Garg, Sourav, Suenderhauf, Niko, & Milford, Michael
(2019)
Semantic-geometric visual place recognition: a new perspective for reconciling opposing views.
International Journal of Robotics Research.

https://doi.org/10.1177/0278364919839761

Journal Title
XX(X):1–24
©The Author(s) 2019
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Semantic-Geometric Visual Place Recognition: A New Perspective for Reconciling Opposing Views

Sourav Garg[1], Niko Suenderhauf[2] and Michael Milford[3]

Abstract

Human drivers are capable of recognizing places from a previous journey even when viewing them from the *opposite direction* during the return trip under radically different environmental conditions, without needing to look back or employ a 360 degree camera or LIDAR sensor. Such navigation capabilities are attributed in large part to the robust semantic scene understanding capabilities of humans. However, for an autonomous robot or vehicle, achieving such human-like visual place recognition (VPR) capability presents three major challenges: 1) dealing with a limited amount of commonly observable visual content when viewing the same place from the opposite direction, 2) dealing with significant lateral viewpoint changes caused by opposing directions of travel taking place on opposite sides of the road, and 3) dealing with a radically changed scene appearance due to environmental conditions like time of day, season and weather. Current state-of-the-art place recognition systems have only addressed these three challenges in isolation or in pairs, typically relying on appearance-based, deep-learnt place representations. In this paper we present a novel, semantics-based system that for the first time solves all three challenges simultaneously. We propose a *hybrid* image descriptor that *semantically* aggregates salient visual information, complemented by appearance-based description, and augment a conventional coarse-to-fine recognition pipeline with *keypoint correspondences* extracted from within the convolutional feature maps of a pre-trained network. Finally, we introduce descriptor normalization and local score enhancement strategies for improving the robustness of the system. Using both existing benchmark datasets and extensive new datasets that for the first time combine the three challenges of opposing viewpoints, lateral viewpoint shifts and extreme appearance change, we show that our system can achieve practical place recognition performance where existing state-of-the-art methods fail.

1 Introduction

Visual Place Recognition (VPR) is a key component of a visual SLAM system (Cadena et al. 2016) and one of the key capabilities of a mobile robot or autonomous vehicle. The major challenges of the problem revolve around reliably recognizing places under both a wide variety of environmental conditions that non-uniformly affect the appearance of the scene and a wide range of viewpoints. Because of both improving camera technology and its relevance to many mobile autonomy applications, VPR has been an area of growing research attention over the past decade (Lowry et al. 2016). A large number of methods have been proposed to deal with different aspects of the problem like large-scale localization, appearance variations due to changing environmental conditions, and viewpoint variations, using approaches incorporating temporal information or learning robust place representations.

Appearance-based methods like FAB-MAP (Cummins and Newman 2011) involve a large-scale localization framework utilizing a Bag of Visual Words (BoVW) (Sivic and Zisserman 2003) approach, forming an integral part of visual SLAM methods like ORB-SLAM (Mur-Artal et al. 2015) and LSD-SLAM (Engel et al. 2014). However, the original BoVW approaches were often based on hand-crafted keypoint features like SIFT (Lowe 2004), SURF (Bay et al. 2008) and ORB (Rublee et al. 2011), and as a consequence inherited the underlying feature type's limited robustness to changing environmental appearance.

While the earlier attempts at learning-based local keypoint description (Philbin et al. 2010; Brown et al. 2011; Simonyan et al. 2014) helped improve the BoVW pipeline, more promising description methods for appearance-invariance were initially based on whole-image descriptors. Such methods, for example, patch-normalized images in SeqSLAM (Milford and Wyeth 2012) and HoG (Dalal and Triggs 2005) within Network-Flow (Naseer et al. 2014), have leveraged temporal information inherent within mobile robotic applications. This general approach has resulted in a large body of research focusing on robust methods for utilizing sequential information (Hansen and Browning 2014; Pepperell et al. 2016; Lynen et al. 2017). These approaches have generally exhibited a different weakness to

[1,2,3] are with the Australian Centre for Robotic Vision at the Queensland University of Technology. This work was supported by an ARC Centre of Excellence for Robotic Vision Grant CE140100016.
[1] is supported by QUT Postgraduate Research Award.
[3] is supported by an Australian Research Council Future Fellowship FT140101229.

Corresponding author:
Sourav Garg, PhD Student, ACRV at QUT, Brisbane, Australia
Email: sourav.garg@hdr.qut.edu.au

feature-based techniques: viewpoint-dependence, as well as being of limited utility in wide-baseline stereo matching.

The recent advent of deep-learning based convolutional neural networks (CNNs) (Krizhevsky et al. 2012) have paved the way for developing robust image representations that offer the potential for both viewpoint- and appearance invariance (Sunderhauf et al. 2015), with performance found to depend on the choice of network layer (Garg et al. 2018a). The end-to-end training of VLAD (Jégou et al. 2010) in NetVLAD (Arandjelovic et al. 2016) for visual place recognition developed deep-learnt whole-image representation with both viewpoint- and appearance-invariance and is a good representation of the current state-of-the-art.

In this paper, we improve the versatility of visual place recognition beyond the current state-of-the-art by taking on the additional challenge of place recognition from *opposing viewpoints* whilst simultaneously dealing with lateral viewpoint change and extreme appearance change. The new additional challenge of opposing viewpoints is illustrated in Figure 1: while alone it might be addressable using a robust feature representation, current state-of-the-art techniques are insufficient to address it *whilst also addressing* drastic appearance and lateral viewpoint change as well. Instead, current methods performing bi-directional visual place recognition predominantly rely on a panoramic camera (Arroyo et al. 2014) or LIDAR (Wolcott and Eustice 2017) sensing setup. Whilst this is an acceptable engineering solution in some application domains, the fact that humans are able to do it without these full field of view sensors renders it a scientifically interesting challenge (Ackermann 1996; Hegarty and Waller 2004; Kozhevnikov et al. 2006). In addition, as the initial wave of self-driving car technology development starts to mature, it is becoming clear that there are still significant discrepancies between how highly engineered autonomous driving car systems navigate and how humans navigate. Further improving the capabilities of semantic-based navigation also has potential applications outside of the road-based domain to all mobile robots operating in and interacting with real-world environments and the humans that reside within those environments.

To address these challenges of viewpoint and appearance variations, our approach introduces a number of new techniques based on visual semantics. We exploit advances in the field of semantic scene understanding based on dense semantic image segmentation (Long et al. 2015; Lin et al. 2017) to provide salient cues for both representing and matching the places. Furthermore, we develop an approach using the visual semantic information captured by higher-order layers of the deep network for temporally segmenting the environment into meaningful chunks. Such a segmentation of the environment is useful when appearance variations may occur *both* within the environment and across multiple traverses of the same environment, and can be thought of as an updated, semantically-informed version of the simple chunking used in SeqSLAM (Milford and Wyeth 2012). The semantic information is used to inform and adapt the sequence-based place recognition methods to deal with such variations. We also improve upon the limitations of current deep-learning-based place recognition techniques, which typically provide a whole-image description which cannot be employed for estimating the scene-structure similarity between the matching pair of images. The spatial arrangement of visual landmarks in the image is a vital piece of information that can help reduce false positives (Oliva and Torralba 2001; Gálvez-López and Tardos 2012; Noh et al. 2017). We propose CNN-based keypoint matching that uses semantic filtering and dense descriptor weighing as a part of *fine* place search procedure to improve candidate match selection.

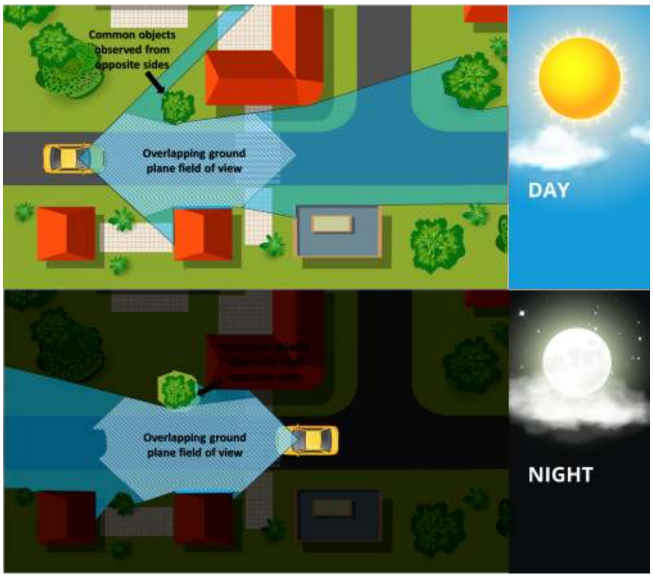


Figure 1. Recognizing places from opposite viewpoints under varying environmental conditions is a very difficult problem due to the additional challenge of limited visual overlap available for image matching, on top of the normal problems of appearance change and lateral viewpoint change.

This paper builds upon earlier work (Garg et al. 2017, 2018a,b) but adds a large number of new contributions in terms of approach, experimental evaluation and the capability of the system. The key contributions include:

- a robust visual place recognition pipeline that leverages visual semantic information *simultaneously* at three different levels: semantic segmentation at database level, semantically-salient description at image level, and semantically-consistent image matching at pixel-level,
- a *hybrid* whole-image place descriptor that complements the strength of an end-to-end learnt representation with a *viewpoint-invariant* and *explicit* aggregation of visual semantics-based description,
- an implicit keypoint correspondence method for spatial layout verification that does not require cross-matching between the keypoint descriptors of a given pair of images,
- demonstration of the limitations of state-of-the-art image representations to recognize places when there are transitions in the environment such that the appearance variations occur both within *and* across the traversals, even with similar viewpoints,
- a *transition-aware* semantic environment segmentation method for local enhancement of place matching

scores to deal with perceptual aliasing caused by transitions in environmental conditions, such as different lighting conditions,

- performance benchmarks on new datasets achieving significantly higher precision-recall performance than existing state-of-the-art systems, including the most challenging VPR datasets to date which combine:
    - opposite-viewpoints with very limited visual overlap between the images captured at the same place,
    - significant lateral viewpoint change due to different lanes of travel during forward and reverse traverses, and
    - large appearance variation due to changing environmental conditions caused by factors such as day-night cycles.
- a collection of new custom datasets for research and development spanning over 800 km of forward-reverse journeys along urban, rural and forested roadways under a range of environmental conditions including rain and day-night cycles.

In particular, we extend our previous work in the following ways. We improve the system's overall robustness by introducing a new *hybrid* whole-image descriptor based on semantic aggregation (Garg et al. 2018b) and appearance-based end-to-end training (Arandjelovic et al. 2016). The original semantic environment segmentation proposed in (Garg et al. 2017) is modified to make use of $fc6$ layer activations and *transition-aware* segmentation in order to account for false matches during the transition. We conduct a new investigation into the mutual influence of whole-image descriptor normalization (Garg et al. 2018a) and local score enhancement (Garg et al. 2017) on system performance using state-of-the-art image description techniques. We also establish the backward-compatibility of our proposed approach as applied to traditional front-view only VPR and showing significant performance gains for day-night image matching. Finally, for the first time in the literature, we demonstrate practical visual place recognition performance on datasets that combine the three challenges of opposing viewpoints, lateral viewpoint shifts, and extreme appearance change, using only a limited field-of-view monocular camera.

The paper proceeds as follows: Section 2 discusses the relevant literature with respect to the components of our proposed approach, Section 3 describes all the modules of the proposed VPR pipeline, Section 4 describes the datasets used, ground truth, and evaluation methods, Section 5 shows performance evaluation of our system against the state-of-the-art methods, Section 6 discusses some key insights related to ground truth of front-rear image matching and PCA (Principal Component Analysis) visualization of normalized image descriptors, and finally Section 7 concludes the paper outlining the scope for future work.

2 Related Work

In this section, we review the existing body of work related to specific components of our proposed approach and the kind of VPR challenges addressed in literature. This mainly includes robust representation of images, use of semantics for VPR, methods dealing with CNN-based keypoint correspondences, and the extent of appearance- and viewpoint-invariance addressed by state-of-the-art methods.

2.1 Image Representations

The early methods based on deep-learnt image representations leveraged activations of different layers of pre-trained CNNs as off-the-shelf image descriptors for image retrieval (Razavian et al. 2014) as well as place recognition (Chen et al. 2014). Further to that, a variety of pooling methods have since been developed for improving these image representations, for example, sum pooling (Babenko and Lempitsky 2015), cross-convolutional pooling (Liu et al. 2015), VLAD pooling (Yue-Hei Ng et al. 2015), integral max-pooling (Tolias et al. 2016), and multi-scale orderless pooling (Gong et al. 2014). End-to-end training for visual place recognition based on deconvolutions (Mukherjee et al. 2017), multi-scale encoding (Chen et al. 2017a), and generalized VLAD (Arandjelovic et al. 2016) has helped in developing even more robust place representations.

Furthermore, recent work has demonstrated that higher-order contextual information and all-in-one monolithic CNNs can lead to even more robust and efficient representations. Kim et al. (2017a) use semi-global context to learn weights for spatial activations. Sarlin et al. (2018b) use a teacher-student network to efficiently localize using a hierarchical framework and further extend their work by learning local keypoints and descriptors using a monolithic CNN for accurate 6-DoF localization (Sarlin et al. 2018a). The SAANE method proposed by Seymour et al. (2018) learns to fuse appearance and higher-order layers of CNN to learn novel embeddings for matching places under a wide variety of environmental conditions. The aforementioned techniques, however, indirectly learn to embed higher-order contextual information and can be used as a drop-in replacement for the end-to-end learnt representation (NetVLAD) used in this paper; we demonstrate that such a learnt representation can be complemented with explicit semantic information-based description.

2.2 Visual Semantics for Retrieval

Visual semantic information has often been used for improving image retrieval and visual place recognition. Semantic masking of images for appearance-invariance (Naseer et al. 2017) and semantic graph matching for multi-view localization (Gawel et al. 2018) have been proposed for improving the robustness of visual place recognition and localization. However, environment-specific training and the requirement of accurate semantic labels tend to limit the potential use of these methods respectively. The Semantic Segment Layout Descriptor (SSLD) is proposed by Castaldo et al. (2015) for the challenging problem of cross-view matching, for example, from an aerial top-view to the ground vehicle front-view. However, the system has only been demonstrated to work for different sensor modalities and even then only with images captured under ideal environmental conditions. Mousavian and Kosecka (2016) combine visual semantics with plane estimation to infer unique building facades in the scene in order to localize under wide viewpoint variations.

