RGB-DI Images and Full Convolution Neural Network-Based Outdoor Scene Understanding for Mobile Robots

Zengshuai Qiu, Yan Zhuang, *Member, IEEE*, Fei Yan, *Member, IEEE*,
Huosheng Hu, *Senior Member, IEEE*, and Wei Wang, *Senior Member, IEEE*

***Abstract*— This paper presents a multisensor-based approach to outdoor scene understanding of mobile robots. Since laser scanning points in 3-D space are distributed irregularly and unbalanced, a projection algorithm is proposed to generate RGB, depth, and intensity (RGB-DI) images so that the outdoor environments can be optimally measured with a variable resolution. The 3-D semantic segmentation in RGB-DI cloud points is, therefore, transformed to the semantic segmentation in RGB-DI images. A full convolution neural network (FCN) model with deep layers is designed to perform semantic segmentation of RGB-DI images. According to the exact correspondence between each 3-D point and each pixel in a RGB-DI image, the semantic segmentation results of the RGB-DI images are mapped back to the original point clouds to realize the 3-D scene understanding. The proposed algorithms are tested on different data sets, and the results show that our RGB-DI image and FCN model-based approach can provide a superior performance for outdoor scene understanding. Moreover, real-world experiments were conducted on our mobile robot platform to show the validity and practicability of the proposed approach.**

***Index Terms*— Full convolution neural network (FCN), mobile robots, multisensor data fusion, outdoor scene understanding, semantic segmentation.**

I. Introduction

MOBILE robots have been widely applied in many real-world applications. Their control and decision making are based on data acquisition, data analysis, outdoor scene reconstruction, and scene classification. In this paper, we study the problem of outdoor scene understanding for mobile robots based on laser point clouds and monocular image data. There are three tasks in our research. The first one is how to implement multisensor data fusion among laser scanner, monocular vision, and IMU. The second one is how to find an optimal projection algorithm to transform the irregularly distributed point clouds to 2-D images. The third one is how to design a better semantic scene understanding framework based on the multisensor fusing data.

Vision and laser sensors are widely used to accomplish scene understanding tasks in autonomous navigation of mobile robots. In order to improve the accuracy of vision and laser data fusion, Hu *et al.* [1] proposed a calibration method that is able to solve the problem of simplified perspective-three-point and perspective-three-line, respectively. The key technique in this paper is to transform the estimation of laser range finder pose into a simplified perspective-three-point problem. A vehicle localization method was proposed in [2], which can fuse data from multiple sensors such as a stereoscopic system, a laser range finder, and GPS. For more accurate laser-based vehicle motion estimation, an outlier-rejection invariant closest point method was proposed to reduce the matching ambiguities of scan alignment [2]. A new algorithm to perform registration from unordered point clouds was proposed in [3], which is an automatic and model free one. More important, this algorithm does not rely on any prior information about the objects in the scene.

Vision-based outdoor scene understanding is also a hot issue in the field of mobile robots. A task of image semantic segmentation with deep learning was accomplished by Chen *et al.* [4]. They proposed a kind of spatial pyramid pooling algorithm to robustly segment objects at multiple scales and improve the localization of object boundaries by combining methods from deep convolutional neural networks (DCNNs) and probabilistic graphical models so as to combine the responses at the final DCNN layer with a fully connected conditional random field (CRF). In recent years, with the development of deep learning theory, the segmentation results of image semantics have made a great progress.

In [5], convolutional networks were trained end-to-end, pixels-to-pixels, and made the improvement on the previous best result in semantic segmentation. The key insight of this paper was to build “fully convolutional” networks that took the input of arbitrary size and produced corresponding output with efficient inference and learning. They defined and detailed the space of fully convolutional networks, and explained their application to spatially dense prediction tasks, and drawn connections to prior models. An efficient framework

Manuscript received October 22, 2017; revised April 21, 2018; accepted April 21, 2018. This work was supported by the National Natural Science Foundation of China under Grant 61503056, Grant 61375088, and Grant U1608253. The Associate Editor coordinating the review process was Dr. Huang-Chen Lee. *(Corresponding author: Yan Zhuang.)*

Z. Qiu, Y. Zhuang, F. Yan, and W. Wang are with the School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: ai_qzs@mail.dlut.edu.cn; zhuang@dlut.edu.cn; fyan@dlut.edu.cn; wangwei@dlut.edu.cn).

H. Hu is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, U.K. (e-mail: hhu@essex.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIM.2018.2834085

to perform recognition and grasp detection of objects from RGB-D images of real scenes was presented in [6]. They proposed a novel method to encode an RGB-D point cloud into a representation that facilitated the use of large convolution neural networks to extract discriminative features from RGB-D images.

In order to obtain the better result of the 3-D scene understanding in outdoor scenes, the semantic segmentation of 3-D point cloud is the fundamental task especially for mobile robots. Zermas *et al.* [7] proposed pipeline aiming to solve the problem of 3-D point cloud segmentation for data received from a LIDAR in a fast and low complexity manner for real-world applications [7]. An adaptive surface model approach was proposed for the segmentation of 3-D point clouds into geometric surfaces [8]. In [9], a two-layer classification model was proposed, in which the first layer consists of a Gaussian mixture model and the second layer consists of semisupervised classifier trained in a large of data set of manual labeling. Many scholars try to combine the classification results of 2-D images and 3-D point clouds to improve the classification accuracy of scene. A fast and efficient segmentation algorithm for 2-D images and 3-D point clouds of building facades trained a sequence of boosted decision trees using autocontext features [10]. An algorithm for detecting the interest regions of object's surfaces in images and point clouds has been proposed [11], which can accomplish application-viewpoint selection so as to provide the most descriptive presentation of the object's surface.

The aim of this paper is to solve the problem of semantic segmentation of the RGB-DI point clouds (including color, depth, and intensity) generated by multiple sensors on our mobile robot. Considering the irregular and unbalanced distribution of point data, the location representation of adjacent points in point cloud is much difficult than that of image. In order to extract the deep feature representation from the multiattribute point clouds and make the distribution of point clouds be more standardized, a novel projection algorithm is proposed to generate RGB-DI images (including color, depth, and intensity) from RGB-DI point clouds so that the semantic segmentation of multiattribute cloud points with complex structure is transformed into RGB-DI image semantic segmentation. A full convolution neural network (FCN) model is proposed, which is suitable for solving the semantic segmentation of RGB-DI images, and then the semantic segmentation results of the RGB-DI image are mapped back to the original point clouds to realize the scene understanding.

The rest of this paper is organized as follows. Section II describes the system framework proposed for our mobile robot to conduct the outdoor scene understanding, which has three subsystems. In Section III, a new projection algorithm is proposed to generate RGB, depth, and intensity (RGB-DI) images so that the outdoor environments can be optimally measured with a variable resolution. Section IV introduces the FCN-based system architecture for semantic segmentation on RGB-DI images. In Section V, the proposed algorithms are tested on different data sets and the real experiments are conducted to show that our RGB-DI image and FCN model-based approach can provide a superior performance for outdoor scene understanding. Finally, a brief conclusion and future work are given in Section VI.

