Improving Visual Place Recognition Performance by Maximising Complementarity

Maria Waheed, Michael Milford, Klaus McDonald-Maier, and Shoaib Ehsan

***Abstract*—Visual place recognition (VPR) is the problem of recognising a previously visited location using visual information. Many attempts to improve the performance of VPR methods have been made in the literature. One approach that has received attention recently is the multi-process fusion where different VPR methods run in parallel and their outputs are combined in an effort to achieve better performance. The multi-process fusion, however, does not have a well-defined criterion for selecting and combining different VPR methods from a wide range of available options. To the best of our knowledge, this paper investigates the complementarity of state-of-the-art VPR methods systematically for the first time and identifies those combinations which can result in better performance. The letter presents a well-defined framework which acts as a sanity check to find the complementarity between two techniques by utilising a McNemar's test-like approach. The framework allows estimation of upper and lower complementarity bounds for the VPR techniques to be combined, along with an estimate of maximum VPR performance that may be achieved. Based on this framework, results are presented for eight state-of-the-art VPR methods on ten widely-used VPR datasets showing the potential of different combinations of techniques for achieving better performance.**

***Index Terms*—Visual place recognition, localization, navigation, complementarity, multi-process fusion.**

I. Introduction

VISUAL place recognition is a fundamental yet challenging task in the field of mobile robotics [1]. It may be defined as the ability of a robot to recognize a previously visited location. Viewpoint changes [2], [3], seasonal variations [4], [5], presence of dynamic objects [6], [7] and illumination changes [8], [9] encountered in real world scenarios make this apparently simple task non-trivial [4], [10]. Several techniques have been presented to solve this problem (such as [11]–[14]), however, every VPR method has its own pros and cons [15]–[18], and there is no universal technique that may be used in all conditions and scenarios.

Recently, a new approach named multi-process fusion has been introduced that combines several image processing methods and negates the requirement of multiple sensors to improve VPR performance [19], [20]. The concept comes from the empirical data which suggests that some VPR methods are more suitable for certain types of environments and scenarios than others [10]. Hence, utilising multiple VPR techniques simultaneously may compensate for each other's weaknesses. Although the systems presented in [19], [20] exhibit promising results, they do not provide a well-defined criterion for selection of VPR techniques based on complementarity out of the available options. Supposing that the fused VPR methods will complement each other in all cases is not a valid assumption and may have detrimental effects on performance and computation. For example, if the VPR techniques that are combined are redundant, they will not achieve higher performance and will only add to the computational cost which may not be suitable for resource-constrained systems. Hence, complementarity information is vital and can enable a multi-process fusion based system to make an informed decision regarding selection of VPR techniques from available options.

To the best of our knowledge, complementarity of VPR methods has not been studied systematically so far. Through this paper, we attempt to bridge this gap and intend to design a framework that can be used as a sanity check for the selection of complementary pairs of VPR techniques for multi-process fusion systems. Our proposed framework is based on a McNemar's test-like approach [21], [22] that categorizes each VPR outcome from a technique as either success or failure (considering ground truth information). The framework allows estimation of upper and lower complementarity bounds for the VPR techniques to be combined, along with an estimate of maximum VPR performance that may be achieved. This framework is then employed for eight state-of-the-art VPR methods to identify highly complementary pairs on widely used VPR datasets.

The rest of this paper is organized as follows. Section II provides an overview of related work. Section III presents the framework for computing complementarity between VPR techniques, and for estimating upper and lower complementarity bounds along with an assessment of maximum achievable VPR performance. Section IV describes the experimental setup. The results based on the proposed framework are presented in Section V. Finally, conclusions are given in Section VI.

Manuscript received December 13, 2020; accepted June 2, 2021. Date of publication June 17, 2021; date of current version June 29, 2021. This letter was recommended for publication by Associate Editor F. Ramos and Editor S. Behnke upon evaluation of the reviewers' comments. This work was supported by the U.K. Engineering and Physical Sciences Research Council through Grants EP/R02572X/1, EP/P017487/1, and in part by the RICE project funded by the National Centre for Nuclear Robotics Flexible Partnership Fund. Michael Milford is partially supported by the QUT Centre for Robotics. *(Corresponding author: Shoaib Ehsan.)*

Maria Waheed, Klaus McDonald-Maier, and Shoaib Ehsan are with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, United Kingdom (e-mail: mw20987@essex.ac.uk; kdm@essex.ac.uk; sehsan@essex.ac.uk).

Michael 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.2021.3088779

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

II. Related Work

This section provides an overview of the related work in the domain of visual place recognition. The methods used for VPR may be divided into three categories: handcrafted feature descriptor-based techniques, deep-learning-based methods, and Region-of-interest-based approaches. All these categories have their own strengths and weaknesses that influence the selection of any methods from among them. Some state-of-the-art handcrafted feature descriptors used for VPR are Scale Invariant Feature Transform (SIFT) [23], Speeded-Up Robust Features (SURF) [24], and GIST [25]. Convolutional Neural Networks (CNNs) have turned out to be revolutionary in the field of VPR and provide significant improvement in performance [26] even under extreme environmental variations. Some of the widely used techniques include NetVLAD [11], AMOSNet [12], and HybridNet [12]. Region-of-interest-based VPR techniques make use of the static and definite regions of images to perform place recognition, such as Regions of Maximum Activated Convolutions (R-MAC) [27].