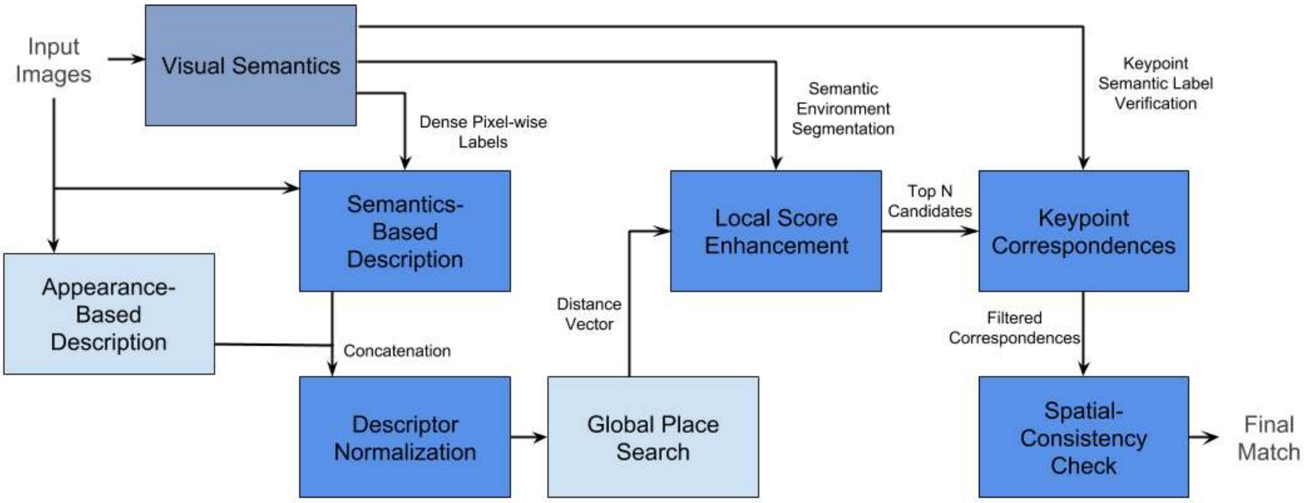


Figure 2. Proposed Visual Place Recognition Pipeline

While the use of buildings-based semantics for localization leads to robustness against extreme variations in scene appearance, other semantic classes can also aid in improving the robustness of a semantics-based system (Naseer et al. 2017; Gawel et al. 2018; Garg et al. 2018b).

Furthermore, many of the existing image retrieval methods based on visual semantics (Johnson et al. 2015; Schuster et al. 2015; Lu et al. 2017; Huang et al. 2018) have not been demonstrated to work for changing environmental appearance due to day-night or seasonal cycles.

2.2.1 Semantics Within Descriptors: Singh and Košecká (2016) and Toft et al. (2017) propose a grid-based image descriptor constructed by concatenating a semantic histogram defined per grid cell. However, this concatenation leads to a viewpoint-dependent description of images and cannot be employed for matching images from opposing viewpoints. Yu et al. (2018) propose a semantic edges-based VLAD description that encodes the probability of observing a semantic edge at a pixel location, unlike our proposed LoST descriptor that uses the probability of belonging to a semantic class in order to weight the dense convolutional descriptors at those pixel locations. Further, we use keypoint locations in a subsequent step to verify the spatial layout between the pairs of images, whereas Yu et al. (2018) use these 2D keypoint locations as a part of their descriptor which might lead to viewpoint-dependence.

2.2.2 Object Semantics and Depth for Localization: Another existing body of work that uses visual semantic information for improving localization is based on object-level semantics. Ardeshir et al. (2014) combine GIS data and object semantics to improve object detection and geospatial localization Atanasov et al. (2016) use windows and cars as objects for localization, however, the utility of cars for long-term localization has not been discussed. Furthermore, depth information along with object semantics (Atanasov et al. 2016; Salas-Moreno et al. 2013; Schreiber et al. 2013) has commonly been used to improve localization where Salas-Moreno et al. (2013) demonstrate an object-based SLAM system for indoor environments and Schreiber et al. (2013) demonstrate the use of lane markings for precise localization. Although depth as an additional information signal often leads to an improved place recognition and localization system (Garg et al. 2019) , object-semantics based systems can lack the ability to model semantic entities that do not have a precise bounding box, for example, roads, vegetation, walls, fences, and building facades that may span across multiple images in urban regions. This is also the motivation for our proposed approach for using dense pixel-level semantic segmentation to guide the image description. Towards this end, (Schönberger et al. 2018) propose a 3D descriptor that combines depth and semantics to achieve robustness against both appearance and viewpoint variations, whereas our proposed system does so by using only monocular imagery. Moreover, our benchmark datasets test the proposed system for extreme appearance and viewpoint variations *simultaneously* and not in isolation.

2.3 CNN-based Keypoint Correspondences

The use of keypoint correspondences based on deep-learnt CNNs has received significant attention. In the field of object/human parts recognition, a number of methods exist for keypoint localization/prediction (Long et al. 2014; Hariharan et al. 2015; Zhang et al. 2015). Some of the recent methods include spatial feature pyramid based correspondences (Ufer and Ommer 2017), local self-similarity learning (Kim et al. 2017b), and geometric-matching based training (Rocco et al. 2017; Han et al. 2017). However, most of these methods have been developed for object-centric images, unlike place-centric (Zhou et al. 2017) scenes where there is no particular object in focus. Moreover, visual place recognition additionally requires dealing with dynamic objects in the scene for which we demonstrate the use of semantic label verification and dense descriptor weighing as useful techniques for filtering correspondences. Although Rocco et al. (2017) demonstrate generalization of their approach by learning the geometric warping from street-view imagery, the experiments are conducted on datasets with limited variations in the camera viewpoint.

2.3.1 Keypoint Detection and Description: The use of CNNs for generating keypoint correspondences for the task of image retrieval and place recognition includes learning of keypoint detectors (Mishkin et al. 2018; Savinov et al. 2017) as well as robust local descriptors (Yi et al. 2016; Mishchuk et al. 2017; Zamir et al. 2016; Tolias et al. 2016). Brown et al. (2011) use a variety of pooling methods to learn discriminative local descriptors and Noh et al. (2017) learn both keypoints and descriptors using a novel attention mechanism but do not incorporate any form of semantic or global context for appearance-invariance. Taira et al. (2018) use CNN-based keypoint descriptors extracted densely from the feature maps of different layers of the network, but like the aforementioned methods, they require cross-matching of these descriptors to obtain correspondences. (Tolias et al. 2016) use regions within the CNN response maps to match and localize a query bounding box object within the database images. While their approach forgoes the cross-matching of regions/patches (Razavian et al. 2014) , a search procedure is still required to find a matching region for their query object. This is not the case with our approach as we assume that these correspondences are implicitly encoded by individual feature maps of the higher-order CNN layers; keypoint localization can simply be performed by an $argmax$ operation on corresponding feature maps. Radenović et al. (2018) employ a similar strategy to visualize keypoint correspondences after filtering them based on their distance measure, however, they do not use this correspondence information for any form of spatial verification. Furthermore, they only consider the extraction of correspondences for pairs of images, whereas our technique extracts keypoints and descriptors independently for a single image based on maximally-activated locations within the feature maps and the correspondences are later established at the matching stage.

2.3.2 Semantics for Keypoint Matching: Apart from the previously described bodies of research on robust semantics-based description and keypoint correspondences, there exist some methods that combine the strengths of both by matching the semantic information at keypoint level. Arandjelović and Zisserman (2014) propose SemanticSIFT descriptor using a semantic vocabulary that seeks to match both appearance and semantics simultaneously, however, in our approach, we first filter the semantically-inconsistent keypoints and then use their appearance descriptors to bias the geometric matching. Mousavian et al. improve the BoW-based retrieval by filtering keypoints in a similar manner as ours, however, they only consider man-made structures for this task (Mousavian et al. 2015; Mousavian and Košecka 2015). Kobyshev et al. (2014) compute binarized semantic histograms for each of the keypoints to reduce the search space for keypoint matching, however, the implicit correspondences used in our approach do not require such search procedures. Toft et al. (2018) propose a method to compute semantic consistency scores (SCS) in order to guide the sampling procedure of RANSAC, however, their approach requires 3D database models to do so. Moreover, in their comparisons with a baseline similar to our filtering strategy, the latter will be at a disadvantage because RANSAC like methods are affected by the inlier to outlier ratio and therefore false negatives (missed inliers) can significantly affect performance. In our case, this problem is handled to some extent by the use of appearance descriptors to bias the geometric matching.

2.4 Extent of Appearance-Invariance

Appearance-invariant visual place recognition techniques have mainly focused on either robust place representations (Sünderhauf et al. 2015; Naseer et al. 2017; Chen et al. 2017b) or image-sequence matching (Milford and Wyeth 2012; Vysotska and Stachniss 2016). However, the extent of variations in appearance has only been limited to matching multiple traverses of the environment with changing appearance *across* these traverses, for example, matching day vs night traverse or winter vs autumn traverse. The challenging scenario of matching places when appearance variations occur *both within* and *across* the traverses has not been addressed in the literature. The major difficulty posed in this case is the affinity of images captured under similar environmental conditions (say night time) to match with themselves (another night-time image) than matching with their counterparts captured under different environmental conditions (day time). This situation generally occurs when there are transitions in the environment from, say, outdoor to indoor areas with changes in lighting conditions both within a traverse and across multiple traverses of the same environment. We demonstrate that the existing state-of-the-art place representations (Arandjelovic et al. 2016), even with sequence-based matching, cannot achieve high performance and that the use of proposed descriptor normalization and local score enhancement is useful for enhancing the performance.

2.5 Extent of Viewpoint-Invariance

The currently prevalent visual place recognition methods have developed solutions for both appearance- and viewpoint-invariance. However, the extent of viewpoint variations has only been limited to changes in lateral displacement, orientation, and scale relative to the reference 6-DoF camera pose in the real world. Variations as extreme as opposite-viewpoint matching under changing appearance of the environment is a challenging problem due to limited visual overlap between a given pair of images taken at the same place from opposite directions. The research areas exploring the use of disjoint field-of-view cameras include camera calibration (Heng et al. 2015), motion estimation (Kawasaki et al. 2015), and mapping (Tribou et al. 2015). However, for VPR, the existing methods capable of opposite-viewpoint matching use panoramic cameras (Murillo et al. 2013; Arroyo et al. 2014) or LIDARs (Wolcott and Eustice 2017); developing a human-like viewpoint-invariance capability (Ackermann 1996) using a limited field-of-view monocular camera is currently an unsolved challenge.

2.6 Datasets

The research efforts in the direction of appearance-invariant visual place recognition have led to a number of new benchmark datasets, but there are limitations in the extent of variety they offer. The Alderley Day-Night dataset (Milford and Wyeth 2012) and Aachen Day-Night dataset (Sattler et al. 2018) , though useful for evaluation under extreme appearance variations, lack variety in varying environmental conditions and only use a single limited field-of-view camera. On the other hand, the Oxford Robotcar Dataset (Maddern et al. 2017) is comprised of repeated traversals of urban regions of Oxford under a variety of environmental conditions such as different times of the day, season, and weather. Further, the multiple-camera setup provides front, sideways, and rear-view imagery and is therefore suitable for evaluating place recognition performance for simultaneous variations in viewpoint and appearance. Similarly, the University of Michigan North Campus Long-Term Vision (NCLT) dataset (Carlevaris-Bianco et al. 2016) provides imagery captured using an omnidirectional camera with varying scene appearance both indoors and outdoors but lacks night time traversals. However, in both the Oxford Robotcar and NCLT datasets, there is no lateral viewpoint shift as the traverses are often repeated in the same direction and hence in the same lane. Our proposed datasets comprise forward as well as reverse journeys along urban, rural, and forested roadways under varying environmental conditions due to different weather conditions or times of day.

3 Proposed Approach

3.1 Overview of the VPR pipeline

An overview of the proposed visual place recognition pipeline is presented in Figure 2. We use the conventional coarse-to-fine matching process, often employed in literature (Cummins and Newman 2011; Mur-Artal and Tardós 2017), for high-precision loop closure detection. The *coarse* place recognition is mainly comprised of robust whole-image description, descriptor normalization, and a cosine distance-based global place search. These place matching scores are locally-enhanced based on semantic environment segmentation with the top N matching candidates being selected. These top candidates are used to perform *fine* place matching based on keypoint correspondences extracted from the pre-trained CNN. These correspondences are finally used to perform a weighted keypoint matching to validate spatial-layout consistency and select the best match. The visual semantic information, as shown in Figure 2, is employed within different components of the system, namely image description, local score enhancement, and keypoint correspondence filtering.

All the modules of our proposed pipeline are described in details in subsequent sections, covering in order: Robust Whole-Image Description, Local Score Enhancement, Keypoint Extraction and Correspondence, and Spatially-Consistent Weighted Keypoint Matching.

3.2 Robust Whole-Image Descriptor

The use of deep-learnt whole-image descriptors for visual place recognition (Naseer et al. 2017) and image retrieval (Arandjelovic et al. 2016) has gained popularity because of their robustness towards both viewpoint and appearance variations, as compared to the representations (Torii et al. 2015) based on hand-crafted features (Lowe 2004). However, the choice of these deep-learnt descriptors has been based on the desired amount of invariance towards changes in appearance and viewpoint of the scene (Sünderhauf et al. 2015).

The whole-image descriptor LoST (Local Semantic Tensor) proposed in (Garg et al. 2018b) describes a place explicitly with respect to different semantic classes (roads, buildings, and vegetation). For this, state-of-the-art semantic segmentation network RefineNet (Lin et al. 2017), trained on the Cityscapes dataset (Cordts et al. 2016), is used to provide dense pixel-wise semantic labels. Further, the convolutional feature maps are extracted from the $conv5$ layer of the network. The feature maps extracted from a CNN form a tensor of size $W \times H \times D$, where W, H, and D are the width, height and depth (or count) of the feature maps. The ResNet101 (He et al. 2016) based architecture in RefineNet leads W_{c5} and H_{c5} to be $1/32^{th}$ of the input image with D_{c5} as 2048 feature maps for the $conv5$ layer. Similarly, the semantic label scores so obtained have W_S and H_S as $1/4^{th}$ of the input image* and D_S is 20 which corresponds to the number of semantic categories in the Cityscapes dataset.

In order to deal with the noise associated with dense semantic labels and inspired from the success of image description methods like VLAD (Vector of Locally Aggregated Descriptors) (Jégou et al. 2010), the Local Semantic Tensor L (LoST) is defined for an image using the convolutional feature maps and semantic label probability:

$$L_s = \sum_{i=1}^{N} m_{is} |\mu_s - x_i| \quad (1)$$

$$\mu_s = \frac{\sum\limits_{x_s \in X_s} X_s}{\mathbf{card}(X_s)} \quad (2)$$

$$X_s = \{x_i \mid l_i = s \quad \forall \quad i \in [1, N]\} \quad (3)$$

$$l_i = \arg\max_s m_{is} \quad (4)$$

where l_i is the semantic label of a D_{c5}-dimensional descriptor x_i at location i within the feature map as shown in Figure 3. μ_s is the mean descriptor computed for a semantic class s by considering only those pixel locations that have their semantic label same as s. While the mean is computed using the most likely labels for the pixels, the image descriptor L_s is computed using the probability m_{is} of a pixel location i to belong to a semantic class s. The m_{is} is computed by L1-normalization of the label scores' tensor across its depth D_S.

Each semantic descriptor L_s forms an aggregation of the residual description from that particular semantic class s and the noisy contributions from the remaining classes, weighted by their semantic label probability. The final image descriptor L is a concatenation of L2-normalized L_s using only three semantic classes, that is, road, building and vegetation. For the reference database, as available beforehand, we modify the Equation 2 to replace the mean computed from a single image by the mean computed over a sliding window of 15 frames centered at the given reference image.

*The semantic score tensor is resized to $conv5$'s resolution when used in conjunction with its feature maps.

One of the limitations of state-of-the-art semantic segmentation networks is that they are generally trained on image data captured under ideal environmental conditions. This particularly leads to inaccurate labeling of night time imagery. This limitation can be overcome by using a *hybrid* descriptor that exploits the benefits of end-to-end VPR training as well as the semantic saliency of LoST. Therefore, we concatenate the state-of-the-art appearance-based descriptor NetVLAD (Arandjelovic et al. 2016) with LoST and refer to it as L' having dimension D'.

