Cogn Comput (2018) 10:272–281
https://doi.org/10.1007/s12559-017-9524-y

Segmentation of Drivable Road Using Deep Fully Convolutional Residual Network with Pyramid Pooling

Xiaolong Liu[1] · **Zhidong Deng**[1]

Received: 28 May 2017 / Accepted: 31 October 2017 / Published online: 27 November 2017

Abstract In recent years, the self-driving car has rapidly been developing around the world. Based on deep learning, monocular vision-based environmental perceptions of either ADAS or self-driving cars are regarded as a feasible and sophisticated solution, in terms of achieving human-level performance at a low cost. Perceived surroundings generally include lane markings, curbs, drivable roads, intersections, obstacles, traffic signs, and landmarks used for navigation. Reliable detection or segmentation of drivable roads provides a solid foundation for obstacle detection during autonomous driving of the self-driving car. This paper proposes an RPP model for monocular vision-based road detection based on the combination of fully convolutional network, residual learning, and pyramid pooling. Specifically, the RPP is a deep fully convolutional residual neural network with pyramid pooling. In order to greatly improve prediction accuracy on the KITTI-ROAD detection task, we present a new strategy through an addition of road edge labels and an introduction of an appropriate data augmentation so as to effectively handle small training samples contained in the KITTI road detection. The experiments demonstrate that our RPP has achieved remarkable results, which ranks second in both unmarked road and marked road tasks, fifth in multiple-marked-lane task, and third in combination task. In this paper, we propose a powerful 112-layer RPP model through the incorporation of residual connections and pyramid pooling into a fully convolutional neural network framework. For small training sample problems such as the KITTI-ROAD detection, we present a new strategy through an addition of road edge labels and data augmentation. It suggests that addition of more labels and introduction of appropriate data augmentation can help deal with small training image problems. Moreover, a larger size of crops or combination with more global information also benefit improvements in road segmentation accuracy. If regardless of restricted computing and memory resources for such large-scale networks like RPP, the use of raw images instead of any crops and the selection of a large batch size are expected to further increase road detection accuracy.

Keywords CNN · Drivable road · Semantic segmentation · Self-driving car

Introduction

In recent years, self-driving car has rapidly been developing around the world. Based on deep learning, monocular vision-based cognitive computation of either ADAS or self-driving cars are regarded as a feasible and sophisticated solution for achieving human-level performance at a low cost. Recognition of surroundings generally includes lane markings, curbs, drivable roads, intersections, obstacles, traffic signs, and landmarks used for navigation. Reliable drivable road segmentation or detection provides a solid foundation for obstacle detection during autonomous driving of the self-driving car. For road cognitive computation, sensors

✉ Zhidong Deng
michael@tsinghua.edu.cn

Xiaolong Liu
xllau@126.com

[1] State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science, Tsinghua University, Beijing 100084, China

like monocular, stereo, and infrared cameras, as well as LiDAR, are popularly adopted. As an active vision, LiDAR can detect depth information and is not susceptible to external changes in illumination. But it has relatively low resolution and expensive quotation. In contrast, camera characterizes high resolution and rich information, e.g., grayscale, color, and texture, which is undoubtedly able to deliver more raw data for drivable road cognitive computation. For segmentation of roads that are hard to discriminate using LiDAR, such as almost no height difference between roadway and sideway, image-based detection is probably a sole viable alternative.

Since 2012 [14], significant advances in the deep convolutional neural network (CNN) for specific scenes or problems, e.g., visual object recognition, human face recognition, speech recognition, and natural language processing, have been made. For many benchmark problems such as ImageNet [21, 31], LFW [3], Switchboard and CallHome [29], and WMT'14 [27], CNN-based recognition systems reach or surpass the human level. Unbelievable capabilities of CNN could basically be attributed to automated feature extraction of multiple granularities, which resembles visual cortex pathway of mammals in a sense. Although a variety of powerful CNN models, e.g., VGG [23], GoogleNet [24], ResNet [10], and DenseNet [11], are presented and new records have been being updated, there exist many problems to be resolved. For example, CNN architecture has too much parameters [7, 8]. It is too deep to efficiently train [10]. In particular, traditional deep CNN requires supervised learning, which means that it solely makes available to complex problems with labeled big data, instead of small samples or unlabeled data scenes.

This paper proposes a residual network with pyramid pooling (RPP) model for monocular vision-based road cognitive computation based on a combination of fully convolutional network, residual learning, and pyramid pooling. Specifically, the RPP is a deep fully convolutional residual neural network with pyramid pooling. In order to greatly improve prediction accuracy on the KITTI-ROAD task, we present a new strategy through an addition of road edge labels and an introduction of an appropriate data augmentation so as to effectively handle small training samples of the KITTI-ROAD. Our experimental results show that the proposed RPP achieves highly competitive results, which ranks second in both unmarked and marked road tasks, fifth in multiple-marked-lane task, and third in combination task.

