CoHOG: A Light-Weight, Compute-Efficient, and Training-Free Visual Place Recognition Technique for Changing Environments

Mubariz Zaffar, Shoaib Ehsan, Michael Milford, and Klaus McDonald-Maier

***Abstract*—This letter presents a novel, compute-efficient and training-free approach based on Histogram-of-Oriented-Gradients (HOG) descriptor for achieving state-of-the-art performance-per-compute-unit in Visual Place Recognition (VPR). The inspiration for this approach (namely CoHOG) is based on the convolutional scanning and regions-based feature extraction employed by Convolutional Neural Networks (CNNs). By using image entropy to extract regions-of-interest (ROI) and regional-convolutional descriptor matching, our technique performs successful place recognition in changing environments. We use viewpoint- and appearance-variant public VPR datasets to report this matching performance, at lower RAM commitment, zero training requirements and 20 times lesser feature encoding time compared to state-of-the-art neural networks. We also discuss the image retrieval time of CoHOG and the effect of CoHOG's parametric variation on its place matching performance and encoding time.**

***Index Terms*—SLAM, visual place recognition, autonomous vehicle navigation, computer vision for automation.**

I. Introduction

FOR A ROBOT to operate autonomously, it needs to be able to remember previously visited places. This ability to remember places has been discussed and widely researched (surveyed by Lowry *et al.* [1]) as the sub-domain of visual-SLAM (Simultaneous Localization and Mapping), namely Visual Place Recognition (VPR). VPR is a well-defined, albeit a highly challenging problem since places change their appearance rapidly due to varying viewpoints and conditions. Other than environmental variations, texture-less and low-informative scenes also pose difficulty to place matching. We show examples of all these challenges taken from public VPR datasets [2], [3], [4] in Fig. 1. Given a query image, the task of a VPR system is to retrieve the best matched image of the same place with robustness to viewpoint and conditional variations under the constraints of run-time memory, processing power and/or pre-deployment training needs.

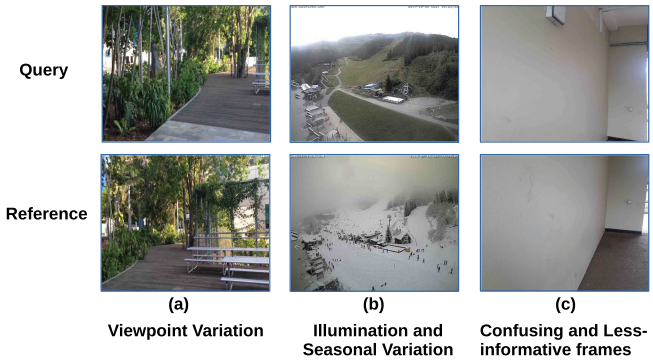

Fig. 1. Difference in the appearance of places under viewpoint and seasonal changes and confusing less-informative scenes.

Prior to the use of neural network based techniques, VPR research was primarily based on local and global handcrafted feature descriptors. Local feature descriptors extract and describe keypoints (areas of interest) from an image, therefore they are primarily viewpoint invariant but suffer from illumination variation. Global feature descriptors, on the other hand, suffer from translational and/or rotational viewpoint change but they are moderately illumination invariant. Moving away from handcrafted feature descriptors, the application of Convolutional Neural Networks (CNNs) to VPR was first studied by Chen *et al.* [5]. Since then, different CNNs with and without architectural modifications have incrementally shown state-of-the-art VPR performance. However, CNNs (and Convolutional Auto-encoders as in [6]) require significant model training with their deployment accuracy directly linked to the size, inter-sample variance and nature of the training dataset. Training of VPR-specific CNNs requires large-scale labelled datasets of places from a multitude of environments, which is a practical limitation. Moreover, training of these CNNs requires dedicated Graphical Processing Units (GPUs) with training time usually ranging from a few days to a few weeks. One key limitation of neural network based techniques is their intense computational nature requiring significantly higher run-time memory and feature encoding-time compared to handcrafted feature descriptors. Thus, while the success of these recent CNN-based techniques from the perspective of place matching is evident, their practical deployment in field is restricted. More specifically, such computational intensiveness raises concerns

Manuscript received September 9, 2019; accepted December 31, 2019. Date of publication January 28, 2020; date of current version February 11, 2020. This letter was recommended for publication by Associate Editor C. Cadena Lerma Editor T. Sattler and upon evaluation of the reviewers' comments. This work was supported by UK Engineering and Physical Sciences Research Council through Grants EP/R02572X/1 and EP/P017487/1 and in part by the RICE project funded by the National Centre for Nuclear Robotics Flexible Partnership Fund. *(Corresponding author: Mubariz Zaffar.)*

M. Zaffar, S. Ehsan, and K. McDonald-Maier are with the School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ U.K. (e-mail: mubariz.zaffar@essex.ac.uk; sehsan@essex.ac.uk; kdm@essex.ac.uk).

M. Milford is with the School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4000, Australia (e-mail: michael.milford@qut.edu.au).

Digital Object Identifier 10.1109/LRA.2020.2969917

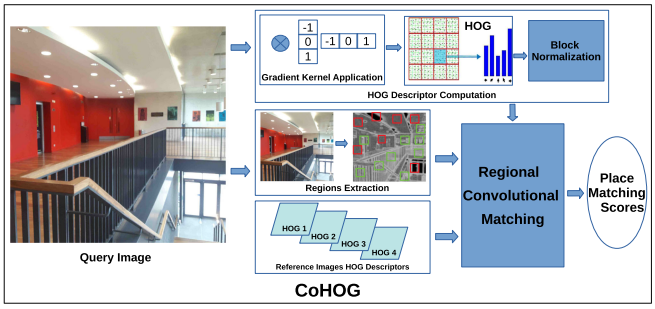

Fig. 2. The developed technique (CoHOG) is explained here. Each query image goes through ROI extraction and HOG computation, which are then fed to the convolutional matching block. This block outputs a similarity score against each reference image in the robot map. The green squares in region extraction block represent salient regions while the red squares are less-informative confusing regions.

for deployability on resource-constrained platforms (including battery-powered aerial, micro-aerial and ground vehicles) as identified in [7], [8].

