation Workshop Series' in Computer Vision and Pattern Recognition Conference (CVPR), 'Visual Place Recognition in Changing Environments Workshop Series' in IEEE International Conference on Robotics and Automation (ICRA), 'Large-Scale Visual Place Recognition and Image-Based Localization Workshop' in IEEE International Conference on Computer Vision (ICCV 2019) and 'Visual Localisation: Features-based vs Learning Approaches' in European Conference on Computer Vision (ECCV 2018).

Due to the multi-domain application nature of VPR, the salience of the problem (and its challenges), advances in deep-learning-based computer vision and the minimal hardware requirements for investigation; VPR has drawn huge interest from the research community, leading to a large number of VPR techniques proposed over the past many years. All of these techniques have claimed state-of-the-art performance, however, due to the large variety of evaluation datasets, difference of metrics employed for evaluation and the limited comparison with contemporary techniques, the correct state-of-the-art remains ambiguous, and additionally the field lacks a formal approach that quantifies viewpoint and appearance change. Before presenting our analysis on VPR evaluation, we acknowledge that no universally best technique at the fronts of all types of conditional variations, computational needs and storage requirements exists or is expected from the research community through this work. The objective of this work is instead to provide an open-source implementation of an evaluation/quality-control framework and a pre-established go-to strategy for employing (or integrating) a variety of metrics, datasets and popular VPR techniques for all new evaluations, thereby identifying the strengths and weaknesses of any future VPR techniques on a common-ground. An overview of our framework is shown in Fig. 1.

This work is a major extension of our previously presented works (Zaffar et al. (2019a), Zaffar et al. (2019b)) at IEEE International Conference on Robotics and Automation-Workshop on Database Generation and Bench-marking (DGB-ICRA 2019) and at IEEE International Conference on Robotics and Automation-Workshop on Aerial Robotics, respectively. Following-up on (Zaffar et al. (2019b)), we received several inquiries for assistance with evaluations, gaps in theoretical understanding and implementation complications, which partly served as a motivation for this work. We then undertook an extensive review which revealed many underlying issues in the research and evaluation landscape. In Zaffar et al. (2019b), we had also identified that the place matching performance improvement is not temporally consistent over the past 10-15 years and that there are irregularities in between datasets and techniques, as shown in Fig. 2. More recently, Ferrarini et al. (2020) proposed that the widely employed Area-Under-the-Precision-Recall-Curves (AUC) metric for evaluating VPR matching performance is not desirable, as it does not regard 'Recall at 100% Precision' and therefore, proposed 'Extended-Precision' as a new evaluation metric for place matching performance. While these existing works are similar in spirit to our presented open-source framework, there are several new insights and improvements in this work that address previously untouched areas of investigation. Firstly, this research is not a snapshot performance evaluation unlike existing evaluations, but instead a comprehensive open-source framework designed in a modular way, such that any VPR researcher can

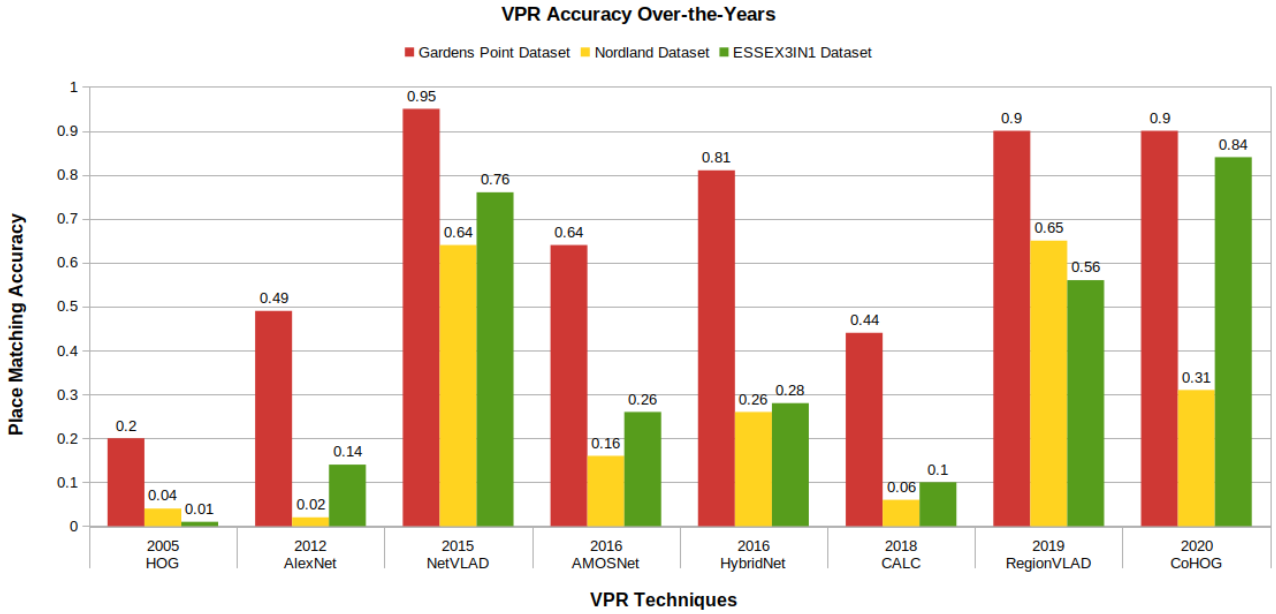

Fig. 2 Accuracy of several VPR techniques on Gardenpoint dataset (Sünderhauf et al. (2015)), Nordland dataset (Skrede (2013)) and ESSEX3IN1 dataset (Zaffar et al. (2018)) is shown here in a chronological order. The trends show irregularities in between techniques and datasets, while the increase in accuracy is also not temporally consistent. These datasets and techniques have been discussed later in our paper. Please note that this graph is not intended to reflect the utility of these techniques, as some less-precise techniques have significantly lower computational requirements and can process more place-recognition (loop-closure) candidates.

easily determine the efficacy of their technique and retrieve detailed performance evaluation, thereby reducing the time overhead and maintaining consistency with the several years of existing VPR research. The modular design and convenient templates enable regular updates to the framework by integrating newly proposed techniques, datasets and/or metrics over time, while also ensuring that the framework does not break. In Zaffar et al. (2019b), we had only used 3 evaluation datasets representing outdoor environment and AUC as the only evaluation metric, however, this work employs significantly more datasets from both indoor and outdoor environments and a range of different evaluation metrics.

A key contribution of this research is attempting to quantify the invariance of VPR techniques to viewpoint and illumination changes. In this respect, we utilise the detailed variation-quantified Point Feature dataset (Aanæs et al. (2012)) and integrate it into our framework to numerically and visually interpret the invariance of techniques, instead of the usual, qualitative invariance terms of 'mild', 'moderate', 'high' and 'extreme' etc. This quantified variation is obtained by taking images of a fixed scene from various angles and distances, under different illumination conditions, as explained later in sub-section 3.5. We devise our analysis based on the decrease in matching scores between images of the same scene (place) as the viewpoint and illumination conditions are varied. We then draw these matching scores of the same-but-varied scene along with the matching scores of different scenes observed from the same viewpoint and under the same illumination. These graphs helps to identify the variation levels where the same-but-varied scenes get matching scores as low as different scenes, which may lead to false positives. We then analyse these graphs and devise metrics that may help to further analyse this invariance.

A principal expectation from an article of this nature is to ask core, practical and veristic questions specific to the research problem. In this respect, the core issues and questions addressed in this research are:

1. Which evaluation metric is the correct choice or is it application dependent and why?
2. Is viewpoint-invariance actually required and why/why not?
3. How to quantify the viewpoint and illumination-invariance of VPR techniques?
4. Can acceptable ground-truth manipulation change the top-performing technique?
5. Do the current performance metrics actually reflect the functional utility of a VPR system in deployed systems?
6. What is a good image retrieval time and what can be classified (and modeled) as a real-time VPR technique?

7. In a real-world, real-time scenario; are moderately precise but fast techniques desirable in comparison with highly precise but slow techniques and is there an application-dependent clustering of techniques?

The remainder of the paper is organized as follows. In Section 2, a comprehensive literature review regarding VPR state-of-the-art is presented. Section 3 presents the details of the evaluation setup employed in this work. Section 4 puts forth the results and analysis obtained by evaluating the contemporary VPR techniques on public VPR datasets, along with insights into invariance quantification. Finally, conclusions and future directions are presented in Section 5.

2 Literature Review

The detailed theory behind Visual Place Recognition (VPR), its challenges, applications, proposed techniques, datasets and evaluation metrics has been thoroughly reviewed by Lowry et al. (2015).

Before diving deep into the core VPR literature review, it is important to co-relate and distinguish VPR research from closely related topics including visual-SLAM, visual-localisation and image matching (or correspondence problem), to set the scope of our research. A huge body of robotics research in the past few decades has been dedicated to the problem of simultaneously localising and mapping an environment, as thoroughly reviewed by Cadena et al. (2016). Performing SLAM with only visual information is termed as visual-SLAM and Davison et al. (2007) were the first to fully demonstrate this. The localisation part of visual-SLAM can be broadly divided into 2 tasks: 1) Computing change in camera/robot pose while performing a particular motion, using inter-frame(s) co-observed information, 2) Recognising a previously seen place to perform loop-closure. The former is usually referred to as visual-localisation and Nardi et al. (2015) developed an open-source framework in this context for evaluating visual-SLAM algorithms. The latter is essentially an image-retrieval problem for the computer vision community and within the context of robotics has been referred to as Visual Place Recognition, as reviewed by Lowry et al. (2015). Image matching (also referred to as keypoint matching or correspondence problem in some literature) consists of finding repeatable, distinct and static features in images, describing them using condition-invariant descriptors and then trying to locate co-observed features in various images of the same scene. It is primarily targeted for visual-localisation, 3D-model creation, Structure-from-Motion and geometric-verification, but can also be utilised for VPR. Some recent advances include SuperPoint (DeTone et al. (2018)) and D2-net (Dusmanu et al. (2019)). Jin et al. (2020) developed an evaluation framework along these lines for matching images across wide baselines.

VPR, however, is a purely image retrieval problem and is not focused on the geometric location of the features in an image or on computing the pose change between two consecutive camera frames. That said, it is possible to combine VPR and local-feature (image) matching to perform accurate localisation, as shown by Camara et al. (2019) and Sarlin et al. (2019). The existing literature in VPR can largely be broken down into: 1) Handcrafted feature descriptors-based VPR techniques, 2) Deep-learning-based VPR techniques, 3) Regions-of-Interest-based VPR techniques. All of these major classes have their trade-offs between matching performance, computational requirements and approach salience.

Handcrafted feature descriptors can be further subdivided into 2 major classes: local feature descriptors and global feature descriptors. The most popular local feature descriptors developed in the vision community include Scale Invariant Feature Transform (SIFT Lowe (2004)) and Speeded Up Robust Features (SURF Bay et al. (2006)). These descriptors have been used for the VPR problem by Se et al. (2002), Andreasson and Duckett (2004), Stumm et al. (2013), Košecká et al. (2005) and Murillo et al. (2007). A probabilistic visual-SLAM algorithm was presented by Cummins and Newman (2011)), namely Frequent Appearance-based Mapping (FAB-MAP), that used SURF as the feature detector/descriptor and represented places as visual words. Odometry information was integrated into FAB-MAP by Maddern et al. (2012) to achieve Continuous Appearance Trajectory-based SLAM (CAT-SLAM) by using RaoBlackwellised particle filter. CenSurE (Center Surround Extremas Agrawal et al. (2008)) is also a popular local feature descriptor and has been used for VPR by Konolige and Agrawal (2008). FAST (Rosten and Drummond (2006)) which is a popular high-speed corner detector has been used in combination with the SIFT descriptor for SLAM by Mei et al. (2009). Matching of local feature descriptors is a computationally intense process which has been addressed by Bag of visual Words (BoW Sivic and Zisserman (2003)) approach. BoW collects visually similar features in dedicated bins (pre-defined or learned by training a visual-dictionary) without topological consideration, enabling direct matching of BoW descriptors. Some of the techniques using BoW for VPR, include the works of Angeli et al. (2008), Ho and Newman (2007), Wang et al. (2005) and Filliat (2007).