Our main contributions are listed below:

1. We propose a deep fully convolutional residual network with pyramid pooling (RPP) model.
2. A new strategy through an addition of road edge labels as well as data augmentation based on underexposure compensation is presented.
3. RPP wins the second places on both unmarked and marked road tasks and third place in combination task.

The paper is organized as follows: In "Related Work," the related work is reviewed. "RPP" proposes our large-scale RPP model. "Experimental Results" provides training or test details of RPP and gives our experimental results on the KITTI road detection benchmark. Finally, "Conclusion and Perspectives" concludes this paper with a brief summary.

Related Work

In this section, we get involved in a literature review of two aspects: road detection in self-driving and image-based semantic segmentation of roads or scenes.

For many years, commonly used approaches on road detection in self-driving car are generally based on LiDAR, camera, or combination of both. Compared to the monocular camera, expensive LiDAR is capable of acquiring three-dimensional information. But it has low resolution and loses rich appearance attributes such as color and texture information, which is generally viewed to play a critical role in real road segmentation.

With significant progress in deep learning, pixel-level semantic segmentation of roads or scenes recently conducted based on convolutional neural networks remarkably outperforms previous traditional methods.

Paper [9] investigates detection problem on drivable road area based on homography estimation and driving model. In their work, on the basis of a stereo camera, the surface of structured road area is yielded through homography estimation. In [1], color consistency characteristic is used for segmentation of road area, although it is only available to a very clean road surface. Based on monocular vision, real-time detection algorithm of the drivable image area is proposed by [19]. They generate a region of interest (ROI) both from dynamic threshold search and from drag process. Some researchers are working on the cognitive computation of the lane boundary. Literature [4] presents and discusses a lane boundary detection technique that is necessary for the task of autonomous driving. For more reviews on camera-based approaches, also see [25, 26] for details.

There are several colored-based methods for road segmentation. Due to the complexity of real road surfaces, e.g., distinct colors for identical roads, heavy shadows, and crossing railways, such methods may be not practicable solutions.

From image-based classification to pixel-based classification, deep convolutional neural network methods have surprisingly surpassed traditional machine learning models for many benchmark problems through end-to-end training since [14]. In recent years, it is also increasingly employed for image-based segmentation of objects [24].

A more viable method for pixel-level semantic segmentation seems to be a fully convolutional neural network (FCN) proposed by [16, 22]. In FCN, the fully connected layer as the classifier of classical CNNs is replaced with the deconvolutional layer or upsample operations. This gives rise to the size of output image being just equal to that of raw input images. From this work on, many variants of such methods have emerged. For example, conditional random fields are used for greatly improving prediction accuracy in [2].

Seemingly, the increase of convolutional layers and combination of global information often benefits performance of deep neural networks. In [10, 23], very deep neural networks are investigated and the effects of increasing the number of learn-able layers are discussed.

Moreover, for image-based segmentation, the use of dilated convolution layer [30] can significantly enhance prediction accuracy. It is proven that the integration of global information and the combination of local receptive fields usually improve the accuracy of prediction. In [28] and [2], the conditional random field is based on global information for scene segmentation. Literature [17] uses echo state networks for microscopic cellular image segmentation. In particular, paper [32] exploits pyramid pooling operations to extract global features and achieves very competitive classification results.

The earliest version of FCNs [16] is based on VGG16. Compared to VGG16, ResNet proposed by [10] is demonstrated to have stronger capabilities of feature representation, the depth of which can even reach up to 1001 learnable layers. It won the 1st places on ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation in ILSVRC 2015 and COCO 2015 competitions [10]. Many lately published deep learning architectures use ResNets as extractors of a feature hierarchy because of their exceptional generalization performance. Literature [18] also implement fully convolutional work on KITTI-ROAD dataset. They achieve this by using a network-in-network (NiN) architecture and by transferring this model into an FCN after training. Paper [20] proposes several architecture refinements which provide the best trade-off between runtime and segmentation quality. This is achieved by a mapping between filters and classes at the expansion side of the network. Moreover, paper [15] combines the best of both unsupervised and supervised methods to reduce human annotation effort and make training scalable. Actually, they all get the state-of-the-art result at that time. The algorithm of RPP has got a better result than all of them in prediction accuracy.

Our RPP adopts residual networks the backbone which guarantees a stronger representation than VGG. Moreover, RPP implements a pyramid pooling structure that can grab global context effectively. Through augmenting KITTI-ROAD data set, RPP achieves a fairly good result on such a small dataset.