In this letter, we propose a novel technique based on handcrafted feature descriptors delivering state-of-the-art (or close to state-of-the-art) VPR performance with no training requirements compared to CNNs. Our technique has significantly lower feature encoding time and RAM commitment while delivering comparable place matching performance on challenging viewpoint- and conditionally-variant datasets. The inspiration for our approach is drawn from the following:

1) By design, CNNs are able to scan an entire image for a particular feature and irrespective of the location of that feature in an image, the same CNN filter (layer activations) will fire.
2) CNNs trained/fine-tuned for VPR have the ability to extract regions-of-interest (ROI) which are informative and distinct.
3) CNNs trained on condition-variant VPR datasets can internally learn representations of places/images which are immune to seasonal and illumination variations.

From the above list, both 1 and 2 contribute towards viewpoint invariance. This is further improved by manually introducing viewpoint variation in training datasets. Conditional invariance is predominantly the result of 3, not user-defined and essentially a black-box.

By deriving motivation from this behavior of CNNs, our technique first computes the entropy map of an image and extracts information-rich regions from it. Each of these ROI are then locally described by dedicated HOG-descriptors. Secondly, we use convolutional matching of regional HOG-descriptors that provides viewpoint invariance. This regional-convolutional matching is based on standard matrix multiplication and is therefore compute-efficient. For illumination invariance, block normalization of HOG-descriptors is used, which shows acceptable performance on conditionally-variant datasets. Our choice of HOG-descriptor is based on its reliable performance across illumination and seasonal variation as shown by McManus *et al.* [9], and its utility as an underlying feature descriptor for training a VPR-specific convolutional auto-encoder in [6]. The image retrieval scheme of CoHOG can be summarized by Fig. 2.

The remainder of the paper is organized as follows. In Section II, a comprehensive literature review regarding VPR state-of-the-art is presented. Section III presents the details of the CoHOG technique developed in this work. Section IV puts forth the results and analysis obtained by evaluating our work and contemporary VPR techniques on public VPR datasets. Finally, conclusions and future directions are presented in Section V.

II. Literature Review

An extensive review of the challenges faced by VPR, published work and directions of research has been performed by authors in [1]. VPR research can be broken down into two main stages i.e. handcrafted feature descriptors based VPR and neural networks based VPR.

The use of handcrafted features in VPR can subsequently be classified into two main streams: local and global feature descriptors. Scale Invariant Feature Transform (SIFT [10]) and Speeded Up Robust Features (SURF [11]) are one of the most widely used local descriptors and have been applied to VPR problem in [12], [13], [14], [15], [16]. FAB-MAP (Frequent Appearance Based Mapping [17]) is a probabilistic visual-SLAM algorithm that represents places as visual words and uses SURF as the underlying interest point detector. An extension to FAB-MAP is presented by utilizing odometry information in CAT-SLAM [18]. Center Surround Extremas for real-time feature detection and matching (CenSurE [19]) has been used for VPR in [20]. FAST [21] is a high-speed corner detector for real-time image processing that has been used for SLAM by Mei *et al.* [22], coupled with SIFT descriptor. One common drawback to all these keypoint based approaches is the extensive matching requirements, which has been addressed by Bag of visual Words (BoW [23]) approach. BoW collects visually similar features in dedicated bins (pre-defined or learned by training a visual-dictionary) without topological consideration, enabling direct matching of BoW descriptors. Different research works have used BoW for VPR, including [24]–[27].

Global feature descriptors like Gist [28] use Gabor filters to create the signature of an entire image and have been used for VPR with panaromic images by Murillo *et al.* [29] and Singh *et al.* [30]. BRIEF [31] descriptor due to its lower encoding requirements and faster matching time is combined with Gist by Sünderhauf *et al.* [32] to perform large scale visual-SLAM. Whole-Image SURF (WI-SURF) is a global variant of SURF and has been used for visual localization by Badino *et al.* [33]. Seq-SLAM [34] does not perform feature extraction but uses normalized pixel-intensity matching (in global fashion) between a sequence of camera frames to achieve VPR in highly challenging environments. SMART [35] extends Seq-SLAM, by considering the variable speed of a robotic platform. McManus *et al.* [9] have proposed an approach similar in spirit to our work, where scene signatures are extracted and described by dedicated HOG descriptors. However, their approach assumes prior knowledge of the approximate location/environment of the robot such that the signatures are pre-learnt for that environment.

Features extracted from CNNs showed promising results on condition- and viewpoint-variant datasets, leading to a paradigm shift in VPR research from traditional handcrafted feature descriptors to neural network activations-based descriptors. Chen *et al.* [5] used features from all layers of Overfeat Network [36] and integrated it into the spatial filtering scheme of Seq-SLAM. Improving upon CNN-based VPR, Chen *et al.* [37] trained two neural networks on Specific Places Dataset (SPED), namely AMOSNet and HybridNet. AMOSNet was trained from scratch

