This is the Pre-Published Version.

The following publication C. -T. Li, W. -C. Siu and D. P. K. Lun, "Semi-Supervised Deep Vision-Based Localization Using Temporal Correlation Between Consecutive Frames," 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 1985-1989 is available at https://dx.doi.org/10.1109/ICIP.2019.8803131.

SEMI-SUPERVISED DEEP VISION-BASED LOCALIZATION USING TEMPORAL CORRELATION BETWEEN CONSECUTIVE FRAMES

Chu-Tak Li, Wan-Chi Siu, Life-FIEEE, Daniel P.K. Lun, SrMIEEE

Centre for Multimedia Signal Processing, Department of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong

ABSTRACT

Vision-based localization is a temporal informative task in which we can obtain information about the ego-motion of a vehicle from the historical information via examining consecutive frames. Sufficient temporal information helps to reduce the search space of the next location. Hence, both efficiency and accuracy of the localization system can be enhanced. This paper presents a semi-supervised deep vision-based localization algorithm, using a novel tubing strategy to find the starting location of a vehicle. We group different number of consecutive frames as sets of tubes based on their temporal correlation to achieve pair searching with variable tube sizes. We also enhance an off-the-shelf network model with our modified training data generation method to improve the discrimination power of the features given by the model. Experimental results show that our proposed temporal correlation based initialization module can confidently localize the starting location of a vehicle (for a certain journey), and achieve 40% precision improvement over that of the conventional CNN approaches.

Index Terms— Visual localization, temporal correlation, scene recognition, autonomous driving, deep learning

1. INTRODUCTION

Visual scene recognition (or visual ego-localization) has been studied for decades and it is crucial to the development of autonomous driving [1,2]. Nowadays, there are numerous applications and services in which Global Navigation Satellite Systems (GNSS) acts as the basis of ego-localization module. Nevertheless, one common problem is that GNSS suffers from the masking and reflection of the GNSS signal on dense trees and concrete buildings [1]. Under the circumstances, many researchers look for other possible techniques to give ego-localization information so as to generalize various types of localization systems. Vision-based localization has been a popular topic because of the richness and cost-effectiveness of visual information. In this paper, we will show how we can accomplish a reliable system through using only visual information.

Visual localization methods can be categorized into two main streams, namely single image-based and sequence-based. For the first category, finding the global best match to a query image among all database images (single nearest neighbor search) is the most straightforward method which merely relies on the discrimination power of the image descriptors. Milford and Wyeth [3] presented the first sequence-based method called SeqSLAM. With the use of sequences of images as a performance boosting technique, SeqSLAM attains 100% precision at around 60% recall rate. [4-8] propose methods for visual localization via using the temporal correlation given by consecutive frames. Here, we define the temporal correlation as the information obtained from the comparisons of previous frames.

Recently, Convolutional Neural Networks (CNN) have outperformed many conventional methods in various computer vision tasks [9]. CNN features (or deep features) have been proven to be more robust than traditional hand-crafted features. [10-14] applied deep features to the localization problem and evaluated the corresponding performance. Heavy computational cost and difficulty in constituting a large amount of training data are two obvious weaknesses of CNN-based methods. [10] proposes a shallow convolutional autoencoder network to speed up the feature extraction stage and a way to generate training data automatically for easing the collection of training data. They achieved satisfactory performance in small databases via using only single nearest neighbor search.

In this paper, our main contributions include (i) we propose a high confident initialization stage which makes use of the temporal correlation between consecutive frames; (ii) we improve the network proposed by [10] and enhance the way of generating training data by considering also the changes in appearance, illumination and viewpoints, as well as fine-tuning the model on parts of a localization dataset, Nordland dataset [15]. Hence, the discrimination power of the deep features from the fine-tuned model is generally improved. Extensive experimental results, including on three challenging datasets with different practical issues, show the temporal correlation between consecutive frames can boost the initialization performance in large databases.

