Loop Closure Detection Algorithm Based on Attention Mechanism

Zhangfang Hu, Wenhao Wang, Kuilin Zhu, Hongyao Zhou, Jiangtao Chen

***Abstract*—Loop closure detection is a key component of visual simultaneous localization and mapping (VSLM), which can effectively reduce the cumulative error of the system and improve the accuracy of mapping. The current loop closure detection method using deep learning can obtain accurate scene descriptions, but it is difficult to cope with the challenge of scene changes. This paper obtains complementary feature maps by fusing the shallow and deep image features of the ResNet50 network, and we incorporate a multiscale channel attention mechanism and a spatial attention mechanism after each layer of the ResNet50 network. The model can effectively extract discriminative scene landmarks, suppress the effects of irrelevant local features on similarity and be more robust to the problem of scene change. The method in this paper has been tested on several publicly available datasets and compared with mainstream methods. The experimental results show that the proposed method significantly outperforms mainstream methods in terms of accuracy-recall performance.**

***Index Terms*—Deep learning loop closure detection, channel attention, spatial attention**

I. Introduction

VISION simultaneous localization and mapping (SLAM) is the ability of a robot to explore unknown areas while maintaining the ability to construct reliable maps [1]. However, robots can generate cumulative errors during motion that can seriously affect the performance of SLAM systems. Loop closure detection is the process by which the robot identifies previously visited positions with vision sensor information during navigation, filtering out erroneous closed loops to correct incremental positional drift problems and reducing mapping errors.

Traditional loop closure detection algorithms include the scale-invariant feature transform (SIFT) [2], oriented FAST and rotated BRIEF (ORB) [3] and generalized search tree (GIST) [4]. These are based on keypoint matching; however, these descriptors and the visual features of the bag-of-words (BoW) [5] model are hand-designed and are not capable of representing the complex texture structure in an image. They are subject to dynamic environments and are sensitive to changes in illumination and thus have low success in detecting closed loops.

With the development of deep learning, researchers have used convolutional neural networks (CNNs) for loop closure detection, and complex features extracted from convolutional neural networks show high recognition ability and better robustness than traditional methods [6-7]. Hou et al. [8] used a place convolutional neural network (PlaceCNN) to extract image depth features, which addressed the problem of traditional methods being sensitive to illumination, but the high dimensionality of the extracted scene features made it difficult to meet the real-time requirements of loop closure detection, and the extracted features could hardly cope with changes in viewpoint in complex scenes. Traditional methods use hand-designed features that ignore some useful information. In response to the large dimensionality of the feature descriptors extracted by existing neural networks, it is difficult to meet the requirement of real-time performance. Guo et al. [9] improved the triple constraint loss function and performed feature extraction based on the DarkNet network framework to obtain a feature descriptor with lower dimensionality and better discrimination and combined it with a self-encoder to improve the detection speed, which could cope with scenes with significant illumination changes. However, the proposed method is less robust to viewpoint changes and dynamic environments. Zuo et al. [10] used pretrained convolutional neural networks for loop closure detection and compared several deep network models on publicly available datasets, including ResNet50, ResNet101 and ResNet152. The results showed that the ResNet50 network model had the best performance. However, it had difficulty in meeting the requirement of real-time performance. Conventional convolution neural networks lack scale feature extraction. Chen et al. [11] proposed a loop closure detection strategy for multiscale deep feature fusion using spatial pyramid pooling (SPP) to extract multiscale features on the basis of the AlexNet model [12]. To solve the problem of different input image sizes, SPP fuses the extracted features at different scales to compensate for the loss of image information caused by cropping the input image in the early stage, thus having a high accuracy rate. Although it is suitable

Manuscript received September 7, 2022; revised February 27, 2023. This work was supported in part by the National Natural Science Foundation Youth Fund Project (Grant No. 61703067), the Chongqing Basic Science and Frontier Technology Research Project (Grant No. Cstc2017jcyjAX0212), and the Chongqing Municipal Education Commission Science and Technology Research Project (KJ1704072).

Zhangfang Hu is a Professor at the Key Laboratory of Optical Information Sensing and Technology, School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail: 3565207151@qq.com)

Wenhao Wang is a graduate student of the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (corresponding author phone: 177-823-32706; e-mail: 1547889468@qq.com)

Kuilin Zhu is a graduate student of the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail: 1104569847@qq.com)

Hongyao Zhou is a graduate student of the Media Arts, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail: 1196940317@qq.com)

Jiangtao Chen is a graduate student of the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail: 1185007137@qq.com)

for applications with significant illumination changes, it has equal difficulty coping with viewpoint changes.

