ISSN:0975-9646

Comb-Net: A New Architecture of CNNs for Vision-based Indoor Localization

Zeyad Farisi, Lian-Fang Tian, Xiang-Yang Li, Bin Zhu

***Abstract*—Aimed at improving the localization accuracy of current vision-based indoor localization algorithm, a new vision based indoor localization method using a creative convolutional neural network architecture (Comb-Net) and a novelty training method is proposed. Comb-Net is composed by an intact U-Net, two first 13 layers of VGG16Net, a fully connection layers of VGG16Net and an ArcFace classifier. U-Net is applied to extract semantically segmented image from RGB image, two first 13 layers of VGG16Nets are used to extract location features from RGB images and semantically segmented images, respectively. These location features are then combined together by the fully connection layers of VGG16Net, ArcFace classifier is applied to obtain the final classification results. What's more, a multi-layer transfer learning training method for complex convolutional neural networks is designed, transfer learning decreases the number of training set and the layered strategy makes the model easy to be trained. Experimental results show that the proposed algorithm can localize indoor mobile robot accurately, compared to RGB image based method and semantic image based method, the accuracy of our method increased by 10.7% and 11.8%, respectively.**

***Index Terms*—CNN, Indoor Localization, semantic segmentation, transfer learning.**

I. INTRODUCTION

With the development of artificial intelligence technology, various types of robots have been widely used. In the application of mobile robots, real-time detecting and monitoring the location of robots is the prerequisite for better service to human beings. Therefore, the wireless positioning technology for mobile robots has gradually become a research hotspot. In outdoor environment, global positioning system (GPS), plough navigation systems and cellular positioning technology can meet most of the positioning needs, but these methods are not suitable for indoor positioning.

For indoor localization assignments, Wi-Fi based method[1], Bluetooth based method[2] and Radio Frequency identification technology[3] were proposed and widely used. However, bottlenecks exist in these methods. Wi-Fi-based methods are vulnerable to multi-path effects, Bluetooth-based methods exist mutually interference and RF-based methods require expensive equipment support.

Vision-based methods[4][5] which can realize real-time positioning only by a normal RGB camera, avoid all these bottlenecks mentioned above and provide a new way for indoor positioning. Traditional vision-based methods usually apply image matching strategy. However, these methods are vulnerable to shallow layer features such as shooting angle, illumination changes, interfered by non-fixed content in scenes and so on. With the popularization of deep learning technology, many scholars use DCNNs (Deep Convolutional Neural Networks) to extract the deep position features of images. However, the localization accuracy of these methods are low and location features of adjacent location points are difficult to distinguish.

In order to overcome the shortcomings and deficiencies of existing vision-based indoor positioning technology, a new vision-based indoor localization algorithm is proposed. Besides the RGB image location feature, the length, shape and area of paths are extracted by semantic segmentation, these path features are then combined with the RGB image location feature to locate the positioning of the camera. For extracting both location features of RGB image and semantic segmentation image, a new convolution neural networks architecture (Comb-Net) is designed. Comb-Net is composed by an intact U-Net[6], two first 13 layers of VGG16Net[7], a fully connection layers of VGG16Net[7] and an ArcFace classifier[8]. U-Net is applied to extract semantically segmented images from RGB images, two first 13 layers of VGG16Nets are used to extract location features from RGB images and semantically segmented images, respectively. These location features are then combined together by the fully connection layers of VGG16Net and ArcFace classifier is applied to get the final classification results. For training the complicated Comb-Net, a training method based on multi-layer transfer learning is designed, transfer learning decreases the number of training set and the layered strategy makes the model easy to be trained. Based on path features, the location of an image to be classified can be determined by the length, area and shape of paths, which makes up the shortcomings of traditional vision-based indoor

This work is supported by Guangdong Key R&D Program (2019B020214001), Guangdong Key R&D Program (2018B010109001), Guangzhou's Major Industrial Technology Tackling Plan (2019-01-01-12-1006-0001).

