applied sciences

MDPI

Article

Loop Closure Detection in RGB-D SLAM by Utilizing Siamese ConvNet Features

Gang Xu [1,2], Xiang Li [2], Xingyu Zhang [2], Guangxin Xing [2] and Feng Pan [1,*]

[1] Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214000, China; xugang@ahpu.edu.cn
[2] Anhui Key Laboratory of Detection Technology and Energy Saving Devices, Anhui Polytechnic University, Wuhu 241000, China; 2200220139@stu.ahpu.edu.cn (X.L.); zhangxingyu@ahpu.edu.cn (X.Z.); x959484550@163.com (G.X.)
* Correspondence: pan_feng_63@163.com

Abstract: Loop closure detection is a key challenge in visual simultaneous localization and mapping (SLAM) systems, which has attracted significant research interest in recent years. It entails correctly determining whether a scene has previously been visited by a mobile robot and completely establishing the consistent maps of motion. There are many loop closure detection methods that have been proposed, but most of these algorithms are handcrafted features-based and perform weak robustness to illumination variations. In this paper, we investigate a Siamese Convolutional Neural Network (SCNN) to solve the task of loop closure detection in RGB-D SLAM. Firstly, we use a pre-trained SCNN model to extract features as image descriptors; then, the L2 norm distance is adopted as a similarity metric between descriptors. In terms of the learned features for matching, there are two key issues for discussion: (1) how to define an appropriate loss as supervision (utilizing the cross-entropy loss, the contrastive loss, or the combination of two); and (2) how to combine the appearance information in RGB images and position information in depth images (utilizing early fusion, mid-level fusion or late fusion). We compare our proposed method of different baseline by experiments carried out on two public datasets (New College and NYU), and our performance outperforms the state-of-the-art.

Keywords: Siamese convolutional neural network; SLAM; loop closure detection; RGB-D

Citation: Xu, G.; Li, X.; Zhang, X.; Xing, G.; Pan, F. Loop Closure Detection in RGB-D SLAM by Utilizing Siamese ConvNet Features. *Appl. Sci.* **2022**, *12*, 62. https://doi.org/10.3390/app12010062

Academic Editor: Shengzong Zhou

Received: 1 November 2021
Accepted: 1 December 2021
Published: 22 December 2021

1. Introduction

Visual simultaneous localization and mapping (SLAM) is one of the fundamental problems in robotics with numerous important applications, such as robot motion [1,2], trajectory planning [3], driving recorder [4], etc. The algorithms for solving SLAM can be roughly classified into two categories: filter-based SLAM and graph-based SLAM. Since traditional filter-based methods, such as Kalman filter [5,6] and Particle filter [7], would accumulate errors when mapping, it cannot be applicable in the process of large-scale map creation. Therefore, recent studies have concentrated on the graph-based method.

Loop closure detection is a key challenge in the task of graph-based visual SLAM. It aims at determining whether a robot has visited a location (such as an office, corridor, or library as we can see in Figure 1) previously arrived at, and is vital for the generation of a consistent map by correcting errors that accumulate overtimes. The problem of visual loop closure detection shares similar ideas with image retrieval; however, significant distinctions exist between these two visual tasks. The purpose of image retrieval is to find out the highest similarity image with the reference image. In comparison, loop closure detection searches identical images to the current scene, without considering object occlusion and light variation. Although the task of loop closure detection has been approached from various angles, all solutions are based on matching and share a common framework:

extracting features as image descriptors, followed by measuring the similarity between two descriptors.

Figure 1. Example of different scenes encountered by the robot.

Recent developments in computer vision have motivated many image descriptors for feature representations, such as Scale Invariant Feature Transform (SIFT) [8], Speeded Up Robust Features (SURF) [9], Oriented FAST and Rotated BRIEF (ORB) [10], and GIST [11], etc. These approaches are based on keypoints matching and have been widely utilized in loop closure detection. However, all the above descriptors are hand-crafted and have weak robustness to the illumination variations. For example, we applied SIFT to extract keypoints and generate descriptors for matching. Figure 2 shows matching maps of a scene with angle variations and illumination variations. As expected, the matching accuracy in Figure 2b is much lower than Figure 2a. However, the key point-based descriptor is usually invariant to affine transformation and illumination.