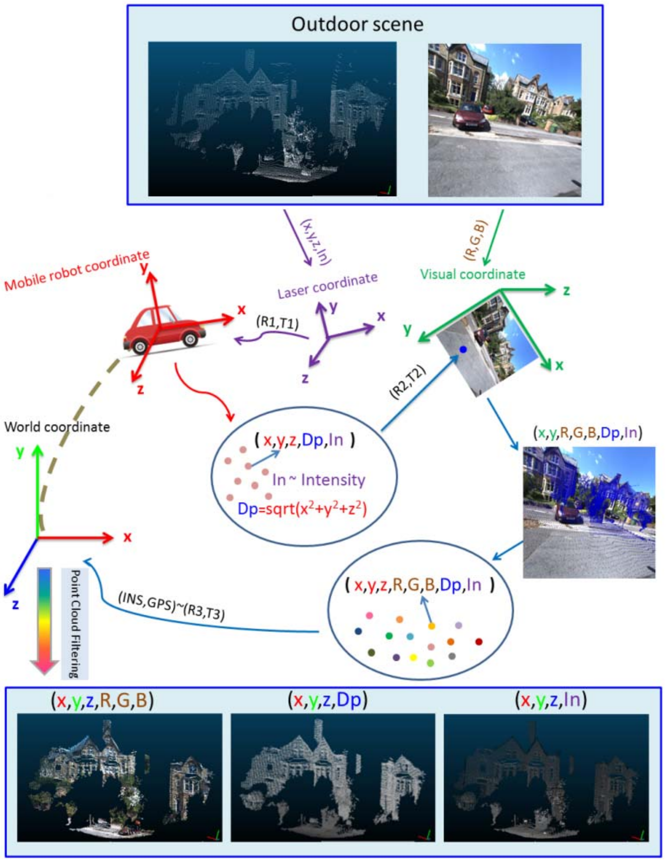


Fig. 1. Data collection and data fusion of mobile robot outdoor scene.

II. System Framework

Figs. 1–3 show the system framework proposed for outdoor scene understanding of mobile robots, which conducts data collection, data fusion, and outdoor scene understanding tasks. More specifically, the system consists of three subsystems.

1) Fig. 1 shows a multisensor data fusion subsystem that integrates data from multiple sensors for better understanding of outdoor scene.
2) Fig. 2 shows a subsystem for the RGB-DI image generation, in which an effective projection model is used to convert 3-D point clouds to 2-D images so that the semantic segmentation of 3-D point clouds is converted to the semantic segmentation of 2-D RGB-DI images.
3) Fig. 3 shows a subsystem of FCN that is applied to RGB-DI image semantic segmentation and the RGB-DI image semantic segmentation results are mapped back to the laser point cloud so as to obtain labeled point cloud.

As shown in Fig. 1, a number of sensors have been deployed for data collection and data fusion of outdoor scene, namely, laser sensors, vision sensors, INS, and GPS. To fuse the data from these sensors, a number of coordinates have been defined,

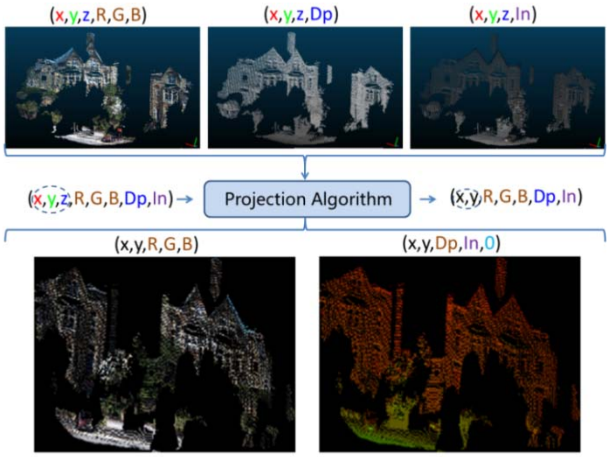


Fig. 2. RGB-DI image is generated by our projection algorithm.

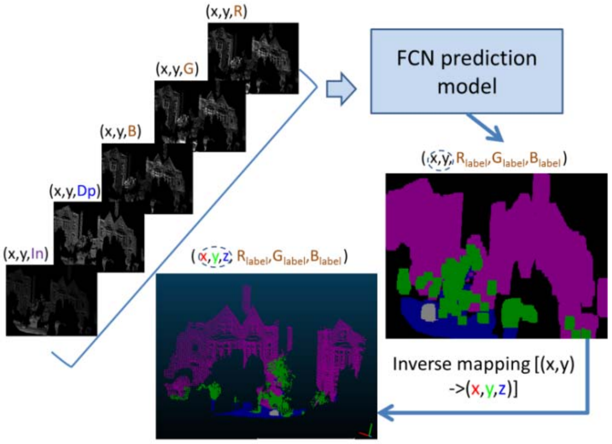

(x,y,R)
(x,y,G)
(x,y,B)
(x,y,Dp)
(x,y,In)
FCN prediction model
(x,y,R_label,G_label,B_label)
(x,y,z,R_label,G_label,B_label)
Inverse mapping [(x,y) ->(x,y,z)]

Fig. 3. Prediction process of outdoor scene understanding by RGB-DI image and FCN model.

namely, world coordinate, mobile robot coordinate, INS coordinate, laser coordinate, and vision coordinate. Considering that relative position between mobile robot, laser scanner, and vision sensor are fixed, we use the rotation and translation matrix (marked R1 and T1 in Fig. 1) to calibrate the laser coordinate system to the mobile robot coordinate system. The depth value (marked Dp in Fig. 1) is calculated in the mobile robot's coordinate system. In order to obtain the data of the point cloud with color information, we use the rotation matrix and translation matrix (marked R2 and T2 in Fig. 1), to calibrate the mobile robot coordinate system to the vision sensor coordinates and generate a (*x, y, R, G, B,* Dp, In) image (marked blue points in Fig. 1).

Based on the relationship between mobile robot coordinate system and visual coordinate system, it is able to obtain point clouds including RGB values, depth values, and intensity values [marked (*x, y, z, R, G, B,* Dp, In) in Fig. 1]. Then, the rotation matrix R3 and the translation matrix T3 for INS and GPS data fusion are used to generate the final results of RGB-DI point clouds in the world coordinate system (see the results labeled by the blue box in the bottom of Fig. 1). To reduce the noisy scanning data acquired by the laser scanner, a point cloud filtering algorithm is adopted to generate a better point cloud to represent the real-world outdoor scene.

In order to solve the problem of semantic segmentation of 3-D point cloud, a novel projection algorithm is proposed to project 3-D laser point cloud into 2-D image in our work, which will be introduced in Section III. As shown in Fig. 2, (*x, y, z, R, G, B*) represents the point cloud with color information, (*x, y, z,* Dp) represents the point cloud with the depth value, (*x, y, z,* In) represents point cloud with the intensity value. It should be noted that the projection algorithm is only to accomplish the task of coordinate transformation from 3-D point cloud to 2-D image, and the values of each pixel (RGB-DI value) in 2-D image are always associated with the original 3-D scanning point. The final result of RGB-DI image generation can be illustrated in two images [(*x, y, R, G, B*) and (*x, y,* Dp, In, 0) in the bottom of Fig. 2].

Now, we can perform outdoor scene understanding using RGB-DI image and FCN model. In practice, the resolution of the RGB-DI image is affected by the range of 3-D point cloud in different scenes. If the range of the 3-D point cloud is large, the resolution of the RGB-DI image is increased accordingly. In our experiments, according to the distribution of the actual laser scanning points, the image resolution is usually set at 180 × 240, 240 × 320, or 300 × 380. It should be noted that the system framework of FCN model can work well for semantic segmentation of outdoor scenes if the resolution of the input RGB-DI images for training and predicting of FCN model is equal to the resolution of the output labeled images.

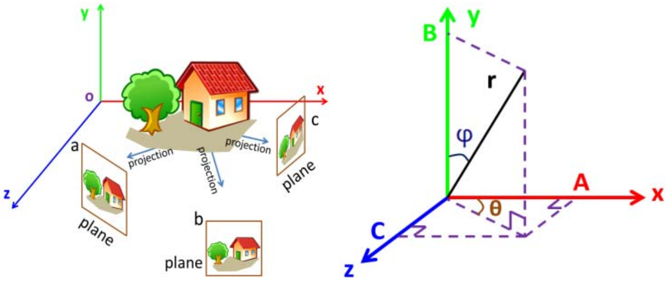
Fig. 3 illustrates the prediction process of outdoor scene understanding by RGB-DI image and FCN model. Semantic segmentation results can be obtained from these output labeled images. Finally, the results of FCN semantic segmentation are mapped to the 3-D point cloud according to the exact correspondence between each 3-D point and each pixel in an RGB-DI image, which are provided to the path planning unit of mobile robots to accomplish the task of autonomous navigation.