3.3 Descriptor Normalization

For boosting the appearance-invariance of the image descriptors described above, we perform a descriptor normalization (Garg et al. 2018a) as below:

$$L''_j = \frac{(L'_j - \mu_{db})}{\sigma_{db}} \quad \forall \quad j \in [1, N_{db}] \tag{5}$$

where μ_{db} is the mean descriptor of dimension D' and similarly, σ_{db} is the D'-dimensional standard deviation of the image descriptors, both computed using the entire database of N_{db} image descriptors. The reference database, available beforehand, is normalized using all the images within the database. The query database is, however, processed in an online manner as the images become available during the traverse, which means μ_{db} and σ_{db} for the query database are updated with every new query image. The normalization parameters of the reference database are not used for normalizing the query descriptors because these parameters often correspond to different environmental conditions, for example, a day-time reference traverse compared to a night-time query traverse. However, the intra-descriptor distributions can be expected to be similar after independent normalization of both the databases.

3.3.1 Global Place Search:

The normalized descriptors are used to compute the cosine distance of the query image from all the reference images:

$$d_{jk} = 1 - \frac{L''_j \cdot L''_k}{\|L''_j\|_2 \|L''_k\|_2} \quad \forall \quad j \in [1, N_{db}] \tag{6}$$

where d_{jk} are the distance values for a query image k, L''_j and L''_k are the normalized descriptors of the reference and query images, j and k respectively, and N_{db} is the total number of images in the reference database. The distance values d_{jk} then undergo neighborhood normalization, as explained in the next section, in order to find the best match based on a locally-enhanced global minimum as opposed to directly obtaining a global minimum.

3.4 Locally-Enhanced Global Minima

Visual place recognition (Naseer et al. 2014; Neubert and Protzel 2016; Sünderhauf et al. 2015) as well as image retrieval (Babenko and Lempitsky 2015; Arandjelovic et al. 2016) methods, in general, search for a best matching candidate by directly obtaining a global minimum of the distance values. However, for visual place recognition methods, the use of local enhancement was emphasized in SeqSLAM (Milford and Wyeth 2012) that employed a sequence-based matching of patch-normalized images. Their proposed local minima-based match selection primarily helped in dealing with perceptual aliasing caused due to the affinity of an image to match with another image obtained under similar environmental conditions than to its true match from different environmental settings. For example, a night-time image may tend to closely match with another night-time image than with its day-time counterpart under a naive matching scheme.

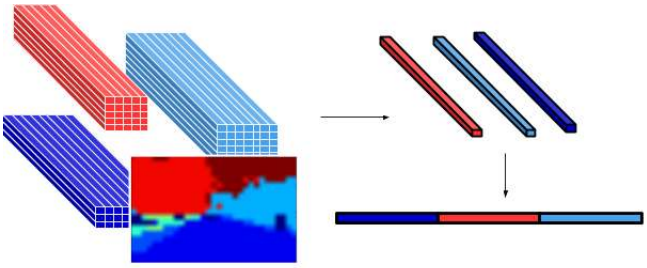


Figure 3. Local Semantic Tensor (LoST): The dense descriptors from $conv5$ layer are extracted with respect to the semantic labels of the image. The dense descriptors are then aggregated with respect to three semantic classes (road, vegetation, and buildings, represented by blue, red, and cyan respectively) and concatenated to form the LoST descriptor.

In this work, we first demonstrate that most of the existing image description techniques find it challenging to deal with this perceptual aliasing and that locally-enhanced global minima-based search can aid in appearance-invariant place recognition, especially when sequence-based methods are employed. However, the definition of *local* in SeqSLAM was based on an arbitrarily-defined neighborhood zone across a given reference image and required parameter-tuning for reaching peak performance. We propose an environment segmentation approach based on visual semantic information captured by higher-order layers of the CNN. The segments of the environment so obtained are therefore considered to be the ideal neighborhood for the reference images and the distance values within these neighborhood zones are then locally-enhanced by performing a normalization operation similar to that defined in SeqSLAM.

3.4.1 Semantic Environment Segmentation:

The environment segmentation is performed on the reference database using the $fc6$ layer from the Places-365 CNN[†] trained on Places data (Zhou et al. 2014). The different layers of any convolutional neural network capture semantic information as we move towards the higher-order layers, for example, from $conv3$ to $fc6$ and then to the final layer (Zhou et al. 2017; Zeiler and Fergus 2014). The use of $fc6$ layer provides a good compromise between capturing the visual semantics and discriminating between different places (Garg et al. 2018a) as compared to the $conv3$ or final layer. Therefore, the $fc6$ layer activation scores can be used to segment the reference database into meaningful semantic neighborhood zones. As the images in the reference traverse follow a temporal order, these activation scores for the entire database form a continuous sequence, where each activation score vector is represented as c_j and is of size 4096 (that is, the number of units in $fc6$ layer).

[†] https://github.com/metalbubble/places365

In order to temporally segment the environment, we use a Hidden Markov Model (HMM) to estimate the sequence of internal hidden states that generated the sequence of observed variables c_j. Given that we are interested only in estimating a transition within the environment which is binary by default, we set the number of hidden states to be 2. The implementation of HMM used in this work is available here[‡]. The sequence of hidden states so obtained is used to find the transition points and therefore different semantic neighborhood zones within the reference database.

$$R = \bigcup_{j_t}^{N_J - 1} \{(max(0, j_t - \delta), min(j_{t+1} + \delta, N_{db}))\} \quad (7)$$

where $J = \{0, j_{t_1}, \ldots, j_{t_{N_J - 1}}, N_{db}\}$ is the set of transition indexes within the reference traverse where the hidden state changes its value and N_j represents the cardinality of set J. These transition indexes with respect to the reference database are shown in Figure 4 with red hyphens across the vertical axis of the distance matrix. R is the set of index pairs to keep a record of the neighborhood zone in which a given reference image j lies. The parameter δ allows flexible borders for the semantic segments and is equal to the number of frames processed during the transition, that is, during the movement from one type of environment to the other, therefore adding *transition awareness* to the segmentation system. This is shown with green round braces in Figure 4 along with the images corresponding to the transition points and the transition window. As it is not trivial to classify the images near the transition point into one of the semantic segments, δ accounts for any such misclassification which would otherwise create false positive zones in a non-matching segment for a given query image. This parameter can be easily determined from the camera speed and frame processing rate.

3.4.2 Neighborhood Normalization: The semantic neighborhood zones segment the environment in such a way that images with similar environmental conditions are chunked together in a single temporal segment. For example, in one of the datasets used in our experiments (Residence Indoor Outdoor), the camera moves from the outdoor daylight to unlit indoor areas of the house as shown in Figure 4. Therefore, the first and second part of the traverse are automatically divided into two different segments as a result of semantic environment segmentation.

The distance values obtained from Equation 6 are modified with respect to the neighborhood zones obtained in Equation 7 according to the following normalization:

$$d'_{jk} = \frac{d_{jk} - \mu_{R_j}}{\sigma_{R_j}} \quad \forall \quad j \in [1, N_{db}] \quad (8)$$

where μ_{R_j} and σ_{R_j} are the mean and standard deviation of the distance values within the neighborhood zone defined by the corresponding region R_j in the set R. and d'_{jk} is the locally-enhanced distance for a query image k with respect to the reference image j. Figure 4 shows raw and locally-enhanced scores (bottom left graphs) for a query image (#25) along with their global minima. The raw scores generate a false global minimum (cross) whereas local enhancement leads to a correct match (circle). These locally-enhanced distance values are finally used to find the top-N matching candidates in the reference database.

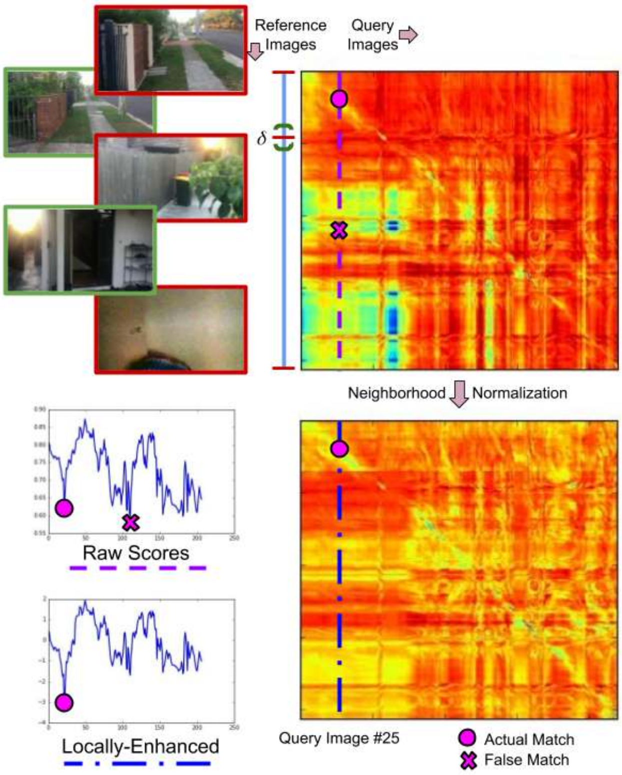


Figure 4. Local Enhancement of Scores: The semantic environment segmentation divides the reference imagery (vertical axis) into two regions as shown by the red hyphens (top). The green hyphens surrounding the transition point form a transition window (δ); the corresponding images are shown with red and green boundaries. The neighborhood normalization is performed according to the transition points and δ value. The effects of the local enhancement can be observed from the corresponding scores shown at the bottom-left for query image 25. The actual match (large circle) gets overshadowed due to perceptual aliasing for the raw scores whereas local enhancement leads to a correct global minimum.

3.5 Keypoint Extraction

The global place search procedure based on whole-image descriptors, though suitable for matching large databases by either using hashing (Vysotska and Stachniss 2017; Sünderhauf et al. 2015), aggregation (Lowry and Andreasson 2018) or random projection (Naseer et al. 2017) techniques, is not immune to false positives. Therefore, we use a second stage of matching the top-N candidates from the global place search to find the final best match. This is achieved by considering the dense convolutional descriptors derived from the feature maps of one of the convolutional layers of the network.

The dense convolutional descriptors have often been employed in the literature by flattening and concatenating all the feature maps of the convolutional layer (Chen et al. 2014; Sünderhauf et al. 2015; Vysotska and Stachniss

[‡] http://hmmlearn.readthedocs.io/en/latest/index.html

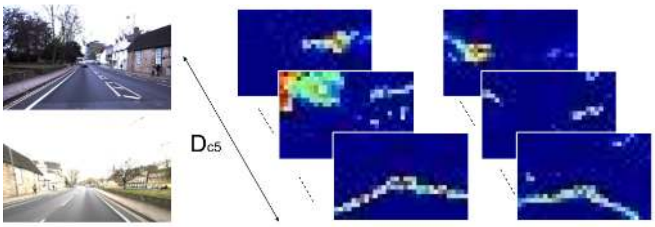


Figure 5. Correspondences inherent in Convolutional Feature Maps: For a matching pair of images (left) captured from opposite viewpoints, feature maps from the $conv5$ layer of the network are shown in the right. The activation patterns tend to be similar for a given feature map when a similar image is encountered, for example, the top and the bottom feature maps capture buildings and sidewalks information in a similar manner for the two given images.

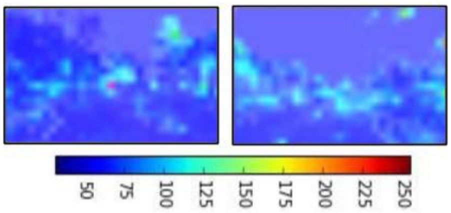


Figure 6. Keypoint Locations: The pair of images shows an activation frequency map overlaid on the two sample images shown in Figure 5. Each of the feature maps of $conv5$ provides a 2D location within the feature map where the maximum activation occurred. The activation frequency shows here the total number of instances when a 2D location attained maximum activation. It can be observed that road and sky categories do not have as many keypoints as buildings, vegetation, and sidewalk.

2016). However, these representations obscure the spatial information about the image which could otherwise be used for matching the scene-structure. We propose to extract salient keypoints from these convolutional feature maps based on their activation locations in the feature maps. Figure 5 shows some feature maps of $conv5$ layer for a pair of images. For each feature map in the $conv5$ layer of the network, we find the location within the feature map where maximum activation occurs. Therefore, for the $conv5$ layer of size $W_{c5} \times H_{c5} \times D_{c5}$, we obtain a total of D_{c5} keypoints. Given that the number of keypoints (D_{c5}) is quite high as compared to the resolution of the feature maps ($W_{c5} \times H_{c5}$), some of the keypoint locations frequently exhibit maximum-activation within different feature maps as shown in Figure 6.

3.6 Keypoint Correspondences

Convolutional neural networks tend to learn semantically meaningful concepts as we go deeper in the network architecture (say from $conv3$ to $conv5$) as established in (Zeiler and Fergus 2014; Zhou et al. 2017). As each convolutional filter within a layer learns a specific concept, it can generally be assumed that the activation patterns within the feature maps would be similar for a matching pair of images. Hence, the keypoints so obtained in the previous section for a given query image can be directly associated with those obtained for a reference image based on the index of the underlying feature map as shown in Figure 5. Therefore, we obtain D_{c5} keypoint correspondences directly from the D_{c5} keypoints. It may be noted that it is not necessary for these correspondences to occur at similar locations within the two feature maps. For example, a traffic-light detected in the left side of one image can correspond to a traffic-light detected in the right of another image. These correspondences can be observed in Figure 5 where the building- and vegetation-based activations occur within different regions of the feature maps for a pair of images having an opposite viewpoint for the same place.

However, these keypoint correspondences cannot be directly used in a spatial-consistency check because they contain a number of false correspondences. The false correspondences arise due to the following reasons: 1) dynamic objects causing activations at different locations within the image, 2) multiple instances of a similar semantic concept, say a pole, leading to cross-correspondence, and 3) misfiring within a feature map that is trained to detect a semantic concept not currently available in the scene, for example, a traffic-light sensitive feature map tested on an indoor image. Therefore, it is essential to filter these correspondences.

3.6.1 Semantic Label Verification: In order to filter the correspondences, we perform semantic label verification for all the correspondences. The semantic labels tensor is resized to $conv5$'s resolution to generate a semantic label for each of the keypoint locations. For a given pair of images, only those correspondences are retained that have the same semantic label. This step filters out a large number of undesired correspondences, especially those related to random activations caused by convolutional filters that detect concepts/features not available in a particular scene.

3.7 Spatially-Consistent Matching

The filtered correspondences are finally matched using their keypoint locations, weighted by the dense descriptor matching value for each of these correspondences. The D_{c5}-dimensional $conv5$ descriptor x_i is extracted for a given keypoint i. For all the D_{c5} keypoint correspondences, we then calculate the cosine similarity as below:

$$a_i = \frac{x_i \cdot x'_i}{\|x_i\|_2 \|x'_i\|_2} \quad \forall \quad i \in [1, D_{c5}] \tag{9}$$

where x_i and x'_i are the D_{c5}-dimensional descriptors for the i^{th} correspondence and a_i is the cosine similarity between these descriptors. These descriptor similarity values are used to perform a weighted keypoint matching as below:

$$r_c = \sqrt{\sum_{i}^{D_{c5}} \frac{a_i (p_i - p'_i)^2}{\|a\|_2}} \quad \forall \quad c \in [1, N] \tag{10}$$

where p_i and p'_i are the x-coordinates (flipped from left to right for rear-view images) of the keypoint locations and r_c is the matching score for a candidate c from the top N candidates retrieved by global place search. The candidate with the lowest score is considered the final best match.