The rest of the paper is organized as follows. Section 2 briefly reviews the architecture of the used network and describes our suggested improvement on training data generation. Section 3 presents the proposed initialization stage. Section 4 provides comparisons of several state-of-the-art methods. Finally, Section 5 draws a conclusion of this paper.

2. REVIEW ON CONVOLUTIONAL AUTOENCODER NETWORK FOR LOOP CLOSURE

In 2012, AlexNet [9] outperformed the conventional methods in the classification competition. [12] evaluated the performance of deep features obtained from AlexNet in localization tasks, for which the structural features of a scene are of critical importance. For scene recognition, moving and standstill objects can be regarded as noise which affects our decision making. We can infer

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

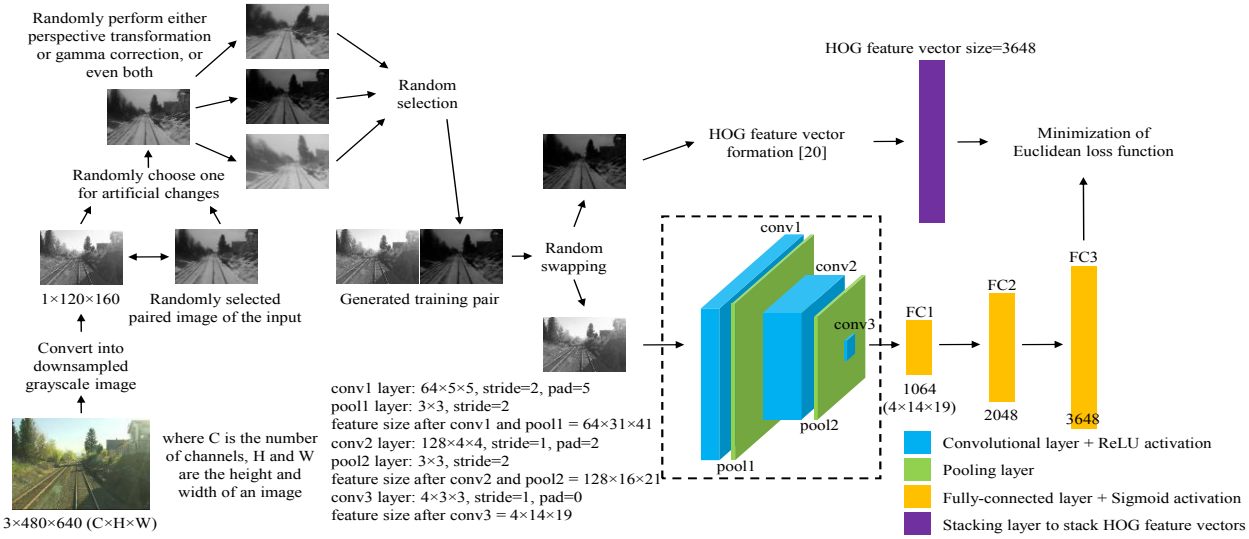

Fig. 1. Illustration of the CALC network architecture and its training process [10] with our modifications in the training data generation. Note that the training pair covers the problems of changes in appearance, illumination, and also viewpoints

that the middle convolutional layer (conv3) is more suitable for localization tasks as structural features of a scene is about understanding the gist of an image. Their experimental results showed that conv3 features give the best performance in localization tasks. However, high dimensional feature vector (conv3 features $\in \mathbb{R}^{384\times13\times13} \in \mathbb{R}^{64896}$) causes the problem of slow pair matching especially for linear full search strategy.

[10] resolved the above mentioned problems via constructing a shallow convolutional autoencoder network and preparing the training data with a self-transformed manner. Autoencoder model is able to extract robust features because of its denoising property [16]. Generally, the model learns to identify the salient features of the input and use them to generate a "denoised" version of the input as output to us. For training data, [10] applied a random perspective transformation to each input such that each input is automatically paired up with its self-perspective transformed version as the ground truth label. As a result, their network is tailored for a particular issue – changes in viewpoints.

