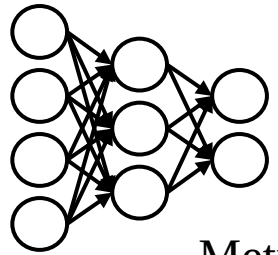
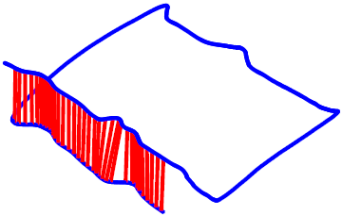
MetricNet: A Loop Closure Detection Method for Appearance Variation using Adaptive Weighted Similarity Matrix

Journal:	*IEEE Sensors Journal*
Manuscript ID	Draft
Manuscript Type:	Regular Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Li, Ying; University of Electronic Science and Technology of China, School of Information and Communication Engineering Zhu, Ran; University of Electronic Science and Technology of China, School of Information and Communication Engineering Yang, MingKun; University of Electronic Science and Technology of China, School of Information and Communication Engineering Xiao, Zhuoling Zhang, Yuhan Yan, Bo; University of Electronic Science and Technology of China, School of Information and Communication Engineering
Keywords:	DATP

images



MetricNet



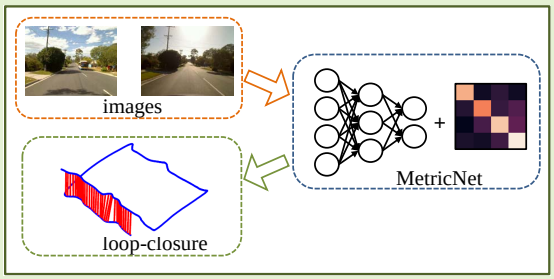
loop-closure

IEEE Sensors Council IEEE SENSORS JOURNAL, VOL. XX, NO. XX, XXXX 2017 1

MetricNet: A Loop Closure Detection Method for Appearance Variation using Adaptive Weighted Similarity Matrix

Ying Li, Ran Zhu, Mingkun Yang, Zhuoling Xiao*, Yuhan Zhang, Bo Yan,

***Abstract*— Loop Closure Detection (LCD), also known as the place recognition of pre-visited areas, is a significant optimization module in visual simultaneous localization and mapping (vSLAM) to reduce the accumulative error over time. This paper presents a novel end-to-end framework (MetricNet) for LCD to enhance the detection performance on complex and changeable scenes. Specifically, MetricNet introduces adaptive weighted similarity matrix by combining the feature extraction module and the similarity measurement module to focus on changing appearance over time. Experiments on three typical open datasets demonstrate that the proposed MetricNet outperforms state-of-the-art learning-based methods in terms of precision by up to 15% - 40% and recall rate by 10% - 30%, proving its promising generalization ability, applicability, and suitability for real-world applications.**



***Index Terms*— Simultaneous localization and mapping, Loop closure detection, Appearance variation, Convolutional neural network, Adaptive weighted similarity matrix**

I. Introduction

VISUAL simultaneous localization and mapping (vSLAM) that simultaneously recovers camera pose and scene structure from video, as one of the key autonomous positioning and navigation technologies in areas where GPS fails or cannot be covered, is gaining importance in robotic applications such as autonomous cars or unmanned aerial vehicles [1]. Loop Closure Detection (LCD), considered one of the essential parts in the visual SLAM system, is designed to recognize pre-visited areas by an autonomous mobile robot according to the image information collected by the visual sensors during the moving so also known as visual place recognition. Accurate LCD methods offer precise pose estimation by introducing extra constraints to correct the trajectory drift over time, improving the system performance [2]. However, there are still two common challenges: 1) the same place has different appearances at different times due to change of illumination and weather; 2) different scenes look similar for reasons such as sharing common objects. Therefore, an excellent LCD method needs to resolve these two problems to detect more correct loops.

In the field of vSLAM, the appearance-based methods treating LCD as an image matching problem compare the similarity between the current image and previous images. If the similarity between them is sufficiently high to exceed a given threshold, we can regard it as a loop closure. As one of the most popular LCD approaches, it can be divided into two crucial steps: feature generation and similarity measurement [3].

In the vSLAM system, image feature extraction forms the basis of a series of tasks, such as keyframe extraction, tracking, positioning, and map construction, which have a decisive influence on the robot's autonomous positioning. The traditional appearance-based methods mainly follow the visual bag-of-words (BoWs) [2], [4] model, which uses a clustering procedure on a training sample of local features and quantizes the descriptor space into Visual Words (VWs). However, this approach uses hand-crafted traditional features. Most of these features discard certain geometric and structural information, making it difficult to cope with the challenges such as camera motion and illumination changes. Moreover, BoWs relies on specific environments and has poor robustness to different application scenarios. In recent years, deep learning has made significant breakthroughs in the field of computer vision. Much related research demonstrates that the deep features learned by convolutional neural networks (CNNs) can provide more robust image representations in changeable environmental conditions, especially when illumination change and viewpoint variation [5]. Besides, the network models trained for specific tasks can be transferred to other tasks successfully. Some classic pre-training network models' availability makes

This work was supported by National Natural Science Foundation of China Grant No. 61703076 and No. 61973056.

The authors are with the Department of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: {yingli, ranzhu, mingkunyang}@std.uestc.edu.cn, zhuolingxiao@gmail.com, 2017200603025@std.uestc.edu.cn, yanboyu@uestc.edu.cn.)

it more convenient to complete various feature extraction tasks [6]. Based on the above reasons, researchers began to apply CNNs to LCD [7]–[9]. Although good results have been achieved, these works paid little attention to similarity measurement strategies, which also play a crucial role in image matching tasks. Fixed pre-specified distance metrics such as Euclidean distance or cosine similarity are commonly used. Moreover, feature extraction exists independently of similarity measurement, its effectiveness severely restrains the similarity measurement's effect. For these reasons, further improvement of detection precision is hindered.

This paper proposes a two-branch end-to-end network called MetricNet that performs LCD based on the adaptive weighted similarity matrix, jointly optimized with feature extraction, to address the issues mentioned earlier. In summary, the key contributions are as follows:

- Adaptive feature selection and similarity matrix: By using the channel weighting to remove the irrelevant background information and a weighted similarity matrix to adaptively select spatial information, MetricNet has been proved to be more robust and accurate than competitive models in both theory and practice.
- An end-to-end LCD framework: this research also proposes a novel LCD framework that encompasses both feature extraction and similarity measurement to extract more effective representation. The learning-based MetricNet can achieve a recall rate up to 47.08% under 100
- Extensive multi-dataset validation: MetricNet has proved its excellent performance on three typical open datasets under drastic illumination and seasonal variations. Outstanding improvements in detection accuracy and generalization ability are also further demonstrated.

