2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

# Siamese-ResNet: Implementing Loop Closure Detection based on Siamese Network

Kai Qiu[1], Yunfeng Ai[1], Bin Tian[2], Bin Wang[3] and Dongpu Cao[4]

*Abstract*— **Deep learning has made significant breakthroughs in the tasks of image classification, detection, segmentation, etc. However, the application of deep learning in robotics is still scarce. SLAM is a fundamental problem in robotics and loop closure detection is an important part of SLAM. This paper attempts to use supervised learning methods to solve the loop closure detection problem in vision SLAM. We proposed Siamese-ResNet network, which combines Siamese network with ResNet to detect loop closure. To show the effectiveness of Siamese-ResNet, we evaluate Siamese-ResNet and FabMap2.0 on several open published datasets, like TUM SLAM dataset and FabMap SLAM dataset. Compared with FabMap2.0, Siamese-ResNet shows higher accuracy, better robustness and shorter time-consuming.**

## I. INTRODUCTION

A SLAM algorithm aims to map an unknown environment while simultaneously localizing the robot. Loop closure detection is a fundamental issue of SLAM research [1,2]. Loop closure detection is the problem of determining whether a mobile robot has returned to a previously visited location. It is critical for building a consistent map of the environment by correcting errors that accumulated over time [3]. In large-scale environments, the problem is more serious and the error of detection may be greater [1,2,4]. Wrong results of loop closure detection will have a serious impact on the consistency of the map, which will lead to the wrong positioning, thus affecting the operation of SLAM.

To develop a loop closure detection algorithm, one class of popular and successful techniques is based on matching the current view of the robot with those in the robot map that correspond to previously visited locations. In this case, the problem of loop closure detection is essentially one of image matching. The traditional method of the Bag-of-Words [1,2,5] model is to construct a dictionary manually, and then describe the sequence of the images collected by the robot with the words in the dictionary. Finally, the result of the loop closure detection is obtained by judging whether the descriptions of the two images are similar or not. This method first constructs the dictionaries by a large amount of manual operations. If the robot comes to a new environment, the constructed dictionaries may not be suitable and the new dictionaries need to be reconstructed, which may result in poor robustness. At the same time, because of the need to extract a large number of key-points in the process of constructing the dictionary, extracting the key-points requires a huge amount of computation, which makes it difficult to achieve real-time loop closure detection.

Deep Learning has made major breakthroughs in the fields of image and natural language processing. In the field of image, deep learning has shown very good results in the tasks of image classification [7,8,9,23,24,26,27], detection [32,33,34], image segmentation, video tracking and even overstep human beings in some tasks. However, deep learning has not received enough attention and application in robot vision SLAM. Deep neural networks trained by a large scale training sets can extract useful features. Deep Neural Networks can extract the feature of images through multi-layer neural network, and judge whether the two images are similar or not by calculating the distance between the features extracted from the two images. At the same time, deep neural networks, once trained and deployed, can extract features very quickly to achieve real-time loop closure detection.

This paper attempts to apply the latest research in deep learning to SLAM robotic vision, hoping to help robots to realize loop closure detection more accurately and quickly by means of the powerful expression ability of deep neural networks. The main contributions are: (1) We propose a loop closure detection method based on Siamese network. In order to improve the ability to extract useful features from the images, we replace the subnet in Siamese network with ResNet, thus our network is called Siamese-ResNet for simplicity. (2) We design a method to construct positive and negative samples to train Siamese-ResNet. (3) We evaluate Siamese-ResNet and FabMap2.0 on several open published datasets and show that Siamese-ResNet can outperform FabMap2.0 in both accuracy and runing time.

This article is organized as follows: Section 3 focuses on the architecture of Siamese-ReNet network and the loss function used in Siamese-ResNet network. Section 4 focuses on the training of Siamese-ResNet network. Section 5 describes the TUM SLAM datasets and FabMap SLAM datasets used in our experiment and three evaluation metrics. Section 6 summarizes the entire article.

*This work was supported in part by the National Natural Science Foundation of China under Grant 61503380 and in part by the Natural Science Foundation of Guangdong Province, China under Grant 2015A030310187.

