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

A Novel Approach for Detecting Road Based on Two-Stream Fusion Fully Convolutional Network

Xin Lv, Ziyi Liu, Jingmin Xin and Nanning Zheng

Abstract— **Road detection is one of the most basic tasks of autonomous driving systems. At present, researches on this issue mainly take two kinds of data as input, *i.e.*, LIDAR point clouds and RGB images from cameras. To make best use of the advantages and bypass the disadvantages of these two kinds of data, we propose a novel network, namely two-stream fusion fully convolutional network (TSF-FCN), which can take advantage of both the accurate location information from LIDAR point clouds and rich appearance information from RGB images. One stream of this network is LIDAR stream which aggregates multi-scale contextual information from LIDAR point clouds. The other stream is RGB stream which is used for extracting features from RGB images. To fuse the two streams, the feature maps of RGB stream are converted to a bird-view representation to concatenate with that of LIDAR stream. In this way, the two kinds of data can complement each other for detecting road. To verify the efficacy of our TSF-FCN, experiments are carried on KITTI-ROAD benchmark and competitive performance is achieved compared with state-of-the-art methods.**

I. INTRODUCTION

Road detection aims at guiding automated vehicles to avoid obstacles, which is an important part of autonomous driving systems. An effective algorithm for road detection is a guarantee of safe driving for automated vehicles. The research on this issue is becoming a hot spot with the popularity of self-driving cars. Since the effect of deep learning has been proved in other computer vision fields [1]–[3], more and more scholars are trying to solve this problem utilizing deep learning. Although some of algorithms [4], [5] have achieved good results, most of them use only LIDAR point clouds or only RGB images without fusing them, which leads to that the scenes they can handle is limited. RGB images contain dense pixel information, whereas they are susceptible to external environment such as illumination and weather conditions. LIDAR point clouds obtained by the laser scanner are robust to these problems and it can provide accurate 3D location information of objects. Unfortunately, they are sparse and lack of appearance information. Obviously, both LIDAR point clouds and images have their own defects and their information is complementary.

Motivated by this fact, we propose a two-stream fusion fully convolutional network (TSF-FCN), in which not only the accurate location information of LIDAR points but also the dense pixel information of RGB images are employed for road detection. The overview of our approach is described as follows. LIDAR points are transformed to a bird-view grid map firstly so that it can be fed into the network. The LIDAR grid maps and front-view RGB images are fed into the LIDAR stream and RGB stream of the network, respectively. Then, the output of RGB stream is projected to grid map coordinate frame to fuse with the corresponding feature maps of LIDAR stream. Finally, the map reflecting the probabilities of each pixel to be regarded as road is obtained from the network. The structure of our model is shown in Fig. 1.

The main contributions of this paper are:

- We propose a novel network named as TSF-FCN which can fuse images captured by monocular camera and LIDAR point clouds to improve the accuracy and robustness of road detection, especially in challenging scenarios.
- The feature maps of RGB stream are projected to LIDAR grid map coordinate frame which can make use of richer information of images than utilizing the RGB information corresponding to LIDAR points directly.

This paper is organized as follows. Section II presents previous work related with our method. In Section III, the details of our approach are presented including the method for LIDAR grid maps, the architecture of TSF-FCN and the concrete structure of fusion layer. Experimental results and discussions are described in Section IV. Section V presents conclusion and future work.

II. RELATED WORK

Over the past few years, many methods have been proposed to perform road detection, which differ from each other mainly in the data they used, such as stereo data [6], [7], images from monocular camera [8]–[11] and LIDAR point clouds [12], [13]. Since the generality of the approach relying on the information from any single sensor is not enough, road detection based on multi-sensor fusion has received a lot of attention. In [14], the author proposed to fuse LIDAR point clouds and images in the framework of conditional random field which can get a balanced result by jointly optimizing them. [15] proposed a multi-modal based approach in which illumination-invariant features from images and ground points obtained from LIDAR point clouds are fused for detecting road. [16] presented a road estimation method that made use of the local spatial-relationship of the LIDAR points projected to the corresponding image to avoid the common assumption in other literatures. The methods detecting road with data fusion we mentioned above are all

*This research was partially supported by the National Natural Science Foundation of China (No. 61773312, 61790563), the Programme of Introducing Talents of Discipline to University (No. B13043).