Fig.1 shows the training process of their proposed network with our modifications in the training data generation. We follow their training process[1] under the Caffe framework [17] to improve the model. Histograms of Oriented Gradients (HOG) [18,19] is a well-known hand-crafted feature which is robust to changes in illumination because of its local contrast normalization. [10] adopted HOG for their network training in which it benefits from (i) smaller size of the extracted features to achieve reasonable data compression (HOG feature vector $\in \mathbb{R}^{3648}$); (ii) illumination invariance property of HOG to handle changes in lighting conditions; (iii) perspective transformed training data to further enhance the features for localization tasks. Their proposed network aims to reproduce the same HOG feature vector of an image pair as the prior knowledge is that an image pair always represents the same scene. For the loss function, they simply employed Euclidean loss function ($L2$ norm) [18] to minimize the difference between the HOG feature vector and the deep feature vector given by the network. For the online use, only the convolutional layers are kept (as shown in the dashed box in Fig.1) for further quicken the feature extraction. Hence, the network learns to map an input image $I \in \mathbb{R}^{120\times160}$ to a feature space $\in \mathbb{R}^{1064}$.

Our main suggested improvement in the training data generation is that we add more variations in the paired training data. Apart from the random perspective transformation, we also employ a random gamma correction to the input images and a random selection of the corresponding image for the data generation. Eqn.1 shows the formula for gamma correction. Note that γ is randomly chosen from 0.1 to 2.5.

$$I_c = \left(\frac{I}{255}\right)^{\gamma} \times 255 \tag{1}$$

where I is the input, I_c is the gamma corrected image and γ is the gamma used to correct the brightness of the input via using non-linear mapping of pixel values. If $\gamma < 1$, the corrected image will be darker than the input. For $\gamma > 1$, the opposite observation is made. We are not restricted to use the input image to generate the ground truth label for training. Our procedure for generating a training image pair is as follows. (i) We randomly select one of the paired images of the input from the dataset or use the input directly as an image pair. For example, if a scene has been recorded in 4 different time slots in the dataset, we will have possible 4 image pairs of this scene. We randomly pick one of them and use it to create a training image pair. Note that all the images input to the network is in grayscale and down-sampled to 120×160 as the same in [10]. (ii) From each training image pair, we randomly choose one image to perform either the perspective transformation or the gamma correction (Eqn.1), or even both of them. With our data generation strategy, we can then resolve the three practical problems, namely changes in appearance, illumination, and viewpoints.

3. PROPOSED METHOD

3.1. Temporal correlation based initialization

[1] Source code and pre-trained model are available online: https://github.com/rpng/calc. Please refer to it for the details.

Our previous work [8] found that only datasets for situations with identical speed can benefit from long image sequences. Practical situations with varying speeds are burdened with the excessive consideration to the historical information. This induces an important question to us. What is a suitable size of image sequence (or we refer to *tube size* later)? This means that how much temporal correlation we have to consider in order to benefit from it. To address this concern, we propose a weighted sum of the searching (or matching) similarity scores (S) of consecutive frames (we refer this strategy to *tubing* later). The weight of each searching similarity score depends on the differences between the current querying frame and the previous frames. Fig.2 shows the pipeline of our proposed temporal correlation based initialization graphically.

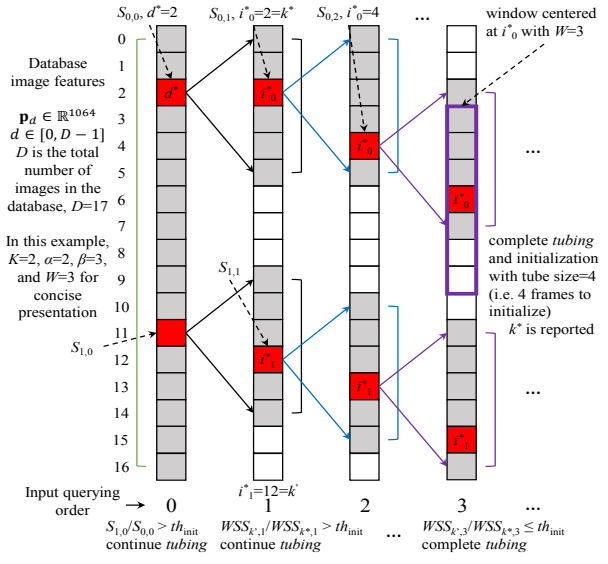

Fig. 2. Our proposed temporal correlation based initialization stage

For the 1st query frame, we perform the linear full search to find the match pair as shown in Eqn.2.

