IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Aggregating Multiple Bio-Inspired Image Region Classifiers For Effective And Lightweight Visual Place Recognition

Bruno Arcanjo[1], Bruno Ferrarini[1], Maria Fasli[1], Michael Milford[2], Klaus D. McDonald-Maier[1] and Shoaib Ehsan[1,3]

***Abstract*— Visual place recognition (VPR) enables autonomous systems to localize themselves within an environment using image information. While VPR techniques built upon a Convolutional Neural Network (CNN) backbone dominate state-of-the-art VPR performance, their high computational requirements make them unsuitable for platforms equipped with low-end hardware. Recently, a lightweight VPR system based on multiple bio-inspired classifiers, dubbed DrosoNets, has been proposed, achieving great computational efficiency at the cost of reduced absolute place retrieval performance. In this work, we propose a novel multi-DrosoNet localization system, dubbed RegionDrosoNet, with significantly improved VPR performance, while preserving a low-computational profile. Our approach relies on specializing distinct groups of DrosoNets on differently sliced partitions of the original images, increasing model differentiation. Furthermore, we introduce a novel voting module to combine the outputs of all DrosoNets into the final place prediction which considers multiple top reference candidates from each DrosoNet. RegionDrosoNet outperforms other lightweight VPR techniques when dealing with both appearance changes and viewpoint variations. Moreover, it competes with computationally expensive methods on some benchmark datasets at a small fraction of their online inference time.**

I. Introduction

Visual place recognition (VPR) is an essential component of mobile robotics, as it allows the system to localize itself in the runtime environment using only image data [1]. The affordability and variety of camera sensors makes VPR localization particularly attractive for hardware restricted robotic platforms, which are common in mobile robotics [2]. Nevertheless, VPR is a complicated task and proposed solutions must deal with several visual challenges. The same place can appear vastly different when visited under different illumination [3], seasonal weather conditions [4], viewpoints [5] and dynamic elements entering and leaving the scene [6].

This work was supported by the UK Engineering and Physical Sciences Research Council through grants EP/Y009800/1, EP/X015955/1, EP/V000462/1, and by the Business and Local Government Data Research Centre BLG DRC through grant ES/S007156/1.

[1]B. Arcanjo, B. Ferrarini, M. Fasli, K. D. McDonald-Maier and S. Ehsan are with the School of Computer Science and Electronic Engineering, University of Essex, United Kingdom (email: bq17319@essex.ac.uk; bferra@essex.ac.uk; mfasli@essex.ac.uk; kdm@essex.ac.uk; sehsan@essex.ac.uk)

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

[3]S. Ehsan is also with the school of Electronics and Computer Science, University of Southampton, United Kingdom (email: s.ehsan@soton.ac.uk)

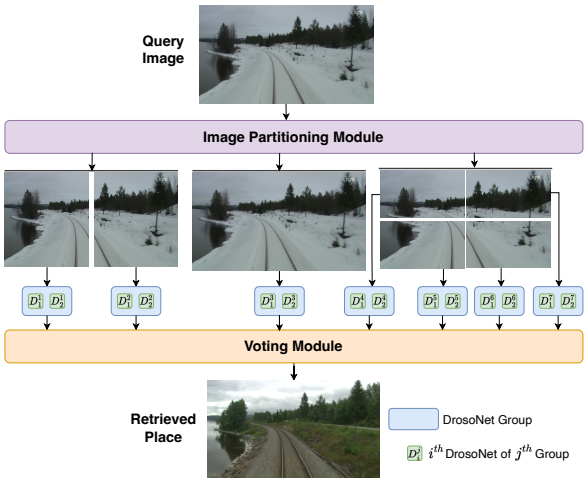


Fig. 1: The query image is divided into multiple heterogeneous regions. Each region is then fed as input into a specialized DrosoNet group which was trained only on that particular region from the training set images. Finally, the output of each group is aggregated in the voting module and a reference place is retrieved.

As alluded, mobile robotic platforms often operate under low-end hardware, often due to physical size or monetary budget, making computational cost an added important consideration when designing VPR techniques [7]. VPR methods based on Convolutional Neural Networks (CNNs) architectures have become increasingly popular due to their impressive performance. Indeed, visual features extracted from CNN layers achieve strong resilience against several of the visual challenges intrinsic to VPR [8]. However, as these networks grow deeper and more complex to achieve higher quality VPR, they also become less suitable for robotic setups equipped with heavily constricted hardware. Moreover, even if the hardware is able to support the use of an expensive CNN model in realtime, a lower computational demand is still valuable in saving power, allowing a mobile platform to operate for longer.

Recently, the authors proposed a lightweight VPR system [9] based on multiple bio-inspired voting units. Each unit, dubbed DrosoNet, is a compact neural network model inspired by the odour processing abilities of Drosophila Melanogaster (the common fruit fly) [10]. The approach relies on the inherent randomness of DrosoNet's initialization and training process, allowing for moderate unit differentiation, and its extremely low computational profile, allowing

for a multi-DrosoNet system which is brought together with a voting mechanism attuned to VPR. Despite strong VPR performance relative to its computational efficiency, the absolute VPR quality of the system makes it unreliable in many of the tested environments, particularly when dealing with strong viewpoint variations.