on SPED while HybridNet initialized weights from top-5 convolutional layers of Caffe-Net. Different off-the-shelf feature encoding methods have been used to create the signature of an image from CNN activations; including cross-pooling [38], holistic pooling [39] and multi-scale pooling [37]. However, in Net-VLAD [40], authors introduce a new VLAD (Vector-of-Locally-Aggregated-Descriptors [41]) layer into the CNN architecture for end-to-end VPR-specfic training, achieving excellent results. Recently, CNN-based description of images/places using only regions of interest (ROI) showed enhanced performance compared to whole-image description. The work in [42], namely R-MAC (Regions of Maximum Activated Convolutions) uses max-pooling on cropped areas in CNN layers' features to extract ROI. Chen *et al.* [43] in Cross-Region-BoW used the CNN layers behaving as high-level feature extractors to identify salient regions in an input image which were subsequently described by low-level feature encoding convolutional layers. This work was followed-up with a flexible attention-based model for region extraction [2]. Khaliq *et al.* [44] combine VLAD with ROI-extraction to show significant robustness to appearance and viewpoint variation. A convolutional auto-encoder network is trained in an unsupervised fashion by Merrill *et al.* [6], utilizing HOG-descriptors of images and synthetic viewpoint variation. While most of these works have been explored specifically for visual-localization, some recent techniques including SuperPoint [45] and D2-net [46] propose generic, deep-learnt, sparse descriptors that are robust across various conditional changes. Authors in [47], [48] have formulated visual localisation as a two-stage process: 1) global matching-based, less-intensive place matching candidates selection 2) local features-based, intensive final candidate selection with focus on spatial constraints. Other interesting approaches to place recognition have also been adopted, including semantic segmentation-based visual localisation (as in [49], [50], [51]) and object proposals-based place recognition [52].

While handcrafted feature descriptors suffer from viewpoint and conditional variation, neural networks, on the other hand, require significant training, computational power, physical memory and their performance is specific to the size and environment explored in the training dataset. In this work, we therefore propose a novel technique utilizing handcrafted feature descriptors for VPR that achieves state-of-the-art or comparable place matching performance on public VPR datasets. Our technique intrinsically does not need any training, performs feature encoding in real-time up to 50 frames-per-second (FPS) and has a lower run-time memory (RAM) requirement at deployment in contrast to deep learning-based techniques.

III. Methodology

This section presents the methodology adopted in our work that constitutes CoHOG. The proposed technique can be broken down into 7 primary blocks (as shown in Fig. 3) for end-to-end VPR image retrieval. The query image can be any incoming RGB camera frame which is converted to grayscale and resized to $W1 \times H1$. The robot map consists of pre-computed HOG-descriptors of reference images. Please note that we have used 'vanilla' HOG in this work, but our implementation computes HOG in the regional sense instead of the usual global fashion. A sub-section is dedicated to each of three crucial computational tasks of our technique, namely HOG-descriptor computation, ROI extraction and regions-based convolutional matching.

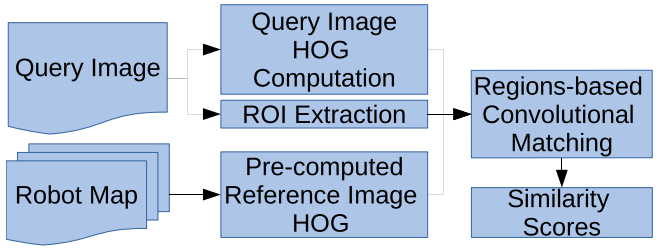

Fig. 3. The block-level overview of CoHOG is shown here.

Algorithm 1: Computing Entropy Map.

$Entropy\_Map = Zeros\_Matrix(W1, H1)$
$Max\_Entropy\_Value = log_2\,(256) = 8$
$Create\ a\ Histogram\ of\ 256\ Pixel\ Intensity\ Bins$
$Radius\ =\ UserDefinedConstant$
for all $Pixels\ in\ Image$ **do**
  $Origin\ =\ Pixel\_Index$
  $Local\_Neighbourhood\ =\ Circle(Origin, Radius)$
  $Local\_Neigh\_List\ =$
  $Append(Local\_Neighbourhood)$
end for
for all $Elements\ in\ Local\_Neigh\_List$ **do**
  for all $Valid\ Pixels\ in\ Local\_Neighbourhood$ **do**
    if $Current\_Pixel\_Intensity\ lies\ in\ BinX$ **then**
      $Items\_in\_BinX = Items\_in\_BinX + 1$
    end if
  end for
  $Entropy\_Map(i, j) =$
  $log_2\,(No.\ of\ Filled\ Histogram\ Bins)$
  $Clear\ all\ Histogram\ Bins$
end for
$Normalize\ Entropy\_Map\ with\ Max\_Entropy\_Value$

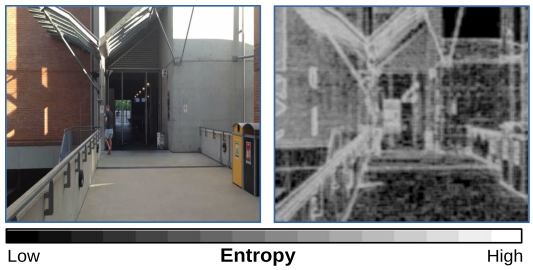

Fig. 4. Example of query image [left] with its corresponding entropy map [right] is shown here. Texture-less walls and floors get filtered out as lower entropy areas which is consistent with our motivation to discard such regions.

A. ROI Extraction

Regions-of-interest based image matching has recently been the subject of significant VPR research [43], [2], [44]. In CoHOG, we use regions in an image that are information-rich. Firstly, we compute the entropy map for each query image using the following algorithm.

The entropy map has the same dimensions as the query image i.e. $W1 \times H1$ and example query images with entropy maps computed using our algorithm are shown in Fig. 4. We now define a region in an image as a $W2 \times H2$ image patch. Thus, a $W1 \times H1$ image with regions/patches of size $W2 \times H2$

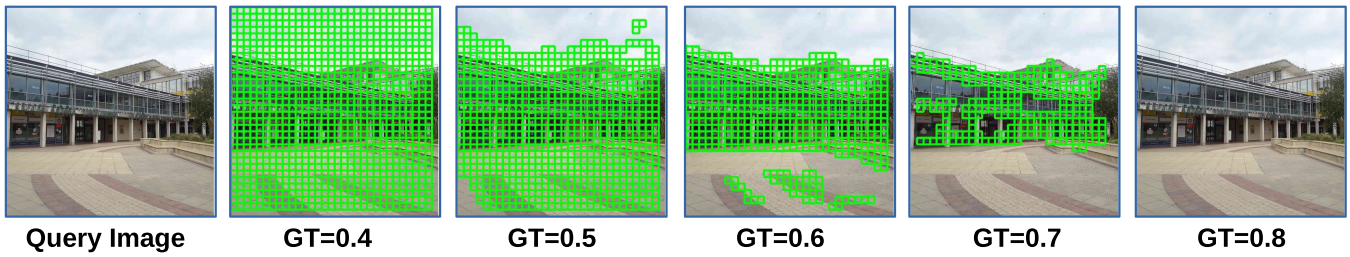

Fig. 5. ROI extracted by CoHOG are shown here with varying GT. Each good region is represented by a green colored square. Increasing GT reduces the number of regions selected by our technique. A clear range exists between $GT = 0.5 - 0.7$, where confusing and low-informative regions coming from sky and texture-less walls/floors are filtered out, while maintaining a reasonable number of regions for subsequent regional convolutional matching.