The authors are with Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shannxi, P.R.China

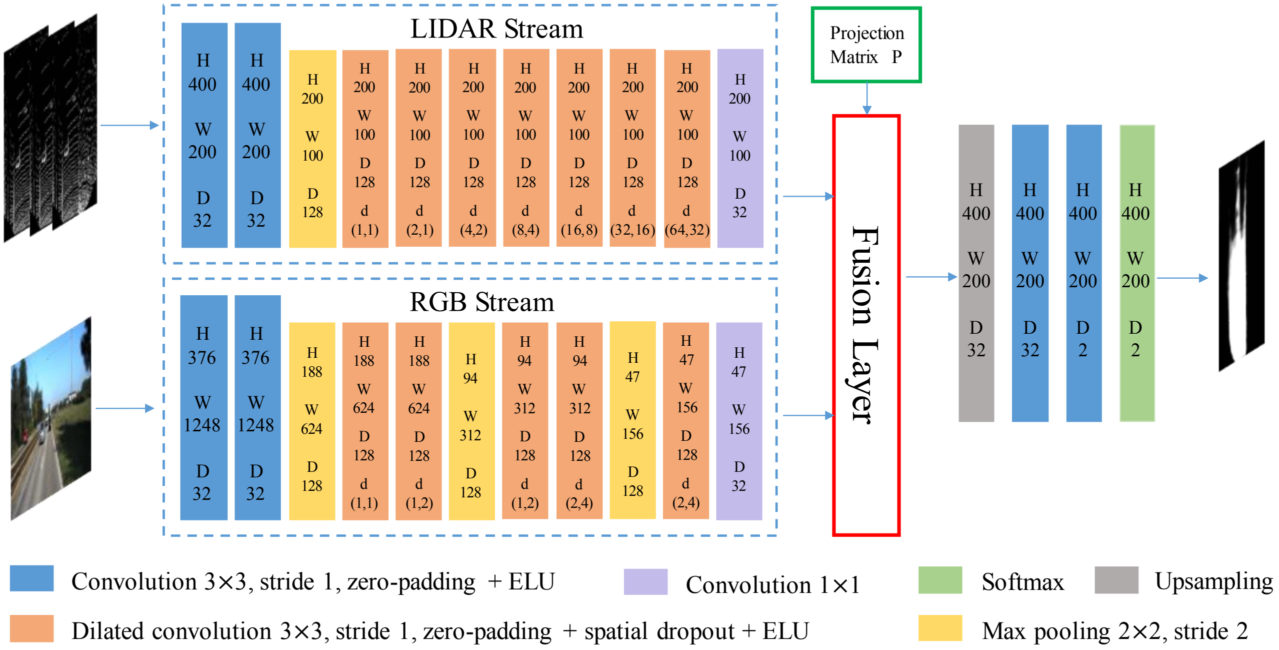

Fig. 1. The architecture of our proposed two-stream fusion fully convolutional network (TSF-FCN). The input of LIDAR stream is the grid map with three channels introduced in Section III-A, while the input of RGB stream is the corresponding RGB image. The outputs of both streams are fused in the Fusion Layer, which will be detailed in Section III-D. The output of TSF-FCN is a map denoting the probability of every pixel in the LIDAR grid map to be regarded as road. W, H and D represent width, height and number of feature maps, respectively. And d denotes the dilation of dilated convolution.

hand-crafted and their effects are not satisfying although they are based on multi-sensor fusion.

With the promotion of deep learning, some deep learning based approaches are proposed to achieve this challenge. [17] put forward a network combining CNNs with deep deconvolutional neural networks which can learn higher order image structure to detect road more effectively. [18] presented convolutional patch networks which incorporated position information of patches for pixel-wise labeling. [19] proposed a cognitive model combined with neural network to detect the free space of the current and adjacent lanes. Since the proposal of fully convolutional network (FCN) [20], which is an end-to-end network for segmentation, has brought image segmentation to a new stage, more and more road detection methods based on FCN yield outstanding performance on KITTI dataset [21]. In [22], the author presented a siamesed fully convolutional network (s-FCN-loc) which not only used the RGB image but also embedded semantic contour and location prior into the network. [23] transformed unstructured LIDAR point clouds to bird-view images used as the input of FCN by creating a grid map. In spite of the remarkable results these methods achieved on KITTI-ROAD benchmark, their performance in various challenging scenarios can not be guaranteed because they all use information from single sensor.