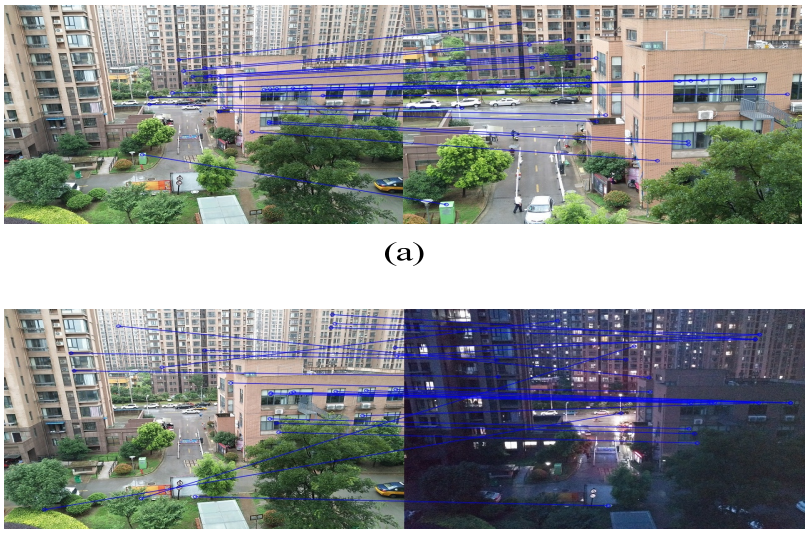


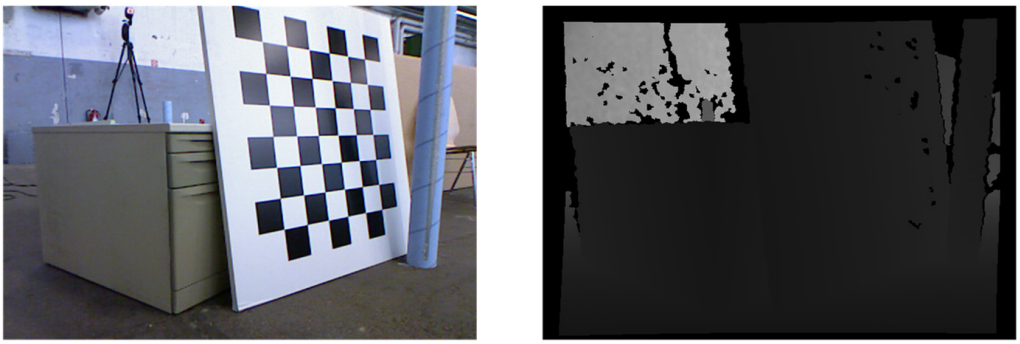
Figure 2. Examples of SIFT keypoints matching.

To overcome the limitations of key points-based matching, in recent years, Bag-of-Visual-Words (BoVW) have been widely used in loop closure detection for visual SLAM [12–16]. These methods store the scene as a visual dictionary and use the K-Means algorithm to generate clusters of the extracted features, where each cluster represents a "word". The image is described by the histogram of each visual vocabulary instead of key point-based descriptors. Fast Appearance-Based Map (FAB-MAP) [12] is one of the typical applications of the BoVW and has become a standard baseline in loop closure detection algorithms. Fisher Vector (FV) [17,18] and vector of locally aggregated descriptors (VLAD) [19,20] are derived from the basic idea of BoVW and become mainstream algorithms for visual SLAM [21,22]. Fisher vectors utilize a Gaussian Mixture Model to generate codebook and local descriptors are vector-quantized to visual words. Compared with BoVW, FV contains richer information. VLAD can be viewed as a simplification of Fisher vector and reduce the computational redundancy. In addition to the above mentioned, GIST as a global image descriptor has also been extensively applied to loop closure detection [23–25]. The GIST descriptor is computed by the Gabor filter on a whole image and the generated descriptor is no more than 1000 dimensions

Inspired by the outstanding performance of deep learning architecture in various computer vision tasks, deep learning-based descriptors have recently come into usage in the problem of loop closure detection [26–29] and the effectiveness and superiority of deep features have been validated on multiple public datasets.

Gao et al. [27,28] firstly trained an unsupervised stacked auto-encode mode to extract feature representations in RGB images and then, the loop is judged by the similarity matrix. To speed up feature processing, there are many methods, such as [30,31], but PCANet [30], which is a simple deep learning architecture, has also been used in loop closure detection. Xia et al. [29] used PCANet to extract feature descriptors and then, the similarity of two descriptors is measured by the cosine distance. As expected, their algorithm costs the shortest computational time of 0.012 s per image. The convolutional neural network structure has gained more and more attention in multiple computer vision tasks, including image and video classification [32,33], object and face detection [34,35], action recognition [36,37], bioacoustics [38], etc. The remarkable achievements of ConvNet on computer vision are largely contributed to its excellent data generalization ability. Thus, it is advisable to apply the power of ConvNet structure to the specific task of loop closure detection instead of using traditional image descriptors. Hou et al. [26] creatively proposed the ConvNet-based method, which utilized AlexNet to automatically learn feature descriptors and then the similarity between descriptors was measured by the L2-norm distance for visual loop closure detection and achieved the state-of-the-art performance. In their implementations, they utilized AlexNet [32], which made a breakthrough success in the Large Scale Visual Recognition Challenge 2012 (ISVRC2012) to extract features and then, the L2 norm distance was used for matching. However, the standard ConvNet architecture was essentially designed for image classification tasks, and descriptors learned by softmax entropy loss may be suboptimal for matching tasks. Thus, there is space for improvements by improving the ConvNet architecture.

Based on the above analysis, in this paper, we develop a new Siamese Convolutional Neural Network (SCNN) for the problem of loop closure detection. Firstly, we use the SCNN model to extract features as image descriptors; then, the L2 norm distance is adopted as a similarity metric between descriptors. The proposed method is validated on two open datasets. Among them, New College [12] is widely used in visual SLAM research; the NYU dataset [39] consists of RGB-D image pairs with both color and depth images collected by a high-resolution Kinect camera, as shown in Figure 3.



(a) RGB image (b) Depth image

Figure 3. An example of an RGB-D image collected by a Kinect camera.

The main contributions of our study can be summarized into three aspects: (1) We utilized a novel Siamese Convolutional Neural Network (SCNN) combining RGB stream and depth stream in a late fusion way to extract features as image descriptors and used the L2 norm distance as a similarity measure between descriptors for the specific task of loop closure detection, which can improve the detection accuracy. (2) We investigated three strategies (early fusion, mid-level fusion, and late fusion) to combine RGB stream and depth stream for RGB-D loop closure detection. This was combined with the experimental selection of late fusion to achieve high accuracy and recall, which is most suitable in the RGB-D loop closure detection task. (3) Our proposed approaches achieved state-of-the-art performance on both the RGB dataset and RGB-D dataset, which shows that the method applies to monocular in addition to RGB-D cameras.

The rest of this paper is organized as follows: the proposed method and the overall framework are detailed in Section 2; the experiments and comparisons of results are summarized in Section 3; and finally, we conclude this paper in Section 5.

2. Materials and Methods

In this section, we first introduce the proposed Siamese ConvNet architecture specialized for loop closure detection and then, explore three fusion strategies (early fusion, mid-level fusion, and late fusion) of combining RGB stream and depth stream for RGB-D loop closure detection. Our sole aim in this paper is to extract representative features to apply the loop closure detection tasks, and to learn better similarity metrics.

2.1. Siamese ConvNet Architecture for Loop Closure Detection

For the Siamese ConvNet architecture, the proposed Siamese Convolutional Neural Networks (SCNN) takes a similar structure to AlexNet [32], which is composed of five convolutional layers and three fully connected layers. However, in contrast to the classical Convolutional Neural Networks (CNNs), it takes a pair of images as input and then feeds forward by two identical branch convolutional structures, which share the same parameters. An illustration of the enhanced deep architecture can be seen in Figure 4 and the parameters of each layer are listed in Table 1.

