Object Classification Using CNN-Based Fusion of Vision and LIDAR in Autonomous Vehicle Environment

Hongbo Gao, Bo Cheng, Jianqiang Wang, Keqiang Li, Jianhui Zhao, and Deyi Li

***Abstract*—This paper presents an object classification method for vision and light detection and ranging (LIDAR) fusion of autonomous vehicles in the environment. This method is based on convolutional neural network (CNN) and image upsampling theory. By creating a point cloud of LIDAR data upsampling and converting into pixel-level depth information, depth information is connected with Red Green Blue data and fed into a deep CNN. The proposed method can obtain informative feature representation for object classification in autonomous vehicle environment using the integrated vision and LIDAR data. This method is also adopted to guarantee both object classification accuracy and minimal loss. Experimental results are presented and show the effectiveness and efficiency of object classification strategies.**

***Index Terms*—Autonomous vehicle, convolutional neural network (CNN), object classification, sensor fusion.**

I. Introduction

IN THE past decades, as one of the most fascinating technology trends in automotive industry, autonomous vehicles have received increasingly significant attention due to their significant potential in enhancing vehicle safety and performance, traffic efficiency [1], and energy saving [2]. Research topics over automotive industry have already received substantial attentions from both academia and industry; some notable programs include Dickmanns and VaMP [3], ARGO project, EUREKA PROMETHEUS project [4], DARPA Grand Challenge [5], Google's autonomous vehicle [6], the annual "Intelligent Vehicle Future Challenge" organized by National Natural Science Foundation of China since 2009 [7]. Hundreds of teams from all over the world participate to compete and demonstrate technological achievements on autonomous vehicles, and to maximize car-following fuel economy and fulfill requirements of intervehicle safety. Especially, Hu *et al.* proposed an optimal look-ahead control method that is based on a model predictive fuel-optimal controller, which uses state trajectories of the leading vehicle from V2V/V2I communication [2]. Autonomous vehicles should be instantaneous, accurate, stable, and efficient in computations to produce safe and acceptable traveling trajectories in numerous urban to suburb scenarios and from high-density traffic flow to high-speed highways. In real-world traffic, various uncertainties and complexities surround road and weather conditions, whereas a dynamic interaction exists between objects and obstacles, and tires and driving terrains. An autonomous vehicle must rapidly and accurately detect, recognize, and classify and track dynamic objects with complex backgrounds and posing technical challenges.

At present, research on object classification for autonomous vehicle can be divided according to two research methods. The first research type is based on Red Green Blue (RGB-D) application. Imran *et al.* first combined an RGB image with a depth image collected with Kinect and trained a convolution network with four-channel data flow [8], [9]. Silberman and Fergus showed an average accuracy of 64.5% for indoor semantic segmentation, and video testing yielded satisfactory reliability and accuracy [10]. Gupta *et al.* proposed a heterogeneous neural network by combining two convolutional neural networks (CNNs) and a support vector machine model for RGB-D-based object detection and segmentation [11]. Eitel *et al.* built a multimodel deep learning architecture to process RGB-D images for object recognition [12]. Wang *et al.* proposed a multimodel based on deep learning [13]. Kosaka *et al.* proposed a method for detecting vehicles from a nighttime driving scene taken by an in-vehicle monocular camera [14]. Cheon *et al.* proposed a vision-based vehicle detection system using symmetry vectors of histograms of gradient orientation; this system consists of a hypothesis generation step and a hypothesis verification step [15]. Chavez-Garcia and Aycard proposed a perceived model of

Manuscript received January 15, 2018; revised March 19, 2018; accepted March 20, 2018. Date of publication April 4, 2018; date of current version September 4, 2018. This work was supported in part by the National Key Research and Development Plan of China under Grant 2017YFB0102603, in part by the Junior Fellowships for Advanced Innovation Think-Tank Program of China Association for Science and Technology under Grant DXB-ZKQN-2017-035, in part by the Project Funded by China Postdoctoral Science Foundation under Grant 2017M620765, and in part by the National Natural Science Foundation of China under Grant 51625503. Paper no. TII-18-0121. *(Corresponding author: Jianqiang Wang.)*

H. Gao, B. Cheng, J. Wang, and K. Li are with the State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing 100084, China (e-mail: ghb48@mail.tsinghua.edu.cn; chengbo@tsinghua.edu.cn; wjqlws@tsinghua.edu.cn; likq@tsinghua.edu.cn).

J. Zhao and D. Li are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: zhaojh14@mails.tsinghua.edu.cn; leedeyi@tsinghua.edu.cn).

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/TII.2018.2822828

the environment to classify four objects of interest: pedestrian, bike, car, and truck [16].

The second research type is based on fusion of vision and light detection and ranging (LIDAR) data application. Navarro-Serment *et al.* used several SICK laser line scanners to build a LIDAR array for pedestrian detection and tracking in indoor environment [17]. Dolson *et al.* designed an accelerated Gaussian interpolation algorithm to upsample camera data and laser scan in real time using high-resolution depth images in computer vision application [18]. Premebida *et al.* analyzed the sparsity of three-dimensional (3-D) laser range sensors (i.e., the Veloyne HDL-64E LIDAR) to compare RGB data and Kinect [19]. Schlosser *et al.* explored several aspects in LIDAR and RGB image fusion for CNNs for pedestrian detection [20]. Zarzoso proposed a convolutional learning system for classification of segmented objects represented in 3-D as point clouds of laser reflection [21]. Wu *et al.* presented a super voxel-based approach for automated localization and extraction of street light poles in point clouds acquired by a mobile LIDAR system [22].