Utilizing deep learning to achieve sensor fusion for road detection needs to be further explored, whereas it's indeed a feasible approach to improve the accuracy and robustness of the detection of road in some challenging scenes. Owing to the complementarity of LIDAR point clouds and RGB images, we propose a novel network called TSF-FCN which is employed for fusing LIDAR data and images to make use of their advantages.

III. ROAD DETECTION BASED ON TSF-FCN

In this section, we detail the proposed two-stream fusion fully convolutional network (TSF-FCN), which performs road detection utilizing information from both LIDAR point clouds and RGB images. LIDAR points are transformed to a bird-view grid map firstly, after which LIDAR grid maps and the corresponding RGB images are fed into the network. The output of this network is a map representing the probability of each pixel to be regarded as road, based on which we can generate the final areas belonging to road.

A. LIDAR Grid Map

Since the input of a convolutional network should be images, the LIDAR point clouds need to be transformed to suit it. A suitable method is converting them to bird-view images [24], called LIDAR gird maps. First, a grid in x-y plane of LIDAR point clouds is created. Then every point has a corresponding grid cell. The statistics, chosen according to our needs, of a cell are viewed as pixel values. To better evaluate our approach on benchmark, we adopt the region covers $-10m$ to $10m$ in lateral direction and $6m$ to $46m$ in longitudinal direction, that is $y \in [-10, 10]$, $x \in [6, 46]$. The size of a grid cell is $0.1m \times 0.1m$. We compute three statistics of each cell, namely mean height, gap between maximum height and minimum height and occupancy (if there is at least one point in a cell, the occupancy of it is one, otherwise it's zero). If the mean height of a cell is zero, we can't tell if there is no point in it or the mean height is really zero,

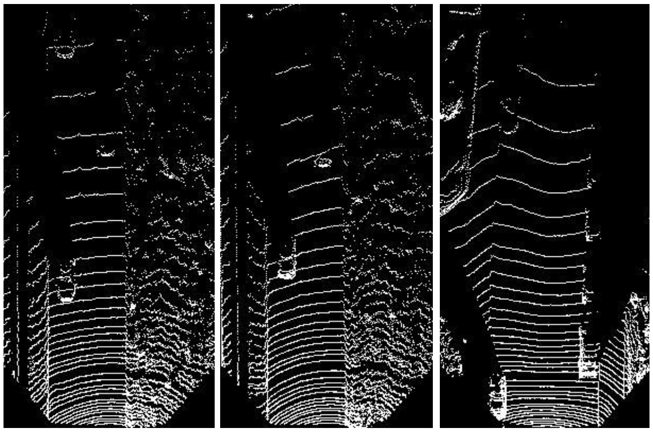

Fig. 2. Three examples of occupancy map. White denotes that there is at least one LIDAR point in this grid cell and black means there is no point here.

so does to the gap between maximum height and minimum height. Therefore, the occupancy map (as shown in Fig. 2) is necessary. In this way, a bird-view grid map with three channels of size 400×200 is obtained.

B. Fixed Bird-view Image

After converting LIDAR point clouds to grid maps, it is necessary to convert the images into a bird-view representation, in order to fuse them. Inverse perspective mapping (IPM) is a frequently used solution to this problem. Nevertheless, it is only effective on flat roads without obstacles. As a result, a bird-view image derived from IPM cannot match with the corresponding LIDAR grid map exactly. In order to solve this problem, a new method is proposed to obtain fixed bird-view image, which can match the LIDAR grid map better. By projecting LIDAR points to the image coordinate frame using calibration information, RGB information corresponding to every LIDAR point can be gotten. Then we can generate a fixed bird-view image with three channels just like the grid map introduced in Section III-A. The only difference is replacing the statistics of a cell with RGB information. Fig. 3 demonstrates that the fixed bird-view image matches the LIDAR grid map more exactly than that generated from IPM.