III. RGB-DI Image Generating

A. Projection of Point Cloud

For different objective functions of projection, different algorithms are created to perform the projection from 3-D data to 2-D data, including principal component analysis (PCA), latent dirichlet allocation (LDA), and multidimensional scaling (MDS). In order to generate an optimal RGB-DI image for semantic segmentation of an outdoor scene, the algorithm that conducts the projection from 3-D point cloud to the 2-D plane should keep contour information of the original 3-D point cloud in the 2-D image format as much as possible.

Fig. 4(a) illustrates the proposed projection algorithm. As can be seen, the same objects are projected onto different planes, but the contour of objects and the layout of the scene will be different in 2-D images. We aim to find an optimal plane to represent an outdoor scene with geometric and semantic information. Taking the three images [labeled as (a)–(c)] in Fig. 4(a) as example, (b) is a more optimal than



(a) (b)

Fig. 4. (a) Illustration of the projection of objects from 3-D space to 2-D plane. (b) Representation of the normal vector.

others since it is easier to distinguish different objects and reduce the mutual occlusion between objects.

B. Algorithm

Fig. 4(b) defines the parameter of normal vector for accurate algorithm derivation, in which φ is the angle between the normal vector and y-axis; θ is a rotating angle in xoz coordinates; and r is the norm of normal vector. The plane equation and linear equation is defined as follows:

$$A(x - x_0) + B(y - y_0) + C(z - z_0) = 0 \tag{1}$$

$$x = x_{\text{pcd}} + At, \quad y = y_{\text{pcd}} + Bt, \; z = z_{\text{pcd}} + Ct \tag{2}$$

$$A = r \sin\varphi \cos\theta, \quad B = r\cos\varphi, \; C = r\sin\varphi\sin\theta \tag{3}$$

where $r = 1$, (x_0, y_0, z_0) represents viewpoint and $(x_{\text{pcd}}, y_{\text{pcd}}, z_{\text{pcd}})$ represents point clouds.

It should be noted that t is a formal parameter of linear equation and it can be derived by (1) to (3) as follows:

$$t = \sin\varphi\cos\theta(x_0 - x_{\text{pcd}}) + \cos\varphi(y_0 - y_{\text{pcd}}) + \sin\varphi\sin\theta(z_0 - x_{\text{pcd}}). \tag{4}$$

Point (x_p, y_p, z_p) is the projection point from 3-D point cloud to 3-D plane and it can be written as (5) to (7)

$$x_p = x_{\text{pcd}} + t\sin\varphi\cos\theta \tag{5}$$

$$y_p = y_{\text{pcd}} + t\cos\varphi \tag{6}$$

$$z_p = z_{\text{pcd}} + t\sin\varphi\sin\theta \tag{7}$$

where (x_p, y_p, z_p) represents a projection point.

The center of projection point is written as in the following equation:

$$\overline{x_p} = \frac{1}{n}\sum_{i=1}^{n} x_{\text{pi}} \quad \overline{y_p} = \frac{1}{n}\sum_{i=1}^{n} y_{\text{pi}} \; \overline{z_p} = \frac{1}{n}\sum_{i=1}^{n} z_{\text{pi}} \tag{8}$$

where $(\overline{x_p}, \overline{y_p}, \overline{z_p})$ represents center of projection point and n is total number of points.

The objective function is given as follows:

$$\begin{aligned}(\varphi^*, \theta^*) &= \arg_{(\varphi,\theta)\in\Omega} \max \operatorname{dis}(\varphi,\theta)\\ \operatorname{dis}(\varphi,\theta) &= \frac{1}{n}\sum_{i=1}^{n}\Big[(x_{\text{pi}} - \overline{x_p})^2 + (y_{\text{pi}} - \overline{y_p})^2 + (z_{\text{pi}} - \overline{z_p})^2\Big]^{\frac{1}{2}}\\ \text{s.t. } \Omega &= \{(\varphi,\theta) | \varphi_0 \le \varphi \le \varphi_1, \theta_0 \le \theta \le \theta_1\}\end{aligned} \tag{9}$$

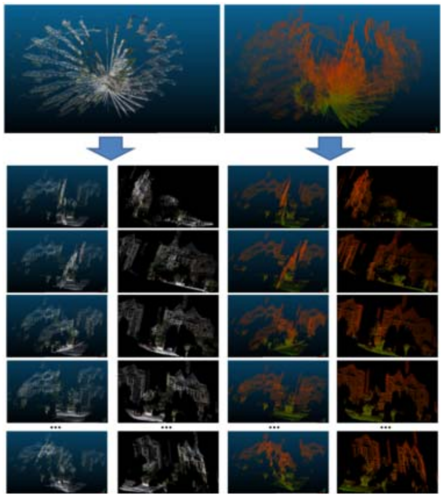


Fig. 5. When a point on the ground is selected as viewpoint, the result of RGB-DI image generating is displayed in each subfigure.

where φ_0 and φ_1 are set at $\pi/3$ and $2\pi/3$, respectively. θ_0 and θ_1 are set at 0 and π in our experiments.

As it is difficult to solve the gradient expression of the objective function, the Monte Carlo method is adopted to solve the optimal solution of the objective function [12]. It is easy to obey the uniform distribution to generate N random points (φ, θ) in a feasible domain. Finally, according to the N solutions, it is easier to find a solution (φ^*, θ^*) that maximizes the objective function, which is approximately equal to the optimal solution.

Fig. 5 shows the process of calculating the optimal solution of the objective function (9) using Monte Carlo method. It is can be seen that many projection planes determined by different parameters have been randomly given by Monte Carlo method. It searches an optimal projection plane to maximize the objective function (9) from those projection planes. In order to ensure that the projection algorithm is free from noise interference, the noise reduction of point cloud should be performed before the projection algorithm is used. A statistical-outlier-removal filter has been applied to remove the noise in point cloud and it is helpful for the semantic segmentation of RGB-DI images. The detailed description of the algorithm is given at the website URL: http://pointclouds.org/documentation/tutorials/statistical_outlier.php#statistical-outlier-removal.

The plane equation can determine a point on the plane and it is defined as the viewpoint. Since each plane has infinite extensibility and the normal vector of each point is the same, it is clear that the projected RGB-DI images are the same when the two planes have different viewpoints and the same normal vector, as shown in Fig. 6. Therefore, the changing viewpoint does not need to be considered and the normal vector obtained from the projection plane is the crucial step to generate RGB-DI images. The code of the proposed projection algorithm is given at the website of URL: https://github.com/ZhuangYanDLUT/pcl/tree/master/projection_algorithm(RGB-DI%20Image).

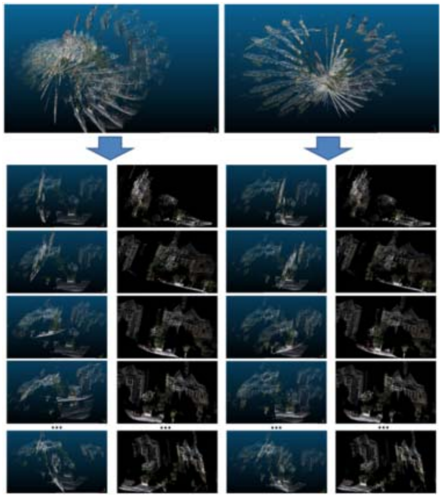


Fig. 6. Performance of RGB-DI image is displayed on condition that different viewpoints are selected.