contains N regions in total, whose goodness is represented by a matrix R;

$$\text{Where;} \quad N = (H1/H2) \times (W1/W2)$$

$$R = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{mn} & r_{mn} & \dots & r_{mn} \end{bmatrix}$$

$$\text{Where;} \quad m = H1/H2 \quad n = W1/W2$$
$$r_{ij} \in \mathbb{Z}_2 | \mathbb{Z}_2 = [0, 1]$$
$$r_{ij} = 1 \ | \text{Region} = \text{Good}$$
$$r_{ij} = 0 \ | \text{Region} = \text{Confusing}$$

For evaluating the goodness r_{xy} of each region in R, the entropy map is represented as a matrix E. This entropy matrix has a size of $W1 \times H1$ with element values between $0 - 1$ (with 1 being the ideal value) and has the below shape.

$$E = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1W_1} \\ e_{21} & e_{22} & \dots & e_{2W_1} \\ \vdots & \vdots & \ddots & \vdots \\ e_{H_1W_1} & e_{H_1W_1} & \dots & e_{H_1W_1} \end{bmatrix}$$

$$\text{where;} \quad \{e_{ij} \in K \ | \ K \subseteq \mathbb{R} \wedge K = \{0, \dots, 1\}\}$$

The goodness r_{xy} of each region is calculated by thresholding the average entropy values of a $(W2 \times 2) \times (H2 \times 2)$ block size i.e. each block containing 4 regions (each region of size $W2 \times H2$), where all 4 regions have a common corner. Such a block-level evaluation provides consistency with HOG-descriptor computation, as shown later in sub-section III-B. The stride of this block-level goodness evaluation is $Stride = W2 = H2$ and hence the total number of regional blocks for evaluation is $M = n - 1 \times m - 1$. All G regions which have an entropy score e_{xy} greater than or equal to the goodness threshold GT are selected for matching. Therefore, G is a variable depending on the scene being represented in an image and may vary from one query image to another. Selecting regions in this manner compared to the conventional Top-G (where G is a constant) regions selection provides more saliency and computational advantages. If an image has more confusing regions, only a few salient regions are selected. This helps in successfully matching low-textured images and is not possible with Top-G regions selection. Discarding confusing regions before regional convolutional matching also leads to lesser computational intensity. Fig. 5 shows examples of good regions extracted with varying GT.

B. HOG-Descriptor Computation

Histogram-of-Oriented-Gradients (HOG) [53], [54] is a well-established handcrafted computer vision technique used originally for object detection. The end-to-end HOG-descriptor computation is quickly summarized as follows:

1) A gradient map is computed for an input grayscale image of size $W1 \times H1$.
2) A histogram-of-oriented-gradients (HOG) is created and computed for all N regions in the image, where every region has a size of $W2 \times H2$. Each regional-histogram has L bins, such that a bin is identified by a range of gradient angles assigned to it.
3) HOG computed previously is L2-normalised at a block level of size $(W2 \times 2) \times (H2 \times 2)$. This results in a descriptor of depth $4 \times L$ with the total number of block-level HOG-descriptors equal to M. Refer to sub-section III-A, each ROI now has a corresponding HOG-descriptor of depth $4 \times L$ which is illumination invariant and can be easily indexed/retrieved.

C. Regions Based Convolutional Matching

After HOG-descriptor computation, a query image is essentially converted into M regions with each region described by a vector of length $4 \times L$. Based on ROI evaluation, these M regions are reduced to G salient regions. This allows us to shape the query image HOG-descriptor as a 2-dimensional matrix A with dimensions $[G, 4 \times L]$. The reference image is also composed of M regions with descriptors of depth $4 \times L$. Thus, the reference image is shaped as a matrix B with dimensions $[M, 4 \times L]$. Next, standard matrix multiplication is performed between A and B^T yielding a matrix C of dimensions $[G, M]$. Each row of matrix C represents a query image region and every column in C represents the cosine-matching scores against all reference image regions.

We employ max-pooling across all rows of matrix C to find the best matched reference image candidate region for every query image region, which yields a vector D having length G. Finally, we take the arithmetic-mean of vector D giving us the similarity score of a query and reference image in the range of $0 - 1$. A query image is matched with all reference images, such that the reference image with the highest matching score is selected as the best match.

IV. Results and Analysis

This section first discusses the experimental setup used in our analysis including the VPR datasets, VPR techniques and evaluation metric used for assessing CoHOG's performance. We then present a detailed qualitative and quantitative comparison of CoHOG with state-of-the-art VPR techniques on the fronts of image matching, feature encoding time and run-time memory requirements. We also discuss the image retrieval performance of CoHOG and show the effect on computational and matching performance by varying different parameters to give the reader an insight into the selection of thresholds.

