Why-So-Deep: Towards Boosting Previously Trained Models for Visual Place Recognition

M. Usman Maqbool Bhutta[1], Yuxiang Sun[2], Darwin Lau[3], and Ming Liu[4]

arXiv:2201.03212v1 [cs.CV] 10 Jan 2022

Abstract—Deep learning-based image retrieval techniques for the loop closure detection demonstrate satisfactory performance. However, it is still challenging to achieve high-level performance based on previously trained models in different geographical regions. This paper addresses the problem of their deployment with simultaneous localization and mapping (SLAM) systems in the new environment. The general baseline approach uses additional information, such as GPS, sequential keyframes tracking, and re-training the whole environment to enhance the recall rate. We propose a novel approach for improving image retrieval based on previously trained models. We present an intelligent method, *MAQBOOL*, to amplify the power of pre-trained models for better image recall and its application to real-time multiagent SLAM systems. We achieve comparable image retrieval results at a low descriptor dimension (512-D), compared to the high descriptor dimension (4096-D) of state-of-the-art methods. We use spatial information to improve the recall rate in image retrieval on pre-trained models. Material related to this work is available at https://usmanmaqbool.github.io/why-so-deep.

Index Terms—Localization, Visual Learning, Recognition

(a) Recall test on Tokyo 24/7 (b) Recall test on Pitts250k

Fig. 1: Our system performance at a lower feature dimension (512-D) is comparable to NetVLAD [6] performance at a higher feature dimension (4096-D). We use NetVLAD best trained network ($VGG-16, f_{VLAD}$) for evaluation.

I. INTRODUCTION

VISUAL place recognition is extensively used in the robotics industry with application in all kinds of SLAM systems for the loop closure detection [1], [2]. The top candidate in the image retrieval helps a lot in the multiagent SLAM system for creating large-scale 3D maps [3]–[5]. Convolutional neural network (CNN)-based approaches such as NetVLAD [6] and DGC-Net [7] produce promising results in the image retrieval.

There are several types of common solutions to enhance the localization performance further. Each of the methods involves retraining the whole network in addition to the data. The first and very common approach in visual place recognition is based on attention-seeking regions in the images. Work-related to landmarks distribution and partitioning of pictures based on regions is also present in the literature [8]–[11]. [12] incorporated NetVLAD and made a features pyramid of the top landmarks in the image to achieve better results. [13]

[1]M. Usman Maqbool Bhutta and [2]Darwin Lau are with the C3 Robotics Lab, Department of Mechanical and Automation, CUHK, Hong Kong. usmanmaqbool@outlook.com ; darwinlau@cuhk.edu.hk

[2]Yuxiang Sun is with the Department of Mechanical Engineering, PolyU, Kowloon, Hong Kong. sun.yuxiang@outlook.com

[4]Ming Liu is with the Department of Electronic and Computer Engineering, HKUST, Hong Kong. eelium@ust.hk

introduced a multi-layered region-based method that extends DGC-Net and incorporates NetVLAD. Even though the better real-time performance of DGC-Net [7], sometimes, the system faces difficulty due to the incorporation of objects changing with time in the images.

The second most used idea is to increase the depth of the neural network by adding several additional layers with the default network. For instance, a multi-layered region-based framework is introduced in [14] which also uses NetVLAD and performs dense pixel matching to achieve better place recognition. For large-scale image correspondence matching, scientists have also introduced several upgrades to the network. [15] presented HD-CNN, a hierarchical deep CNN scheme, while [16] and [17] similarly partitioned the spatial information in higher layers. But the utilization of these methods in a real-time SLAM system is still very challenging due to the computation time and large size of feature dimension.

The third widely used method in the robotics industry is done by integrating 3D-depth information. The corresponding 2-D images, along with the 3D maps, enhance the system performance in loop closure detection [18]–[21]. Moreover, GPS information [22], semantics graph matching [19], and attention-seeking approaches [9], [13], [23] have shown good results for image retrieval tasks. In robotics, the place recognition module should not be tightly coupled with GPS or 3D data. Otherwise, it will become harder for the server of the multiagent SLAM system to handle. Despite their benefits, all the above-described approaches require retraining the complex

networks to enhance system accuracy.

In our proposed study, our system at a lower feature dimension (i.e., 512-D) is able to achieve accuracy similar to a higher feature dimension (i.e., 4096-D) while tested with the same pre-trained model. A detailed comparison is shown in Fig 1. In our work, instead of following a cascading training pipeline or to go deeper in terms of feature dimension, we probabilistically enhance the image retrieval performance based on a pre-trained model for a more reliable place recognition. So, we introduce this as the Multiple AcQuisitions of perceptiBle regiOns for priOr Learning (MAQBOOL) approach. These significant correspondence regions will help in probabilistic landmarks elevation by splitting the full spatial information into multiple regions and estimate their descriptor ℓ_2 distance co-relations, which significantly increases the power of the pre-trained models. In addition, training new models to be utilized in new places involves intensive computation of deep learning models each time.