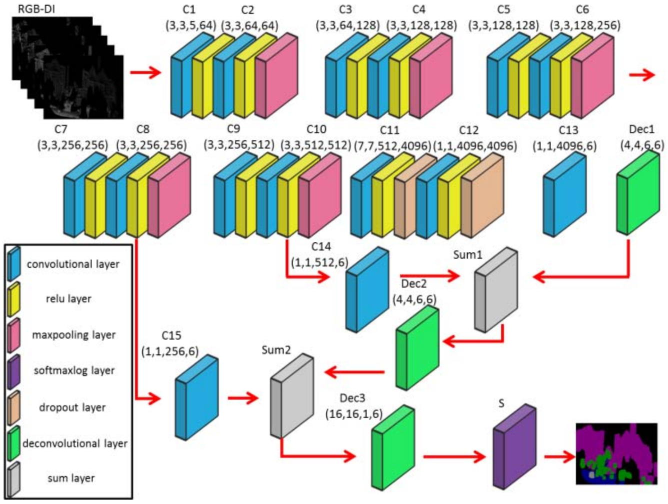


Fig. 7. Proposed FCN architecture is applied to the semantic segmentation of RGB-DI image.

IV. FCN-Based Semantic Segmentation on RGB-DI Images

Fig. 7 describes the architecture of our FCN, which is derived from the revised VGG16 and achieved the state-of-the-art performance in ImageNet [13]. It can be seen that the architecture of our FCN model is composed of convolutional layers, relu layers, pooling layers, dropout layers, deconvolutional layers, softmaxlog layers and sum layers. For the convolutional layers, the size of convolution kernel is set to 3×3, 1×1, or 7×7 and marked in Fig. 7. The relu layer represents the relu activation function. Multiple maxpooling layers are contained in the down-sampling, path, each maxpooling layer performing a 2×2 pooling operation with a stride of 2. The dropout layers prevent the overfitting of the model so as to improve the performance of the neural network, when training samples are less. Sometimes the deconvolutional layer is also called the up sampling layer, which is backwards stride convolution.

Interpolation is another way to connect coarse outputs to dense pixels. In this paper, simple bilinear interpolation is chosen to initialize the deconvolution kernel, which needs to compute each output from the nearest four inputs by a linear map. The method only relies on the relative positions of the input and output cells. If the deconvolution kernel size is set to 4×4, the output of feature maps in length and width, respectively, is 2 times that of the feature input maps in length and width, and the other cases are also marked in Fig. 7.

Up sampling is performed in-network for end-to-end learning by backpropagation from the pixel wise loss. The softmaxlog layer is the classification layer, which is the last layer of the network. During the testing process, it outputs the predicted results. During the training process, it is the starting layer for error generation. In order to improve the performance of deconvolution and obtain more accurate prediction results, the function of sum layer is the sum of the intermediate results specified in the red arrow in Fig. 7.

The VGG16 network is modified here for the semantic segmentation of image. The RBG-DI images are inputs to our network. We change the first layer of the VGG16 network from (3, 3, 3, 64) to (3, 3, 5, 64), remove some of the convolution layers and relu layers of the VGG16 network, adjusted the position of the pooling layers, and added two dropout layers, three deconvolutional layers and three convolutional layers. We defined six categories: columnar objects, plants, buildings, cars, roads, and others for the scenes understanding.

To ensure that the fusion related intermediate layer results match the segmentation requirements of scene understanding, we added three convolution kernels C13, C14, and C15, respectively, set to [1, 1, 4096, 6], [1, 1, 512, 6], and [1, 1, 256, 6]. In this paper, the performance of network is improved by fusing the output of multiple intermediate network layers and multiple up sampling layers rather than adding a deconvolutional layer at the end, which can ensure the consistency of the resolution of the input and output images of FCN model. The residual of network in this paper cannot disappear in backpropagation.

V. Experiments and Analysis

A. Experimental Data Sets and Platforms

In order to fully verify the versatility and practicability of the algorithm, we have used two data sets in this paper. The first one is the Oxford robot car *data set* that is a common public data set for unmanned vehicle research, and the second one is the DLUT data set that is collected by our mobile robot. The training and testing of related algorithms in our experiments are performed using both data sets.

Oxford robot car *data set* is public on the URL (http://robotcar-data set.robots.ox.ac.uk/). The data collection spans the period of May 6, 2014 to December 13, 2015, and consists of 1010.46 km of recorded driving in central Oxford, U.K. The vehicle was driven manually throughout the period of data collection; no autonomous capabilities were used. Traversal times were chosen to capture a wide range of conditions, including pedestrian, cyclist and vehicle traffic, light and heavy rain, direct sun, snow, and dawn, dusk and

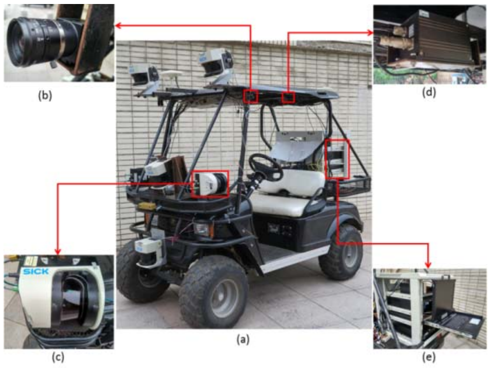


Fig. 8. Smart-Cruiser, a home-developed mobile robot equipped with multiple lasers and monocular camera, etc.

night. The total uncompressed size of the data collected is 23.15 TB. Labels for each condition have been added to each traversal, and traversals can be grouped by labels for easy collection of a particular condition.

DLUT data set built by our laboratory is public on the URL (http://pan.baidu.com/s/1o85fc4E). The data set was collected by our mobile robot named Smart-Cruiser at Dalian University of Technology Campus. The data set includes buildings, trees, vehicles, pedestrians, grasslands, poles, and others. The data collection was conducted within the period of August 26, 2016 to August 28, 2017. *Data set* has color point clouds, depth point clouds, intensity point clouds, and mobile robot's pose, and its size is 9.18 GB.

Fig. 8 shows our mobile robot platform used for implementing the algorithms. It was developed in house, and equipped with multiple lasers, industrial PCs, GPS, monocular camera, and inertial navigation system. The sensors used in this experiment are as follows.

1) 1 × CCD Fly Capture Flea3, 3.2 million pixels (2080 × 1552), and a frame rate of 60 frames/s as shown in Fig. 8(b).
2) 1 × SICK LMS-151 2-D LIDAR, 270○FoV, 50 Hz, 80 m range, 0.5○resolution, as shown in Fig. 8(c).
3) 1 × XW-ADU5600 ALIGN inertial and GPS navigation system, 6 axes, 50 Hz, GPS/GLONASS, dual antenna, as shown in Fig. 8(d).
4) 2 × Advantech IPCs with InterCorei7 CPU, 24 GRAM, GTX970 graphics card, DDR3 SSD, and one D-Link DKVM-L708H Fig. 8(e).

B. Results of RGB-DI Image Generating

There are many projection methods that could 3-D point clouds onto 2-D planes. In this paper, we compare our projection algorithm with the other existing projection methods in terms of image semantic segmentation for outdoor scene understanding. Five commonly used projection algorithms are used here, namely, PCA, Kernel PCA (KPCA) [14], MDS [15], LDA [16], and autoencoder [17]. The first three are linear methods, and the latter two are nonlinear methods. In order to observe the projection results of each algorithm more intuitively, a representative example of the experiment has been given and shown in Figs. 9 and 10.