[1]Kai Qiu and Yunfeng Ai (corresponding author) are with School of Aritificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, 100049 kai.qiu@mails.ucas.ac.cn, aiyunfeng@ucas.ac.cn

[2]Bin Tian is with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100190, and also with the Cloud Computing Center, Chinese Academy of SciencesDongguan, China, 523808 bin.tian@ia.ac.cn

[3]Bin Wang is with School of Software Engineering, University of Science and Technology of China, HeFei, China, 230026 sa615168@mail.ustc.edu.cn

[4]Dongpu Cao is with Department of Mechanical & Mechatronics Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, CANADA dongpu.cao@uwaterloo.ca

## II. RELATED WORK

### A. Loop Closure Detection

SLAM is a key technology for autonomous mobile robots and has received extensive attention and research over the past few decades [3]. This article focuses on the visual SLAM loop closure detection. The simplest method of loop closure detection is to estimate loop closure by matching keyframes. Once a loop closure is found, the global trajectory can be optimized and updated. This method requires large amount of computation to extract and match the feature descriptor of each key frame.

Other loop closure detection methods include odometry-based methods [4] and observation-based methods [1]. The odometey-based approach first assumes that the calculated trajectory is sufficiently accurate. Under such assumptions, we can select a keyframe that is closer in distance to the current frame to find the loop closure. However, in reality, it is difficult to ensure that the hypothesis is right. Appearance-based method does not care about the trajectory. Loop closure is calculated directly through the similarity between key frames to determine the loop closure.

The appearance-based loop closure detection methods have a long history [25,28] and have been successfully applied in practical SLAM systems [2, 29]. Many state-of-the-art appearance-based loop closure detection methods now use the Bag-of-words model [3].FabMap [30] is a visual SLAM method in large scene proposed by Mark Cummins and Paul Newman in 2008. Each frame in the sequence is calculated of the similarity of the previous key frame to determine the loop. SeqSLAM [31] was proposed by Michael J. Milford and Gordon. F. Wyeth in 2012. Instead of calculating the similarity between images, SeqSLAM calculates the similarity between sequence of images. The authors make innovative use of image sequences instead of single-frame images for matching. The author no longer considers the problem of image recognition as finding a uniquely matched image, but rather finding a sequence that best matches the current sequence. ORBSLAM [20,21] is a relatively complete monocular SLAM algorithm proposed by Ral Mur-Artal, Juan D. Tards, JMM Montiel and Dorian Glvez-Lpez in 2015. ORB refers to a rotation invariance feature. The whole algorithm are based on ORB features, unlike SLAM which is based on dense or semi-dense maps. ORBSLAM is a SLAM based on key-point maps.

However, the bag-of-words model has some limitations. First, the vector on the dictionary can only describe the number of occurrences of the word but can not describe the relative relationship between the features in the scene. If the same target appears in both scenes but with different relative positions (indicating that these are two different scenes), the bag-of-words model may decide that the two scenes are the same scene, which will affect the trajectory optimization. Second, the dictionaries in the bag-of-words model are constructed by hand, and such dictionaries are usually only suitable for certain scenes and difficult to adapt to emerging scenes.

There have been attempts to increase the robustness of visual techniques to environment change by performing sensor fusion with lasers. However, these approaches require sensor-to-sensor calibration to ensure that features are represented in a common reference frame, and are unsuitable in unstructured environments or changing conditions where geometry or features are not reliable.

### B. Siamese Network

With the development of deep learning, more and more learning-based visual SLAM metheds are proposed in the last few years [3,10,13,14,15,16,17,19]. [1] and [10] are unsupervised methods to extract features from images. [13] uses well-trained models to extract features and show that features learned by deep neural network is preferable to hand-crafted descriptors. [14] combines features extracted from neural network to precess place recognition task. [15,16,19] use binary features extracted from neural network. [17] uses Siamese network to extract features and match image patches. However, to our best knowledge, there is no work focus on loop closure methed based on Siamese network.

Siamese Network[6] was proposed by Yann LeCun in 2005. It is a classic measure method of similarity. The Siamese networks have achieved good results in tasks such as signature verification [6] and face verification [18]. Through training on Siamese network, similar images will get closer in the feature space and non-similar images will get away from each other in the feature space. [17] proposed several variants of Siamese network and concluded that the 2channel-2stream Siamese network performs best to predict the similarity between image patches. However, the image size used in these experiments are usually very small. In the scene of loop closure detection, image size is often large (eg. 640*480) and the content of images are often complicated.