As multiagent SLAM is the next interest for computer vision researchers, we strongly believe that our contributions will help the computer vision community in many multiagent SLAM scenarios. The contribution of this research is four-fold, as follows:

- Our main contribution is to enhance recall accuracy of previously trained models.
- We introduce a probabilistic layer to constrain the local representation such as prominent regions, along with the global representation of images. The global description of the pre-trained model and the local consistency introduced in our work enables the system to yield good performance at a large-scale.
- Our system shows comparable accuracy at a low descriptor dimension (512-D) compared to the high descriptor dimension (4096-D) of the current state-of-the-art [6].
- In our results, we show good performance of our system at low-dimensional features with a previously trained model while tested in the new environment. It enables the SLAM system to detect the loop closure at good accuracy everywhere.

II. Related works

This work is related to improving image retrieval, which plays a big part in visual place recognition for vision-related applications and any kind of SLAM system. To improve image retrieval, scientists have worked towards making robust global descriptors, and some have utilized local features along with a global representation, as mentioned in previous section. We discuss closely related work below.

Researchers developed a geometric image correspondence-based system, which shows good performance after utilizing dense geometric information [14]. It selects the top candidates from the database using NetVLAD and performs dense pixel matching with the query image. Their geometric model was created by fitting the planer homography to the 3D information, SIFT features and CNN descriptor. This was done with alteration to DGC-Net [7], which does not deal with the 3D structure. After modifying DGC-Net, [14] utilized a unified correspondence map decoder (UCMD) for the dense matching between the top candidates and the query image. It processes a multi-resolution feature pyramid and CNN layers. At the end, the authors used a neighbourhood consensus networks (NCNet), meaning we can summarize their system as NetVLAD-DGC-NC-UCMD-Net. This long computation pipeline makes their system complex, and it requires intense computation for each query image and all database images.

[21] also considered 3D information and performed extensive nearest-neighbour explorations in the descriptive space. In addition, they used coarse correspondence estimation, while [14] used a learned convolutional decoder. Another pyramid-aggregation-based method is found in [16]. It also uses previous NetVLAD training results along with enhancing the accuracy compared to NetVLAD, and introduces weighted triplet loss for updating the weights after each epoch while training the new model.

[24] introduced APANet, which uses principal component analysis (PCA) power whitening along with pyramid aggregation of attentive regions. APANet selects the features of key attention-seeking regions and performs sum pooling, and its results are on 512-D features trained using the AlexNet [25] and VGG-16 [26] networks. Their work is similar to [14] for multi-scale region aggregation to build the pyramid. In addition, their use of PCA power whitening makes it more complex for the system to create a full descriptor and inconvenient to use in real-time applications.

The training-free approach from [10] uses edge boxes [27] for the top landmarks selection in a given scene. This method scales down each landmark feature from a 65K vector to a 1K-dimensional representation using Gaussian random projection. Based on the landmark features, cosine distances between all the landmark proposals are calculated to find the similarity. No training is involved, for better utilization of the landmarks. An unsupervised approach [28] presents a method to re-ranking the NetVLAD top-20 candidates. Their method utilized global as well as local features for improving the recall rate.

III. Multiple acquisitions of perceptible regions for prior learning

An overview of our MAQBOOL system is shown in Fig. 2. Given a query image, we retrieve the nearest candidates in the database descriptor space. We use NetVLAD, which is a fast and scalable method, for the image retrieval application.

Our proposed method consists of three parts. Part I explains the selection of the top regions for the local representation from the images; Part II describes the corresponding spatial information processing using NetVLAD; and Part III shows the probabilistic manipulation, in a pairwise manner, of the query image with initially retrieved database images for more reliable place recognition. These three parts are explained in the following subsections.

A. Top Regions Selection

In Section II, we explained that APANeT [24] uses the top regions and applies single and cascaded blocks to achieve better performance. After taking inspiration from the top
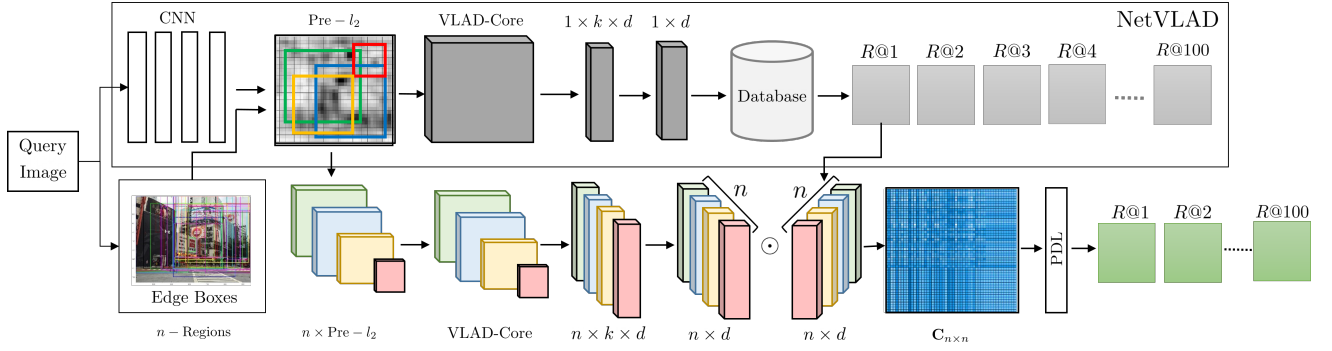
Fig. 2: Overview of the MAQBOOL system. For each image, the spatial $Pre - l_2$ layer is partitioned according to the boxes detected by edge boxes. All the regions are cropped and then normalized for the feature vectors. k is the number of clusters, and d is the feature dimension. The probabilistic decision layer (PDL) is trained using the correlation distance matrix $\mathbf{C}_{n \times n}$ and the ground truth of the ToykoTM dataset.