RPP

In the following sections, a novel RPP architecture is proposed and a new strategy for dealing with small training sample problems is introduced. The basic idea behind our RPP is to incorporate large-scale residual connections and pyramid pooling into an FCN framework for monocular vision end-to-end semantic segmentation.

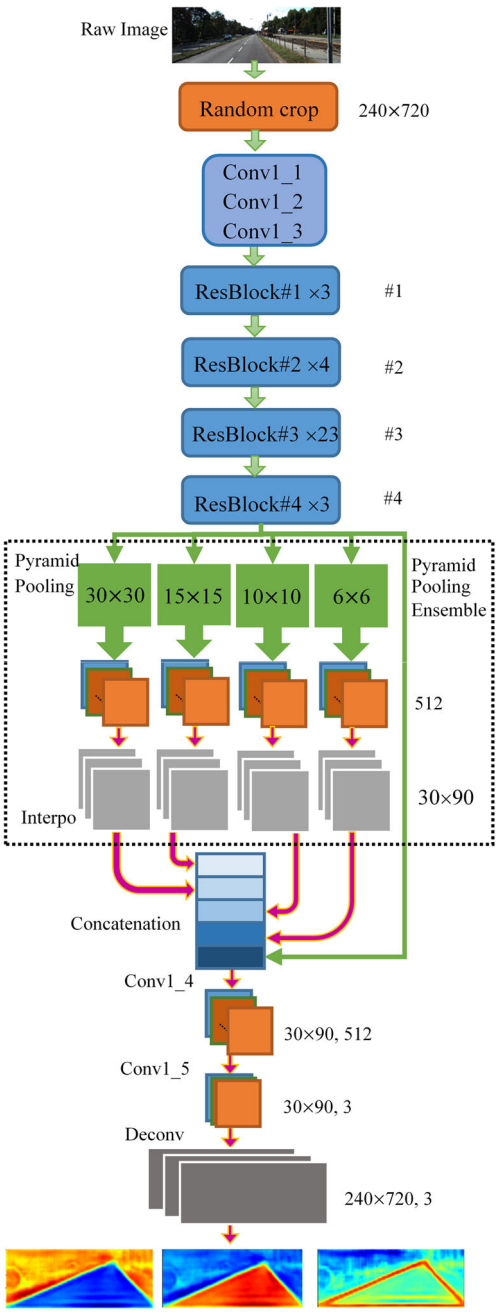


Fig. 1 The fully convolutional residual network with pyramid pooling (RPP)

RPP Architecture

In general, classical CNNs have fully connected layers as the classifier. For semantic segmentation problem, an FCN was proposed by [16], which is similar to autoencoder with encoding and decoding in a sense. In this paper, we propose a fully convolutional residual network with pyramid pooling, i.e., RPP, whose architecture is shown in Fig. 1.

Overall, our RPP architecture consists of four residual ensembles, one pyramid pooling ensemble, and one deconvolutional layer. First, raw input images are randomly cropped as large as possible on account of our limited memory resource. In resulting crops, the size of short sides could be fixed according to prior knowledge of road surface which always appears in the lower half of the image. While long sides could be done randomly with the size as large as possible. We have done experiments on different crop sizes. Second, we design a unique pyramid pooling ensemble after residual ensemble #4, although residual ensembles #1–#4 themselves have similar structures to those of ResNet. As shown in Table 1 and Fig. 1, each residual ensemble comprises several building blocks, which are given in brackets and stacked with the numbers of blocks. Within the dotted box is a pyramid pooling ensemble, which contains a pyramid pooling layer, a convolutional layer of 11 kernels, and an interpolation layer. Note that interpolation layer can be used to resize feature maps to a larger size. Third, three different categories of hot maps, which correspond to the drivable road, road edge, and non-drivable road, are generated using 33 kernels for feature representation outputted by Convl_4. Fourth, three hot maps are further fed to deconvolutional layer so as to give rise to feature maps of 240 × 720. Note that batch

Table 1 RPP architecture for the KITTI road detection benchmark. Building blocks are shown in brackets and stacked with the numbers of blocks