Fusing multiple sensors to improve place recognition performance has been the focus of several research works [28]–[30]. Although multi-sensor approaches help boost performance, they do carry certain disadvantages, such as expensive and bulky sensors, and potential significant increase in computation. To overcome these shortcomings, the concept of fusing multiple VPR techniques has gained popularity. The authors of [31] combined multiple image processing methods into a merged feature vector using a convex optimization approach to decide the best match from the sequence of images generated. The effort did generate some promising results over multiple datasets but had limited overall performance due to the absence of sequential information. Similarly, a multi-process fusion system is introduced in [19] which combines multiple VPR methods using a Hidden Markov Model (HMM) to identify the optimal estimated location over a sequence of images. The authors of [20] have presented a three-tier hierarchical multi-process fusion system which is customizable and may be extended to any arbitrary number of tiers. A different place recognition method is used in each tier to compare the query image with the provided sequence of images.

III. Proposed Framework

This section presents the framework for computing complementarity, for establishing the upper and lower complementarity bounds, and for estimating the maximum achievable VPR performance by a muti-process fusion system. This framework may be employed on an arbitrary number of VPR methods to determine the optimal pairing from among the pool of techniques available. It may also be utilised as a sanity check on whether the VPR techniques that a multi-process fusion system has assembled for integration are even viable. The framework employs a McNemar's test like approach to perform a case-by-case analysis of each VPR technique to compute the complementarity of the given technique with other available methods.

Precision-recall curves, F-scores and accuracy percentage [9] are usually utilised as performance metrics for VPR methods. Although viable for some applications/scenarios, these performance metrics do not provide the specific information that tells where exactly does a VPR method succeeds or fails, and do not show the whole picture. For example, two VPR methods compared over a dataset of 100 images using these performance metrics may appear to have same performance if they both are able to match 70 images (out of 100). However, it is highly likely that the set of 70 images successfully matched by the first VPR method is not the same set that is also correctly matched by the second VPR technique. We believe that this neglected piece of information is critical for determining complementarity of different VPR methods, and is vital knowledge to have specifically when dealing with multi-process fusion systems.

Model stacking [33], a method for combining multiple predictors into one through ensemble learning, holds some similarity to this new proposed approach. Model stacking works based on combining heterogeneous weak learners and aims to capture distinct regions in the data where each model performs the best which is somewhat similar to our approach in the sense that we also target to combine heterogeneous VPR techniques through the proposed framework. However, as opposed to model stacking, the use of the McNemar's test-like approach here allows a case-by-case analysis of the data while avoiding having to divide the training set into several pieces like you would do in k-folds cross validation. Further, our proposed framework focuses on a pair wise approach and allows the identification of pairs with high complementarity likelihood with a quantitative value generated for comparing their compatibility.

McNemar's test is a form of chi-squared test with one degree of freedom that evaluates the performance of two algorithms based on their outcomes on a case-by-case basis over the same dataset. For utilizing McNemar's test, a criterion is needed to determine whether a test case results in success or failure. Our proposed framework is loosely inspired by the McNemar's test as we do pairwise analysis of VPR methods on a case-by-case basis over the same dataset. The two VPR methods in question would produce results in the form of correct or incorrect matches verified using ground truth. This data may then be divided into four possible cases as shown in Fig. 2: first being the number of images where both algorithms are able to match the images correctly, second where the first algorithm matched correctly while the second produced an incorrect match, then vice versa and finally where both algorithms failed and produced incorrect matches. For computing complementarity, our prime focus remains on case two and three as these hold the number of images where the two algorithms perform differently and can help boost each other's performance.

Computing complementarity. Let A be our primary VPR technique. Let B be a VPR method that may be combined with A in a multi-process fusion system to enhance VPR performance over an image dataset D. VPR performance is defined as the ratio of number of images of D that are correctly matched (verified by groundtruth) to the total number of images of D. The complementarity is calculated by the following equation:

$$CBA = \frac{T}{M} \tag{1}$$

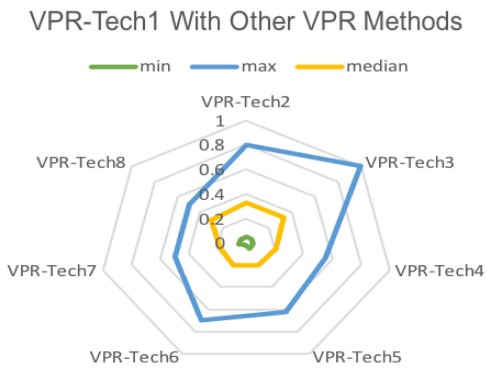

Fig. 1. Sample output of the proposed complementarity framework. Here, VPR-Tech1 is the primary VPR technique which may be combined with other available secondary VPR methods (VPR-Tech2, VPR-Tech3 etc). The green line (**min**) shows the lower complementarity bound of VPR-Tech1 with other methods; the blue line (**max**) depicts the maximum complementarity bound; the yellow line (**median**) shows the median complementarity bound.

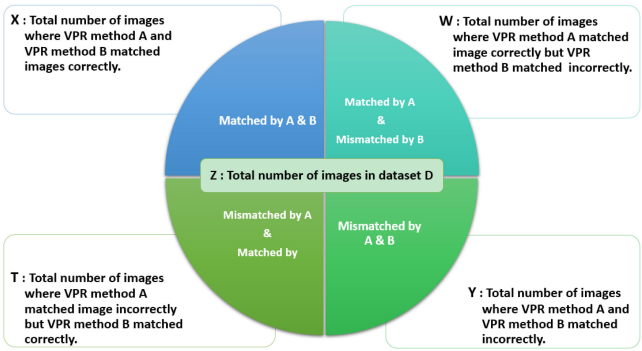

Fig. 2. Possible outcomes of pairwise analysis of VPR methods on a case-by-case basis over the same dataset.

