2019 16th International Conference on Ubiquitous Robots (UR)
Jeju, Korea, June 24-27, 2019

Neural Network-Based Long-Term Place Recognition from Omni-Images

Jongwon Lee[1] and Ayoung Kim[2]*

Abstract— **In robotics perception tasks, visual place recognition has drawn attention as a significant research topic on the grounds of its agile applications without using the global positioning system such as mobile robot navigation, augmented reality, and self-driving vehicles. Owing to the great performance improvement in most computer vision challenges based on deep learning, visual place recognition follows this trend. In this paper, we handle long-term visual place recognition. The long-term visual place recognition can be simplified by substituting it for a conventional supervised classification problem using a convolutional neural network. The proposed network is learned through only a single fisheye-formed illumination-invariant image, captured on Google Street View, for each class. Afterward, sequences of omnidirectional photographs measure how well the network performs. Even though a four-year gap exists between the two datasets, it seems that the proposed network discriminates well against challenges stemming from extreme visual changes.**

I. Introduction

Visual place recognition is a robotics task that focuses on the perceptional ability of an agent to identify places already visited [1]. It has drawn attention as a significant research topic due to its agile applications such as navigation with the lack of blglobal positioning system information and automatic geotagging. With the recent performance improvement that has occurred in most computer vision tasks based on deep learning, visual place recognition has also benefited from this trend, as it can be simplified into an image retrieval problem.

However, unlike typical image retrieval problems, place recognition is usually performed for long time intervals. Thus, the task can easily undergo severe difficulties stemming from changes in scenery over time. For instance, the flow of time brings about lighting changes along with diverse weather conditions as well as the seasonal changes of a landscape. Not only temporal changes but also spatial changes, such as the construction of a building, cause hardships during a task. These difficulties become critical when the time interval extends to different seasons or even a couple of years.

In this paper, we suggest a new method that directly classifies images into predefined labels based on where the data were taken. The end-to-end process is based on a deep learning technique that uses a convolutional neural network(CNN) model suggested by [2]. Our approach makes use of omnidirectional images inspired by [3], because they are not affected by viewing direction and encompass many visual contexts compared with conventional rectangular images due to their wide fields of vision. All omnidirectional images are converted into illumination-invariant images [4] so that they can endure inconsistency over time. The training data are procured from Google Street View, and the evaluation is performed with a spherical camera mounted on a carlike vehicle. Following the training of the CNN in our task, the accuracy of the learning is appraised.

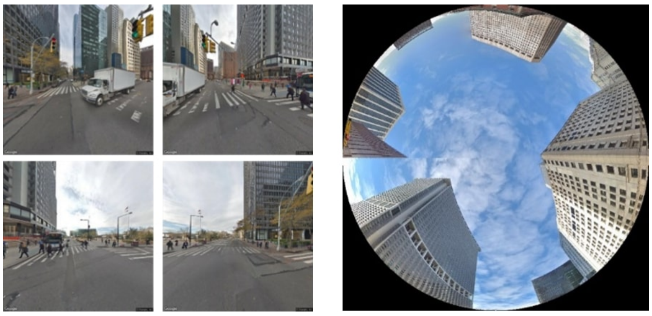


Fig. 1: Comparison of two different types of images. The left is a group of archetypal rectangular images and the right is an omnidirectional fisheye image. Both of them are taken at the same pinpoint. Note that the right one is free from its viewing angle and captures more static information (e.g. buildings and skylines), whereas the left one is not.

II. Related Work

Visual place recognition tasks feature image retrieval problems. For this reason, many efforts have been made to represent and match images effectively. For example, Murillo *et al.* [5] suggest a place recognition system using panoramic images based on GIST, one of the most common hand-crafted global descriptors. GIST is also used in [6] for the purpose of detecting locations being revisited. Nonetheless, global features, including GIST, are susceptible to effects over time, such as partial occlusions and change in illumination. Contrary to global features, local features are less reliant on a scale, the change of light, and computational cost. Cummins *et al.* [7] and Zamir *et al.* [8] did their work based on local features. Cummins *et al.* [7] substitute place recognition for probabilistic retrieval problem, also known as "bag-of-words," by representing visual data as local features. On the other hand, Zamir *et al.* [8] cope with the task by narrowing down the search scope in a hierarchical structure. However, although a number of approaches, including the above examples, to visual place recognition have been proposed, two major questions remain: *How can we actively handle the change that the temporal difference causes? Does a more simple and elaborate method exist?*