Global feature descriptors create a holistic signature for an entire image and Gist (Oliva and Torralba (2006)) is one of the most popular global feature descriptor. Working on panoramic images, Murillo and Kosecka (2009), Singh and Kosecka (2010) used Gist for VPR. Sünderhauf and Protzel (2011) combined Gist with BRIEF (Calonder et al. (2011)) to perform large scale visual-SLAM. Badino et al. (2012) used Whole-Image SURF (WI-SURF), which is a global variant of SURF to perform place recognition. Operating on sequences of raw RGB-images, Seq-SLAM (Milford and Wyeth (2012)) uses normalized pixel-intensity matching in a global fashion to perform VPR in challenging conditionally-variant environments. The original Seq-SLAM algorithm assumes constant speed of robotic platform, thus, Pepperell et al. (2014) extended Seq-SLAM by considering the variable speed of the robotic platform. McManus et al. (2014) extract scene signatures from an image by utilising some *a priori* environment information and describe them using HOG-descriptors. A more recent usage of traditional handcrafted feature descriptors for VPR was presented in CoHOG (Zaffar et al. (2020)) by using entropy-rich regions in an image and using HOG as the regional descriptor for convolutional-regional matching.

Similar to other domains of computer vision, deep-learning and especially Convolutional-Neural-Networks (CNNs) served as a game-changer for the VPR problem by achieving unprecedented invariance to conditional changes. By employing off-the-shelf pre-trained neural nets, Chen et al. (2014) used features from the Overfeat Network (Sermanet et al. (2014)) and combined it with the spatial filtering scheme of Seq-SLAM. This work was followed up by Chen et al. (2017b), where two neural networks (namely AMOSNet and HybridNet) were trained specifically for VPR on the Specific Places Dataset (SPED). AMOSNet was trained from scratch on SPED, while the weights for HybridNet were initialised from the top-5 convolutional layers of Caffe-Net (Krizhevsky et al. (2012)). An end-to-end neural-network-based holistic descriptor is introduced by Arandjelovic et al. (2016) (namely Net-VLAD), where a new VLAD (Vector-of-Locally-Aggregated-Descriptors (Jégou et al. (2010))) layer is integrated into the CNN architecture achieving excellent place recognition results. A convolutional auto-encoder network is trained in an unsupervised fashion by Merrill and Huang (2018), utilizing HOG-descriptors of images and synthetic viewpoint variations for training. Chancán et al. (2020) draw their inspiration from brain architectures of fruit flies, train a sparse two-layer neural-network and combined it with Continuous-Attractor-Networks to summarise temporal information.

Researchers have used Regions-of-Interest (ROIs) to introduce the concept of salience into VPR, so as to ensure that static, informative and distinct regions are used for place recognition. Regions of Maximum Activated Convolutions (R-MAC) are used by Tolias et al. (2016b), where max-pooling across cropped areas in CNN layers' features define/extract ROIs. High-level features encoded in earlier neural-network layers are used for region-extraction and the following low-level features in later layers are used for describing these regions in the work of Chen et al. (2017a). This work is then followed-up with a flexible attention-based model for region extraction by Chen et al. (2018). Khaliq et al. (2019) draw their inspiration from NetVLAD and R-MAC, thereby combining VLAD description with ROI-extraction to show significant robustness to appearance- and viewpoint-variation. Other interesting approaches to place recognition have also been adopted, including semantic-segmentation-based VPR (as in Stenborg et al. (2018), Schönberger et al. (2018), Naseer et al. (2017)) and object-proposals-based VPR (Hou et al. (2018)).

For images containing repetitive structures, Torii et al. (2013) proposed a robust mechanism for collecting visual words into descriptors. Synthetic views are utilized for enhanced illumination-invariant VPR in Torii et al. (2015), which shows that highly condition-variant images can still be matched, if they are from the same viewpoint. In addition to image retrieval, significant research has been performed in semantic mapping to select images for insertion into a metric, topological or topometric map as nodes/places. Semantic mapping techniques are usually annexed with VPR image retrieval techniques for real-world Visual-SLAM, as quoted and extensively reviewed in the survey performed by Kostavelis and Gasteratos (2015). Most of these semantic mapping techniques are based on bayesian-surprise (Ranganathan (2013), Girdhar and Dudek (2010)), coresets (Paul et al. (2014)), region proposals (Demir and Bozma (2018)), change-point detection (Topp and Christensen (2008), Ranganathan (2013)) and salience-computation (Zaffar et al. (2018)).

While the VPR literature consists of a large number of VPR techniques, we have currently integrated 8 of these techniques into the VPR-Bench framework. We plan to increase this number over time due to the modular nature of our framework with the help of the VPR community. In the following section, we explain the framework implementation details, available techniques, metrics and datasets for evaluations in detail.

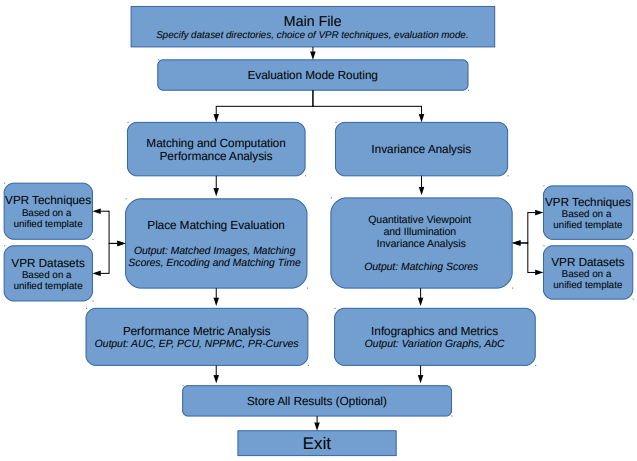

Fig. 3 The code structure of the VPR-Bench framework is shown here.

3 VPR-Bench Framework

This section introduces the details of the VPR-Bench framework, including the datasets module, VPR techniques module, the evaluation metrics module and the invariance quantification module. Along with explaining the generic templates created for these modules, we also explain the currently available several datasets, techniques and metrics in our framework.

3.1 Framework Design

The entire framework has been designed with 2 key focuses: a) A holistic, fully-integrated and easy-to-use framework for VPR performance evaluation at all fronts, b) Modularity and convenient templates for regular updates and future consistency. In this respect, while the modularity, template design and available content within the modules, are explained individually for each of the modules in their respective dedicated sub-sections; this sub-section presents the overall framework structure and implementation details. The code structure of our framework has been described in Fig. 3.

The entry to the framework is a convenient main file, where the choice of evaluation datasets, VPR techniques and evaluation mode can be specified. At present there are 2 evaluation modes: 1) VPR Performance Evaluation and 2) Invariance Analysis. The former yields the place matching performance of different VPR techniques on a specified dataset using different metrics related to precision and computation. The latter tries to present the invariance of these techniques to quantified viewpoint- and illumination-variations. There are 10 evaluation datasets available in the framework from both indoor and outdoor environments. We have integrated 8 different VPR techniques by modifying the open-source codes as per our templates or self-implementing in cases where open-source codes were not available. As we are focused on providing flexibility and ease for integrating new VPR techniques and datasets into VPR-Bench, we have briefly summarised the required steps for both of these changes below.

For integrating a new dataset into VPR-Bench, no change in the framework code is required. You need to setup the dataset as per our unified template, which has been explained in sub-section 3.3 and then set the directory path for this dataset in the main file. In order to integrate a new VPR technique, the main file for this respective technique needs to implement 3 functions, as per the template described in sub-section 3.2. Once these functions have been implemented, you only need to import these functions in the file 'VPR_system.py'. This is a straight-forward process and all other functions and modules will implicitly be integrated for this technique.

The VPR-Bench framework is written fully in Python (2.7), which has been the most used programming/scripting language for VPR research. Our framework does not have a dedicated Graphical-User-Interface (GUI), because the framework is targeted for developers/researchers who are assumed to have basic knowledge of the domain. GUIs also make future improvements much complex and limit the flexibility of an application. The open-source code has been tested on a Ubuntu 18.04.2 LTS system. By default, the framework does not need a GPU (Graphical Processing Unit) for any of the evaluations and all evaluations in this work have been performed using an Intel(R) Core(TM) i5-7500 CPU @ 3.40GHz with a 16GB RAM. Therefore, a huge percentage of VPR researchers, academics and developers, from a broad range of robotic applications ranging from self-driving cars to drones, can conveniently use our framework.

3.2 VPR Techniques

3.2.1 Generic Template

Each VPR technique has a different approach to the problem, which may include neural-network models or traditional feature descriptors. There may be added functionality, like ROI-extraction. However, there is always a common pattern to the input and output fields.

Let Q be a query image and M_R be a list/map of R reference images. The feature descriptor(s) of a query image Q and reference map M_R can be denoted as F_Q and F_M, respectively. If a technique uses ROI-extraction, F_Q will hold within it all the required in-

formation in this regards, including location of regions, their descriptors and corresponding salience as a multi-dimensional list. For a query image Q, given a reference map M_R, let us denote the best matched image/place by a VPR technique as P (where, $P \in M_R$) with a matching score S. The matching score S can be defined as $S \subset \mathbb{R} \; \forall \; \mathbb{R} \in [0 - 1]$. Based on these notations, the following 3 functions need to be implemented in the main file of a VPR technique.

Algorithm VPR Technique Required Template

def *compute_query_desc* (Q)
  Function Body
  return F_Q
def *compute_map_features* (M_R)
  Function Body
  return F_M
def $perform\_VPR\,(F_Q, F_M)$
  Function Body
  return P, S

The definitions (names) of these functions remain the same for all VPR techniques and our framework performs technique-aware selective re-imports of these functions to maintain consistency and ease-of-integration.

3.2.2 HOG Descriptor

Histogram-of-oriented-gradients (HOG) is one of the most widely used handcrafted feature descriptor, which actually performs very well for VPR compared to other handcrafted feature descriptors. It is a good choice for a traditional handcrafted feature descriptor in our framework, based upon its performance as shown by McManus et al. (2014) and the value it presents as an underlying feature descriptor for training a convolutional auto-encoder in Merrill and Huang (2018). We use a cell size of 16×16 and a block size of 32×32 for an image-size of 512×512. The total number of histogram bins are set equal to 9. We use cosine-matching between HOG-descriptors of various images to find the best place match.

3.2.3 AlexNet

The use of AlexNet for VPR was studied by Sünderhauf et al. (2015), who suggest that *conv3* is the most robust to conditional variations. Gaussian random projections are used to encode the activation-maps from *conv3* into feature descriptors. Our implementation of AlexNet is similar to the one employed by Merrill and Huang (2018), while the code has been restructured as per the designed template.

3.2.4 NetVLAD

The original implementation of NetVLAD was in MATLAB, as released by Arandjelovic et al. (2016). The Python port of this code was open-sourced by Cieslewski et al. (2018). The model selected for evaluation is VGG-16, which has been trained in an end-to-end manner on Pittsburgh 30K dataset Arandjelovic et al. (2016) with a dictionary size of 64 while performing whitening on the final descriptors. The code has been modified as per our template.

3.2.5 AMOSNet

This technique was proposed by Chen et al. (2017b), where a CNN has been trained from scratch on the SPED dataset. The authors have presented results from different convolutional layers by implementing spatial-pyramidal pooling on the respective layers. While the original implementation is not fully open-sourced, the trained model weights have been shared by authors. We have implemented AMOSNet as per our template using *conv5* of the shared model. L1-match has been originally proposed by the authors, which is normalised for a score between $0 - 1$.

3.2.6 HybridNet

While AMOSNet was trained from scratch, Chen et al. (2017b) took inspiration from transfer learning for HybridNet and re-trained the weights initialised from Top-5 convolutional layers of CaffeNet (Krizhevsky et al. (2012)) on SPED dataset. We have implemented HybridNet as per our template using *conv5* of the shared HybridNet model. L1-match has been originally proposed by the authors, which is normalised for a score between $0 - 1$.

3.2.7 RegionVLAD

Region-VLAD has been introduced and open-sourced by Khaliq et al. (2019). We have modified it as per our template and have used AlexNet trained as Places365 dataset as the underlying CNN. The total number of ROIs has been set to 400 and we have used 'conv3' for feature extraction. The dictionary size is set to 256 visual words for VLAD retrieval. Cosine similarity is subsequently used for matching descriptors of query and reference images.

3.2.8 CALC

The use of convolutional auto-encoders for VPR was proposed by Merrill and Huang (2018), where an auto-

encoder network was trained in an unsupervised manner to re-create similar HOG-descriptors for viewpoint-variant (cropped) images of the same place. We use model parameters from $100,000$ training iteration and adapt the open-source technique as per our template. Cosine-matching is used for descriptor comparison.

3.2.9 CoHOG

CoHOG is a recently proposed handcrafted feature-descriptor-based technique, which uses image-entropy for ROI extraction. The regions are subsequently described by dedicated HOG-descriptors and these regional descriptors are convolutionally matched to achieve lateral viewpoint-invariance. It is an open-source technique, which has been modified as per our template. We have used an image-size of 512×512, cell-size of 16×16, bin-size of 8 and an entropy-threshold (ET) of 0.4. CoHOG also uses cosine-matching for descriptor comparison.