The remainder of this paper is organized as follows: In Section II, a brief introduction on the feature extraction and similarity measurement are provided. Section III focuses on the proposed architecture and methodology. Section IV presents experimental results and comparisons on three datasets. Finally, the conclusion of this paper is drawn in Section V.

II. Related work

This paper focuses on LCD based on images. Researchers have proposed many effective approaches to implement it, and the simplest one is by matching keyframes through the similarity between them. In this case, detecting a loop is essentially an image matching problem, including feature extraction and similarity measurement. And this section will briefly review representative research in terms of how it is related to our work reported in this paper.

A. Feature Extraction

The study of the feature extraction method has been of interest to researchers for some time. Earlier works on feature extraction tended to adopt features artificially designed by researchers in the field of computer vision, such as SIFT [10], SURF [11], ORB [12] and BRIEF [13], etc. The visual image descriptors are usually divided into two categories: local descriptors and global descriptors, and in which, BoWs is the most successful as mentioned in Section I. It clusters extracted local features into "words" by k-means and describes images in the "words" vector. The BoWs descriptors have been successfully applied to LCD under various scenarios and have achieved outstanding results. Especially Cummins et al. [14] proposed FAB-Map, which represents the high level of the current development of loop detection. It extends the BOWs model and learns a generation model using the data of BOW. This method is not limited to positioning but can determine whether new observations come from places already on the map or not previously seen and can solve the perception bias well. Fisher Vector (FV) [15] constructs a visual dictionary using the Gaussian Mixture Model (GMM). It describes images using the gradient vector of the likelihood function of GMM, where the Gaussian component is similar to the clustering center in BoW. Vector of Locally Aggregated Descriptors (VLAD) [16] is a simplification of FV. Unlike FV, VLAD accumulates the image's residuals on the cluster center and combines them of each cluster center as an image descriptor. BoWs, FV, and VLAD are all based on local descriptors. Different from them, GIST [17] uses Gabor filters to generate low-dimensional global image descriptions. However, based on hand-crafted features, these image descriptors discard certain geometric and structural information, making it difficult to cope with the challenging environment such as intense camera motion and illumination change.

The emergence of deep learning has accelerated the development of relevant technologies of computer vision. Feature extraction methods based on deep learning have achieved great success in image recognition [18], classification [19], and retrieval [20], which provides a new way to address LCD problem and has attracted intensive attention form researchers. To construct a descriptor that can better describe images, Xiang Gao et al. [21] took advantage of Auto-encoder to extract image features and used the similarity matrix to detect closed loops, which improved the impact of illumination variation and got high accuracy on open datasets. Yi Hou et al. [9] proposed to use the convolutional neural network (CNN) to do it. They performed a comprehensive evaluation of the outputs at the intermediate layers of a CNN as image descriptors. The results showed that the abstract high-level features extracted from multi-layer neural networks outperformed state-of-the-art competitors when lighting changes significantly. They were also considerably faster to extract than the state-of-the-art hand-crafted features even on a conventional CPU and are two orders of magnitude faster on an entry-level GPU. To get better recognition results, Kai Qiu et al. [22] proposed the Siamese-ResNet network, which combines the Siamese network with ResNet to detect loop closure. Compared with FabMap2.0, Siamese-ResNet shows higher accuracy, better robustness, and less time-consuming. These above methods all show that CNNs have more potent power to characterize images; the deeper features learned by CNNs are significantly superior to hand-crafted features in visual tasks. They can provide more robust image representation in changeable environmental conditions, especially when illumination change and viewpoint vary.

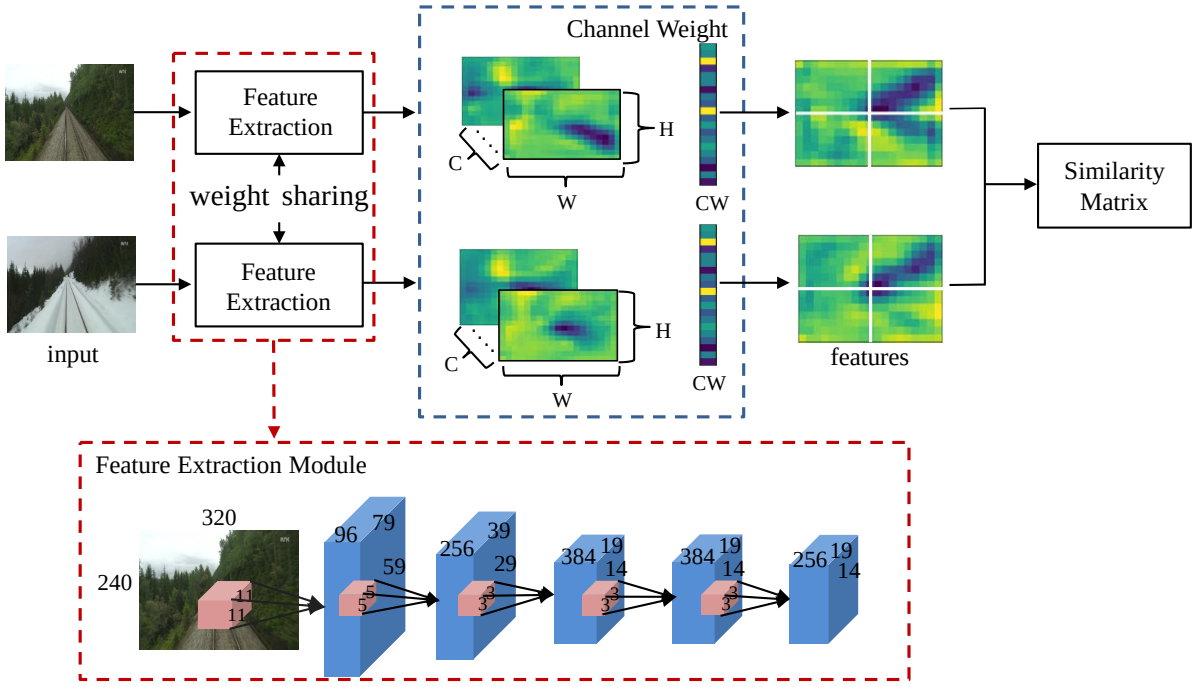


Fig. 1. The pipeline of the proposed MetricNet.

B. Similarity Measurement

Similarity measurement is a significant step in loop closure detection, which is generally measured by the distance between descriptors of images and determines whether the images are captured from the same place. If the distance between the two images is less than a given threshold, then a loop closure is considered to have occurred. There are three commonly used types of distance metrics: Euclidean distance, cosine distance, and hamming distance. To evaluate the effectiveness of different measures, Shahid et al. [23] retrained pre-trained Places-AlexNet using the sizeable open dataset Nordland. They compared several distance measurements: pairwise Euclidean, pairwise cosine, triplet Euclidean, and triplet cosine. The experiment demonstrates that cosine distance has a better performance than Euclidean distance and better works in recognizing scenes. However, the aforementioned fixed distance metrics are based on the given off-the-shelf features and only compare the surface similarity of features at the element level. Also, these methods completely separate feature extraction and similarity measurement so that the measurement results of similarity will be significantly affected by the effectiveness of features. In this paper, unlike the currently available research, we join feature extraction forces with similarity measurement to use the similarities of feature patches to constitute a similarity matrix that contains the spatial information of pictures. Experiments show that the algorithm has better generalization in the similarity measurement step.