Zeyad Farisi is with School of Automation Science and Engineering, South China University of Technology, Guangzhou, China and with College of Community Service Department of Engineering and Science. Tabah University, Medinah, Saudi Arabia (e-mail: z_doo@hotmail.com)

Lian-Fang Tian, Xiang-Yang Li are with the School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, Guangdong , China (e-mail: chlftian@scut.edu.cn, xiangyangli@scut.edu.cn).

Bin Zhu is with the School of Automation Science and Engineering, Jiangxi college of applied technology, Ganzhou 341000, Jiangxi, China (e-mail: 14866266@qq.com).

localization. At last, the ArcFace classifier is employed to locate the positioning of the mobile robot.

Major contributions of our paper are as follows:

1、 A new architecture of convolutional neural network is proposed;

2、 Multilayer transfer learning is proposed to train complicated convolutional neural networks.

The rest of paper is organized as follows. Section 2 introduces the proposed architecture of convolutional neural network. Section 3 describes the network model training method using multilayer transfer learning. The experimental analysis and results are discussed in section 4. Section 5 presents the limitation of our algorithm. Data availability is introduced in section 6 and conclusion is presented in section 7.

II. THE ARCHITECTURE OF PROPOSED COMB-NET

In order to get both location features of RGB image and semantic segmentation image, a new architecture of CNNs is designed. It can be seen in Fig.1, five parts are included in Comb-Net. There are a U-Net, two first 13 layers of VGG16Net (VGG1 and VGG2), a fully connection layers of VGG16Net (VGG3) and an ArcFace classifier. An RGB image is input into both U-Net and VGG2, U-Net is used to generate semantic image of the input RGB image, the semantic image is then applied to extract path regions which is used to get location features using VGG1. VGG2 is used to get location features from the input RGB image directly. These two kinds of features are then combined together by VGG3. At last, we can obtain the localization result using ArcFace classifier.

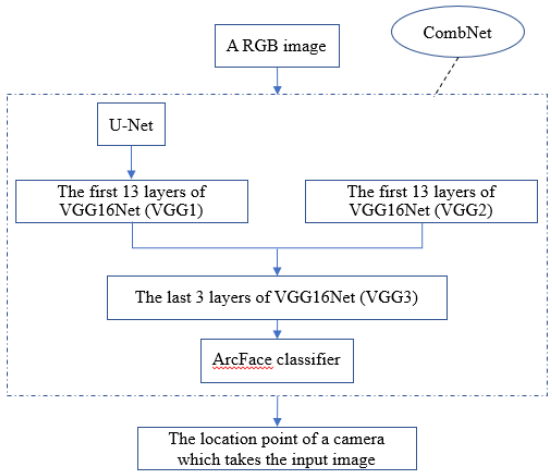

Fig. 1. The flowchart of our algorithm

A. *The structure of the U-Net*

Convolutional neural networks were often applied in classification tasks, where we figure out an image belongs to which class. However, in many visual tasks, such as image segmentation and object detection, a class label is supposed to be assigned to each pixel. U-Net is designed to predict the class label of each pixel. Specifically, using U-Net, we extract path parts in the image, that is to say, the whole image would depart into two parts: path pixels or non-path pixels.

The architecture of U-Net can be seen in Fig.2, totally, it has 23 convolutional layers. In the left side, it is a repeat application of the typical architecture which consists of 3*3 convolutions, batch normalization operation and a rectified linear unit. A 2*2 max pooling with stride 2 is used to down sampling. The typical architecture of CNNs overlaps many times to extract image features from shadow one to moderate one then to deep one. Each feature image is then sent to the right side of the model. Then, for each layer, inversely operation is done to retain a same size of the input and useful information is extracted, useless information is abandoned, all these layers concatenate together to compose the final output. At last, each pixel in the image labeled as required.