The architecture takes an image pair with pixels of $227 \times 227 \times 3$ as input including eight layers. The first five layers are convolutional layers, and the last three layers are the inner product (full connected) layers. Relu, pooling, and normalization follow after the output of each convolutional layer. The first convolutional layer filters the $227 \times 227 \times 3$ input with kernels of size $11 \times 11 \times 3$ with a stride of 4 pixels and then, the output of the first convolutional layer is rescaled to $27 \times 27 \times 96$ by the 3×3 max pooling. The second convolutional layer takes the feature maps of the first layer as input, filters it with 256 kernels of size $5 \times 5 \times 96$, and acquires feature maps with a size of $13 \times 13 \times 256$. The third convolutional layer takes the feature maps of the second layer as input, filters it with 384 kernels of size $3 \times 3 \times 256$, and acquires feature maps with a size of $13 \times 13 \times 384$.

Note that the parameters of other layers can be looked up in Table 1, which are similar to the Alexnet.

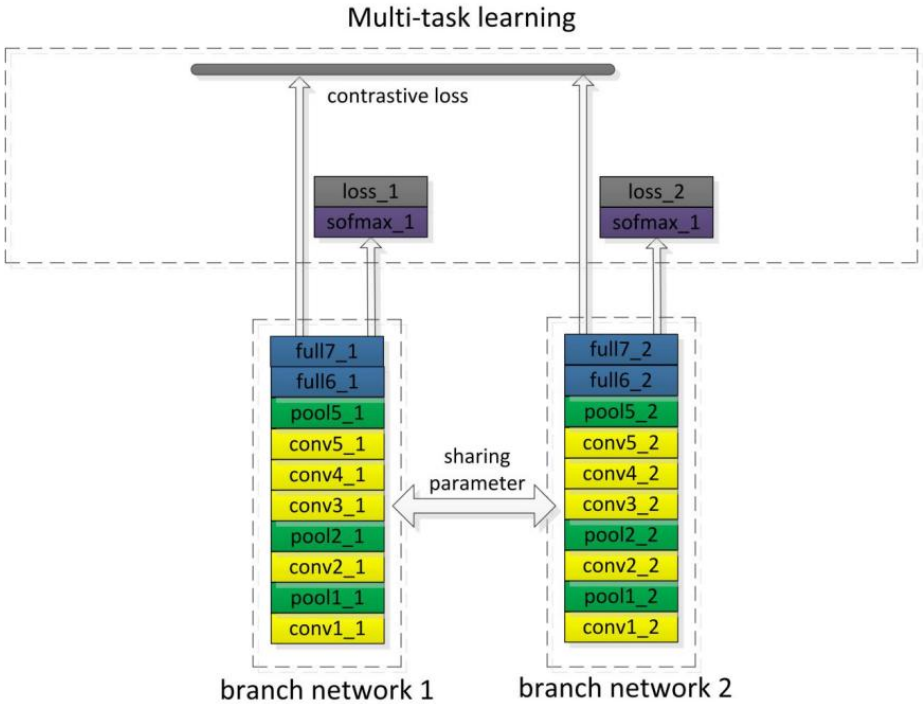


Figure 4. Siamese ConvNet Structure for visual loop closure detection.

Table 1. The relevant layers and their parameters of SCNN.

Layer	Layer Type	Parameter
input1_1, input1_2	input	Image size:227 × 227 × 3
conv1_1, conv1_2	convolution	Filter size:11 × 11, Filter Num:96, Stride: 4
pool1_1, pool1_2	pooling	Pooling method: Max, Kernel size:3 × 3, Stride: 2
conv2_1, conv2_2	convolution	Filter size:5 × 5, Filter Num:256, Stride: 1
pool2_1, pool2_2	pooling	Pooling method: Max, Kernel size:3 × 3, Stride: 2
conv3_1, conv3_2	convolution	Filter size:3 × 3, Filter Num:384, Stride: 1
conv4_1, conv4_2	convolution	Filter size:3 × 3, Filter Num:384, Stride: 1
conv5_1, conv5_2	convolution	Filter size:3 × 3, Filter Num:256, Stride: 1
pool5_1, pool5_2	pooling	Pooling method: Max, Kernel size:3 × 3, Stride: 2
full6_1, full_6_2	fully connected	Neurons output:4096
full7_1, full_7_2	fully connected	Neurons output:2048
softmax_1, softmax_2	fully connected	Neurons output:205
feat1_1, feat1_2	fully connected	Neurons output:1000

- Pre-train

The Siamese ConvNet on the Place dataset [40] contains more than 2.5 million images of 205 scene classes. Formally, we feed image pairs with a size of 227 × 227 × 3 as input and define training samples of $x_1^k = \left\{x_1^k, x_2^k, l_1^k, l_2^k\right\}$, where x_1^k and x_2^k represent the k-th input pairs, and l_1^k and l_2^k are its corresponding labels. Then, the architecture can automatically extract 2048 dimensions feature vectors $\mathbf{f}^k = \left\{\mathbf{f}_1^k, \mathbf{f}_2^k\right\}$:

$$\mathbf{f}^k = Conv(x^k \Big| \theta_{net}) \tag{1}$$

where $Conv(\cdot|\theta_{net})$ is the feature extraction function defined by the Siamese ConvNet, and θ_{net} is the shared parameter to be learned.

Then, the problem can be reduced to select an appropriate loss function as supervision. Here, we attempt three strategies for feature learning: softmax supervision, contrastive supervision, and multi-task supervision.

