ETH zürich

ETH Library

Learning Context Flexible Attention Model for Long-Term Visual Place Recognition

Journal Article

Author(s):
Chen, Zetao; Liu, Lingqiao; Sa, Inkyu; Ge, Zongyuan; Chli, Margarita

Publication date:
2018-10

Permanent link:
https://doi.org/10.3929/ethz-b-000282829

Rights / license:
In Copyright - Non-Commercial Use Permitted

Originally published in:
IEEE Robotics and Automation Letters 3(4), https://doi.org/10.1109/lra.2018.2859916

Funding acknowledgement:
157585 - Collaborative vision-based perception for teams of (aerial) robots (SNF)
644128 - Collaborative Aerial Robotic Workers (SBFI)
644227 - Aerial Data Collection and Analysis, and Automated Ground Intervention for Precision Farming (SBFI)

This page was generated automatically upon download from the ETH Zurich Research Collection.
For more information, please consult the Terms of use.

Learning Context Flexible Attention Model for Long-term Visual Place Recognition

Zetao Chen, Lingqiao Liu, Inkyu Sa, Zongyuan Ge and Margarita Chli

***Abstract*— Identifying regions of interest in an image has long been of great importance in a wide range of tasks, including place recognition. In this paper, we propose a novel attention mechanism with flexible context, which can be incorporated into existing feed forward network architecture to learn image representations for long-term place recognition. In particular, in order to focus on regions that contribute positively to place recognition, we introduce a multi-scale context-flexible network to estimate the importance of each spatial region in the feature map. Our model is trained end-to-end for place recognition and can detect regions of interest of arbitrary shape. Extensive experiments have been conducted to verify the effectiveness of our approach and the results demonstrate that our model can achieve consistently better performance than the state-of-the-art on standard benchmark datasets. Finally, we visualize the learned attention maps to generate insights into what attention the network has learned.**

***Index Terms*— Localization, Deep Learning in Robotics and Automation, Visual-Based Navigation.**

I. INTRODUCTION

PLACE recognition involves estimating a match between the query and the previously visited places [1]. Long-term visual place recognition poses a fundamental challenge in robot navigation because a single place can undergo significant appearance changes due to illumination, weather or seasonal changes. Recently, the success of deep learning in computer vision has triggered a range of investigations into how to generate a feature representation from Convolutional Neural Networks (CNNs) that is robust to such variations [2, 3]. In this paper, we investigate learning attention for place recognition using CNNs in an end-to-end fashion.

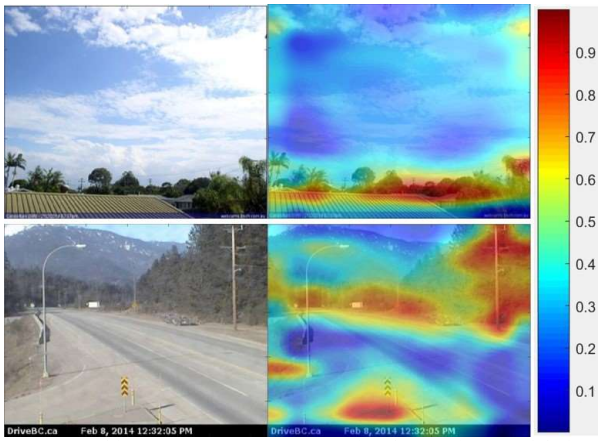

Figure 1 In this paper, we propose a visual attention model for place recognition, which learns to automatically focus on regions that are most discriminative in defining a place. Two examples are shown here where the left is the original image and the right is the super-imposed heat map

Attention plays a crucial role in the human perception process and finding regions of interest relevant to the task has long been of great importance in computer vision and robotics communities. There have been attempts in utilizing attention in many tasks, such as image classification [4], image retrieval [5] or segmentation [6], while attention can also play an important role in visual place recognition [7]. Visual cues that are relevant to place recognition are generally not uniformly distributed across an image, therefore focusing on important regions, as opposed to irrelevant or confusion areas, is key to improve the place recognition performance. For example, when recognizing a street scene, using features extracted from time-varying objects, such as moving cars or pedestrians, as opposed to those extracted from static structures, such as buildings or road signs, can introduce misleading information into place recognition.

To address this problem, there have been attempts to automatically select informative regions for recognizing places. The approaches proposed in [8, 9] first identify salient regions and then extract features to match image patches over extreme condition- and viewpoint-variations. However, these methods interpret saliency selection and feature extraction as two separate processes, while our model can optimize both processes simultaneously in an end-to-end fashion. The approach in [10] leverages on high-level information from semantic segmentation to restrict features extracted only from man-made structures, such as buildings. However, this manually-defined attention is limited and is not flexible enough to capture all relevant context.

Manuscript received: February 24, 2018; Revised: April 30, 2018; Accepted: July 3, 2018. This paper was recommended for publication by Editor Dongheui Lee upon evaluation of the Associate Editor and reviewers' comments. This research was supported by the Swiss National Science Foundation (SNSF, Agreement no. PP00P2 157585), EC's Horizon 2020 Programme under grant agreement n. 644128 (AEROWORKS), the European Union's Horizon 2020 research and innovation Programme under grant agreement No 644227 (Flourish) and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 15.0029.