$$d^* = \arg\max_{d \in [0, D-1]} C(\mathbf{q}_t, \mathbf{p}_d) = \arg\max_{d \in [0, D-1]} (\mathbf{q}_t \cdot \mathbf{p}_d) \quad (2)$$

where $C(\mathbf{q},\mathbf{p})$ is the Cosine similarity defined as the cosine of the angle difference between normalized vectors $\mathbf{q}$ and $\mathbf{p}$. In this paper, $\mathbf{p}$ and $\mathbf{q}$ represent the normalized deep feature vectors of the database images and query images respectively extracted by the network model discussed in Section 2. D is the total number of frames in the database, t is the input query order, t=0 for the 1st query frame. Neighbor d^* is the single nearest neighbor to the 1st query frame among the database with the highest similarity score ($S_{0,t=0} = C(\mathbf{q}_t,\mathbf{p}_{d^*})$) and d^* is regarded as the match pair for output. If the ratio of $S_{0,t=0}$ to $S_{1,t=0}$ (the second highest similarity score found outside the window centered at d^*, with window size W) is smaller than a threshold, th_{init}, the confidence of this match pair is high and the initialization can be done with only the 1st query frame. Otherwise, we record the top T% of the nearest neighbors with the similarity scores denoted as S_k, $k \in [0,K\text{-}1]$ and $K=D\times T\%$. The search range of the next query frame is based on the K nearest neighbors and i_k denote the location of the kth neighbor in the database. Starting from the 2nd query frame, the weighted similarity score is computed by Eqns.3 and 4.

$$i_k^* = \arg\max_{i \in [i_k - \alpha, i_k + \beta]} C(\mathbf{q}_t, \mathbf{p}_i) \quad (3)$$

$$S_{k,t} = C(\mathbf{q}_t, \mathbf{p}_{i_k^*}) \times (1 - C(\mathbf{q}_t, \mathbf{q}_{t-1})) \quad (4)$$

where α and β are the upward and downward search offset respectively. For each i_k, it has its own search range and we find the corresponding single nearest neighbor i^*_k. If $i^*_k\text{-}\alpha < 0$, we start to search from the 1st database frame; if $i^*_k+\beta \ge D$, we stop the search once reaching the last database frame. $S_{k,t}$ is the weighted similarity score of the kth neighbor in time t. We weight the score using the difference between the current and previous query frames. If the two frames are very similar, this means that we can neglect the current score as there is very little new information given by the current query frame. Note that i_k is updated for each query frame based on i^*_k, hence we only keep the K nearest neighbors to each query frame. We report the match pair of the current query frame (k^*) using Eqn.5.

$$k^* = \arg\max_{k \in [0, K-1]} (S_{k,t} + \sum_{a=0}^{t-1} S_{k,a} \times C(\mathbf{q}_t, \mathbf{q}_a)) \quad (5)$$

$$WSS_{k^*,t} = S_{k^*,t} + \sum_{a=0}^{t-1} S_{k^*,a} \times C(\mathbf{q}_t, \mathbf{q}_a) \quad (6)$$

if t=1, $S_{k,0}$ is obtained from the linear full search of the 1st query frame (i.e. $S_{k,t=0}=\{S_{0,0}, S_{1,0},\ldots, S_{K\text{-}1,0}\}$). When the current query frame is different from the past query frames, the vehicle goes far away from the past locations and the influence of the historical information on the current decision making should be diminished. The weighted sum of scores of location k^* for the current query frame is calculated using Eqn.6. Similar to the case of the 1st query frame, we compute the ratio of $WSS_{k^*,t}$ to $WSS_{k',t}$ (k' is location which has the second highest weighted sum of scores found outside the window centered at k^*, with window size W, the purple box in Fig.2). If the ratio is smaller th_{init}, high confidence of k^* is observed and the initialization is done with tubing.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Dataset for fine-tuning and module parameters