Many loop closure detection methods only aim at static and single environments, and while they can cope with changes in lighting, they still face significant challenges in dealing with changes in scenes caused by moving objects and changes in viewpoint. To address the above issues, we propose a loop closure detection method with a hybrid attention mechanism. It consists of two modules: one is a multiscale channel attention module, and the other is a spatial attention module. The multiscale channel attention module uses multiple convolutional kernels and channel attention layers to make effective use of multiscale information and feature relationships between channels. The channel attention layer is responsible for selecting and reweighting the most salient features from the input feature map, giving them higher weights. The spatial attention layer highlights more discriminative areas. The features learned are refined through a hybrid attention module. Simultaneously, to make use of features from different layers and compensate for information loss, we improve the performance of loop closure detection by fusing features from shallow and deep layers. Experiments show that our method can give higher weights to features with high differentiation in the scene, thus better coping with scene problems such as moving objects and local occlusions.

The organizational structure of this paper is as follows: Section II briefly introduces the relevant work in the field of closed-loop detection. Section III introduces the details of our method. Section IV shows the experimental results of our method on three public datasets. Section V summarizes the content of this paper.

II. Related Work

Early loop closure detection methods such as Cummins et al. [13] proposed fast appearance-based mapping (FAB-MAP) for identifying static scenes, which was considered the most popular offline method at that time. Lowry et al. [14] proposed a viewpoint-invariant place recognition method that is more robust to changes in the environment. To cope with complex scene changes, ConvNet was proposed in [15] to extract the landmark regions in images to achieve potential landmark matching, judging the similarity of the whole image by the similarity of the landmark regions, which has better robustness under partial occlusion and major changes in viewpoint. Chen et al. [16] performed a deeper study on the selection of landmarks and the representation of relative regions. Significant regions are identified from deep convolutional layers, and then ConvNet features are obtained directly from these significant regions. Ahmad Khaliq et al. [17] proposed a lightweight approach to visual place recognition, first extracting significant features on a CNN model with a small number of network layers to reduce memory and computational costs and then combining it with VLAD, a local aggregation feature descriptor with better performance in image retrieval tasks, to identify and extract the landmark regions present in the features, achieving good results in scenes with significant changes in viewpoint and appearance.

Attention mechanisms play an important role in robotic tasks. In the work of Chen et al. [18], the authors proposed an attention mechanism that, in combination with existing feedforward network architectures, could detect arbitrarily shaped regions of interest for long-term place recognition. Reference [19] used convolutional neural networks to extract local information for appearance-based loop closure detection and used an attention module when extracting deep local features, allowing the most relevant features to be assigned higher scores. Xu et al. [20] incorporated a second-order attention module into the lightweight network EfficientNetB0 to learn correlations between different spatial location features to improve the global feature performance. Huang et al. [21] combined the pretrained ShuffleNetV2 network with the SE attention model, which has a higher accuracy than the traditional method and the VGG16 network. Mao et al. [22] proposed an attention model with a multiscale feature pyramid from which the attention model was learned to select distinguished features for place recognition and demonstrated that the multiscale feature fusion network obtains better visual features than the unfused network.

Multiscale features have been widely used in deep learning networks. Xin et al. [23] combined global and local information to generate multiscale landmarks and proposed a useful similarity measure to cope with the changing environment. Considering the spatial distribution of landmarks in the similarity measure can improve the robustness of viewpoint changes. In addition, other methods use multiscale feature extraction methods to generate features that are more robust to viewpoint changes [24-25]. Shallow features tend to contain detailed information such as corner points or edges, while deeper features focus on abstract semantic information. Features extracted from different layers of the CNN can capture different semantic structures [26]. Similarly, to make full use of features from different layers, Liu et al. [27] extracted more useful semantic information by fusing multiple layers of features in the network model to improve the accuracy of loop closure detection.

III. Proposed Method

In this section, we describe our loop closure detection method in detail. As shown in Fig. 1, first, we take the residual network ResNet50 as the backbone network and improve the original structure on this basis. A multiscale channel attention mechanism and a spatial attention mechanism are embedded after each layer to adaptively select and add important features. Finally, the shallow and deep features of ResNet50 are fused to realize the complementary strengths of the features in each layer. An improved residual network is used to improve the accuracy of loop closure detection.

A. The process of loop closure detection

The process of loop closure detection is shown in Figure 2. Our network model extracts the features of the current image and the features of the historical image, then compares the similarity and determines whether the current similarity is greater than or equal to the set threshold. If it is greater than or equal to the set threshold, the loop is determined to be closed; otherwise, the next image from the historical image sequence is selected to recalculate the similarity. The end of the loop is marked by finding the location of the closed loop or traversing the historical image sequence.