Where *CBA* is the complementarity of *B* with *A*; *T* is the number of images of *D* which are incorrectly matched by *A* but correctly matched by *B* when the two methods are run; *Y* is the number of images of *D* that are incorrectly matched by *A* and *B* when the two methods are run; *M* is the summation of *T* and *Y*, and thus is the total number of images of *D* that are incorrectly matched by *A* when run. A large value of *CBA* implies that *B* complements *A* well on dataset *D* and will result in potential increase in VPR performance. On the other hand, a small value of *CBA* means that *B* does not complement *A* well. In other words, *A* and *B* are redundant and combining *A* with *B* will increase computational cost without any substantial increase in VPR performance.

Establishing complementarity bounds. It is interesting to further explore the upper and lower extremities of complementarity of *B* with *A*. Let *K* be the set of *n* individual datasets on which *A* and *B* are run.

$$K = \{D_1, D_2, D_3, ...D_n\} \tag{2}$$

Let *J* be the set of complementarity scores (*B* with *A*) computed over *n* dataset in *K*.

$$J = \{CBA_1, CBA_2, CBA_3...CBA_n\} \tag{3}$$

The upper complementarity bound is then established as

$$U = max\{CBA_1, CBA_2, CBA_3...CBA_n\} \tag{4}$$

The lower complementarity bound is estimated as

$$L = min\{CBA_1, CBA_2, CBA_3...CBA_n\} \tag{5}$$

The median of complementarity of *B* with *A* is computed as

$$Q = median\{CBA_1, CBA_2, CBA_3...CBA_n\} \tag{6}$$

Estimating maximum achievable performance. It is beneficial to estimate the maximum achievable VPR performance of a multi-process fusion system over a dataset at an early stage. This is estimated as follows:

$$MAPE = \frac{(T + W + X)}{Z} \tag{7}$$

Where *MAPE* is the maximum achievable VPR performance estimate for the multi-process fusion system over a dataset *D*; *T* is the number of images of *D* which are incorrectly matched by *A* but correctly matched by *B* when the two methods are run; *W* is the number of images of *D* which are correctly matched by *A* but incorrectly matched by *B* when the two methods are run; *X* is the number of images of *D* which are correctly matched by both *A* and *B* when the two methods are run; *Z* is the total number of images of *D*.

IV. Experimental Setup

This section provides details of the experimental setup used for obtaining results by utilising the proposed framework. Table I lists the widely used VPR datasets [31] that are used for our experiments, namely GardensPoint, 24/7 Query [34], Essex3in1 [35], SPEDTest, Cross-Seasons [36], Synthia [37], Corridor, 17-Places, Living room, and Nordland [38]. The implementation details of the eight state-of-the-art VPR techniques that are utilised in the experiments are given below.

AlexNet: The use of AlexNet for VPR was studied by [41], who suggested that *conv3* is the most robust to conditional variations. Gaussian random projections are used to encode the activation-maps from conv3 into feature descriptors. Our implementation of AlexNet is similar to the one employed by [42].

NetVLAD: The original implementation of NetVLAD was in MATLAB, as released by [11]. The Python part of this code was open-sourced by [39]. The model selected for evaluation is VGG-16, which has been trained in an end-to-end manner on Pittsburgh 30 K dataset [11] with a dictionary size of 64 while performing whitening on the final descriptors.

AMOSNet: This technique was proposed by [12], where a CNN was trained from scratch on the SPED dataset. The authors presented results from different convolutional layers by implementing spatial pyramidal pooling on the respective layers. While the original implementation is not fully open-sourced, the trained model weights are shared by authors.

HybridNet: While AMOSNet was trained from scratch, [12] took inspiration from transfer learning for HybridNet and re-trained the weights initialised from Top-5 convolutional layers

TABLE I
VPR-BENCH DATASETS [32]

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

of CaffeNet [40] on SPED dataset. We have implemented HybridNet using 'conv5' of the shared HybridNet model.

RegionVLAD: This technique is introduced and open-sourced by [14]. We have used AlexNet (trained on Places365 dataset) as the underlying CNN. The total number of regions of interest is set to 400, and we have used 'conv3' for feature extraction. The dictionary size is set to 256 visual words for VLAD retrieval. Cosine similarity is subsequently used for matching descriptors of query and reference images

CALC: The use of convolutional auto-encoders for VPR was proposed by [13], where an auto-encoder network was trained in an unsupervised manner to re-create similar HOG descriptors for viewpoint variant (cropped) images of the same place. We use model parameters from 100 000 training iteration. Cosine-matching is used for descriptor comparison.

HoG: Histogram-of-oriented-gradients (HoG) is one of the most widely used handcrafted feature descriptor, which actually performs very well for VPR compared to other handcrafted feature descriptors. We use a cell size of 16×16 and a block size of 32×32 for an image-size of 512×512 for our implementation. The total number of histogram bins are set equal to 9. We use cosine-matching between HOG-descriptors of various images to find the best place match.

CoHOG: It is a recently proposed handcrafted feature descriptor-based technique, which uses image-entropy for region-of-interest extraction. The regions are subsequently described by dedicated HoG descriptors and these regional descriptors are convolutionally matched to achieve lateral viewpoint-invariance. It is an opensource technique and we have used an image size of 512×512, cell size of 16×16, bin-size of 8 and an entropy-threshold (ET) of 0.4. CoHOG [43] also uses cosine-matching for descriptor comparison.