[1] Department of Mechanical Engineering, KAIST, Republic of Korea (E-mail: jongwon.lee@kaist.ac.kr)

[2] Department of Civil and Environmental Engineering, KAIST, Republic of Korea (E-mail: ayoungk@kaist.ac.kr)

978-1-7281-3232-7/19/$31.00 ©2019 IEEE

To deal with these difficulties, a few approaches have been suggested. Maddern *et al.* [4] and McManus *et al.* [9] make use of color gamut mapping called illumination-invariant color space to improve many aspects of visual localization, especially ensuring color consistency over time. This characteristic is first introduced by [4], who examined the zero-mean normalized cross correlation (ZNCC) between images taken at the same place over 24 hours. By virtue of its robustness against the effect of luminosity, both Maddern *et al.* [4] and McManus *et al.* [9] could substantially improve the performance of an outdoor visual navigation system. In a similar way, Chen *et al.* [10] demonstrate a place recognition technique based on a CNN model by utilizing downsized grayscale image sets converted from three channel images. After the feature vectors of images are extracted by the network, the pairings between training and testing images are reported subsequent to cross-check through a spatial and sequential filter.

In this work, we examine leveraging the illumination-invariant transform suggested by [4], [9] so that an agent could overcome long-term changes in a landscape. Inspired by [3], each image is acquired as an omnidirectional format taken in the vertical-upward direction. As one can see in Fig. 1, this is advantageous due to the rotational independence and the abundance of features attributed to its wide visual field. Moreover, it is less sensitive to ephemeral objects on the ground, such as pedestrians or vehicles. The end-to-end place recognition is performed by a customized CNN architecture based on VGG16 [2], one of the most qualified CNNs due to its powerful performance.

III. Methodology

A. *Illumination-Invariant Color Space*

Previously, McManus *et al.* [9] gave promise of improvement in long-term navigation and in robust navigation outdoors by harnessing lighting-invariant image transforms. This paper insists that it can reduce hardship, which may be precipitated by modifications in the lighting condition. In this sense, we transform our visual dataset into an illumination-invariant chromaticity space to dampen the effects of light. The fundamental assumption is that the image sensor of a camera has an infinitely narrow-band spectral response, and the spectrum of sunlight can be approximated by a blackbody. In this way, the response of linear image sensor R is simplified into a few components: reflectivity of materials, geometry, illumination, wavelength, and color temperature. Further procedures are described in [4] and [9].

Through the simplification, we can obtain an illumination-invariant intensity $\mathcal{I}$, instead of using three color channels $\{R_1, R_2, R_3\}$ from a raw image. Illumination-invariant intensity $\mathcal{I}$ is denoted as:

$$\mathcal{I} = \log(R_2) - \alpha \log(R_1) - (1-\alpha)\log(R_3) \quad (1)$$

where parameter α satisfies the following constraint:

$$\frac{1}{\lambda_2} - \frac{\alpha}{\lambda_1} - \frac{1-\alpha}{\lambda_3} = 0 \quad (2)$$

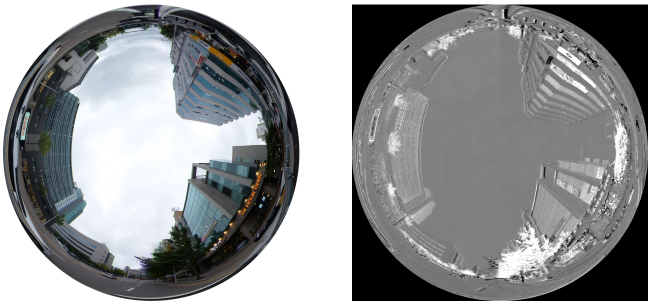

(a) RGB-image (b) Illumination-invaraint image

Fig. 2: Comparison between three-channel and illumination-invariant color space. The right picture is converted from the left one. Parameter α is set to 0.5.