The traditional computer vision of an optical camera can be easily modified to satisfy the on-board requirement to detect vehicles, pedestrians, or traffic signals. The images obtained contain rich semantic information, but graphic computation consumes computer resources. Additionally, optical cameras are sensitive to illumination and lighting angle. The lack of depth information from RGB data is the most serious problem that occurs when detecting overlapping objects on roads [20]. Acoustic radar and LIDAR can offer accurate distance information in short slots, but none of the available active LIDARS can acquire high-density environment information. Veloyne HDL-64E, which is the most frequently used LIDAR sensor, can only produce sparser point clouds [19]. Although these RGB-D-based methods perform well, the hardware used in them are designed for near-distance scenarios, such as indoor environment. Equipment, such as Kinect, cannot process long-distance object detection outdoors [18]. However, the insufficient depth information has resulted in bottlenecks in accuracy, efficiency, and timeliness of detection, recognition, tracking, and segmentation techniques based on traditional RGB images [11], [23], [24]. To solve the shortcomings of the above research problems, this paper proposes sensor fusion of camera RGB data and LIDAR point clouds. RGB-D data that are suitable for long-distance object detection in outdoor environment can be obtained by combining pure RGB data with depth information from long-distance sensible LIDAR point cloud. The two key technologies of object recognition classification based on multisensor fusion include data fusion and classification methods. The data fusion method based on upsampling is simple and efficient and suits practical applications. Deep CNN (DCNN) can achieve remarkable results on a highly challenging image dataset using purely supervised learning. Hence, if we combine DCNN and upsampling fusion method for object classification, the merits of both methods can be inherited. On the other hand, shortages of each method can be avoided. The scientific contributions of this paper are briefly described as follows.

1) A simple but powerful DCNN and upsampling fusion method is proposed to handle object classification problems in autonomous vehicle environment efficiently.

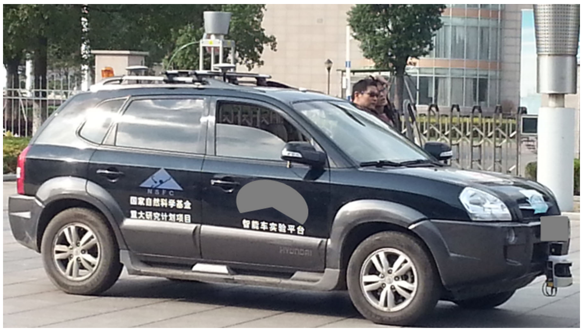


Fig. 1. Mengshi autonomous vehicle.

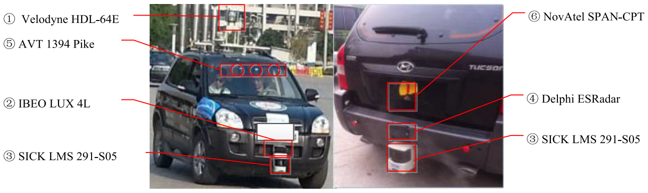


Fig. 2. Sensor deployment of Mengshi.

2) Using the integrated vision and LIDAR data, the proposed method can obtain informative feature representation for object classification in autonomous vehicle environment. By making the point cloud of LIDAR data upsampling and converting into pixel-level depth information, depth information is connected with RGB data and fed into a DCNN.

3) A DCNN and upsampling fusion method is adopted to guarantee both object classification accuracy and low loss to improve processing efficiency.

The rest of this paper is organized as follows. Section II describes the system architecture of Mengshi autonomous vehicle. Section III presents the object classification approach, including sparse data upsampling range, dataset description, and object classification method. Section IV presents the experimental results and result analysis. Section V concludes the paper.

II. System Architecture of Mengshi Autonomous Vehicle

The autonomous vehicle, Mengshi, was cooperatively designed and developed by Tsinghua University under the leadership of Prof. D. Li and Prof. K. Li. Fig. 1 shows the outlook of Mengshi.

Fig. 2 illustrates sensor deployment of Mengshi, which consists of five radar sensors, three vision sensors, and one integrated position/attitude sensor. The radar sensors include two single laser radar (SICK LMS 291-S05), one four-line laser radar (IBEO LUX 4L), one 64-line laser sensor (Velodyne HDL 64E), and one mm-wave radar (Delphi ESR). The vision sensors include three cameras (AVT 1394 Pike F-100c) that are evenly equipped on the back of the frontal mirror. The integrated position/attitude sensor includes global positioning system (GPS) and inertial navigation system, which originated from NovAtel. Table I provides detailed descriptions of each sensor.