(a) Top-ranked image correspondence proposals are marked on the query image.

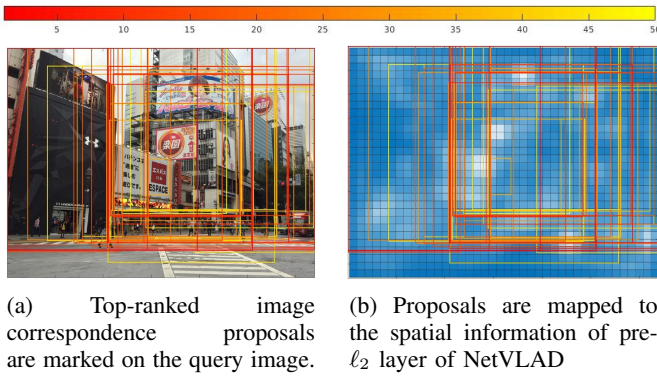(b) Proposals are mapped to the spatial information of pre-ℓ_2 layer of NetVLAD

Fig. 3: Query image from the Tokyo 24/7 dataset. Top region proposals are selected using edge boxes [27]. The intensity bar shows the top-scoring region selections on the image.
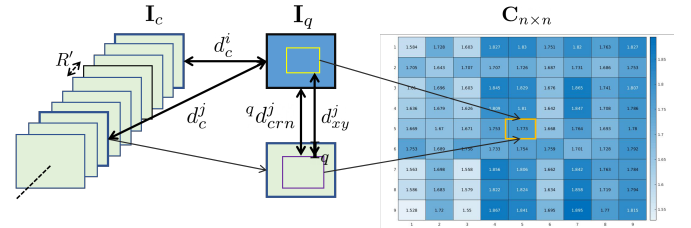

Fig. 4: The query image is compared with all the top candidates. R' is the distance span within adjacent image proposals. The correlation distance matrix $\mathbf{C}_{n \times n}$ is calculated between the query image and the candidate images by taking the feature distance between all top landmarks. Each value of this matrix represents the feature distance between specific landmarks shown by a yellow box, and their distance is represented as d_{xy}^j.

regions selection in [10], we also choose edge boxes [27] instead of using the simple regions selection techniques from [29] and [30]. Edge boxes is a fast and unsupervised method that detects the likelihood of an object based on the points on the edges.

B. Spatial Landmarks Estimation

We process the landmarks information at the pre-ℓ_2 layer of NetVLAD. All the top regions' enclosed boxes are mapped to this pre-ℓ_2 layer, and we crop the corresponding spatial information to create the NetVLAD 512-D and 4096-D feature vectors of all the boxes. These feature vectors are used for the probabilistic landmarks elevation in the next section. Essential regions are calculated in the query image and candidate images from the database. Fig. 3 shows the top proposals based on the objectness scores and their mapping at the pre-ℓ_2 layer of the NetVLAD pipeline.

For a query image $\mathbf{I}_q$, NetVLAD predicts top-100 matches from the database, based on their ℓ_2 distances in the descriptor space. Let's assume, for this query image $\mathbf{I}_q$, $\mathbf{I}_c$ are the top candidates from the database and d_c^j is the corresponding ℓ_2 feature distance of jth candidate image from the query image. We further denote the feature distance $^q d_{crn}^j$ between query

and different regions of jth candidate image pictures by a double-sided solid arrow, as shown in Fig. 4. We calculate relative distances R' of two adjacent candidate proposals by taking the derivative of the original distance set d_c.

Instead incorporating multi-scale pyramid aggregation or estimating sets of cyclically consistent dense pixel matches, we simply performed probabilistic spatial landmarks elevation followed by the correlation distance for the re-ranking of the top retrieved candidates. For the d-dimensional feature vector and n (number of spatial regions plus one considering the full image), we define the whole image representation by $\mathbf{F}$, which has a size of $n \times d$. We estimate the correlation distance matrix $\mathbf{C}_{n \times n}$ of the query image $\mathbf{I}_q$ with the $j^{th} \in \mathbf{I}_c$ database images $\mathbf{I}_c^j$ as

$$\mathbf{C}_{n \times n}^j = \mathbf{F}_{q(n \times d)} \times \mathbf{F}_{c(n \times d)}^{j \ T}, \tag{1}$$