- Softmax supervision:

Softmax supervision. The softmax_1 and softmax_2 layers output the probability of 205 scene classes. We compute the cross-entropy of once iteration and select the softmax loss as supervision, which can be represented as follows:

$$L_{cls}(\mathbf{f}^k, l^k, \theta_{cls}) = -[\sum_{i=1}^{n} 1\left\{i = l^k\right\} \log P(i = l^k, \theta_{cls})] \tag{2}$$

where $\mathbf{f}^k = \left\{\mathbf{f}_1^k, \mathbf{f}_2^k\right\}$ is the k-th extracted feature vectors, $l^k = \{l_1^k, l_2^k\}$ denotes the class label, θ_{cls} is the parameter of softmax classifier, and $1\{\cdot\}$is the indicator function:

1{a true statement} = 1 and
1{a flase statement} = 0.

$P\left\{\cdot \Big| \mathbf{f}^k, \theta_{cls}\right\}$ represents the predicted probability of each class.

Softmax supervision has been widely used in various image classification tasks, but the learned features are not well suited for our application of matching tasks, which will be validated by the experiment in Section 4.

- Contrastive supervision:

The full7_1 layer and full7_2 layer outputs 2048-d feature vectors of $\mathbf{f}^k = \left\{\mathbf{f}_1^k, \mathbf{f}_2^k\right\}$. We apply the L2-normal distance to quantify the similarity between pairs of feature vectors and define a contrastive loss as supervision, which can be represented as follows:

$$L_{cts}(\mathbf{f}^k, l^k, \theta_{cst}) = \begin{cases} \frac{1}{2}\left|\left|\mathbf{f}_1^k - \mathbf{f}_2^k\right|\right|_2^2 & if\ l_1^k = l_2^k \\ \frac{1}{2}\max\left\{0, m - \left|\left|\mathbf{f}_1^k - \mathbf{f}_2^k\right|\right|_2\right\}^2 & if\ l_1^k \neq l_2^k \end{cases} \tag{3}$$

Optimizing the contrastive loss equals minimizing the L2-norm distance between feature vectors, when $\mathbf{f}_1^k$ and $\mathbf{f}_2^k$ are from the same class ($l_1^k = l_2^k$), and maximize the L2-norm distance between feature vectors, when $\mathbf{f}_1^k$ and $\mathbf{f}_2^k$ are from the different class($l_1^k \neq l_2^k$).

θ_{cst} = {m} is the parameter to be learned in the contrastive loss function.

Although contrastive supervision can be applicable in matching tasks, we consider it to be a suboptimal solution due to the lack of regularization items.

- Multi-task supervision:

We can jointly learn the softmax entropy loss and contrastive loss by optimizing the multi-task loss function. We can define multi-task supervision as follows:

$$L(x^k, l^k, \theta) = L_{cls} + \lambda \cdot L_{cst} + \left|\left|\theta_{net}\right|\right|_2^2 + \left|\left|\theta_{cst}\right|\right|_2^2 + \left|\left|\theta_{cls}\right|\right|_2^2 \tag{4}$$

where L_{cls} is the softmax entropy loss, L_{cst} is the contrastive loss, and λ is the hyperparameter balance of the softmax entropy loss and the contrastive loss. θ_{net}, θ_{cls}, and θ_{cst} are the parameters to be learned. We use the stochastic gradient descent algorithm to update the related parameter and the proposed multi-task learning strategy is summarized in Algorithm 1.

Algorithm 1 learning strategy for loop closure detection

Input: training set $X = \left\{x^1, x^2, x^3, \ldots, x^k\right\}$, $x^k = \left\{x_1^k, x_2^k, l_1^k, l_2^k\right\}$, initialized parameters θ_{net}, θ_{cls} and θ_{cst}. learning rate $\eta(t)$, $t \leftarrow 0$, times of iteration N, batch size k.
While $t \neq N$ **do**
$t \leftarrow t+1$ sample M training samples from X
$\mathbf{f}^i = Conv(x^i \left| \theta_{net} \right.)$
$\nabla\theta_{cls} = \sum_{i=1}^{M} \frac{\partial L_{cls}(\mathbf{f}^i, l^i, \theta_{cls})}{\partial \theta_{cls}}$
$\nabla\theta_{cst} = \lambda \cdot \sum_{i=1}^{M} \frac{\partial L_{cst}(\mathbf{f}^i, l^i, \theta_{cst})}{\partial \theta_{cst}}$
$\nabla\mathbf{f}^i = \frac{\partial L_{cls}(\mathbf{f}^i, l^i, \theta_{cls})}{\partial \mathbf{f}^i} + \lambda \cdot \frac{\partial L_{cst}(\mathbf{f}^i, l^i, \theta_{cst})}{\partial \mathbf{f}^i}$
$\nabla\theta_{conv} = \sum_{i=1}^{M} \frac{\partial Conv(x^i, \theta_{conv})}{\partial \theta_{conv}}$
Update $\theta_{conv} = \theta_{conv} - \eta(t) \cdot \nabla\theta_{conv}$, $\theta_{cst} = \theta_{cst} - \eta(t) \cdot \nabla\theta_{cst}$, $\theta_{cls} = \theta_{cls} - \eta(t) \cdot \nabla\theta_{cls}$
End while
Output θ_{conv}, θ_{cls} and θ_{cst}

- Detecting loops;

After training the Siamese networks, we feed the test image to the trained architectures and then extract the features of the full7 layer as image descriptors. Given two arbitrary scenes (x^i, x^j), we compute the L2-norm distance of the generated descriptors:

$$D(i,j) = \left|\left| \mathbf{f}^i - \mathbf{f}^j \right|\right| \tag{5}$$

By this definition, if an item $D(i, j)$ is smaller than the threshold T, we can say the two scenes are matching and vice versa.

2.2. RGB-D Fusion for Loop Closure Detection

For RGB-D SLAM, we employ a Microsoft Kinect camera to capture location and depth information [41] as complementary in our application. The Siamese ConvNet architecture can feed-forward and generate image descriptions for loop closure detection. However, how to combine the RGB stream network and the depth stream network needs further discussion. There are three available strategies for fusion: early fusion, mid-level fusion, and late fusion.

