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

Real-time Traffic Scene Segmentation Based on Multi-Feature Map and Deep Learning

Linhui Li, Weina Zheng, Lingchao Kong, Ümit Özgüner, Wenbin Hou, Jing Lian*

Abstract—**Visual-based semantic segmentation for traffic scene plays an important role in intelligent vehicles. In this paper, we present a new real-time deep fully convolution neural network (FCNN) for pixel-wise segmentation with six channel inputs. The six channel inputs include the RGB three channel color image, the Disparity (D) image generated by stereo vision sensor, the image to describe the Height (H) of each pixel above road ground, and the image to describe the Angle (A) between each pixel normal direction and the predicted direction of gravity, which are defined as a RGB-DHA multi-feature map. The FCNN is simplified and modified based on AlexNet to meet the real-time requirements of intelligent vehicle for environmental perception. The proposed algorithm is tested and compared in Cityscapes dataset, yields global accuracies 73.4% and 22ms for 400×200 resolution image with one Titan X GPU.**

Index Terms—Intelligent vehicle, traffic scene segmentation, multi-feature map, deep learning

I. Introduction

Traffic scene segmentation is a fundamental task for intelligent vehicles in detecting obstacles, planning paths, and navigating autonomously. Semantic segmentation, also known as image parsing or image comprehension [1], aims to divide the image into predefined non-overlapping regions and translate them into abstract semantic information. In recent years, with the rapid development of computer hardware, especially the Graphics Processing Unit (GPU), the emergence of large-scale markup data, and the application of deep Convolutional Neural Networks (CNNs) in image classification and object detection have been rapidly developed and have become the current mainstream image segmentation methods. Recently, most studies have been focused on improving the accuracy of semantic segmentation by making the network deeper and larger. However, increasing the parameters often comes at the expense of the memory of computers and leads to the network being slower. So, how to improve accuracy under the premise of ensuring real-time functionality is one of the most important tasks in deep learning.

The advent of depth sensors makes it possible to obtain depth information, which contain more positional information than the RGB image. There are two ways to apply the depth map to the image semantic segmentation: one is to combine the raw depth images and the RGB images into a four-channel RGB-D image as a CNNs input [2]-[4], and the other is to input the images containing richer depth information and the RGB images into two CNNs, respectively [5]-[7]. Specifically, with the help of rich information about the object relationships that is provided in the depth images, both methods can achieve a better performance than using only the RGB image. However, entering data into two CNNs will increase the number of parameters causing the network slow down. Therefore, in this paper, to improve the accuracy, the Disparity, Height, and Angle maps (DHA) are fused with RGB images into 6-channel RGB-DHA map and used directly as input data.

This paper focuses on building a fast functioning semantic segmentation network with good performance, especially for the road targets that drivers are more concerned about. As a result, a new network architecture is proposed, and then the depth map and its derived height and norm angle maps are added to train the network for higher accuracy. The main work can be stated as follows:

- A fully convolution neural network called D-AlexNet network is developed based on AlexNet [8], which has a simple structure containing several convolutional layers to increase forward speed of the network.
- The proposed D-AlexNet achieves 2.2x+ speedup of reference and reduces parameters by 39+ times.
- The 6-channel RGB-DHA map can achieve a better result in semantic segmentation than only using RGB images as input, especially for identifying road targets in a traffic scene, such as pedestrians and cars.

II. Related Work

A. RGB Semantic Segmentation

A Fully Convolutional Network (FCN) [9] replaces the last fully-connected layers of the traditional neural network with the convolution layers, which lays the foundation for FCN to be applied to semantic segmentation. Deeplab [10], proposed by L. C. Chen et al., obtained better results by reducing the stride, using the hole algorithm, and the conditional random field to fine-tune the network. SegNet [11], [12] achieves the pixel-level semantic segmentation by using encoder-decoder structures to restore the feature maps from the higher layers with spatial information from the lower layers. In [13], [14], multi-scale feature ensembles are used to increase performance. The PSPNet [15] complete the prediction by aggregating context information.

*Resrach supported by the National Natural Science Foundation of China (Grant Nos. 51775082, 61473057 and 61203171) and the China Fundamental Research Funds for the Central Universities (Grant Nos. DUT17LAB11 and DUT15LK13).