III. SYSTEM MODEL

In this session, we first describe the overview of the proposed method's framework and then introduce each module in detail.

This paper uses the SAES metric layer from Chenyang Zhao et al. [24] as a reference that introduced an end-to-end network into loop closure detection. The proposed improved network called MetricNet, mainly involves two modules: a feature extraction (FE) module and a similarity metric (SM) module, as shown in Fig. 1. The neural network input is an image pair composed by current image I_1 and previous I_1 with label 1 or 0, which indicates whether this pair is a loop closure or not. There are two identical branches in the FE module with shared weights, which map the image pair into the feature space and divide them into four equal patches, respectively. And then the feature of a single image sent into the SM module can be represented as $(f_1,\ f_2,\ f_3,\ f_4)$. The following SM module will give a value in the range of 0 to 1 as the verdict on the image pair's similarity. Finally, we can determine whether a loop closure occurs on the output of the SM module.

A. Feature Extraction

We employed the AlexNet [25], a state-of-the-art framework for computer vision tasks, as our feature extraction module, as shown in Fig. 1. There are 5 convolution layers containing ReLU activation function and max-pooling followed by three fully connected layers and a subsequent soft-max layer. While the capability of extracted representations is enhanced as the neural network goes deeper, it has been demonstrated in [9] that the features extracted from the fully connected layers cannot make ideal representations due to the loss of spatial information in the image. Therefore, our module only preserves the 5 convolution layers (5CONVs), discarding the subsequent fully connected layers and the softmax layer.

Considering the equal treatment towards feature extracted from the part with different distinguishability could lead to more false positives in loop closure detection, weights for scenes whose distinguishability is relatively small, such as the sky and ground, should be reduced. Inspired by [26], we adopt the self-adaptive channel weight (CW), which is similar to the inverse documentary frequency (IDF) in BOW, to draw the attention of our module to the low-frequent features, which are more differentiated. The CW is defined as:

$$T_c = \frac{\sum_{X_{h,w}>0} 1}{H \times W} \tag{1}$$

$$CW_c = \begin{cases} \log(\frac{\sum_{c=1}^{C} T_c}{T_c}), T_c > 0 \\ 0, T_c = 0 \end{cases} \tag{2}$$

where H, W denotes the size of a feature map. $X \in R^{H \times W}$ represents one feature map. c, h, w is the location of the feature. T_c denotes the average response of the feature of the c-th channel. CW_c denotes the weight of the c-th channel.

Then we can calculate the weighted feature-maps F_{weight} as follows:

$$F_{weight} = F \times CW \tag{3}$$

where $F \in R^{(C \times H \times W)}$ denotes the 3-dimension features of the 5CONVs.

B. The Adaptive Weighted Similarity Matrix

To use the spatial information of image pairs to improve the ability to cope with the appearance changes caused by illumination or seasons, we divide the obtained feature map into four equal patches and calculate each patch's similarity to form the similarity matrix. It takes advantage of the natural properties of convolutional network sliding Windows, which is that there are some overlaps between the windows and preserves the context at the patches' boundary. The final similarity between the two images is determined by the data distribution in the similarity matrix.

For a given image pair, the corresponding features obtained by dividing the feature map are F^1_{weight} and F^2_{weight}:

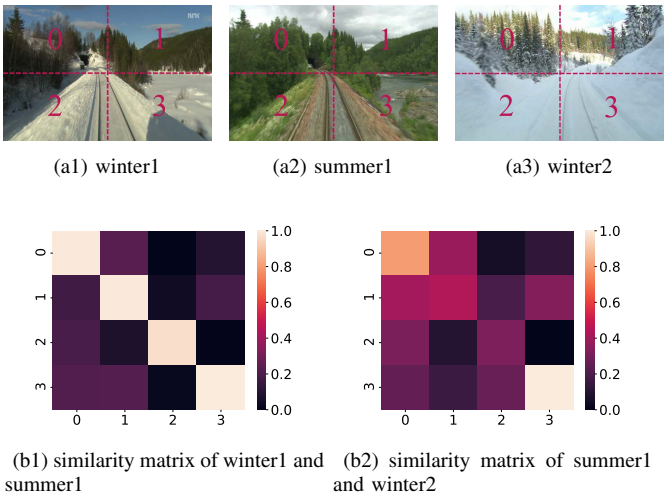
$$F^i_{weight} = \{f_{i1}, f_{i2}, f_{i3}, f_{i4}\}, i \in [1, 2] \tag{4}$$

where f_{ij} represents the features generated from the j-th feature map patch of the i-th image. The cosine between feature vectors can express the similarity between image patches:

$$SM_{ij} = \cos < f_{1i}, f_{2j} > \tag{5}$$

where f_{1i} and f_{2j} are the feature descriptors of 2 feature map patches, respectively. SM_{ij} is the similarity between feature map patches. The larger the cosine is, the more similar the feature patches are. Finally, the whole similarity matrix SM is given by:

$$SM = F^1_{weight} \cdot (F^2_{weight})^T = \{SM_{ij}, 0 \le i \le 3, 0 \le j \le 3\} \tag{6}$$



(a1) winter1 (a2) summer1 (a3) winter2

(b1) similarity matrix of winter1 and summer1 (b2) similarity matrix of summer1 and winter2

Fig. 2. The input images and the corresponding similarity matrixes. Winter1 in (a1) and summer1 in (a2) are a positive pair, summer1 in (a2), and winter2 in (a3) is a negative pair.

The matrix reflecting the similarity between feature patches of image pairs are visualized in Fig.2 in which the gray value represents the extent of the similarity. In Fig. 2(b1), the values on diagonal are obviously overall higher than those on off-diagonal, while values do not show such a pattern in Fig. 2(b2). It means that positive pairs and negative pairs can be distinguished with a resort to such characteristics.