A. Experimental Setup

In order to evaluate CoHOG, we have utilized 5 VPR datasets that represent all the challenges in VPR (as identified in Section I). For viewpoint variation, we have used Gardens Point dataset [55] consisting of 210 query images and 210 reference images. Frame-to-frame ground-truth is available for this dataset. Secondly, we use the ESSEX3IN1 dataset which was first introduced in [4]. This dataset consists of highly confusing and challenging images of places with viewpoint variations. The total number of query images is 210 with one-to-one ground-truth reference. Thirdly, we use the SPEDTest dataset which has been introduced in [5] and is a sub-set of the original Specific Places Dataset [37]. It comprises of 607 query and 607 reference images representing conditional variation resulting from changes in seasons and times of day. We also employ the synthetically created Synthia [56] dataset, which consists of city-like traversal during Winter and Spring seasons. The number of query and reference images are 959 and 947, respectively. Finally, we use the low-quality, highly dynamic and blurry Cross-Seasons dataset [57] consisting of 206 sunny query images and 202 dusk reference images. Results on this dataset present the failure-cases of CoHOG and identify important directions for future research.

For comparison with CoHOG, we use all contemporary VPR techniques reviewed by authors in [58]. The implementation details, selected parameters and evaluation platform have all been kept similar to the setup of [58] for a fair comparison, except that we use AlexNet for the Region-VLAD approach instead of HybridNet. We have also reported the performance for using Top-G (at G = 200, 400 and 800) based regions selection with CoHOG. As our evaluation-metric, we examine the place matching performance per compute unit of all VPR techniques. The extensive review performed by Lowry *et al.* in [1] identifies high precision to be a desirable characteristic of a VPR system due to the advent of false-positive prediction systems (as in [59], [60], [61]). On the other hand, authors in [6], [44], [7] and [58] have identified feature encoding time (t_e) to be a crucial computational metric. Therefore, by combining precision at 100% recall with encoding time per image, we define the Performance-per-Compute-Unit (PCU) as below.

$$PCU = \text{Precision} \times \log\left(\frac{\text{Max Feature Encoding Time}}{\text{Feature Encoding Time}} + 9\right)$$

In the above equation, higher precision directly leads to higher PCU. However, for feature encoding time t_e, we compute the logarithmic encoding time boost for a given VPR technique to provide a reasonable combination of precision and encoding time metrics. Thus, only exponential increase in encoding time for a highly precise VPR technique leads to increase in PCU. Maximum feature encoding time (t_{e_max}) belongs to the most computationally intensive VPR technique, which in our case is Cross-Region-BoW with the highest feature encoding time of 0.83 seconds. A scalar '9' is added to ensure that $PCU = Precision$ for the technique with $t_e = t_{e\_max}$, instead of $PCU = 0$, thus providing an interpretable scale.

B. Performance Evaluation

This section provides a detailed comparison of CoHOG with state-of-the-art VPR techniques on the frontiers of performance-per-compute-unit and run-time memory requirements. The reported performance is for $GT = 0.5, W1 = H1 = 512, W2 = H2 = 16$ and $L = 8$.

1) Place Matching Performance: This sub-section presents the PCU of CoHOG in comparison with other VPR techniques. While Fig. 7 shows the PCU of all techniques, the absolute values of precision at 100% recall and feature encoding time are listed in Table I for reader's reference.

CoHOG achieves state-of-the-art PCU on all the 5 datasets utilised in our work as shown in Fig. 7. We also report state-of-the-art precision on ESSEX3IN1 dataset and comparable precision on other datasets (except cross-seasons dataset), as listed in Table I. The viewpoint variation in ESSEX3IN1 dataset is catered for by CoHOG's regional convolutional matching while confusing frames (and/or regions within) are handled by our entropy-based region extraction. This matching performance is qualitatively shown in Fig. 6. We achieve close-to-ideal place matching precision on Gardens Point dataset and Fig. 6 shows samples of places correctly matched by our technique despite the viewpoint variation. The nature of challenges handled in SPEDTest and Synthia datasets is also depicted in Fig. 6, where we show that under notable seasonal and illumination changes, CoHOG can still retrieve correct place matches. However, the cross-seasons dataset consisting of low-quality images with motion blur and significant dynamic objects identifies important limitations of our gradient-based technique, that can intrinsically be handled by neural network-based techniques. Please note that the average number of regions employed by CoHOG are 730, 790, 780 and 750 on SPEDTest, Synthia, Gardens Point and ESSEX3IN1 datasets, respectively, but it still achieves better matching performance than Top-800, similar to our motivation in Subsection III-A.

The precision-recall curves for CoHOG are presented in Fig. 8. In an environment-aware VPR system, conditional variations are predictable [62] and can either be avoided or the VPR system be switched accordingly. Thus, given the lower computational and zero training requirements, CoHOG presents the best overall utility for a computationally-efficient VPR system in changing environments.

2) Run-Time Memory Requirements: Due to their intense computational requirements, neural network-based techniques have significantly higher run-time memory consumption which is an important factor for resource-constrained and battery-powered robotic platforms that are usually running multiple tasks simultaneously. We report the run-time memory consumption of all VPR techniques in Table I, which shows that CoHOG is light-weight compared to the rest of VPR techniques. This is because CoHOG intrinsically does not

TABLE I
PLACE MATCHING PRECISION, FEATURE ENCODING TIME, AND RAM COMMITMENT

Performance Metric	VPR Techniques (Platform: Intel(R) Xeon(R) Gold 6134 CPU @ 3.20GHz with 32 cores, 64GB RAM, No GPU)												
	HOG	AlexNet	AMOSNet	HybridNet	CALC	Cross-R-BOW	NetVLAD	R-VLAD	RMAC	Top-200	Top-400	Top-800	CoHOG
Precision ESSEX3IN1	0.01	0.14	0.26	0.28	0.1	0.62	0.76	0.56	0.12	0.75	0.79	0.82	0.84
Precision Gardens	0.2	0.49	0.64	0.81	0.44	0.81	0.95	0.9	0.42	0.74	0.82	0.87	0.9
Precision SPEDTest	0.02	0.03	–	–	0.02	0.5	0.74	0.54	0.6	0.4	0.44	0.48	0.51
Precision Synthia	0.37	0.89	0.91	0.92	0.76	0.89	0.95	0.86	0.92	0.67	0.83	0.91	0.92
Precision CrossSeasons	0.5	0.85	0.93	0.96	0.67	0.9	0.97	0.87	0.83	0.68	0.75	0.43	0.65
Encoding Time(sec)	0.007	0.67	0.36	0.36	0.027	0.83	0.77	0.46	0.47	0.02	0.02	0.02	0.02
RAM Consumption(MBs)	0.02	47.04	4.22	4.22	2.3	0.58	1.21	4.22	0.58	0.06	0.06	0.06	0.06