TABLE I
MENGSHI INTELLIGENT VEHICLE SENSOR CONFIGURATION

Sensor Type	No.	Property
Velodyne HDL-64E	1	Range: 70m Horizontal Angle:360 deg Angular Resolution:0.1 deg Updating: 50 ms
IBEO LUX 4L	1	Range:200m Horizontal Angle:110 deg Angular Resolution:0.5 deg Updating:80ms
SICK LMS 291-S05	2	Range:80m Horizontal Angle:180 deg Angular Resolution:0.5 deg Updating:80ms
Delphi ESR	1	Range:174m Horizontal Angle:+/-10 deg Angular Resolution:0.25 deg Updating:50ms
AVT 1394 Pike-100c	3	Resolution Ratio:1000×1000 pixels Recognition Distance:80 m Angular Resolution:62.5 deg Updating:25-40 ms
NovAtel SPAN-CPT	1	Location Accuracy:1 cm Velocity Accuracy:0.02 m/s Attitude Accuracy: 0.05 deg (Pitch/Roll), 0.1 deg (Azimuth Angle) Frequency:5 Hz

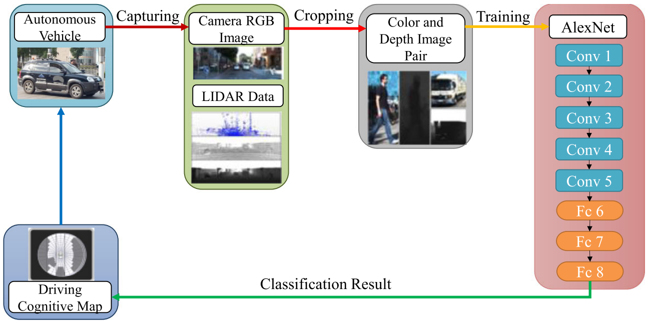


Fig. 3. Pipeline used in our approach.

III. OBJECT CLASSIFICATION APPROACH

Fig. 3 summarizes the pipeline used in this paper. We first capture the sparse-depth map by rotating Velodyne laser-point cloud data from the KITTI database to the RGB image plane using the calibration matrix [25]. Then, we upsample the sparse-depth map to high-resolution depth image. We extract four objects (pedestrian, cyclist, car, and truck) from each image by considering the ground truth from KITTI [19]. We build three image datasets according to these objects. One database is for the pure RGB image of the four kinds of object, one for the gray-scale image with gray level corresponding to actual distance information from LIDAR point clouds, and the third one is an RGB-LIDAR image dataset consisting of the former two information. Each dataset comprises 6843 labeled objects. Finally, we present a structure based on CNN to train a classifier for detecting the four kinds of objects on the road. These classification results are provided to the driving cognitive module for vehicle decision-making and control [26].

This approach has been successfully used in our Mengshi autonomous vehicle. Mengshi is an autonomous vehicle independently developed by Tsinghua University and other research institutes. Velodyne HDL-64E LIDAR lies on top of the vehicle and collects cloud point data. AVT F200C camera is located under the windshield and captures color image. GPS-real-time kinematic (RTK) occupies the trunk and records location data. These sensors are all used for multimodal fusion experiments. This vehicle achieves full autonomous driving and won the second prize in Future Challenge 2016.

A. Sparse Data Upsampling Range

In this study, a novel method of upsampling LIDAR range inputs is employed to align depth with RGB images. In this method, we compute dense depth maps just from the original range data instead of using information from RGB images.

We formulate upsampling using bilateral filtering formalism in our method to generate the dense map D (output image) from a noisy and sparse-depth image I [19]. Assuming that input I is coordinated in pixel units and features calibration w.r.t.$w.r.t.$ a high-resolution camera, pixel positions in I are nonintegers owing to the uncertainty of calibration parameters and data sparsity. According to the intensity value of a pixel p on the depth map, expressed as lower index $()_P$ and its N neighborhood mask, the pixel value lies on the same position of output map D_p, as shown in the following equation:

$$D_p = \frac{1}{W_p}\sum_{q\in N} G_{\sigma_r}\left(|I_q|\right) I_q G_{\sigma_s}\left(\|p - q\|\right) \qquad (1)$$

where G_{σ_r} penalizes the influence of points q caused by their range values, G_{σ_s} weighs inversely to the distance between position p and location q, and W_p functions as the normalized factor, which ensures that the sum of weights are equal to one. In (1), we set G_{σ_s} to be inversely proportional to the Euclidean distance $(\|p - q\|)$ between pixel position p and location q.

B. Dataset

Considering the inherent uncertainty in sensor returns of Velodyne HDL-64E S2, we can assume that the instrument yields a 0.002-rad beam divergence and a 2.5-cm root-mean-square error range in average. Additionally, with increasing distance from LIDAR, these uncertainties will be magnified quickly, indicating that the wider distance of the object from the image I results in more errors in the range positions. With the influence of inherent properties, the value of $G_{\sigma_r}\left(|I_q|\right)$ becomes proportional

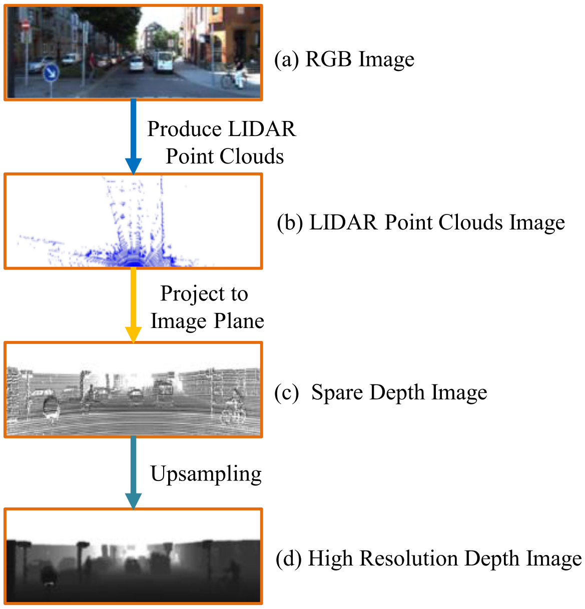

(a) RGB Image
Produce LIDAR Point Clouds
(b) LIDAR Point Clouds Image
Project to Image Plane
(c) Spare Depth Image
Upsampling
(d) High Resolution Depth Image