In this work, we propose a novel multi-DrosoNet localization pipeline which achieves increased VPR performance across various visual challenges, while maintaining a low computational profile. The core of the approach, dubbed RegionDrosoNet, relies on introducing additional model differentiation by training specialized DrosoNet groups on different regions of the training images. At inference time, as can be observed in Fig. 1, each partition of the query image is served as input to its respective group and each DrosoNet produces its reference place confidences. The training and inference process is tailored to DrosoNet, taking full advantage of its peculiarities: it's extremely fast and compact, allowing for the use of multiple units; it's a neural network classifier, not requiring storage of an image descriptor for every map location as a reference for image matching; DrosoNet groups trained on different image regions benefit from additional model differentiation induced by different training data, while units within each group continue benefiting from DrosoNet's inherent differentiation.

The outputs of all DrosoNets are then aggregated using a novel voting module which considers multiple top place candidates from each DrosoNet, allowing the system to converge on the most generally agreed upon reference place, mitigating the individual DrosoNets failing to realize a correct match.

We present a general setup for our proposed system which outperforms other lightweight VPR techniques across several benchmark datasets, while taking less time to retrieve a match. Furthermore, we also compare our results against high-performing but computationally expensive VPR methods to better situate this work in the literature.

The rest of this paper is organized as follows. Section II provides an overview of VPR literature with a focus on lightweight methods. Section III details our methodology, starting from a short DrosoNet overview, followed by the image partitioning module, training and inference processes, and finishing with the voting aggregation method. Section IV explains our experimental setup, providing insight into the benchmark datasets, evaluation metrics and model settings. Results are presented and discussed in Section V. We conclude in Section VI by summarizing our findings, highlighting key system limitations and possible future work.

II. Related Work

As the appearance of a place can vary substantially due to a wide variety of environmental and navigation factors, computing an image representation resilient against such changes becomes foundational for autonomous long-term navigation. Nevertheless, the computation, storage, and search of place representations should remain computationally efficient when the target robotic platform cannot afford to carry high-end hardware.

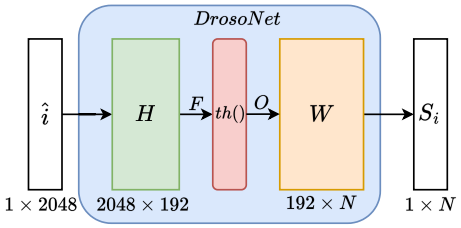


Fig. 2: DrosoNet model diagram.

The first image descriptors used for VPR were based on handcrafted methods such as Histogram-of-oriented gradients (HOG) [11], which has been successfully used as a global image descriptor for VPR [12]. Moreover, when combined with image region-of-interest detectors such as [13], [14], HOG acted as a local feature descriptor for VPR.

Machine learning techniques have became increasingly popular in the computer vision community over recent years, and CNN-based methods have crept into VPR applications, achieving high performance when dealing with both appearance changes [15] and viewpoint variations [16]. The image descriptors produced by the inner layers of CNNs, even when the model was trained for a different task, are effective in matching place images [17]. When trained specifically for the VPR problem [18], such as HybridNet and AMOSNet [19], these CNN-based descriptors achieve even higher VPR performance. With the continuous focus on absolute VPR reliability, these techniques have become increasingly complex. NetVLAD [20] separates the processes of CNN feature extraction and aggregation into two stages. Patch-NetVLAD [21] introduces yet another stage during descriptor matching. While these algorithmic variations and additions do result in increased VPR reliability, the computational cost of such methods prohibits their use with mobile robotics equipped with resource-constrained hardware. Several computationally efficient VPR methods have been proposed to address the shortcomings of CNNs. CoHOG [22] was proposed as an efficient and trainless algorithm for VPR. It finds regions-of-interest within an image and computes a HOG descriptor for each found region. CNN adaptations have been proposed to lower their computationally requirements. CALC [23] is a lightweight CNN-based VPR method which presents lower computational requirements. MobileNets [24] introduces depth-wise convolutions to lower overall computational requirements. Quantization of neural networks [25] into lower bit precisions has also been shown to improve computational profiles. These concepts have been bridged to VPR, with binary neural networks combined with depthwise convolutions [26] showing great computationally efficiency when paired with specialized hardware.

Efficient bio-inspired VPR methods are designed to mimic the neural activity of small animals, which exhibit incredible navigation capabilities relatively to the size of their brains [27], [28]. RatSLAM [29] takes inspiration from the neural activations of rats to perform navigation. FlyNet [30]

takes inspiration from the brain of the fruit fly [31] and its odour processing to perform highly efficient VPR by creating a small, binary image representation. Similarly, [32] also produces a binary image representation by applying a random projection and binarization step to the input image, a process inspired by the human neocortex. In the authors' previous work, a new algorithm also inspired by the fruit fly was introduced, dubbed DrosoNet [9], using multiple of these small models as voting units to perform highly lightweight VPR. [33] also proposes a multi-model approach for performing lightweight VPR, where individual units are small, region-specialized spiking neural networks.