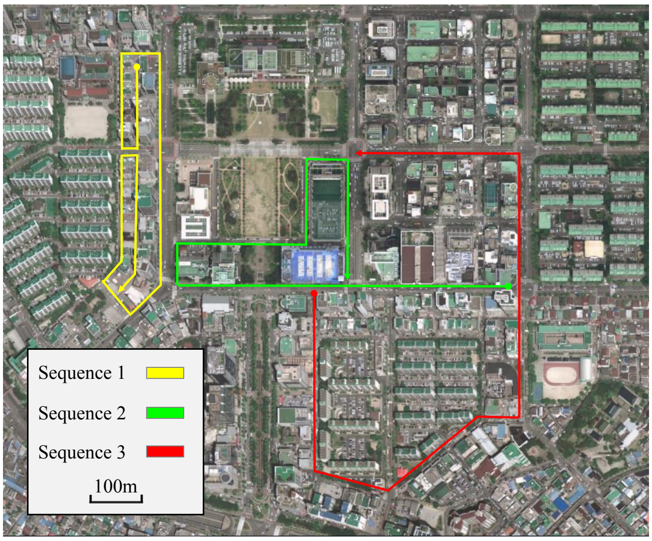


Fig. 3: Three trajectories where we obtained dataset are overlaid on a satellite image.

TABLE I: Information of Dataset

Sequence No.	Traveled Distance (km)	The Number of Images
1	1.64	137
2	1.47	136
3	1.61	132

with a set of wavelengths $\{\lambda_1, \lambda_2, \lambda_3\}$ corresponding to the peak sensitivities of a image sensor. An example of the actual application of an illumination-invariant color space is shown in Fig. 2.

B. *Dataset*

To deal with the long-term place recognition task, two sets of omnidirectional image data taken along the same trajectories in an urban area were required. In this section, we explain how we took the visual data.

1) Training: Google Street View sufficiently meets the conditions we mentioned above. It is an expansive database made up of complete 360° panoramic images with a distance of 12m or so between consecutive frames. Because it provides a number of images taken at almost all streets and roads, we can readily plan a path and extract images of the places we are interested in.

2) Evaluation: Owing to a flood of affordable vision sensors lately, omnidirectional images can be obtained via a variety of hand-held devices with ease. Ricoh Theta V is an exemplary device in terms of not only its high-quality panorama views but also its ease of use. By using Ricoh Theta V, we captured sequences of evaluation data for testing.

Both the training and evaluation datasets were collected along the same paths. The former was from Google Maps as panorama images captured in November 2014, whereas the other was manually taken from Ricoh Theta V in October 2018. Because the images from the former were placed at about 12m intervals, we regarded each location where the image was taken as a single class and labeled each element of the latter one-by-one. Subsequently, we projected them onto a spherical form and then converted them into illumination-invariant color space. Details of the two datasets used are summarized in Fig. 3 and TABLE I.

IV. Experiment and Results

A. Training the Network

We used a pre-trained CNN model as a feature extractor. VGG16 [2] is one of the representative CNNs well known not only for its simple and intuitive architecture but also for its powerful performance. Before the main training process, the CNN was trained by a large-scale scene recognition database called places365 [11]. Empirically, we found that the CNN trained by places365 dataset worked better than any other pre-trained network for visual place recognition. This implies that the pre-trained network is specialized in recognizing and extracting features from given omnidirectional urban images. Feature vectors printed out from the CNN then flowed into the three units of batch normalization, dropout, and fully connected layers. In the end, the softmax layer whose units were one-hot encoded with each label of a location gave the probability that the input image corresponded to each label. The overall architecture of the network is depicted in Fig. 4. We utilized an Adam optimizer [12] with default parameters from Keras [13], which minimized the categorical cross entropy of each label. A batch of 75 and a dropout of 0.4 were used.

All input data were encoded into a spherical form, shrank to the size of 224×224, and mapped into illumination-invariant color space, as mentioned in Section III-A and Section III-B. Because we do not know the actual light sensitivities and spectral responses of both vision sensors used in capturing training and evaluation data, parameter α, related to the peak spectral responses of each sensor channel, simply ranged from 0.1 to 0.9 at 0.1 intervals to augment the data for training, while being set at 0.5 in the evaluation stage. In addition, the training dataset was augmented via rotating at 30 degrees from 0 to 360 degrees to offset effects stemming from directional differences between the training and evaluation data. Each training datum was labeled according to its corresponding place index, whereas each test datum was estimated rather than learned. That is to say that we simply considered the place recognition to be a conventional supervised classification problem.