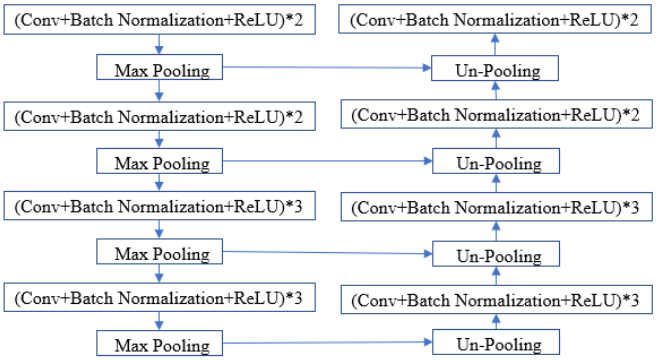

Fig. 2. The architecture of the U-Net

B. *The architecture and configuration of VGG1 and VGG2*

VGG16Net is a typical vision-based classifier, where the input is an image, the output is a single class label. The first 13 layers of VGG 16Net are CNN layers which are used to feature extraction, the last 3 layers of VGG16Net are fully connection layers which are used to concatenate features and generate classification. The architecture of VGG1 can be seen in Fig.3, it is the first 13 layers of the typical VGG16Net. The input of VGG1 is a semantic image. It is used to extract location features in semantic image.

All the input (training or testing) should be adjusted to a fixed-size. For highlight the difference, the mean RGB value is subtracted for each pixel before training or testing. The image is then processed by a stack of convolutional layer with 3×3 convolution kernel (which is the smallest size of a filtering window). The stride of convolutional operation is fixed to 1 pixel, the spatial resolution is preserved after convolutional operation by the spatial padding with 1 pixel for 3×3 convolutional layers. The width of the first convolutional layers is 64, then increasing by a factor of 2 after each section and it reaches 512 in the end. Meanwhile, the size of the feature map will reduce by half after each max-pooling.

Five max-pooling layers are designed to down-sampling for decrease the size of the feature map. Not all the convolutional layers are followed by max-pooling, the location of the max-pooling layer can be seen in Fig. 3. Max-pooling is performed over a 2×2 pixel window, with stride 2. Besides max-pooling, all hidden layers are equipped with the rectification non-linearity (ReLU) and batch normalization. ReLU is used to transform the linear propagation net into Nonlinearity. The batch normalization is applied to limit large variances into a reasonable scope.

The architecture and configuration of VGG2 is the same as VGG1. The only difference is the input and function. The input

of VGG2 is an RGB image. It is used to extract location features in RGB image.

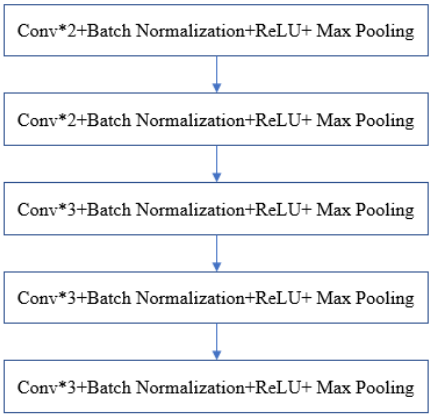

Fig. 3. The architecture of VGG1 and VGG2

C. *The architecture and configuration of VGG3*

The architecture of VGG3 is the last 3 layers of VGG16Net. It is the three Fully-Connected (FC) layers which is used to receive location features from both semantic image and RGB image. Its output is the classification of the location. It can be seen in Fig.4. the first two layers is consist of affine layer, batch normalization, ReLU and Dropout, the last layer only include affine layer. The first affine layer has 4096 channels to accept both location features, the second affine layer has 1024 channels to make location features accumulated. channels of the third affine layer depend on the number of the classification, i.e. the number of the location point.

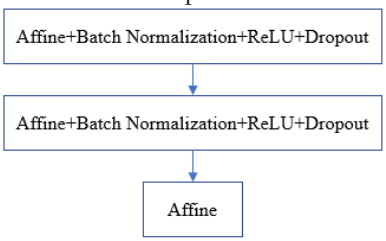

Fig. 4. The architecture of VGG3

D. *ArcFace classifier*

After fully-connection layers, a classifier is connected to the end of the network. Traditionally, the softmax classifier is frequently used. But the Softmax classifier would cause wrong classification results in vision-based indoor localization, especially for adjacent location regions where features of different location regions are similar. In order to solve this problem, an ArcFace classifier is selected in this paper. The loss function of Softmax is:

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \qquad (1)$$

Where x_i is the *i-th* feature of the y_i*-th* class, W_j is the weight of the *j-th* node in the last fully connected layer, b_j is the bias value, N is the mini-batch size and n is the class number. To simplify the formula, we set $b_j = 0$ and $W_j^T x_i$ can rewrite to $\|W_j\|\|x_i\|\cos\theta_j$, θ_j is the angle between vector W_j and vector x_i, using L_2 normalization, $\|x_i\|$ can be rewrite to s and $\|W_j\|=1$. Then the loss function can be depicted as follow:

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}} \qquad (2)$$

The loss function of ArcFace classifier is designed to enhance intra-class compactness and inter-class discrepancy by adding angular margin penalty m into the angular between W_j and x_i:

$$L_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}} \qquad (3)$$

Fig. 5 shows decision plans of softmax loss and ArcFace loss under binary classification. no gap between two classes for softmax classifier which means classification confusion would show up on the edge of decision plan and the ArcFace loss can afford evident gap between classes where classification confusion can be avoided.

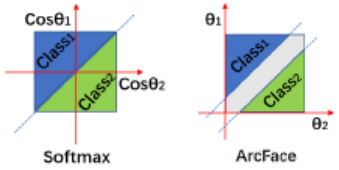

Fig. 5. The decision plan of two loss functions under binary classification

III. MULTILAYER TRANSFER LEARNING BASED TRAINING METHOD

The architecture of our net is complicated, many convolutional layers and many parameters are included in. It is a common sense that the size of the considered networks is directly proportional to the size of the required training sets. According to traditional training method, large amount of training sets is needed in our model, compared to the complexity of the model, our training samples are not enough. In order to solve this deficiency, a multilayer transfer learning training method is proposed. public database and special dataset are both employed of the training process. A strategy from local to global is used, specific process is as below:

1. Pre-training U-Net, VGG1 and VGG2 with ImageNet dataset first to improve the generalization ability of the model, respectively.
2. Transfer learning of the whole U-Net is carried out by images labeled path pixel or non-path pixel for each pixel. After this procedure, the network can generate a path-only image from an RGB image.
3. VGG1 is strengthened by these path-only images labeled with location categories. After transfer learning, the network can obtain location features from the path-only image.
4. VGG2 is strengthened by RGB images labeled with location categories. After transfer learning, the location features of RGB images can be obtained.
5. Combining U-Net, VGG1, VGG2, VGG3 and ArcFace classifiers as a whole, the whole network is retrained again by RGB images labeled with location categories. In this stage, the weights of U-Net will not be adjusted and weight parameters of VGG1, VGG2 and VGG3 will be fine-tuned.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Experimental basis. We go on experiments in a factory environment, the whole scene is segmented into 18 regions and each region has a location center (in the author's former work [9], we segment the same factory into 9 region). The intention of our algorithm is to predict the position of a mobile robot in real time by images taking at this position. Fig.6 shows the factory plan with 18 location centers. The experimental platform is a mobile robot which is equipped with an RGB camera, mobile control devices and a mini-computer in which placed our trained Comb-Net. The experimental platform can be seen in Figure 7. For testify the superiority of our algorithm, several algorithms are used for comparation, there are RGB image + VGG16Net + softmax (method1), semantic image + VGG16Net + softmax (method2), Comb-Net + Softmax (method3) and our proposed algorithm: Comb-Net + ArcFace(method4).

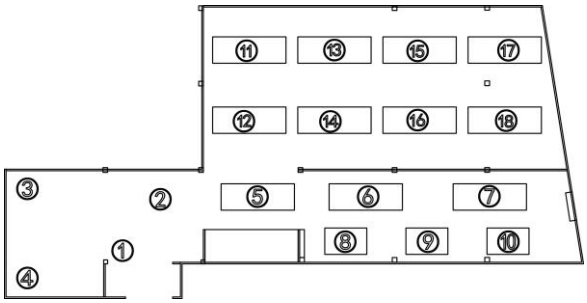

Fig. 6. Floor plan of the experimental scene

Fig. 7. The experimental platform

Dataset. 1.3M images in ImageNet dataset are used for pre-train. 100 images taken from each location point and their corresponding semantic images are used to construct our dataset, in which 90 images are applied for training and 10 images are applied for testing. Fig. 8 shows RGB images taken from different locations and their corresponding semantic images.