Layer name	Output size	# channels	112-layer
Random crop	240 × 720	1	
Conv1_1	120 × 360	64	3 × 3, stride 2/BN
Conv1_2	120 × 360	64	3 × 3, stride 1/BN
Conv1_3	120 × 360	128	3 × 3, stride 1/BN
ResEnsemble #1	60 × 180		$\begin{bmatrix} 1\times1 & 64 \\ 3\times3 & 64 \\ 1\times1 & 256 \end{bmatrix} \times 3$
ResEnsemble #2	30 × 90		$\begin{bmatrix} 1\times1 & 128 \\ 3\times3 & 128 \\ 1\times1 & 512 \end{bmatrix} \times 3$
ResEnsemble #3	30 × 90		$\begin{bmatrix} 1\times1 & 256 \\ 3\times3 & 256 \\ 1\times1 & 1024 \end{bmatrix} \times 3$
ResEnsemble #4	30 × 90		$\begin{bmatrix} 1\times1 & 512 \\ 3\times3 & 512 \\ 1\times1 & 2048 \end{bmatrix} \times 3$
Pyramid pooling	1 × 3	512	30 × 30, stride 30,
	2 × 6	512	15 × 15, stride 15,
	3 × 9	512	10 × 10, stride 10,
	5 × 30	512	6 × 6, stride 6
Conv_pyramid	1 × 3	512	1 × 1, stride 1/BN
	2 × 6	512	1 × 1, stride 1/BN
	3 × 9	512	1 × 1, stride 1/BN
	5 × 30	512	1 × 1, stride 1/BN
Interpo layer	30 × 90	512	30 × 30, stride 30,
	30 × 90	512	15 × 15, stride 15,
	30 × 90	512	10 × 10, stride 10,
	30 × 90	512	6 × 6, stride 6
Concatenation	30 × 90	4,096	
Convl_4	30 × 90	512	3 × 3, stride 1/BN/DropOut
Convl_5/hotmaps	30 × 90	3	3 × 3, stride 1/BN
Deconv layer	240 × 720	3	15 × 15, stride 8
Ouput layer	240 × 720	3	

normalization (BN) [12] and ReLU follow each convolutional layer in our RPP. Finally, through taking maximization, three 240×720 softmax output maps of different categories labels are produced, which have the same size as above crops of raw input images and can be employed to compute loss. By re-assembling crops during an inference phase, we can accomplish road segmentation for raw images.

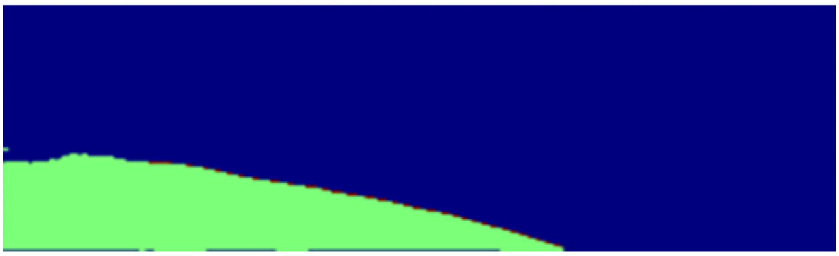
Addition of Road Edge Label and Data Augmentation

In order to deal with such small number of training samples as in the KITTI-ROAD road detection, we present a new data augmentation strategy by introducing the road edge label.

For training datasets, we add a pixel-level annotation of road edge through a few lines of code automatically. As a result, all the KITTI-ROAD training images have three categories of labels, including drivable road, road edge, and non-drivable road. In fact, road edges provide information even more critical than road areas themselves. Because road edges directly separate drivable roads from non-drivable roads. It could be demonstrated to play an important role in improvements of performance in the semantic segmentation of drivable road.

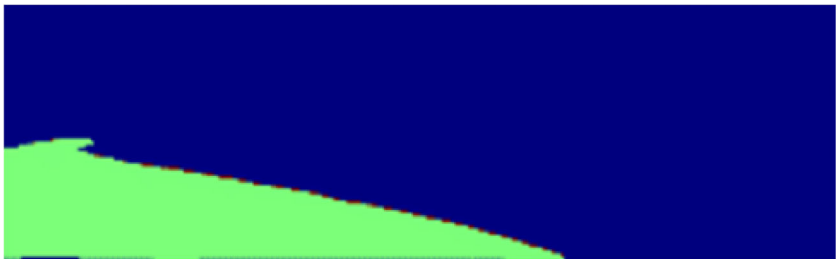
To further address challenges of small training samples, we present a data augmentation strategy based on underexposure compensation procedure [6]. In underexposure compensation, the histogram of dark areas, instead of whole images, is moved up. In contrast to histogram equalization, underexposure compensation only increases the contrast in the low-illumination area. As shown in Fig. 2, this leads to

(a) Raw image



(b) Results yielded through detecting road surface in (a).

(c) Results obtained using underexposure compensation on (a).



(d) Results achieved through detecting road surface in (c).