L. Li, W. Zheng, L. Kong, W. Hou and J. Lian are with the School of Automotive Engineering, Faculty of Vehicle Engineering and Mechanics, Dalian University of Technology, Dalian 116024, China. And J. Lian is the corresponding author. (e-mail: lilinhui@dlut.edu.cn; zhengweina_1993 @mail.dlut.edu.cn; 31703177klc@mail.dlut.edu.cn; houwb@dlut.edu.cn; lianjing@dlut.edu.cn).

Ümit Özgüner is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, 43210 USA (e-mail: ozguner.1@osu.edu).

To perform segmentation in real-time on existing hardware. Some of the methods have been used to speed up the network. SegNet [12] improved forward speed by reducing the number of layers in the network. A. Chaurasia et al. [16] linked the encoder blocks to the corresponding decoder directly to decrease the processing time. Z. Hengshuang et al. [17] proposed compressed-PSPNet-based image cascade network that incorporates multi-resolution branches under proper label guidance to yield real time inference.

B. Semantic Segmentation with Depth Information

Compared to the single RGB images, depth maps contain more location information that is a benefit to semantic segmentation. In [18], the raw depth image was simply treated as a one-channel image, and CNNs were then applied to extract features for indoor semantic segmentation. In [5], depth information was used as three channels: horizontal disparity, height above ground, and norm angle. Qi et al. [19] proposed a 3D Graph Neural Network (3DGNN) that builds a k-nearest neighbor graph, and finally boosted the prediction. The above works prove that using more characteristic information as input to train the network helps to improve the accuracy of semantic segmentation.

III. NETWORK ARCHITECTURE

In general, a deeper network structure will result in better semantic segmentation although it often comes at the expense of having many training parameters and a longer running time, which can't meet the real-time requirements of intelligent driving. To tackle this problem, intuitively, we believe that reducing network parameters and simplifying the network model can speed up the network, and, moreover, adding the depth information can improve network performance. Motivated by AlexNet [8] and N. Hyeonwoo [20], who proposed an encoder-decoder network architecture based on the VGG16 network, the proposed deep fully convolution neural network architecture is shown in Figure 1, and includes 11 convolutional layers, 3 pooling layers, 3 upsampling layers, and 1 softmax layer.

In the new network structure, AlexNet is modified in the following ways to make it suitable for pixel-level semantic segmentation tasks:

- In order to adapt the network to images of different sizes, the full connectivity layer of AlexNet is removed. Then, the stride of the first convolutional layer is changed from 4 to 1 and the kernel size of the maximum pooling layer is changed from 3×3 to 2×2.
- Experimental results showed that the existence of a packet structure in convolutional layer can't improve the accuracy of the final semantic segmentation. Therefore, we removed the second, fourth, and fifth convolutional packets and deleted the two LRN layers.
- The existence of internal covariates will increase the difficulty of deep network training. This paper added a batch normalization layer between each convolution layer and ReLU layer to solve this problem.
- The convolution kernels of all the convolutional layers are unified to be 3×3 in size, and the number of convolution kernel outputs is 96.

With reference to the upsampling method used by Z. D. Matthew et al. [21], we record the maximum eigenvalue position of each pooling window in the pooling process and put it in the corresponding position in the upsampling process. The decoder is the mirror structure of the encoder, except for its sixth convolutional layer where the kernel size is 1×1. The output of the decoder network is K feature maps and then it is fed to the softmax layer to produce a K-channel class probability map, where K is the number of classes. The result of the segmentation is that each pixel of the image corresponds to the class with the largest predicted probability.

IV. MULTI-FEATURE MAP

DHA images can contain richer image feature information compared to using the raw depth information for learning the deep network. This process includes the steps described below.

A. Horizontal Disparity Map

The left and right images obtained from the Cityscapes dataset can be used to generate the disparity map with a stereo matching algorithm. According to the degree of matching, the stereo vision matching algorithm can be divided into three categories: the local matching algorithm, the semi-global matching algorithm, and the global matching algorithm. The global matching algorithm gets the highest matching accuracy, and worst real-time performance. The local matching algorithm is the fastest, but its matching accuracy is very low.