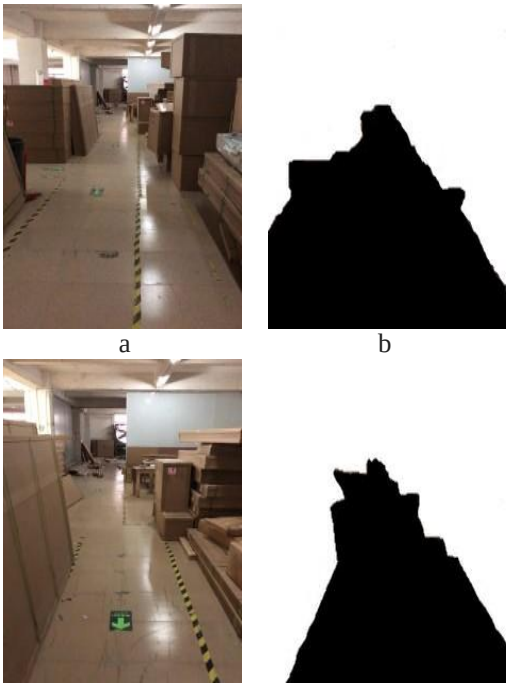

Fig. 8. RGB images and their corresponding semantic images

Training accuracy and Loss. Fig .9 and Fig. 10 show curves of training set accuracy and curves of loss function of four methods, respectively. we can see in fig. 8: when iteration time increase, the training set accuracy of four methods improve. At the beginning, method1 and method2 are better than method3 and method4, that's because compared to complicated Comb-Net, smaller net is easier and faster to find optimal solution. But Comb-Net includes more location features which makes it more sophisticated to distinguish different classes, so when iteration time further increases, performances of method3 and method4 are better than method1 and method2. At last, method4 achieves the highest accuracy among these four methods because compared to Softmax classifier, ArcFace classifier is better to distinguish images with similar features. We can see similar circumstances in Fig 9. At the beginning, the loss value of method1 and method2 are smaller than method3 and method4, but when iteration time increases, method3 and method4 turn to better than method1 and method2. At last, method4 achieves the smallest loss value among these four methods.