3.3 Evaluation Datasets

3.3.1 Generic Template

All the datasets that have been employed to date for VPR evaluation comprise of multiple (mostly 2) views of the same environment that may have been extracted under different seasonal, viewpoint and/or illumination conditions. These views are mostly available in the form of monocular images and are structured as separate folders representing query and reference images. However, these views may have been extracted from a traversal or a non-traversal-based mechanism. For the former, consecutive images within a folder (query/reference) usually have overlapping visual content, while for the latter, images within a folder are independent. Accompanying these folders is usually some level of ground-truth information, which has been represented in various ways (e.g, CSV, numpy-arrays, pickle-files etc.) for different datasets. In some cases, ground-truth is not explicitly provided, as images with the same index/name represent the same place. A key observation is that in most traversal-based datasets, there is no single correct match for a query image, because images which are geographically close-by can be considered as the same place, leading to a range requirement for ground-truth matches instead of a single match/value. For consistency in VPR-research and performance-reporting, it is essential to affix a unified template for all of these VPR datasets.

In order to have a fixed template for all the datasets that are available in (or can be integrated into) VPR-Bench, we design a simplistic, generic template that can accommodate the above understanding and variations. Firstly, the query and reference traverses for a dataset are represented by their dedicated sub-folders, namely 'query' and 'ref'. Images within each of these folders need to be named as integers, which is motivated by a graph structure, such that for a traversal-based dataset, increments or decrements to integer values can represent the geographically next or previous image, respectively. The ground-truth file for each dataset is a numpy-array (.npy), which unlike CSV or Pickle files is not protocol-specific. This numpy-array (integer-type) of ground-truth information has dimensions of $Z \times 3$, where Z is the total number of query images in the dataset. For all Z rows of query images, each column represents the query image index, the minimum ground-truth matching reference image index and the maximum ground-truth matching reference image index. For a non-traversal-based dataset, the minimum and maximum ground-truth indices are equal, i.e., there is only a single correct match.

3.3.2 Outdoor Environment

We have integrated several outdoor datasets in our framework representing different types and levels of viewpoint-, illumination- and seasonal-variations. Details of these datasets have been summarised in Table 1 and sample images are shown in Fig. 4. Each of these datasets has a particular attribute to offer, that lead to its selection and they are briefly discussed below.

The GardensPoint dataset was introduced by Sünderhauf et al. (2015), where two repeated traversals of the Gardens Point Campus of Queensland University of Technology, Brisbane, Australia were performed with varying viewpoints in day and night times. A huge body of VPR research has used this dataset for reporting their VPR matching performance, as it depicts outdoor, indoor and natural environments, collectively. The 24/7 query dataset was proposed by Torii et al. (2015), which consists of 6-DOF (degrees-of-freedom) viewpoint-variations and time-of-day variations. It is one of the most challenging datasets for VPR due to the sheer amount of viewpoint- and conditional-variation. The ESSEX3IN1 dataset was proposed by Zaffar et al. (2018) and is the only dataset designed with focus on perceptual aliasing and confusing places/frames for VPR techniques. The SPEDTest dataset was introduced by Chen et al. (2018) and consists of low-quality, high scene-depth frames extracted from CCTV cameras across the world. This dataset has the unique attribute of covering a huge variety of scenes from all across the world under several different weather, sea-

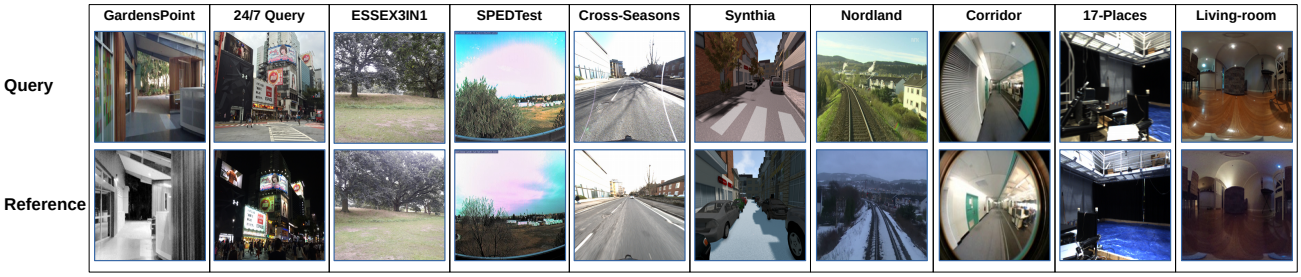

Fig. 4 Sample images from all 10 VPR datasets employed in this work are presented here. These datasets span several different environments, including cities, natural scenery, train-lines, rooms, offices, corridors, buildings, busy-streets and such.

Table 1 VPR-Bench Datasets

Dataset	Environment	Query Images	Ref Images	Viewpoint-Variation	Conditional-Variation
GardensPoint	University Campus	200	200	Lateral	Day-Night
24/7 Query	Outdoor	375	750	6-DOF	Day-Night
ESSEX3IN1	University Campus	210	210	6-DOF	Illumination
SPEDTest	Outdoor	607	607	None	Seasonal and Weather
Cross-Seasons	City-like	191	191	Lateral	Dawn-Dusk
Synthia	City-like (Synthetic)	947	947	Lateral	Seasonal
Nordland	Train Journey	1622	1622	None	Seasonal
Corridor	Indoor	111	111	Lateral	None
17-Places	Indoor	406	406	Lateral	Day-Night
Living-room	Indoor	32	32	Lateral	Day-Night

sonal and illumination conditions. The Synthia dataset was introduced in Ros et al. (2016) and represents a simulated city-like environment in summer and winter conditions. The Cross-Seasons dataset employed in our work represents a traversal from Larsson et al. (2019), which is a subset of the Oxford RobotCar dataset (Maddern et al. (2017)). This dataset represents a challenging real-world car traversal from dawn and dusk conditions. One of the widely employed datasets for VPR is the Nordland dataset (Skrede (2013)), which represents a train journey in Norway during Summer and Winter seasons. As Nordland dataset represents natural (non-urban), outdoor environment, which is unexplored in any other dataset, we have integrated it into VPR-Bench.

3.3.3 Indoor Environment

A significant focus in recent research in VPR has primarily been on evaluation with outdoor datasets: here we incorporate indoor environments which are usually a key area of study within robot autonomy. Therefore, we find it important to integrate a few indoor datasets in VPR-Bench. While indoor datasets, usually do not represent the seasonal variation challenges as outdoor datasets and the level of viewpoint-variation is relatively lesser than outdoor datasets, they do contain dynamic objects like humans, animals or changing setup/environment configurations, less-informative content and perceptual-aliasing. The details of these datasets have been summarised in Table 1 and sample images are shown in Fig. 4. We have briefly discussed the currently available indoor datasets in VPR-Bench, in the following paragraph.

We have integrated the 17-Places dataset introduced by Sahdev and Tsotsos (2016) into VPR-Bench, which consists of several different indoor scenes, ranging from office environment to labs, hallways, seminar rooms, bedrooms and many other. This dataset exhibits both viewpoint- and conditional-variations. We also use the viewpoint-variant Corridor dataset, introduced by Milford (2013), which represents the challenge of low-resolution images (160×120 pixels) for vision-based place recognition. Mount and Milford (2016) introduced the living-room dataset for home-service robots, which represents indoor environment from a highly relevant and challenging viewpoint of cameras mounted close-to-ground level.

3.4 Evaluation Metrics

A trend within current VPR research has shown that a single, universal metric to evaluate VPR techniques that could simultaneously extend to all applications, platforms and user-requirements does not exist. For example, a technique which has a very high-precision,

but a significantly higher image-retrieval time (few seconds), cannot extend to a VPR-based, real-time navigation system, as the localisation module will be much slower (in frames-per-second processed) than the platform dynamics. However, for offline applications, where real-time place matching may not be required, for example, offline loop-closures for map correction, improved-representations and structure-from-motion, high precision at the cost of higher retrieval time may be acceptable. Therefore, reporting performance on a single metric may not fully present the utility of a VPR technique to the entire academic, industrial and research audience.

We have integrated into VPR-Bench, all the different metrics that evaluate a VPR technique on the fronts of matching performance, computational needs and storage requirements. Firstly, the most used metric for matching performance in VPR is the Area-under-the-Precision-Recall-Curves (AUC). Precision-Recall curves are aimed at understanding the loss of precision with increasing recall at different confidence scores, i.e, how many false positives are introduced by reducing the number of false negatives for a particular confidence score. Generally, in VPR the image matching/similarity scores are considered as confidence scores. Precision and Recall are computed using the below equations.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (1)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

Where in terms of VPR, a True Positive (TP) represents an image correctly matched by a VPR technique based on ground-truth information (range-based or single-value). A False Positive (FP) represents an incorrectly matched image based on ground-truth information (range-based or single-value). A False Negative (FN) is any correctly matched image that is rejected because the matching score for that match is lower than a matching threshold, where this matching threshold is user-defined. Please note that in VPR datasets, all correctly matched images that are rejected due to matching scores lower than the threshold are classified as false negatives, because ground-truth matches exist for all images in the datasets. There are no true negatives in the datasets. By selecting different values of the matching threshold, varying between the highest matching score and the lowest matching score, different values of Precision and Recall can be computed. The Precision values are plotted against the Recall, and area under this curve is computed, which is termed as AUC.

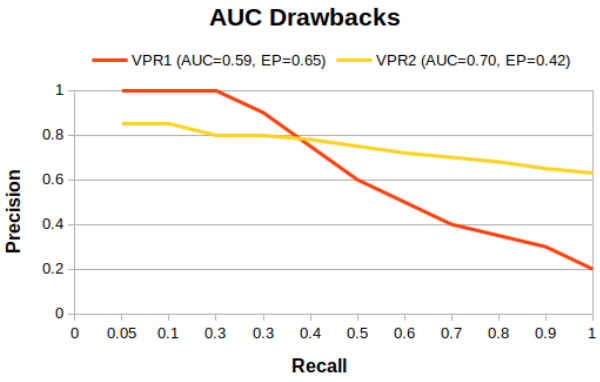

Fig. 5 The PR curves for 2 hypothetical VPR techniques are shown here. The curves show that although VPR2 never reaches 100% precision, its AUC value is higher than VPR1. For applications requiring higher precision at the cost of more false negatives (rejected correct matches), Extended-Precision (EP) presents much better value than AUC.

While AUC gives a good overview of the matching performance, it does not consider Recall at 100 % Precision (R_{P100}) and Precision at 0% Recall (P_{R0}), which may be very crucial for some applications. We use Fig. 5 to further explain this, where PR curves for two hypothetical VPR techniques are drawn. As shown in Fig. 5, the AUC of VPR2 is higher than VPR1, despite the fact that VPR2 never reaches to 100% Precision. For applications, where false positives are catastrophic, it is desirable to use VPR1 than VPR2, however, AUC would suggest otherwise. Ferrarini et al. (2020) proposed a new evaluation metric, namely 'Extended Precision (EP)', that takes into account R_{P100} and P_{R0}. While the original EP metric uses P_{R0}, we identify that a recall value of 0 may not always exist, as the number of true positives can be non-zero for the highest possible matching threshold, i.e, the lowest possible recall value can be non-zero. Thus, we utilise the Precision at minimum Recall P_{Rmin} instead of P_{R0} and compute the value of EP using the below equation 3, where, $R_{P100} = 0$ if the value of P_{Rmin} is not equal to 1. Please note that P_{Rmin} generally represents the maximum possible precision that can be achieved by a VPR technique, however in some cases it is even possible that the maximum precision lies towards the right of the minimum possible recall value in a PR-curve. While this is possible, it is not a desirable behavior of a VPR system and therefore, EP utilises the P_{Rmin} metric instead of the maximum precision at any recall value.

$$EP = \frac{P_{Rmin} + R_{P100}}{2} \quad (3)$$

AUC and EP are focused only on the matching performance and do not accommodate computational intensity of techniques. For real-world, resource-constrained platforms, matching performance needs to be related to computational units. In Zaffar et al. (2020), the precision at 100% recall of a technique (P_{R100}) is combined with feature encoding time per image (t_e), to define the Performance-per-Compute-Unit (PCU), which is computed as below:

$$PCU = P_{R100} \times \ log(\ \frac{t_{e\_max}}{t_e} + 9) \tag{4}$$

In the above equation 4, higher precision directly leads to higher PCU. However, for t_e, the logarithmic encoding time boost is computed for a given VPR technique to provide a reasonable combination of precision and encoding time metrics. Thus, only exponential increase in encoding time for a highly precise VPR technique leads to increase in PCU. Maximum feature encoding time (t_{e_max}) belongs to the most computationally intensive VPR technique in VPR-Bench, i.e, PCU is a relative performance metric and not absolute. A scalar '9' is added in equation 4 to ensure that $PCU = P_{R100}$ for the technique with $t_e = t_{e\_max}$, instead of $PCU = 0$, thus providing an interpretable scale.