Based on the similarity matrix's data distribution, this paper proposes an adaptive weighted similarity measurement method and defines the overall similarity S of the image pair as:

$$S = \alpha \sum_{i=0}^{3} \omega_i SM_{ii} \tag{7}$$

where SM_{ii} is the similarity of patches with the same index, α represents the probability that the image pair come from the same location, and ω_i is the weight of the diagonal elements and satisfies the following formula:

$$\sum_{i=0}^{3} \omega_i = 1 \tag{8}$$

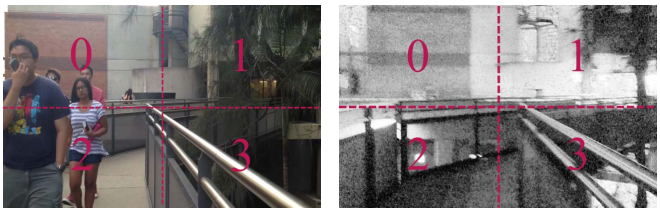
According to the previous analysis, whose values on diagonal in the similarity matrix are significantly greater than off-diagonal, it is very likely to be a positive pair. Therefore, the likelihood alpha of the same location it is defined as:

$$\alpha = \begin{cases} \frac{e^d - 1}{e - 1}, & d \ge 0 \\ 0, & d < 0 \end{cases} \tag{9}$$

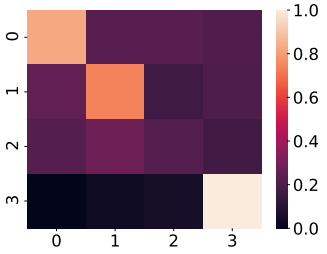
where d represents the difference between diagonal and off-diagonal values linked to the average value of diagonal similarities S_{dia} and off-diagonal similarities S_{off} in the matrix.

$$d = S_{dia} - S_{off} \tag{10}$$

$$S_{dia} = \frac{1}{4} \sum_{i=0}^{3} SM_{ii} \tag{11}$$



(a1) image captured during day (a2) image captured during night



(b) Similarity Matrix of (a1) and (a2)

Fig. 3. An example of a location that varies greatly at different times.

$$S_{off} = \frac{1}{12}\sum_{i \neq j} SM_{ij} \tag{12}$$

In this case, if $S_{dia} \gg S_{off}$, α approaches to 1, it means there's a high probability of having a positive pair, so the final similarity is the weighted average of diagonal patch similarities. On the contrary, the closer S_{dia} is to S_{off}, the smaller α is. So, if $S_{dia} < S_{off}$, it is highly possible that the two images are captured from different places, thus setting α to 0 directly.

In terms of some places, such as the pavements, the same location may have relatively large differences at different times. As shown in Fig.3, the two images captured from the same location have distinct differences in the lower left part due to the pedestrians blocking. Hence the diagonal value (SM_{22}) calculated by these patches is no longer significantly larger than that of the off-diagonal value associated with it, which is called the SM_{22} outlier. By giving it less weight ω_i, we can reduce its influence on the overall similarity. These weights can be formulated as:

$$\gamma_i = \frac{(\sum_{i,j \neq i} SM_{ij} + \sum_{j,i \neq j} SM_{ij})}{6} \tag{13}$$

$$k_i = \begin{cases} SM_{ii} - \gamma_i, SM_{ii} > \gamma_i \\ 0, SM_{ii} < \gamma_i \end{cases} \tag{14}$$

$$\omega_i = \frac{k_i}{\sum_i k_i} \tag{15}$$

where γ_i is the average value of the off-diagonal values of the similarity matrix calculated from the image pair's i-th patch. k_i is the distance of diagonal and off-diagonal values generated by the i-th patch. ω_i denotes the weight of the diagonal elements. These weights are used to reduce the influence of outliers and ensure the overall similarity.

TABLE I
DETAILS OF THE TESTING DATASETS

Dataset	Environment	Appearance Variation	Viewpoint Variation
Gardens Point	Campus	Day-Night	Moderate
Nordland	Train journey	Winter-Summer	Small
Stlucia	Suburban	Morning-Afternoon	Moderate

C. Training

We constituted the training dataset by sampling image pairs from the SPED 900 dataset [27]. The PyTorch framework implements the network on an NVIDIA Geforce Titan XP GPU. To speed up network training, we used the parameters of AlexNet 5CONVs pre-trained by the ImageNet dataset [28] as initial parameters in the feature extraction module. Adam [29] with β_1 =0.9, β_2 = 0.99 is used as the optimizer to train the network with batch size of 64 due to the limit of the memory. The initial learning rate is set to 0.001 and reduced by 0.1 times every 30 epochs. Besides, early stopping technologies are introduced to prevent the model from overfitting. In our experiments, the scale of RGB images is resized to 320×240×3 before fed into the network, and the ground truth is processed to a binary classification with the label space 0, 1. We chose binary cross-entropy (BCE) as the loss to train the model, which can be defined as:

$$Loss = -\frac{1}{n}\sum_i y_i * \log S_i + (1 - y_i)\log(1 - S_i) \tag{16}$$

where S_i and y_i correspond to the similarity score and the label of the image pair sample, respectively. The label of positive pairs is 1 and of negative is 0.

IV. EVALUATION AND ANALYSIS

In this section, we first present the details of three publicly available datasets used in the experiments. Then, comparisons performed against several state-of-the-art approaches, such as SAES [24], CALC2.0 [30] and Place-ResNet [31] to verify the feasibility and effectiveness of the proposed method. Finally, we evaluate and analyze the proposed framework in terms of feature extraction and similarity measurement methods.

A. Datasets

To evaluate the robustness of the proposed system in response to appearance changing conditions, we choose three public datasets as the test set (i.e., GardensPoint dataset [32], Nordland dataset [33], and Stlucia dataset [34]). Details of these datasets are shown in Table I, and some examples of the real scene are shown in Fig. 4.

The Gardens Point dataset is collected at Queensland University of Technology campus by traversing the walkways in the daytime (along both sides) and the night (only along the right side). The day-right and night-right pairs are adopted as

(a1) DAY (a2) NIGHT

(b1) SUMMER (b2) WINTER

(c1) SL0845 (c2) SL1545

Fig. 4. Some sample images of the image sequences we use in the experiments.

DAY-NIGHT to assess the performance tackling the significant illumination variation. Besides, the method for generating its ground truth is to select correspondences of the frame manually.

Nordland dataset is produced from a TV documentary that records a train journey covering four seasons' appearance variations. By arranging images from different seasons following their positions, the ground truth can be constructed based on image indexes. Winter-summer images (denoted as WINTER and SUMMER) with the most significant appearance variations are selected as part of the test set.

For the Stlucia dataset, the images are captured in the suburbs at five different times (8:45, 10:00, 12:10, 14:10, 15.45) of the day with significant appearance changes due to illumination and GPS logs obtain its ground truth. In our evaluation experiment, the SL0845 and SL1410 collected at 8:45 and 14:10, respectively, are selected to form the test set because they have the most significant contrasts among all pairs.