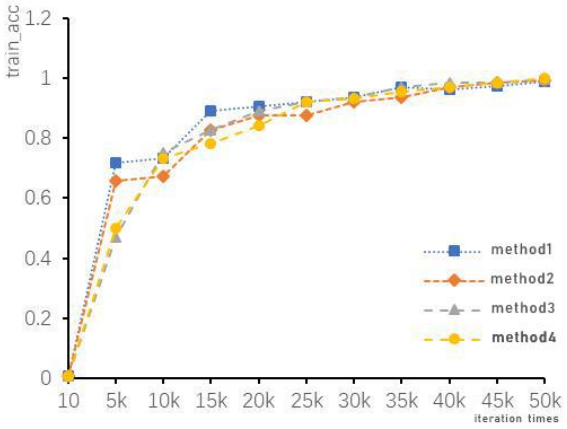

Fig. 9. Curves of training set accuracy for four methods

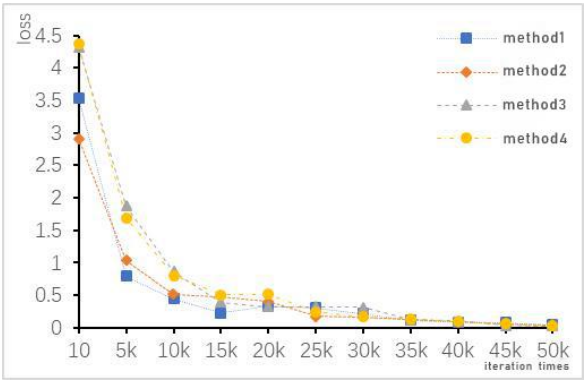

Fig. 10. Curves of loss function for four methods

Actual experiments. After trained the Comb-Net with ImageNet dataset and our own dataset, our model is used for vision-based location classification, just input an RGB image and we can get the location classification which the camera belongs to. Actual experiments were done using four methods mentioned above, 30 images of each location region were used for testing these algorithms and experiment results depicted by confusion matrix were shown in Fig. 11, Fig. 12, Fig. 13 and Fig. 14. We can see in Fig. 11 that correct classification time of method1 is 444 and wrong classification time is 96, the accuracy of method1 is 82.2%, more errors happened in middle regions of the building, that is because location features taken from RGB images in these nearby location points are similar and hard to distinguish. In Fig. 12, correct classification time of method1 is 438 and wrong classification time is 102, the accuracy of method2 is 81.1%, compared to method1, errors in middle regions decreased, but errors in side regions increased. That is because semantic images use path information as location features which can identify nearby middle location region, but these features in side regions are not obvious. Our proposed Comb-Net in this paper combined both location features of RGB image and semantic image to identify the location and errors in each region decreased, we can see the result in Fig. 13, correct classification time of method3 is 473 and wrong classification time is 67, the accuracy of method3 is 87.5%. we get the highest accuracy of classification in Fig. 14 where correct classification time of method4 is 502 and wrong classification time is 38, the accuracy is 92.9%.

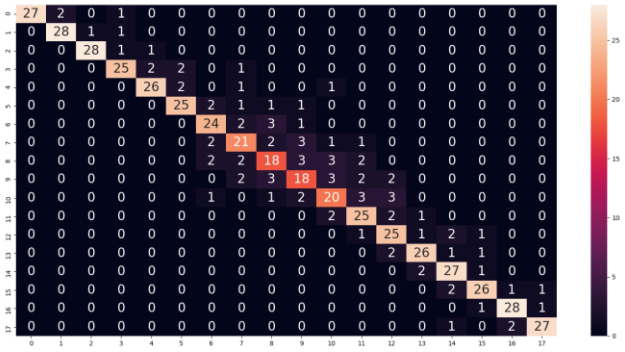

Fig. 11. Experimental result of method1

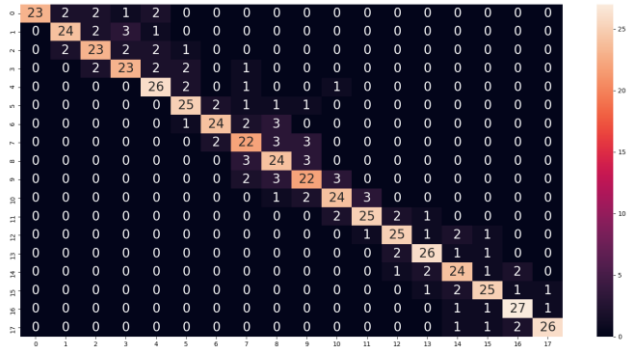

Fig. 12. Experimental result of method2

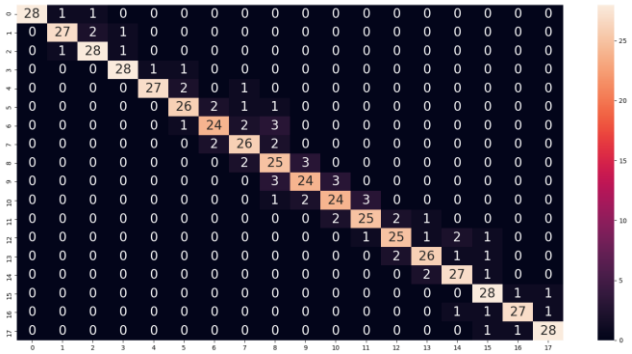

Fig. 13. Experimental result of method3

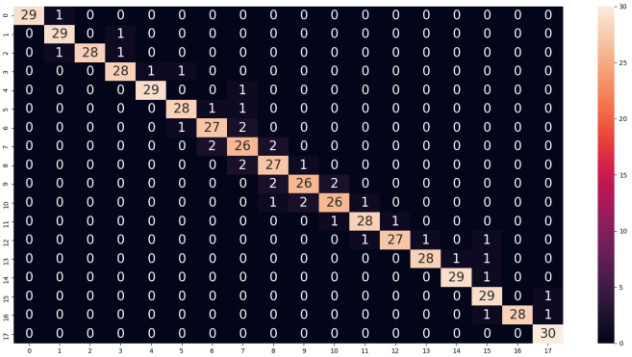

Fig. 14. Experimental result of method4

V. LIMITATION

The accuracy of actual experiment is a little low than the accuracy by test dataset. That is because images taking from camera in real time would be affected by environment changes. Besides, some limitations exist in our algorithm.