Despite the efforts in developing lightweight VPR techniques, the absolute VPR performance of such methods remains unreliable. In this work, we propose a new approach to a multi-DrosoNet localization system, dubbed Region-DrosoNet, which aims to substantially improve absolute VPR reliability while remaining computationally efficient.

III. Methodology

In the interest of self-containment, this section starts by providing a technical background into the DrosoNet model. Following, we detail the proposed image partitioning module, which produces several heterogeneous image regions. The DrosoNet training and inference processes are then described. Finally, the voting module, responsible for aggregating the outputs of all DrosoNets into a final place prediction, is detailed.

A. DrosoNet

DrosoNet is a compact and fast neural network image classifier where each of the environment's total N places is a different class. We use the same configuration as in [9], which can be seen in Fig. 2. An 64×32 grayscale image is first flattened into a one-dimensional vector, denoted as $\hat{i}$, followed by a matrix multiplication with H, producing vector F. H is a binary, sparse, and randomly initialized matrix, where 10% of each column's elements are initialized to 1 and the remaining to 0. Matrix H is untrained, and thus the random initial values are fixed from its construction. F is then binarized by the function th, where the top 50% of values are set to 1 and the bottom 50% are set to 0, resulting in the binary vector O. W is a fully connected layer which learns to map O to one of the N classes, i.e. reference places. The final output vector s stores the score distribution for each reference place, and the DrosoNet's prediction is the index of the largest score in s.

While DrosoNet is a fast algorithm, its standalone VPR performance is too unreliable. Moreover, due to the randomness of its H matrix initialization and supervised training, different DrosoNets exhibit high variance in their VPR performance. Combining multiple DrosoNets was hence proposed as an avenue to improve overall VPR performance, relying only the native stochastic behaviour of the models for differentiation [9].

This work increases DrosoNet differentiation by training distinct models on different partitions of the original images, producing region specialized DrosoNets. Moreover, by training multiple DrosoNets on each image region, we continue taking advantage of the randomness associated with the initialization and training processes.

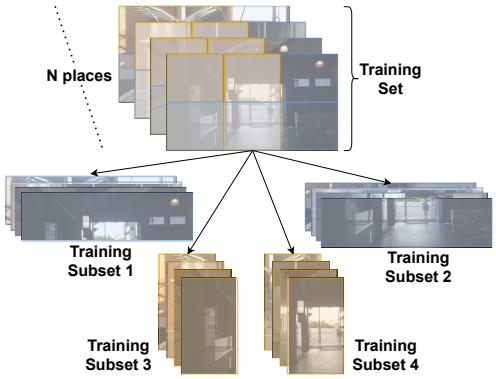


Fig. 3: A training subset is produced for each grid position. In this example, the grids $[(2\times1),(1\times3)]$ are used, with the blue regions highlighting the 2×1 grid and the yellow regions the 1×3 grid (the last column was omitted for visibility). The total number of regions is 5.

B. Image Partitioning

The image partitioning module receives as inputs an image i and grid dimensions (r, c), where r represents the number of rows and c the number of columns, outputting rc image regions. As detailed, DrosoNet operates with grayscale images with a resolution of 64×32, thus the produced regions are converted to grayscale and resized to the correct dimensions. In Section IV-D, we show how different grid setups can significantly impact the VPR performance of the overall system. Since it is not possible to predict which grid layout is best for the deployment environment without access to ground-truth information, we propose the use of multiple, heterogeneous image regions. In this arrangement, the partitioning process is simply repeated for G different grid settings. The total number of image partitions P can thus be computed as follows:

$$P = \sum_{g=1}^{G} r_g c_g \tag{1}$$

where r_g and c_g represent the number of rows and columns associated with grid setup g, respectively.

C. Training and Inference

Each dataset contains N image, one per place, in their training traversal. Before the training process, we construct P training subsets, each corresponding to one of the desired regions (Fig. 3). Each subset therefore also contains N image partitions.

A group of Z DrosoNets is assigned for each of the P training subsets, with each group being trained only on their respective grid position. The total number of DrosoNets in the system T is therefore given as:

$$T = PZ \tag{2}$$

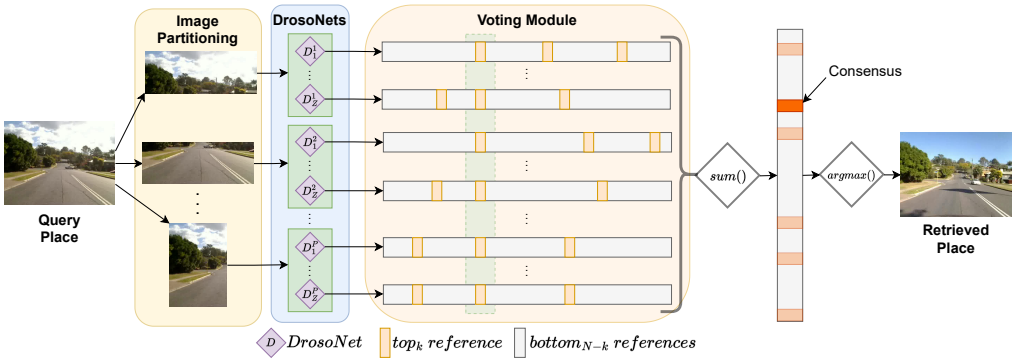


Fig. 4: The voting module receives all score vectors produced by each DrosoNet, with the largest K values being considered (in this case $K = 3$) and all remaining $N - K$ values being discarded.