While the above discussed metrics (AUC, EP and PCU) try to quantitatively summarise the matching performance of a VPR technique, their extension to real-world scenarios is ambiguous. That is, although higher AUC/EP/PCU may reflect that a technique retrieves mostly correct matches, real-world factors like image-matching time, platform-speed and trajectory-length are neglected. A real-world application may require that a VPR technique must retrieve K potential matching candidates in a trajectory-length of L meters, when the platform is moving at V meters-per-second (mps) speed. We propose a new metric inspired from this need, that represents the number of prospective place matching candidates (NPPMC). This NPPMC, as the name suggests, depends on both the matching performance of a technique, but also on the computational performance of a technique given platform characteristics.

Let the retrieval-time of a VPR technique be denoted as t_R, where this t_R represents the time taken (in seconds) by a VPR technique to encode an input query image and match it with images in the reference map of Z images to output a potential place matching candidate. We model this t_R as in equation 5.

$$t_R = t_e + O(Z) \times t_m \tag{5}$$

Where, $O(Z)$ represents the search mechanism for image matching and could be linear, logarithmic or other depending upon the employed neighbourhood selection mechanism (e.g., linear search, approximate nearest neighbour search etc.). Additionally, t_e represents the feature encoding time and t_m represents the time required to match the the feature descriptors of 2 images. Thus, the total query frames that can be matched by a VPR technique, when the platform covers a trajectory of length L at a speed V is denoted as TMF and computed as below;

$$TMF = \frac{1}{t_R} \times \frac{L}{V} \tag{6}$$

These matched frames correspond to consecutive frames acquired every meter (constant-distance-based place sampling), but because this is a linear relation in equation , it can be extended to other values of fixed-distance-based sampling, e.g, frames matched every 5 meter and such. Out of these total matched frames (TMF), some will be correct matches and others will be false positives. This probabilistic distribution can be modelled to an acceptable level by accuracy A. The accuracy of a technique is the percentage of query images correctly matched in a given dataset by that technique. We therefore scale TMF by A to represent the total number of prospective place matching candidates (NPPMC) as below:

$$NPPMC = A \times TMF = A \times \frac{1}{t_R} \times \frac{L}{V} \tag{7}$$

It is possible to take the ratio (L/V) constant in equation 7, because it remains the same for all VPR techniques and one could argue that it can be neglected. However, we propose that it is an important attribute to model, which if even taken constant, can assist to determine the real-world usage of a VPR technique from an apparent zoo of techniques, thus to make informed choices. An important application of NPPMC is that although the accuracy A is computed on the entire dataset (which intrinsically supports using the accuracy A for probability modelling of prospective correct matches as all the dataset images are observed), TMF transfers this A to represent the performance in a real-world scenario. One of the assumptions in computing NPPMC is that the camera FPS (frames-per-second) is equal to or higher than the retrieval performance of a technique, which is valid in most realistic scenarios. All of these metrics, including AUC, EP, PCU and NPPMC are fully integrated into the VPR-Bench framework and easily accessible.

One of the key focus while implementing this framework was to ensure that t_e and t_m are computed

in a fashion, where all subsequent dependencies, pre-processing and preparations of a VPR technique are included in the timings. Therefore, as per the template presented in sub-section 3.2.1 and the design presented in sub-section 3.1, t_e and t_m are computed in a technique-independent port fashion. Additional to the metrics discussed previously, we also compute and report the feature descriptor size of all VPR techniques to reflect the storage requirements, which are highly relevant for large-scale maps.

3.5 Invariance Quantification Setup

A significant body of VPR research (as reviewed in Section 2) has been focused on proposing techniques that are invariant to viewpoint, illumination and seasonal variations, which are the 3 major challenges in VPR. While seasonal variations are difficult to quantify, viewpoint and illumination variation can be modelled by quantitative metrics. In this regard, Aanæs et al. (2012) proposed a well-designed and highly-detailed dataset, namely Point Features dataset, where a scene is captured from 119 different viewpoints, under 19 different illumination conditions. While the original dataset consists of different scenes, some of which are irrelevant to VPR, we utilise a subset of the dataset that represents scenes of synthetically-created places. We have integrated this subset of the Point Features dataset in our framework and this sub-section is dedicated to explaining the details of the dataset. Fig. 6 shows various components of the dataset.

The Point Features dataset can be broadly classified to have 3 variations: 1) Viewpoint, 2) Illumination and 3) Scene. We fully use the former two variations in our work, while only relevant scenes are utilised from the latter. The authors (Aanæs et al. (2012)) achieve viewpoint-variation by mounting the scene facing camera on a highly-precise robot arm, where this robot arm is configured to move across and in-between 3 different arcs, that amount to a total of 119 different viewpoints, as depicted in Fig. 7. Their setup used 19 LEDs that varied from left-to-right and front-to-back to depict a varying directional light source. This directional illumination setup has been reproduced in Fig. 8, while the azimuth (ϕ) and elevation angle (θ) of each LED is listed in Table 2.

In order to utilise the densely-sampled viewpoint and illumination conditions in the Point Feature dataset, we had to devise an analysis scheme where VPR performance variation could be quantified and analysed. This quantification is not possible with the traditional place matching evaluation, where there are only 2 possible outcomes for a given query image, i.e, a correct match or a false match. This is because the mismatch cannot be guaranteed to have resulted from that particular variation and may have resulted from perceptual-aliasing or a smaller map-size. Also, even if an image is matched, it is not guaranteed that increasing the map-size (i.e, the no. of reference images) would not effect the outcome, as the greater the no. of reference images, the greater the chances of mismatch. However, each VPR technique does yield a confidence-score for the similarity of 2 images/places. Ideally, if 2 images represent the same place, then the confidence-score should remain the same, if one of the image of that place is varied with respect to viewpoint or illumination, while keeping the other constant. However, in practical cases, VPR techniques are not fully-immune to such variations and a useful analysis would be to see this effect on the confidence-score.

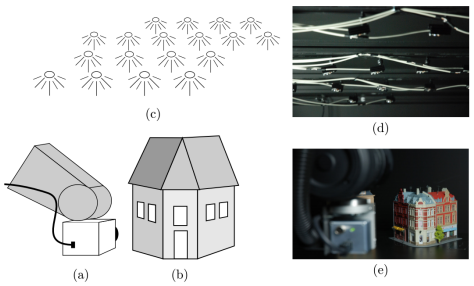

Fig. 6 The schematic setup of Point Features dataset has been reproduced here with permission from Aanæs et al. (2012). The dataset primarily consists of (a) A camera mounted on a robot-arm, (b) Scene, (c) LED arrays for illumination, (d) (e) snapshots of the actual setup.

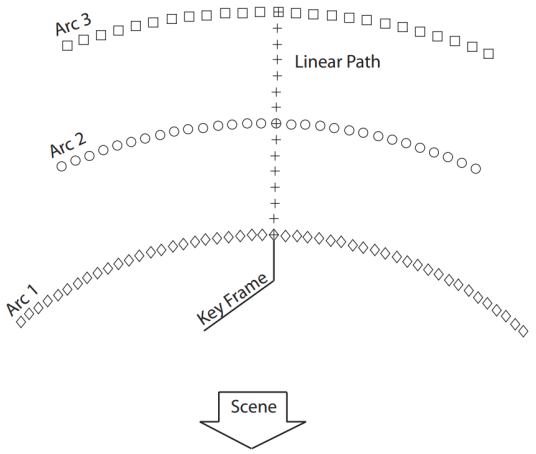

Fig. 7 The 119 different viewpoints in the Point Features dataset have been reproduced here with permission from Aanæs et al. (2012). Camera is directed towards the scene from all viewpoints. Arc 1, 2 and 3 span 40, 25 and 20 degrees, respectively, while the radii are 0.5, 0.65 and 0.8 meters.

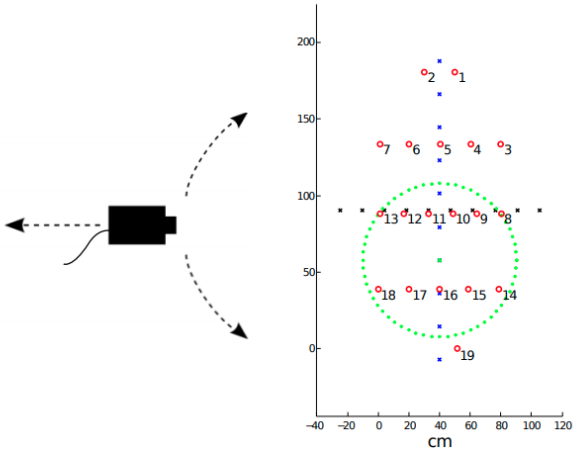

Fig. 8 The distribution of LEDs across physical space is shown as seen from above. Each red circle represents an LED and only a single LED is illuminated at a point in time, yielding 19 different illumination conditions. In the original work, Aanæs et al. (2012), used artificial linear relighting from left-to-right (blue) and front-to-back (black) based on a Gaussian-weighting, as depicted with the green-circle, but in our work we have only used the original 19 single-LED illuminated cases. These 19 cases (red-circles) need to be seen in correspondence with Table 2.

Table 2 The azimuth (ϕ) and elevation angle (θ) of each LED is listed here (in degrees) with respect to the physical table surface that acts as the center of coordinate system.

LED Number	θ	ϕ	LED Number	θ	ϕ
1	264	57	11	28	86
2	277	57	12	10	80
3	227	68	13	6	74
4	245	72	14	125	65
5	270	73	15	109	68
6	297	72	16	89	69
7	314	68	17	69	68
8	174	74	18	53	64
9	170	80	19	97	56
10	152	86			

Therefore, our analysis and the VPR-Bench framework are developed based on the effect of viewpoint- and illumination-variation on the confidence score. This confidence score usually refers to the matching score (L1-matching, L2-matching, cosine-matching etc.) in VPR research and for 2 exactly similar images (i.e, 2 copies of an image) this confidence/matching score is always equal to 1. However, when the image of the same place/scene is varied with respect to viewpoint or illumination, the confidence score decreases. This decrease in matching score by varying images of the same place/scene along the pre-known, numerically-quantified 119 viewpoint- and 19 illumination-levels, presents analytically and visually the limits of invariance of a VPR technique. However, the trends of these variations in-between different VPR techniques cannot be compared solely based on the decrease of confidence scores, due to different matching methodologies. Therefore, for each VPR technique, we draw the confidence score variation trend for the same place along with the trend for a different place/scene. The point at which the matching score for the same place (but viewpoint or illumination varied) approaches near (or below) the matching score for a different place (with the same viewpoint and illumination), identifies the numeric value of viewpoint/illumination change that VPR technique cannot prospectively handle.