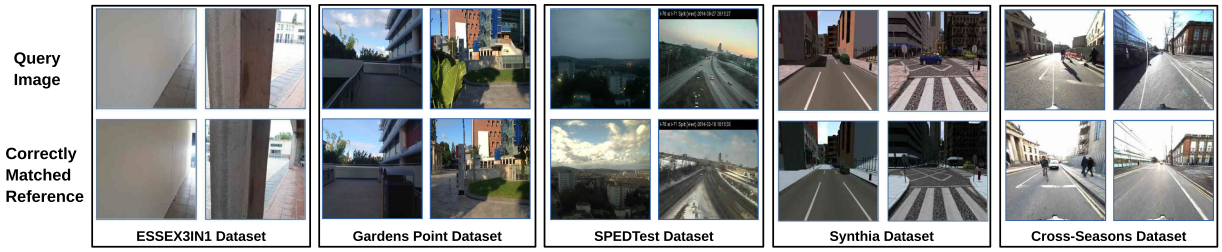

Fig. 6. Samples of correctly matched places by CoHOG on all 5 datasets are shown here. Given the viewpoint variations in ESSEX3IN1 [4] and Gardens Point datasets [55] datasets, CoHOG's regional-convolutional matching scheme can retrieve correct matches from the database. Even with the conditional variation in SPEDTest [2], Synthia [56] and Cross-Seasons datasets [57], our technique is able to correctly match places. More samples of correctly and incorrectly matched places and our open-source technique are provided at https://github.com/MubarizZaffar/CoHOG_Results_RAL2019.

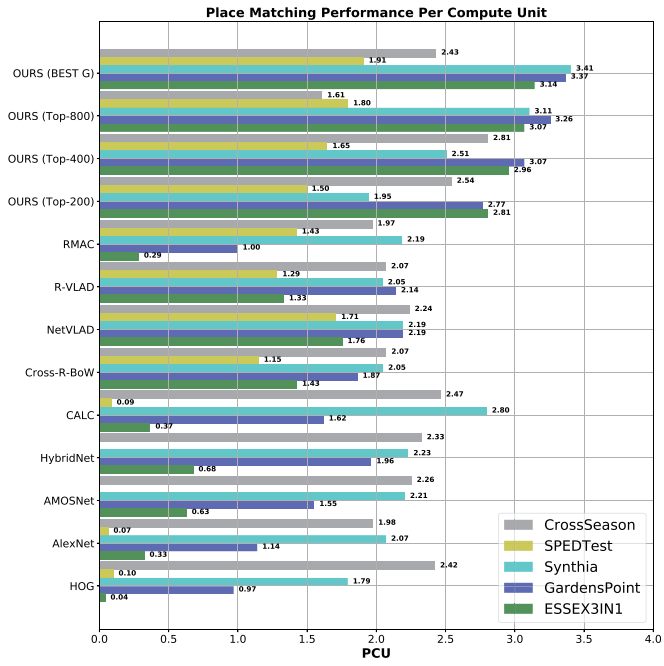

Fig. 7. The PCU of CoHOG is compared with all other VPR techniques. HybridNet and AMOSNet are trained on SPED dataset and thus not included for SPEDTest comparison.

involve loading/deployment of any machine-learning models into RAM for feature extraction/description. The reported RAM commitment is only for encoding a single query image.

3) Descriptor Matching Time: The descriptor matching time (t_m) represents the time required to match the feature descriptors of 2 images and determines the retrieval performance of a VPR system. The image retrieval time (T) for any VPR system can be modelled as $T = t_e + O(Z) \times t_m$. Where, O(Z) represents the total number of prospective candidate matches and could be linear, logarithmic or other depending upon the employed neighbourhood selection mechanism (e.g., linear search, approximate nearest neighbour search etc.). We further model t_m as $t_m = O(D) \times N1 \times N2$, where O(D) is the time required to match 2 descriptors of length D, N1 is the number of query image descriptors and N2 is the number of reference image descriptors. Theoretically, the value of t_m for CoHOG is $t_m = O(4 \times L) \times G \times M$. The values of t_e and t_m for our implementation of CoHOG are 0.02 sec and 0.2 msec, respectively, for the parameters specified in Subsection IV-B, such that the value of T will be $T = 0.02 + 0.0002 \times O(Z)$ sec.

Because it is computationally intractable to have a linear O(Z) for larger values of Z, different approaches exist to cater for this: 1) The total number of images in a map can be limited to a fixed value [63], 2) A spatial context can be introduced to search across images within a particular geographical radius [64], 3) A two-stage approach can be adopted to first extract possible candidate matches, followed by rigorous feature matching [47] [48], 4) Multi-processing systems can be employed to distribute the matching task across several processors. For further timing comparison between the techniques discussed in this work and understanding respective limitations, we would refer the reader to our previous work [58], provided the value of Z and the nature of O(Z) are known.

C. Parameter Sweep

This sub-section presents the effects of changing CoHOG's parameters. The parametric sweep is performed for GT, $W1$ and $W2$ on ESSEX3IN1 dataset. Each of the 3 parameters is first varied within a suitable range while keeping the other 2 constant, where the values of these constants are the same as used in Subsection IV-B. We also show the effect of varying $W1$ and $W2$ with a constant ratio.