At inference time, the query image is partitioned following the same G grids, and each DrosoNet is fed the corresponding region of its group, resulting in T score vectors for the query image. All these vectors are aggregated into a final prediction using the proposed voting module.

D. Voting Module

The voting scheme combines all the output score vectors into a final score vector from which the reference place can be identified. Fig. 4 illustrates the matching process for a single query image.

For each of the T score vectors s, the voting vector $\hat{s}$ is constructed by setting each of the N elements $\hat{s}_n$ as:

$$\hat{s}_n = \begin{cases} s_n & \text{if } s_n \geq top_K(s) \\ 0 & \text{else} \end{cases} \tag{3}$$

where $top_K(S)$ represents the value of the K^{th} largest score in s, with K being an hyperparameter. Fig. 4 shows an example of this operation with $K = 3$, where only the highest 3 scores per DrosoNet are considered and the remaining $N - K$ are set to 0. All the voting vectors are then summed element wise into the final score vector V:

$$v = \sum_{t=1}^{T} \hat{s}^t \tag{4}$$

and the retrieved reference place m is the most voted for index: $m = argmax(v)$.

IV. Experimental Setup

This Section details our experimental setup, starting with a presentation of the benchmark datasets, followed by evaluation metrics, comparison VPR methods and implementation settings of our proposed method.

A. Datasets

1) Nordland Fall & Winter: The Nordland dataset [34] consists of four train traversals with varying seasonal weather conditions. We use the Summer traversal as reference for training, testing on the Fall traversal to assess resilience against moderate appearance changes and on the Winter traversal to assess performance with extreme appearance changes. We use 1000 images per traversal, allowing for a margin for error of 1 frame around the ground-truth location.

2) Gardens Point Day-Right: The Gardens Point dataset [35] consists of three traversals around the Queensland University of Technology. We use the traversal filmed from a left viewpoint during the day as training and the right viewpoint daily traversal as testing, assessing resilience against moderate lateral shifts. The entire 200 images per traversal are utilized, with an error allowance of 2 frames.

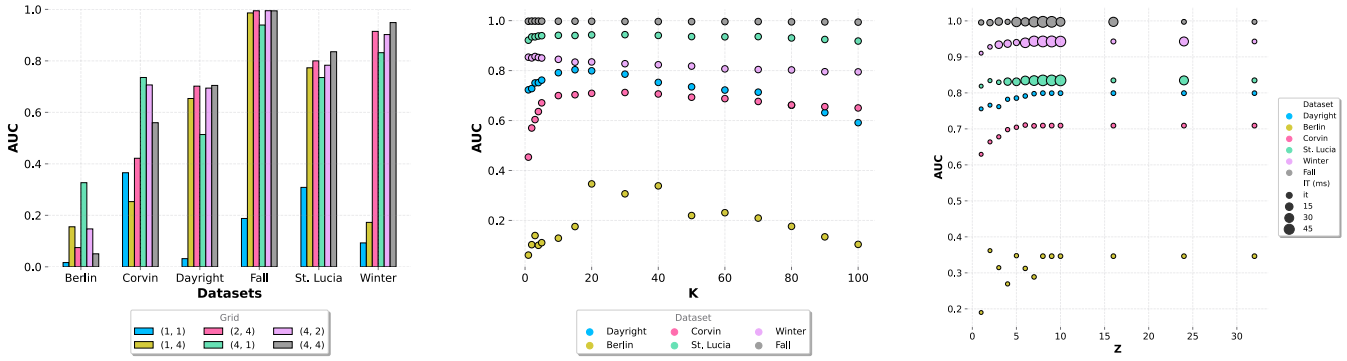
3) St. Lucia: St. Lucia [36] contains a number of car recorded sequences in St. Lucia, Bribane at different day times. The dataset exhibits moderate appearance changes and dynamic elements. We use the morning traversal recorded at 8:45AM (190809_0845) as reference and the afternoon traversal recorded at 2:10PM (190809_1410) as query, with 1150 images per traversal and an error margin of 2 frames around the ground-truth location.

4) Berlin: The Berlin dataset [37] contains traversals over three locations in Berlin: Halense Strasse, Kudamm and A100. The dataset is characterized by moderate to strong point of view variations and significant dynamic elements such as cars and pedestrians. Due to the small number of frames in each traversal, we combine the three locations into a single dataset, utilizing the traverses halensestrasse-2, kudamm-1 and A100-1 as references and halensestrasse-1, kudamm-2 and A100-2 as queries, resulting in a total of 250 images. We allow for an error margin of 1 frame.