The main idea is to map the input image to the target space through a network, and use the Euclidean distance to predict the similarity of features in the target space. In training phase, the loss function values of a pair of samples from the same class are minimized, while the loss function values of a pair of samples from different classes are maximised. The structure of the network consists of two sub-networks. For metric learning, the Siamese network inputs are image pairs (containing two images) and the loss function is a contrastive loss function. Through supervised learning, Siamese networks can make the Euclidean distance of the images with high similarity closer in the target space ,and vice versa. The Siamese networks have achieved good results in tasks such as signature verification [6] and face verification [18]. The essence of loop closure detection of visual SLAM is aim to predict the similarity between images. If the similarity of two images is high, we can think of there is a loop closure.

## III. OUR METHOD

In this section, we describe the architecture of Siamese-ResNet and the loss function used in our experiment.

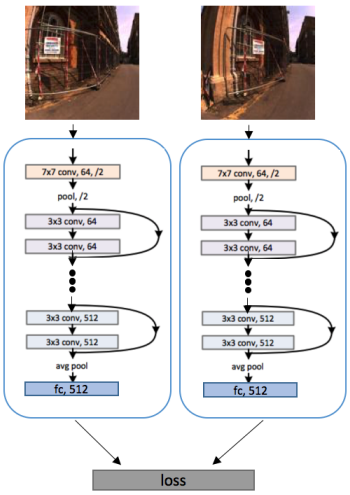


Fig. 1. Architecture of Siamese-ResNet Network.

Siamese-ResNet is an optimization of Siamese network. Compared with Siamese network, Siamese-ResNet can extract better features from input images and show better performance in the task of loop closure detection.

### A. Network Structure

Siamese network was first proposed for signature verification. [17] proposed a 2stream-2channel Siamese Network and shows good performance on matching image patches. However, these works focus on images with small size. In our task, image size is often large and we need to change the network structure of Siamese network.

The original Siamese network consists of two subnets, left-subnet and right-subnet. The subnet is similar to LeNet, which consists of two convolution layers, two pooling layers, tow fully connected layers and one ReLU layer. It shows good performance in the task of signature verification. However, the network scale is too small to extract useful features on images captured in complex environments. On the other hand, ResNet can extract useful features from various images and show good performance in general tasks. An intuitive idea is to replace ResNet with LeNet in Siamese network. In addition, considering that the scale of training dataset in our experiment is not large enough, we choose a ResNet with 27 layers to avoid overfitting. For simplicity, we call our network as Siamese-ResNet. Detail structure of Siamese-ResNet is presented in Figure1. Images are first resized to 224*224. Then a pair of two images are input into the network and flow through the subnets. The last layer is a fully connected layer and output of the fully connected layer is considered to be high level presentation of images. Contrastive loss is used in our experiment which is consistent with the original Siamese network. Weights of left-subnet and right-subnet are shared during training phase and test phase.

Another point worth noting is that the number of neuron in the last fully connected layer is important to the performance. As we increase the number of neuron in the last fully connected layer, we can see a obvious increment of the precision on validation set. Details are given in section IV.

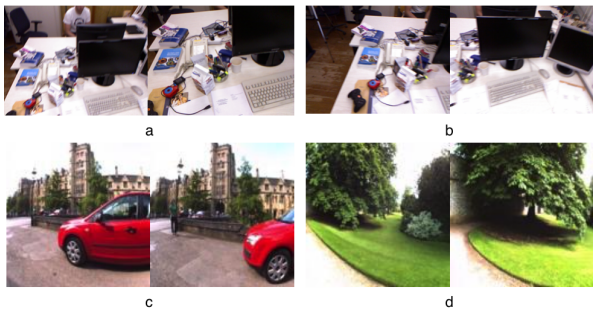


Fig. 2. Images sampled from our dataset. 'a' is a positive sample from TUM dataset. 'b' is a negative sample from TUM dataset. 'c' is a positive sample from FabMap dataset. 'd' is a negative sample from FabMap dataset.