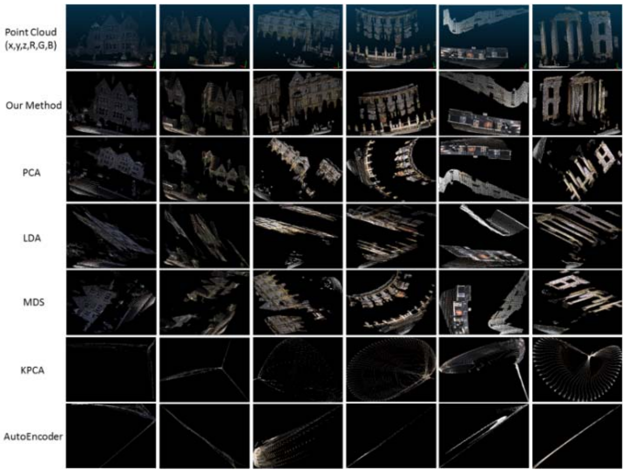


Fig. 9. Results of different projection algorithms using the same group of point cloud (*x, y, z, R, G, B*).

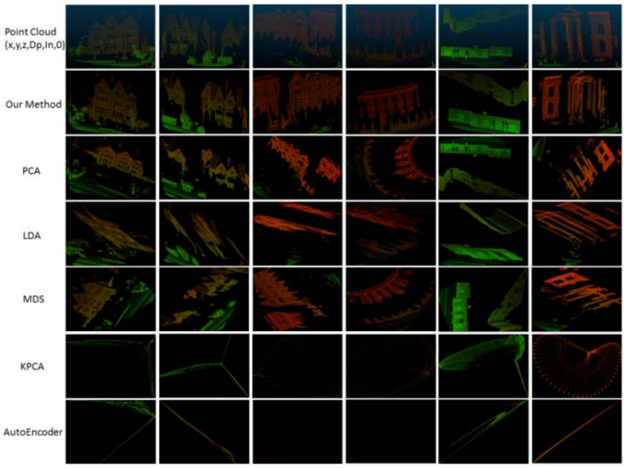


Fig. 10. Results of different projection algorithms using the same group of point cloud (*x, y, z,* Dp, In, 0).

As can be seen from Figs. 9 and 10, the nonlinear projection method loses the semantic information of the point cloud and the morphological information of the scene. Therefore, the nonlinear projection method KPCA and autoencoder do not contribute effectively to the research in this paper. Instead, the linear projection method is superior to the nonlinear projection method in terms of results. Although PCA, LDA and MDS maintain the morphological and semantic information of the original point cloud, the context of the projected object is greatly changed. The LDA algorithm maps a number of 3-D points to 2-D points, resulting in greater loss of information. Compared with MDS and PCA, the loss of information is insignificant, but there is still a great change in the location of the objects and different degrees of rotation.

In order to further analyze the effect of different projection methods on the semantic segmentation of RGB-DI images, the two-order CRF method, which takes into account the pixel position relationship [18], is applied to the semantic segmentation of RGB-DI images generated by PCA, LDA, MDS, and our method. In experiments, the RGB-DI images

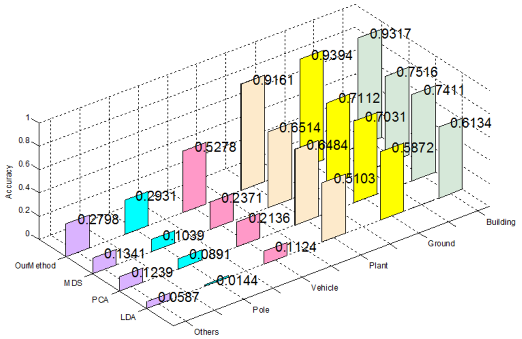


Fig. 11. Accuracy of semantic segmentation with different algorithms (our method, MDS, PCA, and LDA) in Oxford data set. There are totally six classes in the semantic segmentation, which are building, ground, plant, vehicle, pole, and others.

generated by different projection algorithms have a resolution of 240 × 320. Color histogram features are extracted from the neighborhood of a pixel in RGB-DI images.

The classification results of RGB-DI images generated by different algorithms are mapped to point clouds, and the histogram statistics is done according to the classification accuracy of point cloud, as shown in Fig. 11. The accuracy in Figs. 11 and 13–15 is based on ground truth of RGB-DI point cloud. The definition of accuracy is as follows:

$$\text{accuracy} = \frac{TP}{TP + FP} \tag{10}$$

where true positives is the number of samples that are actually positive samples and divided into positive samples by the classifier, and false positives is the number of samples that are actually negative samples and are divided into positives samples by the classifier.

As shown in Fig. 11, compared with different projection methods, our method can improve the recognition accuracy of objects in the scene, but no matter what projection methods do not significantly improve the accuracy of objects that are not obvious such as poles or others.

It is clear the results of our method are substantially better than those of PCA, LDA, and MDS. The results of MDS and PCA are superior to LDA. We can draw a conclusion that the shape of the original point cloud is consistent with one of projection RGB-DI images, and the rotation range of the projected objects is reduced. The proposed algorithm takes into account the constraints of the correlation angle and the maximization of object projection, which ensures that the generated RGB-DI images have better semantic segmentation results so as to obtain better classification results of point cloud.

C. Semantic Segmentation With Different Input Values

The purpose of this experiment is to analyze the influence of different input values of point cloud on the semantic segmentation results. It should be noted that the classifier and the feature extraction algorithm are the same in all of the experiments of this section. According to different input values of point clouds, five groups of input data in our experiments are (x, y, z), (x, y, z, R, G, B), $(x, y, z, R, G, B, \text{Dp})$, $(x, y, z, R, G, B, \text{In})$, and $(x, y, z, R, G, B, \text{Dp}, \text{In})$. The feature extraction of cloud points is composed of two parts. In the first part, Fast Point Feature Histograms (FPFH) is applied to compute the features of point clouds based on (x, y, z) [19]. In the second part, color histogram feature of neighborhood of a point in point cloud is used. The specific process of feature extraction is shown in Fig. 12.

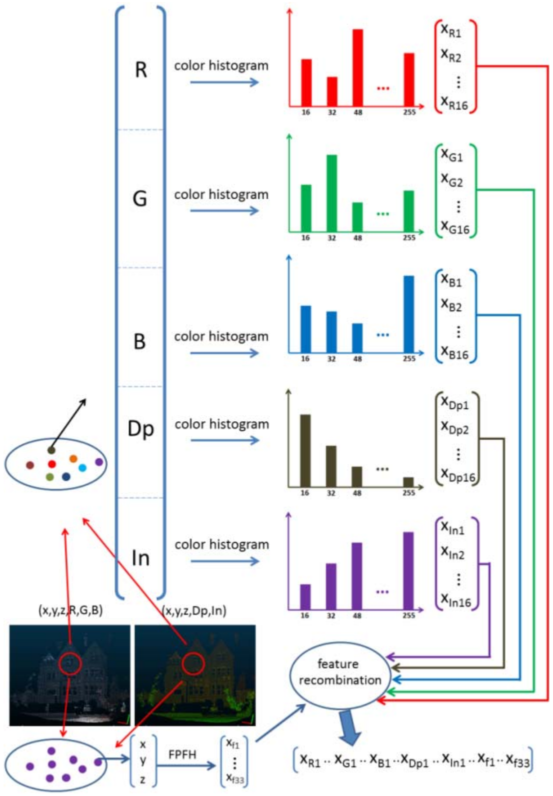


Fig. 12. Process of feature extraction on point cloud $(x, y, z, R, G, B, \text{Dp}, \text{In})$.