5) Corvin 30 Degrees: Corvin [38] is a synthetic dataset recorded using flight simulation around the Corvin Castle, focusing on strong viewpoint and scale variations. We use 1000 images per traversal, with the one filmed at a 0 degree angle for training and the 30 degree traversal for testing, allowing for a ground-truth error margin of 20 frames. Corvin is a challenging dataset and a large error allowance is required to make results for all techniques conclusive [9].

B. Evaluations Metrics

1) Area Under The Precision-Recall Curve (AUC): AUC is a widely used metric for assessing VPR performance [39]. In our experiments, we compute Precision-Recall pairs by varying the confidence threshold for which a technique considers a match correct [40]. There is usually an inverse



(a) (b) (c)

Fig. 5: AUC impact of the region grid (5a), the top K voted places (5b) and the number of DrosoNets per region Z (5c).

relationship between Precision and Recall, and thus the area under the plotted curve is a strong indicative of VPR performance [41]. A high AUC value is most useful for applications where retrieving enough possible correct matches is more important than assuring every retrieved match is absolutely correct [42].

2) Extended Precision (EP): The Recall at 100% Precision (R_{P100}) metric [43] computes how many correct matches are retrieved before an incorrect one is introduced. It is useful for applications where a single incorrect match would result in catastrophic failure but does not consider the lower performance bound of the technique. EP [44] combines R_{P100} with the Precision at Minimal Recall, providing a more balanced performance view for such applications.

3) Inference Time (IT): We measure IT as the time elapsed from the technique receiving a query image to a match being computed. This includes the time required for any runtime image pre-processing, descriptor computation and descriptor matching. We compute IT on the St. Lucia dataset, taking the average of 1100 inferences. We compute these results on an Intel 12900k processor, running Ubuntu 20.03. The tests are purposely ran without a GPU, as many lower performance robots do not carry an on board dedicated GPU.

C. Comparison VPR Techniques

We compare RegionDrosoNet to several VPR techniques which claim computational efficiency as one of their main strengths: CALC [23], CoHOG [22], and Voting [9]. Moreover, to better situate our work, we additionally include comparison against the computationally expensive VPR algorithms HybridNet [19] and Patch-NetVLAd [21]. We use the implementations in [40] for CALC, CoHOG and HybridNet, and [42] for Patch-NetVLAD. For Voting, we test both the implementation given in [9] with 32 DrosoNets and an additional setup with 82 to match the same number of DrosoNets as our proposed setup.

D. Ablation Studies & Implementation Details

RegionDrosoNet has three main hyperparameters: the grid setups used to construct image regions, the number of DrosoNets per region Z, and the number of top_K voted places per DrosoNet. We conduct ablation studies to find optimal settings with the aim of providing a general setup that performs strongly across all datasets, rather than fine-tuning the system for each scenario. The results of these studies can be seen in Fig. 5.

As can be seen in Fig. 5a, different grid settings significantly impact VPR performance, and the optimal individual grid setting varies from dataset to dataset. As such, we use a combination of all tested partitioning grids:

$$[(1,1),(1,4),(4,1),(2,4),(4,2),(4,4)] \tag{5}$$

resulting in a total of 41 partitions, following the example scheme in Fig. 3.

The choice for K also has a substantial impact on VPR performance, as can be seen in Fig. 5b. We set $K = 20$ as it presents the best overall AUC performance across all datasets.

Finally, the number of DrosoNets per region Z has a significant impact on both AUC performance and inference time, observable in Fig. 5c. We set the system to $Z = 2$, as there are heavily diminishing AUC returns with higher Z values, even lowering VPR performance on Corvin and Berlin. With the choice of grids described above, the total number of DrosoNets in the system becomes 82.

Each DrosoNet is trained for 200 epochs using the Adam optimizer [45] and with a learning rate of 0.001.

V. Results

This section presents and discusses our results, firstly with a comparison of RegionDrosoNet versus other computationally efficient VPR techniques, followed by a comparison against expensive methods and finalizing with a per-region performance analysis.

A. VPR Performance VS Lightweight Methods

In Fig. 6 we observe the VPR performance in terms of AUC for all tested techniques. RegionDrosoNet outperforms every other lightweight algorithm on all appearance-based datasets (Winter, Fall and St. Lucia). The performance advantage on the Winter dataset over other efficient methods is the most notable, with RegionDrosoNet more than doubling