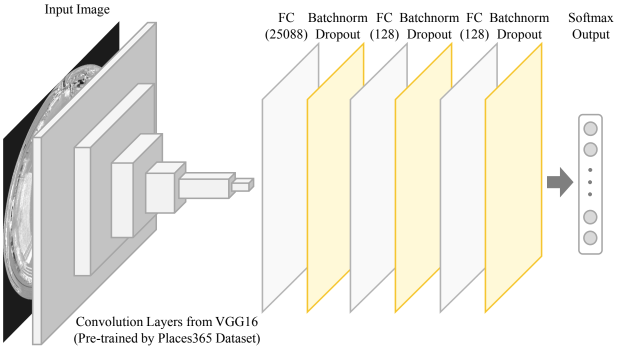


Fig. 4: Our network's architecture. We made use of convolution layers from VGG16 [2], pre-trained by places365 [11] dataset. The output vector then goes into three units of fully connected layer, batch normalization, and dropout. Finally, classification probability vector of places are printed out.

B. Expanded Top-5 Accuracy

As shown in Fig. 5, in our dataset, subtle differences exist between training and evaluation image pairs, as the trajectory path for gathering evaluation data cannot precisely traverse the route for training. At this point, it is necessary to introduce a metric that offsets inaccuracy resulting from inconsistencies between the training and evaluation databases before we appraise our work's performance. Now, we expand the concept of top-5 accuracy, a widely used metric for measuring image classification problems.

Given that the network outputs an ordered list of candidate locations for each query image, the top-k accuracy metric counts how many query images are localized correctly within k top-ranked accuracy.

Assuming that query image I_i ($i \in [1,n]$) is an ith image in sequence $S = \{I_1, I_2, \cdots, I_n\}$ of evaluation data, we define an *adjacent matrix around* I_i as

$$adj(I_i) = \{I_{i-1}, I_i, I_{i+1}\}$$

If a *top-k candidate of an ith image* I_i is denoted as $C(I_i) = \{c_1, c_2, \cdots, c_k\}$, *expanded top-k accuracy of an ith image* I_i $Acc(I_i)$ is

$$Acc(I_i) = \begin{cases} True, & \text{if } adj(I_i) \cap C(I_i) \neq \emptyset \\ False, & \text{otherwise} \end{cases}$$

The average expanded top-k accuracy of sequence $S = \{I_1, I_2, \cdots, I_n\}$ is calculated as

$$\overline{Acc(I_i)} = \frac{\sum_{k=1}^{n} Acc(I_k)}{n}$$

It seems to be a quite reasonable metric in that query image I_i is correctly recognized and localized if 1 out of k accuracy lies inside a tolerance range $adj(I_i)$. In this paper, we adopt the concept of the modified top-k accuracy in lieu of the typical one.

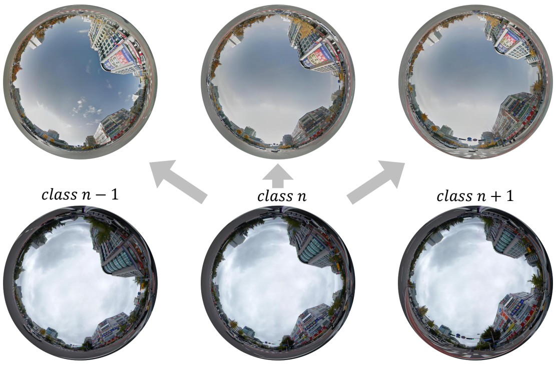


Fig. 5: In the dataset, there are not only delicate distinctions between training and evaluation image pairs, which have the same label, but also ambiguities among images marked as different places. To cancel out these unwanted effects, we simply expand the concept of the top-k accuracy.

TABLE II: Expanded Top-k Accuracy

Sequence No.	Top-1 (%)	Top-5 (%)
1	64.23	75.91
2	52.21	63.97
3	49.24	58.33

C. Performance Evaluation

In this section, we demonstrate how our method successfully recognized their places well. As TABLE II indicates, the network performed best on sequence 1. The expanded top-1 and top-5 accuracies on the sequence consisted of 137 images marking 64.23% and 75.91%, respectively. This is a noticeable result considering that the time interval between training and evaluation data exceeds four years.