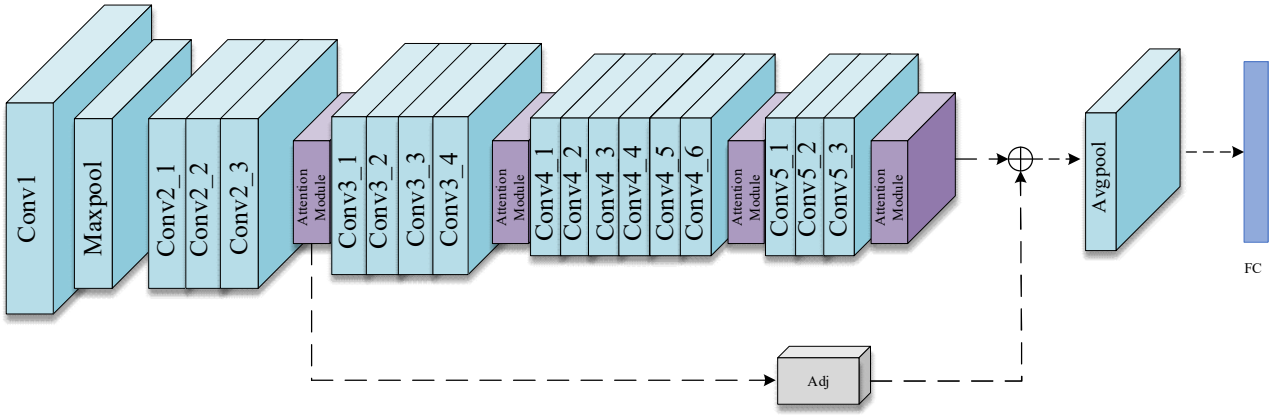

Fig. 1. Architecture of the network structure for feature fusion

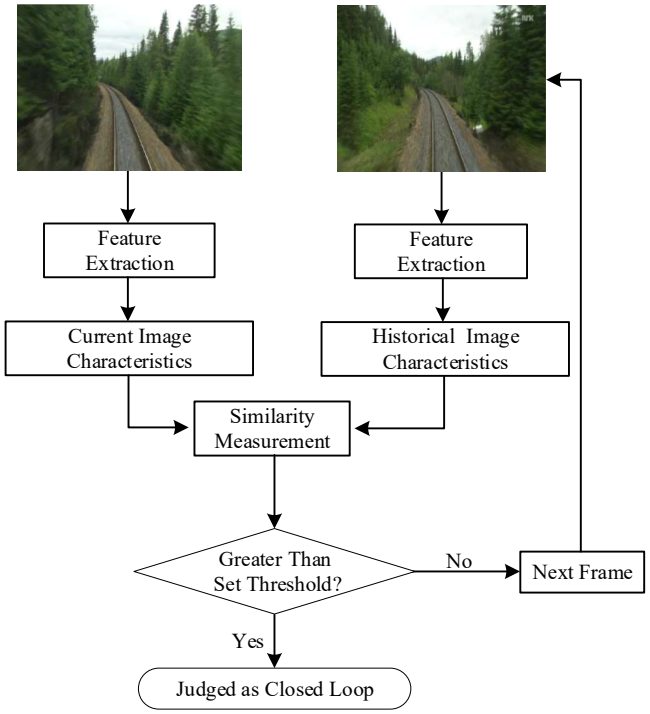

Fig. 2. The process of loop closure detection

B. *The module of multiscale channel attention*

Channel attention is used to select and reweight the most significant and highly differentiated features from the input feature map. Unlike the previous channel attention, to handle richer features, based on the idea of the Inception network [28], we use different convolution kernel sizes of 3×3, 5×5 and 7×7 to generate different feature maps. This allows spatial information of different sizes to be obtained when aggregating information, increasing the diversity of features. Subsequently, the group convolution module (GCM) and global average pooling module (GAPM) are then used to generate multiscale channel attention feature maps, which are then stitched together and sent to the next fully connected layer, where the most relevant features are given higher weights, and in this way, the weights of each feature map are finally obtained. Our multiscale channel attention module is shown in Fig. 3. We describe this in more detail below.

Assume an intermediate feature map of the input image $F \in R^{H\times W\times C}$, where H represents the height of the feature map, W represents the width of the feature map and C represents the number of channels in the feature map. The intermediate feature maps are passed through different convolution kernels of sizes 3×3, 5×5 and 7×7 to obtain multiscale information. The max pooling layer collects another important cue about unique object features to infer finer channel attention, so the max pooling output and the average pooling output are used together to greatly improve the representation of the network [29]. To reduce the number of parameters and the amount of computation involved in the convolution process, we have added the group convolution module. The GCM (Fig. 4) first performs max pooling and average pooling towards the feature maps to change the size of the two feature maps to $\frac{H}{2}\times\frac{W}{2}\times C$, where the max pooling layer and the average pooling layer have a size of 2×2 and a stride of 2. The feature maps from the same channel are then concatenated together. Since the use of group convolution reduces the number of parameters and computation in the