### B. Loss Function

Cross entropy loss is very popular in deep neural networks, especially in the task of classification. In the field of metric learning, contrastive loss function is more often used. By using contrastive loss, similar images will get closer in the feature space and non-similar images will get away from each other in the feature space. The input of the comparison loss function is the Euclidean distance between the features extracted from left and right sub-network. If the label of the two input images is 1, the loss function will reduce the Euclidean distance and vice versa. Euclidean distance can be replaced by other distance, like vector distance. In this paper, we are consistant with [6].In the following formula, L represents loss, N stands for batch size, y represents the label of the sample, d represents the Euclidean distance between two features, and margin is an artificially set threshold, generally set to 1.

$$L = 1/2N \sum_{n=0}^{N} (yd^2 + (1-y)max(margin - d, 0)^2)$$

## IV. TRAINING

Training is processed on TUM dataset. TUM has many image sequences captured from different scenes. It provides a relatively large scale dataset so we choose it as training set. We first divide TUM dataset into three parts: training set, validation set and test set. We use training set to train our models and calculate precision on validation set. The model which gets highest precision on validation set is choosed to compare with FabMap2.0.

Training set has 30k pairs of images, of which half is positive pairs and half is negative pairs. Validation set has 6k pairs of images, also half positive and half negative. In TUM dataset, every image has a corresponding location and orientation. The location is presented as x, y, z coordinate while the orientation is presented by a quaternion. However, the groundtruth of loop closure is not provided and that's why we need to define the loop closure by ourselves. Based on experience, we make a rule to judge whether two images selected from the same sequence come from the same scene:

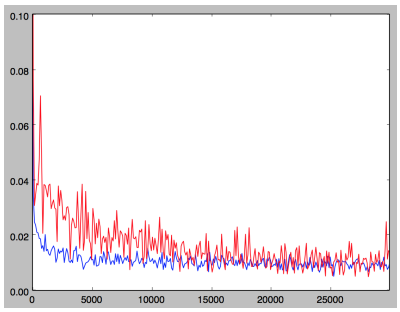


Fig. 3. Loss curve during training. The blue line is the loss curve of Siamese network. The red line is the loss curve of Siamese+ResNet network. Note that Siamese+ResNet network has more neural in the last innnerproduct layer. The distance between outputs of Siamese+ResNet network will be larger and the loss will also be larger.

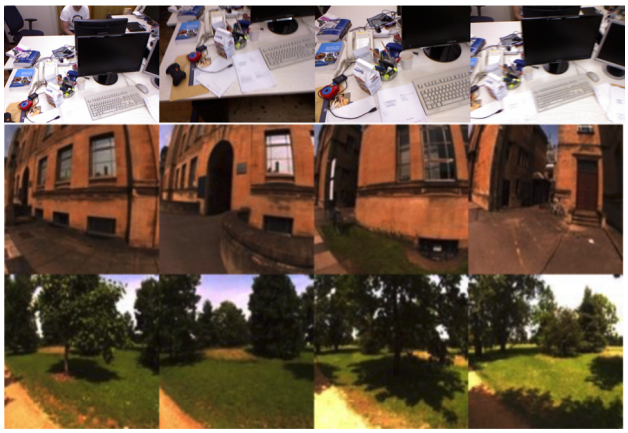

Fig. 4. Datesets used in our experiment. The first line is sampled from TUM dataset and used as training set. The second line and third line is sampled from FabMap dataset and they are both used as test set.

if the distance between two location is less than 2 meters and the angle between two orientation is less than 15°, then the two corresponding images are considered to be a positive pair; otherwise a negative pair. If two images come from diffirent sequences, then they are considered to be a negative pair. In FabMap dataset, the groundtruth of loop closure is given and we can easily compare our experiment results with the groundtruth. Some image samples are presented in Figure 2.

Our training is processed within Caffe [12]. We use TITAN Xp GPU to train our model. To accelerate training, we convert images to lmdb format. Due to the limit of memory, batch size is set to 256. The initial learning rate is set to 0.01. The maximum iteration is set to 30k and learning rate is multiply by 0.1 every 10k iterations. We use SGD to update weights. Momentum is set to 0.9. Loss curve on validation set is show in Figure3.

## V. PERFORMANCE EVALUATION

### A. Datasets and Evaluation Metrics

Two open datasets were used in our experiments, the TUM SLAM dataset and the FabMap dataset. The TUM dataset mainly collects images of indoor scenes while the

TABLE I

| Network | precision |
|---|---|
| Siamese(8) | 69.5% |
| Siamese(128) | 82.0% |
| Siamese + ResNet(256) | 85.0% |
| Siamese + ResNet(512) | 87.7% |

Precision of different networks on validation dataset.