We filter the irrelevant distances score from $\mathbf{C}_{n \times n}$ after subtracting the maximum distances $d_c^{max} \in d_c$ and using sign function:

$$\mathbf{C}_f^j = \mathbf{C}_{n \times n}^j - d_c^{max}. \tag{2}$$

$$\mathbf{D}^j = \begin{cases} sgn(\mathbf{C}_f^j), & \text{for } c_f < 0 \\ 0, & \text{for } c_f \geq 0, \end{cases} \tag{3}$$

where $c_f \in \mathbf{C}_f^j$. By using $\mathbf{D}^j$, we drop large distances landmarks from $\mathbf{C}_{n \times n}^j$ as follows:

$$\mathbf{C}_{n \times n}^j = |\mathbf{D}_{qj}^j| \odot \mathbf{C}_{n \times n}^j. \tag{4}$$

C. Probabilistic Spatial Landmarks Elevation

Firstly, we determine the information $\mathbf{S}^j$ based on the size of the boxes of both the query and database image:

$$s_{xy}^j = \beta * P_b^q * P_b^c * e^{-(d_c^{min} + d_{xy}^j)} * e^{-R_j'}, \tag{5}$$

where $s_{xy}^j \in \mathbf{S}^j$, $x, y \in \{1, ...n\}$, $\beta = 10$, $d_{xy}^j \in \mathbf{C}_{n \times n}^j$, and $R_j' = \frac{d}{dx}(d_c^j)$ is the candidates' relative distances' difference. Furthermore, b_w and b_h are the width and height, respectively, of the landmarks' bounding box. We estimate the probabilities of the landsmarks in the query and candidate images as follows:

$$P_b^q = e^{-\frac{b_w^q * b_h^q}{q_w * q_h}} \text{ and } P_b^c = e^{-\frac{b_w^c * b_h^c}{c_w * c_h}}. \tag{6}$$

The reason for taking the negative exponential of the weight of the landmarks' over the full image is that a smaller bounding box results in a lower probability of a good match. For instance, in night images, most regions are black, so there exists a higher chance of accumulating dark region effects.

We use information matrix $\mathbf{S}^j$ and filtered correlation distance matrix $\mathbf{C}_f^j$ to produce the probabilistic correlation matrix $\mathbf{P}_{SC}^j$ as follows:

$$\mathbf{P}_{SC}^j = \mathbf{S}^j \odot \mathbf{C}_f^j. \tag{7}$$

Each column of $\mathbf{P}_{SC}^j$ corresponds to a probabilistic match of a particular landmark in the query image with all landmarks in the jth candidate image. We further shrink $\mathbf{P}_{SC}^j$ to 10×10 by sorting each column of $\mathbf{C}_{n \times n}$ and index matching with $\mathbf{P}_{SC}^j$.

P_{SM}^j processes the original ℓ_2 distances of all retrieved candidates from the query image. We use softmax for estimating the probability controlled by the c_{min}:

$$P_{SM}^j = e^{-c_{min}^j} \cdot \underbrace{\left(\frac{e^{-d_c^j}}{\sum_i e^{-d_c^i}} \right)}_{\text{softmax}} = e^{-c_{min}^j}.\text{softmax}(e^{-d_c^j}), \tag{8}$$

where c_{min}^j is the minimum value of $\mathbf{C}_{n \times n}^j$. We pass this estimated probability information to the regression process to create the predictive model.

We estimate the probability $\mathbf{M}^j$ by utilizing the P_{SM}^j and $\mathbf{P}_{SC}^j$:

$$\mathbf{M}^j = P_{SM}^j * \mathbf{P}_{SC}^j. \tag{9}$$

We also consider incorporation of the feature distance d_c^j between the query image and the candidate image, as well as the distances $^q d_{crn}^j$ between the top regions of the candidate image, with the whole query image, as follows

$$\mathbf{C}_{qc}^j = [R_j', {}^q d_{cr1}^j, {}^q d_{cr2}^j, {}^q d_{cr3}^j, ..., {}^q d_{crn}^j], \tag{10}$$

In this work, we choose the top 10 feature distances $^q d_{crn}^j$ of the candidate image's top regions from the whole query image.

D. Prediction Model and Probabilistic Distance Update

We design the probabilistic decision layer (PDL) using the ground truth based on the TokyoTM validation dataset. The information estimated in the previous subsection is used in creating the model as follows:

$$P_M^j = f(d_c^j, \mathbf{C}_{qc}^j, \mathbf{M}^j, Y^j). \tag{11}$$

We train P_M^j with the ground truth Y^j of about 250 images and create the model. We choose bootstrap aggregation in the decision tree (DT), which allows the tree to grow on an independently drawn bootstrap, duplicate of the input. This reduces the variance and increases accuracy. We trained bootstrap-aggregated decision trees of sizes 50 (DT-50) and 100 (DT-100) for the testing. After creating the model, we apply it to work like the prior distribution to predict the response. In this manner, we update all the distances of the top candidates for re-ranking the retrieval images as follows:

$$d_{new}^j = |d_c^j - \alpha \log^{(P_M^j)}|, P_M^j \in [1, 2]. \tag{12}$$

We keep the model response binary. The predicted value '1' corresponds to an irrelevant match, while '2' indicates the nearest match. The regularizing variable α controls the weight of the probabilistic response. We choose $\alpha = 1.15$ while working at the 512-D feature vectors, while at the higher dimension, i.e., 4096-D, we use $\alpha = 0.31$. The main motivation for using $\alpha = 0.31$ is to minimize the effect of the regularizing variable. We observed that the feature distance between two adjacent images is small at a higher dimensional feature space. So the impact of the regularizing variable should also be low for the high-dimensional features.