Z. Chen and M. Chli are with the Vision for Robotics Lab, ETH Zurich, Switzerland, chenze@ethz.ch; chlim@ethz.ch.

L. Liu is with school of Computer Science, University of Adelaide, Australia, lingqiao.liu@adelaide.edu.au.

I. Sa is with the Autonomous Systems Lab, ETH Zurich, Switzerland, inkyu.sa@mavt.ethz.ch.

Z. Ge is with E-research Centre, University of Monash, Australia, Zongyuan.Ge@monash.edu.

Digital Object Identifier (DOI): See top of this page.

Context provides crucial cues in determining a region's importance. For example, a painting on a building may be useful for place recognition, while the same one on a bus can be misleading. A range of investigations has been conducted to learn context-based attention maps automatically for place recognition. Similarly to our work, the approach in [7] proposes to learn an attention model to weight the importance of each spatial location in the feature map. However, the contextual area around each region was manually defined as a rectangular bounding box. We show in the experiments that this manually defined local contextual region is not powerful enough to capture all relevant context. Instead, here, the attention for each spatial region is estimated automatically by a global adaptive contextual area, whose shape depends only on image features.

Inspired by recent studies which show that features extracted from different layers of CNNs capture different semantic structures [11-13], in this paper we also estimate the attention map using multi-semantic contexts. Features from lower convolutional layers generally respond to low level image features, such as corners or edges, while those from higher layers focus on structures that are more semantically meaningful, such as human faces or buildings. Our attention map is estimated by fusing context computed from different convolutional layers in order to capture multi-level semantics.

The main contributions of this work are as follows:

- we propose a novel contextual-flexible attention model that is able to adaptively select relevant contexts for each target attention region, as opposed to the fixed contexts used by most other approaches,
- we conduct intensive (ablation) experiments on three public benchmarks, and achieve consistently better performance than other state-of-the-art approaches,
- we introduce a new benchmark dataset called SPEDTEST for long-term place recognition.

II. RELATED WORK

In this section, we briefly review previous approaches on CNNs for place recognition, attention-based place recognition and multi-scale contexts in deep neural networks.

A. Visual Place Recognition with CNNs

The first step in visual place recognition is to derive image representation that is discriminative in defining a place. Previous approaches either operate directly on raw pixels [14] or utilize a fixed set of handcrafted descriptors [15-18]. Recently, inspired by the success of deep learning in computer vision, a range of studies have been conducted on the applicability of deep learning for place recognition. [19] was the first work to introduce CNNs for visual place recognition and [20] conducted a detailed analysis of the utility of deep learnt features for place recognition. However, these studies utilized networks that were pre-trained on tasks different in nature from place recognition. Later, [2, 7, 13] trained CNN models particularly for place recognition and demonstrated that doing so can further improve the performances. In [21], a CNN was trained for a different but relevant task of camera pose estimation. The aforementioned approaches generally define a manual context region to calculate the attention mask. In this paper, we propose a context flexible attention-modulated CNN model particularly for place recognition.

Rather than recognizing specific places, networks have also been trained for recognizing the types of places [22]. However, this scene recognition task is different in nature from visual place recognition: image under the same scene category can come from different places.

B. Attention Model for Place Recognition

Attention, or so called saliency detection, has been shown to play an important role in a wide variety of computer vision and robotics tasks [4, 6, 23-25]. Despite their different application scenarios, such works utilize a neural network to learn to automatically locate task-relevant regions.

Attention also plays an important role in visual place recognition. This is because not all content in the image is relevant in representing a place, and identifying salient regions, as opposed to useless or misleading areas, is crucial to improve the recognition performance. The approach in [10], for example, leverages on supervised information from semantic segmentation to extract features only from man-made structures, such as buildings or roads. Despite its simplicity and efficiency, this manually defined attention, as opposed to the learning-based approach we propose in this paper, is not flexible enough to define all relevant regions. Similarly, in [26] the attention mechanism was explored to identify salient landmarks for mobile robot localization. Approaches in [8, 9] identify salient regions using an external landmark detector. However, their landmark detectors are based on networks that are trained on tasks different in nature from place recognition. In contrast, our attention model is trained end-to-end specifically for place recognition.

Different approaches have been proposed to adaptively estimate a task-dependent attention map. In particular, the context information, which is generally addressed by selecting image areas surrounding the localized, target image regions [11], have been shown to provide useful cues to estimate the attention. However, the contextual areas around each target region are usually defined manually in the shape of rectangular bounding boxes [6, 7, 11, 27]. Such manually defined focus areas often may be suboptimal, since the actualy contextual area can vary under different situations. In this paper, we propose a context estimation mechanism, which can adaptively define contextual areas in any shape, depending only on the image features.

C. Multi-scale Features in Deep Networks

Multi-scale features have been shown to provide effective use of context information [28-30]. In the context of deep neural networks, multi-scale features refer to activations extracted from different layers of the networks. It has been shown that features extracted from lower layers most often correspond to low-level image features, such as edges or corners, while activations from higher layers respond to more semantically meaningful shapes, such as trees or faces [13]. Inspired by this success of multi-scale features [6, 11, 31], here we propose to learn attention by fusing context extracted from multiple layers of the network.