- Early fusion:

In early fusion strategy, appearance information and depth information are combined at the beginning of the network. RGB images and depth images are concatenated together, giving rise to a four-channel input (three channels from RGB stream and one channel from depth stream). Then, the Siamese network generates the description for the final decision. Figure 5 shows the flowchart of the early fusion pipeline.

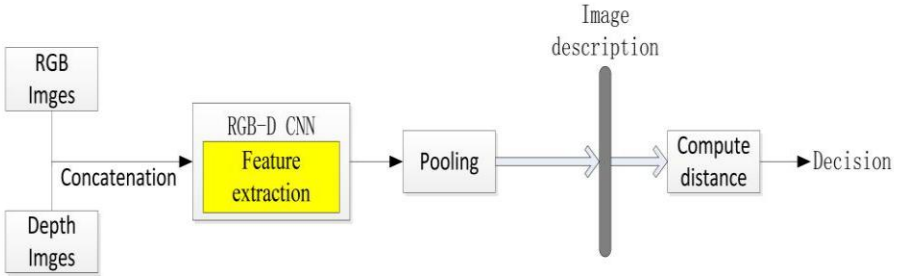


Figure 5. Combining RGB-D information by early fusion pipeline.

- Mid-level fusion:

In contrast to the early fusion, RGB images and depth images are respectively fed to the RGB stream and depth stream, and two networks are fused at intermediate layers with the same size of feature maps. Finally, the combined architecture outputs the description for detecting loops. Figure 6 shows the flowchart of the mid-level fusion pipeline.

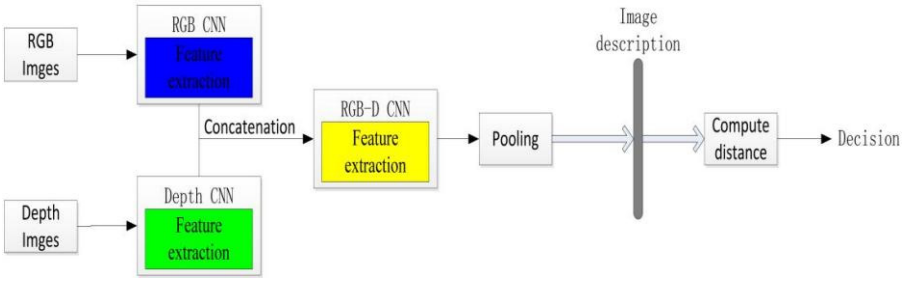


Figure 6. Combining RGB-D information by mid-level fusion pipeline.

- Late fusion:

In the late fusion strategy, the two streams are discriminatively trained for feature extraction. The generated RGB descriptions and depth descriptions are combined by averaging for the final decision, as shown in Figure 7.

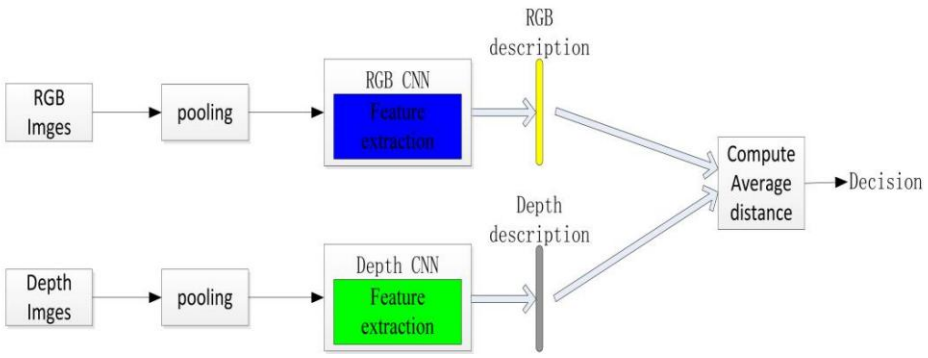


Figure 7. Combining RGB-D information by late fusion pipeline.

We experiment to compare the three-fusion strategy in the next Section 3. By quantifying the experiment results, we conclude that late fusion is best suited in our application of RGB-D loop closure detection.

3. Results

In our experiment, we use the open-source toolkit Caffe to implement the proposed Siamese CNN architecture for feature extraction and matching. We perform on a workstation with Intel Core I7, NVIDIA GTX TITAN X GPU, and the operating system of Ubuntu 16.04. We utilize the Stochastic Gradient Decent (SGD) to update the parameters of each layer with the mini-batch size of 100, the momentum of 0.9, and the initial learning rate of 0.01. The training procedure is maintained until the validation accuracy remains unchanged for 10 consecutive epochs.

3.1. Experiment Setup

Our proposed method is evaluated on two public loop closure detection datasets (New College and NYU). New College is an RGB SLAM dataset, which consists of 1073 image pairs with a resolution of 640 $\times$ 480, recorded by a motion robot with two symmetrical cameras on the right and left side while moving through the outdoor environment, and all ground truth loops are annotated for evaluating the performance of RGB loop closure detection. NYU is an

RGB-D SLAM dataset that contains 1449 RGB-D image pairs with a resolution of 512 × 424 including both color and depth images, captured by Microsoft Kinect.

Here, to quantify the comparative experiment results, we use two evaluation criteria: precision rate and recall rate, which can be denoted as the following formulations:

$$precision = \frac{N_{TruePositive}}{N_{TruePositive} + N_{\mathbf{f}alsePositive}} \tag{6}$$

$$recall = \frac{N_{TruePositive}}{N_{TruePositive} + N_{\mathbf{f}alseNegative}} \tag{7}$$

Precision rate refers to the probability that all loop closures extracted by the algorithm are true loops. Recall is the probability of being correctly detected in all true loop closures. The precision–recall curve is utilized to intuitively reflect the performance of different implementations in the next subsection.

3.2. Experiment Results on New College Dataset

We report the experimental results of our proposed Siamese convolutional neural network-based method and its comparisons with both traditional methods and other deep learning-based methods. A precision–recall curve of comparisons on the New College dataset can be seen in Figure 8.