Fig. 4. (a) Example of an intensity image from KITTI database. (b) Illustrated sparse LIDAR point clouds from KITTI database (bird's-eye view). (c) Sparse-depth map obtained by projecting point cloud data to image plane. (d) High-resolution depth image generated using our proposed method.

to the range value and decreases linearly, penalizing returns as function of their measured distance from LIDAR. Our filter is implemented to normalize weight G_{σ_r} by the maximum range value of $I_q \in n$. This upsampling method can resemble a spatial filter, in which I is "convolved" with a kernel (mask) of fixed size (e.g., 5 × 5). Although kernel size is fixed, the number of pixels $q \in N$ depends on 3-D cloud sparsity and is not constant. Fig. 4(a)–(d) shows an example of an RGB image, the sparse LIDAR point clouds from KITTI database (bird's-eye view), sparse-depth map obtained by projecting point cloud data to image plane, and high-resolution depth image obtained after applying our smoothing filter to (1), respectively.

RGB images and the 3-D point clouds from KITTI are used as object benchmarks [27] to classify objects, such as cars, pedestrians, trucks, and cyclists. RGB color images are captured by the left color video camera (10 Hz, resolution: 1392 × 512 pixels, opening: 90° × 35°), whereas the 3-D point clouds are produced by a Velodyne HDL-64E unit and projected back in image forms. As one of the few available sensors that provide depth information, Velodyne system can generate accurate 3-D data from moving platforms. This system can also be applied in outdoor scenarios and long sensing range compared with structured light systems such as Microsoft Kinect [25], [28].

We crop the objects in 7418 RGB images and upsampled depth images according to the ground truth from KITTI datasets. Our benchmarks comprise 6843 RGB images and upsampled depth data pairs (4 types, 1750 cars, 1750 pedestrians, 1643 trucks, 1700 cyclists, maximum pixels: 600 × 365, minimum pixels: 30 × 30; examples are shown in Fig. 5), which consist of 5475 training images and 1368 test images with corresponding labels.

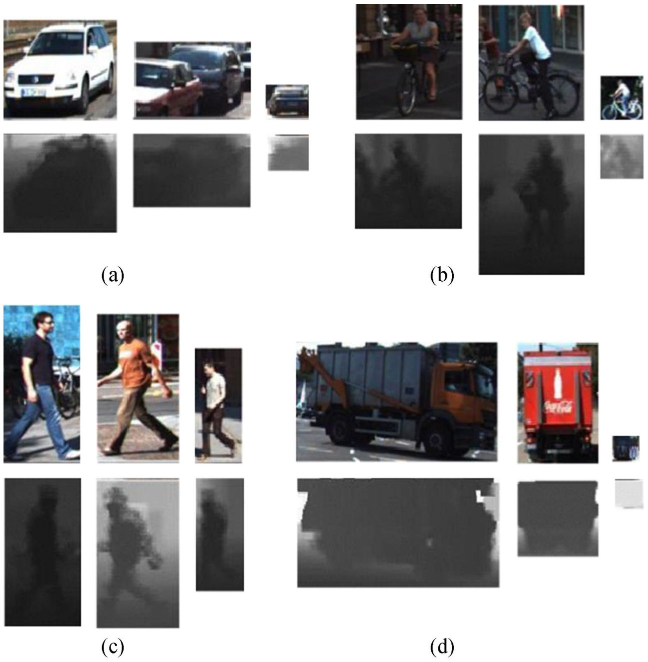


Fig. 5. Datasets used in this paper. (a) Car. (b) Cyclist. (c) Pedestrian. (d) Truck.

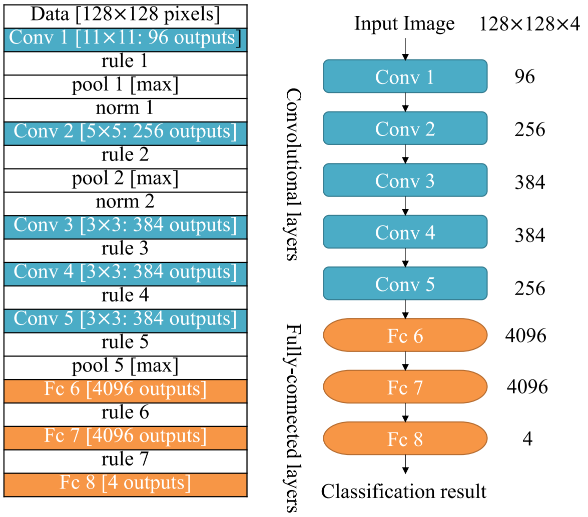


Fig. 6. Structure of AlexNet.

C. Object Classification

For object classification, we classified images from KITTI into cars, cyclists, pedestrians, and trucks. Then, we adopt the AlexNet model as our CNN architecture [29]. AlexNet comprises five convolutional layers (named conv1–conv5) and three fully connected layers (named as fc6, fc7, and fc8), as shown in Fig. 6. Each convolutional layer contains multiple kernels, and each kernel represents a 3-D filter connected to the outputs of the previous layer. For fully connected layers, each layer comprises multiple neurons, and each neuron contains a positive value and is connected to all neurons in the previous layer. We resize images captured from Section III-B to 128 × 128 resolution for valid input and then passed them into AlexNet.