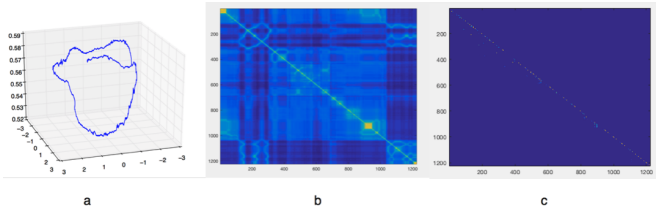


Fig. 5. Results on TUM dataset. 'a' is the odometry of image sequence. 'b' is the similarity matrixs of Siamese-ResNet. 'c' is the similarity matrixs of FabMap2.0.

FabMap dataset is consists of outdoor images. We divide the TUM SLAM dataset into three parts: training set, validation set, and test set. In FabMap dataset, the groundtruth is provided so we can conveniently compare the performance of different methods. Figure 4 shows some images sampled from our datasets. In training set, 30k positive samples and 30k negative samples are constrcted from TUM dataset. In validation set, 6k positive samples and 6k negative samples are constructed from TUM dataset. TUM test set consists of 1225 images. In FabMap City Centre dataset, 1000 images are selected as test set. In FabMap New College dataset, 2500 images are selected as test set.

Evaluation metrics is consistent with [3]. We use two metrics to evaluate our methed and FabMap2.0. Firstly, we draw the similarity matrix of the sequence of images to compare the results of different method with the groundtruth. Secondly, we draw the precision-recall curve to compare different methods. In the image of precision-recall curves, the curve which is closer to the upper right corner has better performance. In addition, we calculate the time consumed by different method.

### B. Results on TUM dataset

Results on TUM consist of two parts: validation set and test set. The precision calculated on validation set is shown in Table 1. The number in the brackets represent the number of neuron in the last fully connected layer. As we can see, more neuron in the last layers will result in better performance. As the images in our task is complex and diverse, we need enough neuron in the last layer to represent high level feature. Siamese-ResNet is more powerful in extracting valid high level features than original Siamese network. Siamese-ResNet gets the best precision(87.7%) on validation set.

Results on TUM test set is show in Figure5. The groundtruth of odometry is shown in Figure 5a. As we can see, there is a loop closure during the movement. Figure 5b is the similarity matrix produced by Siamese-ResNet. Although there are some noise along the diagonal, we can