Fig. 2 Result comparison between one original image and the image after augmentation

increased luminance and contrast of dark image patches and makes their distribution more even, even if the existence of shadows in the glare of the sun. It is easy to see from Fig. 2 that detection performance of drivable road has significantly been improved after completing data augmentation based on underexposure compensation.

Discussions

In this paper, we propose a high-performance 112-layer RPP model that combines residual connections with pyramid pooling under an FCN framework. Residual learning is allowed to have much deeper layers and seems more likely to achieve excellent generalization performance for classification/segmentation problems [10]. Meanwhile, pyramid pooling method presented by [32] is capable of enlarging receptive fields and extracting global features. For so large models but a small amount of training image, any global information, together with BN and Dropout, is crucial to avoid over-fitting. In addition, FCN framework [16] is naturally well suited to pixel-level semantic segmentation, whose deconvolutional layers function as bilinear interpolation and can resize feature maps to the same size as input image crops. As a result, it would be reasonable to believe that our RPP architecture should achieve competitive performance on the KITTI road detection benchmark.

In fact, to reach this goal, we have to further tackle the challenging difficulties brought about by small training samples the KITTI road detection data set provides. KITTI-ROAD dataset is composed of 289 training and 290 test images totally [5]. For this sake, we introduce three strategies. First, different crops of raw images give rise to different detection performance. Small size crops may lose global information. For instance, the sidewalk would be misclassified as roadway if trained with a smaller size crops due to their similar appearance. But through the combination of more global information, e.g., taking image crops as large as possible under memory constraints, such situation is less likely to happen. Second, owing to the fact that road edges directly separate drivable road from the non-drivable road, detection performance of drivable road surfaces is expected to become better after adding road edge labels. But a small proportion of such category samples over whole samples may cause unbalanced training, although our experimental results illustrated that it actually has less impact on detection accuracy. At last, it is fairly obvious that data augmentation based on underexposure compensation technique indeed improves road segmentation performance significantly. Because it makes the histogram of low-illumination patches more even without any change to normal-illumination area.

Experimental Results

We evaluate the proposed RPP on the KITTI road detection benchmark data set [5] and it won the second places on the two categories of road scenes: UM_ROAD and UU_ROAD.

KITTI-ROAD Dataset and Evaluation Criteria

The KITTI-ROAD detection data set contains four different categories of road scenes below: UU_ROAD-urban unmarked (98/100), UM_ROAD-urban marked (95/96), UMM_ROAD-urban multiple marked lanes (96/94), and URBAN_ROAD—combination of the three above, where (x/y) represent x training images and y test ones, respectively [5]. Labels or ground truth are provided for training images only.

It is known that one of the biggest challenges posed by this benchmark lies in small training samples themselves. Additionally, a portion of road scenes contained in such benchmark is rather complex. For example, some road boundaries look like very unclear and irregular. In some images, there are strong shadows, overexposures, or heavy road surface interference that is resulting from railway or crosswalk crossing. At the same time, pedestrian and road images may have very similar gray levels or colors. But single road surface probably possesses distinct materials. In particular, there may occur a variety of occlusions from pedestrians or vehicles due to the fact that all raw images are collected from real urban traffic scenarios.

For comparison, we use the same measure metrics as [5] for evaluating the performance of such classical pixel-based road detection benchmark. It follows:

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$F - \mathrm{Measure} = \frac{(1 + \beta^2)\mathrm{PRE} \times \mathrm{REC}}{\beta^2(\mathrm{PRE} + \mathrm{REC})} \tag{3}$$

$$\mathrm{MaxF} = \mathrm{argmax}_{\tau} F - \mathrm{measure} \tag{4}$$

$$\mathrm{AP} = \frac{1}{11} \sum_{r(1,0.1,\cdots 1.0)} \max_{f>r} \mathrm{PRE}(\tilde{r}) \tag{5}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN},} \tag{6}$$

Table 2 The influences of input size, road edge label, and data augmentation on the RPP road detection performance

Model	Overall accuracy (%)	Mean accuracy (%)	Mean IU (%)	FW_{IU}(%)
RPP_half	93.91	93.90	93.89	94.02
RPP_whole	95.93	95.70	95.62	95.95
RPP_whole_c3	95.99	75.24	70.04	94.60
RPP_whole_c3_aug	96.23	87.73	83.37	98.59

where TP = true positive, FP = false positive, TN = true negative, and FN = false negative. PRE stands for precision, REC recall, AP average precision, and ACC accuracy. τ indicates classification threshold and r recall value. Note that MaxF and AP reflect the optimal and average performance of an algorithm, respectively [5].