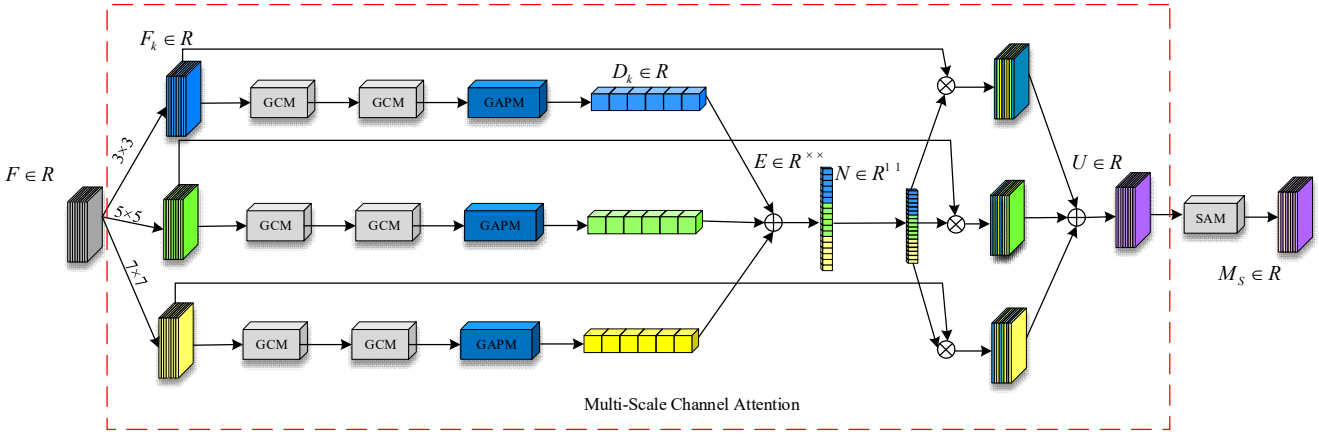

Fig. 3. Architecture of the multiscale channel attention module

convolution process, the feature maps of each channel are converged using $3\times3\times2$ group convolution to obtain a feature map with a size of $\frac{H}{2}\times\frac{W}{2}\times C$. The group convolution module is calculated as shown in (1).

$$f_{gcm}(F)=f_{g\_conv}^{3\times3\times2}([MaxPool(F);AvgPool(F)]) \quad (1)$$

Fig. 3. Architecture of the multiscale channel attention module

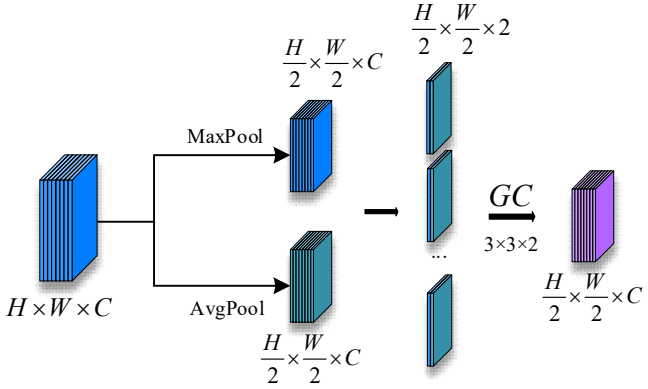

Fig. 4. Architecture of the group convolution module (GCM)

where F is the intermediate feature map of the input image. $f_{g\_conv}^{3\times3\times2}$ is the group convolutional layer with a filter size of $3\times3\times2$, and both the stride and padding are 1. $f_{gcm}(\cdot)$ denotes the group convolution module. After two group convolution modules, the feature map size is reduced to $\frac{H}{4}\times\frac{W}{4}\times C$.

The GAPM consists of a global average pooling layer and a sigmoid function. Global average pooling enforces the correspondence between feature maps and categories and is more suitable for convolutional structures, while the network has fewer parameters and avoids overfitting problems [30]. The computation of the channel attention map is performed through the global average pooling module, as shown in (2).

$$M_C(F)=\sigma(f_{gap}(f_{gcm}(f_{gcm}(F)))) \quad (2)$$

where f_{gap} is the global average pooling layer. σ is the sigmoid function. $M_C(F)$ is the channel attention map. To make full use of the multiscale channel attention map to generate the whole channel information, we integrate the channel attention maps of multiple branches and then send them to the fully connected layer to obtain the weight vector of size $1\times1\times C$. Finally, the input feature map and the weight vector are multiplied channel by channel to obtain the final feature map. The final feature map $U\in R^{H\times W\times C}$ is calculated from the mathematical formulas below.

$$U=\sum_{k=1}^{3}F_k\cdot N \quad (3)$$

C. The module of spatial attention

Spatial attention emphasizes the spatial location of the most salient features rather than treating the whole image as equally important. The spatial attention module is shown in Fig. 5.

Its input is the final feature map after the channel attention module, which is assumed to be $U\in R^{H\times W\times C}$. Two 2D feature maps U_{avg}^{s} and $U_{\max}^{s}$ of size $H\times W\times1$ are obtained after average pooling and max pooling. After stitching them into a feature map of size $H\times W\times2$, two convolutional layers *Conv*1 and *Conv*2 of size $3\times3\times2$ and $1\times1\times1$ are used for convolution, and the final spatial attention map is obtained by the sigmoid function, as shown in (4).

$$M_s(F)=\sigma(f_{conv}^{1\times1\times1}(f_{conv}^{3\times3\times2}([U_{\max}^{s};U_{avg}^{s}]))) \quad (4)$$

where $f_{conv}^{1\times1\times1}$ is the convolutional layer with a filter size of $1\times1\times1$, and both the stride and padding are 1. $f_{conv}^{3\times3\times2}$ is the convolutional layer with a filter size of $3\times3\times2$, and both the stride and padding are 1. σ is the sigmoid function. $M_s(F)$ is the spatial attention map. U_{avg}^{s} and $U_{\max}^{s}$ represent 2D feature maps after average pooling and max pooling.