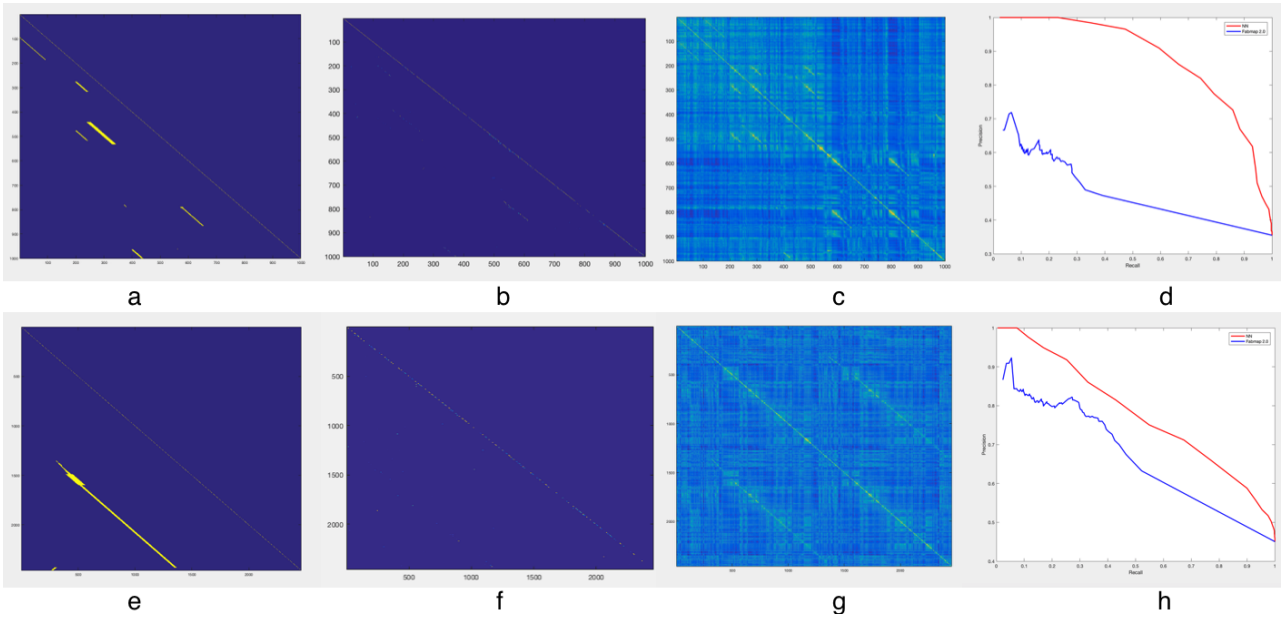


Fig. 6. Resutls on FabMap dataset. The first line is the result of FabMap City Centre dataset. The second line is the result of FabMap New College dataset. 'a' is the similarity matrix of groundtruth of City Centre. 'b' is the similarity matrix of result of FabMap2.0. 'c' is the similarity matrix of result of Siamese-ResNet. 'd' is the precision-recall curve. The red line stands for Siamese-ResNet and the blue line stands for FabMap2.0. 'e' is the similarity matrix of groundtruth of New College. 'f' is the similarity matrix of result of FabMap2.0. 'g' is the similarity matrix of result of Siamese-ResNet. 'h' is the precision-recall curve. The red line stands for Siamese-ResNet and the blue line stands for FabMap2.0.

see the bright area at the lower left corner. This bright area corresponds to the loop closure part. We argue that the noise along the diagonal is produced by the similarity of environment. Images may look very similar in different position and orientation. Figure 5c is the result produced by FabMap2.0. As we can see, FabMap2.0 almost doesn't give any imformation about the loop closure. The only bright area is along the diagonal.

### C. Results on FabMap dataset

Results on FabMap test set is show in Figure6. Figure 6a is the groundtruth similarity matrix of City Centre. Figure 6b is the similarity matrix of FabMap2.0 in City Centre dataset. Figure 6c is the similarity matrix of City Centre dataset of Siamese-ResNet. There are six loop closures in the groundtruth. FabMap2.0 can only detect four of them. Meanwhile, Siamese-ResNet can detect all the six loop closures. The similar phenomenon can be found in FabMap New College dataset. While Siameese-ResNet detected the loop closure in the groundtruth, FabMap2.0 could not detect it.

Precision-recall curve on City Centre dataset is shown in Figure 6d. The red line stands for Siamese-ResNet and the blue line stands for FabMap2.0. As we can see, precision of Siamese-ResNet calculated at every recall value is higher than FabMap2.0. The precision-recall curve is very close to the upper right corner and very smooth, which shows good bobustness. On the other hand, FabMap2.0 is more close to the lower left corner and not smooth, which shows bad robustness. Precision-recall curve on New College dataset is similar with the curve on City Centre.

### D. Time Consuming

Forward the Siamese-ResNet network takes an average time of 0.23s to extract features of one image on an Intel i5 CPU. Calculating the similarity matrix of Siamese-ResNet features of 1000 images takes only 2.8s. However, Fabmap2.0 takes about 22 minutes to compute the similarity matrix for 1000 images on the same device. It is obviously that Siamese-ResNet is much faster and more likely to achieve real-time performance.

## VI. CONCLUSIONS

This paper focuses on the loop closure detection of vision SLAM systems. We propose a loop closure detection method based on Siamese-ResNet network. First, rules for judging whether two images come from the same scene are defined. According to this rule, positive and negative pairs for training are generated from open published SLAM data sets. Through multiple iterations of training, Siamese-ResNet network can make images from similar scenes closer in the target space, while images from non-similar scenes far away from each other in the target space. Finally, by calculating the distance between the two images in the target space, we can predict whether the two images come from the same scene, so as to realize loop closure detection. The whole training process is a supervised learning process. Siamese-ResNet shows good performance on both indoor and outdoor images. Our method is an end-to-end learning process compared to the traditional loop closure detection method based on the bag-of-words model (e.g FabMap2.0) and don't need to construct feature descriptors manually. By training on a large amount of data, Siamese-ResNet can extract more abstract and useful features

than handmade features and such Siamese-ResNet network have better robustness. In addition, the Siamese-ResNet network extract features faster than the classic SIFT, SURF, and is more likely to realize real-time robot vision SLAM. However, deep learning method relies heavily on large scale of datasets. Large distribution difference between training data and test data will lead to poor performance on the test set. We argue that the scale of our training set is too small and don't achive the best performanc of Siamese-ResNet. By expanding the training set and optimizing the network structure, it is expected to further improve the robustness of Siamese-ResNet. In addition, combining our method with the latest SeqSLAM is expected to achieve better performance. This will also be the focus of our research in the future.