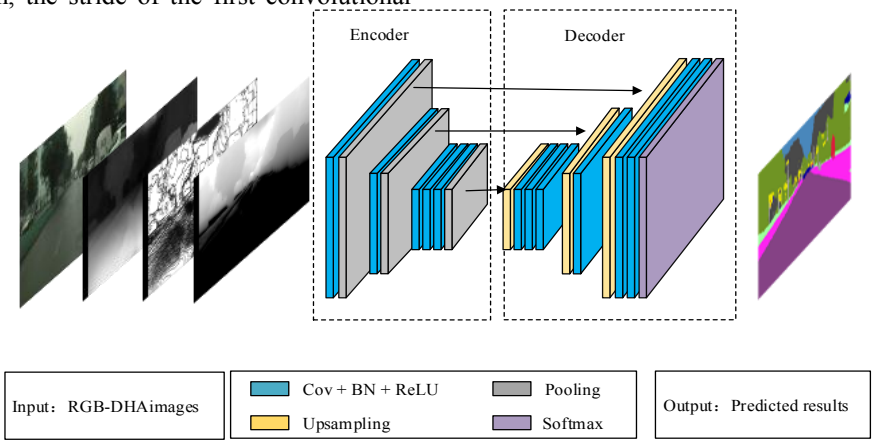


Figure 1. The structure of D-AlexNet network.

The semi-global matching algorithm can better match the accuracy and real-time computing needs, so for this paper we chose this method for obtaining the disparity map.

An edge preserving smoothing method proposed by M. Dongbo [22] is used to improve the segmentation accuracy by optimizing the coarse disparity map and making the disparity value more continuous.

B. Height Above Ground

Based on the obtained parallax maps, the $P(x, y, z)$ points in the world coordinate system corresponding to $P'(u, v)$ pixels in the image coordinate system can be obtained by equations (1) and (2),

$$z = \frac{fb}{d} \tag{1}$$

$$y = \frac{v - c_y}{f_y} \times z \tag{2}$$

where x and y are the coordinates of point P in the world coordinate system. z is the distance between point P and the camera. f and b are the focal length of the camera and the baseline length of the two cameras, respectively. f_y and C_y are the internal parameters of the camera, and y is the height of the pixel. A correction is required since the camera's installation does not guarantee complete parallelism with the ground plane. A part of the ground area in the parallax map is selected, and the least squares method is used to fit the ground. By assuming that the fitted ground plane equation is $Y = aX + bZ + c$, the value of a, b, and c can be obtained by equation (3). After correcting the ground, the actual pixel height can be obtained by equation (4).

$$\begin{bmatrix} \sum X_i^2 & \sum X_i Z_i & \sum X_i \\ \sum X_i Z_i & \sum Z_i^2 & \sum Z_i \\ \sum X_i & \sum Z_i & n \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum X_i Y_i \\ \sum Z_i Y_i \\ \sum Y_i \end{bmatrix} \tag{3}$$

$$h = y - \left(aX + bZ + c \right) \tag{4}$$

In the height map, the sky, the building, and the tree correspond to a large height value, while the more important objects such as vehicles and pedestrians correspond to a relatively small height value. To highlight important goals, equation (5) is used to transform the height value corresponding to each pixel to generate the height image whose height value is between 0 and 255.

$$\begin{cases} h' = 255 \times \log_2 \dfrac{h}{0.1} / \log_2 \dfrac{15}{0.1}, & h > 0.1 \\ h' = 0\ , & h <= 0.1 \end{cases} \tag{5}$$

C. Surface Normal

For urban traffic scenes, generally, the road surface is horizontal, and the surface of objects, such as buildings, traffic signs, vehicles, and so on, are in a vertical direction. According to these characteristics, an algorithm can be used to find the direction which is the most aligned to or most orthogonal to the locally estimated surface normal directions at as many points as possible. Hence, to leverage this structure, the algorithm proposed by G. Saurabh et al. [5] is used to determine the direction of gravity.

Finally, by calculating the angle between the pixel normal direction and the predicted direction of gravity, the required angle information can be obtained.

V. Experiments and Analysis

The experiments were conducted based on the learning platform Caffe. In addition, all our experiments were performed on the software and hardware shown in Table I.