The temporal discrepancy causes significant visual changes between two datasets, and they make the proposed network hard to train. For instance, newly built structures along the street are one of the major factors that severely affects the incorrect rate, not to mention the transition in the color of the leaves with the seasons.

Even though we have introduced the illumination-invariant imaging technique to deal with the possible inconsistency of the color temperature of visual data with respect to the shooting time, place recognition often fails. This seems to result from the naive estimation of the spectral response of two camera sensors and excessive idealization with regard to the outdoor environment. Moreover, as shown in Fig. 5, the discrepancy in the viewpoints of image pairs originating from the discordances of two trajectory pairs may bring about errors. The region where the task fails is depicted in Fig. 6.

In a nutshell, the dominant cause of difficulty in our long-term place recognition problem can be categorized into three cases: (i) newly-built buildings, (ii) the similarity of skylines among various scenes, and (iii) the lack of features in an image. For more details, consult Fig. 7 as a comprehensive case of failures.

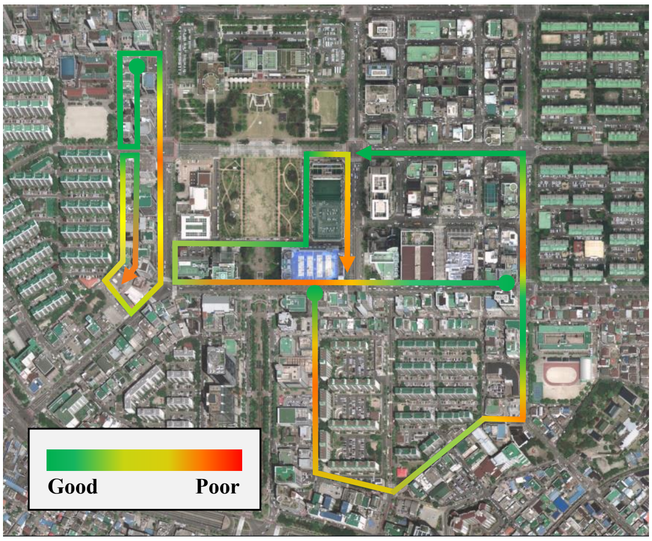


Fig. 6: Illustration of test results along three paths of sequences. The green color indicates that the network successfully recognizes its position, whereas the red does not.

V. Conclusion

In this study, we demonstrated a long-term visual place recognition task in conjunction with a few assumptions and approaches. We used only single training - single test pairs of samples per each location in an urban canyon. In spite of the lack of data and visual disharmony on several image pairs, which even made it hard for people to distinguish themselves, our method showed quite intriguing performance. This approach may shed light on the advent of cost-effective mobile autonomy, which could be worked on long term for vision-based devices in the future.

VI. Acknowledgement

This work was supported by the Department of Civil and Environment Engineering, KAIST through 'CEE-URP Program' and by KAIST through 'Undergraduate Research Participation Program'.