In the experiment, the feature dimension of the FPFH algorithm is 33 and the feature dimension of color histogram feature is 16. As color histogram statistics ranges from 0 to 255, histogram features are intercepted in units of 16 in order to eliminate redundancy. The features of five groups of input data (x, y, z), (x, y, z, R, G, B), $(x, y, z, R, G, B, \text{Dp})$, $(x, y, z, R, G, B, \text{In})$ and $(x, y, z, R, G, B, \text{Dp}, \text{In})$, are $(Xf1, \ldots, Xf33)$, $(Xf1, \ldots, Xf33, XR1, \ldots, XR16, XG1, \ldots, XG16, XB1, \ldots, XB16)$, $(Xf1, \ldots, Xf33, XR1, \ldots, XR16, XG1, \ldots, XG16, XB1, \ldots, XB16, XDp1, \ldots, XDp16)$, $(Xf1, \ldots, Xf33, XR1, \ldots, XR16, XG1, \ldots, XG16, XB1, \ldots, XB16, XIn1, \ldots, XIn16)$, $(Xf1, \ldots, Xf33, XR1, \ldots, XR16, XG1, \ldots, XG16, XB1, \ldots, XB16, XDp1, \ldots, XDp16, XIn1, \ldots, XIn16)$, respectively. Random Forest classifier is applied to semantic segmentation of point clouds since it is directly applicable to multiclass problems and yield good results in reasonable time on large point clouds [20].

The idea of Random Forest classifier is repeatedly and randomly to extract K samples from the original N training samples and then to generate K decision trees based on the K samples to form Random Forests. The classification results of the test data are based on the vote of multiple decision trees.

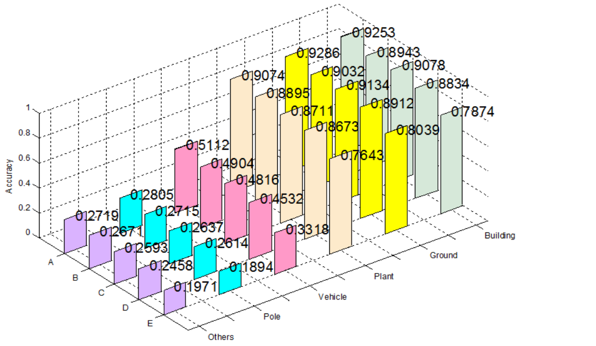


Fig. 13. Accuracy of semantic segmentation with different input values in Oxford data set [Row A: (*x, y, z, R, G, B,* Dp, In), Row B: (*x, y, z, R, G, B,* In), Row C: (*x, y, z, R, G, B,* Dp), Row D: *(x, y, z, R, G, B)*, and Row E: *(x, y, z)*]. There are totally six classes in the semantic segmentation, which are building, ground, plant, vehicle, pole, and others.

The classification ability of a single decision tree may be very limited. However, after a large number of decision trees are generated randomly, a test sample can be classified according to the statistical results of each tree, so as to get the most probable classification results.

In these five groups, the accuracy of semantic segmentation with different input values is counted in Fig. 13. Considering that the extracted features are based on data collected by sensors of mobile robot which are independent, the extracted features are also independent so as to avoid curse of dimensionality. As shown in Fig. 13, in order to analyze the accuracy of the overall recognition, the results of the sum of accuracy are 3.8249, 3.716, 3.6969, 3.6023, and 3.0739 for five input values A, B, C, D, and E in Oxford data set, respectively.

With the increase of the input values of point clouds, the accuracy of semantic segmentation is also improved. Therefore, we draw a conclusion that the performance of outdoor scene understanding can be improved by making full use of multisensor data fusion of mobile robots.

D. Semantic Segmentation Results With Different Algorithms

The task of outdoor semantic segmentation is often guided by two kinds of research ideas. The first one extracts features and designs classifier directly from point clouds. The second one uses RGB-DI images to perform sematic segmentation, and return the segmentation results to 3-D point clouds. In our experiments, the data point clouds (*x, y, z, R, G, B,* Dp, In) are used to test the performance of different algorithms. There are five kinds of semantic segmentation algorithms used in this experiment.

For the first idea, FPFH, and color histograms are used to extract features and specific process is shown in Fig. 12. The random forest classifier [21] and the CRF classifier [18] are selected. For the second idea, the RGB-DI image is generated by the proposed method and color histogram features are extracted from the neighborhood of a pixel in RGB-DI images. The random forest classifier, the CRF classifier and the proposed FCN model are applied to the semantic segmentation of RGB-DI images and classification results are mapped onto the point cloud. The detailed description of the five algorithms is shown in Table I.

TABLE I

FIVE ALGORITHMS COMPARED IN OUR EXPERIMENTS

Algorithm	Data source	Feature Extraction	Classifier
A1	RGB-DI images	color histogram	random forest
A2	RGB-DI images	color histogram	CRF
A3 (ours)	RGB-DI images	FCN	FCN
B1	RGB-DI point cloud	color histogram + FPFH	random forest
B2	RGB-DI point cloud	color histogram + FPFH	CRF

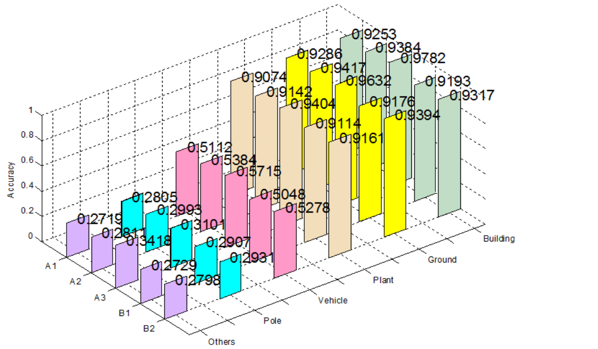


Fig. 14. Accuracy of semantic segmentation with different algorithms on Oxford robot car data set.

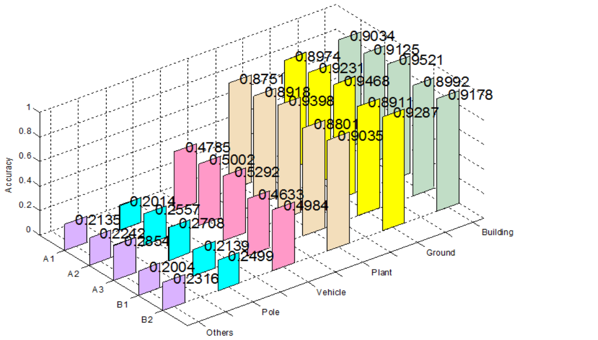


Fig. 15. Accuracy of semantic segmentation with different algorithms on DLUT data set.

Both A1 and A2 are based on the first idea, and the feature extraction methods are the same and the classifiers are different. The random forest classifier is used by A1 and the CRF classifier is used by A2. B1, B2 and A3 are based on the second idea, the feature extraction methods of B1 and B2 are the same, and the classifiers are different. The random forest classifier is used by B1 and the CRF classifier is used by B2. A3 is based on the proposed FCN model. Fig. 14 shows the accuracy of semantic segmentation with different algorithms on the Oxford data set. Fig. 15 shows the accuracy of semantic segmentation with different algorithms on the DLUT data set.

The random forest algorithm achieved similar accuracy on both point clouds and RGB-DI images, respectively, and

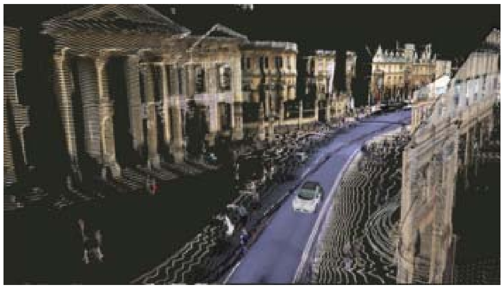


Fig. 16. Point cloud processed after multisensor data fusion in Oxford data set.