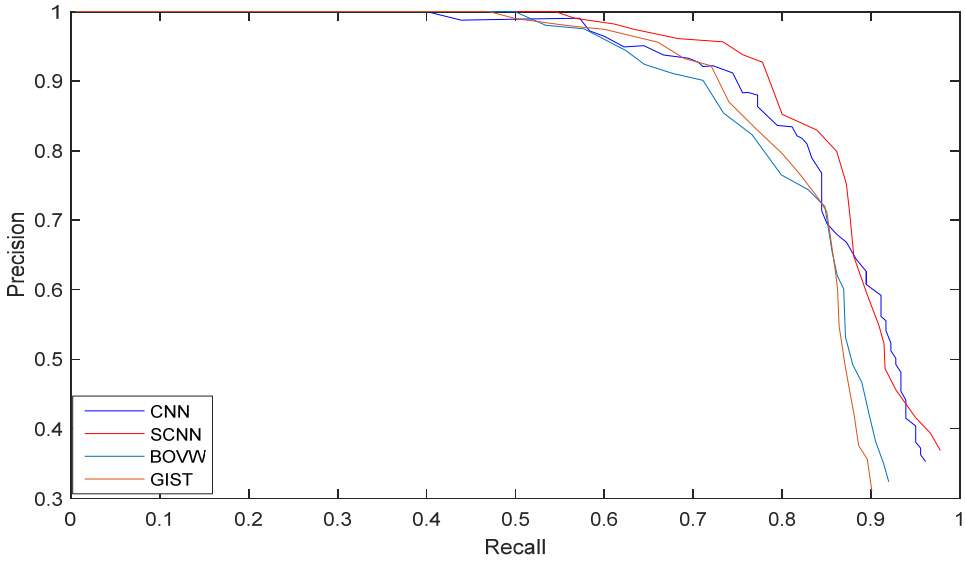


Figure 8. Precision–recall curves for different methods on the New College dataset.

The yellow curve shows the experiment results of the global image descriptors GIST, which is proposed by Singh et al. [24]. In their implementations, they utilized the GIST features to generate global descriptors, and then the Manhattan distance is adopted as a similarity metric for matching. Here, we repeated their feature extraction approach and matching strategy. However, it obviously can be observed that the global descriptor-based method is barely satisfactory in the loop closure detection tasks.

Cummins et al. [11] proposed a baseline method, in which the Bag-of-Visual-Words (BoVW) model was employed to generate codebook representations for matching. Here, we use the green curve to represent the BoVW-based method and we can discover that the BoVW-based method slightly outperforms the method of [24].

So far, the latest algorithm of loop closure detection is deep learning based. Hou et al. [26] utilized AlexNet to automatically learn feature descriptors and then the similarity between descriptors is measured by the L2-norm distance. We use the purple curve to represent the performance and we observe that the standard CNN achieves comparable experiment results to that of the FV and the VLAD-based method.

We test our proposed Siamese CNN-based method on the New College dataset and show its performance with the red curve. Three learning strategies (softmax supervision, contrastive supervision, and multi-task supervision) are annotated with different line shapes. As we can see in Figure 8, the performance of single-task supervision is similar to the standard CNN method. However, Siamese CNN with the multi-task learning strategy yields a more than 90% precision rate with an 83% recall rate, significantly increasing the previous loop closure detection approaches and achieving the state of the art.

3.3. Experiment Results on NYU Dataset

In this subsection, we test our proposed method on an RGB-D loop closure detection dataset NYU. Firstly, Siamese CNN is utilized to extract descriptors in both RGB stream and depth stream; and then, we investigate the performance of three fusion strategies for the RGB and depth stream at different locations of the SCNN, which are early fusion, mid-level fusion, and late fusion.

In early fusion, RGB images and depth images are combined at the beginning of the SCNN. In mid-level fusion, two streams are fused at the intermediate layers of the networks while in late fusion, two streams learn feature descriptors respectively, and then are combined by score averaging. The experiments use three different fusion methods respectively combined with SCNN on the NYU dataset.

Here, we depict the performance of three fusion strategies in Figure 9. In mid-level fusion, the fusion experiments were chosen to be performed in different intermediate layers of the networks and precision–recall curves were obtained. W list the average precision of mid-level fusion from the different intermediate layers in Table 2. We can observe that mid-level fusion at the Conv2 layers achieves the highest average precision score of 0.8111. However, as shown in Figure 9, the late fusion strategy achieves a 92% precision rate with an 82% recall rate, which is best suited for the RGB-D loop closure detection task.

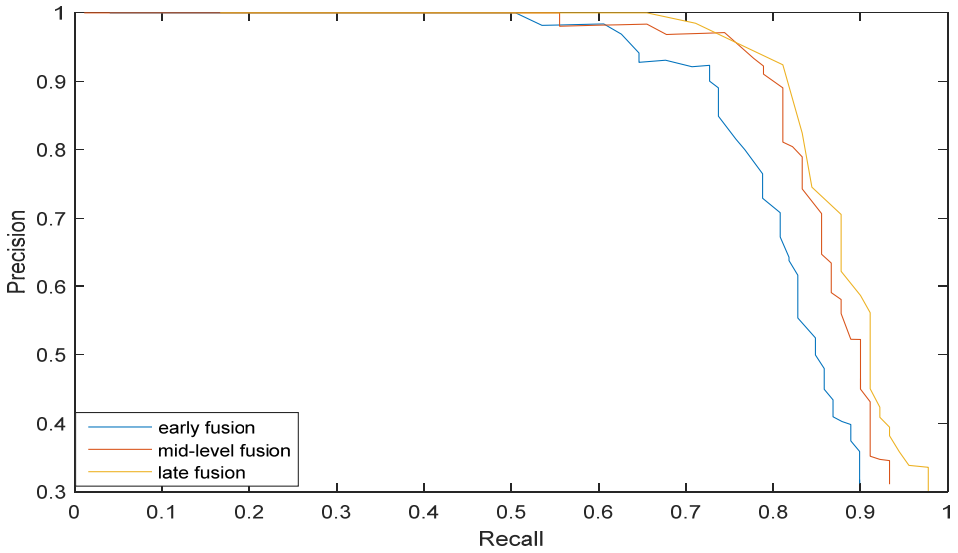


Figure 9. Precision–recall curves for different fusion strategies on the NYU dataset.