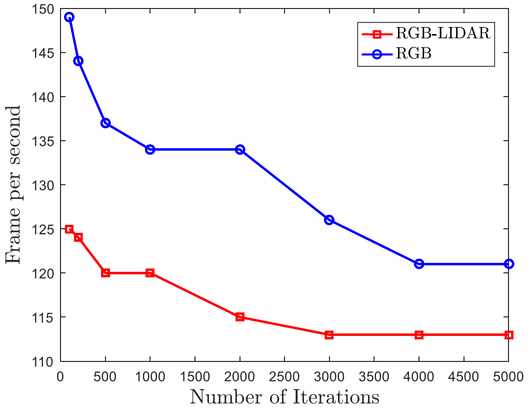


Fig. 7. FPS of RGB-LIDAR-based CNNs.

AlexNet is trained for 1000 classes. We change the size of fc8 layer from 1000 to 4 to match our dataset with four classes. The parameters from layer conv1 to layer fc6 are fixed to prevent overfitting.

This RGB-LIDAR-based method notably improves average precision on classification of four categories in the KITTI datasets (see Section IV). We use the same dataset for training and testing models.

IV. Experiment and Analysis

A. Experimental Setup

We use the RGB-LIDAR dataset to train CNNs. This dataset contains four categories of objects, namely, pedestrians, cyclists, cars, and trucks. Each object consists of four channels, the traditional RGB channel, and an additional depth channel. We call this dataset as RGB-LIDAR for convenience. The class labels correspond to KITTI benchmarks.

The hardware CNN training platform is NVIDIA GeForce GTX Titan X with a Core (TM) i7-5930K (3.5 GHz) and two GPUs, and the hardware CNN testing platform is NVIDIA Jetson TX1 with ARM A57 CUPs and one GPU (1 TFLOP/s 256-core Maxwell). The software development platform comprises convolutional architecture for fast feature embedding and NVIDIA CUDA8.0, and the operation system is Ubuntu16.04.

B. Experimental Results and Analysis

1) Processing Time: We focus on on-road detection of different objects given that processing time is a critical metric for autonomous vehicles. Losing any key frame may influence subsequent control decision, regardless of the object being a pedestrian or a car. When CNN cannot handle information in real time, delay will accumulate and affect the whole on-board network.

We first train the whole network with different numbers of iteration (e.g., 100, 200, 500, 1000, 2000, 3000, 4000, and 5000) and randomly select 1400 images from the test set for testing. This process is repeated 20 times, and the average result is calculated.

Fig. 7 shows the average frames per second (FPS) of the trained dataset under different iterations using RGB-LIDAR and RGB method. Average FPS of the trained network continuously decreases from 100 to 3000 iterations and levels off after 3000. However, these values considerably change and reach approximately 110 FPS using RGB-LIDAR method, but these values change and reach approximately 120 FPS using RGB method. Cameras recently used on autonomous vehicles feature a common standard of approximately 30 FPS. The only difference is that the camera contains a much larger pixel than the input of our net, but this difference can be fixed by scaling. We can obtain an efficient network for parallel processing of three cameras in autonomous vehicle environment using 3000 iterations in the training process. If we train the network with 100 iterations, the final net can handle four cameras simultaneously.

2) Accuracy: To show the performance of additional depth information, we then compared the traditional RGB image and our four-channel RGB-LIDAR image under AlexNet. We compared the average loss and accuracy under the two conditions by changing the size of the training set and increasing iterations from 100 to 4000 for training.

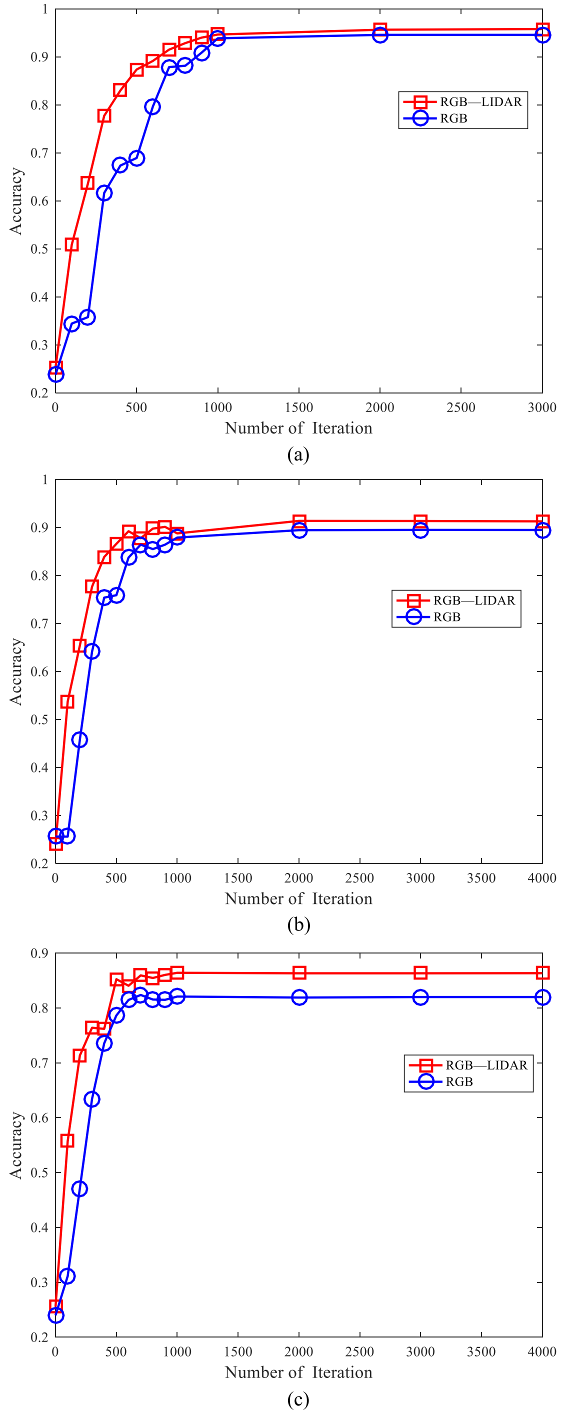
The three different scales of training sets and testing sets are as follows.