4 Results and Analysis

4.1 VPR Performance Evaluation

In this section, we present the results obtained by executing the VPR-Bench framework given the attributes presented in Section 3. Firstly, the precision-recall curves for all 8 VPR techniques on the 10 indoor and outdoor datasets are presented in Fig. 9. From the perspective of place matching precision, VPR-specific deep-learning techniques generally perform better than handcrafted feature descriptors, with the exception of CoHOG, which always performs better than AlexNet and CALC. While CoHOG can handle lateral viewpoint-variation, it cannot handle 6-DOF viewpoint-variation as present in the 24/7 Query dataset. NetVLAD can handle 6-DOF viewpoint-variation better than any other technique, because the training dataset for NetVLAD contained 6-DOF viewpoint-variations. HybridNet and AMOSNet can handle only moderate viewpoint-variations, but perform well under conditional variations due to training on highly conditionally-variant SPED dataset. Please note that the SPED dataset and SPEDTest dataset do not contain the same images, therefore the state-of-the-art performance of HybridNet and AMOSNet on SPEDTest dataset advocates for the utility of deep-learning techniques in environments similar to training environments (which in this case is the world from a CCTV's point-of-view). HOG and AlexNet usually lie on the lower-end of matching capabilities for all viewpoint-variant datasets, but perform acceptably on moderately condition-variant datasets that have no viewpoint variation. A notable exception here is the state-of-the-art performance of HOG compared to all other techniques on the Living Room dataset, which consists of high-quality images of places under indoor illumination variations. CALC cannot handle conditional variations to the same level as other deep-learning-based techniques, as the auto-encoder in CALC is

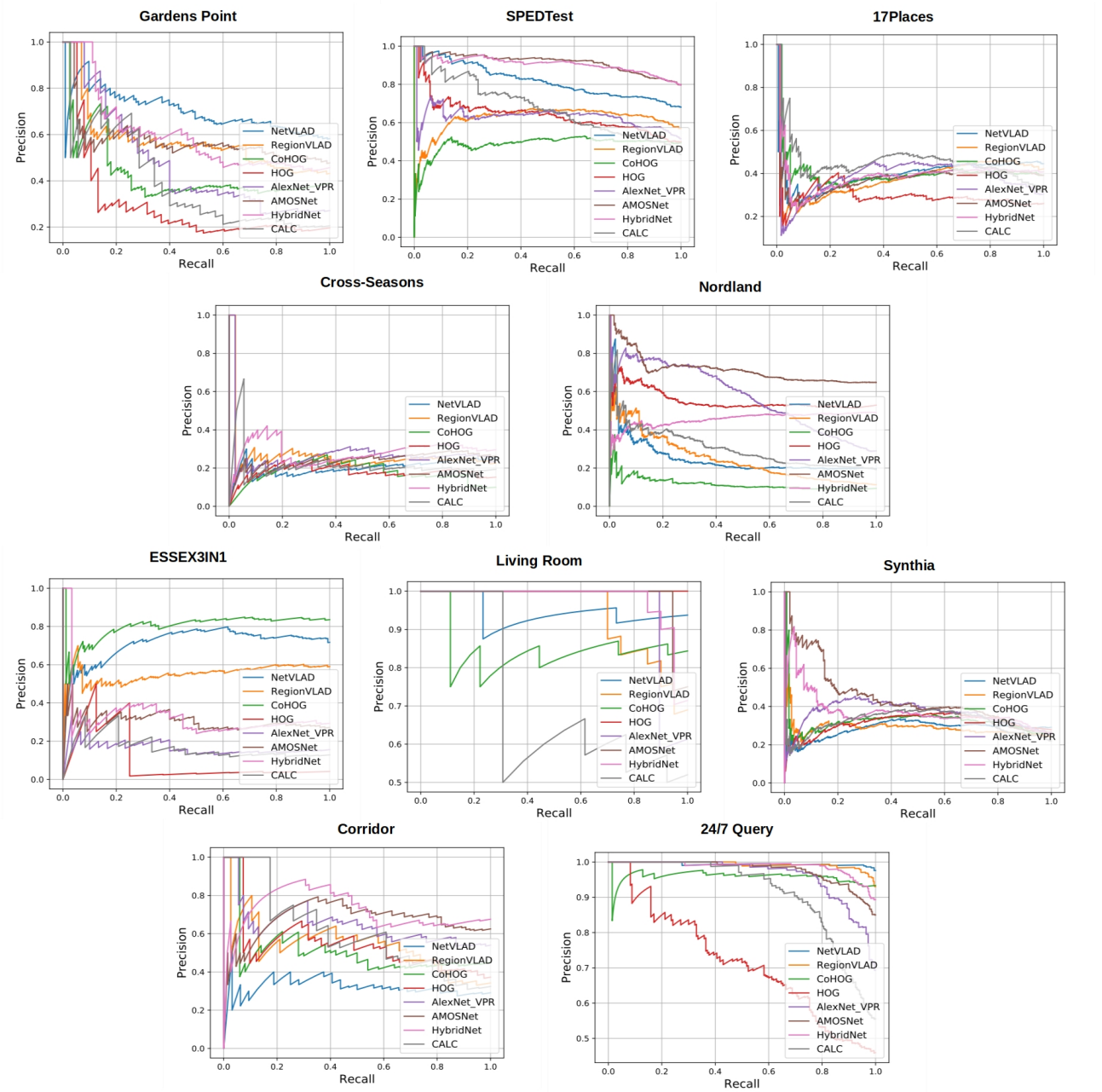

Fig. 9 The Precision-Recall curves for all 8 VPR techniques generated on the 10 datasets by VPR-Bench framework are presented here.

only trained to handle moderate and uniform illumination changes. Region-VLAD also performs in the same spectrum as NetVLAD, but cannot surpass it on most datasets. All techniques perform poorly on the 17 Places dataset that represents a challenging indoor environment, suggesting that the outdoor performance success of techniques cannot be extended to an indoor environment. The perceptual-aliasing of datasets like Cross-Seasons and Synthia also presents significant challenges to VPR techniques. The AUC of HOG comes out as 1 for the Living Room dataset, because a threshold exists above which all images are correct matches (17 out of 32) and below which (15 out of 32) all images are incorrect matches. The values of AUC for all techniques have been listed in Table 3.

As previously discussed, AUC does not reflect the trend of PR-Curve and the matching performance considering computational requirements. Thus, the values of Extended-Precision (EP) and Performance-per-Compute-Unit (PCU) have also been computed by our

Table 3 The values of AUC, PCU and EP are listed here in the respective order for all the techniques on the 10 datasets. The bold values in each row represent the state-of-the-art technique for each dataset for the corresponding metric.

Dataset Name	NetVLAD	RegionVLAD	CoHOG	HOG	AlexNet	AMOSNet	HybridNet	CALC
Gardens Point	**0.70**,0.58,0.50	0.56,0.49,0.53	0.42,**0.76**,0.51	0.28,0.56,0.52	0.47,0.32,0.54	0.57,0.58,0.52	0.59,0.56,**0.55**	0.38,0.45,0.51
SPEDTest	0.81,0.68,**0.51**	0.61,0.57,0.0	0.48,0.76,0.0	0.63,**1.20**,0.50	0.63,0.53,0.50	0.91,0.84,**0.51**	**0.94**,0.84,**0.51**	0.67,0.80,**0.51**
Nordland	0.24,0.19,0.0	0.24,0.12,0.0	0.11,0.16,0.0	0.55,**1.37**,**0.50**	0.57,0.31,**0.50**	**0.71**,0.72,**0.50**	0.45,0.54,**0.50**	0.30,0.37,0.0
Living Room	0.94,0.93,0.61	0.94,0.91,0.85	0.85,2.03,0.55	**1.0**,**3.32**,**1.0**	0.95,0.84,0.94	0.95,1.08,0.97	0.97,1.03,0.92	0.70,1.35,0.65
Synthia	0.28,0.29,0.0	0.28,0.29,**0.50**	0.32,0.56,**0.50**	0.31,**0.87**,0.0	0.36,0.33,0.0	**0.44**,0.36,**0.50**	0.37,0.36,0.0	0.32,0.53,**0.50**
17Places	0.39,0.44,**0.50**	0.38,0.45,**0.50**	0.40,**0.70**,**0.50**	0.29,**0.70**,**0.50**	0.39,0.33,**0.50**	0.37,0.44,**0.50**	0.39,0.46,**0.50**	**0.45**,0.66,**0.50**
Cross-Seasons	0.20,0.25,0.0	0.26,0.27,**0.51**	0.17,0.22,0.0	0.17,**0.48**,0.0	0.27,0.30,**0.51**	0.24,0.35,0.0	**0.28**,0.40,0.0	0.20,0.47,0.0
Corridor	0.31,0.29,0.0	0.53,0.34,0.51	0.50,0.65,0.53	0.54,**0.84**,0.53	0.64,0.54,0.52	0.66,0.64,0.0	**0.71**,0.69,0.0	0.60,0.53,**0.58**
24/7 Query	**0.99**,0.97,0.63	**0.99**,1.28,**0.73**	0.95,**2.30**,0.50	0.71,1.57,0.54	0.96,1.01,0.71	0.97,1.33,0.70	0.98,1.40,0.64	0.91,1.47,0.69
ESSEX3IN1	0.71,0.71,0.0	0.55,0.66,0.0	**0.80**,**1.63**,0.50	0.09,0.11,0.0	0.16,0.17,0.0	0.30,0.32,0.0	0.32,0.35,**0.51**	0.16,0.28,0.0

Table 4 The values of accuracy (A) and encoding time (sec) t_e are listed here in the respective order for all the techniques on the 10 datasets used in this work. The values of t_e are averaged over the entire dataset, i.e, over the total number of query images. The values of encoding times vary between the datasets due to varying input image size. The bold values in each row represent the state-of-the-art technique for each dataset for the corresponding metric.

Dataset Name	NetVLAD	RegionVLAD	CoHOG	HOG	AlexNet	AMOSNet	HybridNet	CALC
Gardens Point	**0.56**,5.72	0.43,1.15	0.39,0.06	0.19,**0.007**	0.25,0.85	0.47,0.69	0.45,0.69	0.17,0.03
SPEDTest	0.67,1.77	0.56,1.32	0.49,0.06	0.49,**0.007**	0.51,1.01	**0.79**,0.69	**0.79**,0.69	0.42,0.02
Nordland	0.19,2.94	0.11,1.19	0.09,0.06	0.52,**0.007**	0.28,0.86	**0.64**,0.68	0.48,0.68	0.19,0.03
Living Room	**0.93**,16.37	0.62,1.34	0.84,0.06	0.53,**0.007**	0.59,1.12	0.56,0.85	0.62,0.85	0.40,0.04
Synthia	**0.29**,10.11	0.24,1.33	0.25,0.06	0.27,**0.007**	0.25,0.95	0.26,0.70	0.26,0.70	0.21,0.04
17Places	**0.44**,3.69	0.39,1.27	0.38,0.06	0.22,**0.007**	0.30,1.01	0.39,0.77	0.40,0.79	0.30,0.03
Cross-Seasons	0.24,10.95	0.21,1.31	0.09,0.06	0.15,**0.007**	0.23,1.10	0.25,0.79	**0.29**,0.79	0.18,0.03
Corridor	0.28,0.93	0.34,1.25	0.45,0.06	0.36,**0.007**	0.48,1.12	0.58,0.79	**0.67**,0.79	0.20,0.03
24/7 Query	**0.96**,20.48	0.92,1.37	0.93,0.06	0.45,**0.007**	0.67,1.01	0.84,0.72	0.89,0.72	0.53,0.04
ESSEX3IN1	0.7,5.62	0.59,1.33	**0.82**,0.06	0.03,**0.007**	0.14,1.11	0.26,0.79	0.28,0.80	0.11,0.03

Table 5 NPPMC values and matching time t_m (msec) are listed here for all the techniques on the Gardens Point dataset. Because the matching time remains the same for all datasets, it is only specified for a single dataset. The values of t_m are averaged over the entire dataset, i.e, over the total number of query images. Using data in Table 4, NPPMC values for other datasets can be computed as well. The last row shows feature descriptor sizes of all 8 VPR techniques in Kilo-Bytes (KBs) for a single image. The bold values in each row represent the state-of-the-art technique for the corresponding metric.

Dataset Name	NetVLAD	RegionVLAD	CoHOG	HOG	AlexNet	AMOSNet	HybridNet	CALC
t_m (msec)	0.003	0.07	0.60	0.02	**0.002**	0.01	0.01	**0.002**
NPPMC (Z=10)	9	37	512	**2451**	29	67	64	455
NPPMC (Z=100)	9	36	281	**1868**	29	67	64	453
NPPMC (Z=1000)	9	34	50	**553**	29	66	62	433
NPPMC (Z=5000)	9	28	10	133	28	60	57	**362**
Descriptor Size (KBs)	16.38	786	123	14.59	**4.25**	61.4	61.4	**4.25**

framework and listed in Table 3. Techniques (e.g, CoHOG, CALC) which have lower encoding times and reasonable matching precision, achieve higher values for *PCU*. On the other hand, techniques (e.g, HOG, AlexNet) that have very low precision, but are computationally very efficient, still get lower *PCU* due to the poor matching performance. Highly-precise but computationally expensive techniques (e.g, NetVLAD, RegionVLAD, HybridNet) lie in between the 2 extremes for PCU performance. VPR techniques with higher EP values suggest that these techniques should be employed in applications where false positives are catastrophic. However, evidently most techniques have a low EP value on all the datasets except the Living Room and 24/7 Query datasets. This proposes that contemporary VPR techniques are not immune to false positives even under low recall and therefore, presents a significant room for improvement on this front. VPR techniques that cannot reach 100% precision at any recall value are extremely penalised by the EP metric and get a score equal to zero, while VPR techniques with an EP score of 0.5 identify techniques that can achieve a 100% precision at minimum recall.