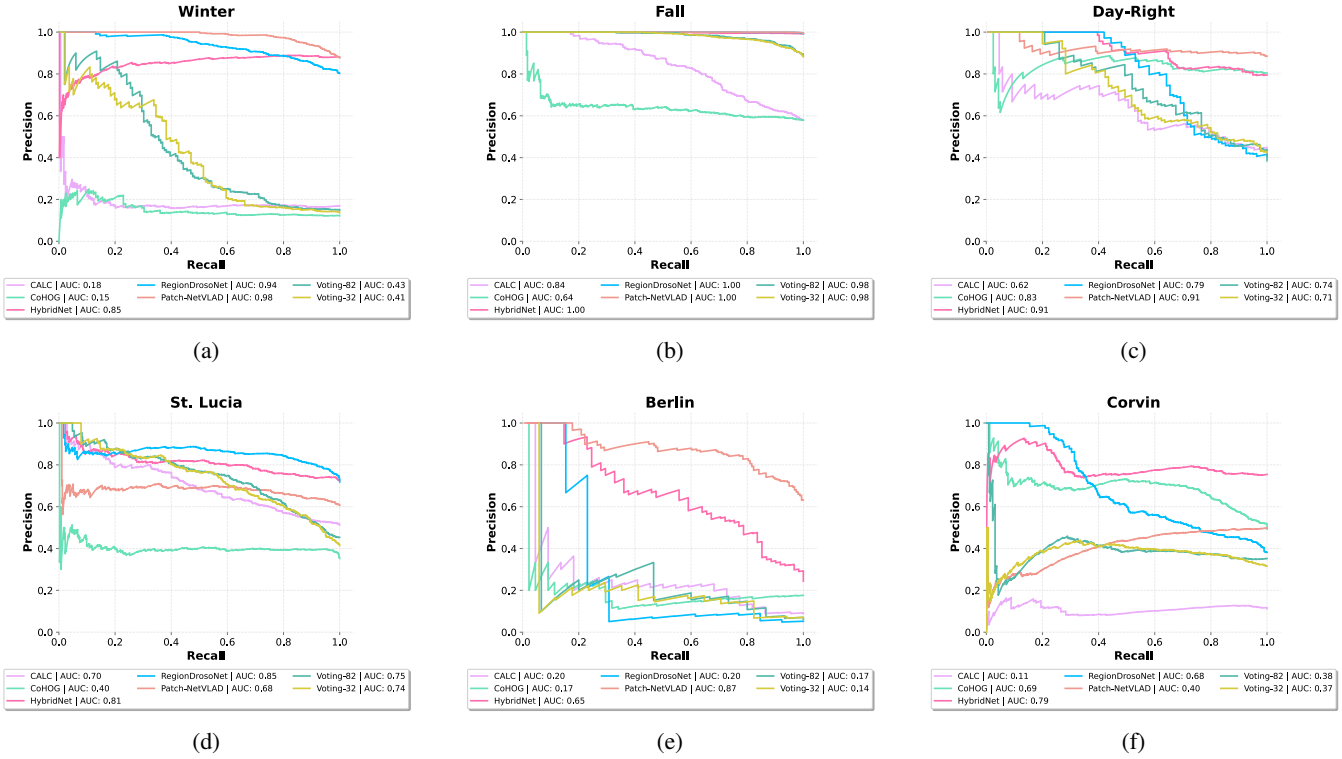


Fig. 6: Precision-recall curves and respective AUC

the AUC of the second best efficient technique (Voting-82). Viewpoint performance on the Corvin dataset is also commendable, with RegionDrosoNet achieving the highest EP result (Fig. 7) and matching CoHOG in AUC. While all lightweight techniques perform poorly on the Berlin dataset, our method achieves the highest EP amongst them and ties with CALC for the highest AUC. The VPR performance of Voting-32 and Voting-82 is functionally indistinguishable, showing that simply increasing the number of DrosoNets does not contribute significantly to place matching. Conversely, the use of 82 units in the proposed pipeline provides significant improvements in VPR, as demonstrated by the performance gap between RegionDrosoNet and Voting-82. Table I shows the inference times at runtime for every tested technique. RegionDrosoNet is the third-fastest method, second only to Voting-32 and Voting-82, the latter due to the extra image pre-processing required by RegionDrosoNet. Nevertheless, it achieves substantially higher VPR reliability on both viewpoint and appearance-based visual challenges while remaining 18 times faster than CALC and over two orders of magnitude faster than CoHOG.

Despite these efficiency advantages, it is worth noting that different methods can offer various benefits over each other. CoHOG, while requiring the reference traversal images for the reference map computation, is a trainless technique. CALC, while trained and also requiring the reference place images for the descriptor database, does not require environment specific training. RegionDrosoNet, while achieving better VPR performance and efficiency, does require environment specific training due to its dependency on DrosoNet.

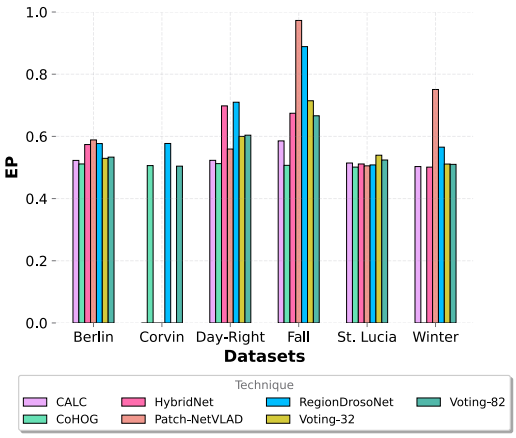


Fig. 7: Extended precision (EP) comparison.