III. METHOD

In this section, we present the details of our multi-scale attention learning system. We first describe how local features can be extracted from a convolutional feature map to represent an image region at a spatial location. Then we illustrate our global adaptive context system, which is used to estimate the attention score at each spatial location of the image. We demonstrate how we learn attention at different convolutional layers and in the end, we discuss how to automatically merge the attention from each layer to a final attention mask. The schematic illustration of the proposed system is shown in Figure 2.

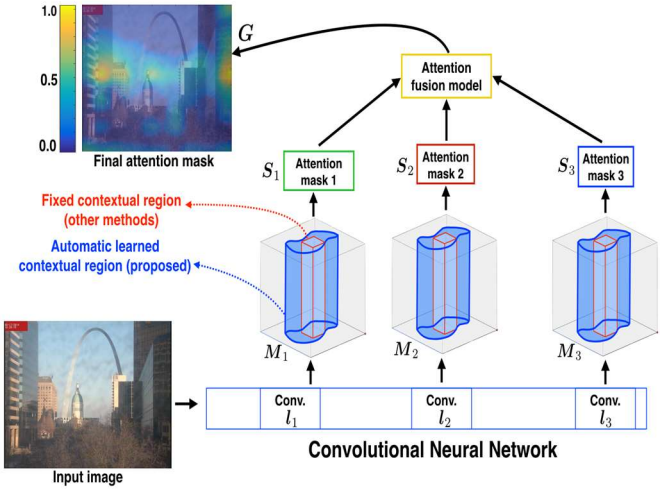

Figure 2 Schematic illustration of our attention learning system. We propose a multi-scale context-flexible attention learning model. Attention is estimated from context information extracted at different layers of the network and then the multi-scale attention maps (S_1, S_2, S_3) are fused to generate a final attention mask G.

A. Local Feature Extraction from Convolutional Activations

Given a pre-trained CNN, our attention-learning model takes as inputs its convolutional activations and outputs an attention mask to weight the importance of each spatial location. Similar to [9], we treat activations at a certain convolutional layer as a tensor of size $\mathrm{H} \times \mathrm{W} \times \mathrm{C}$, which is considered as a set of C-dimensional local features in $\mathrm{H} \times \mathrm{W}$ spatial location. Formally, the convolutional layer activations $\mathrm{X} \in R^{H \times W \times C}$ can be expressed as:

$$\mathrm{X} = \{x_i \in R^C | \mathrm{i} \in \{1, \dots, \mathrm{H} \times \mathrm{W}\}\}. \quad (1)$$

B. Adaptive Global Context Feature

We propose the fusion of local descriptor and global latent context to estimate the attention score at each spatial location. For each local descriptor x_i created in the previous step, a corresponding global latent context $c_i, i \in \{1, \dots, H \times W\}$ is created by adaptively pooling all local descriptors, such that:

$$c_i = \sum_j r_j^i x_j, i = 1, \dots H \times W,$$
$$r_j^i = {x_i}^T \Omega x_j, \quad (2)$$

where c_i has the same dimension as x_i, r_j^i is a scalar that measures the importance of feature x_j when constructing the context feature c_i. When calculating the relevance r_j^i, we construct the correlation matrix: $\Omega \in R^{C \times C}$, which is a trainable matrix that can be automatically adapted to relate x_i and x_j. Compared to many other existing attention learning approaches, which manually define the contextual region as rectangular bounding boxes [6, 7, 11, 27] (denoted in red in Figure 2), our approach adopts r_j^i to adaptively learn the contextual area for each spatial region. In theory, this can define an adaptive contextual area of any shape. We argue that this is a more powerful context than a manually defined rectangular bounding box. Considering the case of looking at the corner of a window, it is when the surrounding areas are observed that we know that this is actually a corner of a window and can potentially be discriminative when recognizing a place. However, if we can have a global context, which tells us that there are many windows in the image with the same structure, the saliency of such corners gets reduced.

C. Saliency Estimation by Observing Local Appearance and Global Context

While x_i denotes the local appearance at the i^{th} location, c_i represents the awareness of its global context. By looking at both its local appearance and its global context, the attention mask can be estimated more accurately. This process is also more similar to the human perception process, which uses the global high-level context information to guide the attention spent on each region. We construct another context feature map $\mathrm{C} \in R^{H \times W \times C}$ from all the context feature c_i, such that:

$$\mathrm{C} = \{c_i \in R^C | \mathrm{i} \in \{1, \dots, \mathrm{H} \times \mathrm{W}\}\}. \quad (3)$$

Since C and X have the same dimension, we can merge them together either by element-wise summation or by concatenating them. Concatenation allows for more flexibility but requires more memory for training. We experimented both and found that both methods achieve relatively similar performance. As a result, we adopt the element-wise summation approach, which is more memory-efficient, to create a fused feature map M:

$$\mathrm{M} = \mathrm{C} + \mathrm{X}. \quad (4)$$

Based on the fused feature map, we apply an attention model to learn a soft attention mask over all spatial locations. The attention model is parameterized by a CNN, which takes the feature map M as an input. Formally:

$$\mathrm{T} = \mathrm{f}(\mathrm{K} * \mathrm{M} + \mathrm{b}), \quad (5)$$

where $*$ denotes the convolution operator, K represents the convolution filter, b is the bias, and f is the ReLU nonlinear activation function. Since we use a 3×3 kernel and output only a single channel, K is in the shape of $\mathrm{C} \times 3 \times 3 \times 1$. The output $\mathrm{T} \in R^{H \times W}$ is the soft attention mask that summarizes information of all channels in M.