Since the matching precision may not reflect the real-world usage of a VPR technique, we utilise the accuracy A values, feature encoding times and descriptor matching times (as listed in Table 4 and Table 5) to estimate the NPPMC values for all techniques. As NPPMC values are linearly related to L and V, we do not show the trend of NPPMC variation for different values of L

and V. However, by assuming a linear search metric for O(Z) (i.e, a query image is matched against all reference images in the map), we compute the NPPMC for different values of Z, because this trend may or may not be similar across VPR techniques due to equation 5. For a trajectory-length (L) of 1000 meters and a platform-speed (V) of 10 mps, the NPPMC values are listed for different map-sizes (no. of reference images Z) in Table 5 for the Gardens Point dataset. The extensive information provided in Table 4 allows the reader to compute NPPMC values for other datasets as well, which has been skipped in this manuscript to avoid redundancy. Examples of images matched/mismatched by all VPR techniques on the 10 datasets are shown in Fig.10 for a qualitative insight.

Some of the key findings from this analysis can be summarised as below:

1. Unlike previous evaluations, where state-of-the-art AUC performance was almost always achieved by NetVLAD, this paper shows that state-of-the-art AUC performance is widely distributed among all the techniques across the 10 datasets.
2. State-of-the-art technique on a particular dataset is metric-dependent and therefore, application-specific. A computationally-restricted application may find metric like PCU relevant, while computationally-powerful platforms may only utilise AUC. On the other hand, false-positive sensitive systems may find EP useful.
3. Computationally-efficient and moderately precise techniques can present much more place matching candidates than highly precise but computationally-expensive techniques, as evidenced by NPPMC performance. This does come at the cost of proportionally more false-positives, and systems that are either robust to false-positives or can either predict or discard such false-positives should consider NPPMC performance.
4. Contrary to existing beliefs, simple hand-crafted place recognition techniques can also achieve state-of-the-art performance. This paper shows how HOG and CoHOG have achieved state-of-the-art performance for all metrics on at least one dataset.
5. Applications where the explored environment is small (e.g, a home service robot as in the Living Room dataset) and the variations are moderate, it is better to use a handcrafted computationally-efficient technique.
6. VPR techniques are far from ideal performance based on the EP metric, which presents significant room for improvement.
7. Because state-of-the-art performance is distributed across the entire set of VPR techniques, an ensemble-based approach presents more value to VPR than a single-technique-based VPR, provided that the high computational and storage requirements of an ensemble can be afforded.
8. Image retrieval time is dominated by descriptor matching time at large values of Z and by feature encoding time at small values of Z. Therefore, any application-specific selection of VPR techniques will depend on the size of the map.
9. A perfect AUC score (i.e, equal to one) does not mean that a technique has correctly matched all the images in the dataset, but instead that a matching score threshold exists above which all images were correctly matched and below which all images were mismatched. Thus, it is important that the accuracy (A) of VPR techniques is also reported in addition to AUC. See for example the AUC and accuracy of HOG on the Living Room dataset.

4.2 Acceptable Ground-truth Manipulation

An important finding from the analysis performed for sub-section 4.1 was that the matching performance also varies depending on the ground-truth place matching information in a VPR dataset. It is possible that the ground-truth is slightly modified such that the new ground-truth is usually acceptable to the reviewing audience, but it also leads to a change of state-of-the-art technique on a particular dataset. For example, the matching performance varies if the query and reference traverses are inter-changed, especially for conditionally-variant datasets. We show this in Fig. 11 for the Nordland and GardensPoint dataset. This is important when the matching performance changes are observed in reference to each other for all the VPR techniques, as the rise/decline in performance is not necessarily the same in magnitude and direction for all techniques.

Moreover, in most of the traversal-based VPR datasets, there is always some level of overlap in visual content in between consecutive frames. Thus, techniques which are viewpoint-invariant may get benefits if the ground-truth identifies such frames as correct matches. On the other hand, if the ground-truth only considers frame-to-frame matches (i.e, one query frame has only one correct matching reference frame), such viewpoint-invariant techniques may not get the same matching performance (in the form of AUC, PCU, EP, Accuracy etc), because their viewpoint invariance will actually lead to false positives. Examples of these consecutive frames with visual overlap are shown in Fig. 12. We report this effect of changing ground-truth range on the AUC of various VPR techniques in Fig. 13. One

Fig. 10 Exemplar images matched/mismatched by VPR techniques are shown here for a qualitative insight. Red bounded images are incorrect matches (false positives) and green-bounded images are correct matches (true positives). An image is taken from each of the 10 datasets. For the convenience of reader's reference, all of the exemplar images are selected as the first query image in each dataset. An important insight here is that some images are matched by all of the techniques, irrespective of the technique's complexities and abilities. This figure also suggests that because all of the images are matched by at least 1 technique, an ensemble-based approach can significantly improve matching performance of a VPR-system.

could argue that a correct ground-truth must regard such viewpoint-variant images of the same place as true positives, however, a contrary argument exists for applications that utilise VPR as the primary and only module for localisation, as discussed further in subsection 4.5. This sub-section demonstrates that different state-of-the-arts (i.e, top performing techniques) can be created on the same dataset by manipulating the ground-truth information accordingly.

4.3 Invariance Analysis

One of the key aspects of the VPR-Bench framework as explained in Section 3 is the quantification of viewpoint- and illumination-invariance of a VPR technique. In sub-section 4.1, we had utilised the traditional VPR analysis schema, where datasets are usually classified based on the qualitative severity of a particular variation. However, in this section, we utilise the Point Features dataset presented in sub-section 3.5 and utilise

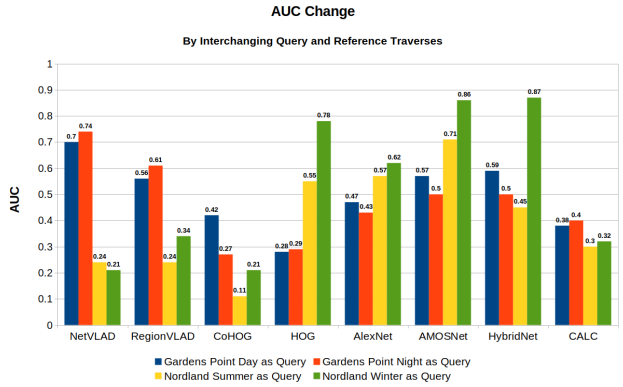

Fig. 11 The effect on AUC performance of techniques by inter-changing the query and reference traverses is shown here for the Gardens Point dataset and Nordland dataset.

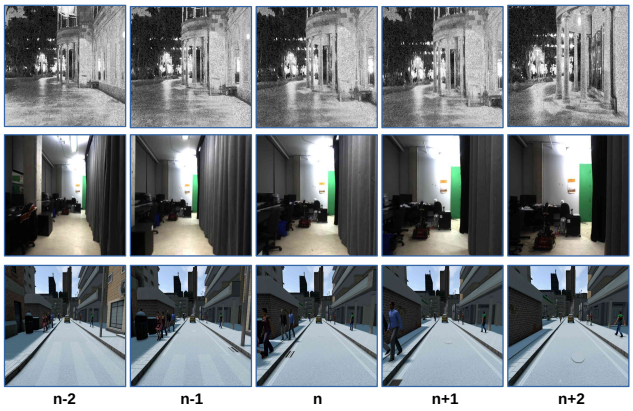

Fig. 12 The overlap between visual information among subsequent images in traversal-based datasets is shown here. Depending on what level of ground-truth true positive range is acceptable, benefits will be distributed among the techniques based on their viewpoint-invariance.

the quantitative information presented in Fig. 7, Fig. 8 and Table 2.

There are a total of 119 different viewpoint positions and 19 different illumination levels. We consider the illumination case 1 in Fig. 8 and the left-most point on Arc 1 as our keyframe(s) for viewpoint- and illumination-invariance analysis, respectively. For each analysis and each VPR technique, the key-frame is matched with itself to provide an ideal matching score, i.e, 1. For viewpoint-variation analysis, we keep the illumination level constant, move along Arc 1 in a clock-wise fashion and compute the matching scores between the keyframe and viewpoint-varied (quantified) images. The same is repeated for Arcs 2 and 3, where the keyframe remains the same i.e, the left-most point on Arc 1. This yields a total of 119 different matching scores for each of the 119 different viewpoint positions. The change in matching score along these arcs is

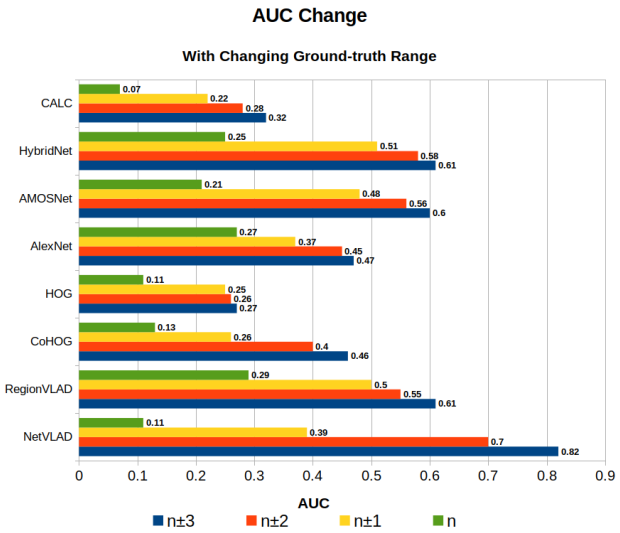

Fig. 13 The effect on AUC performance of techniques by changing the range of ground-truth true positive images is shown here for the Gardens Point dataset and Nordland dataset.

shown in Fig. 15 for all the techniques. There is clear decline in matching scores as the viewpoint is varied both along the arcs and in-between the arcs. A key insight is that moving along the arcs has more effect (negative) on the matching score than jumping between the arcs (i.e, moving towards or away from the scene). From a computer vision perspective, this means that a change in the scale of the world (zooming-in, zooming-out) has lesser effect on matching scores than the change in 3D-appearance of the scene. Because the decline in matching score itself does not provide too much insight, we draw the matching scores for the same scene in Point Features dataset, along with the matching scores when the reference scene is a different place (i.e, the query/keypoint frame and reference frame are different places). The viewpoint position and illumination level are the same for the curve of different place/scene, as they are for the same place/scene.

Ideally, the matching scores for the same scene/place should be equal to 1 for the range of variation a technique can handle and the matching score for a different scene/place should be 0. However, in practice, all techniques give lower than 1 matching scores, when 2 images of a scene have a particular variation in-between them, while giving higher than 0 scores to places that are different. The point at which the matching score for the same-but-varied place is equal to or lower than 'any' of the the matching scores for different place, represents the absolute limits for that VPR technique. Please note, that the 2 curves (same-but-varied place and different place)

Fig. 14 The change in appearance of a scene for 19 different illumination levels is shown here from the Point Features dataset.

should not be compared point-to-point, but instead point-to-curve, because the matching score for the same-but-varied place should not be less than 'any' of the matching scores for different place. Thus, while it may appear that the 2 curves for NetVLAD do not intersect under any viewpoint positions, the matching score for the same-but-varied place for positions $110 - 119$ is almost equal to the matching score for different place at position 0, which will lead to false positives. A conclusive remark from this viewpoint-variation analysis is that none of the 8 VPR techniques in this work is immune to all levels of viewpoint-variation. We have also computed the Area-between-the-Curves (AbC), where the 2 curves represent matching scores for the same-but-varied and different places, for each of the techniques, which have been reported for all the techniques. Higher value of AbC represents that a technique can distinguish well between the same-but-varied place and a different place. The ideal value of AbC is equal to the number of variations (x-axis), as the matching score should remain 1 along the entire x-axis in an ideal scenario. Please note that the AbC does not reflect the absolute matching performance of a VPR technique, and should not be compared with AUC/EP/PCU, because the analysis in only based on 2 places/scenes.

A similar analysis is performed for the 19 different matching scores given the 19 quantified illumination variations, as shown in Fig. 15. While the 119 different viewpoint positions represented in Fig. 7 are intuitive for analysis, the nature and level of illumination change in Table: 2 is not obvious. We have presented these 19 different cases qualitatively in Fig. 14, so that the illumination-variance curves in Fig. 15 can be further understood. It can be seen that uniform or close to uniform changes do not have much effect on the matching score. However, directional illumination changes that lead to the partitioning of a scene between highly-illuminated and low-illuminated portions has the most dramatic effect. An interesting insight is that some basic handcrafted VPR techniques (HOG-based) are able to distinguish between the same-but-illumination-varied places and different places, under all 19 scenarios (i.e, no point on the same-but-varied place curve is lower than any point on the different place curve). While, contemporary deep-learning-based techniques struggle with such illumination-variation.