Table 2. Average precision score of mid-level fusion at different intermediate layers.

Layers	Conv1	Pool1	Conv2	Pool2	Conv3	Conv4	Conv5	Pool5
AP	0.7444	0.7644	0.8111	0.7778	0.7222	0.7444	0.6667	0.6333

3.4. Computational Time

Another important way to evaluate our approach is computational efficiency. We use Matlab to extract traditional hand-crafted features and deep learning-based features with a 3.40 GHz CPU and 32 GB memory. We also use the Caffe for testing the deep learning-based descriptor extraction time on a PC with NVIDIA TITAN X GPU. The average computational time for different descriptors is listed in Table 3. The reported time here is based on the average of over 400 images, which only contains the feature extraction time. We can find the SCNN descriptors are faster than all hand-crafted descriptors but slower than CNN descriptors on CPU. For the GPU-based extraction, the SCNN descriptors only take 0.029 s per image on average, and it approximates the average extraction time with CNN descriptors.

Table 3. Average computational time for different approaches.

Feature	BoVW	GIST	CNN		SCNN	
Time(s)	1.517	0.524	CPU 0.142	GPU 0.021	CPU 0.208	GPU 0.029

4. Discussion

In this paper, we conduct a comparative study of loop closure detection based on traditional methods and neural network methods. As traditional methods, such as visual Bag-of-Visual-Words and feature descriptor methods, are susceptible to environmental factors, deep learning-based methods are gradually coming into the focus of researchers. This paper focuses on the extraction of image features by convolutional neural networks and applies them to loop closure detection.

The SCNN method proposed in this paper is compared and analyzed with another deep learning method (CNN) and two other artificially designed features (BOVW, GIST) in terms of accuracy and time performance. The experimental results show that, compared with the traditional feature extraction method, the neural network method has significantly higher efficiency and real-time performance. Moreover, the SCNN method greatly improves the efficiency and accuracy of image feature extraction.

In addition, we make full use of the depth information and divide it into three fusion strategies according to the different fusion positions of RGB stream and depth stream in the neural network, which are called early fusion, mid-level fusion, and late fusion, and the experiments prove that the late fusion strategy has a higher accuracy and recall rate.

In the study of loop closure detection using deep learning algorithms, although the accuracy and computational efficiency have been improved, how to combine the improved algorithm with the back-end optimization of visual SLAM requires further research.

5. Conclusions

We introduce Siamese Convolutional Neural Networks (SCNNs) in this paper, which make full use of depth information and use a late fusion strategy to fuse the rgb stream with the depth stream to solve the loop closure detection task in RGB-D slam. We used SCNN-based features from a pre-trained model for scene classification. The experiment results on the public dataset show that SCNN-based image descriptors perform better than CNN and hand-crafted descriptors. In addition, the SCNN-based image descriptors can be applied to RGB-D slam in loop closure detection. Our study can be applied to most indoor scenes. However, for the special task of loop closure detection, an SCNN model suitable for outdoor scenes needs to be trained.

Author Contributions: All named authors initially contributed a significant part to the paper. The experimental model is built by G.X. (Gang Xu) and X.L. Analyses were carried out by G.X. (Gang Xu) and X.Z. The organization of data was led by G.X. (Gang Xu) and G.X. (Guangxin Xing). The descriptions of text use were assisted by X.L. and F.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National First-Class Discipline Program of Light Industry Technology and Engineering, grant number LITE2018-17. And it was also funded by the Open Research Fund of Anhui Key Laboratory of Detection Technology and Energy Saving Devices, Anhui Polytechnic University, grant number 2017070503B026-A01.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: "New College Dataset" at https://ori-drs.github.io/newer-college-dataset/; "NYU dataset" at https://cs.nyu.edu/~silberman/datasets/.