By introducing the attention module, these attention maps are of the same size as the input feature maps. They focus on key regional features of the scene to be attended to during closed-loop detection, rather than differentiating the scene by the entire image information, aiming to improve the representational power of the network model. At the same time, feature fusion is used to compensate for the problem of information loss as the number of layers in the network increases.

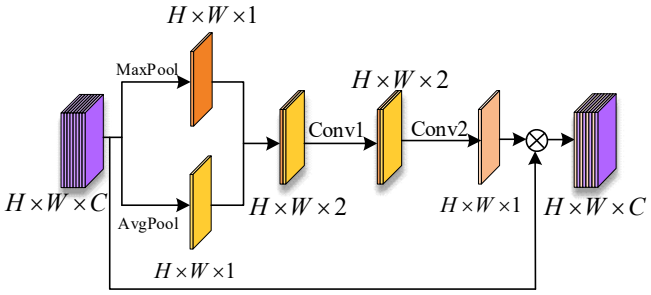

Fig. 5. Architecture of the spatial attention module

D. Feature fusion module

Many existing deep learning methods only use the extracted single layer features as image descriptors [31]. To make full use of the different layers of features, we use the fused shallow and deep features as image descriptors. The deep layer features contain rich semantic information, and the feature maps are more abstract and can cope with viewpoint changes. The shallow features contain detailed features such as edge lines. They can cope with interference from similar objects and increase the accuracy of location. We exploit the complementary nature of the shallow and deep features to generate features with greater expressive power. We adjust the number of channels and size of the feature map prior to feature fusion using the adjustment module, which consists of max pooling and 1×1 convolution. The max pooling layer is used to change the size of the feature map while preserving important information such as the background and texture of the image. The 1×1 convolution is used to facilitate feature fusion by varying the number of feature channels. The adjustment module is shown in Figure 6. With the shallow feature map size of $256\times56\times56$, after a maximum pooling of size 8×8, the stride of 8 and a convolution of 1×1 are used to obtain an output feature map size of $2048\times7\times7$.

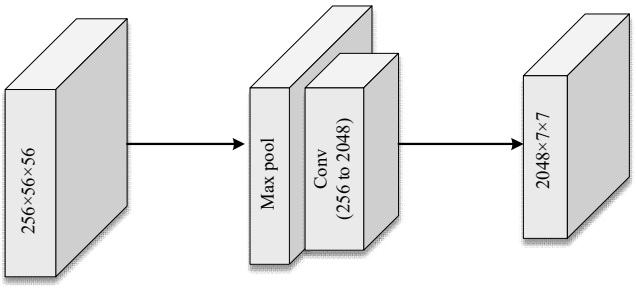

Fig. 6. Architecture of the adjustment module

IV. Experimental

The host configuration used for this experiment was: Intel Xeon(R) CPU E5-2678 v3 at 2.50 GHz, RTX2080Ti GPU, Ubuntu 18.04 LTS, and PyTorch 1.7.1 deep learning framework.

A. Dataset

To evaluate the performance of our proposed method, we conducted separate performance evaluations on widely used datasets, including New College (NC) [32], City Centre (CC) [13] and KITTI 00 and KITTI 02. The New College and City Centre datasets are large urban scene images captured every 1.5 metres forwards by a mobile robot containing left and right monocular cameras. These two datasets were originally collected for the evaluation of FAB-MAP and were later used primarily for the evaluation of the loop closure detection module. In addition, the datasets provide realistic loop closure information. As they contain many dynamic objects (such as cars and pedestrians) and repetitive structures, they are used to evaluate the performance of the method proposed in this paper in the face of dynamic objects and partial occlusion problems. The KITTI dataset is obtained with a stereo camera device mounted on a moving vehicle and contains real image data collected from scenes such as urban, rural and motorway scenes [33]. We chose two sequential scenarios, KITTI 00 and KITTI 02, which were initially used for visual odometry, to evaluate our proposed method. Real loop closure information for the KITTI dataset was provided by the authors of reference [19]. Fig. 7 shows an example of correct loop closure for the CC dataset in the case of viewpoint changes and partial occlusion scenes.