TABLE I. Training Software and Hardware Conditions

Project	Content
CPU	Intel Xeon E5-2620
RAM	32GB
GPU	GeForce GTX TITAN X
Operating System	Ubuntu 14.04 LTS
Cuda	Cuda7.5 with Cudnn v5
Deep Learning Framework	Caffe

A. Dataset and Evaluation Metrics

We applied our system to the recent urban scene understanding data Cityscapes, which contains 5,000 finely and 20,000 coarsely annotated images. In addition, the dataset provides left and right views captured by a stereo camera, providing the chance of obtaining parallax and depth maps. In this paper, 5,000 finely annotated images were selected, and were split into training, verification, and a test set. These sets contained 2,975, 500, and 1,525 images, respectively. The image size was converted to 200 × 400 to shorten training time and reduce memory consumption. To mark the significant traffic information, traffic scenarios were classified into 11 categories, including roads, road borders, buildings, poles, traffic signs, trees, lawns, skies, people, cars, and bicycles or motorcycles. Both the global accuracy rate and network forward time were used for evaluation.

B. Training Process

In the training process, the weight of the convolution layers were initialized in the same way as AlexNet, and the method used by H. Kaiming et al. [23] was applied to initialize the weight of the batch normalization layers. Cross-entropy was employed as a loss function for training the network and calculating the loss value. In the back-propagation phase, stochastic gradient descent was adopted to optimize the network weights. The initial learning rate and momentum were set to 0.01 and to 0.9, respectively. In addition, the weight decay was set to 0.0005 to prevent overfitting of the network. It is noteworthy that to maintain the purity of the data and simplify the training process, we trained our network without data augmentation and no pre-trained model with other datasets was used.

For every 300 training times, we conducted an accuracy assessment on the validation set and saved a snapshot. The validation accuracy, training loss value curves based on RGB-DHA images are shown in Figure 2. More iterations may mean higher accuracy. However, when the accuracy and

the loss start to converge, it is feasible to stop the training. Therefore, the network was iteratively trained 10,000 times, and the Caffe model with the highest accuracy was selected as the model that was finally used for scene segmentation.

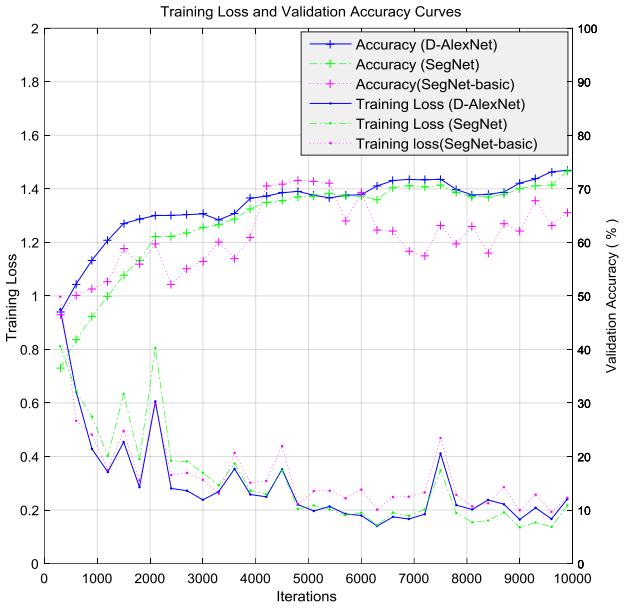


Figure 2. Training loss and Accuracy curves of different networks.

C. Comparison and Analysis

We first evaluated how our proposed network was useful in speeding up semantic segmentation taking SegNet[11] and SegNet-basic[12] as the baseline. When taking RGB images and RGB-DHA images as the input data, the performance results of the networks are shown in table II. Our proposed network structure was 2.2 times faster than SegNet and 1.8 times faster than SegNet-basic. From Figure 2 and table II we can find that our proposed architecture can achieve better real-time results with competitive segmentation results. Furthermore, for each network frame, the validation accuracy obtained using RGB-DHA images is higher than that obtained using RGB images, which also indicates that more characteristic information is useful to improve the performance of networks.