Acknowledgments: The work was supported by National First-Class Discipline Program of Light Industry Technology and Engineering (LITE2018-17). And it was also sponsored by the Open Research Fund of Anhui Key Laboratory of Detection Technology and Energy Saving Devices, Anhui Polytechnic University (2017070503B026-A01).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Klingensmith, M.; Sirinivasa, S.S.; Kaess, M. Articulated Robot Motion for Simultaneous Localization and Mapping (ARM-SLAM). *IEEE Robot. Autom. Lett.* **2016**, *1*, 1156–1163. [CrossRef]
2. Pan, H.Z.; Zhang, J.X. Extending RRT for Robot Motion Planning with SLAM. *Appl. Mech. Mater.* **2012**, *151*, 493–497. [CrossRef]
3. Valencia, R.; Andrade-Cetto, J.; Porta, J.M. Path planning in belief space with pose SLAM. *IEEE Int. Conf. Robot. Autom.* **2011**, *43*, 78–83.
4. Lee, K.H.; Hwang, J.N.; Okapal, G.; Pitton, J. Driving recorder based on-road pedestrian tracking using visual SLAM and Constrained Multiple-Kernel. In Proceedings of the IEEE International Conference on Intelligent Transportation Systems, Qingdao, China, 8–11 October 2014; pp. 2629–2635.
5. Panzieri, S.; Pascucci, F.; Ulivi, G. Vision based navigation using Kalman approach for SLAM. *Int. Conf. Adv. Robot.* **2001**.
6. Huang, G.P.; Mourikis, A.I.; Roumeliotis, S.I. Analysis and improvement of the consistency of extended Kalman filter-based SLAM. In Proceedings of the IEEE International Conference on Robotics & Automation, Pasadena, CA, USA, 19–23 May 2008; pp. 473–479.
7. Montemerlo, M.; Thrun, S.; Roller, D.; Wegbreit, B. FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges. *Int. Jt. Conf. Artif. Intell.* **2003**, *133*, 1151–1156.
8. Lowe, D.G. *Distinctive Image Features from Scale-Invariant Keypoints*; Kluwer Academic Publishers: New York, NY, USA, 2004; Volume 60, pp. 91–110.
9. Bay, H.; Tuytelaars, T.; Gool, L.V. SURF: Speeded Up Robust Features. *Comput. Vis. Image Underst.* **2006**, *110*, 404–417.
10. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. *IEEE Int. Conf. Comput. Vis.* **2011**, *58*, 2564–2571.
11. Oliva, A.; Torralba, A. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [CrossRef]
12. Cummins, M.; Newman, P. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *Int. J. Robot. Res.* **2008**, *27*, 647–665. [CrossRef]
13. Filliat, D. A visual bag of words method for interactive qualitative localization and mapping. In Proceedings of the IEEE International Conference on Robotics and Automation, Roma, Italy, 10–14 April 2007; pp. 3921–3926.
14. Cummins, M.; Newman, P. Highly Scalable Appearance Only SLAM-FAB-MAP 2.0. In *Robotics: Science and Systems (RSS)*; MIT Press: Cambridge, MA, USA, 2009.
15. Kim, A.; Eustice, R.M. Combined visually and geometrically informative link hypothesis for pose-graph visual SLAM using bag-of-words. In Proceedings of the International Conference on Intelligent Robots & Systems, San Francisco, CA, USA, 25–30 September 2011; pp. 1647–1654.
16. Zhang, H.; Liu, Y.; Tan, J. Loop Closing Detection in RGB-D SLAM Combining Appearance and Geometric Constraints. *Sensors* **2015**, *15*, 14639–14660. [CrossRef]
17. Perronnin, F.; Dance, C. Fisher Kernels on Visual Vocabularies for Image Categorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 18–23 June 2007; pp. 1–8.
18. Perronnin, F.; Sanchez, J.; Mensink, T. Improving the Fisher Kernel for Large-Scale Image Classification. In Proceedings of the European Conference on Computer Vision (ECCV), Heraklion, Greece, 5–11 September 2010; Volume 6314.
19. Jegou, H.; Douze, M.; Schmid, C.; Perez, P. Aggregating Local Descriptors into a Compact Image Representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
20. Arandjelovic, R.; Zisserman, A. All about VLAD. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 1578–1585.

21. Liu, Y.; Zhang, H. Indexing visual features: Real-time loop closure detection using a tree structure. In Proceedings of the IEEE International Conference on Robotics & Automation, Saint Paul, MN, USA, 14–18 May 2012; Volume 20, pp. 3613–3618.
22. Korrapati, H.; Uzer, F.; Mezouar, Y. Hierarchical visual mapping with omnidirectional images. *Int. Conf. Intell. Robot. Syst.* **2013**, *8215*, 3684–3690.
23. Singh, G.; Kosecka, J. Visual Loop Closing using Gist Descriptors in Manhattan World. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) Omnidirectional Robot Vision Workshop, Kobe, Japan, 3–7 May 2010.
24. Sunderhauf, N.; Protzel, P. BRIEF-Gist-Closing the Loop by Simple Means. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), San Francisco, CA, USA, 25–30 September 2011; pp. 1234–1241.
25. Liu, Y.; Zhang, H. Visual Loop Closure Detection with a Compact Image Descriptor. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Algarve, Portugal, 7–12 October 2012; pp. 1051–1056.
26. Hou, Y.; Zhang, H.; Zhou, S. Convolutional Neural Network-Based Image Representation for Visual Loop Closure Detection. In Proceedings of the 2015 IEEE International Conference on Information and Automation, Lijiang, China, 8–10 August 2015.
27. Gao, X.; Zhang, T. Loop closure detection for visual slam systems using deep neural networks. In Proceedings of the Control Conference (CCC), 2015 34th Chinese, Hangzhou, China, 28–30 July 2015; IEEE: Piscataway, NJ, USA, 2015.
28. Gao, X.; Zhang, T. Unsupervised learning to detect loops using deep neural networks for visual SLAM system. *Auton. Robot.* **2017**, *41*, 1–18. [CrossRef]
29. Xia, Y.; Li, J.; Qi, L.; Fan, H. Loop Closure Detection for Visual SLAM Using PCANet Features. In Proceedings of the International Joint Conference on Neural Networks, Vancouver, BC, Canada, 24–29 July 2016.
30. Chan, T.H.; Jia, K.; Gao, S.; Lu, J.; Zeng, Z.; Ma, Y. PCANet: A Simple Deep Learning Baseline for Image Classification. *Image Process. IEEE Trans.* **2015**, *24*, 5017–5032. [CrossRef] [PubMed]
31. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006.
32. Krizhevsky, A.; Ilya, S.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
33. Zeiler, M.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 818–833.
34. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
35. Sun, Y.; Wang, X.; Tang, X. Deep learning face representation from predicting 10,000 classes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 1891–1898.
36. Hu, Y.C.; Chang, H.; Nian, F.D.; Wang, Y.; Li, T. Dense crowd counting from still images with convolutional neural networks. *J. Vis. Commun. Image Represent.* **2016**, *38*, 530–539. [CrossRef]
37. Zhang, C.; Li, H.; Wang, X.; Yang, X. Cross-scene crowd counting via deep convolutional neural networks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
38. Smirnov, E. North Atlantic right whale call detection with convolutional neural networks. In *ICML Workshop on Machine Learning for Bioacoustics*; Citeseer: Atlanta, GA, USA, 2013.
39. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. *Indoor Segmentation and Support Inference from RGBD Images. Computer Vision–ECCV 2012*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 746–760.
40. Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; Oliva, A. Learning Deep Features for Scene Recognition using Places Database. In *Advances in Neural Information Processing Systems (NIPS)*; Neural Information Processing Systems Foundation: San Diego, CA, USA, 2014; pp. 487–495.
41. Scherer, S.A.; Kloss, A.; Zell, A. Loop closure detection using depth images. *Eur. Conf. Mob. Robot.* **2014**, *10*, 100–106.