1) The training set consists of 5475 images: 1400 cars, 1360 cyclists, 1400 pedestrians, and 1315 trucks. The test set consists of 1368 images: 328 trucks, 350 cars, 340 cyclists, and 350 pedestrians.
2) For the exchange training set and testing set, 1368 images are selected as training set, and the testing set consists of 5475 images.
3) A total of images are randomly selected from the total set as training set, and the testing set includes 5475 images.

Fig. 8 shows the corresponding average accuracies. All curves show similar tendencies, and average accuracies plateaus after approximately 1000 iterations. However, the AlexNet with RGB-LIDAR constantly performs better, especially when training with a small set (400 images). Using the proposed method, with the increase of the depth of information provided, the multimodel RGB-LIDAR data show approximately 5% higher accuracy than pure RGB data when training iterations reach higher than 1000. RGB-LIDAR images consistently present better accuracies than RGB-based training sets when using iterations lower than 1000.

3) Loss: We list the corresponding average losses of combinations of datasets shown in Fig. 9(a)–(c). Average loss continuously drops as the number of iterations increases. The loss of AlexNet drops to zero when using number of iterations higher than 500 and when a small set of 400 images is trained. The multimodel RGB-LIDAR consistently converges more rapidly than the RGB training set.

4) Classification Prediction Result: We offer a confusion matrix of classification prediction results based on RGB-LIDAR method in Table II. The values in the main diagonal are the percentage of the correctly classified items, the rest are the miss classified items and corresponding the percentage of the error. We found that the main error happens when “others” is classified as “truck,” and a “truck” is classified as “others.” We think these two class objects are very alike in the current dataset, and their backgrounds are very similar. This will be the next step we need to solve the problem.

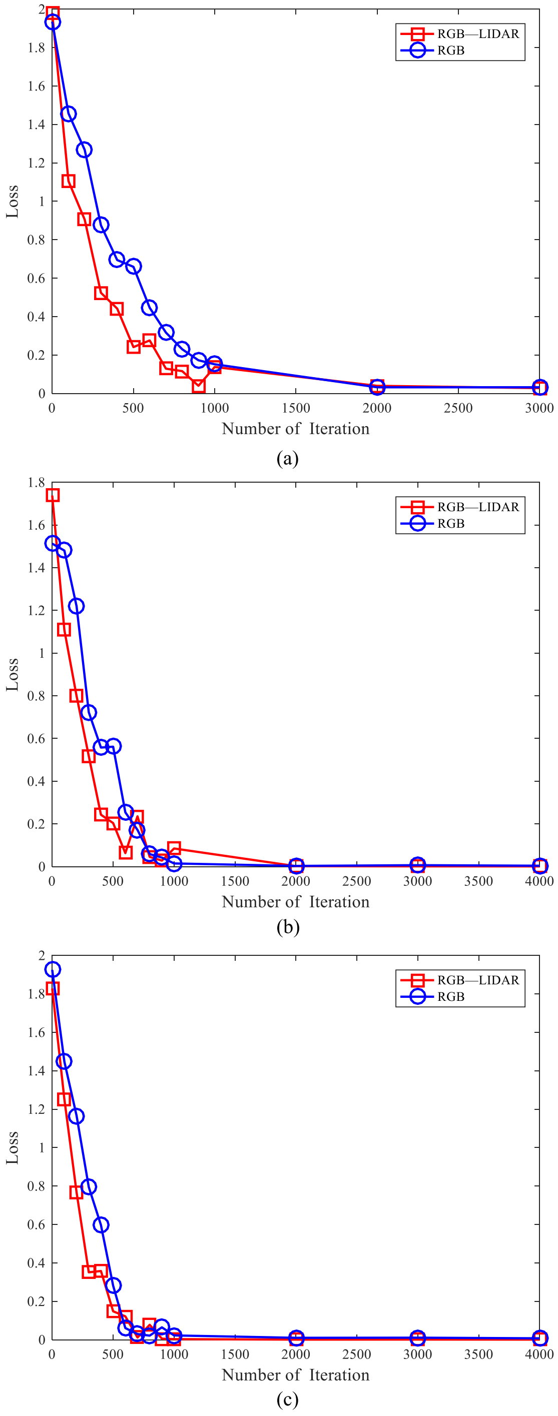


(a)

(b)

(c)

Fig. 8. Average accuracies of different training sets: (a) 5475 image set, (b) 1000 image set, and (c) 400 image set.



(a)

(b)

(c)

Fig. 9. Average losses of different training set: (a) 5475 images, (b) 1000 images, and (c) 400 images.