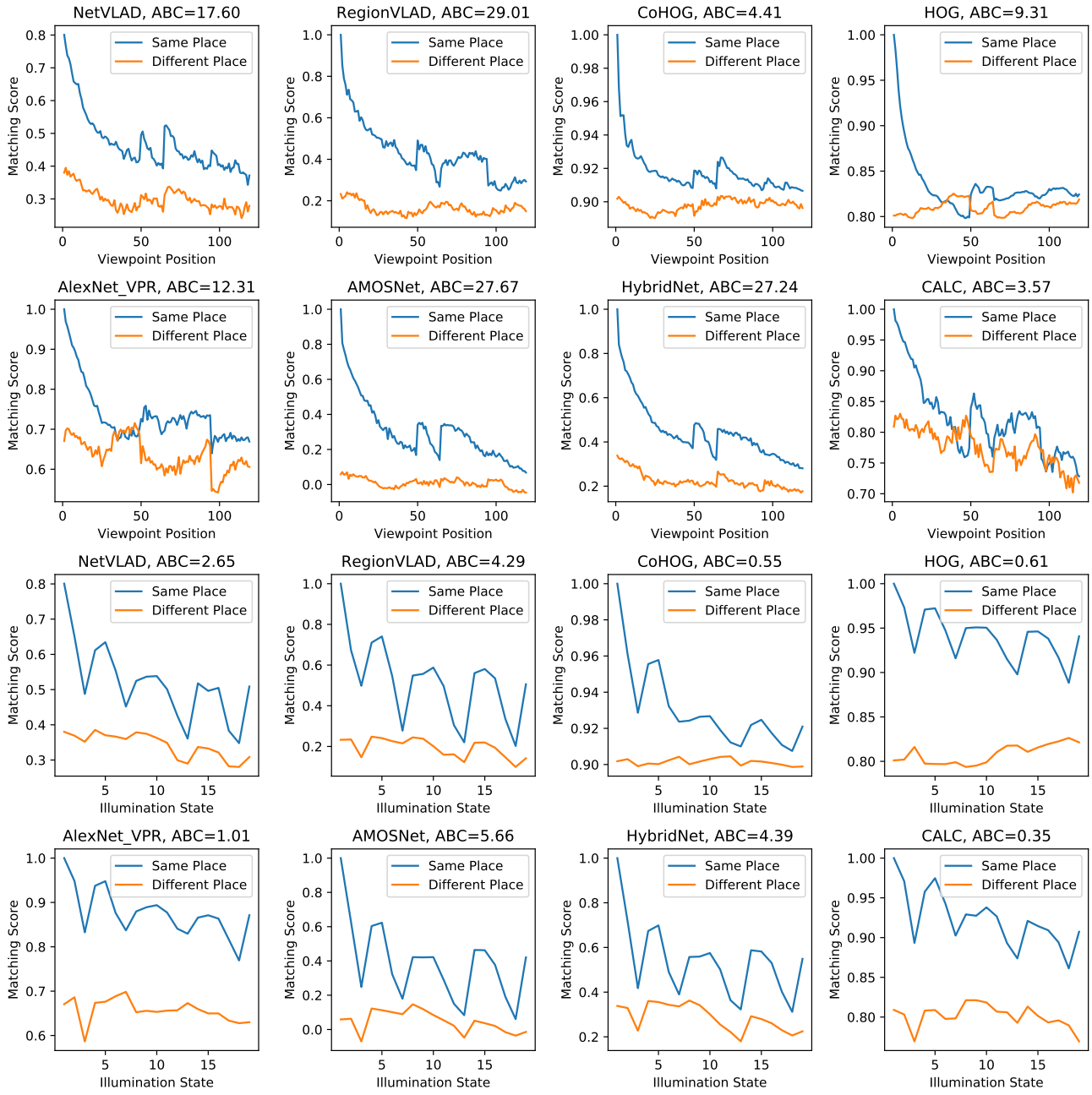

Fig. 15 The change in matching score for quantified viewpoint and illumination variations is shown here on the Point Features dataset. The first two rows contain changes for all techniques with 119 viewpoint positions, while the bottom two row show these changes for 19 different illumination levels. Please see accompanying text for analysis.

4.4 Retrieval Time vs Platform Speed

One of the question that we wanted to address through this manuscript is, 'What is a good image-retrieval time?'. This is important because most VPR research papers (as covered in our literature review) that claim real-time performance consider anything between 5-25 frames-per-second (FPS) as real-time. However, there are 2 important caveats to such performance. Firstly, the retrieval performance for a VPR application is dependent on the size of the map. It is therefore very important that the size of the map is addressed either by presenting the limits for the map-size or by proposing methodologies to affix the map-size. Secondly, the retrieval performance is directly related to the platform speed. A real-time VPR application may require that a

place-match (localisation) is achieved every few meters, while a dynamic platform traverses an environment. In such a case, the utility of a technique will depend upon the speed of the platform, as the faster the platform moves, the lower the retrieval time that is acceptable. We have modelled this as following.

Let us assume that a particular application requires K frames-per-meter (where K could be fractional) and that the platform moves with a velocity V. Also, let the size of the map (no. of reference images) be Z. Then, the required FPS retrieval performance given the values of K and V is denoted as FPS_{req} and computed as;

$$FPS_{req} = K \times V \tag{8}$$

The retrieval performance of a VPR technique will depend on the number of reference images and can be denoted as FPS_{VPR}. This FPS_{VPR} has been modelled previously in equation 5, such that $FPS_{VPR} = 1/t_R$. Therefore, to understand the limits of real-time performance of a VPR technique given the application requirements (V, K and Z), we draw the retrieval performance of all techniques along the platform speed for different values of Z in Fig. 16, assuming $K = 0.5$ frames-per-meter. The curves for FPS_{VPR} are straight-lines for constant values of Z and the range of horizontal-axis (Speed V) for which FPS_{VPR} is less than or equal to FPS_{req} represents the range of platform speed (for that map-size) that a technique can handle. The VPR-Bench framework enables the creation of these curves conveniently and therefore, presents value to address the subjective real-time nature of a technique's retrieval time for VPR.

4.5 Variance vs Invariance

A generic perception among the VPR research community, as evident from the recent trend in developing highly viewpoint-invariant VPR techniques, is that the more viewpoint-invariant a technique is, the more utility it has to offer. Through this sub-section, we take the opportunity to address that this may not always be the case. In fact, viewpoint-variance may actually be required in some applications, instead of viewpoint-invariance. A key example here is the applications where VPR techniques act as the primary localisation module and where, there is no image-to-image, epipolar-geometry-based motion estimation (location refinement) module. For example, Zeng et al. (2019) extend the concept of VPR for precise localisation in mining environments. Similar extensions of VPR as the only module for precise-localisation are possible in several applications, where an accurate geo-tagged image database of the environment exists, e.g, in factory/plant environments or outdoor applications which can afford to create an *a priori* accurate appearance-based metric/topo-metric map of the environment. For such applications, VPR techniques are required to have viewpoint-variance, so that even if the 2 images of the same place are viewpoint-varied, the VPR technique can distinguish between them to perform metrically-precise localisation. If a viewpoint-invariant technique is utilised in this scenario, the inherent viewpoint-invariance will lead to discrepancies in localisation estimates and eventually cause a system failure.

Thus, a key area to investigate within VPR research should be controlled viewpoint-variance. In subsection 4.3, we presented a methodology to estimate the viewpoint-invariance of a technique, however, there is no control parameter for any technique that could govern and tune its invariance to viewpoint changes. We believe that this is an exciting research challenge and should be a topic for VPR research in the upcoming years. Nevertheless, our proposal is that both viewpoint-variance and invariance are desirable properties, depending upon the underlying application and should be regarded/investigated accordingly.

5 Conclusions and Future Work

In this paper, we presented a comprehensive and variation-quantified evaluation framework for visual place recognition performance. This open-source framework, namely VPR-Bench, integrates 10 different indoor and outdoor datasets, each representing a unique challenge. Along with 8 contemporary VPR techniques, the framework provides several different evaluation metrics to assess the performance of techniques on various fronts. We also propose a new metric to bridge the gap between dataset-based evaluation metrics and their extension to real-world applications. The framework design is modular and permits future integration of datasets, techniques and metrics in a convenient manner. We utilised the variation- and illumination-quantified Point Features dataset to evaluate and analyse the level and nature of variations that a VPR technique can handle.

Using our framework, we provide several useful insights about the nature of challenges that a particular technique can handle. We identify that no universal state-of-the-art technique exists for place matching precision and discuss the reasons behind the success/failure of these techniques from one dataset to another. We also propose that the utility of VPR techniques is highly divergent based on the employed evaluation metric and that the corresponding utility is application-dependent.

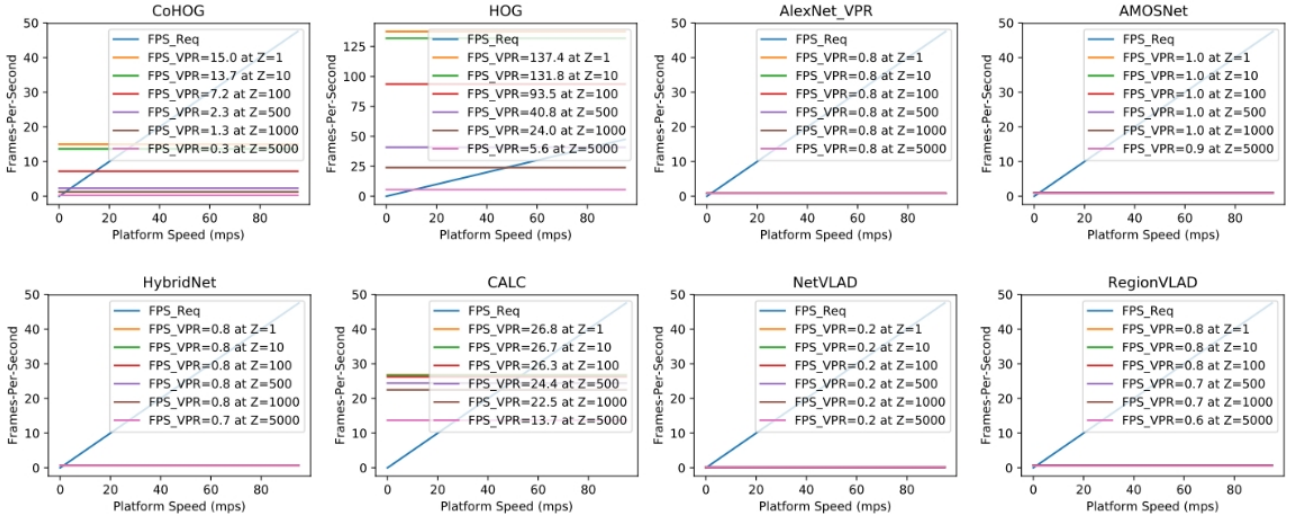

Fig. 16 The retrieval performance of techniques is drawn for different map-sizes (Z) across the platform speed. Depending upon the value of frames required per meter (K) for an application, these curves will scale linearly according to equation 8.

The analysis also shows that there is no one-for-all evaluation metric for VPR research and that only a combination of these metrics presents the overall utility of a technique.

We develop our analysis around the Point Features dataset for viewpoint and illumination-invariance quantification and integrate this analysis within the framework for ease-of-use of VPR researchers. Additionally, we also present other useful insights for the VPR research community, including the effects of acceptable ground-truth manipulation, variance vs invariance and the subjective real-time nature of a technique's retrieval performance. Because we have employed several different datasets, techniques and metrics, the dimensions of comparative performances enabled by VPR-Bench is very high and we have only discussed/analysed a few of these comparisons to limit the scope. It would be useful to further investigate, for example, the trends of NPPMC variation between datasets, techniques and even based on the bottle-necks caused by encoding times and linear scaling of matching times with the number of reference images. Further insights can also be presented by evaluating in-depth, how different metrics yield different state-of-the-art VPR techniques on the same dataset.

We hope that this work proves useful for the VPR community and that all subsequent evaluations and/or newly proposed techniques employ our framework to present a detailed comparison with the state-of-the-art techniques, on the many datasets, using the various evaluation metrics. We are also very keen on integrating more open-source VPR techniques into the VPR-Bench framework and would be very encouraging towards any such feedback, collaborations and suggestions.

Acknowledgements Our work was supported by the UK Engineering and Physical Sciences Research Council through grants EP/R02572X/1, EP/P017487/1 and EP/R026173/1 and in part by the RICE project funded by the National Centre for Nuclear Robotics Flexible Partnership Fund. Michael Milford was supported by ARC grants FT140101229, CE140100016 and the QUT Centre for Robotics.