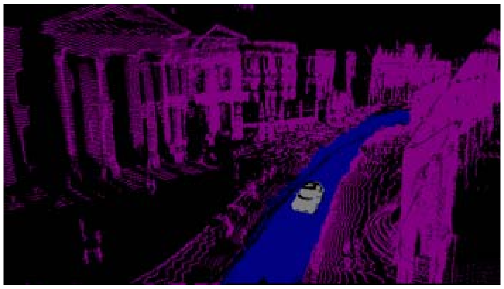


Fig. 17. Performance of semantic segmentation using FCN model in Oxford data set.

the same is CRF. However, the accuracy of CRF is much better than the one of the random forest algorithm for both point clouds and RGB-DI images. The reason is that CRF took the classification effects of adjacent points into account. The classification effects of the proposed FCN model are superior to the other four methods. The experimental results show that the FCN model can extract much valuable features when the sensor information is fully used Figs. 16 and 17 show the performance of the FCN model.

E. Generalization Analysis of FCN Model

We obtained the FCN model by modifying the VGG16 model whose parameters are used to initialize the parameters of the training FCN model. Hence, a relatively good initial value of the neural network is obtained at the start of training so that the stochastic gradient descent method [22] is used to obtain the optimal parameters in the training network. The cross validation process has been used in the experiment. In the experiment, our RGB-DI images have been divided into three parts to be defined as train set, validation set, and test set. The train set contains 500 RGB-DI images. The validation set only contains 100 RGB-DI images which are not mixed by train set. In the same way, test only contains 100 RGB-DI images which are not mixed by train set and validation set. The loss results of train set have been shown in Fig. 18. The results of training FCN model in test set have been shown in Fig. 19. The results of training FCN model in train set have been shown in Fig. 20. The results of training FCN model in validation set have been shown Fig. 21.

Finally, in order to verify the accuracy of the trained model in (epoch = 350), 100 RGB-DI images, which are not intersecting with train set, test set, and validation set, are generated by random sampling algorithm. These 100 | RGB-DI images were predicted by a trained model. The prediction accuracy of the model is shown in Fig. 22. In experiment, epoch is 400, and batch size is 20, and learning rate is 0.5. In order to optimize the model, the learning rate is adjusted dynamically according to the accuracy of the real-time validation set. The accuracy is based on ground truth of RGB-DI images. The accuracy in FCN model is the average accuracy of multiple object recognition in a RGB-DI image. When the accuracy is stable in multiple epochs, the learning rate is reduced by 0.8 times so as to make loss continue to decline. The tensor flow framework is applied to training FCN model.

As shown in Fig. 22, the FCN model obtained by modifying the VGG16 is trained on the train set and the optimal parameters can be obtained from validation set. In Figs. 18 and 21, when epoch is greater than 350, the results of test and validation are approximately the same. It can be seen from the results of cross validation that the FCN model obtained by modifying the VGG16 has a better robustness.

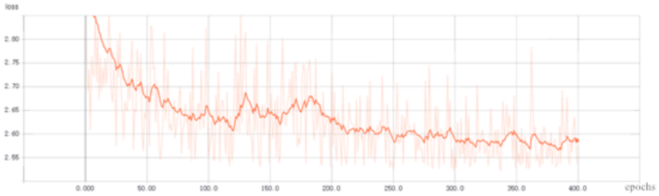


Fig. 18. Performance of training FCN model in loss function in train set.

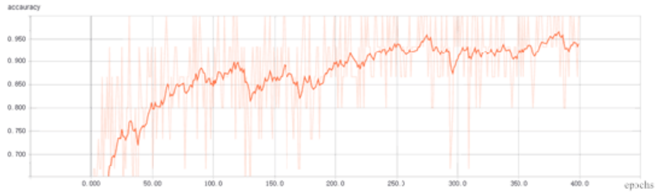


Fig. 19. Performance of training FCN model in test set.

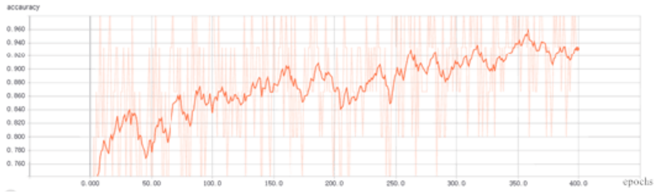


Fig. 20. Performance of training FCN model in train set.

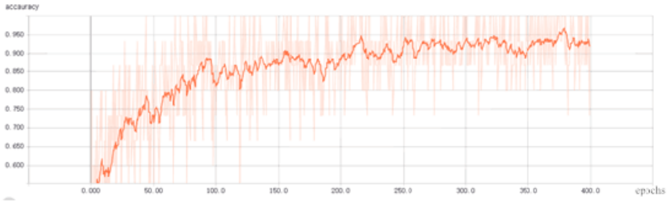


Fig. 21. Performance of training FCN model in validation set.

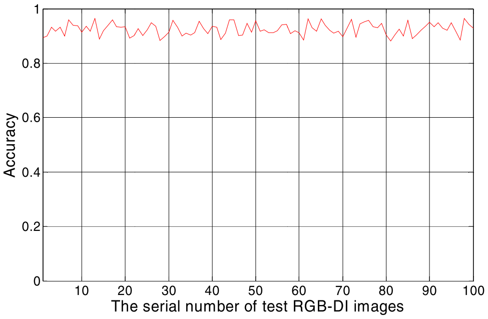


Fig. 22. Accuracy for each image on a random sampled test set.

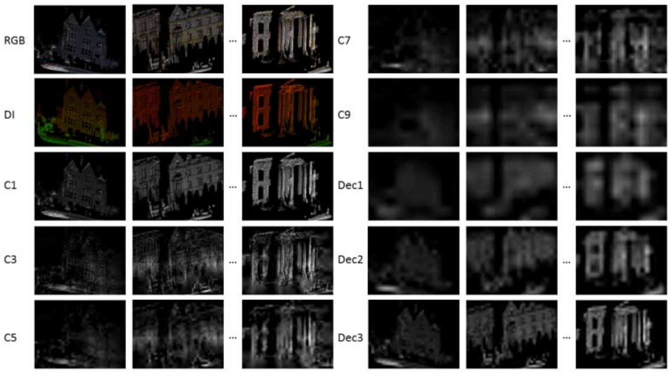


Fig. 23. Output of some convolutional layers and deconvolutional layers in testing process.

The proposed FCN model contains three deconvolutional layers, each of which is stacked with the information of the middle layer. In order to observe the performance of each layer more visually, the output of some convolutional layers and deconvolutional layers are shown in Fig. 23. It can be seen clearly that the convolution and deconvolution of the FCN model realize the encoding and decoding of the input RGB-DI image. The process is able to extract the features, which are more conducive to classification. Therefore, the combination of RGB-DI images and the FCN model is more suitable for solving outdoor scene understanding.

VI. Conclusion

This paper proposes a novel sensor fusion method to solve the problem of outdoor scene understanding for mobile robots. In order to make full use of the information collected by multiple sensors, the image and laser are fused to generate point cloud with color, depth, and intensity value. In order to solve the semantic segmentation of point cloud, a projection algorithm is designed to generate multichannel RGB-DI images. Finally, the FCN model is used to segment the RGB-DI images and the 2-D results based on RGB-DI images are mapped to 3-D point cloud.

Comparing the projection algorithm proposed in this paper with a variety of traditional projection algorithms, we have come to the conclusion that our projection algorithm generates RGB-DI images more suitable for solving outdoor scene understanding of mobile robots. Comparing the FCN model proposed in this paper and the traditional semantic segmentation of point clouds, the experiment shows that the FCN model can make full use of the fused data generated by multiple sensors that are onboard of a mobile robot, so as to obtain better feature representation and classification results. For Oxford and DLUT data sets, the accuracy of trees, buildings and ground have been improved 3% to 5%, and the accuracy of vehicles, poles and others have been improved 4% to 6%. Our future work will be focused on reducing the layer number of FCN model to improve the training and predicting efficiency of FCN model without reducing the accuracy of semantic segmentation.