IV. RESULTS AND DISCUSSION

Our proposed method requires small training of decision tree model, and it works excellently when tested in new surroundings. All the testing results in this section are based on the same decision tree model. We used the TokyoTM validation dataset to train this decision model for the PDL layer. MAQBOOL probabilistically elevates the perceptible regions' distributions for improving the loop closure module of the multiagent SLAM system. Our system performs better image retrieval than schemes that change the network with a complex structure and include additional sensor information or perform repetitive training to get impressive results on challenging datasets.

A. Datasets and Implementation

NetVLAD is mainly evaluated on Pittsburgh [31] and Tokyo 24/7 [32] datasets. We used the same NetVLAD VGG-16-based models for the performance evaluation. We tested our MAQBOOL method on the Pittsburgh and Tokyo 24/7 datasets compared with NetVLAD and APANet [24]. The Pittsburgh 250K dataset consists of 254K perspective images taken from 10.6K Google Street View and 8.2K query images, while the Tokyo 24/7 dataset has 76K database images and 315 query images.

(a) Pitts250k at 512-D (b) Pitts250k at 4096-D (c) Tokyo 24/7 at 512-D (d) Tokyo 24/7 at 4096-D
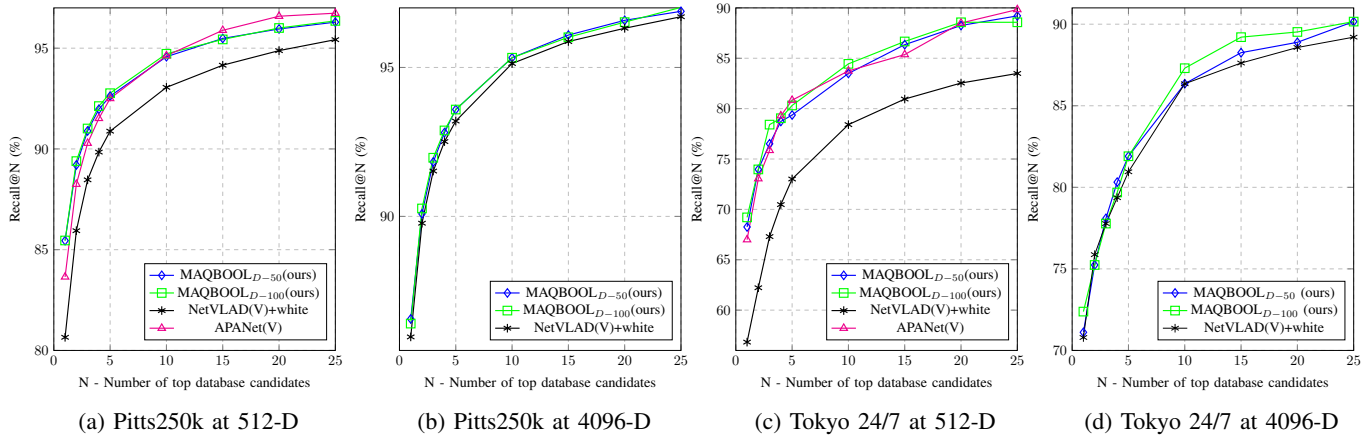
Fig. 5: Recall @ 1-25 on Toyko 24/7 and Pittsburgh 250k datasets. (a) and (b) show the performance of MAQBOOL compared to NetVLAD based on the model pre-trained on VGG Pittsburgh 250k. (c) and (d) show the performance based on the model pre-trained on VGG-16.

B. Testing on Pittsburgh and Tokyo 24/7 Datasets

Fig. 5 shows the comparison of our proposed MAQBOOL strategy while testing on the Pittsburgh and Tokyo 24/7 datasets. In comparison to the feature dimension of 4096, we achieve significant improvement at a feature dimension of 512, as shown in Fig. 5a and 5c.

The top five recall results tested on the Pittsburgh and Tokyo 24/7 datasets are shown in Fig. 7. It is shown in Fig. 8a and 8b that NetVLAD fails to retrieve the nearest match with the query in the first five places of the Tokyo 24/7 dataset, while MAQBOOL successfully adjusts the distances of the retrieved images and re-ranks the closest match to the first position. Similarly, Fig. 7c and 7d show the robustness of our proposed system compared to NetVLAD when tested on the Pittsburgh dataset at feature dimensions of 512 and 4096, respectively.

C. MAQBOOL Representation

In visual place recognition, the standard baseline is to increase the feature dimension. Our work proved that we could make it better by adding probabilistic information. SLAM systems do not recognize the place at 30/60 frames per second (FPS), instead they detect loop closure at nearly 1 second intervals. We perform landmarks-based verification that takes longer than one second. There is always a trade-off between speed and accuracy. Edge Boxes takes nearly 0.37 second in MATLAB. We used $n = 50$ regions in this work, which takes an additional 0.87 sec. Depending on the application, we can reduce the n regions. For instance, the system takes only 0.13 second instead of 0.87 sec for processing five boxes.