C. Comparison of Classification Results

The AlexNet proposed by Krizhevsky *et al.* classified 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into 1000 different classes and the top-1 and top-5 error rates achieved were 37.5% and 17.0%, respectively [29]. The unsupervised multistage features learning proposed by Sermanet *et al.* yielded a competitive error rate of 10.55% in pedestrian detection [30]. The region-based CNN (R-CNN) algorithm proposed by Girshick *et al.* achieved a mean average precision (MAP) of 53.3% on PASCAL VOC dataset [31]. The fast R-CNN algorithm proposed by Girshick achieved a MAP of 66.0% and worked faster than the previous one on VOC [27]. With the increase in the number of images, the average error of the object classification using CNN increases considerably. Experiment results on the benchmark dataset of KITTI demonstrate that our RGB-LIDAR data can reach lower loss, and higher

TABLE II
Classification Prediction Results Based on RGB-LIDAR Method

	Classification prediction results (%)				
	Pedestrian	Cyclist	Car	Truck	Others
Pedestrian	100	0	0	0	0
Cyclist	0	100	0	0	0
Car	0	0	98.6	0	1.4
Truck	0	0	0	88.6	11.4
Others	0	0	0	2.8	97.2

average accuracy at 15% than single RGB-based model. The average accuracy of the final classifier can reach a maximum value of 96%.This means, with the additional high-level LIDAR feature, we can improve the accuracy of classifiers.

V. Conclusion

In this paper, we propose a deep-learning-based approach by fusing vision and LIDAR data for object detection in autonomous vehicle environment. On the one hand, we upsample point clouds of LIDAR data and convert the upsampled point cloud data into pixel-level depth feature map. On the other hand, we convert the RGB together with depth feature map and then fed the data into a CNN. On the basis of the integrated RGB and depth data, we utilize DCNN to perform feature learning from raw input information and obtain informative feature representation to classify objects in the autonomous vehicle environment. The proposed approach, in which visual data are fused with LIDAR data, exhibits superior classification accuracy over the approach using only RGB data or depth data. During the training phase, using LIDAR information can accelerate feature learning and hasten the convergence of CNN on the target task. We perform experiments using the public dataset and display the effectiveness and efficiency of the proposed approach.

In this paper, the camera and LIDAR on the vehicle are used to collect images and point cloud images, and NVIDIA GeForce GTX Titan X and NVIDIA Jetson TX1 are used for offline detection and classification. In our further work, we will perform real-world experiments, and verify the ability of the proposed approach in classifying objects in an autonomous vehicle environment based on vehicle-mounted domain controller.

References

[1] L. Figueiredo *et al.*, "Towards the development of intelligent transportation systems," *Intell. Transp. Syst.*, vol. 88, pp. 1206–1211, 2001.

[2] X. Hu, H. Wang, and X. Tang, "Cyber-physical control for energy-saving vehicle following with connectivity," *IEEE Trans. Ind. Electron.*, vol. 64, no. 11, pp. 8578–8587, Nov. 2017.

[3] E. D. Dickmanns, *Dynamic Vision for Perception and Control of Motion*. New York, NY, USA: Springer, 2007.

[4] Eureka PROMETHEUS Project, 1986. [Online]. Available: http://www.eurekanetwork.org/project/id/45

[5] DARPA Grand Challenge, 2005. [Online]. Available: http://www.darpa.mil/default.aspx

[6] Google's autonomous vehicle, 2010. [Online]. Available: http://googleblog.blogspot.com/2010/10/what-were-driving-at.html

[7] H. B. Gao, X. Y. Zhang, T. L. Zhang, Y. C. Liu, and D. Y. Li, "Research of intelligent vehicle variable granularity evaluation based on cloud model," *Acta Electron. Sin.*, vol. 44, no. 2, pp. 365–374, 2016.

[8] J. Imran and P. Kumar, "Human action recognition using RGB-D sensor and deep convolutional neural networks," in *Proc. Int. Conf. Adv. Comput., Commun. Informat.*, 2016, pp. 144–148.

[9] C. Lv, H. Wang, and D. Cao, "High-precision hydraulic pressure control based on linear pressure-drop modulation in valve critical equilibrium state," *IEEE Trans. Ind. Electron.*, vol. 64, no. 10, pp. 7984–7993, Oct. 2017.

[10] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGB-D images," in *European Conference on Computer Vision*. New York, NY, USA: Springer, 2012, pp. 746–760.

[11] S. Gupta, R. Girshick, P. Arbeldez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *European Conference on Computer Vision*, vol. 8695. New York, NY, USA: Springer, 2014, pp. 345–360.

[12] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proc. 2015 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 681–687.

[13] Z. Wang *et al.*, "Correlated and individual multi-modal deep learning for rgb-d object recognition," arXiv preprint 375, arXiv:01655.1604.

[14] N. Kosaka and G. Ohashi, "Vision-based nighttime vehicle detection using CenSurE and SVM," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2599–2608, 2015.

[15] M. Cheon, W. Lee, C. Yoon, and M. Park, "Vision-based vehicle detection system with consideration of the detecting location," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1243–1252, Sep. 2012.

[16] R. O. Chavez-Garcia and O. Aycard, "Multiple sensor fusion and classification for moving object detection and tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 525–534, Feb. 2016.

[17] L. E. Navarro-Serment, C. Mertz, and M. Hebert, "Pedestrian detection and tracking using three-dimensional LADAR data," *Int. J. Robot. Res.*, vol. 29, no. 12, pp. 1516–1528, 2010.