For the network model fine-tuning mentioned in Section 2, we used parts of the Nordland dataset [15] and removed the tunnel and stop frames. This dataset contains 4 long rail sequences recorded at 4 seasons and it has been time-synchronized such that any frame in one of the sequences represents the same frame in the other sequences. We used the first 10,000 frames of each sequence. After the removal of stop and tunnel frames, there are 7,705 frames for each sequence and 30,820 frames in total for the fine-tuning.

We consider the top 10% of the nearest neighbors for the initialization. α and β are set to 5 and 10 respectively. The window size W is set to 3 and th_{init} is predefined as 0.8.

4.2. Datasets for comparisons

Much experimental work has been done, including 3 challenging datasets with several state-of-the-art approaches, namely CALC [10], and AlexNet conv3 deep feature-based approaches [9,12,20]. CALC is the network model that we fine-tuned for our proposed method. Note that CALC merely focuses on the discrimination

Table. 1. The initialization performance of various approaches (the best scores are in **bold** typeface)

	Fine-tuned CALC									
	Without tubing strategy		With tubing strategy		CALC [10]		AlexNet, Places365[20]		AlexNet, ImageNet[9]	
querying sequence	Precision	Average tube size	Precision	Average tube size	Precision	Average tube size	Precision	Average tube size	Precision	Average tube size
Alderley	0.2	1.0	**0.5**	11.1	0.1	1.0	0.0	1.0	0.1	1.0
Summer	0.7		**1.0**	8.0	0.5		0.5		0.5	
Fall	0.9	1.0	**1.0**	2.5	0.8	1.0	0.6	1.0	0.3	1.0
Winter	0.4		**0.7**	7.6	0.3		0.2		0.0	
LRT2	0.6		**0.7**	54.1	0.2		0.6		0.2	
LRT3	0.6	1.0	**1.0**	80.4	0.6	1.0	0.7	1.0	0.6	1.0
LRT4	0.6		**0.7**	27.4	0.5		**0.7**		0.6	
Average	0.571	1.0	**0.800**	27.3	0.429	1.0	0.471	1.0	0.329	1.0

power of its deep features, the simplest single nearest neighbor search is applied to find the match pair. For AlexNet conv3 deep feature-based approaches, we directly extract the conv3 features from two AlexNets pre-trained on two datasets, ImageNet [9] and Places365 [20]. The former one is for classification tasks and the latter one is for single scene recognition tasks. These approaches also employ the single nearest neighbor search.

4.2.1. Alderley Dataset

This dataset is described in [3] which focuses on extreme changes in weather and lighting conditions. We extracted the first 2,000 frames of the daytime sequence to construct the database and there are 2,069 frames of the nighttime sequence correspond to these 2,000 database images. The ground truth of this dataset is very close to a diagonal line which means situation with identical speed.

4.2.2. Nordland Dataset

We used the last 5,000 frames of each sequence for comparisons. We have ensured that there is no overlap with the training data and we did not remove the tunnel and stop frames for the comparisons. It is because this is the real situation for practical applications, and this is for testing algorithms whether they can perform well with slow moving or even stop frames. The database was formed using the "Spring" sequence.

4.2.3. Light Rail Transit (LRT) Dataset

LRT dataset was captured directly from a public transportation system in Hong Kong. There are 4 sequences of the same route, 3 in the daytime and 1 at the nighttime. The dataset consists of many practical difficulties such as varying speeds, extreme changes in illumination, and blurring. On average, there are 2566 frames for a sequence in this dataset. We used one of the daytime sequences to form the database. As the sequences are not time-synchronized for this dataset, we manually marked the ground truth of all the sequences.