4 Experimental Setup

4.1 Datasets

We used 4 different datasets to evaluate performance of the proposed VPR pipeline: Oxford Robotcar, Multi-Lane Forward-Reverse, Parking Lot, and Residence Indoor Outdoor. The first two datasets are used for opposite-viewpoint VPR experiments; the last two are used to investigate the efficacy of state-of-the-art image description techniques when appearance variations occur both within *and* across the traverses. All datasets are provided here[§].

The Oxford Robotcar dataset is comprised of a number of traverses of an urban environment under different environmental conditions due to varying time of day, weather and seasons. Further, all the traverses have been captured with front- as well rear-view cameras. Therefore, it is a prominent choice for testing the appearance-invariance of a VPR system, especially with opposite viewpoints. The Multi-Lane Forward-Reverse dataset was collected for testing for the first time even harder scenario of opposite-viewpoint matching under changing appearance, that additionally exhibits lateral viewpoint change due to different lanes of travel during the forward and reverse traverses of a route. All the datasets are described in details in subsequent sections.

4.1.1 Oxford Robotcar: The Oxford Robotcar dataset used here is a subset of the publicly available 1000 km data (Maddern et al. 2017). We use front- and rear-view imagery from four traverses exhibiting different environmental conditions, referred to as Overcast Autumn, Overcast Summer, Night Autumn, and Overcast Winter[¶]. These traverses correspond to the initial 2 km of the full traverse; the aerial view of the trajectory is shown in Figure 7. GPS data is used to sample the images at a constant distance of approximately 2 meters that leads to around 600-900 image frames in all the traverses.

4.1.2 Multi-Lane Forward-Reverse: The Multi-Lane Forward-Reverse (MLFD) dataset is comprised of three traverses captured at different times of day, that is, day, dusk, and night. The day and night traverses were captured during a forward journey and the dusk traverse was captured while moving in reverse direction. The aerial view of the trajectory is shown in Figure 7. We extract two segments from the trajectory (shown with green and red terminal markers) referred to as MLFR-DN and MLFR-DD for Dusk-Night and Dusk-Day respectively. The MLFR-DN and the MLFR-DD are approximately 1 km and 1.5 km in length respectively. Unlike the Oxford Robotcar dataset where front-rear matching does not exhibit any lateral viewpoint variation, the MLFR dataset is more challenging due to lateral shift in opposite viewpoints due to different lanes of travel during the forward and reverse journeys.

4.1.3 Parking Lot: The Parking Lot dataset is a pair of 2.5 km traverses of a parking area spread across outdoor and indoor regions of residential apartments. The two traverses, captured during different time of day (morning and night), also exhibit different lighting conditions *within* the traverses. While the daytime traverse transits between well-lit outdoor areas and partially-lit indoor regions, the night-time traverse transits between unlit outdoor areas and artificially-lit indoor regions. The videos were captured using a hand-held mobile device while driving a motor-bike and only every 20th video frame was used in the dataset. The aerial view of the trajectory is shown in Figure 8 with indoor and outdoor areas marked in different colors along with sample images.

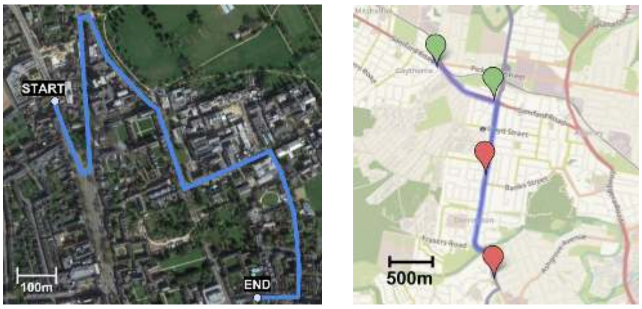


Figure 7. Aerial view of ground truth trajectories for Oxford Robotcar (left) and Multi-Lane Forward-Reverse (right) datasets. We use two pairs of trajectories from the latter as shown with different terminal markers (green and red). Source: Google Maps and Open Street Maps

4.1.4 Residence Indoor Outdoor: The Residence Indoor-Outdoor dataset is a 0.5 km dataset, comprising two traverses of a residential area, captured on foot once during daytime and then at night using a hand-held camera. Similar to the Parking Lot dataset, this dataset also exhibits appearance variations *both* within and across the traverses as camera transits between variably-lit outdoor and indoor regions of the house. The videos were captured using a hand-held camera and only every 10th video frame was used in the dataset. The aerial view of the trajectory is shown in Figure 8 with indoor and outdoor areas marked in different colors along with sample images.

4.2 Ground Truth

The visual place recognition ground truth for Oxford Robotcar dataset was generated using the GPS data. However, the front- and rear-view images from the same GPS location do not form a correct ground truth match as the visual overlap between them becomes maximum only when they are some physical distance apart from each other. We refer to this distance as *visual offset* and present an evaluation of it for one of the traverses of the Oxford Robotcar dataset in Section 6. The ground truth for the other three datasets was manually generated for regular keyframes and then interpolated for the entire traverse. The results shown for the opposite viewpoint matching in subsequent sections do not account for *visual offset* in their localization radius which means that localization accuracy of $40 - 50$ meters for front-rear matching is equivalent to 10 meters of localization if it were to be the front-front VPR[‖].

[§] http://michaelmilford.com/car-datasets/

[¶] Originally 2014-12-09-13-21-02, 2014-12-10-18-10-50, 2015-05-19-14-06-38, 2015-02-03-08-45-10 respectively (Maddern et al. 2017)

[‖] For more details, refer to Section 6 and Figure 18.

4.3 Evaluation

As per standard reporting in this research field, we use maximum F1 scores and Precision Recall curves for evaluating the performance of the system. A threshold on the distance scores calculated in Equation 10 is varied to generate the Precision-Recall values and max-F1 scores. For P-R curves, a match is considered to be correct if it lies within a range of 20 frames of its ground truth match. This value corresponds to 40, 50, 30, and 10 meters for the Oxford Robotcar, MLFR, Parking Lot, and Residence Indoor Outdoor datasets.

4.4 Comparative Study

We compare our approach with different state-of-the-art visual place recognition methods including NetVLAD (Arandjelovic et al. 2016) (their best performing VGG-16 + NetVLAD + whitening, Pittsburgh[**]), which has been demonstrated to work better than off-the-shelf convolutional descriptors (Sünderhauf et al. 2015), max-pooling (Tolias et al. 2016), and DenseVLAD (Torii et al. 2015). The results for the proposed system are presented in three modes: 1) LoST+NetVLAD+KC (where KC stands for Keypoint Correspondences based *fine* place search) represents complete pipeline, 2) LoST+NetVLAD represents the *hybrid* descriptor based global *coarse* place search only, and 3) LoST represents the semantic descriptor based global *coarse* search only. As the LoST descriptor is based on ResNet101 (He et al. 2016) architecture, we include the max-pooled descriptor, similar to the MAC descriptor (Tolias et al. 2016) from the $conv5$ layer, referred to as MaxPool, in order to observe any performance gains solely attributed to the ResNet architecture. We also include and propose *semantic max-pooling* (referred to as SemanticPool) using RefineNet in order to differentiate between a naive use of visual semantics for image description and our proposed aggregated descriptor LoST (Equation 1). The SemanticPool descriptor is constructed in a similar way as the MaxPool but with the max-pooling operation done separately for each semantic class based on the dense semantic labels of neurons in the feature maps. Finally, we include DenseVLAD using the authors' provided source code, off-the-shelf $conv3$ layer descriptors from pre-trained Places365 CNN (Zhou et al. 2014) and Sum of Absolute Difference (SAD) of patch-normalized (PN) images (Milford and Wyeth 2012) in order to observe variations in the performance trends when viewpoint changes from 0° to 180°.

5 Results

We first present results for the proposed visual place recognition system for the opposite viewpoints-based Oxford Robotcar and Multi-Lane Forward-Reverse datasets. Then, we evaluate the Parking Lot and Residence Indoor Outdoor datasets for the efficacy of locally-enhanced global minima selection, along with the influence of descriptor normalization and sequence-based matching.

5.1 VPR with Single Frame Matching

5.1.1 Oxford Robotcar Figure 9 shows the results for front- and rear-view matching of different traverses under

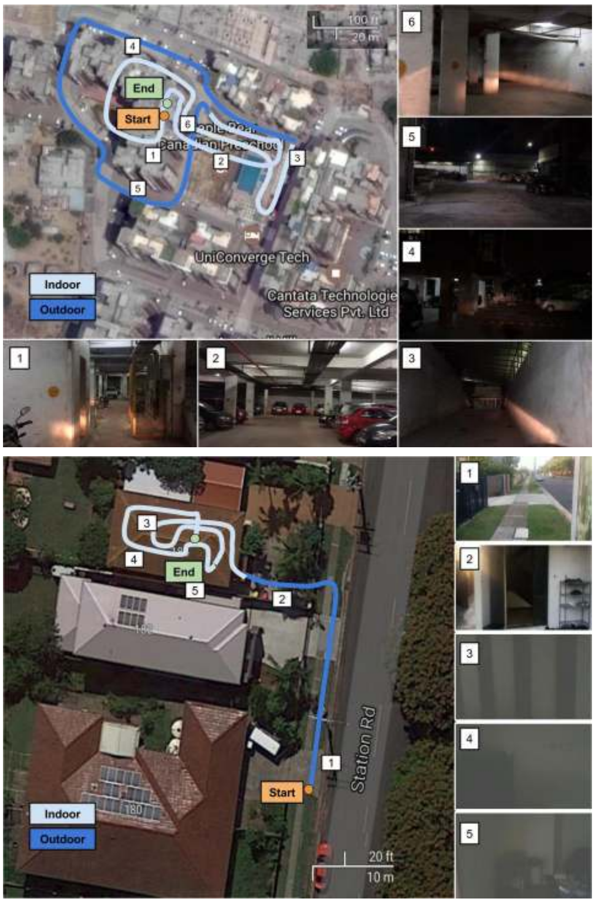
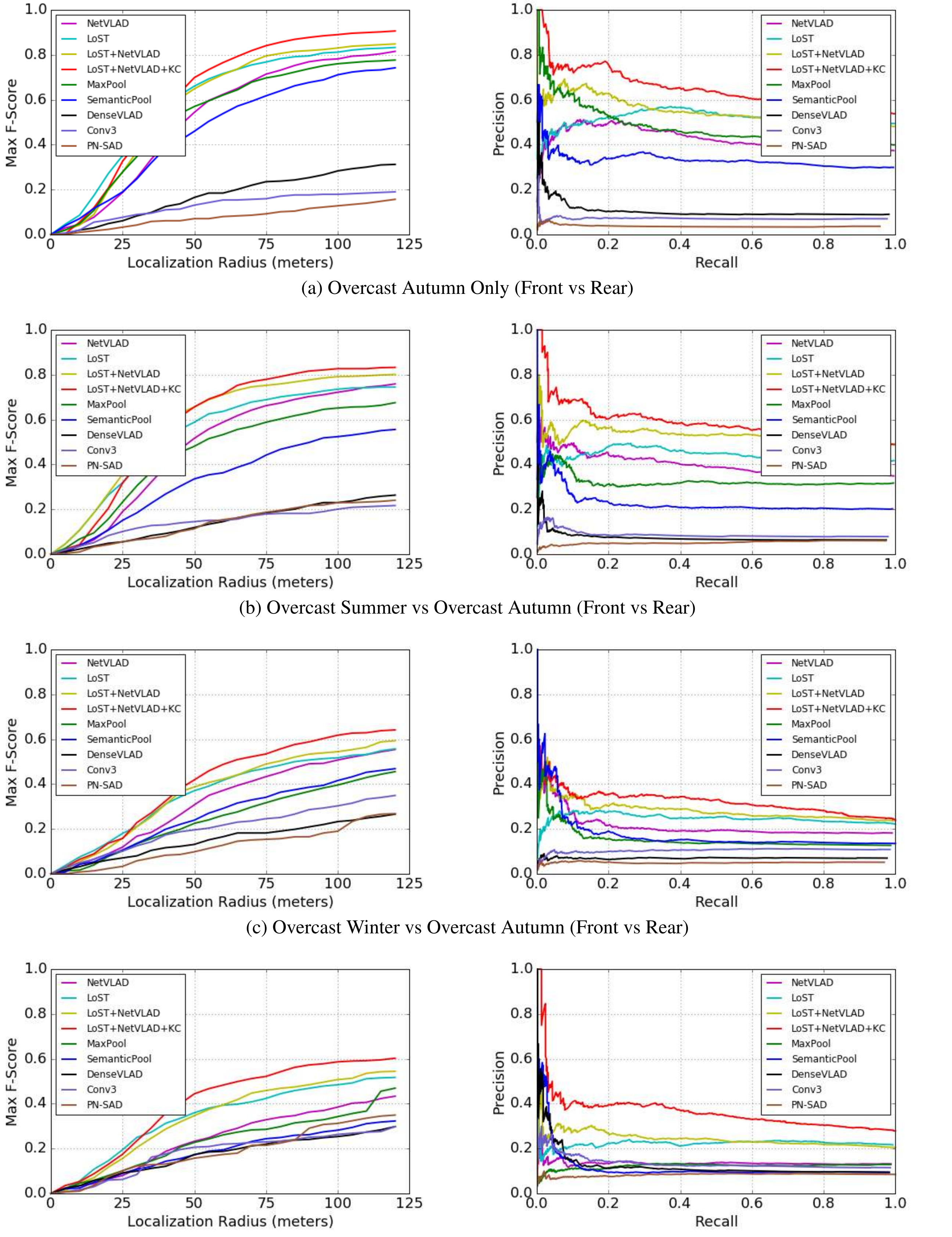


Figure 8. Aerial View of ground truth trajectories for Parking Lot (top) and Residence Indoor Outdoor (bottom) datasets. The indoor and outdoor areas in both the views are marked with light and dark blue color respectively. Further, the sample images from different locations from within one of the traverses of each dataset show the appearance variations occurring within the traverse. Source - Google Maps

varying environmental conditions. It can be observed that LoST-based descriptors consistently improve performance in terms of both Max-F1 scores and Precision-Recall curves. Further, the proposed system, including keypoint correspondences (KC), performs the best in most of the scenarios with one of the exceptions being the cross-season comparison with the Winter traverse where better performance is achieved only in the high-recall regime (see Figure 9 (c) right). It is worth noting that the use of keypoint correspondences helps in attaining a useful recall rate, though small, at 100% precision, where global place search, based on whole-image descriptors only, does not even reach 100% precision. Furthermore, in terms of overall performance across multiple traverses, the performance drops as the environmental conditions become more challenging from similar conditions, seasonal changes to change in time of day as shown in Figure 9 (a), (b)-(c), and (d) respectively.