References

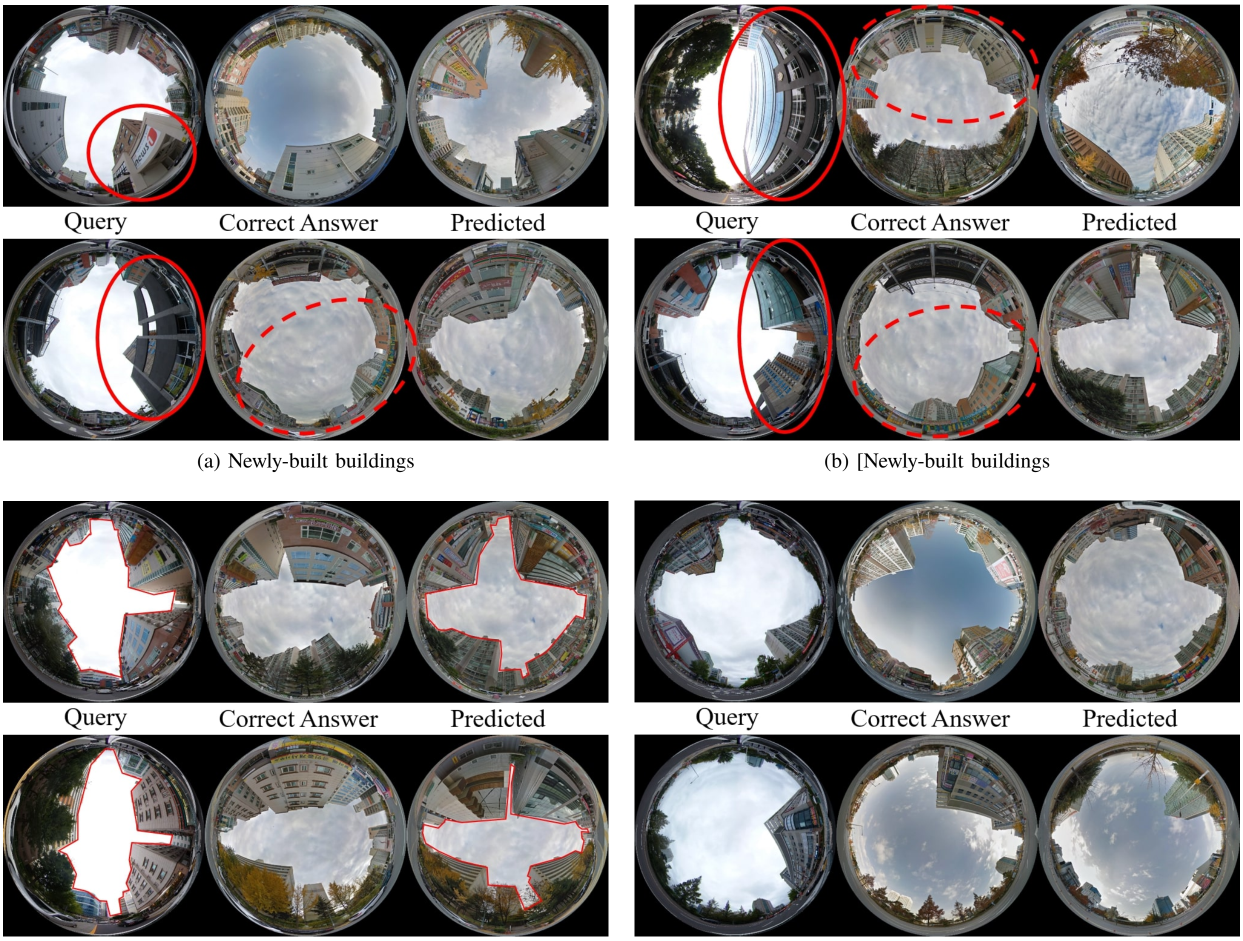
[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

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

[3] S. Ramalingam, S. Bouaziz, P. Sturm, and M. Brand, "Geolocalization using skylines from omni-images," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, IEEE, 2009, pp. 23–30.

[4] W. Maddern, A. Stewart, C. McManus, B. Upcroft, W. Churchill, and P. Newman, "Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles," in *Proceedings of the Visual Place Recognition in Changing Environments Workshop, IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China*, vol. 2, 2014, p. 3.

[5] A. C. Murillo and J. Kosecka, "Experiments in place recognition using gist panoramas," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, IEEE, 2009, pp. 2196–2203.



(a) Newly-built buildings

(b) [Newly-built buildings

(c) Similar skylines

(d) Lack of features

Fig. 7: Three major cases of failures. The left element in each pair is an assigned query image for the network we proposed. The middle one is the correct answer which the network should have predicted, and the rightmost is an image the network actually submitted.

[6] G. Singh and J. Kosecka, “Visual loop closing using gist descriptors in manhattan world,” in *ICRA Omnidirectional Vision Workshop*, 2010.

[7] M. Cummins and P. Newman, “Fab-map: Probabilistic localization and mapping in the space of appearance,” *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[8] A. R. Zamir and M. Shah, “Accurate image localization based on google maps street view,” in *European Conference on Computer Vision*, Springer, 2010, pp. 255–268.

[9] C. McManus, W. Churchill, W. Maddern, A. D. Stewart, and P. Newman, “Shady dealings: Robust, long-term visual localisation using illumination invariance,” in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, IEEE, 2014, pp. 901–906.

[10] Z. Chen, O. Lam, A. Jacobson, and M. Milford, “Convolutional neural network-based place recognition,” *arXiv preprint arXiv:1411.1509*, 2014.

[11] 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.

[12] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[13] F. Chollet *et al.*, “Keras: Deep learning library for theano and tensorflow,” *URL: https://keras. io/k*, vol. 7, no. 8, 2015.