References

Aanæs H, Dahl AL, Pedersen KS (2012) Interesting interest points. International Journal of Computer Vision 97(1):18–35

Agarwal S, Furukawa Y, Snavely N, Simon I, Curless B, Seitz SM, Szeliski R (2011) Building rome in a day. Communications of the ACM 54(10):105–112

Agrawal M, Konolige K, Blas MR (2008) Censure: Center surround extremas for realtime feature detection and matching. In: European Conference on Computer Vision, Springer, pp 102–115

Andreasson H, Duckett T (2004) Topological localization for mobile robots using omni-directional vision and local features. IFAC Proceedings Volumes 37(8):36–41

Angeli A, Doncieux S, Meyer JA, Filliat D (2008) Incremental vision-based topological slam. In: IROS, Ieee, pp 1031–1036

Arandjelovic R, Gronat P, Torii A, Pajdla T, Sivic J (2016) Netvlad: Cnn architecture for weakly supervised place recognition. In: CVPR, pp 5297–5307

Badino H, Huber D, Kanade T (2012) Real-time topometric localization. In: ICRA, IEEE, pp 1635–1642

Bay H, Tuytelaars T, Van Gool L (2006) Surf: Speeded up robust features. In: ECCV, Springer, pp 404–417

Cadena C, Carlone L, Carrillo H, Latif Y, Scaramuzza D, Neira J, Reid I, Leonard JJ (2016) Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE T-RO 32(6):1309–1332

Calonder M, Lepetit V, Ozuysal M, Trzcinski T, Strecha C, Fua P (2011) Brief: Computing a local binary descriptor very fast. IEEE T-PAMI 34(7):1281–1298

Camara LG, Gäbert C, Preucil L (2019) Highly robust visual place recognition through spatial matching of cnn features. ResearchGate Preprint

Chancán M, Hernandez-Nunez L, Narendra A, Barron AB, Milford M (2020) A hybrid compact neural architecture for visual place recognition. IEEE Robotics and Automation Letters 5(2):993–1000

Chen Z, Lam O, Jacobson A, Milford M (2014) Convolutional neural network-based place recognition. preprint arXiv:14111509

Chen Z, Maffra F, Sa I, Chli M (2017a) Only look once, mining distinctive landmarks from convnet for visual place recognition. In: IROS), IEEE, pp 9–16

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

Chen Z, et al. (2017b) Deep learning features at scale for visual place recognition. In: ICRA, IEEE, pp 3223–3230

Cieslewski T, Choudhary S, Scaramuzza D (2018) Data-efficient decentralized visual slam. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 2466–2473

Cummins M, Newman P (2011) Appearance-only slam at large scale with fab-map 2.0. IJRR 30(9):1100–1123

Davison AJ, Reid ID, Molton ND, Stasse O (2007) Monoslam: Real-time single camera slam. IEEE transactions on pattern analysis and machine intelligence 29(6):1052–1067

Demir M, Bozma HI (2018) Automated place detection based on coherent segments. In: 2018 IEEE 12th International Conference on Semantic Computing (ICSC), IEEE, pp 71–76

DeTone D, Malisiewicz T, Rabinovich A (2018) Superpoint: Self-supervised interest point detection and description. In: CVPR Workshops, pp 224–236

Dusmanu M, et al. (2019) D2-net: A trainable cnn for joint description and detection of local features. In: CVPR, pp 8092–8101

Ferrarini B, Waheed M, Waheed S, Ehsan S, Milford MJ, McDonald-Maier KD (2020) Exploring performance bounds of visual place recognition using extended precision. IEEE Robotics and Automation Letters 5(2):1688–1695

Filliat D (2007) A visual bag of words method for interactive qualitative localization and mapping. In: ICRA, IEEE, pp 3921–3926

Fraundorfer F, Engels C, Nistér D (2007) Topological mapping, localization and navigation using image collections. In: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, pp 3872–3877

Girdhar Y, Dudek G (2010) Online navigation summaries. In: 2010 IEEE International Conference on Robotics and Automation, IEEE, pp 5035–5040

Ho KL, Newman P (2007) Detecting loop closure with scene sequences. IJCV 74(3):261–286

Hou Y, Zhang H, Zhou S (2018) Evaluation of object proposals and convnet features for landmark-based visual place recognition. Journal of Intelligent & Robotic Systems 92(3-4):505–520

Jégou H, Douze M, Schmid C, Pérez P (2010) Aggregating local descriptors into a compact image representation. In: CVPR, IEEE Computer Society, pp 3304–3311

Jin Y, Mishkin D, Mishchuk A, Matas J, Fua P, Yi KM, Trulls E (2020) Image matching across wide baselines: From paper to practice. arXiv preprint arXiv:200301587

Johns E, Yang GZ (2011) From images to scenes: Compressing an image cluster into a single scene model for place recognition. In: 2011 International Conference on Computer Vision, IEEE, pp 874–881

Khaliq A, Ehsan S, Chen Z, Milford M, McDonald-Maier K (2019) A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes. IEEE Transactions on Robotics

Konolige K, Agrawal M (2008) Frameslam: From bundle adjustment to real-time visual mapping. IEEE Transactions on Robotics 24(5):1066–1077

Košecká J, Li F, Yang X (2005) Global localization and relative positioning based on scale-invariant keypoints. Robotics and Autonomous Systems 52(1):27–38

Kostavelis I, Gasteratos A (2015) Semantic mapping for mobile robotics tasks: A survey. RAS 66:86–103

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105

Larsson M, Stenborg E, Hammarstrand L, Pollefeys M, Sattler T, Kahl F (2019) A cross-season correspondence dataset for robust semantic segmentation. In: CVPR, pp 9532–9542

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. IJCV, Springer 60(2):91–110

Lowry S, Sünderhauf N, Newman P, Leonard JJ, Cox D, Corke P, Milford MJ (2015) Visual place recognition: A survey. IEEE Transactions on Robotics 32(1):1–19

Maddern W, Milford M, Wyeth G (2012) Cat-slam: probabilistic localisation and mapping using a continuous appearance-based trajectory. IJRR 31(4):429–451

Maddern W, Pascoe G, Linegar C, Newman P (2017) 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research 36(1):3–15

McManus C, Upcroft B, Newmann P (2014) Scene signatures: Localised and point-less features for localisation. Robotics, Science and Systems Conference

Mei C, Sibley G, Cummins M, Newman P, Reid I (2009) A constant-time efficient stereo slam system. In: Proceedings of the British machine vision conference, BMVA Press, vol 1

Merrill N, Huang G (2018) Lightweight unsupervised deep loop closure. arXiv preprint arXiv:180507703, Robotics Science and Systems Conference

Milford M (2013) Vision-based place recognition: how low can you go? The International Journal of Robotics Research 32(7):766–789

Milford MJ, Wyeth GF (2012) Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In: International Conference on Robotics and Automation, IEEE, pp 1643–1649

Mount J, Milford M (2016) 2d visual place recognition for domestic service robots at night. In: 2016 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 4822–4829

Murillo AC, Kosecka J (2009) Experiments in place recognition using gist panoramas. In: ICCV Workshops, IEEE, pp 2196–2203

Murillo AC, Guerrero JJ, Sagues C (2007) Surf features for efficient robot localization with omnidirectional images. In: Proceedings of IEEE ICRA, pp 3901–3907

Nardi L, Bodin B, Zia MZ, Mawer J, Nisbet A, Kelly PH, Davison AJ, Luján M, O'Boyle MF, Riley G, et al. (2015) Introducing slambench, a performance and accuracy benchmarking methodology for slam. In: 2015 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 5783–5790

Naseer T, Oliveira GL, Brox T, Burgard W (2017) Semantics-aware visual localization under challenging perceptual conditions. In: 2017 IEEE ICRA, pp 2614–2620

Odo A, McKenna S, Flynn D, Vorstius J (2020) Towards the automatic visual monitoring of electricity pylons from aerial images. In: 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications 2020, SciTePress, pp 566–573

Oliva A, Torralba A (2006) Building the gist of a scene: The role of global image features in recognition. Progress in brain research 155:23–36

Paul R, Feldman D, Rus D, Newman P (2014) Visual precis generation using coresets. In: 2014 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 1304–1311

Pepperell E, Corke PI, Milford MJ (2014) All-environment visual place recognition with smart. In: 2014 IEEE international conference on robotics and automation (ICRA), IEEE, pp 1612–1618

Ranganathan A (2013) Detecting and labeling places using runtime change-point detection and place labeling classifiers. US Patent 8,559,717

Robertson DP, Cipolla R (2004) An image-based system for urban navigation. In: Bmvc, Citeseer, vol 19, p 165

Ros G, Sellart L, Materzynska J, Vazquez D, Lopez AM (2016) The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3234–3243

Rosten E, Drummond T (2006) Machine learning for high-speed corner detection. In: ECCV, Springer, pp 430–443

Sahdev R, Tsotsos JK (2016) Indoor place recognition system for localization of mobile robots. In: 2016 13th Conference on Computer and Robot Vision (CRV), IEEE, pp 53–60

Sarlin PE, Cadena C, Siegwart R, Dymczyk M (2019) From coarse to fine: Robust hierarchical localization at large scale. In: CVPR, pp 12716–12725

Schönberger JL, Pollefeys M, Geiger A, Sattler T (2018) Semantic visual localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6896–6906

Se S, Lowe D, Little J (2002) Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. IJRR 21(8):735–758

Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2014) Overfeat: Integrated recognition, localization and detection using convolutional networks. In: 2nd International Conference on Learning Representations, ICLR 2014

Singh G, Kosecka J (2010) Visual loop closing using gist descriptors in manhattan world. In: ICRA Omnidirectional Vision Workshop, pp 4042–4047

Sivic J, Zisserman A (2003) Video google: A text retrieval approach to object matching in videos. In: null, IEEE, p 1470

Skrede S (2013) Nordland dataset. `https://bit.ly/2QVBOym`

Stenborg E, Toft C, Hammarstrand L (2018) Long-term visual localization using semantically segmented images. In: 2018 IEEE ICRA, pp 6484–6490

Stumm E, Mei C, Lacroix S (2013) Probabilistic place recognition with covisibility maps. In: IROS, IEEE, pp 4158–4163

Sünderhauf N, Protzel P (2011) Brief-gist-closing the loop by simple means. In: IROS, IEEE, pp 1234–1241

Sünderhauf N, Shirazi S, Dayoub F, Upcroft B, Milford M (2015) On the performance of convnet features for place recognition. In: IROS, IEEE, pp 4297–4304

Tolias G, Avrithis Y, Jégou H (2013) To aggregate or not to aggregate: Selective match kernels for image search. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1401–1408

Tolias G, Avrithis Y, Jégou H (2016a) Image search with selective match kernels: aggregation across single and multiple images. International Journal of Computer Vision 116(3):247–261

Tolias G, Sicre R, Jégou H (2016b) Particular object retrieval with integral max-pooling of cnn activations. arXiv:151105879, ICLR

Topp EA, Christensen HI (2008) Detecting structural ambiguities and transitions during a guided tour. In: 2008 IEEE International Conference on Robotics and Automation, IEEE, pp 2564–2570

Torii A, Sivic J, Pajdla T, Okutomi M (2013) Visual place recognition with repetitive structures. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 883–890

Torii A, et al. (2015) 24/7 place recognition by view synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1808–1817

Wang J, Zha H, Cipolla R (2005) Combining interest points and edges for content-based image retrieval. In: IEEE International Conference on Image Processing 2005, IEEE, vol 3, pp III–1256

Zaffar M, Ehsan S, Milford M, Maier KM (2018) Memorable maps: A framework for re-defining places in visual place recognition. arXiv preprint arXiv:181103529

Zaffar M, Khaliq A, Ehsan S, Milford M, Alexis K, McDonald-Maier K (2019a) Are state-of-the-art visual place recognition techniques any good for aerial robotics? arXiv preprint arXiv:190407967 ICRA 2019 Workshop on Aerial Robotics

Zaffar M, Khaliq A, Ehsan S, Milford M, McDonald-Maier K (2019b) Levelling the playing field: A comprehensive comparison of visual place recognition approaches under changing conditions. arXiv preprint arXiv:190309107, IEEE ICRA Workshop on Database Generation and Benchmarking

Zaffar M, Ehsan S, Milford M, McDonald-Maier K (2020) Cohog: A light-weight, compute-efficient, and training-free visual place recognition technique for changing environments. IEEE Robotics and Automation Letters 5(2):1835–1842

Zeng F, Jacobson A, Smith D, Boswell N, Peynot T, Milford M (2019) Lookup: Vision-only real-time precise underground localisation for autonomous mining vehicles. In: 2019 International Conference on Robotics and Automation (ICRA), IEEE, pp 1444–1450