4.3. High confident initialization with temporal module

In this part, we can show the advantage of using our proposed tubing strategy for the initialization to localize the starting location of the vehicle. For computing the precision, a match pair is regarded as correct if its difference between the ground truth is less than 5 frames. We randomly selected 10 starting points for each querying sequences and the precisions of the initializations are shown in Table.1.

Without our proposed tubing strategy, the initialization is simply done with the single nearest neighbor search. Hence, tube size is always 1.0 as there is no any temporal correlation between consecutive query frames. Obviously, the initialization performance is boosted with the tubing strategy. On average, we get 0.229 (=0.800-0.571) increase in precision. High confident initialization is always the first step in comprehensive localization systems. Compared with the original CALC, we also show our improvement in the performance with our modified training data generation method. Considering only the discrimination power of the deep features, the original CALC get 0.429 in average precision while our fine-tuned CALC attained 0.571. We can also observe that the AlexNet pre-trained on Places365 outperforms the AlexNet pre-trained on ImageNet. On average, these two models achieved 0.471 and 0.329 precision respectively. This implies that one can benefit from the model pre-trained on task-related datasets. We achieve 0.4 (=0.800-(0.471+0.329)/2) precision improvement over that of the two AlexNet conv3 deep feature-based approaches.

In addition, our tubing strategy groups a number of query frames to make a confident initialization decision. Note that we have had larger tube size for LRT2 (54.1) and LRT3 (80.4) sequences. This is due to the fact that we have to handle situations with varying speeds. The random starting points sometimes start at stop-frame locations and this requires more frames to make the decision as stop frames contain very little or even no new information. On average, the tubing strategy requires 27.3 frames to localize the starting location of the vehicle which costs around 1.1 second in a 25 fps system.

5. CONCLUSION

In this paper, we have proposed a high confident initialization module via the use of temporal correlation between consecutive query frames. We also fine-tuned an existing network model with our modified training data generation method to successfully enhance the discrimination power of the deep features. More importantly, we have suggested a way to achieve pair searching with variable tube sizes. For future development, we will focus on the localization systems with the initialization module. We will keep studying how to effectively make use of the temporal information given by the coming query frames so as to maximize its benefit to our proposed localization systems.

6. ACKNOWLEDGEMENT

This work was supported by the Hong Kong Polytechnic University (1-ZE1B), and Mr. LI Chu Tak would like to acknowledge the Postgraduate Studentship offered by the same university.

7. REFERENCES

[1] Guillaume Bresson, Zayed Alsayed, Li Yu, and Sébastien Glaser, "Simultaneous Localization And Mapping: A Survey of Current Trends in Autonomous Driving," *IEEE Trans. on Intelligent Vehicles*, vol.2, no.3, pp. 194-220, Sept. 2017.

[2] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J. Leonard, "Past, Present, and Future of Simultaneous Localization And Mapping: Towards the Robust-Perception Age," *IEEE Trans. on Robotics*, vol.32, no.6, pp. 1309-1332, Dec. 2016.

[3] Michael J. Milford and Gordon F. Wyeth, "SeqSLAM: Visual Route-Based Navigation for Sunny Summer Days and Stormy Winter Nights," *Proceedings*, *IEEE International Conference on Robotics and Automation* (*ICRA*), Saint Paul, Minnesota, USA, pp. 1643-1649, May 2012.

[4] Roberto Arroyo, Pablo F. Alcantarilla, Luis M. Bergasa, and Eduardo Romera, "Towards Life-Long Visual Localization using an Efficient Matching of Binary Sequences from Images," *Proceedings*, *IEEE International Conference on Robotics and Automation* (*ICRA*), Seattle, Washington, pp. 6328-6335, May 2015.

[5] Roberto Arroyo, Pablo F. Alcantarilla, Luis M. Bergasa, and Eduardo Romera, "OpenABLE: An Open-source Toolbox for Application in Life-Long Visual Localization of Autonomous Vehicles," *Proceedings*, *IEEE International Conference on Intelligent Transportation Systems* (*ITSC*), Rio de Janeiro, Brazil, pp. 965-970, Nov. 2016.