V. RESULTS AND DISCUSSION

This section presents the results obtained by utilizing the proposed framework. Fig. 3 and Fig. 4 depict the complementarity scores of different VPR methods with other techniques on various standard datasets that contain two major types of variation, namely conditional and viewpoint. Analysing the results from the point of view of these variations shows an interesting performance pattern and helps identify the best VPR combinations for certain types of changes.The datasets that consist of the conditional variation of day and night changes include GardenPoint, 24/7, 17Places and LivingRoom. Of all the VPR combinations tested over these datasets, the highest complementarity scores belong to pairs consisting of either NetVLAD or RegionVLAD. This pattern can be observed consistently over all the combinations as shown in Fig. 3 and Fig. 4. Hence, it may be concluded that for environments that encounter day and night changes, the best option for VPR are the pairs formed with either NetVLAD or RegionVLAD.

Essex3in1 is the only dataset that mainly deals with illumination changes. CoHOG stands out as the best complementary secondary technique for pairing throughout. Interestingly, pairs formed with CoHOG do not show significantly high complementarity scores over any other datasets except the one with illumination changes. This is a useful piece of information to have when this type of variation can be anticipated in the environment. For seasonal and weather changes that are encountered in SPED, Synthia and Nordland datasets, the VPR pairs with the best complementarity vary between HybridNet, AMOSNet and NetVLAD. Hence when dealing with a dataset where the variation appears to be seasonal, the best pairs for VPR to consider would be from among the above three.The last type of seasonal variation which is dawn-dusk changes is encountered in the Cross-Seasons dataset. The pairs scoring the highest complementarity values in most cases consist of HybridNet or NetVLAD as the secondary VPR technique. These results show that it is sensible to use VPR pairs consisting of HybridNet or NetVLAD when dealing with dawn-dusk variations. Lateral variation in viewpoint is present in several datasets including GardenPoint, Cross-Seasons, Synthia, Corridor, 17Places and Livingroom. A high complementarity score for these datasets can be consistently seen in Fig. 3 and Fig. 4 for VPR pairs containing either NetVLAD, RegionVLAD or AMOSNet. On the other hand, Essex3in1 and 24/7 which comprise of a 6-DOF variation only show high complementarity score between pairs of NetVLAD or CoHOG as evident from Fig. 3 and Fig. 4.

Fig. 5 depicts the lower and upper bounds of complementarity for the different VPR combinations that are discussed above. This allows us to determine the minimum and maximum complementarity a certain pair can have given any type of environment. This information is beneficial for circumstances when the environment or dataset to be used is unknown in which case selection of the pair with the highest lower complementarity bound and highest upper complementarity bound would be the best option. Starting from the pairs of AlexNet, the highest upper

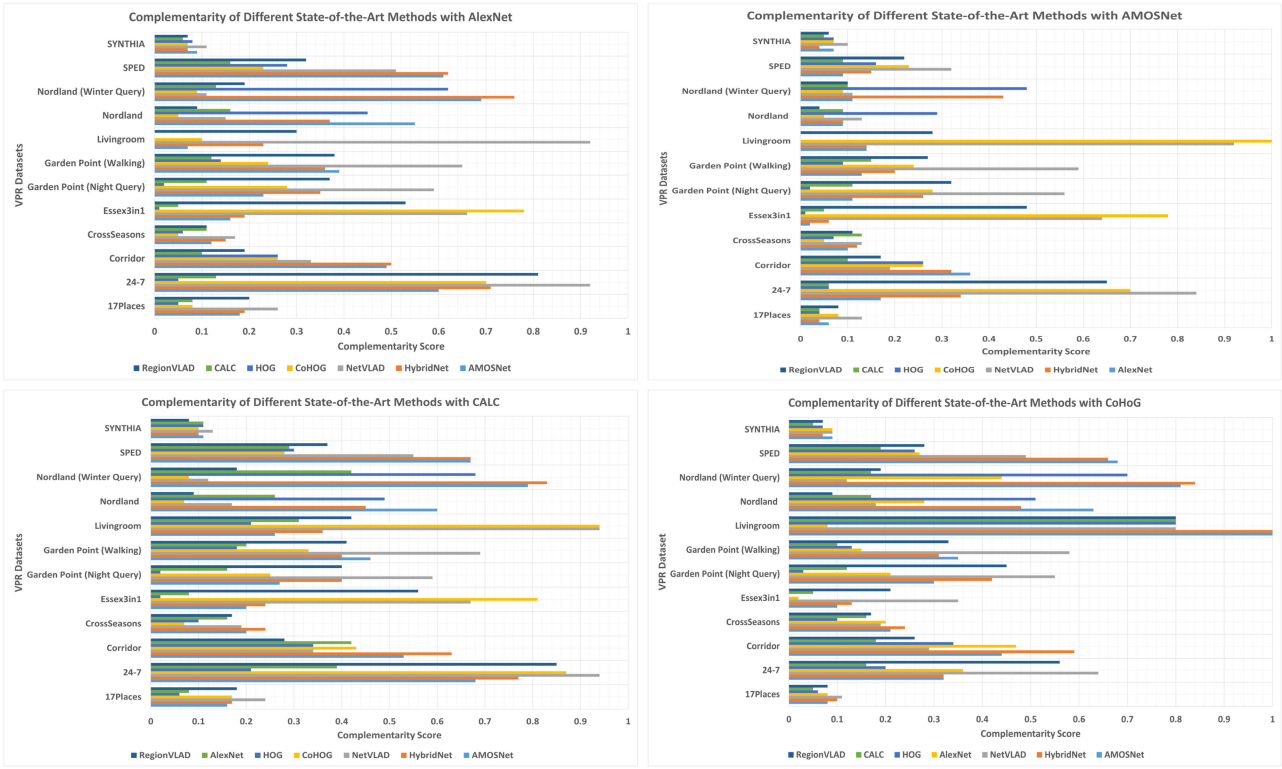

Fig. 3. Complementarity of state-of-the-art VPR methods with: AlexNet (top left); AMOSNet (top right); CALC (bottom left); CoHOG (bottom right).