## ACKNOWLEDGEMENTS

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

## REFERENCES

[1] M. Cummins, P. Newman, "Appearance-only slam at large scale with fab-map 2.0," The International Journal of Robotics Research, 2011, vol.30(9), pp.1100-1123

[2] M. Labbe and F. Michaud, "Appearance-based loop closure detection for online large-scale and long-term operation," IEEE Transactions on Robotics (TRO), vol.29(3), pp.734-745, 2013

[3] X. Gao, T. Zhang, "Unsupervised learning to detect loops using deep neural networks for visual SLAM system," Autonomous Robots, vol.41, pp.1-18, 2017

[4] D. Hahnel, W. Burgard, D. Fox, S. Thrun, "An efficient fastslam algorithm for generating maps of large-scale cyclic environments from raw laser range measurements," Intelligent Robots and Systems, 2003, pp.206-211

[5] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. Romera, "Towards life-long visual localization using an efficient matching of binary sequences from images," IEEE International Conference on Robotics and Automation, 2015, pp.6328-6335

[6] S. Chopra, R. Hadsell and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, vol.1, pp.539-546

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," International Conference on Learning Representations, 2015.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," Neural Information Processing Systems, 2012, pp.1097-1105

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, pp.770-778

[10] X. Gao, T. Zhang, "Loop closure detection for visual slam systems using deep neural networks," Chinese control conference, 2015, pp.5851-5856

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009, pp.248-255

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," ACM Multimedia, 2014, pp.675-678

[13] Y. Hou, H. Zhang, and S. Zhou, "Convolutional neural network-based image representation for visual loop closure detection," International Conference on Information and Automation, 2015, pp.2238-2245

[14] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," arXiv:1411.1509, 2014

[15] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. Romera, "Fusion and binarization of CNN features for robust topological localization across seasons," IEEE/RSJ International Conference on Intelligent Robots and Systems, 2016.

[16] G. Zhang, M. J. Lilly, and P. A. Vela, "Learning binary features online from motion dynamics for incremental loop-closure detection and place recognition," IEEE International Conference on Robotics and Automation, 2016, pp.765-772

[17] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, pp.4353-4361

[18] Q. Cao, Y. Ying, P. Li, "Similarity Metric Learning for Face Recognition," International Conference on Computer Vision, 2013, pp.2408-2415

[19] V. Balntas, L. Tang, K. Mikolajczyk, "Binary Online Learned Descriptor For Efficient Image Matching," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, pp.2367-2375

[20] R. Murartal, J. Montiel, J. Tardos, "ORB-SLAM: a versatile and accurate monocularSLAMsystem," IEEE Transactions on Robotics, 2015, pp.1147-1163

[21] R. Murartal, J. Tardos, "ORB-SLAM2: An Open-SourceSLAMSystem for Monocular, Stereo, and RGB-D Cameras," IEEE Transactions on Robotics, 2017, pp.1-8

[22] J. Sturm and N. Engelhard and F. Endres and W. Burgard and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," Intelligent Robots and Systems, 2012, pp.573-580

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, pp.2818-2826

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," International Conference on Machine Learning, 2015, pp.448-456

[25] G. Dudek, D. Jugessur, "Robust place recognition using local appearance based methods," International Conference on Robotics and Automation, 2000, pp.1030-1035

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, pp.1-9

[27] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," National Conference on Artificial Intelligence, 2016, pp.4278-4284

[28] F. Endres, J. Hess, J. Sturm, D. Cremers, W. Burgard, "3-d mapping with an rgb-d camera," IEEE Transactions on Robotics, 2014, vol.30(1), pp.177-187

[29] Y. Latif, C. Cadena, J. L Neira, "Robust loop closing over time for pose graph slam," International Journal of Robotics Research, 2013, vol.32(14), pp.1611-1626

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

[31] M. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," IEEE International Conference on Robotics and Automation, 2012, pp.1643-1649

[32] Girshick, R. B., Donahue, J., Darrell, T., and Malik, J, "Rich feature hierarchies for accurate object detection and semantic segmentation," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014, pp.580-587

[33] Girshick R B. "Fast R-CNN," International Conference on Computer Vision, 2015, pp.1440-1448

[34] Ren S, He K, Girshick R B, et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, vol.39(6), pp.1137-1149