Assume that T(i) denotes the i^{th} element in T, where i $\in \{1, \dots, H \times W\}$. A softmax operation is applied to T spatially as follows:

$$S(i) = \frac{e^{T(i)}}{\sum_{l' \in L} e^{T(l')}}, \quad (6)$$

where $L \in \{1, \dots, H \times W\}$ and $\sum_i S(i) = 1$. S is the soft attention map learned from the layer where the feature map X is extracted and will be later applied to reweight the feature in X.

D. Multi-scale Attention Fusion

It is observed that activations from early convolutional layers generally fire at lower level features, such as at edges or corners, while feature maps at later convolutional layer are more selective to semantically meaningful structures, such as shapes or objects [13]. To improve the robustness of our learned features, we propose to learn an attention model from different layers of the network and to merge them together softly. In particular, we pick three layers l_1, l_2 and l_3 from a pre-trained network. l_3 is chosen to be the last convolutional layer, while l_1 and l_2 are the two earlier ones. As illustrated in section III.C, we construct M_1, M_2 and M_3 and calculate the soft attention mask S_1, S_2 and S_3, respectively from layers l_1, l_2 and l_3. S_1 and S_2 are resized to have the same resolution as S_3. We then apply a $1 \times 1 \times 3$ convolutional layer to compute a weighted sum of these three masks:

$$G = \sum_{k=1}^{3} w_k S_k + b_k, \quad (7)$$

where w_k and b_k are learnable parameters that denote the convolution weight and the bias applied on the attention mask of the k^{th} layer, respectively, while G is the final attention mask of size $H \times W$, whose value indicate which spatial region of the feature maps are important.

E. Final Feature Representation

In order to generate the final representation, we apply the final attention mask G to modulate the activations from the l_3 convolutional layer X_{l_3}:

$$F(p) = G \circ X_{l_3}(p), \quad p = 1, \dots, C, \quad (8)$$

where p is the index of feature channel and $\circ$ denotes the channel-wise Hadamard matrix product here. $F \in R^{H \times W \times C}$ denotes the feature map after attention modulation. F will be the final representation of the images used during the testing.

Similarly to the approach in [13], we formulate the place recognition task as a classification problem, which has been demonstrated to deliver superior performance. In particular, we train our attention model on a subset collected from the SPED dataset introduced in [13], which consists of images collected from 1136 webcameras around the world. Images captured by the same camera are assigned to the same label. F then goes through another convolutional and fully-connected layer, followed by a $1136-$ way softmax layer to learn a correct label output.

IV. EXPERIMENTAL SETUP

This section describes the benchmark datasets used in the experiment and the acquirement of their ground truth. We also discuss the implementation details of our attention learning model.

A. Training Dataset

The attention learning model is trained on the SPED dataset introduced in [13]. In particular, we manually selected 1136 different cameras from the whole dataset and make sure images generated from these cameras are of good quality, for example, excluding corrupted images or images that are completely black. These cameras are located at different places around the world and the images capture a wide variety of environments. Each camera captured one picture every half an hour for a period of ten years. For our testbed, we randomly select 1000 images from each camera, captured in February 2014 and August 2014, with these time points chosen specifically to exhibit dramatic seasonal and illumination changes. Amongst them, 800 were used for training and the other 200 for validation. Since there are 1136 different cameras, we formulate our task as a $1136-$ class classification problem.

B. Testing Datasets

We evaluated the effectiveness of our proposed system on three benchmarking place recognition datasets. These datasets capture a range of different environments, exhibiting different degrees of variations in conditions and viewpoints. These variations represent typical challenges that a robot can encounter in the real-world. Details of all the datasets are summarized in Table 1, while some example images are illustrated in Figure 3. Each dataset consists of two traverses along the same route with the first traverse used as reference and the second one used for testing.

The St. Lucia dataset [32] was recorded in the suburbs of St. Lucia at different times of day and exhibits medium viewpoint and significant illumination changes. The Synthesized Nordland dataset [33] was captured by a camera mounted on a train. The first traverse was captured in spring and the second one during winter. Because the original dataset exhibits no viewpoint changes, we synthesize the viewpoint variations by cropping the images, such that images from the first and second traverse have an overlap of 75%. To construct the SPEDTEST set, we first manually select 668 new cameras from the SPED dataset that are not used for the training and then for each camera, we pick one image from the winter as the query and three other images from the summer as the reference dataset. We manually verify the images to make sure they all exhibit significant condition and moderate viewpoint variations. Some example images of this newly constructed dataset can be seen in Figure 3. The authors plan to make this dataset public available to the community in the near future.