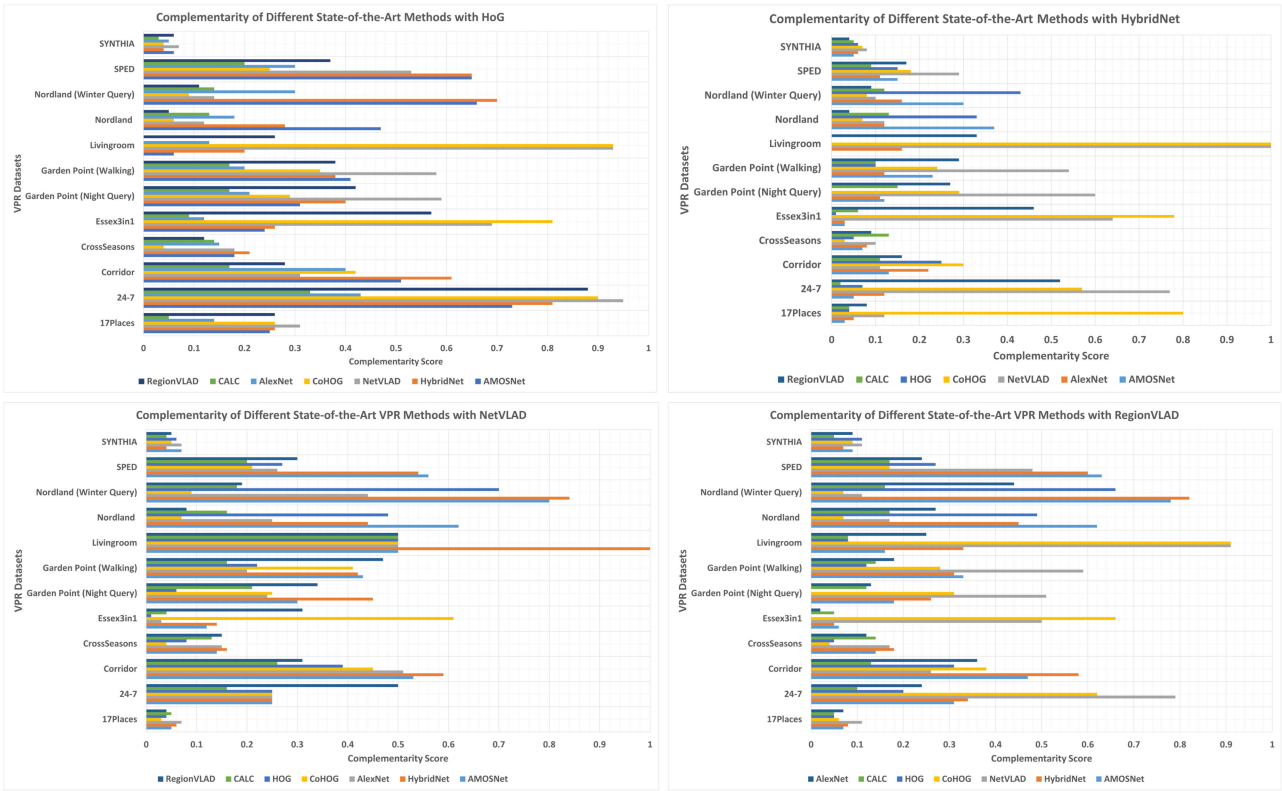

Fig. 4. Complementarity of state-of-the-art VPR methods with: HoG (top left); HybridNet (top right); NetVLAD (bottom left); RegionVLAD (bottom right).