The choice of a VPR technique is highly application dependant and all factors such as data availability, hardware, deployment environment and risk of failure should be taken into account.

B. VPR Performance VS Expensive Methods

As can be seen in Table I, HybridNet and Patch-NetVLAD are significantly slower than the lightweight methods.

Despite its substantially lower computational requirements, RegionDrosoNet is able to compete with these expensive methods, even outperforming them on some datasets. On the Corvin dataset, RegionDrosoNet achieves higher EP

TABLE I: Inference Time (IT) & Frames Per Second (FPS)

Model	IT (ms)	FPS
CoHOG	671	1.49
CALC	166	6.02
Voting-32	3	333.33
Voting-82	8	125.00
HybridNet	3318	0.30
Patch-NetVLAD	2892	0.35
RegionDrosoNet	9	111.11

(Fig. 7). In the challenging Winter dataset, it outperforms HybridNet in both EP and AUC. The highest performance drop from RegionDrosoNet is in Berlin, where it loses substantially in both AUC and EP to the costly techniques.

C. Per-Region Insights

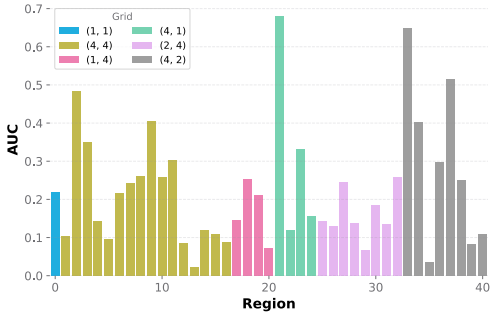
In Fig. 8 we show RegionDrosoNet's AUC per region on the Corvin (8a) and St. Lucia (5b) datasets. As per Eq. 1 and Eq. 5, our setup has a total of 41 regions, each represented by a bar, where the colour code shows the corresponding grid arrangement from which it originated from. It is clear that some regions perform substantially better than others, and region performance is dataset dependant. Furthermore, the region corresponding to the whole query image (region 0, in blue) is not the best performing one.

Looking at Fig. 9, we find visual insights for the large performance discrepancy. On Corvin, region 13 does not have enough visual detail for DrosoNet to specialize on, while 21 contains strong features. Region 13 also performs better than the whole query image, as the former has less non-detailed visual zones and less compression resulting from the image scaling pre-processing. Finally, St. Lucia follows the same pattern with its respective best and worst performing regions.

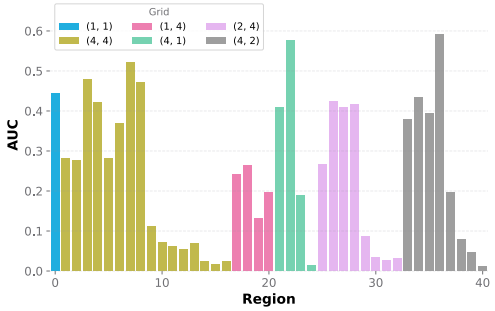
VI. Conclusions and Future Work

In this work, we propose RegionDrosoNet: a novel multi-DrosoNet localization system which significantly improves upon the VPR performance of current lightweight methods while remaining computational efficient. The approach relies on increasing the differentiation of different DrosoNets by training specialized groups on several image partitions. Moreover, the introduce a novel voting method which considers multiple top place candidates from each DrosoNet, allowing a correct consensus to be reached even if individual DrosoNets place an incorrect highest scoring match.

DrosoNet is a neural network classifier which requires training on the reference set of the target environment. While training time is low compared to expensive models, it remains a limitation of this work. Future research could focus on adapting DrosoNet into a descriptor-based method which does not require environment specific training.



(a) Corvin



(b) St. Lucia

Fig. 8: AUC per region, with colour highlighting the associated grid dimensions. Within each grid, regions are placed from left-to-right, top-to-bottom. E.g., for grid $(4, 4)$, its first bar represents $row1$, $column1$, the second bar $row1$, $column2$, the fifth bar $row2$, $column1$, etc.

Fig. 9: Query regions: whole image in blue, best performing region in green, and worse performing region in red.

References

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2015.

[2] F. Maffra, Z. Chen, and M. Chli, "tolerant place recognition combining 2d and 3d information for UAV navigation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2542–2549.

[3] A. Ranganathan, S. Matsumoto, and D. Ilstrup, "Towards illumination invariance for visual localization," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 3791–3798.

[4] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *Twenty-eighth AAAI conference on artificial intelligence*, 2014.

[5] A. Pronobis, B. Caputo, P. Jensfelt, and H. I. Christensen, "A discriminative approach to robust visual place recognition," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2006, pp. 3829–3836.