It can also be noted that three of the methods: DenseVLAD, $conv3$, and patch-normalized image descriptor

[**] http://www.di.ens.fr/willow/research/netvlad/



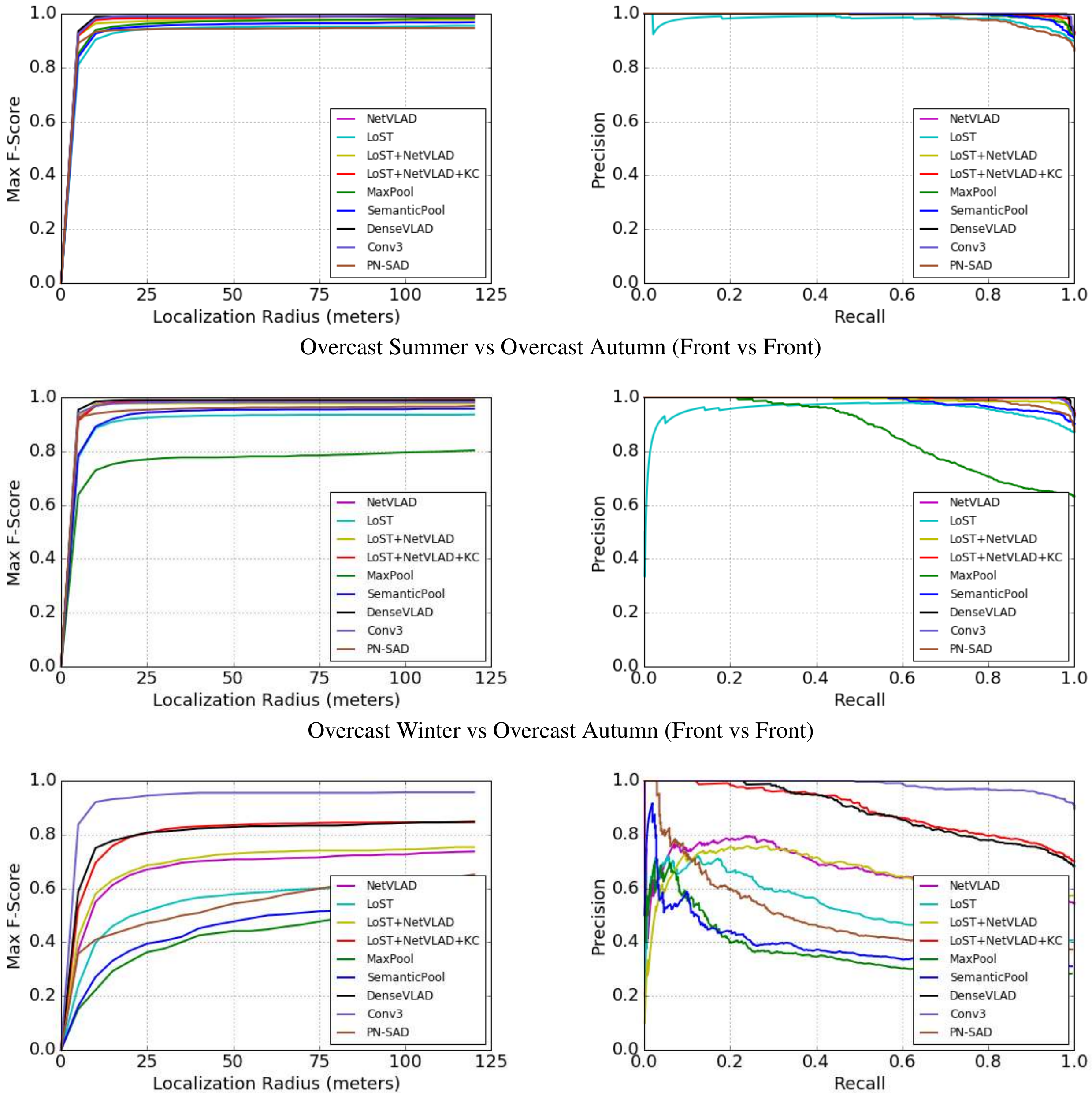
(a) Overcast Autumn Only (Front vs Rear)

(b) Overcast Summer vs Overcast Autumn (Front vs Rear)

(c) Overcast Winter vs Overcast Autumn (Front vs Rear)

(d) Night Autumn vs Overcast Autumn (Front vs Rear)

Figure 9. Oxford Robotcar dataset results with *opposite viewpoints* under changing environmental conditions: The performance curves show high performance using the LoST-based proposed descriptors, especially with the use of Keypoint Correspondences (KC). It is worth noting that the fine place matching using keypoint correspondences helps in attaining a useful recall rate at 100% precision, where a simple global place search never even reaches 100% precision.



Overcast Summer vs Overcast Autumn (Front vs Front)

Overcast Winter vs Overcast Autumn (Front vs Front)

Night Autumn vs Overcast Autumn (Front vs Front)

Figure 10. Oxford Robotcar dataset with *similar viewpoints* under changing environmental conditions: The performance trends for the traditional similar-viewpoint VPR are similar to those observed for the opposite-viewpoint scenario in Figure 9. The proposed use of Keypoint Correspondences (KC) improves performance, especially leading to higher recall at 100% precision along with higher Max-F1 scores for Day-Night matching (bottom row).

(PN-SAD), though known to perform quite well for similar-viewpoint place recognition, almost entirely fail at the task of matching images from opposing viewpoints. The off-the-shelf $conv3$ and the hand-crafted PN-SAD are by design viewpoint-dependent as they encode and match appearance depending on the pixel location which cannot be assumed for wide-baseline image matching including the opposite viewpoints. However, the VLAD architecture encodes appearance information independent of the pixel locations and has been successfully employed in methods like NetVLAD and LoST. Therefore, the performance drop for DenseVLAD in this case can be attributed to the simultaneous effect of at least these two factors: first, the underlying keypoint descriptor SIFT which is not flip-invariant (Ma et al. 2010; Zhao and Ngo 2013) and second, the density of keypoints considered in the image which can be a disadvantage when the visual overlap between two images is very low, for example, when viewed from opposing viewpoints. The performance trend for DenseVLAD is also consistent with the recent work of Schönberger et al. (2018) for $180°$ viewpoint-shift experiments on KITTI dataset (Geiger et al. 2012).

Figure 10 shows performance curves for visual place recognition based on similar-viewpoints under changing environmental conditions. Although, the overall performance is higher for all the comparisons as compared to

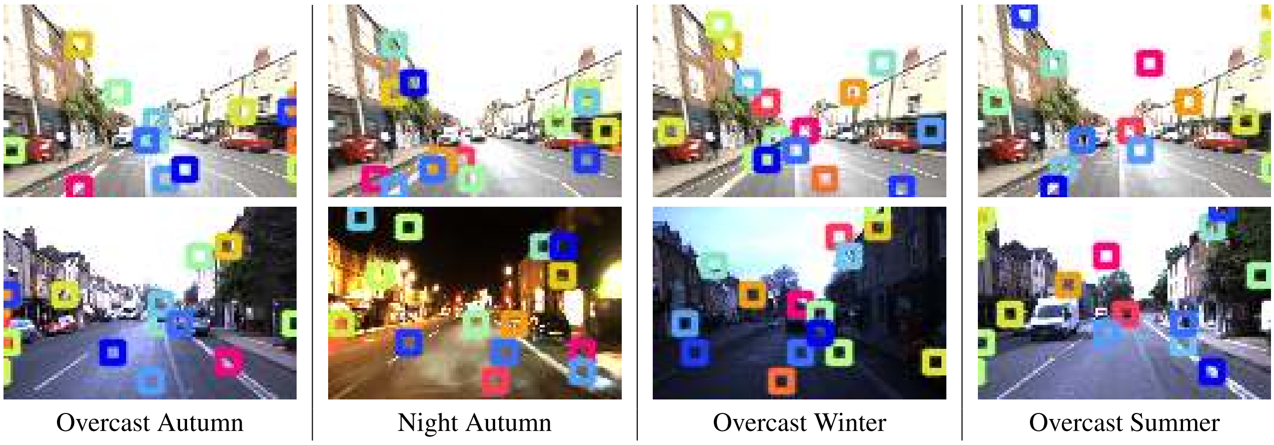


Figure 11. Matched place examples from the Oxford Robotcar Dataset with *opposite viewpoints* under different environmental conditions. The square markers show locations of keypoints correspondences, represented by the same color for a matching image pair in a given column. The top row shows a sample query image from the rear-view camera of the Overcast Autumn with a different set of keypoints matched each time. The bottom row shows front-view images matched from different reference traverses of the dataset. It can be observed that some of the salient keypoint locations repeatedly help in matching places.

the opposite-viewpoint scenario, the performance trends are quite different for different methods. Due to very similar viewpoints between the repeated traverses, viewpoint-dependent methods perform quite well here with $conv3$ setting the state-of-the-art for the day-night comparison followed by DenseVLAD and our proposed approach. This also shows that our proposed approach performs the best among the viewpoint-invariant description techniques that consistently outperformed the viewpoint-dependent methods for the opposing-viewpoint scenario. While NetVLAD and LoST+NetVLAD perform at par with each other, it can be observed that Keypoint Correspondences (KC) significantly improve performance in all the scenarios, especially for the day vs night comparison where a maximum recall of up to 20% is achieved at 100% precision whereas the other viewpoint-invariant methods never even reach 100% precision. The corresponding curves also emphasize the backward-compatibility of the proposed system.

Figure 11 shows some example image matches for a given query image from the rear-view camera of the Overcast Autumn traverse matched to front-view images from different reference traverses. The top 15 correspondences, based on their weighted matching value (extracted from Equation 10), are shown with the same color square markers for each column. It can be observed that some of the salient locations within the image repeatedly help in matching the places, despite changes in appearance or viewpoint.

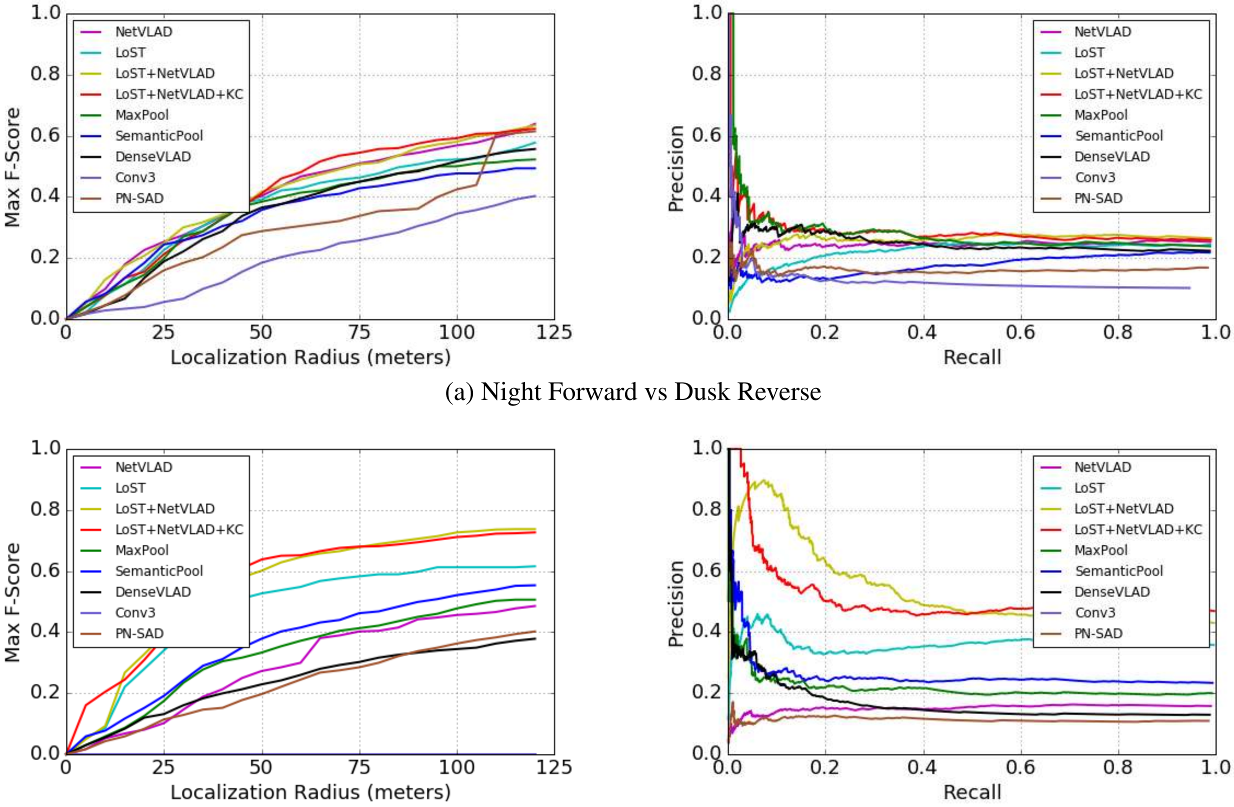
5.1.2 Multi-Lane Forward-Reverse Figure 12 shows performance comparisons for the Multi-Lane Forward-Reverse dataset. This dataset exhibits a more challenging scenario as compared to the Oxford Robotcar dataset in terms of opposite viewpoints, as the vehicle travels in different lanes in the forward and reverse traverses, adding lateral viewpoint change. The performance trends show that all of the methods find it challenging to match images between the night and the dusk traverses. It can also be observed that for the Day-Dusk comparison, keypoint correspondences lead to a higher recall at 100% precision but do not offer a consistent improvement as was observed in the Oxford Robotcar dataset. The overall performance in terms of max-F1 scores is lower as compared to the Oxford Robotcar dataset due to lateral viewpoint change attributed to different lanes of travel, limiting the visual overlap on top of the challenges of opposite viewpoints and changing environmental conditions. Figure 13 shows some example matches using the proposed approach for Day-Dusk and Night-Dusk comparisons.

5.2 VPR with Sequence-based Matching

The single-frame-based matching for opposite viewpoints under extreme appearance variations is a challenging problem, especially when viewpoint also varies laterally due to different lanes of travel during the forward and reverse journeys. The use of sequence-based methods has been proven to significantly improve single-frame based VPR (Cummins and Newman 2011; Hansen and Browning 2014; Naseer et al. 2018) and is feasible in some application domains. Hence, we also evaluated the proposed approach in conjunction with image sequence matching; however the use of sequential matching is not a claimed contribution of this work but rather a standard practice enhancement. We used OpenSeqSLAM (Milford and Wyeth 2012) for our experiments. However, in general, any other sequence-based method can be used (Pepperell et al. 2014; Vysotska and Stachniss 2016).

Figure 14 shows the Precision-Recall curves calculated with a sequence length of 20 frames and a localization radius of 30 frames[††] for the opposite-viewpoints scenario using the proposed approach and NetVLAD (referred to as Ours+Seq and NetVLAD+Seq respectively). In Figure 14 (a)-(d), it can be observed that for the opposite-viewpoints from the same lane, as in Oxford Robotcar dataset, the proposed approach is able to recall a significantly higher number of matches within the high-precision zone as compared to the state-of-the-art for all the comparisons under different environmental conditions. For relatively easier

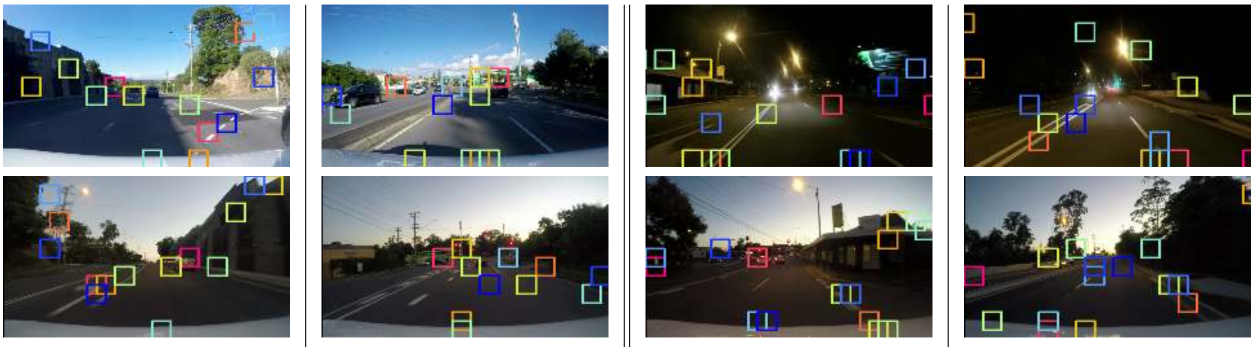
[††] These values correspond to 40 and 50 meters of sequence length and 60 and 75 meters of localization radius (without accounting for *visual offset*) for the Oxford Robotcar and Multi-Lane Forward-Reverse datasets respectively.