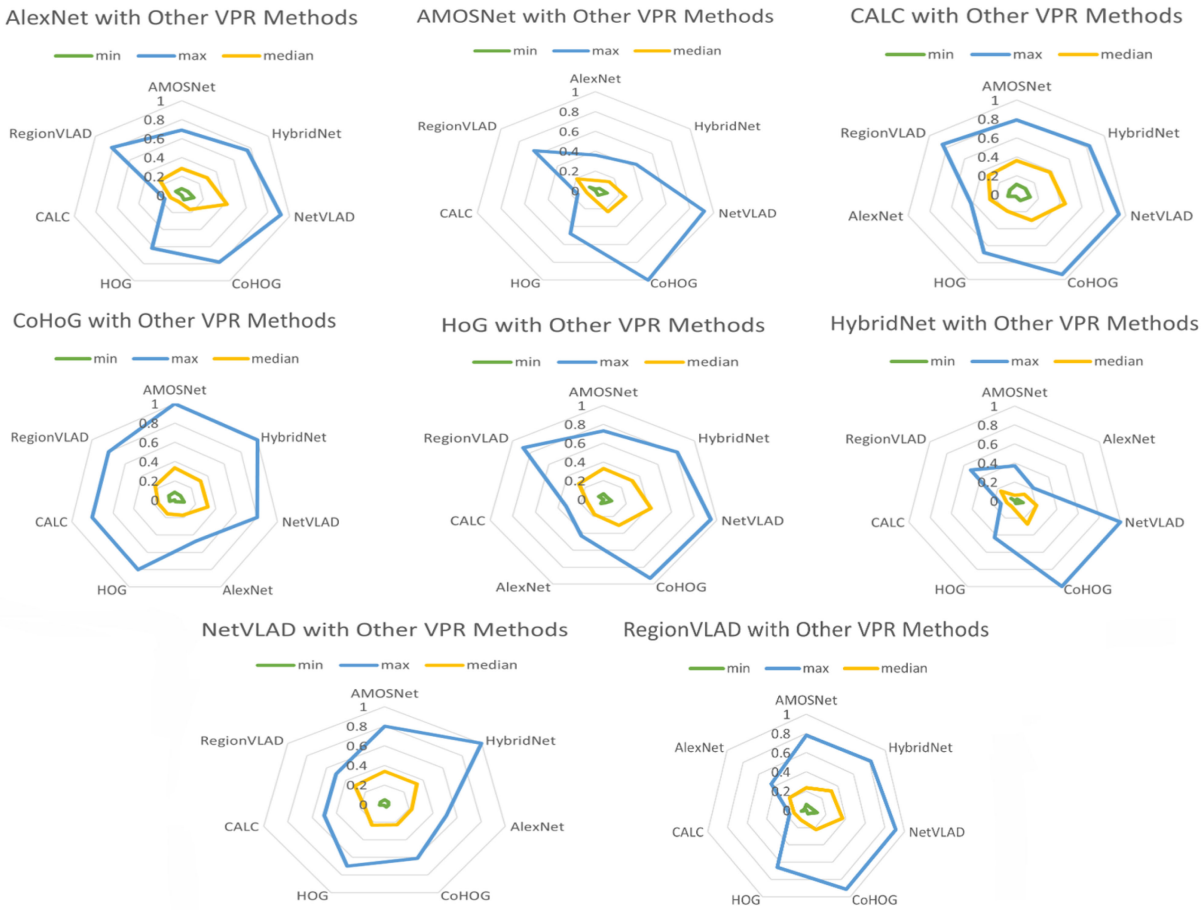

Fig. 5. Max (upper bound), Min (lower bound), and Median complementarity of state-of-the-art VPR methods with: AlexNet (top left); AMOSNet (top centre); CALC (top right); CoHOG (middle left); HoG (middle centre) HybridNet (middle right); NetVLAD (bottom left); RegionVLAD (bottom right).

TABLE II
PERFORMANCE OF DIFFERENT STATE-OF-THE-ART VPR METHODS ON STANDARD DATASETS
(IN PERCENTAGE)

VPR Techniques	Garden Point	24-7	Essex3in1	SPED	Cross-Seasons	SYNTHIA	Nordland	Corridor	17Places	Livingroom
AlexNet	23.1	67.6	14.3	51.6	23.1	25.4	46.6	49.1	30.1	61.2
AMOSNet	33.1	84.7	26.3	79.5	25.2	26.7	81.5	59.1	39.2	58.1
CALC	19.1	53.7	11.4	42.6	18.9	21.7	20.1	20.9	30.3	41.9
CoHOG	39.2	93.5	82.7	49.5	10.0	25.6	10.8	45.4	39.0	87.1
HoG	4.52	45.4	3.82	50.0	15.2	27.8	70.9	37.2	22.4	54.8
HybridNet	43.7	89.5	28.7	79.5	29.4	26.3	85.1	68.1	40.2	64.5
NetVLAD	58.7	97.1	70.3	67.9	24.2	29.2	14.1	29.1	44.2	96.7
RegionVLAD	44.7	92.5	59.3	56.7	22.1	24.2	22.6	34.5	40	64.5

and lower bounds of complementarity is achieved by the pairs of NetVLAD, and then HybridNet. The VPR technique that needs to be avoided for pairing with AlexNet under unknown type of variation is CALC. For VPR pairs that can be formed with AMOSNet, the best option to consider is CoHOG while the second best is NetVLAD. Methods that should be avoided when pairing with AMOSNet are CALC, HoG and AlexNet. For CALC and its pairs, all available options appear to be viable while NetVLAD is slightly better. However, the only pairing to avoid in this case would be with AlexNet. An overall view would suggest that NetVLAD and HybridNet seem to be the feasible option in most cases for VPR pairing while CALC should be the least preferred option.

Table II presents the performance results of state-of-the-art single VPR techniques on standard datasets. The purpose of this table is to provide a clear comparison between the performance of a single VPR technique and the maximum achievable VPR performance values for 28 different combinations of state-of-the-art VPR methods utilizing the proposed framework (presented in Table III).

It is evident that each combination has varying *MAPE* values over each dataset. The highest *MAPE* value by a VPR combination for each dataset is highlighted in Table III. It is interesting to note that all the VPR pairs identified for the highest *MAPE* value for a certain dataset are from among the pairs that were identified with having the highest complementarity for the same

TABLE III
MAXIMUM ACHIEVABLE PERFORMANCE ESTIMATE FOR DIFFERENT COMBINATIONS OF STATE-OF-THE-ART VPR METHODS ON STANDARD DATASETS (IN PERCENTAGE)