Let's assume we are utilizing the SLAM system on Mars, and we choose the model trained in some datasets. In that case, vanilla NetVLAD cannot perform with high accuracy. Fig. 6 shows the MAQBOOL performance on the Tokyo 24/7 datasets with 512-D and 4096-D features compared with the state-of-the-art NetVLAD. Both use the same model trained on the Pittsburgh 30k dataset and tested on the Tokyo 24/7 dataset. We achieved notable improvement at low-dimensional (512-D) feature-based recall, comparable with the 4096-D NetVLAD.

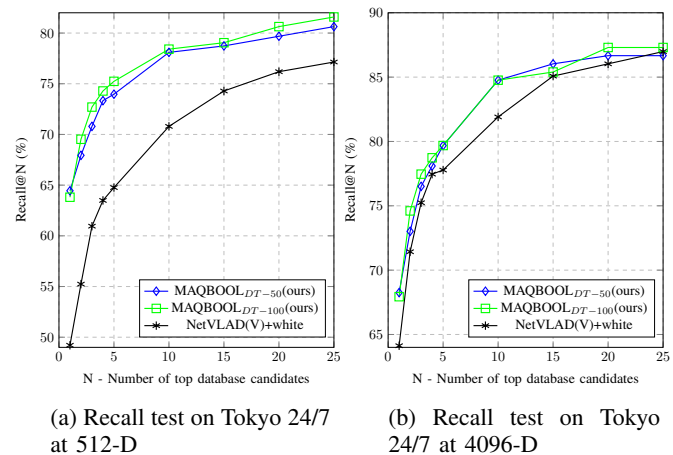(a) Recall test on Tokyo 24/7 at 512-D (b) Recall test on Tokyo 24/7 at 4096-D

Fig. 6: MAQBOOL performance comparison with NetVLAD on Tokyo 24/7 dataset. Their retrieval performances are evaluated using the same VGG model, which is trained on the Pitts30k dataset.

D. Comparison with Power Whitening PCA

While maintaining the same baseline of PCA whitening followed by NetVLAD, MAQBOOL outperforms NetVLAD as well as APANet, as shown in Table I. APANet introduces an additional PCA power whitening concept on different block types and produces better performance than NetVLAD, but by keeping the default PCA whitening, our simplest model delivers better results than APANet. We observe that by increasing the tree size, there is a significant improvement in the accuracy at a high dimension of 4096. Moreover, for a decision tree dimension of 50, MAQBOOL achieves good results compared to APANet on the Tokyo 24/7 and Pittsburgh datasets, as shown in Table I.

E. Ablation Study

As mentioned in the previous section, we use a decision tree model at the PDL. We choose the decision tree and Gaussian

TABLE I: MAQBOOL performance comparison with APANet and NetVLAD at 512-D.

Method	Whitening	Tokyo 24/7			Pitts250k-test		
		Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
Sum pooling	PCA whitening	44.76	60.95	70.16	74.13	86.44	90.18
	PCA-pw	52.70	67.30	73.02	75.63	88.01	91.75
NetVLAD [6]	PCA whitening	56.83	73.02	78.41	80.66	90.88	93.06
	PCA-pw	58.73	74.6	80.32	81.95	91.65	93.76
APANet [24]	PCA whitening	61.90	77.78	80.95	82.32	90.92	93.79
	PCA-pw	66.98	**80.95**	83.81	83.65	92.56	94.70
MAQBOOL (Ours)	PCA whitening + DT-50	68.25	79.37	83.49	85.45	92.62	94.58
	PCA whitening + DT-100	**69.21**	80.32	**84.44**	**85.46**	**92.77**	**94.72**

probability models for use in the PDT layer, and we discuss them as follows.

1) Decision tree model: We trained the default model based on '100' tree size. For the ablation study, we trained a decision tree model of size '50'. We found that by reducing the tree size, the system performance is almost similar. MAQBOOL with a decision tree size of '100' works slightly better at a higher dimension (4096-D) than MAQBOOL with a decision tree size of '50', as shown in Fig. 5. Furthermore, We find that the tree size of '50' has a better recall than NetVLAD and power-whiteing-based APANet. We choose ground truth data of the Tokyo Time Machine validation set to create the model. However, we observed that the model works better than NetVLAD if created using small datasets, such as the Oxford 5k [33] and Paris [34] datasets. These datasets have 55 query images. However, prediction models based on these datasets show similar performance.

2) Gaussian probability model: The Gaussian probability model is also a popular choice in regression studies. We observe that it has a similar performance with the decision tree of size '50'.

V. CONCLUSION

In this paper, we introduce MAQBOOL to improve the accuracy of image retrieval results without retraining a new deep learning model, for better visual place recognition. We elevate essential regions at spatial layers and probabilistically verify the image correspondence efficiently. Our MAQBOOL approach intelligently processes the high-level layer to produce more-reliable top matches than the current state-of-the-art. Without any further training or introducing additional sensors or pieces of ground truth information to the system, our framework outperforms PCA power whitening on APANet, and on NetVLAD. Our method achieves good accuracy on low-dimensional features (i.e., 512-D) with more reliable candidates, which makes it useful in general SLAM applications for loop closure detection.