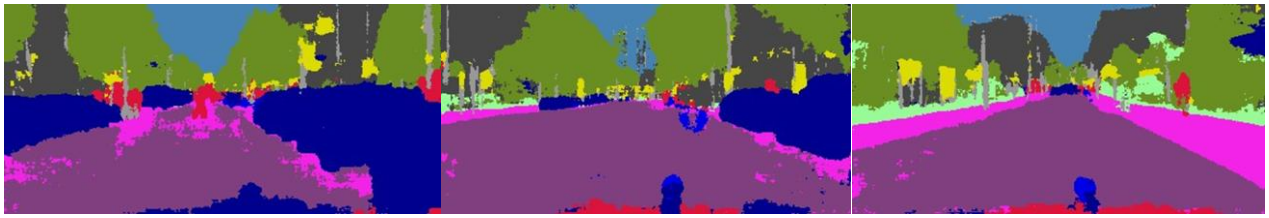
TABLE II. PERFORMANCE OF BASELINE AND D-ALEXNET ON VALIDATION SET OF CITYSCAPES

Project	SegNet	SegNet-basic	D-AlexNet
Accuracy (%) (RGB)	72.6	69.9	69.4
Accuracy（%）(RGB-DHA)	73.2	71.5	73.4
Training time (ms)	140	77	63
Testing time (ms)	48	28	22
Memory(MB)	117.8	5.7	3.0

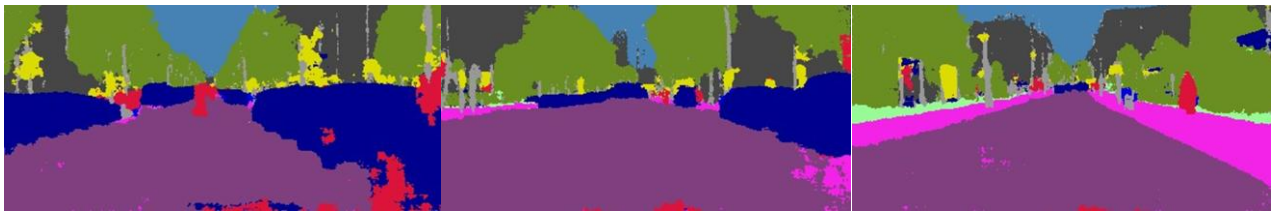
TABLE III. ACCURACY COMPARISON OF DIFFERENT INPUT ON VALIDATION SET OF CITYSCAPES

Input	Road	Sidewalks	Building	Poles	Traffic signs	Trees	Lawns	Sky	Pedestrian	cars	Two-wheelers	Mean accuracy	Global accuracy
RGB(%)	90.5	64	61	72.3	74	80.6	67.3	96.7	79	85.8	62.6	75.8	69.4
RGB-D(%)	91.6	67.7	75.1	74.2	78.6	80.2	72.5	98.1	81.2	88.7	59.3	78.8	73.1
RGB-H(%)	90.3	65	72.9	68.4	76.5	78.7	67.5	98.4	83.8	84.9	61.6	77.1	71.5
RGB-A(%)	91.1	68.5	71.6	75.3	73.6	80.8	77.8	98	77.1	89.9	65.8	79.0	72.4
RGB-DHA(%)	91.2	70.2	76.3	72.8	78	80.7	68.9	99	84	91	52	78.5	73.4

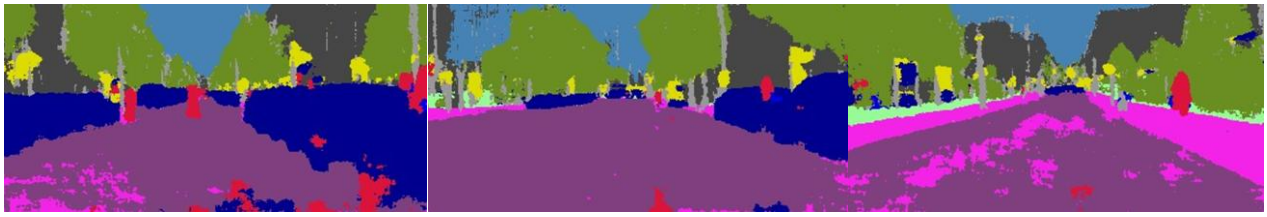
(a) Original set of color images of the test samples



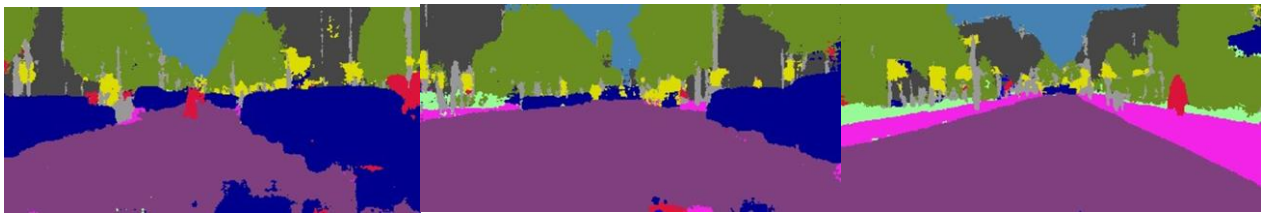
(b) Semantic segmentation results based on RGB images