The qualitative effect of variation in GT is already shown in Fig. 5 and the quantitative effect is reported in Fig. 9 (a). More salient regions and less confusing regions are selected

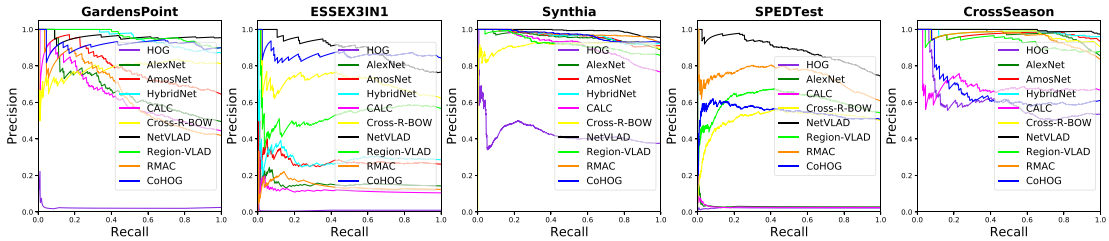

Fig. 8. The Precision-Recall curves for all 10 VPR techniques on the 5 datasets employed in our work are presented.

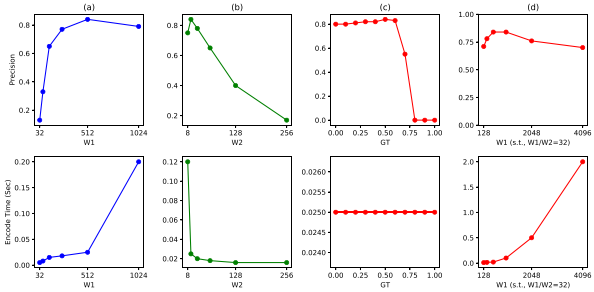

Fig. 9. The impact on CoHOG's performance by sweeping various thresholds within a suitable range is depicted here.

with increasing GT, leading to improved matching performance. The quantitative contribution of GT to place matching performance is inherent to the places being represented in the dataset and may vary. While feature encoding-time is independent of GT, it depends on both $W1$ and $W2$, as reported in Fig. 9. Matching performance improves with increasing image-size [Fig. 9 (b)] as greater number of gradients now contribute to the regional HOG-descriptors, which also results in increased gradient bin-assignment time. Increasing the cell-size reduces viewpoint-invariance as lesser number of regions (N) are now available for regional-convolutional matching, thus reducing matching performance [Fig. 9 (c)]. The key take-away here is the critical ratio of $W1$ and $W2$ that determines the number of regions for entropy-evaluation and regional-convolutional matching. The area represented by each region in an image should be small enough to accommodate for viewpoint variation between adjacent regions and yet large enough for a suitable regional HOG-descriptor (i.e. each region should contain a reasonable number of intensity-change gradients). Fig. 9 (d) shows stable precision under a range of values for W1 and W2, given a constant value for $\frac{W1}{W2}$.

V. Conclusion

We presented a light-weight, compute-efficient and training-free VPR technique (namely CoHOG) based on Histogram-of-Oriented-Gradients (HOG) descriptor that achieves state-of-the-art performance under computational constraints on standard VPR datasets. By evaluating on both viewpoint and appearance variant datasets, the utility of our approach is discussed. We show highly precise place matching performance on viewpoint variant datasets, while comparable precision is achieved on condition variant dataset. With zero training requirements, lower encoding time and lesser run-time memory footprint than neural networks, CoHOG promises better deployability in real-world applications.

The technique presented in this work is agnostic in nature. Although, we have used HOG as our region descriptor but any feature descriptor can be plugged-in for robustness to viewpoint variations. Further exploration into condition invariant descriptors paves the path for future research into hand-crafted descriptors-based visual place recognition.

References

[1] S. Lowry *et al.*, "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2015.
[2] Z. Chen, L. Liu, I. Sa, Z. Ge, and M. Chli, "Learning context flexible attention model for long-term visual place recognition," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4015–4022, Oct. 2018.
[3] N. Sünderhauf *et al.*, "Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free," in *Proc. Robot. Sci. Syst. XII*, 2015.
[4] M. Zaffar, S. Ehsan, M. Milford, and K. M. Maier, "Memorable maps: A framework for re-defining places in visual place recognition," 2018, *arXiv:1811.03529*.
[5] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," in *Proc. Australas. Conf. Robot. Automat.*, vol. 2, 2014, p. 4.
[6] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," in *Proc. Robot. Sci. Syst. Conf.*, 2018.
[7] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, K. Alexis, and K. McDonald-Maier, "Are state-of-the-art visual place recognition techniques any good for aerial robotics?," *IEEE ICRA Workshop Aerial Robot.*, 2019.
[8] F. Maffra, L. Teixeira, Z. Chen, and M. Chli, "Real-time wide-baseline place recognition using depth completion," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1525–1532, Apr. 2019.
[9] C. McManus, B. Upcroft, and P. Newmann, "Scene signatures: Localised and point-less features for localisation," in *Proc. Robot., Sci. Syst. Conf.*, 2014.
[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[11] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vision*, pp. 404–417, 2006.
[12] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *Int. J. Robot. Res.*, vol. 21, no. 8, pp. 735–758, 2002.
[13] H. Andreasson and T. Duckett, "Topological localization for mobile robots using omni-directional vision and local features," *Int. Federation Accountants Proc. Volumes*, vol. 37, no. 8, pp. 36–41, 2004.
[14] E. Stumm, C. Mei, and S. Lacroix, "Probabilistic place recognition with covisibility maps," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 2013, pp. 4158–4163.
[15] J. Košecká, F. Li, and X. Yang, "Global localization and relative positioning based on scale-invariant keypoints," *Robot. Auton. Syst.*, vol. 52, no. 1, pp. 27–38, 2005.
[16] A. C. Murillo, J. J. Guerrero, and C. Sagues, "Surf features for efficient robot localization with omnidirectional images," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2007, pp. 3901–3907.
[17] M. Cummins and P. Newman, "Appearance-only slam at large scale with fab-map 2.0," *Int. J. Robot. Res.*, vol. 30, no. 9, pp. 1100–1123, 2011.
[18] W. Maddern, M. Milford, and G. Wyeth, "Cat-slam: Probabilistic localisation and mapping using a continuous appearance-based trajectory," *Int. J. Robot. Res.*, vol. 31, no. 4, pp. 429–451, 2012.