As presented in [25], the transformation from image coordinate frame to grid map coordinate frame can be represented as a projection matrix P. Suppose the RGB image is of size $H_i \times W_i \times C$, each channel of it can be flatten to a H_iW_i vector I_c. Each channel of the $H_g \times W_g \times C$ fixed bird-view image we want can be flatten to a H_gW_g vector B_c. Projection matrix P is a $H_gW_g \times H_iW_i$ sparse matrix. If there is a LIDAR point in ith cell of B_c and its coordinate in image frame is (u, v) corresponding to jth element of vector I_c, we have

$$P(i, j) = 1; \tag{1}$$

Otherwise,

$$P(i, j) = 0. \tag{2}$$

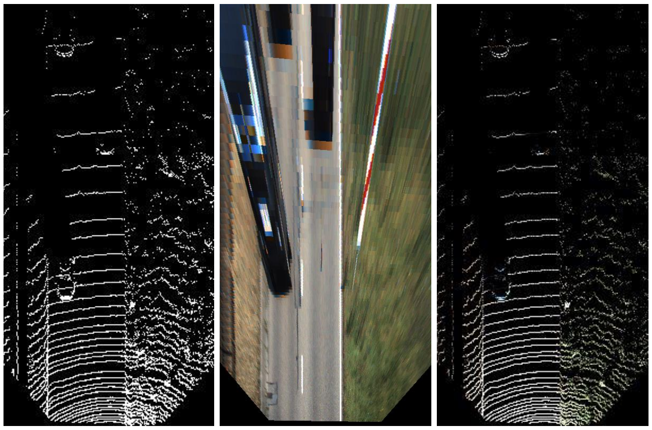

Fig. 3. Comparison of the bird-view image from IPM and the fixed bird-view image. The left image is the grid map presented in Section III-A. The middle image is the bird-view image obtained by IPM and the right one is the fixed bird-view image generated by the method introduced in Section III-B.

Then we can get each channel of the fixed bird-view image by reshaping the vector $B_c = P \times I_c$ to a matrix of size $H_g \times W_g$. The projection matrix P can be pre-calculated and adjusted to adapt to the size of images.

C. Two-Stream Fusion Fully Convolutional Network

Inspired by the fusion idea for object detection in [25], we propose the TSF-FCN based on the network in [23], which is one of the top-performing approaches on the KITTI dataset. As shown in Fig. 1, our network is composed of two streams, namely, LIDAR stream and RGB stream. LIDAR stream uses dilated convolution operator [26] to obtain multi-scale contextual information and its input is the LIDAR grid map described in Section III-A. RGB stream is for deriving the features of RGB images (images are padding to a same size before as an input), which also utilizes dilated convolution operator.

In order to detect the road more accurately, it is necessary to have a large receptive field. One possible way to enlarge the size of receptive field without increasing the number of convolutional parameters is adding pooling layers. However, the feature maps must be upsampled because the output of FCN is with the same size as the input, which will result in the loss of resolution. A solution to this problem is using the dilated convolution, which can expand the receptive field while maintaining the resolution and the number of parameters. To reduce the memory requirement of the TSF-FCN, one and three pooling layers are added in LIDAR stream and RGB stream, respectively.

To make better use of the two kinds of data with different modalities, a two-stream network is a more suitable choice than ordinary FCN. The network in [22] also has two streams whose parameters are shared with each other and updated simultaneously. But in our network, before the fusion layer, the two streams are trained independently to suit the difference in the modality of data. At the fusion layer, the outputs of two streams would be fused. More details are provided in Section

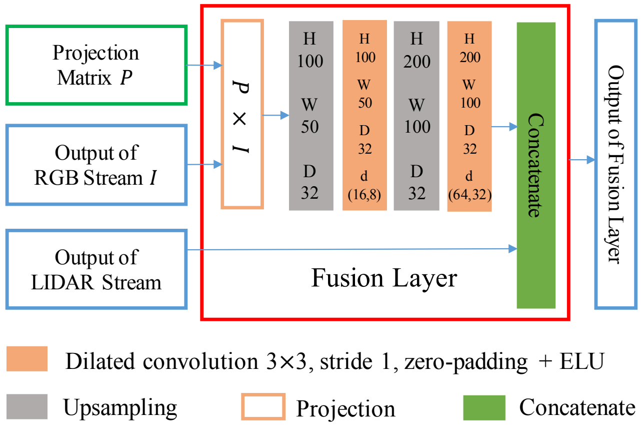

Fig. 4. The structure of fusion layer in which the outputs of two streams are fused.

III-D. After fusion, the feature maps are upsampled to the same size as the input of LIDAR stream and then fed into the following convolutional layers. Finally, the probability map reflecting each grid cell belonging to road is obtained after the cascaded softmax layer.