B. Evaluation Methods

Loop closure, essentially, is a binary classification task with only two outcomes, a loop closure or not a loop closure. Therefore, we utilize the precision-recall curve (PR-Curve), a standard method to evaluate binary classification, to quantify the proposed LCD method's effectiveness. For this problem, the correct detections are considered as true positives(TP), the incorrect detections are known as false positives(FP), and the ground-truth loops undetected are defined as false negatives(FN). In the image of PR-Curve, the curve, which is closer to the upper right corner, has better performance. To produce the PR-Curve of given datasets, we compute the similarity for each pair of images. A threshold on the similarity is then applied to determine if loop closure has occurred, and a precision and recall pair results after all images in the datasets are considered and defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

Precision denotes the ratio between the number of correct positive and all positive results. While the recall is the ratio between correct positive and all the loops in the ground-truth. We calculate different values of precision and recall by varying the threshold so that we can plot a PR-curve. There are many ways to interpret PR-Curve, we mainly use: (1) the average precision (AP), where a high precision over all recall values is desirable, and AP is calculated by using (19); and (2) the maximum recall rate at 100% of precision, denoted by R, where again a higher value is desired. These two criteria are also useful when we need scalar values to characterize the overall performance of loop closure detection.

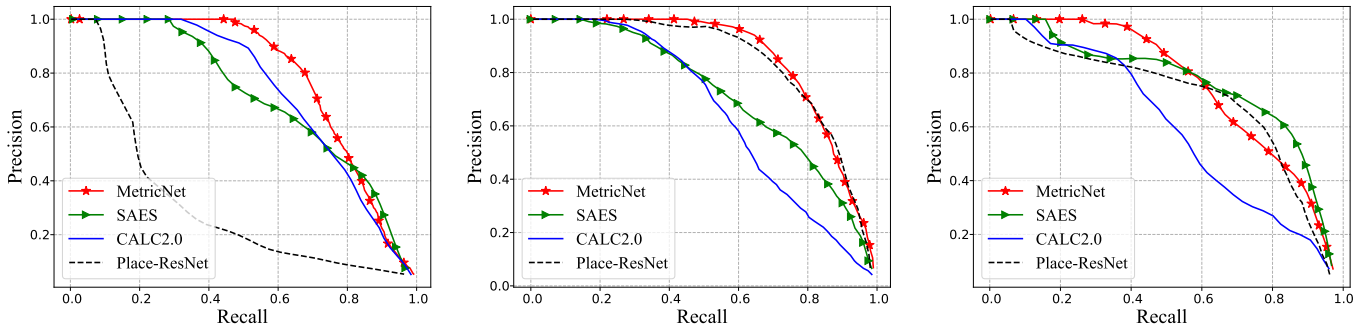
$$AP = \int_0^1 P(r)\mathrm{d}r \tag{19}$$

where $P(r)$ denotes the PR-curve.

C. Experimental Results

1) Comparison with state-of-the-art: Comparisons are made on Gardens Point, Nordland, and Stlucia datasets between our method and SAES, CALC2.0, and Place-ResNet, which are state-of-the-art loop closure detection approaches.

The PR-Curve graphs are drawn based on the experimental results, and the curves are shown in Fig. 5. It can be seen from the figures that with a higher level of similarity threshold, the precision is more heightened and recall is lower for all methods. As the decrease of similarity threshold, precision drops, and recall increases. Therefore, there is a mutual restriction relationship between precision rate and recall rate. In practical application, the similarity threshold must be set according to the actual environment as a balance point to make the precision and recall rate relatively higher. It is also obvious from the figures that the PR-curve of the proposed MetricNet is closer to the upper right corner than other methods. On the one hand, this means that the proposed MetricNet can get relatively higher precision and recall at the same time than other methods. On the other hand, the improvement in three test datasets captured in environments different from the training dataset shows that our approach has better robustness to environmental changes when applied to loop closure detection. These improvements mainly benefit from the consideration of spatial information and the contextual information of the images. The precision rate of MetricNet is lower when the recall rate is high on Gardens Point and Nordland datasets because these two datasets have



(a) GardensPoint (b) Nordland (c) Stlucia

Fig. 5. The PR-Curves of different networks on (a) Gardens Point dataset, (b) Nordland dataset and (c) Stlucia dataset.

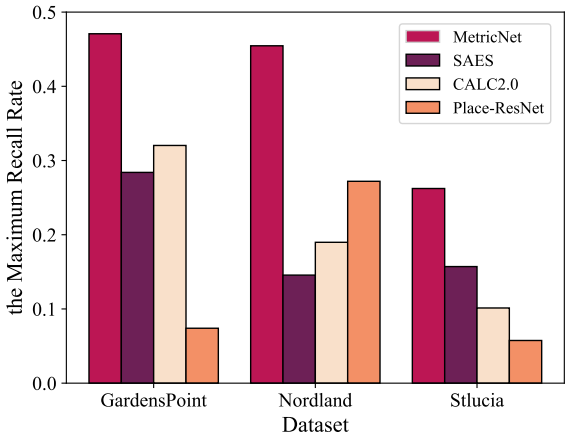


Fig. 6. The maximum recall rate at 100% of precision of different networks.

more serious viewpoint changes and more dynamic objects than the training dataset. Additional training datasets with viewpoint variations and dynamic objects can easily address this in real-world applications.

Fig. 6 shows the maximum recall rate at 100% of precision of our method and the competing approaches. It has demonstrated that the proposed method can almost achieve the highest recall rate with 100% of precision. The recall rate reaches up to 43.97%, 44.49%, and 26.72% on three datasetsrespectively, which are improved by up to 10% - 30% compared with other methods. The result means that fewer real loop closure can be missed leveraging our approach, which is expected in the practical tasks. This improvement is principal because we use the feature block to construct the similarity matrix and assign adaptive weights to each element to enhance positive pairs' characteristics, making positive pairs easier to distinguish. It can also be seen that the proposed MetricNet performs equally well with different challenging datasets and can achieve the highest recall rate with 100% precision.

Table II shows the comparison of average precision on three datasets calculated from the PR-curve of each method. Machine learning theory points out that the larger the area enclosed by the PR-curve and the coordinate axis, the higher the average precision and the better the algorithm's performance.

TABLE II
AP COMPARISON OF DIFFERENT NETWORKS

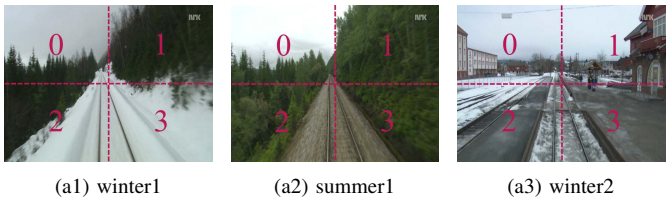
Dataset	SAES	CALC2.0	Place-ResNet	MetricNet
GardensPoint	0.6997	0.7284	0.3048	**0.7831**
Nordland	0.7162	0.6348	0.8209	**0.8458**
Stlucia	0.7098	0.5869	0.6860	**0.7476**

It can be observed that the AP of the proposed method is higher than that of other comparison methods on each dataset, respectively, which demonstrates the satisfactory performance of the proposed method.