C. Ground Truth

For the SPEDTEST set, the ground truth was built by manually parsing the frames and building the frame-level correspondence. For the St. Lucia dataset, we used the GPS annotations provided with the original dataset to build a coarse correspondence followed by a manual step to build more accurate frame-level correspondence. For the Synthesized

Nordland dataset, we used the frame-level correspondence provided with the original dataset.

D. Implementation Details

We find that fine-tuning from a pre-trained network can achieve better performance than training from scratch; therefore, we employed the VGG16 network [34] as our pre-trained network to learn the attention, although other networks, such as ResNet [35], GoogleNet [36] or AlexNet [37] can also be utilized. Each image is first resized to 224×224 before it is fed to CNN for training. During training, we set the learning rate of the pre-trained network to 0.00008 and the learning rate of our newly added attention model to 10 times larger (i.e. 0.0008). We find that this dual learning rate mechanism can achieve better performance than setting a uniform learning rate for all the weights. We chose a batch size of 64 and the learning rate drops by a factor of 10 at a rate of every 30000 iterations. We stop training after 80000 iterations, because no significant performance gain was observed after that. The parameters are set once and used across all experiments.

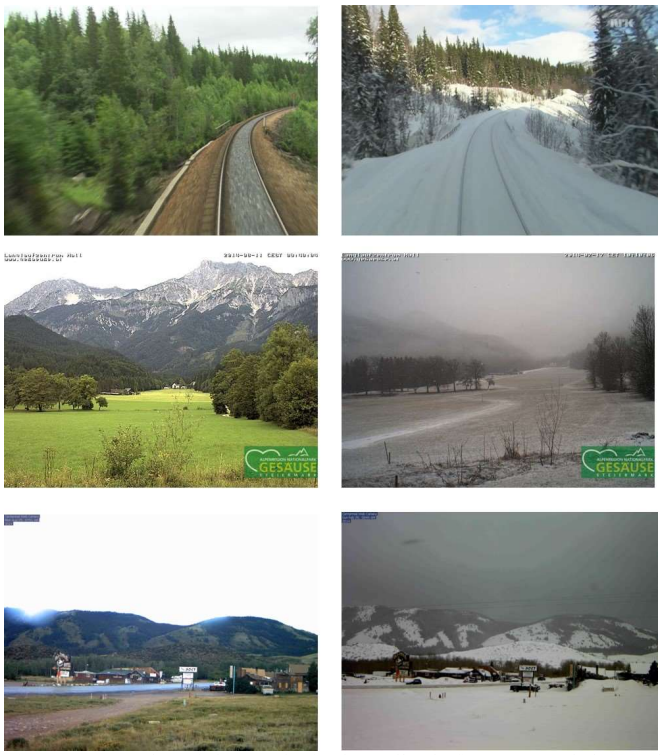

Figure 3 Examples of the Synthesized Nordland (first row) and the newly constructed SPED test set (second and third row). Each row depicts the same place under different conditions.

Dataset	No of frames	Environment	Viewpoint variation	Condition variation
Synthesized Nordland	1800	Train journey	Moderate	Strong
St. Lucia	2557	Suburban	Moderate	Moderate
SPEDTEST	2005	Very diverse	Moderate	Strong

Table 1 DATASET DESCRIPTIONS

E. Comparison Methods

We first compare our full model ("Ours-full") with different ablated versions in Section V.A to investigate the impact of different components in our attention model:

- "No Finetune":We directly take activations from the same convolutional layer of the pre-trained network (which is VGG16 [34] in our case) without any fine-tuning.
- "Finetune+NoAttent": We fine-tune the pre-trained network on our SPED training dataset. This approach performs fine-tuning but no attention model is incorporated.
- "Finetune+Attent+NoContext": We apply attention model when fine-tuning the network. However, the attention score of each spatial location is estimated without looking at its global context vector as defined in (3).

We further compare our method with different state-of-the-art attention models and place recognition algorithms in Section V.B:

- Attentive Attention (AA) [38]: This is a recent attention-based descriptor that achieves state-of-the-art performances on several image retrieval datasets. For fair comparison, we fine-tune the model on our training dataset and remove the geometric verification stage.
- Fixed Context (FC) [7]: This attention-modulated descriptor utilizes a fixed bounding box contextual region when estimating each target attention region. By comparing with this method, we can demonstrate the superiority of having a flexible contextual region as proposed in our approach. For fair comparison, we also fine-tune this model on our training dataset. This method is a state-of-the-art CNN model for visual place recognition.
- Cross Pooling (CP) [39]: This method utilizes a late convolutional layer as fixed attention masks to generate a global descriptor for image classification and retrieval. By comparing with this method, we demonstrate the superiority of our learned attention models over the pre-defined attention masks from a convolutional layer.
- FABMAP [1]: This is a state-of-the-art place recognition algorithm built on top of handcrafted features
- SeqSLAM [14]: This is a sequence-based place recognition approach, which have demonstrated state-of-the-art performances on mapping environments across seasons, weather conditions and different times of a day.
- Places365 [22]: This is a CNN-based scene recognition model trained to recognize 365 scene types. The model was trained on over two million pictures. We extracted activations before the softmax layer as the image representation

V. RESULTS

Our proposed attention-learning model is tested on a variety of scenarios and evaluated against state-of-the-art approaches recording performance in terms of precision and recall.