[18] J. Dolson, J. Baek, C. Plagemann, and S. Thrun, "Upsampling range data in dynamic environments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1141–1148.

[19] C. Premebida, J. Carreira, J. Batista, and U. Nunes, "Pedestrian detection combining RGB and dense LIDAR data," in *Proc. 2014 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2014, pp. 4112–4117.

[20] J. Schlosser, C. K. Chow, and Z. Kira, "Fusing LIDAR and images for pedestrian detection using convolutional neural networks," in *Proc. 2016 IEEE Int. Conf. Robot. Autom.*, 2016, pp. 2198–2205.

[21] V. Zarzoso, P. Comon, and R. Phlypo, "A contrast function for independent component analysis without permutation ambiguity," *IEEE Trans. Neural Netw.*, vol. 21, no. 5, pp. 863–868, May 2010.

[22] F. Wu *et al.*, "Rapid localization and extraction of street light poles in mobile LiDAR point clouds: A supervoxel-based approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 2, pp. 292–305, Feb. 2017.

[23] M. Luber, L. Spinello, and K. O. Arras, "People tracking in RGB-D data with on-line boosted target models," in *Proc. 2011 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2011, pp. 3844–3849.

[24] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *Proc. 2011 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2011, pp. 821–826.

[25] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[26] H. B. Gao, X. Y. Zhang, Y. C. Liu, and D. Y. Li, "Cloud model approach for lateral control of intelligent vehicle systems," *Sci. Program.*, vol. 24, no. 12, pp. 1–12, 2016.

[27] R. Girshick, "Fast R-CNN," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[28] C. Lv, Y. Liu, X. Hu, H. Guo, D. Cao, and F. Y. Wang, "Simultaneous observation of hybrid states for cyber-physical systems: A case study of electric vehicle powertrain," *IEEE Trans. Cybern.*, vol. PP, no. 99, pp. 1–11, 2017.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[30] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. Lecun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3626–3633.

[31] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

Hongbo Gao received the Ph.D. degree in computer science and technology from Beihang University, Beijing, China, in 2016.

He is currently an assistant researcher with State Key Laboratory of Automotive Safety and Energy, Tsinghua University, He has authored or coauthored more than 30 journal papers. He is the coholder of 6 patent applications. His current research interests include intelligent vehicles and robotics, machine learning, deep learning, decision-making and control of intelligent driving.

Bo Cheng received the B.Tech. and M.S. degrees in automotive engineering from Tsinghua University, Beijing, China, in 1985 and 1988, respectively, and the Ph.D. degree in Mechanical Engineering from University of Tokyo, Japan, in 1998.

He is a Professor in automotive engineering with Tsinghua University. He has authored over 100 papers and is a coinventor on 40 patents in China and Japan. His research interests include vehicle dynamics, control for driver-assistance systems, and active vehicle safety. Dr. Cheng has engaged in over 30 sponsored projects, and he was the recipient of 10 awards.

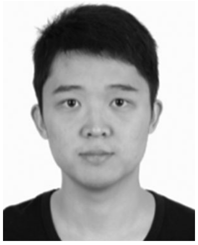
Dr. Chen was the recipient of ten awards.

Jianqiang Wang received the B.Tech. and M.S.degrees in automotive engineering from the Jilin University of Technology, Changchun, China, in 1994 and 1997, respectively, and the Ph.D. degree in automotive engineering from Jilin University, Changchun, in 2002.

He is currently a Professor with the Department of Automotive Engineering, the State Key Laboratory of Automotive Safety and Energy, and Collaborative Innovation Center for Electric Vehicles, Tsinghua University, Beijing, China. He has authored or coauthored more than 120 journal papers. He is the coholder of 80 patent applications. He has been involved in more than ten sponsored projects and was a recipient of nine awards. His active research interests include intelligent vehicles, driving-assistance systems, and driver behavior.

Keqiang Li received the B.Tech. degree in automotive engineering from Tsinghua University, Beijing, China, in 1985, and the M.S. and Ph.D. degrees in automotive engineering from Chongqing University, Chongqing, China, in 1988 and 1995, respectively.

He is a Professor in automotive engineering with Tsinghua University. He has authored more than 90 papers and is a coinventor of 12 patents in China and Japan. His research interests include vehicle dynamics, control for driver assistance systems, and hybrid electrical vehicles. Dr. Li is a senior member of the Society of Automotive Engineers of China. He is on the Editorial Boards of the International Journal of Intelligent Transportation Systems Research and the International Journal of Vehicle Autonomous Systems. He was a recipient of the Changjiang Scholar Program Professor Award and of some awards from public agencies and academic institutions of China.



Jianhui Zhao received the B.E. degree in information engineering from National University of Defense Technology, Hunan, China, in 2006, and M.S. degree in automotive engineering from Military Transportation University, Tianjin, China, in 2011, where he is currently pursuing the Doctor's degree in computer science and technology, Tsinghua University. His current research interests include intelligent driving and cloud computing.

Deyi Li received the Ph.D. degrees in computer science and technology from Nautical watts Watt University, Edinburgh, UK, in 1983. He is a Professor in Department of Computer Science and Technology with Tsinghua University. He has authored over 100 papers and is a coinventor on 30 patents in China. His research interest covers artificial intelligence, deep learning, intelligent vehicle and cloud computing.

Dr. Li is the president of Chinese Association of Artificial Intelligence, and also he is an academician of the Chinese Academy of engineering and anacademician Eurasian Academy of Sciences.