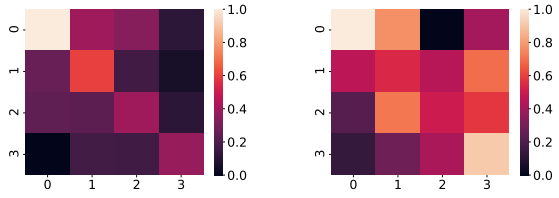
2) Evaluation of Feature Extraction Methods: To evaluate the feature extraction module's performance, we compared our feature Extraction method with the most common approach based on image patch matching. In the contrast experiment, each image is directly divided into four patches and sent to the FE module. Then the features of every patch are extracted respectively to construct the similarity matrix.

Row 1 of Fig. 7 shows example images from the Nordland dataset. And the images in Fig. 7(a1) and Fig. 7(a2) origin from the same place, and their similarity matrices obtained by our method and competing method are shown in Fig. 7(b1) and Fig. 7(c1), respectively. And the images in Fig. 7(a2) and Fig. 7(a3) are captured from different places, and their similarity matrices obtained by these two methods are shown in Fig. 7(b2) and Fig. 7(c2), respectively. Although the competing method and our method retain similar matrix characteristics, compared with the competing method, our method enlarges the difference between the positive pair and the negative pair. For positive pairs, the diagonal values are far more than the off-diagonal values, while the value is even more random regarding negative pairs. This makes the difference between the positive and the negative more prominent and thus easier to distinguish.

To prove this effect more clearly, we calculate the similarity differentials between the mean of diagonal values and the mean of off-diagonal values in the similarity matrix of positive pairs and negative pairs on three datasets obtained by the two methods and define them as d_1 and d_2 respectively. Then we plot the probability density distribution of $\Delta = d_1\text{-}d_2$ for positive pairs and negative pairs of each dataset. As shown in

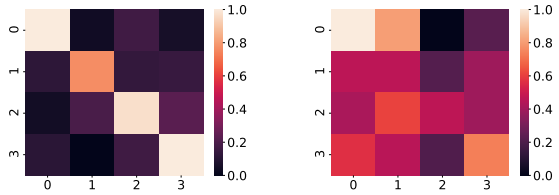


(a1) winter1 (a2) summer1 (a3) winter2



(b1) Similarity matrix of winter1 and summer1 obtained by dividing the images directly

(b2) Similarity matrix of winter2 and summer1 obtained by dividing the images directly



(c1) Similarity matrix of winter1 and summer1 obtained by MetricNet

(c2) Similarity matrix of winter2 and summer1 obtained by MetricNet

Fig. 7. The similarity matrixes obtained by dividing the images directly and obtained by MetricNet.

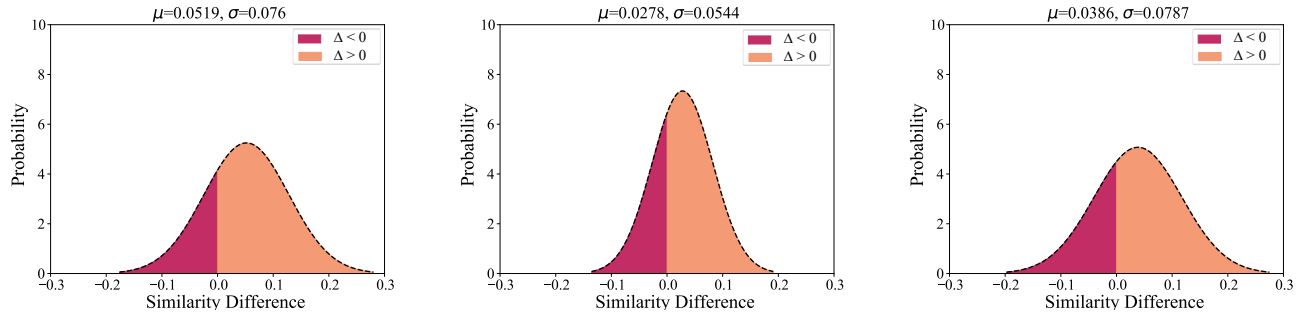
Fig. 8, the horizontal axis represents the similarity difference of Δ, and the vertical axis represents the corresponding probability density. It obeys distributed on all datasets, and the corresponding probability density function can be expressed as:

$$f(\Delta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\Delta-\mu)^2}{2\sigma^2}} \tag{20}$$

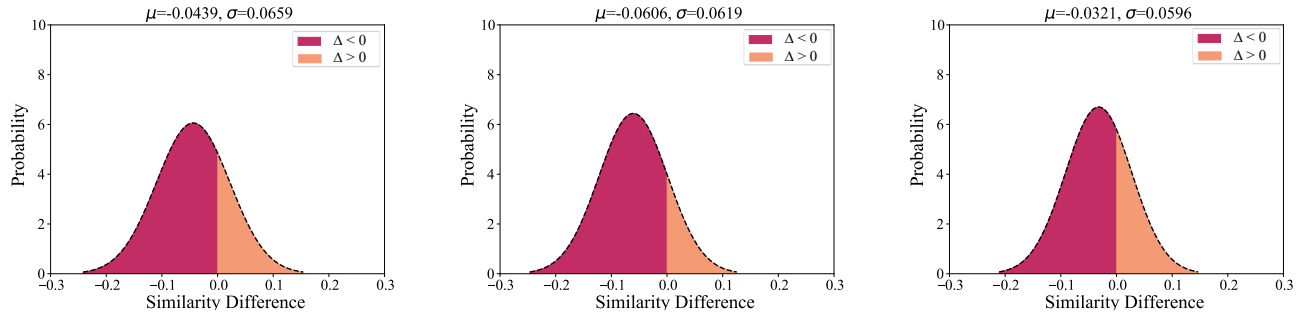
where μ and σ denote the mean and standard deviation of Δ, respectively.

And the area under the curve represents the probability of Δ, if $\Delta>0$, we fill the area under the curve with orange, otherwise fill it with mediumvioletred. As shown in Fig. 8, for the positive pairs of the three datasets, most of the similarity differences obtained by our method are larger than those obtained by the comparison method. While, for the negative pairs, the differences brought by our method are relatively small. The results show that the extracted features enhance the difference between the positive pair and the negative pair of the similarity matrix, making it easier to judge them, which is what we expect. The results show that using the convolutional network's sliding window to consider the context connection between different feature patches is better than directly dividing the image. Simultaneously, removing background information by channel weighting can further remove the redundancy of features and make the extracted features more discriminative.