To make the experiments more persuasive, we also have adopted metrics in regular performance evaluation metric for image segmentation and scene parsing tasks [16]. We reported four metrics: pixel accuracy (P_{acc}), mean accuracy (M_{acc}), mean region intersection upon union (IU) (M_{IU}), and frequency weighted IU (FW_{IU}). Let n_{ij} indicate the number of pixels of class i predicted to belong to class j, where there are n_{cl} different classes, and let $t_i = \sum_j n_{ij}$ indicate the number of pixels of class i. The four metrics are described in Eqs. 7, 8, 9, and 10.

$$P_{acc} = \frac{\sum_i n_{ii}}{\sum_i t_i} \tag{7}$$

$$M_{acc} = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i} \tag{8}$$

$$M_{IU} = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \tag{9}$$

$$FW_{IU} = \frac{1}{\sum_k t_k} \sum_i \frac{t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \tag{10}$$

Implementation Details and Results

Based on the well-known open source frame Caffe [13], we implemented the proposed RPP model as well as the other CNNs like FCNs and ResNet, the two latter networks of which are only used for analysis and comparison.

In our experiments, the learning rate is initially chosen and the weight decay is set to 0.05. During the training phase, the updated law of the learning rate is given below:$(1 - \text{ITER}/\text{MaxITER})^{0.9}$, where ITER (MaxITER) denotes the (maximum) iterations of training and MaxITER is assigned to 10,000.

(a) Raw images in KITTI-ROAD. (1, 242 × 375), the blue rectangle shows small crop used in (b) and the red one is large crop used in (c).

(b) Test results from small crop (240 × 240).

(c) Test result from large crop (240 × 720).

Fig. 3 Influence of crop sizes on RPP detection accuracy. The batch size is set to 8 (left column) and 4 (right column), respectively. The figures (**b**, **c**) are results stitched together from their respective crop setting

At first, we investigate the influences of crop sizes, road edge labels, and data augmentation on RPP road detection performance. Considering that ground truths of test images of the KITTI road detection benchmark are unknown to us, we randomly select 1/3 of training images as test datasets for this specific experiments on performance analysis in terms of measures given in [16] which can refer to Eqs. 7, 8, 9, and 10. The experimental results obtained are listed in Table 2. We have four major observations from Table 2. First, although our RPP network is able to allow any size images as input, the detection performance could get worse when input size is adjusted to 1/2 (i.e., RPP_half) from a full size (i.e. RPP_whole). Second, different crop experiments are shown in Fig. 3. Small crop (240×240) performs worse than bigger crop (240×720). Some of the reasons behind this are that a small size crop may lose global information that benefits discrimination between roadway and sideway. Third, the addition of road edge labels and training with them (i.e. RPP_whole_c3) give rise to improved performance. Finally, the prediction accuracy is considerably increased after performing our data augmentation described above (i.e. RPP_whole_c3_aug).

On the basis of the whole KITTI-ROAD training images, we then train our RPP network. The network is initialized from those trained on the cityscape. As mentioned above, we make crops of raw images, add road edge labels, and carry out data augmentation in order to address small training sample problems. Two different crop sizes of 240×240 and 240×720 experiment are implemented for comparison. On account of the large size of the 112-layer RPP network and memory constraints, the batch size of training dataset is respectively set to 8 and 4. The influence of crop sizes on the RPP is shown in Fig. 3. Apparently, larger crops tend to get better results.

Due to our limited memory resource, we at most utilize a larger size crops of 240×720 to conduct all our RPP experiments on KITTI-ROAD dataset. The RPP road detection results, reported by the test server, are shown in Table 3, where FPR indicates false positive rate and FNR denotes false negative rate. The experimental results demonstrate that the proposed RPP achieves a highly competitive segmentation performance, even the best results before February 6, 2017. Now its latest rankings on KITTI Vision Benchmark Suite[1] are given as follows: UM_ROAD-2/62, UMM_ROAD-5/60, UU_ROAD-2/60, and URBAN_ROAD-3/60, as shown in Tables 4, 5, 6 and 7. All these results were also reported by the test server. Note that such a remarkable performance is achieved based on small training images but so large models. Actually, some drivable roadways contained in the test data set

[1] http://www.cvlibs.net/datasets/kitti/eval_road.php

Table 3 The road detection results of RPP_whole_c3_aug

KITTI (%)	MaxF (%)	AP (%)	PRE (%)	REC (%)	FPR (%)	FNR (%)
UM_ROAD	96.04	89.77	95.61	96.48	2.02	3.52
UMM_ROAD	97.03	92.36	96.36	97.70	4.06	2.30
UU_ROAD	95.47	88.74	95.16	95.77	1.59	4.23
URBAN_ROAD	96.36	90.36	95.85	96.87	2.31	3.13