Furthermore, the learned attentions are visualized in an attempt to provide insights about what networks have learned.

A. Ablation Experiments

In Figure 4~6 (left), the precision-recall (PR) curves are computed on all three test datasets for our proposed attention learning approach against other ablated baselines.

Figure 4 (left) presents the PR curves generated by these methods on the Nordland dataset. It is evident that our approach achieves consistently better performance than all other approaches. The approach (blue dotted) that is fine-tuned with attention model but without context, achieves better performance than those that either directly extract convolutional activations from the pre-trained network (dashdot), or perform fine-tuning without attention mechanisms (dashed), indicating the benefits of using attention models.

Figure 5 (left) presents results on the St. Lucia dataset with our full model (attention+context) still outperforming all other approaches by a small margin. A similar tendency is observed that fine-tuning using attention models but without context (blue dotted) achieves the second best performance. It is also interesting to see that fine-tuning without attention (dashed) does not improve the performance over the one without any fine-tuning (dashdot), indicating the significance of using attention models for improving features' robustness against condition variations.

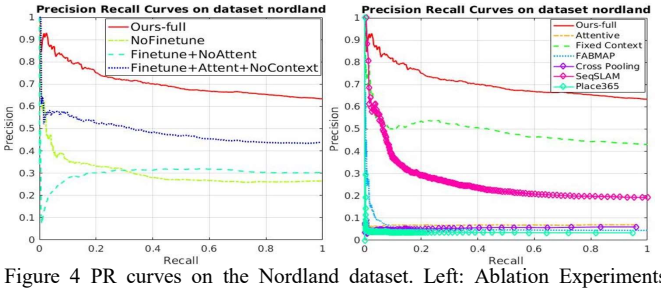

Figure 4 PR curves on the Nordland dataset. Left: Ablation Experiments; Right: Compare with state-of-the-art.

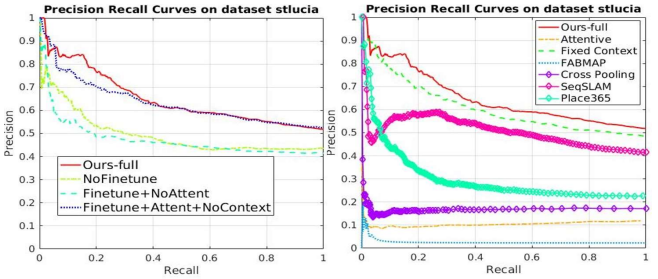

Figure 5 PR curves on the St. Lucia dataset. Left: Ablation Experiments; Right: Compare with state-of-the-art

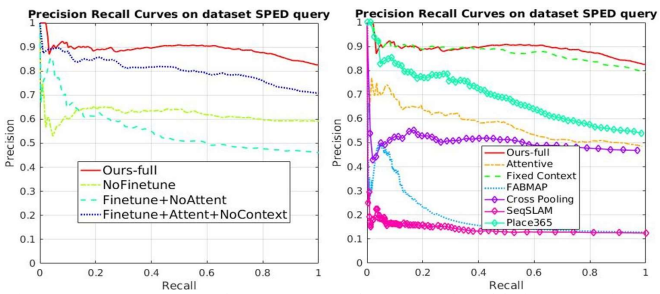

Figure 6 PR curves on the SPED test dataset. Left: Ablation Experiments; Right: Compare with state-of-the-art

Figure 6 (left) illustrates the PR curves on the newly constructed SPEDTEST dataset. Note that even though our network is trained on a subset of the SPED dataset, images in the SPEDTEST dataset are constructed from completely different scenarios and can be used to test the generalization performance of our model. As illustrated in this figure, our full model (attention+context) achieves slightly better performance than the one using attention but without context (blue dotted), indicating the benefits of using context in attention estimation. Besides, fine-tuning on the training dataset without attention (dashed) slightly improves the performance over the one that directly extracts activations from the pre-trained network without any fine-tuning (dashdot). A big gap between the blue dotted and green dashed lines illustrates the benefits of using attention models in feature modulation.

In summary, all the results in the ablation experiments indicate both the benefits of using attention models in feature learning and the context in attention estimation.

B. Comparing with State-Of-The-Art

In Figure 4~6 (right figures), we compared our full model with six other algorithms, including state-of-the-art place recognition approaches, image retrieval methods and scene recognition model (details in Section 0).

In all three results, our method (Ours-full), which utilizes our proposed automatically learned contextual regions to estimate the attention, consistently outperforms the approach that uses a fixed pre-defined contextual region [7] (green dashed). This clearly illustrates the benefits of having a flexible contextual region over a fixed one.

FABMAP [1], a state-of-the-art place recognition algorithm, achieves the worst performance in all three experiments. A closer investigation reveals that SURF [40], the handcrafted feature that FABMAP is built on top of, is not robust against strong condition variations and as a result, FABMAP cannot successfully match the places. This again highlights the necessity of having a deeply learned feature (as proposed in this paper) for long-term place recognition.