3) Comparison of Similarity Measurement Methods: We also compare our adaptive weighted similarity matrix with the other three conventional approaches (e.g., cosine, euclidean, and average similarity) of similarity measurement on three

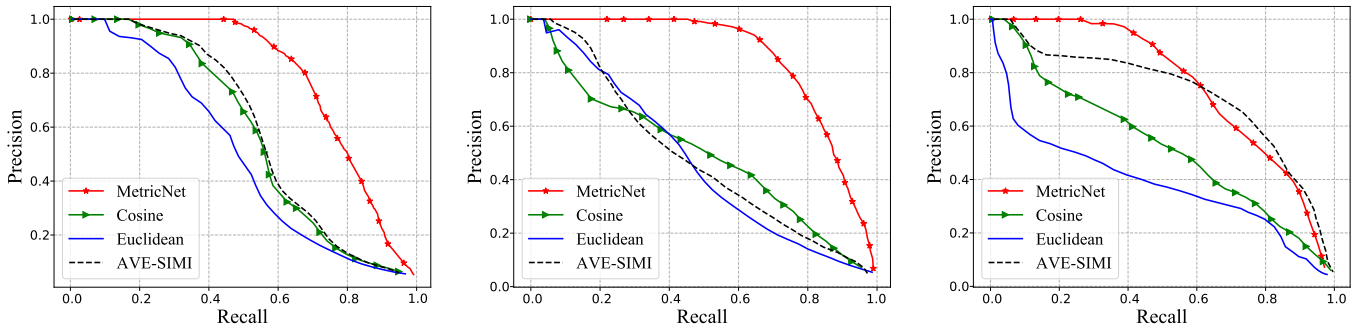


(a1) positive pairs of GardensPoint (b1) positive pairs of Nordland (c1) positive pairs of Stlucia



(a2) negative pairs of GardensPoint (b2) negative pairs of Nordland (c2) negative pairs of Stlucia

Fig. 8. The probability density distribution of the difference of the similarity between our method and the method of dividing image directly.



(a) GardensPoint (b) Nordland (c) Stlucia

Fig. 9. The PR-Curves of different measurement methods on (a) Gardens Point dataset, (b) Nordland dataset and (c) Stlucia dataset.

TABLE III
AP COMPARISON OF DIFFERENT MEASUREMENT METHODS

Dataset	Euclidean	Cosine	AVE-SIMI	MetricNet
GardensPoint	0.4932	0.5663	0.5805	**0.7831**
Nordland	0.4638	0.4885	0.4736	**0.8458**
Stlucia	0.3828	0.5205	0.5659	**0.7476**

datasets. For the methods of cosine and euclidean, the images are fed into feature extractor to generate features without being divided into four parts, and similarities are calculated by features directly. For the average similarity (AVE-SIMI), we directly use the average value of all elements of the similarity matrix as the final similarity.

Fig. 9 shows the performance of the proposed MetricNet and the competing approaches on three datasets. It can be observed that both competing approaches and MetricNet can achieve higher precision at a low recall rate, but in a high recall rate, MetricNet has the best performance. We conclude that our method can obtain more real loops when the recall rate is high, which we expect to see in practice. This similarity between images is reflected in the feature vector's size and the direction of the feature vector. We supplement the direction information by constructing a similarity matrix. For this reason, the proposed method can get much better performance than other methods.

Fig. 10 shows the maximum recall rate at 100 % of precision of each method. Among all of the datasets, compared with other methods, the recall rate can be increased by 10% - 40%, especially compared with the euclidean method. The result means we will miss fewer loop closures, which can bring great help in practical application.

Table III shows the average precision of the proposed MetricNet and comparison method. As we can see, the proposed method can get the maximum average precision, which proves that our method for measuring similarity has relatively high precision in all recall rates. It has the best performance on the whole. Similar results also can be seen in Fig. 11. The good performance on each dataset shows that the proposed MetricNet has good generalization for different appearance variations. All of the above improvements are due to the full

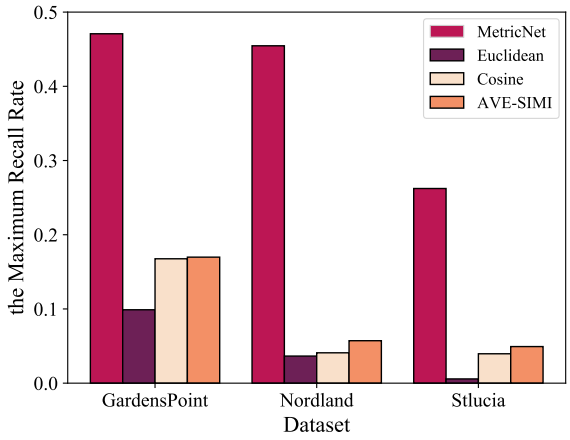


Fig. 10. The maximum recall rate at 100% precision of different measurement methods.

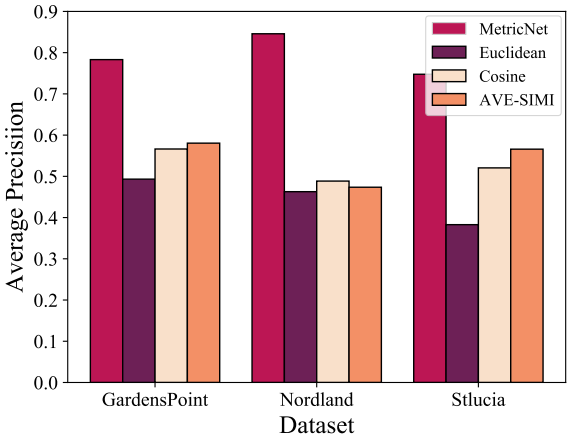


Fig. 11. Ap comparison of different measurement methods.

use of image information, considering the similarity between the appearance features of the image itself and spatial information.

4) Computational performance: The average computational cost required by MetricNet is compared with SAES in terms of the running time for (1) feature extraction and processing

TABLE IV
PROCESSING TIME COMPARISON ON THE NORDLAND DATASET

Algorithm	Processing Time(s)		
	Feature Extraction	Similarity Matrix Construction	Similarity Calculation
MetricNet	0.0370	0.0055	0.0022
SAES	0.0218	0.0065	0.0025

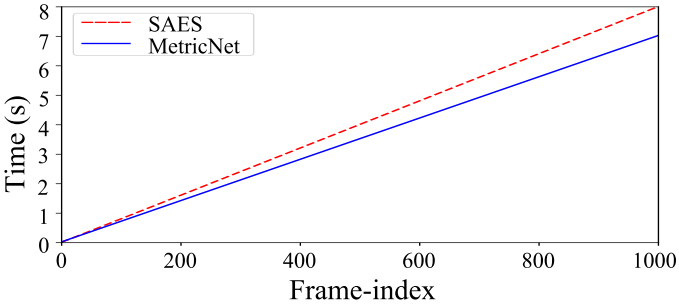


Fig. 12. Comparison of time cost of MetricNet and SAES in loop closure detection.

by the neural networks models, (2) construction of similarity matrix, (3) similarity calculation between the image pair of the current and previous image. The Nordland dataset is used to test the time consumption, as it has the largest database, as these two networks contain the three modules to be compared. Table IV shows the average time consumption of the two tested networks. The algorithms are conducted on an NVIDIA Geforce Titan XP GPU. As shown in Table IV, feature extraction takes most of the execution time due to a large number of computations in the deep networks for these two algorithms, and our method takes more time because of the channel weighting, further optimizations can reduce the execution time. But for the other two modules, our algorithm takes less time.