Table 4 Road estimation evaluation on UM_ROAD

Method	MaxF (%)	AP (%)	PRE (%)	REC (%)	FPR (%)	FNR (%)
UNV	96.69	92.41	97.38	96.01	1.18	3.99
RPP	96.04	89.77	95.61	96.48	2.02	3.52
SAIT	95.65	90.17	96.06	95.25	1.78	4.75
TuSimple	95.64	93.50	95.46	95.83	2.08	4.17
SAIT	95.09	89.58	95.41	94.76	2.08	5.24

Table 5 Road estimation evaluation on UMM_ROAD

Method	MaxF (%)	AP (%)	PRE (%)	REC (%)	FPR (%)	FNR (%)
TuSimple	97.62	95.53	97.41	97.82	2.86	2.18
UNV	97.34	94.23	97.52	97.16	2.71	2.84
SAIT	97.31	92.68	96.71	97.91	3.66	2.09
SAIT	97.15	93.34	97.44	96.86	2.79	3.14
RPP	97.03	92.36	96.36	97.70	4.06	2.30

Table 6 Road estimation evaluation on UU_ROAD

Method	MaxF (%)	AP (%)	PRE (%)	REC (%)	FPR (%)	FNR (%)
UNV	95.71	90.32	95.13	96.30	1.61	3.70
RPP	95.47	88.74	95.16	95.77	1.59	4.23
TuSimple	95.17	92.73	95.97	94.39	1.29	5.61
SAIT	95.06	87.59	93.89	96.25	2.04	3.75
SAIT	95.01	88.58	94.98	95.04	1.64	4.96

Table 7 Road estimation evaluation on URBAN_ROAD

Method	MaxF (%)	AP (%)	PRE (%)	REC (%)	FPR (%)	FNR (%)
UNV	96.74	92.39	96.88	96.60	1.71	3.40
TuSimple	96.41	93.88	96.44	96.37	1.96	3.63
RPP	96.36	90.36	95.85	96.87	2.31	3.13
SAIT	96.27	90.34	95.83	96.72	2.32	3.28
SAIT	96.02	90.72	96.24	95.79	2.06	4.21

Fig. 4 The training loss and learning rate in relation to iterations

look too tiny to surely recognize even for a human. But our RPP is able to reliably detect them. We do believe that the RPP performance could further be improved if more training samples are added, although there are also a few images with a large detection error due to small training image problem.

From Tables 4, 5, 6 and 7, we can find out the most important metric is MaxF which the ranking is arranged according to. Though we have got the 2nd place under MaxF in UM_ROAD, we are the best under REC 96.48% and FNR 3.52%. And in the URBAN_ROAD task, RPP's FNR 3.13% is the best too. The algorithm UNV is the best on UM_ROAD and UU_ROAD. It surpasses RPP by a slight advantage of 0.65 and 0.24% under MaxF. We are the fifth in UMM_ROAD with 0.59% disadvantage than the best algorithm TuSimple. On task URBAN_ROAD, RPP ranks 3rd place with 0.28% lower than the first algorithm UNV. It is worth to mention that other methods listed in Tables 4, 5, 6 and 7 are not publicly available except for the evaluation data, so we cannot discuss more on that. The training loss and learning rate are shown in Fig. 4. Because of the absence of ground truth of test set, we do not plot testing loss in this figure.

Conclusion and Perspectives

In this paper, we propose a powerful 112-layer RPP model through the incorporation of residual connections and pyramid pooling into a fully convolutional neural network framework. For small training sample problems such as the KITTI-ROAD detection, we present a new strategy through an addition of road edge labels and data augmentation. Based on such a benchmark, the experimental results show that our RPP achieves highly competitive results. RPP wins second places on both unmarked and marked road tasks, fifth place in multiple marked lane tasks, and third place in combination task, compared to the other dozens of models.

Our work suggests adding more labels and appropriate data augmentation can indeed help dealing with the problem of small training sample. Moreover, using larger cropping sizes and/or combining more global information also tends to improve the road segmentation accuracy. When discarding the demanding computing and memory resources required by such large-scale networks like RPP, merely utilizing raw images (instead of any cropping operations) and much larger batch sizes could further increase the road detection accuracy. There are additional perspectives that may further evolve our work, such as testing with just the data augmentation but not the new labels, and to assist implementing a real-time application, we need to do more explorations on the reduction of inference time, while preventing a significant drop of the prediction accuracy.

Acknowledgments The authors are grateful to the reviewers for their valuable comments that considerably contributed to improving this paper.

Funding Information This work was supported in part by the National Science Foundation of China (NSFC) under Grant Nos. 91420106, 90820305, and 60775040, and by research fund of Tsinghua University - Tencent Joint Laboratory for Internet Innovation Technology.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Informed Consent Informed consent was not required as no humans or animals were involved.