Surprisingly, the "Attentive" [38] (yellow dashdot) and "Cross Pooling" [39] (purple solid diamond) models, which have been demonstrated to perform quite well in image retrieval or classification task, delivers much worse performances in the experiments. This is probably due to the fact that place recognition is different in nature from other tasks, such as image classification or retrieval, where there is always a single object occupying the largest part of the image. In place recognition, a place can be represented by multiple region elements and the contextual relationship between these elements is important to estimate their attention mask.

By comparing the approach "Fixed Context" with the method "Finetune+attent+NoContext" in Section V.A, we can observe that both methods achieve relatively similar performance, with the "Finetune+Attent+NoContext" achieves marginally superior performance on the Nordland dataset, while the "Fixed Context" performs slightly better on the St.Lucia dataset. This implies that a fixed-shaped contextual region may not be consistently beneficial when recognizing places with strong condition variations, which again highlights the significance of having a flexible context.

C. Combining Different Number of Attentional Layers

In this section, we investigate the influence of combing different number of attentional layers. Specifically, in Figure 7, we plot the Area Under the Curves (AUCs) when one, two and three attentional layers are respectively combined to estimate the attention. It is clear that on both Nordland and SPEDTEST datasets, combining more layers achieve consistently better performance, while such differences are not significant on the St.Lucia dataset. As illustrated in Figure 3 and Table 1, the Nordland and SPEDTEST datasets illustrate much stronger condition variations (across seasons) when compared to the St.Lucia dataset (morning to afternoon). This indicates that merging multiple contextual layers may have more significant benefits when there exist larger appearance changes.

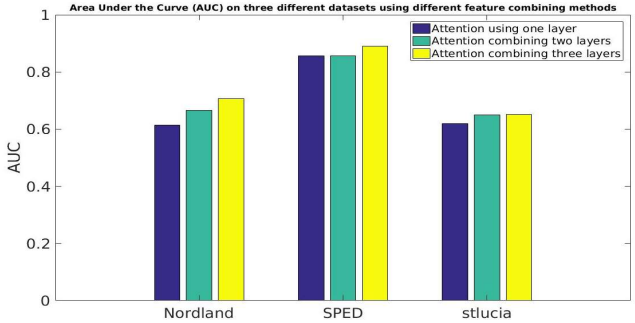

Figure 7 Area Under the Curve (AUC) between combing different number of attentional layers on three different datasets.

D. Attention Map Visualization

To gain an insight into what the attention mechanism our network has learned, we visualize eight example attention maps on unseen scenes that are generated by our model and the method that utilizes fixed context [7]. These attention masks illustrate the areas, where the network pays the strongest attention to. As illustrated in Figure 8, although these images are not used during the network training stage, our attention masks can successfully identify some salient regions in image space, while the other method usually fails. For example, in the example on the second column, the generated attention mask successfully filters out the sky and focusing only on the bottom region of the image, where there are many houses and they are more robust against long-term appearance variations.

VI. Conclusion

Inspired by the success of attention models on other computer vison tasks and the recent boom in deep learning techniques, in this paper, we propose a multi-scale context-flexible attention model for long-term place recognition. Our attention mask is estimated from context information extracted from multiply layers of the network. In contrast to most existing approaches that manually define contextual regions of pre-defined shapes, here we simultaneously learn the attention and contextual shape for a particular place in an end-to-end fashion. The learned attention mask is then used to modulate the feature maps from the original network. Evaluation against the state of the art on several benchmarking datasets reveals the superior performance of our proposed method in place recognition against strong viewpoint and condition variations.

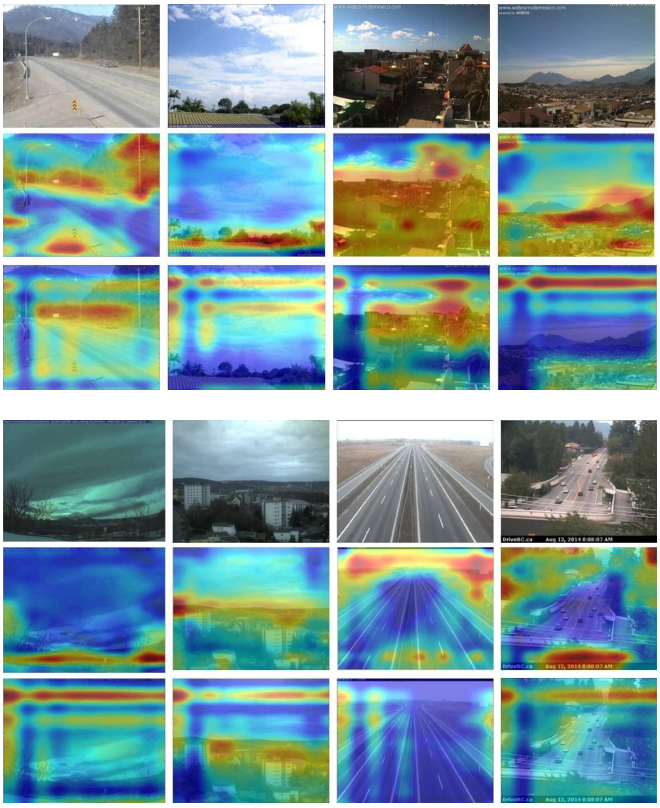

Figure 8 Input images (top row) and the attention maps learned from our model (middle row) and the model that uses fixed context [7] (bottom row) . Note that all these images used here do not appear during the network training stage.