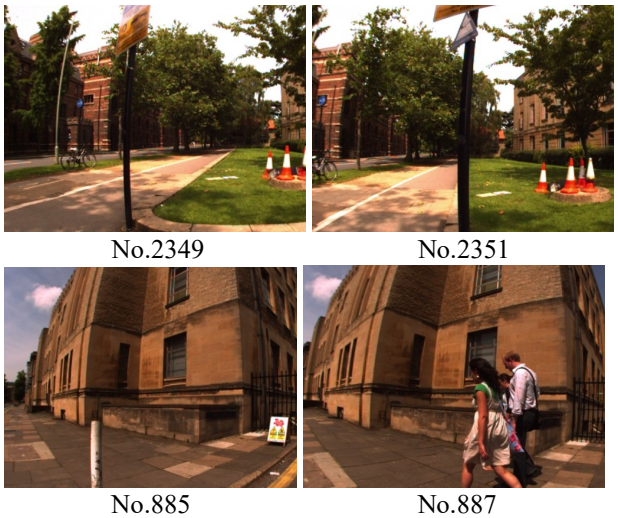

Fig. 7. Examples of the CC dataset

B. Evaluation criteria

We use Precision and Recall to evaluate the performance of our proposed method. Different accuracy and recall rates are obtained by varying the threshold size of similarity and plotting the P-R curves, which are calculated as (5) and (6):

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{6}$$

where TP denotes the number of true positives, i.e., correct loop closure points in the loop closure detected by the algorithm, FP denotes the number of false positives, i.e., incorrect loop closure points in the loop closure detected by the algorithm, and FN denotes the number of false negatives, i.e., loop closure points in the correct loop closure case that are not detected. The accuracy rate is the ratio between the number of correct loop closures and the number of loop closures detected by the loop closure algorithm. The recall is the probability of all correct closed loops being detected correctly. For SLAM systems, a high accuracy rate is important. False positives can eventually lead to biased location results and degrade the performance of the system. Therefore, we set the accuracy to 100% and evaluate the recall with 100% accuracy. For the similarity calculation, we use the cosine similarity calculation.

C. Experimental analysis

As shown in Table 1, we compared our method with various state-of-the-art and classical loop closure detection algorithms, including FILD [6], Wang [34], KAZMI [35], FILD++ [19], HTMap [36], FAB-MAP [13], Zhang [37], and DBoW2 [38]. "-" indicates that the algorithm was not performed on the dataset experiments. Our method has a higher recall than other methods on most datasets. Although our method performs worse than those of FILD [6] and Wang [34] on the NC dataset, it is also better than most other methods. It is worth mentioning that our method achieves satisfactory results on the CC dataset because our attention mechanism can assign less weight to these irrelevant features and more weight to key landmarks in the scene when faced with scenes with a large number of dynamic objects such as cars and pedestrians. The NC dataset images have a high visual ambiguity, and many scenes look very close to each other. This situation poses a significant challenge to our attention mechanism. The K00 dataset was collected in a dynamic environment, and our method has a higher recall than other algorithms even when the environment changes. All the algorithms perform poorly on the K02 dataset. This is due to the low image texture of the K02 dataset and the absence of major landmarks, with most scenes having only trees and roads with relatively high similarity, making it difficult to extract similarity features.

We conducted ablation experiments on four separate datasets, as shown in Table 2, comparing the maximum recall at 100% accuracy for the methods with and without the attention module and the feature fusion module. In addition, the accuracy-recall curves for the ablation experiments are shown in Fig. 8-Fig. 11.

TABLE I
RECALL OF DIFFERENT ALGORITHMS AT 100% ACCURACY

Approach	NC	CC	K00	K02
FAB-MAP[12]	51.91	38.50	-	-
FILD[6]	**89.94**	-	91.23	65.11
Zhang[36]	48.79	63.19	95.37	-
FILD++[18]	82.37	90.01	94.92	73.52
KAZMI[34]	51.09	75.58	90.39	79.49
HTMap[35]	73.60	79.68	90.24	-
DBoW2[37]	55.92	30.16	78.42	67.59
Wang[33]	87.41	86.63	96.68	-
Proposed	82.67	**90.20**	**97.16**	**79.84**

TABLE II
COMPARATIVE RESULTS OF ABLATION EXPERIMENTS

Approach	NC	CC	K00	K02
Attention+Fusion	82.67	90.20	97.16	79.84
Attention(Without Fusion)	81.46	87.13	97.10	77.03
Without Attention+Fusion	68.02	64.30	77.08	57.78
Without Attention(Without Fusion)	65.55	63.26	69.41	55.90

The experimental results show that the methods using both the attention mechanism and feature fusion outperform the other methods. The difference between the performance of the method using the attention module and that of the method without the attention module is significant. The attention mechanism is able to learn the correlation between the original features, extract more discriminative features and suppress the interference of irrelevant object feature regions, thus having a greater impact on the final results of the experiments. At the same time, the feature fusion approach further improves the accuracy of loop closure detection.