1. Paths should clean and have not detainees staying on it.
2. The camera should not be covered or occluded, when a human or object stands too close in front of the camera, the localization would fail.
3. The motion of the robot should be slow, stable, smooth and in an illumination sufficient environment, to make sure the stability of the camera when taking images.
4. If the work environment changes, the dataset used for transfer learning should be reconstructed for the new environment.

VI. DATA AVAILABILITY

Image-Net dataset which used in pre-training stage can be found in [10], the public dataset is designed for image classification. It can be used to train our model so that our

model can extract shallow and moderate features in images. Our own dataset can be found in [26]. We constructed our own dataset in a factory environment where we test our algorithm. 18 location points are included in the dataset and 100 samples are take in each location point with different shooting angles. For each sample, the location point is labeled and its corresponding semantic segmentation image in which retain path regions only is prepared. In this project, too many similar samples are useless to the accuracy and 100 samples for each location point are suitable.

VII. CONCLUSION

A creative architecture of convolutional neural networks (Comb-Net) is proposed in this paper to improve positioning accuracy of vision-based indoor localization. For training complicated neural networks with limited dataset, a multi-layer transfer learning training method is designed. The Comb-Net is composed by an intact U-Net, two the first 13 layers of VGG16Net, a last 3 layers of VGG16Net and an ArcFace classifier. U-Net is applied to assign a class label for each pixel of an RGB image which generate a semantic image. two the first 13 layers of VGG16Nets are used to extract location features from RGB images and semantically segmented images, respectively. These location features are then combined together by the last 3 layers of VGG16Net, ArcFace classifier is applied to obtain the final classification results. The combination of location features in semantic image and RGB image make the localization more accurate. Transfer learning decreases the number of training set and the layered strategy makes the model easy to be trained. It is no doubt that the Comb-Net architecture can be used to solve many other tasks.

REFERENCES

[1] Moustafa A, Moustafa E, Marwan T. “WiDeep: WiFi-based Accurate and Robust Indoor Localization System using Deep Learning”, 2019 IEEE International Conference on Pervasive Computing and Communications, Kyoto, Japan, IEEE, March 2019: 1883-1890.

[2] Cabrera E, Camacho D. “Towards a Bluetooth Indoor Positioning System with Android Consumer Devices,” The IEEE International Conference on Information Systems and Computer science, Quito, Ecuador, 2017: 56-59.

[3] Qiu L, Huang Z, Wirstrom N, et al. “3DinSAR: Object 3D localization for indoor RFID applications,” The IEEE International Conference on RFID, Orlando, USA, IEEE, 2016:101-108.

[4] Desai A, Ghagare N, Donde S. “Optimal Robot Localization Techniques for Real World Scenarios”, 2018 Fourth International Conference on Computing Communication Control and Automation, Pune, India, IEEE, Aug, 2018: 1861-1868.

[5] Walch F, Hazirbas C, Sattler T, et al. “Image-based localization using LSTMs for Structured Feature correlation,” The IEEE International Conference on Computer Vision, Venice, Italy, IEEE, 2017: 627-637.

[6] Olaf R, Philipp F, Thomas B. “U-Net: Convolutional Networks for Biomedical Image Segmentation”, International conference on Medical Image Computing and Computer-Assisted Intervention, Los Angeles, USA, Springer, Nov, 2015: 234-241.

[7] Simonyan K, Zisserman A. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. Computer Science, 2014, 35(4): 386-400.

[8] Deng J, Guo S, Zafeiriou C. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” Computer Vision and Pattern Recognition, 2018, 32(6): 1-11.

[9] Zeyad F, Tian L F, Li X Y, Zhu B. “Vision based Indoor Localization Method Via Convolution Neural Network”, International Journal of Advanced Computer Science and Applications, 2019, 10(7): 55-59.

[10] http://www.image-net.org/

[11] https://pan.baidu.com/s/1oR7fg_sZHe_qHtpSH7PUzg, password: ix4q.