(a) Night Forward vs Dusk Reverse

(b) Day Forward vs Dusk Reverse

Figure 12. The most challenging scenario: Multi-Lane Forward-Reverse dataset with *opposite viewpoints* along with *different* lanes of travel: The performance improvement is consistent using LoST-based descriptors, though the overall performance is lower in both (a) and (b) as the dataset offers the most challenging scenarios due to different lanes of travel. The keypoint correspondences (KC) improve the max-F1 score more for (a) than (b), however, the P-R curves show consistent gains in performance.



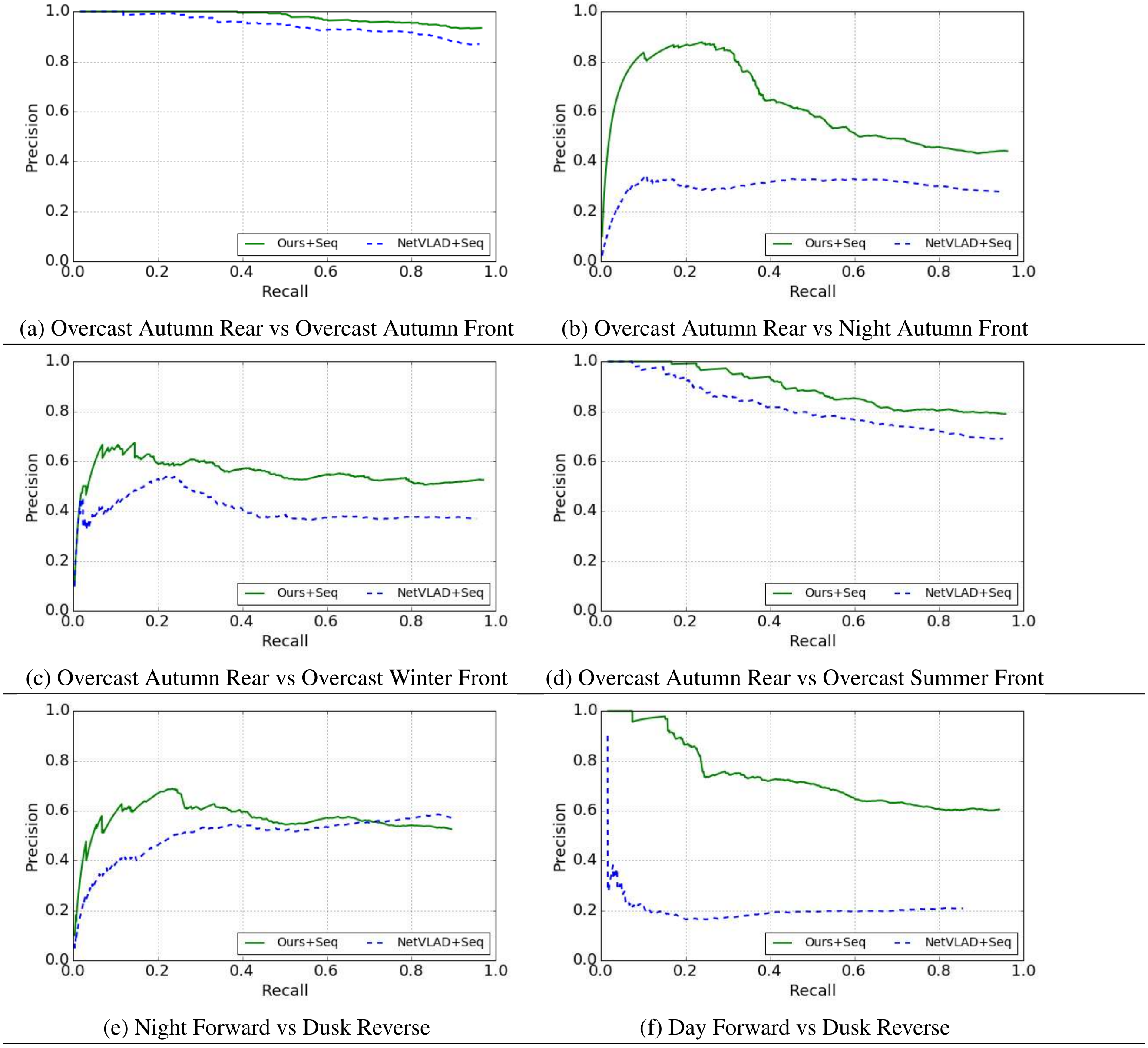
Day Forward vs Dusk Reverse

Night Forward vs Dusk Reverse

Figure 13. Matched place examples from the Multi-Lane Forward-Reverse dataset for Day-Dusk and Night-Dusk comparisons. Each column shows a matched pair of images. The lateral viewpoint change due to use of different lanes while traveling in opposite directions further reduces the visual overlap caused due to opposite viewpoints.

cases of limited appearance variations as shown in Figure 14 (a) and (d), the proposed approach achieves a maximum recall of 40% and 20% respectively at 100% precision as compared to an approximate 10% of maximum recall using NetVLAD. For a more challenging scenario of day-night place recognition as shown in Figure 14 (b), a maximum recall of 30% at 85% precision is achieved where the state-of-the-art method struggles to achieve even half of that precision at any recall level.

In Figure 14 (e)-(f), results are presented for the first time on the even more difficult scenario using the most challenging datasets to date that exhibit lateral viewpoint change along with opposite-viewpoints and changing environmental conditions. The proposed approach outperforms the state-of-the-art with significant marginsfor day and dusk matching (Figure 14 (f)), where a maximum recall of 8% is achieved at 100% precision while the state-of-the-art method never reaches the 100% mark for any recall value. For the more challenging Night-Dusk comparison, both the methods fail to achieve 100% precision. It can further be observed that the proposed approach ceases to perform better in the high-recall regime.



(a) Overcast Autumn Rear vs Overcast Autumn Front

(b) Overcast Autumn Rear vs Night Autumn Front

(c) Overcast Autumn Rear vs Overcast Winter Front

(d) Overcast Autumn Rear vs Overcast Summer Front

(e) Night Forward vs Dusk Reverse

(f) Day Forward vs Dusk Reverse

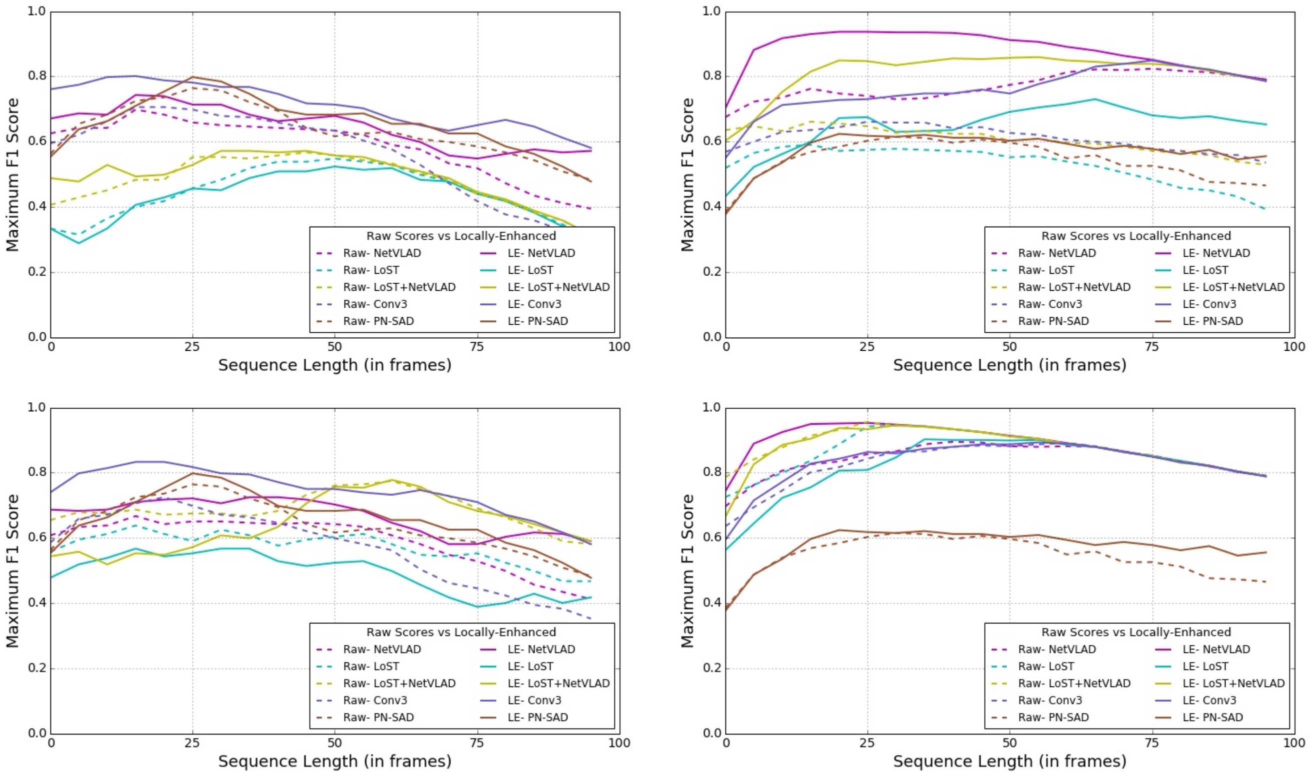
Figure 14. *Sequence-based* matching for *opposite-viewpoints* under varying appearance of the environment. It can be observed that for the opposite-viewpoints from the same lane (a)-(d), the recall rate of the proposed approach within high-precision zones is significantly better than that achieved using the state-of-the-art method, especially for the day-night scenario (b). Further in (e)-(f), for the most challenging scenario of lateral viewpoint shift on top of opposite viewpoints and changing conditions, performance improvements are significant, especially for day-dusk (f) where the proposed approach attains a maximum recall of 8% at 100% precision whereas state-of-the-art never reaches the 100% mark.

Hence, across all implementations, whether single-frame-based or sequence-based, our proposed methods significantly improve performance.

5.3 Effects of Local Score Enhancement

The Parking Lot and Residence Indoor Outdoor datasets were used for evaluating the efficacy of locally-enhanced scores (Equation 7) based global minima selection against a direct selection of global minima. The use of locally-enhanced scores was emphasized in SeqSLAM (Milford and Wyeth 2012) for sequence-based matching of patch-normalized images belonging to different environmental conditions (for example, day vs night). In order to characterize the inter-influence of whole-image descriptor normalization (Equation 5), neighborhood normalization based local enhancement of scores (Equation 7), and the use of sequential-information for VPR, we evaluated the performance of different image description methods on the aforementioned datasets. These datasets exhibit variations in lighting conditions, natural (day vs night) and artificial (indoor), both within *and* across the traverses, as also shown in Figure 8. Figure 4 (top-right) shows a distance matrix computed using off-the-shelf $conv3$ descriptors and demonstrates the difficulty to differentiate between a correct (circle) and a false (cross) match without the proposed local score enhancement.

We used 5 different image description methods for performance comparisons: NetVLAD, LoST, LoST+NetVLAD, $conv3$ (pre-trained on Places365), PN-SAD (as used in SeqSLAM). The first three description methods do not



(a) Residence Indoor Outdoor

(b) Parking Lot

Figure 15. Max-F1 scores with respect to varying sequence length showing effects of *Local enhancement (LE)* for raw (top) and normalized (bottom) descriptors: the performance improves consistently with the use of locally-enhanced scores for most of the descriptors for both the datasets, irrespective of the use of descriptor normalization.

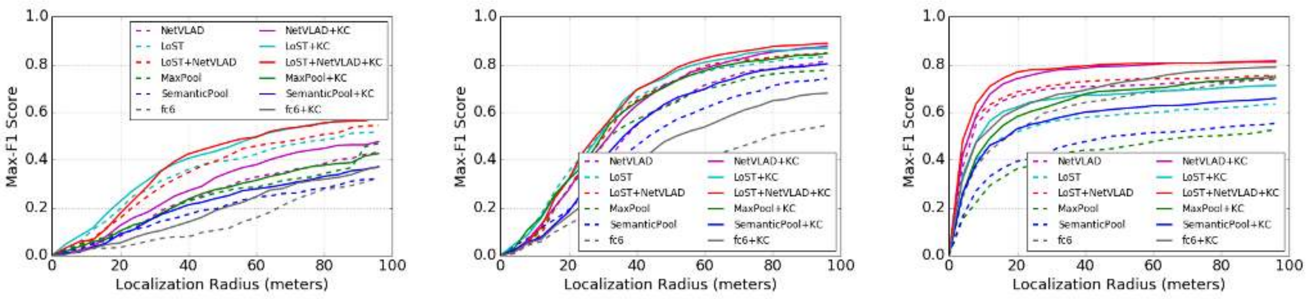
explicitly retain the spatial layout information of the image. $conv3$ layer descriptor is $65K$ in size and is understood to perform very well for appearance variations, especially when there is no significant change in viewpoint (Sünderhauf et al. 2015), which is also the case with both these datasets. The Sum of Absolute Difference of Patch-normalized images (PN-SAD), which are hand-crafted image representations, has been demonstrated to be useful for appearance-invariant place recognition using SeqSLAM (Milford and Wyeth 2012). We excluded comparisons with MaxPool and SemanticPool here for legibility as the other description techniques have already been shown to outperform these in previous sections. For all the different image description techniques used here, OpenSeqSLAM (Milford and Wyeth 2012) was used to match image sequences.

Figure 15 shows max-F1 scores for both the datasets for a varying sequence-length, with and without the descriptor normalization and local score enhancement. It can be observed that for most of the descriptors local enhancement helps improve the performance, irrespective of the descriptor normalization. However, the performance gains are more significant when descriptors are not normalized as descriptor normalization tends to improve the baseline performance, therefore, leaving less scope for local enhancement for any further improvement. Furthermore, the LoST-based descriptors, though attaining higher performance for locally-enhanced raw descriptors (Figure 15 top row), perform at par with the usage of raw scores when normalized descriptors are used (Figure 15 bottom row). The use of road-based semantics for describing LoST, applied to indoor images of these datasets, leads to performance loss for LoST-based descriptors and the aforementioned different behavior. In an ideal scenario, a different semantic segmentation network for indoor and outdoor imagery would be more useful and is an area of future work. However, the purpose of the analysis in Figure 15 is to evaluate the use of local enhancement for state-of-the-art image description techniques, and show improvements in performance for most cases.