In the future work, we will exploit the use of high-level semantic information, from either segmentation or an object detection pipeline, to improve the end-to-end attention learning system. It is expected that items with similar semantic lab should have similar attention score.

Acknowledgement: We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

REFERENCES

[1] M. Cummins and P. Newman, "Highly scalable appearance-only SLAM - FAB-MAP 2.0," in *Robotics: Science and Systems*, Seattle, United States, 2009.

[2] 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.

[3] Z. Chen, A. Jacobson, N. Sunderhauf, B. Upcroft, L. Liu, C. Shen, *et al.*, "Deep Learning Features at Scale for Visual Place Recognition," in *International Conference on Robotics and Automation (ICRA)*, 2017.

[4] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, *et al.*, "Residual Attention Network for Image Classification," *arXiv preprint arXiv:1704.06904,* 2017.

[5] P. Wang, L. Liu, C. Shen, Z. Huang, A. van den Hengel, and H. T. Shen, "Multi-attention Network for One Shot Learning," in *Proceedings of CVPR*, 2017.

[6] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3640-3649.

[7] H. J. Kim, E. Dunn, and J.-M. Frahm, "Learned Contextual Feature Reweighting for Image Geo-Localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[8] N. Sunderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft*, et al.*, "Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free," *Proceedings of Robotics: Science and Systems XII,* 2015.

[9] 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 2017)*, 2017.

[10] A. Mousavian, J. Košecká, and J.-M. Lien, "Semantically guided location recognition for outdoors scenes," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, 2015, pp. 4882-4889.

[11] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," *arXiv preprint arXiv:1702.07432,* 2017.

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

[13] Z. Chen, A. Jacobson, N. Sunderhauf, B. Upcroft, L. Liu, C. Shen*, et al.*, "Deep Learning Features at Scale for Visual Place Recognition," in *International Conference on Robotics and Automation*, 2017.

[14] M. Milford and G. Wyeth, "SeqSLAM: Visual Route-Based Navigation for Sunny Summer Days and Stormy Winter Nights," in *IEEE International Conference on Robotics and Automation*, St Paul, United States, 2012.

[15] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *The International Journal of Robotics Research,* vol. 30, pp. 1100-1123, August 1, 2011 2011.

[16] S. Lowry and H. Andreasson, "Lightweight, Viewpoint-Invariant Visual Place Recognition in Changing Environments," *IEEE Robotics and Automation Letters,* vol. 3, pp. 957-964, 2018.

[17] C. McManus, B. Upcroft, and P. Newmann, "Scene signatures: Localised and point-less features for localisation," 2014.

[18] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2014.

[19] Z. Chen, L. Obadiah, A. Jacobson, and M. Milford, "Convolutional Neural Network based Place Recognition," presented at the Australiaian Conference on Robotics and Automation, Melbourne, Australia, 2014.

[20] N. Sünderhauf, F. Dayoub, S. Shirazi, B. Upcroft, and M. Milford, "On the Performance of ConvNet Features for Place Recognition," presented at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 2015.

[21] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Computer Vision (ICCV), 2015 IEEE International Conference on*, 2015, pp. 2938-2946.

[22] 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,* 2017.

[23] J. Choi, H. Jin Chang, J. Jeong, Y. Demiris, and J. Young Choi, "Visual tracking using attention-modulated disintegration and integration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4321-4330.

[24] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár*, et al.*, "From captions to visual concepts and back," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1473-1482.

[25] X. Chen and C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2422-2431.

[26] C. Siagian and L. Itti, "Biologically inspired mobile robot vision localization," *IEEE Transactions on Robotics,* vol. 25, pp. 861-873, 2009.

[27] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, and T.-S. Chua, "SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning," *arXiv preprint arXiv:1611.05594,* 2016.

[28] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE transactions on pattern analysis and machine intelligence,* vol. 33, pp. 898-916, 2011.

[29] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1134-1142.

[30] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015, pp. 437-446.

[31] D. Yu, J. Fu, T. Mei, and Y. Rui, "Multi-level Attention Networks for Visual Question Answering," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[32] 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 *International Conference on Robotics and Automation*, Anchorage, United States, 2010.

[33] P. Neubert, N. Sünderhauf, and P. Protzel, "Superpixel-based appearance change prediction for long-term navigation across seasons," *Robotics and Autonomous Systems,* vol. 69, pp. 15-27, 2015.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov*, et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-9.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," presented at the Advances in neural information processing systems, 2012.

[38] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-Scale Image Retrieval with Attentive Deep Local Features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3456-3465.

[39] L. Liu, C. Shen, and A. van den Hengel, "Cross-convolutional-layer Pooling for Image Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2016.

[40] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer vision and image understanding,* vol. 110, pp. 346-359, 2008.