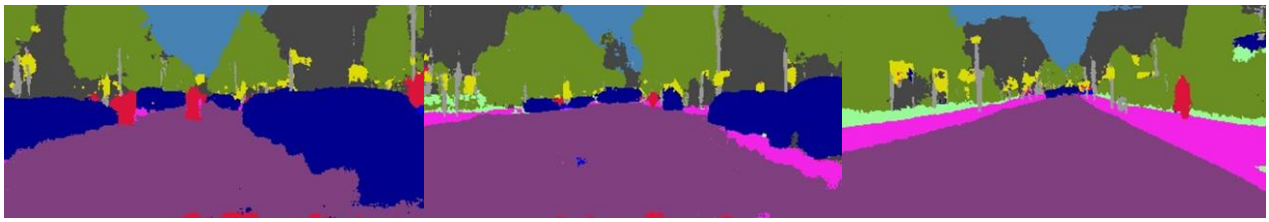
(c) Semantic segmentation results based on RGB-D images



(d) Semantic segmentation results based on RGB-H images



(e) Semantic segmentation results based on RGB-A images



(f) Semantic segmentation results based on RGB-DHA images

Figure 3. Examples of semantic segmentation results in the test set.

To further understand the efficiency gain in each feature map, we first respectively merged three feature maps obtained from Section 4 with the RGB images into 4-channel images, and then all 3 feature images were merged with RGB images into 6-channel images. After that, both 4-channal and 6-channal images were used as input data for training the network. The testing results are shown in Table Ⅲ, from which we can draw conclusion that the segmentation accuracy based on 4-channel and 6-channel images was obviously improved when compared with the one based on 3-channel images. Under the same training parameters, the global accuracies obtained from RGB-D, RGB-H, RGB-A, and RGB-DHA images are 3.7%, 2.1%, 3%, and 4% higher than those obtained from raw RGB images, respectively. With RDB-DHA 6-channel images as input, our proposed system finally achieves a segmentation accuracy of 73.4%.

Figure 3 shows the results of semantic segmentation on the test set of our network model with 3-channels, 4-channels, and 6-channels, respectively, as inputs. As shown, the segmentation results obtained based on RGB images were sometimes rough and there were many wrongly classified pixels on road or around boundary contours of different categories. For example, many pixels in the road surface were misclassified as sidewalks in the left image of Figure 3 (b). The effect based on the four-channel images was generally batter than that based on RGB three-channel images, and the RGB-DHA images can further improve the segmentation accuracy, which shows less error classification points.

In addition, when use RGB-DHA images as net input, road targets such as pedestrians and cars got higher segment accuracy than use RGB image as net input. For example, the pedestrians segment accuracy rise from 79% to 84% and the cars segment accuracy rise from 85.8% to 91%. Some details comparison are shown as Figure 4. It can be seen that pedestrian and car in Figure 4(c) and Figure 4(f) have clearer contours than in Figure 4(b) and Figure 4(e), which will helpful for behavior analysis of different road targets.

(a) Pedestrian image (b) RGB as input (c) RGB-DHA as input

(d) Car image (e) RGB as input (f) RGB-DHA as input

Figure 4. Examples of detail comparison of pedestrians and cars.

VI. CONCLUSION

This paper presents a traffic scene semantic segmentation method based on a novel deep fully convolutional network (D-AlexNet) and multi-feature map (RGB-DHA). The network achieves good real-time performance by 22ms for each 400×200 resolution image on Titan X GPU. Disparity maps, height maps, and angle maps are obtained from the original RGB images and fused into 6-channel images to train the network. Experiments show that using multi-feature map as the input to the network can achieve 4% higher segmentation accuracy compared with using RGB image-as input. In the future, we will focus on more efficient deep network to joint semantic segmentation, targets tracking and parameter identification together.

ACKNOWLEDGMENT

The authors would like to thank Dr. Rencheng Zheng for his contribution to the fruitful discussions.