[6] C.-C. Wang, C. Thorpe, S. Thrun, M. Hebert, and H. Durrant-Whyte, "Simultaneous localization, mapping and moving object tracking," *The International Journal of Robotics Research*, vol. 26, no. 9, pp. 889–916, 2007.

[7] 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 *2019 NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*. IEEE, 2019, pp. 103–108.

[8] Y. Hou, H. Zhang, and S. Zhou, "Convolutional neural network-based image representation for visual loop closure detection," in *2015 IEEE International Conference on Information and Automation*, 2015, pp. 2238–2245.

[9] B. Arcanjo, B. Ferrarini, M. Milford, K. D. McDonald-Maier, and S. Ehsan, "An efficient and scalable collection of fly-inspired voting units for visual place recognition in changing environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2527–2534, 2022.

[10] T. A. Ofstad, C. S. Zuker, and M. B. Reiser, "Visual place learning in drosophila melanogaster," *Nature*, vol. 474, no. 7350, pp. 204–207, 2011.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[12] C. McManus, B. Upcroft, and P. Newmann, "Scene signatures: Localised and point-less features for localisation," in *Proceedings of Robotics: Science and Systems*, Berkeley, USA, July 2014.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[14] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[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," *ICRA 2019 Workshop on Database Generation and Benchmarking of SLAM Algorithms for Robotics and VR/AR*, 2019.

[16] 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 2019 Workshop on Aerial Robotics*, 2019.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[18] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

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

[20] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[21] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 14 141–14 152.

[22] 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 Robotics and Automation Letters*, vol. 5, no. 2, pp. 1835–1842, 2020.

[23] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," in *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018.

[24] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[25] J. Yang, X. Shen, J. Xing, X. Tian, H. Li, B. Deng, J. Huang, and X.-s. Hua, "Quantization networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7308–7316.

[26] B. Ferrarini, M. Milford, K. D. McDonald-Maier, and S. Ehsan, "Highly-efficient binary neural networks for visual place recognition," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 5493–5500.

[27] A. Cope, C. Sabo, E. Yavuz, K. Gurney, J. Marshall, T. Nowotny, and E. Vasilaki, "The green brain project–developing a neuromimetic robotic honeybee," in *Conference on Biomimetic and Biohybrid Systems*. Springer, 2013, pp. 362–363.

[28] A. Narendra, S. Gourmaud, and J. Zeil, "Mapping the navigational knowledge of individually foraging ants, myrmecia croslandi," *Proceedings of the Royal Society B: Biological Sciences*, vol. 280, no. 1765, p. 20130683, 2013.

[29] M. J. Milford, G. F. Wyeth, and D. Prasser, "RatSLAM: a hippocampal model for simultaneous localization and mapping," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, vol. 1. IEEE, 2004, pp. 403–408.

[30] M. Chancán, L. Hernandez-Nunez, A. Narendra, A. B. Barron, and M. Milford, "A hybrid compact neural architecture for visual place recognition," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 993–1000, 2020.

[31] S. Dasgupta, C. F. Stevens, and S. Navlakha, "A neural algorithm for a fundamental computing problem," *Science*, vol. 358, no. 6364, pp. 793–796, 2017.

[32] P. Neubert, S. Schubert, and P. Protzel, "A neurologically inspired sequence processing model for mobile robot place recognition," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3200–3207, 2019.

[33] S. Hussaini, M. Milford, and T. Fischer, "Ensembles of compact, region-specific & regularized spiking neural networks for scalable place recognition," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4200–4207.

[34] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging SeqSLAM on a 3000 km journey across all four seasons," in *Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA)*. Citeseer, 2013, p. 2013.

[35] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 4297–4304.

[36] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, "Fab-map+ ratslam: Appearance-based slam for multiple times of day," in *2010 IEEE international conference on robotics and automation*. IEEE, 2010, pp. 3507–3512.

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

[38] F. Maffra, L. Teixeira, Z. Chen, and M. Chli, "Real-time wide-baseline place recognition using depth completion," *IEEE Robotics and Automation Letters*, 2019.

[39] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[40] M. Zaffar, S. Garg, M. Milford, J. Kooij, D. Flynn, K. McDonald-Maier, and S. Ehsan, "Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *International Journal of Computer Vision*, pp. 1–39, 2021.

[41] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.

[42] S. Schubert, P. Neubert, S. Garg, M. Milford, and T. Fischer, "Visual place recognition: A tutorial," *IEEE Robotics & Automation Magazine*, p. 2–16, 2024. [Online]. Available: http://dx.doi.org/10.1109/MRA.2023.3310859

[43] D. Bai, C. Wang, B. Zhang, X. Yi, and X. Yang, "Sequence searching with cnn features for robust and fast visual place recognition," *Computers & Graphics*, vol. 70, pp. 270–280, 2018.

[44] B. Ferrarini, M. Waheed, S. Waheed, S. Ehsan, M. J. Milford, and K. D. McDonald-Maier, "Exploring performance bounds of visual place recognition using extended precision," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1688–1695, April 2020.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *3rd International Conference for Learning Representations, San Diego, 2015*, 2015.