5.4 Fine Place Search

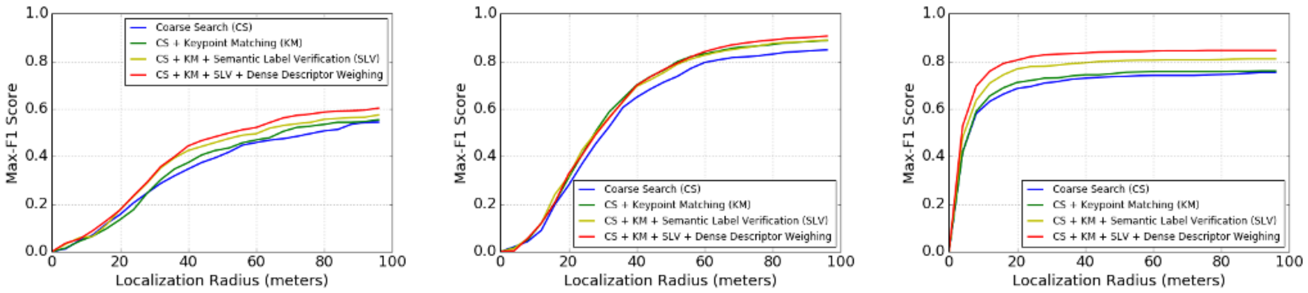
The proposed VPR pipeline is comprised of a coarse-to-fine place search procedure as outlined in Figure 2. Here, we demonstrate that *fine* place searching is in general applicable to different whole-image description techniques and different components of *fine* searching incrementally improve performance and are therefore crucial to the proposed approach. We used two traverses from the Oxford Robotcar dataset with varying appearance (day-night) and opposite viewpoints (front-rear) for this purpose.

5.4.1 General Applicability: Figure 16 shows max-F1 scores for different whole-image descriptors as used in previous sections along with Keypoint Correspondences (KC) based *fine* place search applied to each of them. It can be observed that for all three sets of comparisons for varying appearance and viewpoints, significant performance improvements can be achieved on different *coarse*-search baselines. Further, the gain in performance tends to be larger if the baseline is weaker.



Overcast Front vs Night Rear Overcast Front vs Overcast Rear Overcast Front vs Night Front

Figure 16. *General applicability* of the proposed *fine* place search approach with *coarse* place search performed using different whole-image descriptors. It shows that significant performance improvements can be achieved on different *coarse*-search baselines; further, the gain in performance tends to be larger if the baseline is weaker.



Overcast Front vs Night Rear Overcast Front vs Overcast Rear Overcast Front vs Night Front

Figure 17. *Performance contribution* of different components of the proposed *fine* place search procedure. The consistent gain in performance using the proposed components of the *fine*-search procedure emphasizes their individual utility.

5.4.2 Components' Contribution: Figure 17 shows the max-F1 scores corresponding to the contribution of different components of the *fine*-search procedure towards system performance. The effects of different components are observed using the following:

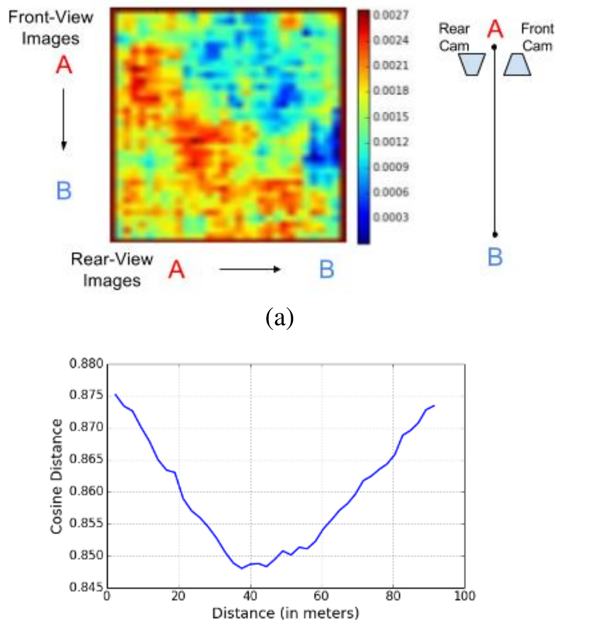
- Coarse Search (CS), that is, without using any component of *fine*-search,
- CS + Keypoint Matching (KM), that is, performing keypoint matching (Equation 10) for spatial layout verification without using any semantic label verification (SLV as in Section 3.6.1) or dense descriptor weighing (Equation 9),
- CS + KM + Semantic Label Verification (SLV), that is, the proposed *fine*-search but without dense descriptor weighing,
- CS + KM + SLV + Dense Descriptor Weighing, that is, the complete pipeline.

In Figure 17, performance curves for different sets of appearance and viewpoint variations show a consistent gain in performance using the proposed components of the *fine*-search procedure. This demonstrates that each of these components is crucial and adds value in terms of system performance.

6 Discussion

6.1 Visual Offset for Opposite Viewpoints

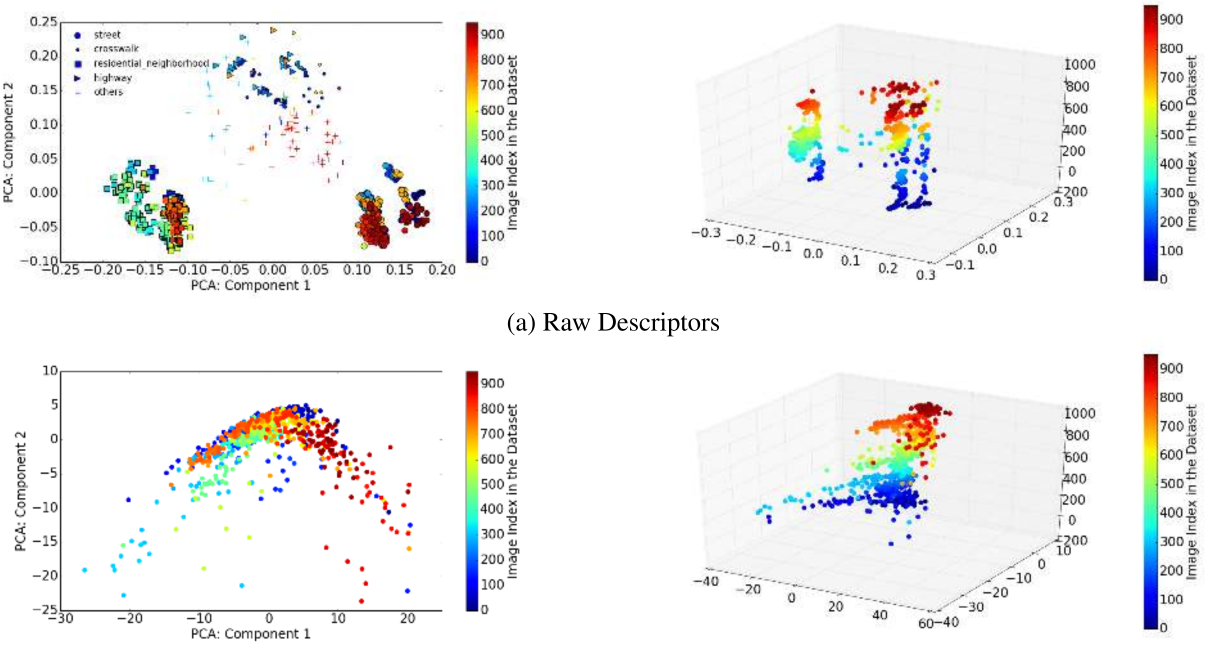
Techniques addressing the visual place recognition problem for front-view only image matching generally only need to deal with limited variations in viewpoint such that the



(a)

(b)

Figure 18. (a) Average distance matrix computed by matching images (front vs rear) captured between locations that are 60 meters apart. (b) Average cosine distance between different pairs of front- and rear-view images plotted against the average physical distance between them. This curve shows that on an average there is approximately 40 meters of *visual offset* between a matching pair of front- and rear-view images for this dataset.



(a) Raw Descriptors

(b) Normalized Descriptors

Figure 19. PCA Visualization: The 2-D PCA projection of final layer descriptors from Overcast Autumn Traverse of Oxford Dataset in Raw (top) and Normalized (bottom) form. The raw descriptors tend to cluster according to their semantic labels, irrespective of the image index in the video, whereas, the normalized descriptors tend to form spatio-temporal clusters, despite the absence of any explicit temporal signal during PCA training. The 3D plot on the right uses image index on vertical axis and shows that the places which are close in time are also close in space after normalization, which is not the case with the raw descriptors.

amount of visual overlap can be directly related to the changes in viewpoint (Lucas and Kanade 1981). However, this is not the case when two images of the same place are captured from opposite viewing directions. For such pairs of images, the physical distance between the cameras that is necessary to provide maximum visual overlap may not always be the same. For example, consider a building spanning 20 meters in length on the left side of the road, two cameras with opposing viewpoints (facing each other) will need to be placed at the end points of the building (that is 20 meters apart) in order to maximize the visual overlap. This distance, referred to as *visual offset*, depends on different factors, for example, the type, size, and position of the visual landmarks in the scene and the camera's field-of-view.

Figure 18 (a) illustrates the visual overlap (in terms of cosine distance between image descriptors) for image sequences captured by two cameras that traveled from position A to B. This setup dictates that images captured by the front-view camera near position B will not match with any of the images captured by the rear-view camera while moving from A to B. This can be observed through the bottom rows of the cosine distance matrix. A similar behavior can be observed for images captured by rear-view camera near position A through the left columns of the distance matrix. As the cameras move from A to B, rear-view images *begin* to match after a certain distance is covered and front-view images *cease* to match after that particular distance is traversed; this distance is equivalent to the *visual offset*. The distance matrix shown in Figure 18 (a) is an average calculation over different pairs of A and B that are 60 meters apart in the Overcast Autumn traverse.

Figure 18 (a) highlights the average image matching trend between front- and rear-view images; Figure 18 (b) shows the average physical distance that would provide maximum visual overlap and therefore minimum cosine distance between the matching pair of images. The latter curve was generated by matching a randomly picked image with its neighboring images within a range of 100 meters. Now for the front-view only image matching, the minimum cosine distance should ideally correspond to the minimum physical distance (assuming the camera orientation is almost fixed). However for the front-rear matching, as can be seen from the graph, the physical distance of approximately 40 meters on average is necessary for maximum visual overlap. This shows that 50 meters of localization accuracy (based on GPS) in our results is actually equivalent to 10 meters after accounting for the *visual offset*. The 40 meters estimate might be inflated for two reasons: 1) any image matching algorithm would not perfectly capture the variation in cosine distance as visual overlap varies between the images, especially when there are dynamic objects in the scene, 2) the average is performed on randomly picked images including the front- and rear-view images captured near the turns which cannot be always associated because of completely unobserved visual content by one of the cameras during the turn. By visually observing a handful of front-rear image samples along with their GPS coordinates, we found the *visual offset* to be approximately 30 meters.

6.2 PCA of Descriptor Normalization

The descriptor normalization performed in Equation 5 modifies the image descriptors such that the query image can better discriminate among the reference descriptors, therefore leading to improved performance. Figure 19 shows two principal components of the raw (top) and normalized (bottom) descriptors in 2D (in left) and 3D (in right with time as vertical axis). Here, we used the $final$ layer of a pre-trained visual place categorization network (Zhou et al. 2017) to represent the images from Overcast Autumn traverse. Different layers of a CNN learn semantically more meaningful concepts as we go higher in their order, that is, from $conv3$, $conv5$, $fc6$ to $final$ (Zhou et al. 2017; Zeiler and Fergus 2014). The descriptor normalization has similar effects on the descriptors derived from higher-order layers (like NetVLAD, LoST, MaxPool from $conv5$), however it is more intuitive to understand this effect through the $final$ layer.

Figure 19 (top-left) shows that the raw descriptors, derived from a network trained for place categorization, tend to form clusters according to their semantic categories. However, the normalized descriptors, as shown in Figure 19 (bottom-right), tend to cluster in such a way that places that are close in time are also close in space, which facilitates visual place recognition process.

7 Conclusion and Future Work

Recognizing places from opposite viewpoints under varying appearance, especially while traveling in different lanes during the forward and reverse traverses is an extremely challenging problem that had only previously been addressed with panoramic sensing arrangements. Our proposed approach employs appearance- and semantics-based robust place representations and semantically-filtered keypoint correspondences to achieve visual place recognition under these challenging conditions whilst only using a limited field of view forward facing camera. On publicly-available and new contributed datasets, we demonstrated that our system attains a maximum recall of 8% and 30% at 100% and 85% precision respectively under situations where the state-of-the-art method struggles to attain even 40% precision at any recall level. In particular the opposite-viewpoint, multi-lane, and varied appearance scenario has not been addressed in previous VPR literature; the maximum recall of 8% at 100% (Figure 14 (f)) precision under such challenging settings, though low, is significantly better than what is achievable using the current state of the art. Further, we showed that our contributions around descriptor normalization and local score enhancement boost VPR performance for the majority of the state-of-the-art image descriptors, which would otherwise struggle to deal with the perceptual aliasing caused by appearance variations *within* and *across* the traverses. Our analysis showed a number of interesting insights into performance under these conditions, including a characterization of the average $30-40$ meters of *visual offset* (in terms of physical distance) that exists between matching image pairs from opposing viewpoints for these types of datasets.

In future, we plan to extend the current work with metric relative pose estimation for recognized places using the keypoint correspondences already generated by our approach. Such a capability will enable integration with current state-of-the-art 6-DoF visual SLAM and Structure-from-Motion (SfM) systems, especially for metric appearance-invariant localization (Sattler et al. 2018). The broader use of visual semantic information for visual place recognition and visual SLAM is a promising avenue (Cadena et al. 2016); we also intend to develop a deep-learning framework that can leverage the semantic aggregation and appearance-based cues proposed by our current system to further improve viewpoint- and appearance-invariant localization performance. Finally, as human-robot interaction research progresses in the domain of autonomous vehicles, we will investigate how the higher similarity in place recognition methodology between our semantics-based system and humans (compared to highly engineered, often predominantly geometric, panoramic sensor-based approaches) could facilitate more effective human-robot or human-autonomous vehicle interaction. Overall, we hope that this body of work serves as a contribution that further progresses the field of semantic-based perception, navigation and localization.

References

Ackermann E (1996) Perspective-taking and object construction: Two keys to learning. In: *Constructionism in practice*. Lawrence Erlbaum Associates, pp. 25–37.

Arandjelovic R, Gronat P, Torii A, Pajdla T and Sivic J (2016) Netvlad: Cnn architecture for weakly supervised place recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 5297–5307.

Arandjelović R and Zisserman A (2014) Visual vocabulary with a semantic twist. In: *Asian Conference on Computer Vision*. Springer, pp. 178–195.

Ardeshir S, Zamir AR, Torroella A and Shah M (2014) GIS-assisted object detection and geospatial localization. In: *European Conference on Computer Vision*. Springer, pp. 602–617.

Arroyo R, Alcantarilla PF, Bergasa LM, Yebes JJ and Gámez S (2014) Bidirectional loop closure detection on panoramas for visual navigation. In: *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*. IEEE, pp. 1378–1383.

Atanasov N, Zhu M, Daniilidis K and Pappas GJ (2016) Localization from semantic observations via the matrix permanent. *The International Journal of Robotics Research* 35(1-3): 73–99.

Babenko A and Lempitsky V (2015) Aggregating local deep features for image retrieval. In: *Proceedings of the IEEE international conference on computer vision*. pp. 1269–1277.

Bay H, Ess A, Tuytelaars T and Van Gool L (2008) Speeded-up robust features (surf). *Computer vision and image understanding* 110(3): 346–359.

Brown M, Hua G and Winder S (2011) Discriminative learning of local image descriptors. *IEEE transactions on pattern analysis and machine intelligence* 33(1): 43–57.