References

[1] Z. Hu, Y. Li, N. Li, and B. Zhao, "Extrinsic calibration of 2-D laser rangefinder and camera from single shot based on minimal solution," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 4, pp. 915–925, Apr. 2016.

[2] L. Wei, C. Cappelle, and Y. Ruichek, "Camera/laser/GPS fusion method for vehicle positioning under extended NIS-based sensor validation," *IEEE Trans. Instrum. Meas.*, vol. 62, no. 11, pp. 3110–3122, Nov. 2013.

[3] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. Wan, "An integrated framework for 3-D modeling, object detection, and pose estimation from point-clouds," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 3, pp. 683–693, Mar. 2015.

[4] 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," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[5] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[6] U. Asif, M. Bennamoun, and F. A. Sohel, "RGB-D object recognition and grasp detection using hierarchical cascaded forests," *IEEE Trans. Robot.*, vol. 33, no. 3, pp. 547–564, Jun. 2017.

[7] D. Zermas, I. Izzat, and N. Papanikolopoulos, "Fast segmentation of 3D point clouds: A paradigm on LIDAR data for autonomous vehicle applications," in *Proc. IEEE ICRA*, Singapore, May/Jun. 2017, pp. 5067–5073.

[8] F. Husain, B. Dellen, and C. Torras, "Consistent depth video segmentation using adaptive surface models," *IEEE Trans. Cybern.*, vol. 45, no. 2, pp. 266–278, Feb. 2015.

[9] A. Maligo and S. Lacroix, "Classification of outdoor 3D LIDAR data based on unsupervised Gaussian mixture models," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 1, pp. 5–16, Jan. 2017.

[10] R. Gadde, V. Jampani, R. Marlet, and P. V. Gehler, "Efficient 2D and 3D facade segmentation using auto-context," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1273–1280, May 2018.

[11] G. Leifman, E. Shtrom, and A. Tal, "Surface regions of interest for viewpoint selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2544–2556, Dec. 2016.

[12] Y. Altmann, N. Dobigeon, and J.-Y. Tourneret, "Unsupervised post-nonlinear unmixing of hyperspectral images using a Hamiltonian Monte Carlo algorithm," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2663–2675, Jun. 2014.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[14] T.-J. Chin and D. Suter, "Incremental kernel principal component analysis," *IEEE Trans. Image Process.*, vol. 16, no. 6, pp. 1662–1674, Jun. 2007.

[15] K. Rajawat and S. Kumar, "Stochastic multidimensional scaling," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 3, no. 2, pp. 360–375, Jun. 2017.

[16] Y. Pang, S. Wang, and Y. Yuan, "Learning regularized LDA by clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2191–2201, Dec. 2014.

[17] K. Zeng, J. Yu, R. Wang, C. Li, and D. Tao, "Coupled deep autoencoder for single image super-resolution," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 27–37, Jan. 2017.

[18] F. Liu, G. Lin, R. Qiao, and C. Shen, "Structured learning of tree potentials in CRF for image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2631–2637, Jun. 2018, doi: 10.1109/TNNLS.2017.2690453.

[19] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *Proc. IEEE ICRA*, Kobe, Japan, May 2009, pp. 3212–3217.

[20] T. Wang, J. Li, and X. An, "An efficient scene semantic labeling approach for 3D point cloud," in *Proc. IEEE ITSC*, Las Palmas, Spain, Sep. 2015, pp. 2115–2120.

[21] Z. Li, L. Zhang, R. Zhong, T. Fang, L. Zhang, and Z. Zhang, "Classification of urban point clouds: A robust supervised approach with automatically generating training data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 3, pp. 1207–1220, Mar. 2017.

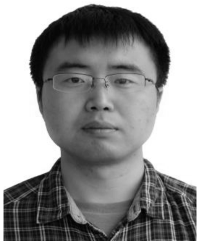
[22] X.-L. Li, "Preconditioned stochastic gradient descent," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1454–1466, May 2018.

Zengshuai Qiu received the bachelor's degree in automation from the City Institute, Dalian University of Technology, Dalian, China, in 2010, and the master's degree in control theory and engineering from the Shenyang University of Technology, Shenyang, China, in 2010 and 2013, respectively. He is currently pursuing the Ph.D. degree with the School of Control Science and Engineering, Dalian University of Technology.

His current research interests include robotics, deep learning, image processing, and semantic scene understanding.

Yan Zhuang (M'11) received the bachelor's and master's degrees in control theory and engineering from Northeastern University, Shenyang, China, in 1997 and 2000, respectively, and the Ph.D. degree in control theory and engineering from the Dalian University of Technology, Dalian, China, in 2004.

In 2005, he joined the Dalian University of Technology, as a Lecturer and became an Associate Professor in 2007, where he is currently a Professor with the School of Control Science and Engineering. His current research interests include mobile robot 3-D mapping, outdoor scene understanding, 3-D-laser-based object recognition, and 3-Dscene recognition and reconstruction.



Fei Yan (M'15) received the bachelor's and Ph.D. degrees in control theory and engineering from the Dalian University of Technology, Dalian, China, in 2004 and 2011, respectively.

In 2013, he joined the Dalian University of Technology, as a Post-Doctoral and became a Lecturer in 2015, where he is currently an Associate Professor with the School of Control Science and Engineering. His current research interests include 3-D mapping, path planning, and semantic scene understanding.

Huosheng Hu (M'94–SM'01) received the M.Sc. degree in industrial automation from Central South University, Changsha, China, in 1982, and the Ph.D. degree in robotics from the University of Oxford, Oxford, U.K., in 1993.

He is currently a Professor with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, U.K., leading the Robotics Research Group. He has authored around 450 papers in journals, books, and conferences in these areas, and received a number of best paper awards. His current research interests include behaviour-based robotics, human-robot interaction, service robots, embedded systems, data fusion, learning algorithms, mechatronics, and pervasive computing.

Dr. Hu is a Founding Member of the IEEE Robotics and Automation Society Technical committee on Networked Robots, a fellow of IET and InstMC, and a Senior Member of ACM. He has been a Program Chair and a member of the Advisory/Organising Committee for many international conferences such as IEEE ICRA, IROS, ICMA, ROBIO, ICIA, ICAL, and IASTED RA, CA, and CI conferences. He currently serves as the Editor-in-Chief of the *International Journal of Automation and Computing* and the *Online Robotics Journal*, and the Executive Editor of the*International Journal of Mechatronics and Automation*.

Wei Wang (SM'01) received the bachelor's, master's, and Ph.D. degrees in industrial automation from Northeastern University, Shenyang, China, in 1982, 1986, and 1988, respectively.

He was a Post-Doctoral Fellow with the Division of Engineering Cybernetics, Norwegian Science and Technology University, Trondheim, Norway, from 1990 to 1992, and a Research Fellow at the Department of Engineering Science, University of Oxford, Oxford, U.K., from 1998 to 1999. He is currently a Professor with the School of Control Science and Engineering, Dalian University of Technology, Dalian, China. He has authored over 200 papers in international and domestic journals. His current research interests include adaptive control, predictive control, robotics, computer integrated manufacturing systems, and computer control of industrial process.

Dr. Wang was a member of the IFAC Technical Committee of Mining, Mineral and Metal Processing in 1999, the Chair of the IFAC Technical Committee on Cost Oriented Automation from 2005 to 2008, the Steering Commission Member of the Asian Control Association in 2011. He was a recipient of the National Distinguished Young Fund of the National Natural Science Foundations of China in 1998.