In the process of loop closure detection, it is necessary to extract features of the current image in real time and construct similarity matrix and calculate the total similarity between the current image and the previous image. With the increase of previous images, the loop closure detection time of each frame will gradually increase. According to Table IV, the loop closure detection time-consuming models of the two algorithms can be obtained respectively. As shown in Fig. 12, with the increase of the previous images, the MetricNet has higher efficiency and better real-time performance.

V. Conclusion

This paper presents a novel framework that solves the detection problem caused by illumination and seasonal variations in the LCD task. In the framework, the feature extraction and similarity measurement are trained and deployed in an end-to-end manner. By introducing channel weighting, the irrelevant background information is removed. Besides, we utilize the spatial information of images by constructing a weighted similarity matrix to measure the overall similarity adaptively. The extensive experiments on three different datasets verify that the MetricNet outperforms many learning-based algorithms and produces a promising generalization in appearance variations image pairs.

In the future, we plan take into account further challenges such as viewpoint changes and apply our algorithm to the real-world vSLAM system. Besides, since we detect loops based on similarity scores in this paper, instead of the current fixed measurement method , we will try to use the learning-based way to model a measurement function in a data-driven manner to pursue higher performance.

References

[1] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial intelligence review*, vol. 43, no. 1, pp. 55–81, 2015.
[2] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.
[3] M. Labbe and F. Michaud, "Appearance-based loop closure detection for online large-scale and long-term operation," *IEEE Transactions on Robotics*, vol. 29, no. 3, pp. 734–745, 2013.
[4] M. Cummins and P. Newman, "Appearance-only slam at large scale with fab-map 2.0," *The International Journal of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.
[5] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, pp. 1–2, Prague, 2004.
[6] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *2012 IEEE International Conference on Robotics and Automation*, pp. 1643–1649, IEEE, 2012.
[7] Y. Kong, W. Liu, and Z. Chen, "Robust convnet landmark-based visual place recognition by optimizing landmark matching," *IEEE Access*, vol. 7, pp. 30754–30767, 2019.
[8] T. Sun, M. Liu, H. Ye, and D.-Y. Yeung, "Point-cloud-based place recognition using cnn feature extraction," *IEEE Sensors Journal*, vol. 19, no. 24, pp. 12175–12186, 2019.
[9] Y. Hou, H. Zhang, and S. Zhou, "Convolutional neural network-based image representation for visual loop closure detection," in *2015 IEEE international conference on information and automation*, pp. 2238–2245, IEEE, 2015.
[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
[11] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*, pp. 404–417, Springer, 2006.
[12] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*, pp. 2564–2571, Ieee, 2011.
[13] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *European conference on computer vision*, pp. 778–792, Springer, 2010.
[14] 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.
[15] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *2007 IEEE conference on computer vision and pattern recognition*, pp. 1–8, IEEE, 2007.
[16] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3304–3311, IEEE, 2010.
[17] W. Tahir, A. Majeed, and T. Rehman, "Indoor/outdoor image classification using gist image features and neural network classifiers," in *2015 12th International Conference on High-capacity Optical Networks and Enabling/Emerging Technologies (HONET)*, pp. 1–5, IEEE, 2015.
[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[19] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral-spatial attention network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[20] 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 international conference on computer vision*, pp. 3456–3465, 2017.
[21] X. Gao and T. Zhang, "Loop closure detection for visual slam systems using deep neural networks," in *2015 34th Chinese Control Conference (CCC)*, pp. 5851–5856, IEEE, 2015.
[22] K. Qiu, Y. Ai, B. Tian, B. Wang, and D. Cao, "Siamese-resnet: implementing loop closure detection based on siamese network," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 716–721, IEEE, 2018.
[23] M. Shahid, T. Naseer, and W. Burgard, "Dtlc: Deeply trained loop closure detections for lifelong visual slam," in *Proceedings, Workshop on Visual Place Recognition, Conference on Robotics: Science and Systems (RSS)*, pp. 1–8, 2016.
[24] C. Zhao, R. Ding, and H. L. Key, "End-to-end visual place recognition based on deep metric learning and self-adaptively enhanced similarity metric," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 275–279, IEEE, 2019.
[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.
[26] C. Yu, Z. Liu, X.-J. Liu, F. Qiao, Y. Wang, F. Xie, Q. Wei, and Y. Yang, "A densenet feature-based loop closure method for visual slam system," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 258–265, IEEE, 2019.
[27] Z. Chen, A. Jacobson, N. Snderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3223–3230, 2017.
[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.
[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[30] N. Merrill and G. Huang, "Calc2. 0: Combining appearance, semantic and geometric information for robust and efficient visual loop closure," *arXiv preprint arXiv:1910.14103*, 2019.
[31] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.
[32] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4297–4304, IEEE, 2015.
[33] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons," in *Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA)*, p. 2013, 2013.
[34] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, "Fab-map+ ratslam: Appearance-based slam for multiple times of day," in *2010 IEEE international conference on robotics and automation*, pp. 3507–3512, IEEE, 2010.

Ying Li received the B.Eng. degree from Qingdao University, Qingdao, China, in 2018. She is currently pursuing the masters degree with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China. Her current research interests include indoor localization, machine learning, and information fusion.

Ran Zhu received the B.Eng. degree from Jilin University, Changchun, China, in 2018. She is currently pursuing the masters degree with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China. Her current research interests include indoor localization, and machine learning techniques for sensor networks and indoor localization.

Mingkun Yang received the B.Eng. degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2018, where he is currently pursuing the masters degree with the School of Information and Communication Engineering. His research interests focus on the application of machine learning techniques in sensor networks and indoor localization.

Zhuoling Xiao is an Associate Professor at the University of Electronic Science and Technology of China. He obtained his Ph.D at University of Oxford, became a postdoctoral researcher at University of Oxford. His interests lie in localization protocols for networked sensor nodes and machine learning techniques for sensor networks and localization. He has several international patent applications and over 30 papers published in leading journals and conferences including several best paper awards from leading conferences including IPSN and EWSN.

Yuhan Zhang is currently pursuing the B.Eng degree with the Glasgow College, University of Electronic Science and Technology of China, Chengdu, China. Her current research interests include indoor localization, and machine learning.

Bo Yan received the master's and Ph.D. degrees from University of Electronic Science and Technology of China(UESTC), Chengdu, China. And now she is a professor in School of Information and Communication Engineering at UESTC. Her current research interests lie in embedded system techenology, FPGA/ASIC design, and AI for Internet of Things(AIoT).