[6] Yongliang Qiao, Cindy Cappelle, and Yassine Ruichek, "Visual Localization across Seasons Using Sequence Matching Based on Multi-Feature Combination," *Journal of Sensors*, vol.17, no.11, pp. 2442-2463, Oct. 2017.

[7] Sayem Mohammad Siam, and Hong Zhang, "Fast-SeqSLAM: A Fast Appearance Based Place Recognition Algorithm," *Proceedings*, *IEEE International Conference on Robotics and Automation* (*ICRA*), Singapore, pp. 5702-5708, May 2017.

[8] Chu-Tak Li, and Wan-Chi Siu, "Fast Monocular Vision-based Railway Localization for Situations with Varying Speeds," *Proceedings*, *APSIPA Annual Summit and Conference 2018* (*APSIPA-ASC 2018*), Hawaii, USA, pp. 2006-2013, Nov. 2018.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Proceedings*, *the 25th International Conference on Neural Information Processing Systems* (*NIPS*), Lake Tahoe, Nevada, USA, pp. 1097-1105, Dec. 2012.

[10] Nate Merrill and Guoquan Huang, "Lightweight Unsupervised Deep Loop Closure," *Proceedings*, *Conference on Robotics: Science and Systems* (*RSS*), Pittsburgh, Pennsylvania, USA, pp. 1-9, Jun. 2018.

[11] Roberto Arroyo, Pablo F. Alcantarilla, Luis M. Bergasa, and Eduardo Romera, "Fusion and Binarization of CNN Features for Robust Topological Localization across Seasons," *Proceedings*, *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Daejeon, South Korea, pp. 4656-4663, Oct. 2016.

[12] Niko Sünderhauf, Sareh Shirzai, Feras Dayoub, Ben Upcroft, and Michael J. Milford, "On the Performance of ConvNet Features for Place Recognition," *Proceedings*, *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Hamburg, Germany, pp. 4297-4304, Sept. 2015.

[13] Tayyab Naseer, Wolfram Burgard, and Cyrill Stachniss, "Robust Visual Localization Across Seasons," *IEEE Trans. on Robotics*, vol.34, no.2, pp. 289-302, Apr. 2018.

[14] Muneeb Shahid, Tayyab Naseer, and Wolfram Burgard, "DTLC: Deeply Trained Loop Closure Detections for Lifelong Visual SLAM," *Proceedings*, *Workshop on Visual Place Recognition*, *Conference on Robotics: Science and Systems* (*RSS*), Ann Arbor, MI, USA, pp. 1-8, Jun. 2016.

[15] Niko Sünderhauf, Peer Neubert, and Peter Protzel, "Are We There Yet? Challenging SeqSLAM on a 3000 km Journey Across All Four Seasons," *Proceedings*, *Workshop on Long-Term Autonomy*, *IEEE International Conference on Robotics and Automation* (*ICRA*), Karlsruhe, Germany, pp. 102-115, May 2013.

[16] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, "Extracting and Composing Robust Features with Denoising Autoencoders," *Proceedings*, *the 25th International Conference on Machine Learning*, New York, USA, pp. 1096-1103, Jul. 2008.

[17] Y Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrel, "Caffe: Convolutional Architecture for Fast Feature Embedding," *Proceedings*, *the 22nd ACM International Conference on Multimedia*, Orlando, Florida, USA, pp. 675-678, Nov. 2014.

[18] Navneet Dalal, and Bill Trigss, "Histograms of Oriented Gradients for Human Detection," *Proceedings*, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR*), San Diego, CA, USA, pp. 886-893, Jun. 2005.

[19] Chu-Tak Li, Wan-Chi Siu, and Daniel P.K. Lun, "Boosting the Performance of Scene Recognition via Offline Feature-Shifts and Search Window Weights," *Proceedings*, *IEEE International Conference on Digital Signal Processing* (*DSP*), Shanghai, China, pp. 1-5, Nov. 2018.

[20] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba, "Places: A 10 million Image Database for Scene Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.40, no.6, pp. 1452-1464, Jul. 2017.