VPR Combinations	Garden Point	24-7	Essex3in1	SPED	Cross-Seasons	SYNTHIA	Nordland	Corridor	17Places	Livingroom
AlexNet + AMOSNet	54.0	87.2	28.0	81.3	32.9	32.2	83.6	73.8	43.1	62.5
AlexNet + CALC	34.0	72.0	19.0	59.6	31.9	30.62	53.8	54.0	36.2	59.3
AlexNet + CoHOG	48.5	95.7	82.8	63.4	28.7	32.7	50.8	71.17	44.0	96.8
AlexNet + HoG	36.0	69.3	15.7	65.2	28.2	32.1	79.9	62.1	33.7	59.3
AlexNet + HybridNet	52.0	90.6	30.1	81.7	35.0	31.3	87.4	74.7	43.3	68.7
AlexNet + NetVLAD	65.5	97.6	70.9	76.2	**36.1**	**34.2**	52.7	65.7	**48.2**	96.8
AlexNet + RegionVLAD	53.5	94.1	60	67.2	31.9	31.2	57.1	58.5	44.5	71.8
AMOSNet + CALC	55.5	85.6	30.0	81.3	35.0	30.4	83.4	63.0	41.8	56.2
AMOSNet + CoHOG	60.5	95.4	84.2	84.1	29.3	32.5	83.1	69.3	44.0	**100**
AMOSNet + HoG	52.5	85.6	27.1	82.8	30.8	32.5	90.3	69.3	42.11	56.2
AMOSNet + HybirdNet	57.9	89.8	30.9	82.5	34.5	29.8	89.5	72.0	42.11	62.5
AMOSNet + NetVLAD	75.5	97.6	73.8	**86.1**	35.0	34.1	83.6	66.6	47.2	96.8
AMOSNet + RegionVLAD	62.0	94.6	61.9	84.0	33.5	31.2	83.4	65.7	44.5	68.7
CALC + CoHOG	45.5	94.3	83.3	59.1	25.1	29.6	26.6	54.9	42.3	96.8
CALC +HoG	33.0	63.4	13.3	60.1	27.2	30.6	75.0	47.7	34.4	53.1
CALC + HybridNet	45.5	94.3	83.3	59.1	25.1	29.6	26.6	54.9	42.3	96.8
CALC + NetVLAD	63.5	97.3	71.4	74.4	34.5	32.1	30.0	47.7	47.0	96.8
CALC + RegionVLAD	51.5	93.0	61.4	64.4	32.9	28.4	35.1	43.2	43.3	65.6
CoHOG + HoG	47.5	94.6	82.3	62.9	19.3	31.2	73.5	63.9	42.8	96.8
CoHOG + HybridNet	58.5	95.4	84.7	83.1	31.9	31.4	86.2	**77.4**	45.0	**100**
CoHOG + NetVLAD	74.5	97.6	**88.5**	74.6	27.2	32.9	22.3	61.2	45.8	96.8
CoHOG + RegionVLAD	59.5	97.0	86.1	64.0	25.6	31.5	28.4	59.4	44.0	96.8
HoG + HybridNet	50.5	90.1	29.5	82.5	33.5	31.3	**91.4**	75.6	42.6	62.5
HoG + NetVLAD	74.5	97.6	**88.5**	74.6	27.2	32.5	22.3	61.2	45.8	96.8
HoG + RegionVLAD	50.0	93.8	59.0	68.6	26.1	32.5	74.3	54.9	43.3	65.6
HybridNet + NetVLAD	61.0	94.9	61.4	83.0	**36.1**	29.7	86.4	72.9	45.0	75.0
HybridNet + RegionVLAD	61.0	94.9	61.4	83.0	**36.1**	29.7	86.4	72.9	45.0	75.0
NetVLAD + RegionVLAD	**77.0**	**98.4**	79.5	77.5	35.6	33.1	31.3	51.3	46.7	96.8

dataset. All other VPR pairs with high *MAPE* values are also the ones with high complementarity scores as depicted in Fig. 3 and Fig. 4. This shows that higher complementarity scores may result in potentially higher performance. For the 17Places dataset, the highest *MAPE* value is scored by AlexNet and NetVLAD, while it is already shown above that over this dataset, NetVLAD has the highest complementarity combined with any other VPR technique. The 24/7 dataset shows high *MAPE* scores with all pairs containing NetVLAD or RegionVLAD while the highest scoring pair being NetVLAD and RegionVLAD itself. The Corridor dataset has a varying range of *MAPE* values that also reflects on the fact that it has extremely varying complementarity scores as well. However, we can observe that it does have better *MAPE* scores for all pairs containing HybridNet. The Cross-Seasons dataset presents similar *MAPE* scores for three different VPR pairs but all contain HybridNet and NetVLAD which are also the most highly complementary pairs identified above. Throughout the remaining presented results in Table III, it is evident that where the *MAPE* scores are very high, these are the pairs with the highest complementarity. On the other hand, where *MAPE* scores are relatively lower, these are the pairs with lower complementarity scores. It is interesting to point out that the higher the *MAPE* scores for a VPR pair, the better it is to be used as a combined system of VPR techniques rather individually for the given dataset.

VI. CONCLUSION

This letter has proposed a well-defined framework for determining the viability of combining different VPR methods for a multi-process fusion system. The complementarity information computed through the proposed framework helps to select the best possible combination of VPR techniques to ensure performance improvement in fused systems. The results obtained utilising the presented framework for eight state-of-the-art VPR methods over ten widely-used VPR datasets provide new insights regarding complementarity of various VPR methods and estimate their maximum performance. This paper has considered only pairs of VPR techniques. A promising future direction is to investigate extension to a combination of three or more VPR techniques.