D. Fusion of RGB Stream and LIDAR Stream

A naive way to fuse LIDAR point clouds and RGB images is feeding the fixed bird-view image to the network together with the LIDAR grid map directly. Unfortunately, since we just extract the RGB information of the place where LIDAR points fell, the fixed bird-view image just makes use of little information of the original RGB image.

In order to take advantage of high level features of RGB images, we prefer to feed raw front-view images into network directly and transform the feature maps of RGB stream to LIDAR grid map coordinate frame using the method explained in Section III-B rather than feed the fixed bird-view images into the network. With this trick, more features beneficial to road detection can be encoded in the fixed bird-view feature maps. Since RGB stream has two fewer convolutional layers but two more max pooling layers than LIDAR stream, the fixed bird-view feature maps are concatenated with LIDAR feature maps after the processing of two dilated convolutional layers and two upsampling layers as Fig. 4 shows.

IV. EXPERIMENT

A. Dataset

The TSF-FCN is trained and evaluated on KITTI-ROAD benchmark which includes 289 training images and 290 testing images. The dataset consists of three categories of images, namely Urban Marked (UM), Urban Multiple Marked (UMM) and Urban Unmarked (UU), which are gotten at different road scenes. Since there are only ground truth annotations for images in training set, we divide the training set into two parts by randomly selecting 10 images from each category as validation set which helps to reflect the performance of our network. The other images are taken as training set to adjust the parameters of network in training stage. A detailed description of this dataset is shown in Table I. LIDAR, GPS and stereo data corresponding to images are also provided by this benchmark. In this work, only LIDAR and monocular color data are used.

TABLE I

DETAILED DESCRIPTION OF KITTI DATASET

Category	Train	Validation	Test
UM (urban marked)	85	10	96
UMM (urban multiple marked)	86	10	94
UU (urban unmarked)	88	10	100

B. Generate More Accurate Ground Truth Annotations

Since the ground truth annotations given by KITTI-ROAD benchmark are in front-view while the output of network is in bird-view, the annotations should be transformed to bird-view space. As demonstrated in Section III-B, bird-view images from IPM cannot match with LIDAR grid maps exactly, which is the same to ground truth annotations (as shown in Fig. 5). Inspired by the method of generating more accurate annotations in [23], the sparse fixed bird-view annotations are obtained by projecting LIDAR points to the raw ground truth annotations. Then, dense bird-view ground truth can be generated based on the following approach. The grid cells of the sparse annotations can be divided into N parts according to the angle α between x-axis and the ray from LIDAR to cells.

Suppose that a grid cell c_i belongs to jth part and the distance from this cell to LIDAR is d_i, find the minimum distance min_j, from LIDAR to the cells where there is at least one LIDAR point corresponding to non-road areas, of jth part. The grid cell is considered as road if it satisfies

$$d_i < min_j. \tag{3}$$

After processing every grid cell where there is no LIDAR point like this, the dense fixed bird-view ground truth (as shown in Fig. 5) can be obtained.

C. Data Augmentation and Training Details

In the case of insufficient training samples, data augmentation has been proven as an effective measure to avoid overfitting and increase the generality of the network. Since we can only train the network with the data in the training set, data augmentation is necessary. In view of the specialty of the network, we just flip the LIDAR grid maps and RGB images which can amplify the dataset by two times.

To initialize the TSF-FCN, we first pre-train the LIDAR stream using LIDAR grid maps and the corresponding ground truth. Then, the whole network is trained by employing SGD optimization algorithm with other layers randomly initialized. The initial learning rate and batch size are set to 0.001 and 4, respectively. If the performance on validation set is not improved in the last two epoches, the learning rate will be reduced by two times. The cross entropy function is utilized as the objective function, which is defined as

$$J = -\frac{1}{N \times W \times H} \sum_{k=1}^{N} \sum_{i=1}^{W} \sum_{j=1}^{H} \log p_{i,j}^{k} \tag{4}$$