Human and Animal Rights This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Alvarez J, Gevers T, LeCun Y, Lopez A. Road scene segmentation from a single image. Computer Vision–ECCV. 2012;2012:376–389.
2. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected CRFS. arXiv:1412.7062. 2014.
3. Ding C, Choi J, Tao D, Davis LS. Multi-directional multi-level dual-cross patterns for robust face recognition. IEEE Trans Pattern Anal Mach Intell. 2016;38(3):518–531.
4. Fang L, Wang X. Lane boundary detection algorithm based on vector fuzzy connectedness. Cogn Comput. 2017:1–12.
5. Fritsch J, Kuhnl T, Geiger A. A new performance measure and evaluation benchmark for road detection algorithms. In: 2013 16th international IEEE conference on intelligent transportation systems-(ITSC). Piscataway: IEEE; 2013. p. 1693–1700.

6. Goldman DB. Vignette and exposure calibration and compensation. IEEE Trans Pattern Anal Mach Intell. 2010;32(12):2276–2288.
7. Goodfellow IJ, Warde-Farley D, Lamblin P, Dumoulin V, Mirza M, Pascanu R, Bergstra J, Bastien F, Bengio Y. Pylearn2: a machine learning research library. arXiv:1308.4214. 2013.
8. Goodfellow IJ, Warde-Farley D, Mirza M, Courville A, Bengio Y. Maxout networks. arXiv:1302.4389. 2013.
9. Guo C, Mita S, McAllester D. Stereovision-based road boundary detection for intelligent vehicles in challenging scenarios. In: IROS 2009. IEEE/RSJ international conference on intelligent robots and systems, 2009. Piscataway: IEEE; 2009. p. 1723–1728.
10. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.
11. Huang G, Liu Z, Weinberger KQ, van der Maaten L. Densely connected convolutional networks. arXiv:1608.06993. 2016.
12. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167. 2015.
13. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: convolutional architecture for fast feature embedding. arXiv:1408.5093. 2014.
14. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–1105.
15. Laddha A, Kocamaz MK, Navarro-Serment LE, Hebert M. Map-supervised road detection. In: Intelligent vehicles symposium (IV), 2016 IEEE. Piscataway: IEEE; 2016. p. 118–123.
16. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 3431–3440.
17. Meftah B, Lézoray O, Benyettou A. Novel approach using echo state networks for microscopic cellular image segmentation. Cogn Comput. 2016;8(2):237–245.
18. Mendes CCT, Frémont V, Wolf DF. Exploiting fully convolutional neural networks for fast road detection. In: 2016 IEEE international conference on Robotics and automation (ICRA). Piscataway: IEEE; 2016. p. 3174–3179.
19. Neto AM, Victorino AC, Fantoni I, Ferreira JV. Real-time estimation of drivable image area based on monocular vision. In: Intelligent vehicles symposium (IV), 2013 IEEE. Piscataway: IEEE; 2013. p. 63–68.
20. Oliveira GL, Burgard W, Brox T. Efficient deep methods for monocular road segmentation. In: IEEE/RSJ international conference on intelligent robots and systems (IROS 2016); 2016.
21. Ouyang W, Wang X, Zhang C, Yang X. Factors in finetuning deep model for object detection with long-tail distribution. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 864–873.
22. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(4):640–651.
23. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014.
24. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 1–9.
25. Wang B, Frémont V, Rodríguez SA. Color-based road detection and its evaluation on the KITTI road benchmark. In: Intelligent vehicles symposium proceedings, 2014 IEEE. Piscataway: IEEE; 2014. p. 31–36.
26. Wijesoma WS, Kodagoda KS, Balasuriya AP. Road-boundary detection and tracking using ladar sensing. IEEE Trans Rob Autom. 2004;20(3):456–464.
27. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, et al. Google's neural machine translation system: bridging the gap between human and machine translation. arXiv:1609.08144. 2016.
28. Xie J, Yu L, Zhu L, Chen X. Semantic image segmentation method with multiple adjacency trees and multiscale features. Cogn Comput. 2017;9(2):168–179.
29. Xiong W, Droppo J, Huang X, Seide F, Seltzer M, Stolcke A, Yu D, Zweig G. Achieving human parity in conversational speech recognition. arXiv:1610.05256. 2016.
30. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122. 2015.
31. Zeng X, Ouyang W, Yang B, Yan J, Wang X. Gated bi-directional CNN for object detection. In: European conference on computer vision. Berlin: Springer; 2016. p. 354–369.
32. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. arXiv:1612.01105. 2016.