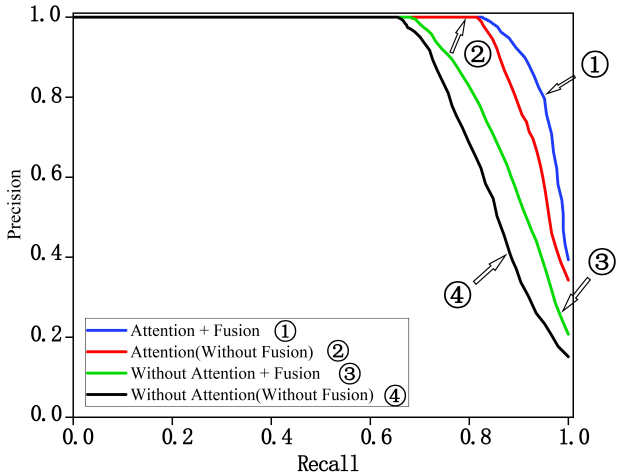

Fig. 8. P-R curve of the NC dataset

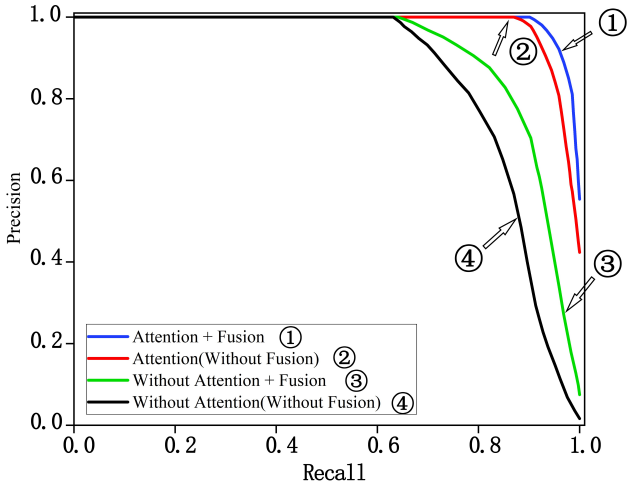

Fig. 9. P-R curve of the CC dataset

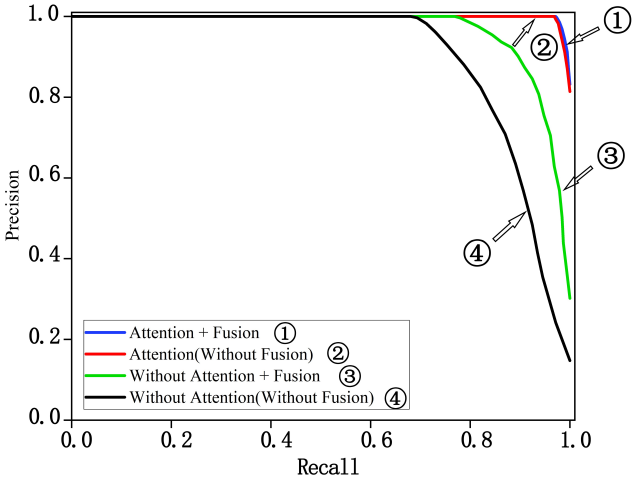

Fig. 10. P-R curve of the K00 dataset

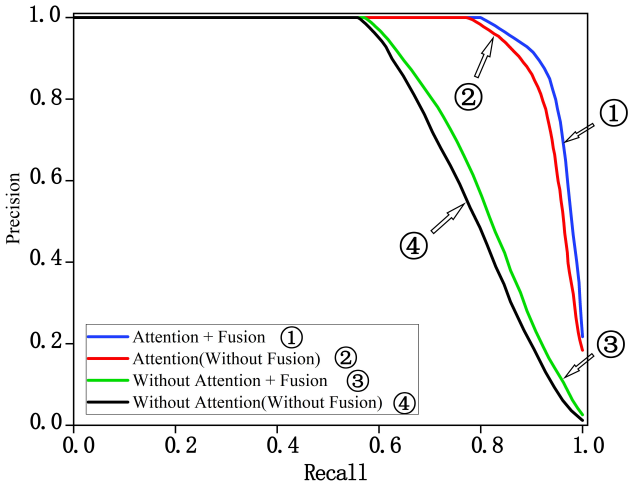

Fig. 11. P-R curve of the K02 dataset

V. CONCLUSION

In this paper, we propose a loop closure detection method that incorporates an attention mechanism. We embed channel attention and spatial attention modules on top of the ResNet50 network. The attention mechanism can better select and reweight the most salient features and focus more on discriminative image regions. Our approach automatically learns the relevance of features in the original feature map and assigns higher weights to regions of the scene that are highly discriminative, thus better meeting the challenge of

scene changes. The feature fusion approach then compensates for the loss of structural information in the deep network, further improving the accuracy of closed-loop detection. Comparisons are made with other methods on the New College, City Centre, and KITTI public datasets. The results show that our method has a higher recall at 100% accuracy. Ablation experiments validate that the introduction of an attention mechanism can better improve the performance of loop closure detection. Compared with other similar methods, our method has higher accuracy and robustness in dealing with scene change problems.

REFERENCES

[1] Garcia-Fidalgo E, Ortiz A. Vision-based topological mapping and localization methods: A survey[J]. Robotics and Autonomous Systems, 2015, 64: 1-20.

[2] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.

[3] Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF[C]//2011 International conference on computer vision. Ieee, 2011: 2564-2571.

[4] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope[J]. International journal of computer vision, 2001, 42(3): 145-175.

[5] Csurka G, Dance C, Fan L, et al. Visual categorization with bags of keypoints[C]//Workshop on statistical learning in computer vision, ECCV. 2004, 1(1-22): 1-2.