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

Carlevaris-Bianco N, Ushani AK and Eustice RM (2016) University of Michigan North Campus long-term vision and lidar dataset. *The International Journal of Robotics Research* 35(9): 1023–1035.

Castaldo F, Zamir A, Angst R, Palmieri F and Savarese S (2015) Semantic cross-view matching. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. pp. 9–17.

Chen Z, Jacobson A, Sünderhauf N, Upcroft B, Liu L, Shen C, Reid I and Milford M (2017a) Deep learning features at scale for visual place recognition. In: *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, pp. 3223–3230.

Chen Z, Lam O, Jacobson A and Milford M (2014) Convolutional neural network-based place recognition. In: *Australasian Conference on Robotics and Automation*, volume 2. p. 4.

Chen Z, Maffra F, Sa I and Chli M (2017b) Only look once, mining distinctive landmarks from convnet for visual place recognition. In: *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE, pp. 9–16.

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B (2016) The cityscapes dataset for semantic urban scene understanding. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3213–3223.

Cummins M and Newman P (2011) Appearance-only slam at large scale with fab-map 2.0. *The International Journal of Robotics Research* 30(9): 1100–1123.

Dalal N and Triggs B (2005) Histograms of oriented gradients for human detection. In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1. IEEE, pp. 886–893.

Engel J, Schöps T and Cremers D (2014) Lsd-slam: Large-scale direct monocular slam. In: *European Conference on Computer Vision*. Springer, pp. 834–849.

Gálvez-López D and Tardos JD (2012) Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics* 28(5): 1188–1197.

Garg S, Babu V M, Dharmasiri T, Hausler S, Suenderhauf N, Kumar S, Drummond T and Milford M (2019) Look No Deeper: Recognizing Places from Opposing Viewpoints under Varying Scene Appearance using Single-View Depth Estimation. In: *IEEE International Conference on Robotics and Automation (ICRA)*.

Garg S, Jacobson A, Kumar S and Milford M (2017) Improving Condition- and Environment-Invariant Place Recognition with Semantic Place Categorization. In: *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE.

Garg S, Suenderhauf N and Milford M (2018a) Don't look back: Robustifying place categorization for viewpoint- and condition-invariant place recognition. In: *IEEE International Conference on Robotics and Automation (ICRA)*.

Garg S, Suenderhauf N and Milford M (2018b) Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics. In: *Proceedings of Robotics: Science and Systems XIV*.

Gawel A, Del Don C, Siegwart R, Nieto J and Cadena C (2018) X-view: Graph-based semantic multi-view localization. *IEEE Robotics and Automation Letters* 3(3): 1687–1694.

Geiger A, Lenz P and Urtasun R (2012) Are we ready for autonomous driving? The Kitti Vision Benchmark Suite. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, pp. 3354–3361.

Gong Y, Wang L, Guo R and Lazebnik S (2014) Multi-scale orderless pooling of deep convolutional activation features. In: *European conference on computer vision*. Springer, pp. 392–407.

Han K, Rezende RS, Ham B, Wong KYK, Cho M, Schmid C and Ponce J (2017) Scnet: Learning semantic correspondence. In: *International Conference on Computer Vision*.

Hansen P and Browning B (2014) Visual place recognition using hmm sequence matching. In: *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*. IEEE, pp. 4549–4555.

Hariharan B, Arbeláez P, Girshick R and Malik J (2015) Hypercolumns for object segmentation and fine-grained localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 447–456.

He K, Zhang X, Ren S and Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778.

Hegarty M and Waller D (2004) A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence* 32(2): 175–191.

Heng L, Furgale P and Pollefeys M (2015) Leveraging image-based localization for infrastructure-based calibration of a multi-camera rig. *Journal of Field Robotics* 32(5): 775–802.

Huang F, Jin C, Zhang Y, Weng K, Zhang T and Fan W (2018) Sketch-based image retrieval with deep visual semantic descriptor. *Pattern Recognition* 76: 537–548.

Jégou H, Douze M, Schmid C and Pérez P (2010) Aggregating local descriptors into a compact image representation. In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, pp. 3304–3311.

Johnson J, Krishna R, Stark M, Li LJ, Shamma D, Bernstein M and Fei-Fei L (2015) Image retrieval using scene graphs. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3668–3678.

Kawasaki A, Saito H and Hara K (2015) Motion estimation for non-overlapping cameras by improvement of feature points matching based on urban 3d structure. In: *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, pp. 1230–1234.

Kim HJ, Dunn E and Frahm JM (2017a) Learned contextual feature reweighting for image geo-localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2136–2145.

Kim S, Min D, Ham B, Jeon S, Lin S and Sohn K (2017b) Fcss: Fully convolutional self-similarity for dense semantic correspondence. In: *Proc. IEEE Conf. Comp. Vision Patt. Recog*, volume 1. p. 8.

Kobyshev N, Riemenschneider H and Van Gool L (2014) Matching features correctly through semantic understanding. In: *3D Vision (3DV), 2014 2nd International Conference on*, volume 1. IEEE, pp. 472–479.

Kozhevnikov M, Motes MA, Rasch B and Blajenkova O (2006) Perspective-taking vs. mental rotation transformations and how they predict spatial navigation performance. *Applied cognitive psychology* 20(3): 397–417.

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

Lin G, Milan A, Shen C and Reid I (2017) Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, volume 1. p. 3.

Liu L, Shen C and van den Hengel A (2015) The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4749–4757.

Long J, Shelhamer E and Darrell T (2015) Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3431–3440.

Long JL, Zhang N and Darrell T (2014) Do convnets learn correspondence? In: *Advances in Neural Information Processing Systems*. pp. 1601–1609.

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60(2): 91–110.

Lowry S and Andreasson H (2018) Lightweight, viewpoint-invariant visual place recognition in changing environments. *IEEE Robotics and Automation Letters* 3(2): 957–964.

Lowry S, Sünderhauf N, Newman P, Leonard JJ, Cox D, Corke P and Milford MJ (2016) Visual place recognition: A survey. *IEEE Transactions on Robotics* 32(1): 1–19.

Lu X, Zheng X and Li X (2017) Latent semantic minimal hashing for image retrieval. *IEEE Transactions on Image Processing* 26(1): 355–368.

Lucas BD and Kanade T (1981) An iterative image registration technique with an application to stereo vision .

Lynen S, Bosse M and Siegwart R (2017) Trajectory-based place-recognition for efficient large scale localization. *International Journal of Computer Vision* 124(1): 49–64.

Ma R, Chen J and Su Z (2010) MI-SIFT: mirror and inversion invariant generalization for SIFT descriptor. In: *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, pp. 228–235.

Maddern W, Pascoe G, Linegar C and Newman P (2017) 1 year, 1000 km: The oxford robotcar dataset. *IJ Robotics Res.* 36(1): 3–15.

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

Mishchuk A, Mishkin D, Radenovic F and Matas J (2017) Working hard to know your neighbor's margins: Local descriptor learning loss. In: *Advances in Neural Information Processing Systems*. pp. 4826–4837.

Mishkin D, Radenovic F and Matas J (2018) Repeatability is not enough: Learning affine regions via discriminability. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 284–300.

Mousavian A and Košecka J (2015) Semantically Aware Bag-of-Words for Localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Mousavian A and Kosecka J (2016) Semantic image based geolocation given a map (author's initial manuscript). Technical report, George Mason University Fairfax United States.

Mousavian A, Košecká J and Lien JM (2015) Semantically guided location recognition for outdoors scenes. In: *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, pp. 4882–4889.

Mukherjee A, Chakraborty S and Saha SK (2017) Learning deep representation for place recognition in slam. In: *International Conference on Pattern Recognition and Machine Intelligence*. Springer, pp. 557–564.

Mur-Artal R, Montiel JMM and Tardos JD (2015) Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics* 31(5): 1147–1163.

Mur-Artal R and Tardós JD (2017) Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics* 33(5): 1255–1262.

Murillo AC, Singh G, Kosecká J and Guerrero JJ (2013) Localization in urban environments using a panoramic gist descriptor. *IEEE Transactions on Robotics* 29(1): 146–160.

Naseer T, Burgard W and Stachniss C (2018) Robust visual localization across seasons. *IEEE Transactions on Robotics* .

Naseer T, Oliveira GL, Brox T and Burgard W (2017) Semantics-aware visual localization under challenging perceptual conditions. In: *IEEE International Conference on Robotics and Automation (ICRA)*.

Naseer T, Spinello L, Burgard W and Stachniss C (2014) Robust visual robot localization across seasons using network flows. In: *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Neubert P and Protzel P (2016) Beyond holistic descriptors, keypoints, and fixed patches: Multiscale superpixel grids for place recognition in changing environments. *IEEE Robotics and Automation Letters* 1(1): 484–491.

Noh H, Araujo A, Sim J, Weyand T and Han B (2017) Large-Scale Image Retrieval with Attentive Deep Local Features. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3456–3465.

Oliva A and Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision* 42(3): 145–175.

Pepperell E, Corke P and Milford M (2016) Routed roads: Probabilistic vision-based place recognition for changing conditions, split streets and varied viewpoints. *The International Journal of Robotics Research* 35(9): 1057–1179.

Pepperell E, Corke PI and Milford MJ (2014) All-environment visual place recognition with smart. In: *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, pp. 1612–1618.

Philbin J, Isard M, Sivic J and Zisserman A (2010) Descriptor Learning for Efficient Retrieval. In: *European Conference on Computer Vision*. Springer, pp. 677–691.

Radenović F, Tolias G and Chum O (2018) Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Razavian AS, Azizpour H, Sullivan J and Carlsson S (2014) Cnn features off-the-shelf: an astounding baseline for recognition. In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*. IEEE, pp. 512–519.

Rocco I, Arandjelovic R and Sivic J (2017) Convolutional neural network architecture for geometric matching. In: *Proc. CVPR*, volume 2.

Rublee E, Rabaud V, Konolige K and Bradski G (2011) Orb: An efficient alternative to sift or surf. In: *Computer Vision (ICCV), 2011 IEEE international conference on*. IEEE, pp. 2564–2571.

Salas-Moreno RF, Newcombe RA, Strasdat H, Kelly PH and Davison AJ (2013) Slam++: Simultaneous localisation and mapping at the level of objects. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1352–1359.

Sarlin PE, Cadena C, Siegwart R and Dymczyk M (2018a) From Coarse to Fine: Robust Hierarchical Localization at Large Scale. *arXiv preprint arXiv:1812.03506* .

Sarlin PE, Debraine F, Dymczyk M and Siegwart R (2018b) Leveraging Deep Visual Descriptors for Hierarchical Efficient Localization. In: *Conference on Robot Learning*. pp. 456–465.

Sattler T, Maddern W, Toft C, Torii A, Hammarstrand L, Stenborg E, Safari D, Okutomi M, Pollefeys M, Sivic J, Kahl F and Pajdla T (2018) Benchmarking 6dof outdoor visual localization in changing conditions. In: *Proc. CVPR*.

Savinov N, Seki A, Ladicky L, Sattler T and Pollefeys M (2017) Quad-networks: unsupervised learning to rank for interest point detection. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schönberger JL, Pollefeys M, Geiger A and Sattler T (2018) Semantic Visual Localization. *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)* .

Schreiber M, Knöppel C and Franke U (2013) Laneloc: Lane marking based localization using highly accurate maps. In: *Intelligent Vehicles Symposium (IV), 2013 IEEE*. IEEE, pp. 449–454.

Schuster S, Krishna R, Chang A, Fei-Fei L and Manning CD (2015) Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In: *Proceedings of the fourth workshop on vision and language*. pp. 70–80.

Seymour Z, Sikka K, Chiu HP, Samarasekera S and Kumar R (2018) Semantically-Aware Attentive Neural Embeddings for Image-based Visual Localization. *arXiv preprint arXiv:1812.03402* .

Simonyan K, Vedaldi A and Zisserman A (2014) Learning Local Feature Descriptors Using Convex Optimisation. *IEEE Trans. Pattern Anal. Mach. Intell.* 36(8): 1573–1585.

Singh G and Košecká J (2016) Semantically Guided Geo-location and Modeling in Urban Environments. In: *Large-Scale Visual Geo-Localization*. Springer, pp. 101–120.

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

Sünderhauf N, Shirazi S, Dayoub F, Upcroft B and Milford M (2015) On the performance of convnet features for place recognition. In: *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, pp. 4297–4304.

Sunderhauf N, Shirazi S, Jacobson A, Dayoub F, Pepperell E, Upcroft B and Milford M (2015) Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. *Proceedings of Robotics: Science and Systems XII* .

Taira H, Okutomi M, Sattler T, Cimpoi M, Pollefeys M, Sivic J, Pajdla T and Torii A (2018) InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 7199–7209.

Toft C, Olsson C and Kahl F (2017) Long-term 3d localization and pose from semantic labellings. In: *ICCV Workshops*, volume 2. p. 3.

Toft C, Stenborg E, Hammarstrand L, Brynte L, Pollefeys M, Sattler T and Kahl F (2018) Semantic match consistency for long-term visual localization. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 383–399.

Tolias G, Sicre R and Jégou H (2016) Particular object retrieval with integral max-pooling of cnn activations. In: *International Conference on Learning Representations*.

Torii A, Arandjelovic R, Sivic J, Okutomi M and Pajdla T (2015) 24/7 place recognition by view synthesis. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1808–1817.

Tribou MJ, Harmat A, Wang DW, Sharf I and Waslander SL (2015) Multi-camera parallel tracking and mapping with non-overlapping fields of view. *The International Journal of Robotics Research* 34(12): 1480–1500.

Ufer N and Ommer B (2017) Deep semantic feature matching. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5929–5938.

Vysotska O and Stachniss C (2016) Lazy data association for image sequences matching under substantial appearance changes. *IEEE Robotics and Automation Letters* 1(1): 213–220.

Vysotska O and Stachniss C (2017) Relocalization under substantial appearance changes using hashing .

Wolcott RW and Eustice RM (2017) Robust lidar localization using multiresolution gaussian mixture maps for autonomous driving. *The International Journal of Robotics Research* 36(3): 292–319.

Yi KM, Trulls E, Lepetit V and Fua P (2016) Lift: Learned invariant feature transform. In: *European Conference on Computer Vision*. Springer, pp. 467–483.

Yu X, Chaturvedi S, Feng C, Taguchi Y, Lee TY, Fernandes C and Ramalingam S (2018) VLASE: Vehicle Localization by Aggregating Semantic Edges. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 3196–3203.

Yue-Hei Ng J, Yang F and Davis LS (2015) Exploiting local features from deep networks for image retrieval. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. pp. 53–61.

Zamir AR, Wekel T, Agrawal P, Wei C, Malik J and Savarese S (2016) Generic 3d representation via pose estimation and matching. In: *European Conference on Computer Vision*. Springer, pp. 535–553.

Zeiler MD and Fergus R (2014) Visualizing and understanding convolutional networks. In: *European conference on computer vision*. Springer, pp. 818–833.

Zhang N, Shelhamer E, Gao Y and Darrell T (2015) Fine-grained pose prediction, normalization, and recognition. *arXiv preprint arXiv:1511.07063* .

Zhao WL and Ngo CW (2013) Flip-invariant SIFT for copy and object detection. *IEEE Transactions on Image Processing* 22(3): 980–991.

Zhou B, Lapedriza A, Khosla A, Oliva A and Torralba A (2017) Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Zhou B, Lapedriza A, Xiao J, Torralba A and Oliva A (2014) Learning deep features for scene recognition using places database. In: *Advances in neural information processing systems*. pp. 487–495.