REFERENCES

[1] S. Lowry *et al.*, "Visual place recognition: A. survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.
[2] S. Garg, N. Suenderhauf, and M. Milford, "Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics," in *Proc. Robotics: Sci. Syst.*, Jun. 2018
[3] F. Maffra, Z. Chen, and M. Chli, "Viewpoint-tolerant place recognition combining 2D and 3D information for UAV navigation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2018, pp. 2542–2549.
[4] 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. Automat.*, 2012, pp. 1643–1649.
[5] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 2564–2570.
[6] M. Bürki *et al.*, "VIZARD: Reliable visual localization for autonomous vehicles in urban outdoor environments," in *Proc. IEEE Int. Veh. Symp.*, vol. 4, 2019, pp. 1124–1130.
[7] C.-C. Wang, C. Thorpe, S. Thrun, M. Hebert, and H. Durrant-Whyte, "Simultaneous localization, mapping and moving object tracking," *Int. J. Robot. Res.*, vol. 26, no. 9, pp. 889–916, 2007.
[8] M. Milford *et al.*, "Sequence searching with deep-learnt depth for condition-and viewpoint-invariant route-based place recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, 2015, pp. 18–25
[9] A. Ranganathan, S. Matsumoto, and D. Ilstrup, "Towards illumination invariance for visual localization," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 3791–3798.

[10] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging SeqSLAM on a 3000 km journey across all four seasons," in *Proc. IEEE Int. Conf. Robot. Autom. Workshop Long-Term Autonomy*, 2013, Art no. 2013.

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

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

[13] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," in *Proc. Robot., Sci. Syst.*, Pittsburgh, PA, USA, 2018.

[14] A. Khaliq, S. Ehsan, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight CNNs for significant viewpoint and appearance changes," *IEEE Trans. Robot.*, vol. 36, no. 2, pp. 561–569, Apr. 2020.

[15] 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," in *Proc. IEEE Int. Conf. Robot. Automat. Workshop Database Gener. Benchmarking SLAM Algorithms Robot. VR/AR*, 2019.

[16] B. Ferrarini, M. Waheed, S. Waheed, S. Ehsan, M. Milford, and K. D. McDonald-Maier, "Visual place recognition for aerial robotics: Exploring accuracy-computation trade-off for local image descriptors," in *Proc. NASA/ESA Conf. Adaptive Hardware Syst.*, 2019, pp. 103–108.

[17] 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.

[18] N. V. Shirahatti and K. Barnard, "Evaluating image retrieval," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2005, vol. 1, pp. 955–961.

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

[20] S. Hausler and M. Milford, "Hierarchical multi-process fusion for visual place recognition," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 3327–3333.

[21] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[22] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical Methods for Rates and Proportions*. Hoboken, NJ, USA: Wiley, 2013.

[23] D. G. Lowe, "Distinctive image features from scale invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[24] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *Proc. IEEE Eur. Conf. Comput. Vis.*, Berlin, Heidelberg: Springer, 2006, pp. 404–417.

[25] A. C. Murillo and J. Kosecka, "Experiments in place recognition using gist panoramas," in *Proc. IEEE 12th Int. Conf. Comput. Vision Workshops*, 2009, pp. 2196–2203.

[26] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutionalneural network-based place recognition," *ACRA*, 2014.

[27] G. Tolias, R. Sicre, and H. Jégou, Particular object retrieval with integral max-pooling of CNN activations, in *Proc. Int. Conf. Learn. Representations*, 2016.

[28] A. Jacobson, Z. Chen, and M. Milford, "Autonomous multi sensor calibration and closed-loop fusion for slam," *J. Field Robot*, vol. 32, no. 1, pp. 85–122, 2015.

[29] A. Jacobson, Z. Chen, and M. Milford, "Leveraging variable sensor spatial acuity with a homogeneous, multi-scale place recognition framework," *Biol. Cybern.*, 2018, pp. 1–17.

[30] J. Collier, S. Se, and V. Kotamraju, "Multi-sensor appearance-based place recognition," in *Proc. Comput. Robot Vis.*, 2013, pp. 128–135.

[31] H. Zhang, F. Han, and H. Wang, "Robust multimodal sequence-based loop closure detection via structured sparsity," in *Robot.: Sci. Syst.*, 2016.

[32] M. Zaffar *et al.*, "VPR-Bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change.," *Int. J. Comput. Vis*., vol. 129, pp. 2136–2174, 2021.

[33] R. Odegua, "An empirical study of ensemble techniques (bagging, boosting and stacking)," in *Proc. Conf.: Deep Learn. IndabaXAt*, 2019.

[34] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 1808–1817.

[35] M. Zaffar, S. Ehsan, M. Milford, and K. D. McDonald-Maier,"Memorable maps: A. framework for re-defining places in visual place recognition," *IEEE trans Intell Transp Syst*, pp. 1–15, 2020, doi: 10.1109/TITS.2020.3001228.

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

[37] 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.

[38] S. Skrede, "Nordland dataset," 2013. [Online]. Available: https://bit.ly/2QVBOym

[39] T. Cieslewski, S. Choudhary, and D. Scaramuzza, "Data-efficient decentralized visual slam," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 2466–2473.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NeurIPS*, 2012, pp. 1097–1105.

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

[42] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," in *Proc. Robot.: Sci. Syst.*, 2018.

[43] M. Zaffar, S. Ehsan, M. Milford, and K. McDonald-Maier, "*CoHOG: A. light-weight*, compute-efficient, and training-free visual place recognition technique for changing environments," *IEEE Robot. Autom. Lett*, vol. 5, no. 2, pp. 1835–1842, Apr. 2020.