VI. APPENDIX

The primary motivation behind this work is to make the multiagent SLAM systems efficacious towards new deployment. Image retrieval is a key part of not only the SLAM system but also of data analytics. This paper suggests an intelligent way to use the top landmarks for better place recognition. If we observe the recall rate of NetVLAD while tested on challenging datasets such as the Tokyo 24/7 dataset [32], Tokyo Time Machine, and Pittsburgh dataset [31]. We found that model should be trained using the same dataset to achieve a good recall rate. Moreover, the top 10-25% candidates from the database have recall rates with accuracy 90% and above. Generally, SLAM systems take the top first candidate from the retrieval for loop closure detection. It means we cannot emphasize the direct usage of NetVLAD for single-loop-closure-based multiagent system [4].

A. Evaluation on Oxford Building and Paris Building datasets

In Fig. 8, NetVLAD failed to put the right matches at the first position for each query image from Oxford 5k building and Paris 6k building datasets. Our work at the lower feature dimension, i.e., 512-D, successfully places the right matches to the first position. That makes it more robust towards using any mapping system. For the localization system, the first recall is very important for a complete global mapping optimization.

B. Recall Improvement at a Lower Features Dimension

Fig. 1 show the quantitative results of MAQBOOL at 512-D compared to NetVLAD at 512-D and 4096-D. In Fig. 9 and 10, we show qualitative results of our approach compared with the recall of NetVLAD. We found that MAQBOOL at 512-D outperformed the NetVLAD at 4096-D for the low light images. For the query image shown in Fig. 9, our MAQBOOL at 512-D with DT-50, not only correct the first match as NetVLAD at 4096-D did, but also it brings the correct match at the third position.

REFERENCES

[1] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[2] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015. [Online]. Available: http://arxiv.org/abs/1502.00956http://dx.doi.org/10.1109/TRO.2015.2463671http://ieeexplore.ieee.org/document/7219438/

[3] T. Schneider, M. Dymczyk, M. Fehr, K. Egger, S. Lynen, I. Gilitschenski, and R. Siegwart, "Maplab: An open framework for research in visual-inertial mapping and localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1418–1425, 2018.

[4] M. U. M. Bhutta, M. Kuse, R. Fan, Y. Liu, and M. Liu, "Loop-Box: Multiagent direct SLAM triggered by single loop closure for large-scale mapping," *IEEE Transactions on Cybernetics*, 2020.

(a) Recall test on Tokyo 24/7 dataset at a features dimension of 512.

(b) Recall test on Tokyo 24/7 dataset at a features dimension of 4096.

(c) Recall test on Pitts250k dataset at a features dimension of 512.

(d) Recall test on Pitts250k dataset at a features dimension of 4096.

Fig. 7: MAQBOOL vs. NetVLAD: (a) and (b) show the image retrieval compared to NetVLAD on the Tokyo 24/7 dataset with feature dimensions 512 and 4096. (c) and (d) show the recall on the Pittsburgh dataset with feature dimensions 512 and 4096.

[5] M. U. M. Bhutta and M. Liu, "PCR-Pro: 3D Sparse and Different Scale Point Clouds Registration and Robust Estimation of Information Matrix For Pose Graph SLAM," in *2018 IEEE 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, 2018, pp. 354–359. [Online]. Available: http://arxiv.org/abs/1808.09693

[6] 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.

[7] I. Melekhov, A. Tiulpin, T. Sattler, M. Pollefeys, E. Rahtu, and J. Kan-

nala, "DGC-Net: Dense geometric correspondence network," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1034–1042.

[8] P. Panphattarasap and A. Calway, "Visual place recognition using landmark distribution descriptors," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 487–502.

[9] Z. Xin, Y. Cai, T. Lu, X. Xing, S. Cai, J. Zhang, Y. Yang, and Y. Wang, "Localizing discriminative visual landmarks for place recognition," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5979–5985.

[10] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell,

(a) Recall test at a features dimension of 512.

(b) Recall test at a features dimension of 512.
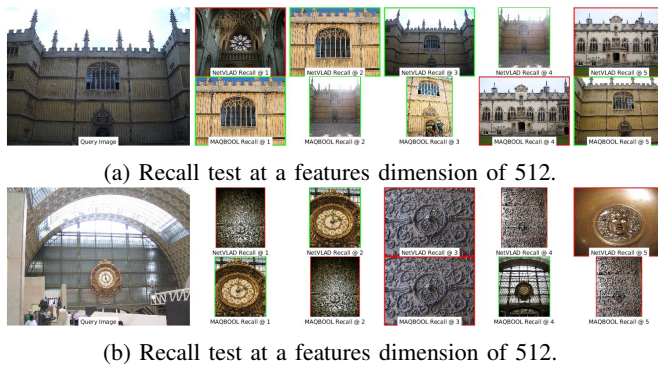
Fig. 8: MAQBOOL vs. NetVLAD: (a) Evaluation is carried out of the Oxford 5k [33] dataset (b) Evaluation is carried out of the Paris 6k [34] dataset.
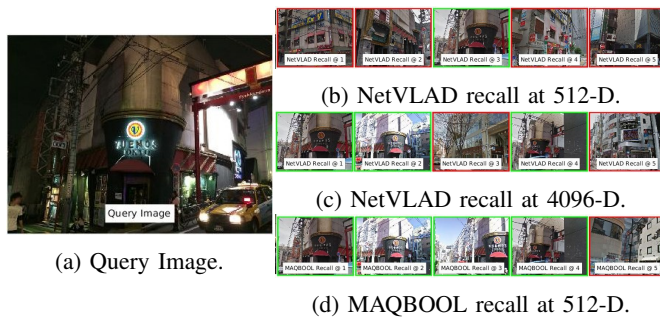