[19] M. Agrawal, K. Konolige, and M. R. Blas, "Censure: Center surround extremas for realtime feature detection and matching," in *Proc. Eur. Conf. Comput. Vision*, 2008, pp. 102–115.

[20] K. Konolige and M. Agrawal, "FrameSLAM: From bundle adjustment to real-time visual mapping," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 1066–1077, Oct. 2008.

[21] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. Eur. Conf. Comput. Vision*, 2006, pp. 430–443.

[22] C. Mei, G. Sibley, M. Cummins, P. Newman, and I. Reid, "A constant-time efficient stereo slam system," in *Proc. Brit. Mach. Vision Conf.*, 2009, vol. 1.

[23] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vision*, 2003, pp. 1470–1477.

[24] A. Angeli, S. Doncieux, J.-A. Meyer, and D. Filliat, "Incremental vision-based topological slam," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 2008, pp. 1031–1036.

[25] K. L. Ho and P. Newman, "Detecting loop closure with scene sequences," *Int. J. Comput. Vision*, vol. 74, no. 3, pp. 261–286, 2007.

[26] J. Wang, H. Zha, and R. Cipolla, "Combining interest points and edges for content-based image retrieval," in *Proc. IEEE Int. Conf. Image Process.*, 2005, vol. 3, pp. III–1256.

[27] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2007, pp. 3921–3926.

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

[29] A. C. Murillo and J. Kosecka, "Experiments in place recognition using gist panoramas," in *Proc. Int. Conf. Comput. Vision Workshops*, 2009, pp. 2196–2203.

[30] G. Singh and J. Kosecka, "Visual loop closing using gist descriptors in manhattan world," *ICRA Omnidirectional Vision Workshop*, 2010, pp. 4042–4047.

[31] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "Brief: Computing a local binary descriptor very fast," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1281–1298, 2011.

[32] N. Sünderhauf and P. Protzel, "Brief-gist-closing the loop by simple means," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 2011, pp. 1234–1241.

[33] H. Badino, D. Huber, and T. Kanade, "Real-time topometric localization," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2012, pp. 1635–1642.

[34] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2012, pp. 1643–1649.

[35] E. Pepperell, P. I. Corke, and M. J. Milford, "All-environment visual place recognition with smart," in *2Proc. IEEE Int. Conf. Robot. Automat.*, 2014, pp. 1612–1618.

[36] P. Sermanet *et al.*, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. 2nd Int. Conf. Learn. Representations, ICLR*, 2014.

[37] Z. Chen *et al.*, "Deep learning features at scale for visual place recognition," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 3223–3230.

[38] L. Liu, C. Shen, and A. van den Hengel, "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 4749–4757.

[39] A. Babenko and V. Lempitsky, "Aggregating deep convolutional features for image retrieval," 2015, *arXiv:1510.07493*.

[40] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 5297–5307.

[41] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2010, pp. 3304–3311.

[42] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," 2015, *arXiv:1511.05879*.

[43] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 2017, pp. 9–16.

[44] A. Khaliq, S. Ehsan, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight cnns for severe viewpoint and appearance changes," *IEEE Trans. Robot.*, pp. 1–9, 2019, doi: 10.1109/TRO.2019.2956352.

[45] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proc. Conf. Comput. Vision Pattern Recognit. Workshops*, 2018, pp. 224–236.

[46] M. Dusmanu *et al.*, "D2-net: A trainable CNN for joint description and detection of local features," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 8092–8101.

[47] L. G. Camara, C. Gäbert, and L. Preucil, "Highly robust visual place recognition through spatial matching of cnn features," *ResearchGate Preprint*, 2019.

[48] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 12716–12725.

[49] E. Stenborg, C. Toft, and L. Hammarstrand, "Long-term visual localization using semantically segmented images," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 6484–6490.

[50] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic visual localization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6896–6906.

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

[52] Y. Hou, H. Zhang, and S. Zhou, "Evaluation of object proposals and convnet features for landmark-based visual place recognition," *J. Intell. Robotic Syst.*, vol. 92, no. 3-4, pp. 505–520, 2018.

[53] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *Proc. Int. Workshop Autom. Face Gesture Recognit.*, 1995, vol. 12, pp. 296–301.

[54] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2005, vol. 1, pp. 886–893.

[55] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 2015, pp. 4297–4304.

[56] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 3234–3243.

[57] M. Larsson, E. Stenborg, L. Hammarstrand, M. Pollefeys, T. Sattler, and F. Kahl, "A cross-season correspondence dataset for robust semantic segmentation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 9532–9542.

[58] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, and K. McDonald-Maier, "Levelling the playing field: A comprehensive comparison of visual place recognition approaches under changing conditions," *IEEE ICRA Workshop Database Generation Benchmarking*, 2019.

[59] E. Olson and P. Agarwal, "Inference on networks of mixtures for robust robot mapping," *Int. J. Robot. Res.*, vol. 32, no. 7, pp. 826–840, 2013.

[60] N. Sünderhauf and P. Protzel, "Switchable constraints for robust pose graph slam," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 1879–1884.

[61] Y. Latif, C. Cadena, and J. Neira, "Robust loop closing over time for pose graph slam," *Int. J. Robot. Res.*, vol. 32, no. 14, pp. 1611–1626, 2013.

[62] P. Neubert, N. Sünderhauf, and P. Protzel, "Appearance change prediction for long-term navigation across seasons," in *Proc. IEEE Eur. Conf. Mobile Robots*, 2013, pp. 198–203.

[63] I. Kostavelis and A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey," *Rajasthan Administ. Service*, vol. 66, pp. 86–103, 2015.

[64] J. Surber, L. Teixeira, and M. Chli, "Robust visual-inertial localization with weak gps priors for repetitive uav flights," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 6300–6306.