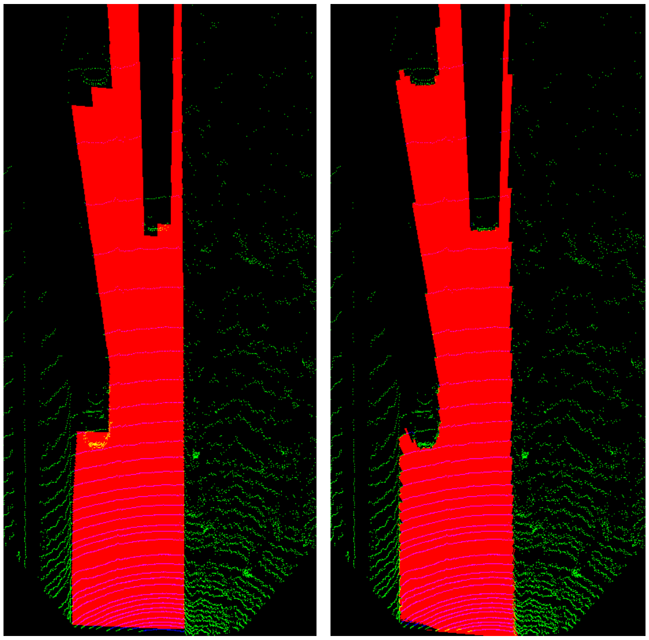

Fig. 5. Comparison of the annotation from IPM and the fixed ground truth annotation, where red denotes road, blue and green represent the LIDAR points corresponding to road and non-road. The left image with mismatch is the annotation from IPM corresponding to the grid map about occupancy shown in Fig. 2 and the right one is about the fixed annotation.

where p represents the probability for right category predicted by the network, and W, H, N denote the width of output, the height of output and batch size, respectively.

The training environment of TSF-FCN is equipped with GeForce GTX TITAN X GPU with 12GB of memory.

D. Results and Discussions

To evaluate our approach, we conduct it on the test set of KITTI-ROAD benchmark mentioned in Section IV-A, which is available online. The results are evaluated with metrics max F-measure (MaxF), average precision (AP), precision (PRE), recall (REC), false positive rate (FPR), and false negative rate (FNR) in BEV on three categories: UM, UMM, UU and a category called urban road which provides an overall performance. For the sake of presenting the effectiveness of our approach, we compare the result with that of other published methods, including, FusedCRF [14], MixedCRF [27], FCN-LC [28], DNN [17], Up-Conv-Poly [4] and LODNN [23].

TABLE II
COMPARISON ON URBAN ROAD KITTI BENCHMARK (IN%) .

URBAN	MaxF	AP	PRE	REC	FPR	FNR
FusedCRF [14]	88.25	79.24	83.62	93.44	10.08	6.56
MixedCRF [27]	90.59	84.24	89.11	92.13	6.20	7.87
FCN-LC [28]	90.79	85.83	90.87	90.72	5.02	9.28
DNN [17]	93.43	89.67	**95.09**	91.82	**2.61**	8.18
Up-Conv-Poly [4]	93.83	90.47	94.00	93.67	3.29	6.33
LODNN [23]	94.07	92.03	92.81	**95.37**	4.07	**4.63**
TSF-FCN(ours)	**94.48**	**93.65**	94.28	94.69	3.17	5.31

TABLE III
COMPARISON ON UM (IN%) .

UM	MaxF	AP	PRE	REC	FPR	FNR
FusedCRF [14]	89.55	80.00	84.87	94.78	7.70	5.22
MixedCRF [27]	91.57	84.68	90.02	93.19	4.71	6.81
FCN-LC [28]	89.36	78.80	89.35	89.37	4.85	10.63
DNN [17]	93.65	88.55	94.28	93.03	2.57	6.97
Up-Conv-Poly [4]	92.20	88.85	92.57	91.83	3.36	8.17
LODNN [23]	92.75	89.98	90.09	**95.58**	4.97	**4.42**
TSF-FCN(ours)	**94.86**	**93.36**	**94.65**	95.08	**2.45**	4.92

TABLE IV
COMPARISON ON UMM (IN%) .

UMM	MaxF	AP	PRE	REC	FPR	FNR
FusedCRF [14]	89.51	83.53	86.64	92.58	15.69	7.42
MixedCRF [27]	92.75	90.24	94.03	91.50	6.39	8.50
FCN-LC [28]	94.09	90.26	94.05	94.13	6.55	5.87
DNN [17]	94.17	92.70	**96.73**	91.74	**3.41**	8.26
Up-Conv-Poly [4]	95.52	92.86	95.37	95.67	5.10	4.33
LODNN [23]	**96.05**	95.03	95.79	**96.31**	4.66	**3.69**
TSF-FCN(ours)	95.42	**95.43**	95.61	95.23	4.80	4.77

TABLE V
COMPARISON ON UU (IN%) .