(a) Query Image.

(b) NetVLAD recall at 512-D.

(c) NetVLAD recall at 4096-D.

(d) MAQBOOL recall at 512-D.

Fig. 9: Extended results of Fig. 1(a). Query image is taken from Tokyo 24/7 dataset. (b), (c) and (d) show the first five recalls from the database.

(a) Query Image

(b) NetVLAD recall at 512-D.

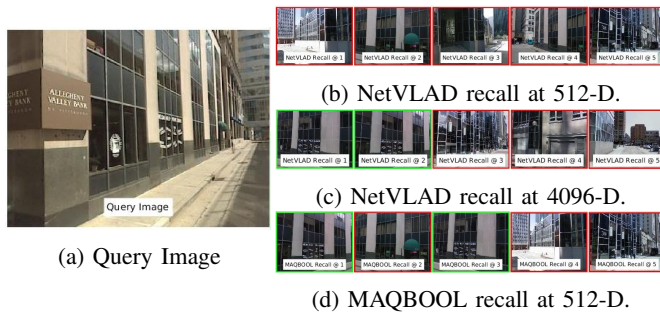(c) NetVLAD recall at 4096-D.

(d) MAQBOOL recall at 512-D.

Fig. 10: Extended results of Fig. 1(b). Query image is taken from Pittsburgh dataset. (b), (c) and (d) show the first five recalls from the database.

B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," *Robotics: Science and Systems*, vol. 11, 2015.

[11] F. Lu, B. Chen, X.-D. Zhou, and D. Song, "STA-VPR: Spatio-Temporal Alignment for Visual Place Recognition," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4297–4304, 2021.

[12] Q. Zhai, H. Cheng, R. Huang, and H. Zhan, "Place clustering-based feature recombination for visual place recognition." *arXiv Computer Science*, pp. 1–9, 2019. [Online]. Available: http://arxiv.org/abs/1907. 11350

[13] Z. Chen, A. Jacobson, N. Sunderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 1, pp. 3223–3230, 2017.

[14] Z. Laskar, I. Melekhov, H. R. Tavakoli, J. Ylioinas, and J. Kannala, "Geometric image correspondence verification by dense pixel matching,"

[15] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. Decoste, W. Di, and Y. Yu, "HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 2740–2748, 2015.

[16] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 2, pp. 661–674, 2019.

[17] L. G. Camara and L. Přeučil, "Spatio-semantic ConvNet-based visual place recognition," in *2019 European Conference on Mobile Robots (ECMR)*. IEEE, 2019, pp. 1–8.

[18] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 716–12 725.

[19] A. Gawel, C. Del Don, R. Siegwart, J. Nieto, and C. Cadena, "X-View: Graph-based semantic multiview localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1687–1694, 2018.

[20] L. Bernreiter, A. Gawel, H. Sommer, J. Nieto, R. Siegwart, and C. C. Lerma, "Multiple hypothesis semantic mapping for robust data association," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3255–3262, 2019.

[21] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "Inloc: Indoor visual localization with dense matching and view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7199–7209.

[22] S. Pillai and J. Leonard, "Self-supervised visual place recognition learning in mobile robots," *arXiv preprint arXiv:1905.04453*, 2019.

[23] A. Pal, C. Nieto-Granda, and H. I. Christensen, "DEDUCE: Diverse scene detection methods in unseen challenging environments," *arXiv preprint arXiv:1908.00191*, 2019. [Online]. Available: http: //arxiv.org/abs/1908.00191

[24] Y. Zhu, J. Wang, L. Xie, and L. Zheng, "Attention-based pyramid aggregation network for visual place recognition," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 99–107.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[27] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision- ECCV 2014*, Springer. Springer International Publishing, 2014, pp. 391–405.

[28] N. V. Keetha, M. Milford, and S. Garg, "A hierarchical dual model of environment- and place-specific utility for visual place recognition," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6969–6976, 2021.

[29] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 842–850.

[30] Y. Yan, B. Ni, and X. Yang, "Fine-grained recognition via attribute-guided attentive feature aggregation," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 1032–1040.

[31] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla, "Visual place recognition with repetitive structures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2346–2359, 2015.

[32] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 257–271, 2018.

[33] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 2007, pp. 1–8.

[34] ——, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.