[6] An S, Che G, Zhou F, et al. Fast and incremental loop closure detection using proximity graphs[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 378-385.

[7] Gordo A, Almazán J, Revaud J, et al. Deep image retrieval: Learning global representations for image search[C]//European conference on computer vision. Springer, Cham, 2016: 241-257.

[8] Hou Y, Zhang H, Zhou S. Convolutional neural network-based image representation for visual loop closure detection[C]//2015 IEEE international conference on information and automation. IEEE, 2015: 2238-2245.

[9] GUO J Z, LIU F L, YANG X Z, et al. The closed loop detection method of vision slam based on deep learning[J]. Journal of Optoelectronics Laser, 2021, 32(6): 9.

[10] Zuo L, Zhang C H, Liu F L, et al. Performance evaluation of deep neural networks in detecting loop closure of visual SLAM[C]//2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). IEEE, 2019, 2: 171-175.

[11] Chen B, Yuan D, Liu C, et al. Loop closure detection based on multi-scale deep feature fusion[J]. Applied Sciences, 2019, 9(6): 1120.

[12] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.

[13] Cummins M, Newman P. FAB-MAP: Probabilistic localization and mapping in the space of appearance[J]. The International Journal of Robotics Research, 2008, 27(6): 647-665.

[14] Lowry S, Andreasson H. Lightweight, viewpoint-invariant visual place recognition in changing environments[J]. IEEE Robotics and Automation Letters, 2018, 3(2): 957-964.

[15] Sünderhauf N, Shirazi S, Jacobson A, et al. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free[J]. Robotics: Science and Systems XI, 2015: 1-10.

[16] Chen Z, Maffra F, Sa I, et al. Only look once, mining distinctive landmarks from convnet for visual place recognition[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017: 9-16.

[17] Khaliq A, Ehsan S, Chen Z, et al. A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes[J]. IEEE transactions on robotics, 2019, 36(2): 561-569.

[18] Chen Z, Liu L, Sa I, et al. Learning context flexible attention model for long-term visual place recognition[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 4015-4022.

[19] An S, Zhu H, Wei D, et al. Fast and incremental loop closure detection with deep features and proximity graphs[J]. Journal Of Field Robotics, 2022, 39(4): 473-493.

[20] Xu Y, Huang J, Wang J, et al. Esa-vlad: a lightweight network based on second-order attention and netvlad for loop closure detection[J]. IEEE Robotics and Automation Letters, 2021, 6(4): 6545-6552.

[21] Huang L, Zhu M, Zhang M. Visual Loop Closure Detection Based on Lightweight Convolutional Neural Network and Product Quantization[C]//2021 IEEE 12th International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2021: 122-126.

[22] Mao J, Hu X, He X, et al. Learning to fuse multiscale features for visual place recognition[J]. IEEE Access, 2018, 7: 5723-5735.

[23] Xin Z, Cui X, Zhang J, et al. Real-time visual place recognition based on analyzing distribution of multi-scale cnn landmarks[J]. Journal of Intelligent & Robotic Systems, 2019, 94(3): 777-792.

[24] Herranz L, Jiang S, Li X. Scene recognition with cnns: objects, scales and dataset bias[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 571-579.

[25] Chen Z, Jacobson A, Sünderhauf N, et al. Deep learning features at scale for visual place recognition[C]//2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017: 3223-3230.

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

[27] Liu J, Xiao M, Lin X, et al. Adaptive Real-Time Loop Closure Detection Based on Image Feature Concatenation[C]//2021 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2021: 1-5.

[28] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826..

[29] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[30] Lin M, Chen Q, Yan S. Network in network[J]. arXiv preprint arXiv:1312.4400, 2013.

[31] McManus C, Churchill W, Maddern W, et al. Shady dealings: Robust, long-term visual localisation using illumination invariance[C]//2014 IEEE international conference on robotics and automation (ICRA). IEEE, 2014: 901-906.

[32] Smith M, Baldwin I, Churchill W, et al. The new college vision and laser data set[J]. The International Journal of Robotics Research, 2009, 28(5): 595-599.

[33] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 3354-3361.

[34] Wang S, Lv X, Liu X, et al. Compressed holistic convnet representations for detecting loop closures in dynamic environments[J]. IEEE Access, 2020, 8: 60552-60574.

[35] Kazmi S M A M, Mertsching B. Detecting the expectancy of a place using nearby context for appearance-based mapping[J]. IEEE Transactions on Robotics, 2019, 35(6): 1352-1366.

[36] Garcia-Fidalgo E, Ortiz A. Hierarchical place recognition for topological mapping[J]. IEEE Transactions on Robotics, 2017, 33(5): 1061-1074.

[37] Zhang X, Wang L, Zhao Y, et al. Graph-based place recognition in image sequences with CNN features[J]. Journal of Intelligent & Robotic Systems, 2019, 95(2): 389-403.

[38] Gálvez-López D, Tardos J D. Bags of binary words for fast place recognition in image sequences[J]. IEEE Transactions on Robotics, 2012, 28(5): 1188-1197.