UU	MaxF	AP	PRE	REC	FPR	FNR
FusedCRF [14]	84.49	72.35	77.13	93.40	9.02	6.60
MixedCRF [27]	85.69	75.12	80.17	92.02	7.42	7.98
FCN-LC [28]	86.27	75.37	86.65	85.89	4.31	14.11
DNN [17]	91.76	86.84	**93.06**	90.50	**2.20**	9.50
Up-Conv-Poly [4]	**92.65**	89.20	92.85	92.45	2.32	7.55
LODNN [23]	92.29	90.35	90.81	**93.81**	3.09	**6.19**
TSF-FCN(ours)	92.35	**91.94**	91.69	93.02	2.75	6.98

Table II presents the six metrics of our approach and the methods used for contrast on Urban Road benchmark. It can be seen from the quantitative results that the proposed TSF-FCN achieves the highest MaxF and AP, which indicates the validity of our approach. Compared with the hand-crafted fusion methods FusedCRF [14] and MixedCRF [27], our approach is more remarkable in all criteria. In addition, MaxF and AP are improved on the basis of outstanding performance of LODNN [23] which only has LIDAR stream. The main reason behind this improvement is that the information from both LIDAR and monocular camera are taken into consideration for detecting road.

As Table III IV V show, our approach gets outstanding results in all the three kinds of scenes, which demonstrates that our method is not only accurate but also robust for different situations. However, the FNR of our network is less competitive than that of LODNN, which may result from that the influence of the external environment on the images isn't completely eliminated when detecting road in some special circumstances. For these three scenarios, the proposed network performs relatively poorly on the UU set. It can be explained that the roads in UU set are much less regular than that of the other two data sets.

Fig. 6 displays three examples of road detection in images from test set. The Green, blue and red denote true positives, false positives and false negatives, respectively. As shown in these examples, false positives and false negatives are easy

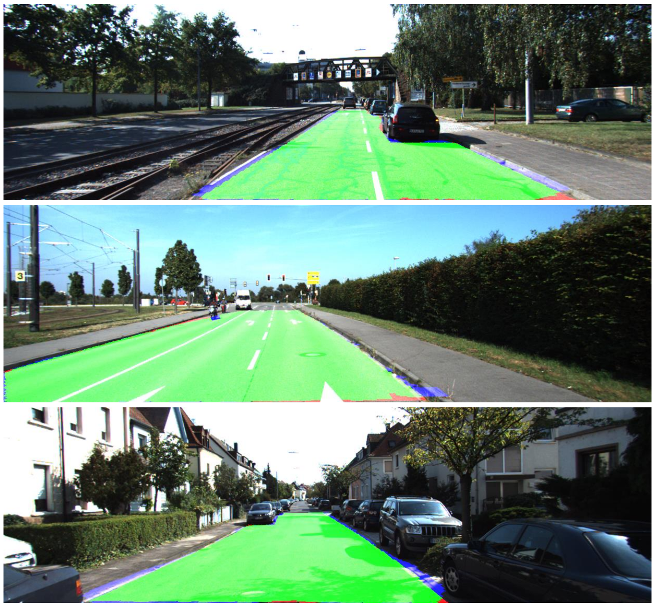

Fig. 6. Examples of road detection on test set. The green corresponds to true positives; blue and red denote respectively false positives and false negatives.

to appear on the edges of the roads and obstacles. False positives are usually due to the fact that the changes in height or color of the edges of the roads and obstacles are not obvious in some cases, while false negatives are generally caused by false alarms, such as roadside leaves.

V. CONCLUSION AND FUTURE WORK

In this paper, a novel neural network structure, namely two-stream fusion fully convolutional network (TSF-FCN) is proposed to fuse rich appearance information of RGB images and accurate location information in LIDAR point clouds for detecting road. LIADR points and images have their own characteristics and are complementary to each other. This paper explored the advantages of fusing them and experimental results verified that it can improve the robustness and accuracy of road detection. In the future, we plan to verify the effectiveness of TSF-FCN in more challenging situations, such as scene with poor illumination, water reflection and so on.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[2] N. Karianakis, T. J. Fuchs, and S. Soatto, "Boosting convolutional features for robust object proposals," *Computer Science*, 2015.

[3] G. Bertasius, J. Shi, and L. Torresani, "Deepedge: A multi-scale bifurcated deep network for top-down contour detection," in *Computer Vision and Pattern Recognition*, 2015, pp. 4380–4389.