REFERENCES

[1] W. Fan, A. Samia, L. Chunfeng and B. Abdelaziz, "Multimodality semantic segmentation based on polarization and color images," *Neurocomputing*, vol. 253, pp. 193-200, Aug. 2017.

[2] L. Linhui, Q. Bo, L. Jing, Z. Weina and Z. Yafu, "Traffic scene segmentation based on RGB-D image and deep learning (Periodical style—Submitted for publication)," *IEEE Transactions on Intelligent Transportation Systems*, submitted for publication.

[3] F. David, B. Emmanuel, B. Stéphane, D, Guillaume, G. Alexander et al, "RGBD object recognition and visual texture classification for indoor semantic mapping," in *IEEE International Conference on Technologies for Practical Robot Applications*, Woburn, 2012, pp. 127-132.

[4] H. Farzad, S. Hannes, D. Babette, T. Carme and B. Sven, "Combining semantic and geometric features for object class segmentation of indoor scenes," *IEEE Robotics & Automation Letters*, vol. 2, no. 1, pp. 49-55, Jan. 2017.

[5] G. Saurabh, G. Ross, A. Pablo and M. Jitendra, "Learning rich features from RGB-D images for object detection and segmentation," *Lecture Notes in Computer Science*, vol. 8695 LNCS, no. PART 7, pp. 345-360, 2014.

[6] G. Yangrong and C. Tao, "Semantic segmentation of RGBD images based on deep depth regression (Periodical style—Submitted for publication)," *Pattern Recognition Letters*, submitted for publication.

[7] E. David and F. Rob, "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Feb. 2015, pp. 2650-2658.

[8] K. Alex, S. Ilya and H. E. Geoffrey, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, June 2017.

[9] S. Evan, L. Jonathan and D. Trevor, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640-651, Apr. 2017.

[10] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs (Periodical style—Submitted for publication)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted for publication.

[11] V. Badrinarayanan, A. Handa and R. Cipolla. "Segnet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *Computer Science*, May 2015.

[12] V. Badrinarayanan, A. Kendall and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for scene segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481-2495, Dec. 2017.

[13] F. Xia, P. Wang, L. C. Chen and A. L. Yuille, "Zoom better to see clearer: human and object parsing with hierarchical auto-zoom net," in *European Conference on Computer Vision*, Switzerland, 2016, pp. 648-663.

[14] C. Liang-Chieh, Y. Yi, W. Jiang, X. Wei and Y. L. Alan, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Las Vegas, July 2016, pp. 3640-3649.

[15] Z. Hengshuang, S. Jianping, Q. Xiaojuan, W. Xiaogang and J. Jiaya, "Pyramid scene parsing network," in *the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 2017, pp. 2881-2890.

[16] A. Chaurasia, and E. Culurciello, "Linknet: exploiting encoder representations for efficient semantic segmentation," *arXiv preprint* arXiv: 1707.03718, 2017.

[17] Z. Hengshuang, Q. Xiaojuan, S. Xiaoyong, S. Jianping and J. Jiaya, "ICNet for Real-Time Semantic Segmentation on High-Resolution Images," *arXiv preprint*, arXiv:1704.08545, 2017.

[18] H. Caner, M. Lingni, D. Csaba and C. Daniel. "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *13th Asian Conference on Computer Vision*, Taipei, Nov. 2016, vol. 10111 LNCS, pp. 213-228.

[19] Q. Xiaojuan, L.Renjie, J. Jiaya, F. Sanja and U. Raquel, "3D Graph Neural Networks for RGBD Semantic Segmentation," in *IEEE International Conference on Computer Vision*, Venice, Oct, 2017, pp. 5209-5218.

[20] N. Hyeonwoo, H. Seunghoon and H. Bohyung, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Feb. 2015, pp. 1520-1528.

[21] Z. D. Matthew and F. Rob, "Visualizing and Understanding Convolutional Networks," in *13th European Conference on Computer Vision*. Sep. 2014, Vol. 8689 LNCS, no. PART 1, pp. 818-833.

[22] M. Dongbo, C. Sunghwan, L. Jiangbo, H. Bumsub, S. Kwanghoon and D. N. Minh, "Fast global image smoothing based on weighted least squares," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5638-5653, Dec. 2014.

[23] H. Kaiming, Z. Xiangyu, R. Shaoqing and S. Jian. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Dec. 2015, pp. 1026-1034.