[4] G. L. Oliveira, W. Burgard, and T. Brox, "Efficient deep models for monocular road segmentation," in *Ieee/rsj International Conference on Intelligent Robots and Systems*, 2016, pp. 4885–4891.

[5] A. Laddha, M. K. Kocamaz, L. E. Navarro-Serment, and M. Hebert, "Map-supervised road detection," in *Intelligent Vehicles Symposium*, 2016, pp. 118–123.

[6] H. Badino, U. Franke, and R. Mester, "Free space computation using stochastic occupancy grids and dynamic programming," *Dynamic Vision Workshop for Iccv*, 2007.

[7] P. Y. Shinzato, D. Gomes, and D. F. Wolf, "Road estimation with sparse 3d points from stereo data," in *IEEE International Conference on Intelligent Transportation Systems*, 2014, pp. 1688–1693.

[8] M. Passani, J. J. Yebes, and L. M. Bergasa, "Fast pixelwise road inference based on uniformly reweighted belief propagation," in *Intelligent Vehicles Symposium*, 2015, pp. 519–524.

[9] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, "Multinet: Real-time joint semantic reasoning for autonomous driving," 2016.

[10] Z. Chen and Z. Chen, "Rbnet: A deep neural network for unified road and road boundary detection," pp. 677–687, 2017.

[11] S. Chen, J. Shang, S. Zhang, and N. Zheng, "Cognitive map-based model: Toward a developmental framework for self-driving cars," in *Intelligent Transportation Systems (ITSC), 2017 IEEE 20th International Conference on*. IEEE, 2017, pp. 1–8.

[12] C. Tongtong, D. Bin, L. Daxue, Z. Bo, and L. Qixu, "3d lidar-based ground segmentation," vol. 32, no. 14, pp. 446–450, 2012.

[13] L. Chen, J. Yang, and H. Kong, "Lidar-histogram for fast road and obstacle detection," in *IEEE International Conference on Robotics and Automation*, 2017, pp. 1343–1348.

[14] L. Xiao, B. Dai, D. Liu, T. Hu, and T. Wu, "Crf based road detection with multi-sensor fusion," in *Intelligent Vehicles Symposium*, 2015, pp. 192–198.

[15] X. Hu, F. S. A. Rodriguez, and A. Gepperth, "A multi-modal system for road detection and segmentation," in *Intelligent Vehicles Symposium Proceedings*, 2014, pp. 1365–1370.

[16] P. Y. Shinzato, D. F. Wolf, and C. Stiller, "Road terrain detection: Avoiding common obstacle detection assumptions using sensor fusion," in *Intelligent Vehicles Symposium Proceedings*, 2014, pp. 687–692.

[17] R. Mohan, "Deep deconvolutional networks for scene parsing," *Computer Science*, 2014.

[18] C. A. Brust, S. Sickert, M. Simon, E. Rodner, and J. Denzler, "Convolutional patch networks with spatial prior for road detection and urban scene understanding," in *Visapp*, 2015, pp. 82–90.

[19] S. Chen, S. Zhang, J. Shang, B. Chen, and N. Zheng, "Brain-inspired cognitive model with attention for self-driving cars," *IEEE Transactions on Cognitive & Developmental Systems*, vol. PP, no. 99, pp. 1–1, 2017.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CVPR (to appear)*, Nov. 2015.

[21] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[22] J. Gao, Q. Wang, and Y. Yuan, "Embedding structured contour and location prior in siamesed fully convolutional networks for road detection," in *IEEE International Conference on Robotics and Automation*, 2017, pp. 219–224.

[23] L. Caltagirone, S. Scheidegger, L. Svensson, and M. Wahde, "Fast lidar-based road detection using fully convolutional neural networks," 2017.

[24] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," 2016.

[25] Z. Wang, W. Zhan, and M. Tomizuka, "Fusing bird view lidar point cloud and front view camera image for deep object detection," 2017.

[26] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015.

[27] X. Han, H. Wang, J. Lu, and C. Zhao, "Road detection based on the fusion of lidar and image data," vol. 14, no. 6, p. 172988141773810, 2017.

[28] C. C. T. Mendes, V. Frmont, and D. F. Wolf, "Exploiting fully convolutional neural networks for fast road detection," in *IEEE International Conference on Robotics and Automation*, 2016.