Road Detection and Centerline Extraction Via Deep Recurrent Convolutional Neural Network U-Net

Article *in* IEEE Transactions on Geoscience and Remote Sensing · May 2019

DOI: 10.1109/TGRS.2019.2912301

6 authors, including:

Xiaofei Yang
Harbin Institute of Technology
19 PUBLICATIONS **204** CITATIONS
SEE PROFILE

Xutao Li
Nanyang Technological University
74 PUBLICATIONS **932** CITATIONS
SEE PROFILE

Raymond Y. K. Lau
City University of Hong Kong
101 PUBLICATIONS **4,266** CITATIONS
SEE PROFILE

Xiaofeng Zhang
Zhejiang Normal University
78 PUBLICATIONS **730** CITATIONS
SEE PROFILE

Road Detection and Centerline Extraction via Deep Recurrent Convolutional Neural Network U-Net

Xiaofei Yang[1,2], Xutao Li[1,2,+], Yunming Ye[1,2,+], Raymond Y. K. Lau[3], and Xiaofeng Zhang[1,2], and Xiaohui Huang[4]
[1]Harbin Institute of Technology, Shenzhen, 518055, China
[2]Shenzhen Key Laboratory of Internet Information Collaboration, Shenzhen, 518055, China
[3]City University of Hong Kong, Hong Kong
[4]School of Information Engineering East China Jiaotong University, China

Road information extraction based on aerial images is a critical task for many applications, and it has attracted considerable attentions from researchers in the field of remote sensing. The problem is mainly composed of two sub-tasks, namely road detection and centerline extraction. Most of the previous studies rely on multi-stage based learning methods to solve the problem. However, these approaches may suffer from the well-known problem of propagation errors. In this paper, we propose a novel deep learning model, recurrent convolution neural network U-Net (RCNN-UNet), to tackle the aforementioned problem. Our proposed RCNN-UNet has three distinct advantages. First, the end-to-end deep learning scheme eliminates the propagation errors. Second, a carefully-designed RCNN unit is leveraged to build our deep learning architecture, which can better exploit the spatial context and the rich low-level visual features. Thereby, it alleviates the detection problems caused by noises, occlusions, and complex backgrounds of roads. Third, as the tasks of road detection and centerline extraction are strongly correlated, a multi-task learning scheme is designed so that two predictors can be simultaneously trained to improve both effectiveness and efficiency. Extensive experiments were carried out based on two publicly available benchmark datasets, and nine state-of-the-art baselines were used in a comparative evaluation. Our experimental results demonstrate the superiority of the proposed RCNN-UNet model for both the road detection and the centerline extraction tasks.

***Index Terms*—Recurrent Convolutional Neural Network, U-Net, Road Detection, Road Centerline Extraction.**

I. Introduction

AUTOMATIC road identification is one of the important research topics in the field of remote sensing. It facilitates a great number of applications such as vehicle navigation [1], urban planning [2], and geographical information system upgrading [3], and so on. However, because of the noisy data, occlusions, and complex background of remote sensing images, it is very challenging to effectively complete such a task. Generally speaking, road identification is composed of two subtasks, namely road detection and road centerline extraction. The former accounts for labeling all the road pixels, while the latter aims to find the centerline pixels. Most of the existing road detection methods rely on pixel-level segmentation or classification techniques [4]–[8]. However, due to the occlusions of cars, and the surrounding trees and shadows, the resulting road boundaries identified by existing methods are often not with high accuracy.

For road centerline extraction, morphological thinning algorithms [9]–[11] are widely utilized. Though the algorithms are simple and easy to implement, they often produce small spurs around the centerline, which lead to low accuracy. To address these shortcomings, more advanced models such as regression methods [6], [12] and nonmaximum suppression methods [13], [14] are developed. However, regression methods fail to extract good centerline around the road intersections and nonmaximum supression methods tend to produce wider centerlines than the original centerlines.

As a state-of-the-art machine learning technique, deep learning has recently witnessed great progress in conventional computer vision tasks such as image classification [15], [16], image segmentation [17]–[22], and so on. Recently, some researchers also has started to explore deep learning techniques for solving remote sensing tasks, and they have achieved promising results [4], [23]–[27]. Their pioneer work demonstrates the great potentials of applying deep learning techniques to analyze remote sensing images.

Due to the superiority in modeling complex non-linear relationships among variables, deep learning techniques have been applied to road identification tasks. For example, Mnih and Hinton proposed a path-based deep neural network [4], which incorporated an unsupervised pre-training step and a supervised post-processing phase to enhance the performance of the road detector. In [28], [29], the convolution neural network (CNN) model was applied to detect roads from remote sensing images. Recently, Zhang et al. have extended the U-Net [18] model by utilizing the deep residual unit as the basis for road detection [20]. However, these models were developed to solve one subtask of road identification, namely road detection, and they could not deal with the centerline extraction task. To alleviate the weakness of the aforementioned methods, Cheng et al. proposed a cascade CNN deep network to support the two subtasks simultaneously [28]. Though promising performance was achieved, their model still suffered from two drawbacks: (i) the plain CNN unit utilized in the network may fail to handle the occlusions of cars, and the surrounding trees and

+Corresponding authors: Yunming Ye (email: yeyunming@hit.edu.cn) and Xutao Li (email: lixutao@hit.edu.cn).

shadows of cars; (ii) the cascaded structure fails to leverage the mutual contributions between the two subtasks, that is, the knowledge learnt from the second task (e.g., centerline extraction) cannot be utilized to facilitate the preceding task (e.g., road detection).

In this paper, we propose a novel deep learning model, namely the recurrent convolution network U-Net (RCNN-UNet), to facilitate the road identification tasks. Specifically, we consider the road detection and centerline extraction tasks as pixel level image segmentation tasks, and we apply the U-Net deep learning framework to carry out these tasks. For our proposed model, we design a recurrent convolution neural network unit and incorporate it into the U-Net framework. Because the proposed model can better exploit the spatial context, the occlusion and noisy data problems related to car-based images can be alleviated. Moreover, instead of isolated learning from a two-stage framework which was utilized in previous research, the proposed RCNN-UNet adopts a multi-task learning scheme to simultaneously solve the road detection and centerline extraction tasks. As the road detection and centerline extraction tasks are simultaneously resolved by the proposed multi-task learning framework, domain knowledge learnt from one task can be directly utilized in the other task, which leads to more consistent and better performance. For clarity, we summarize the main contributions of our work as follows.

1) First, we design a novel recurrent convolution neural network unit that is incorporated into the U-Net framework for road extraction. Due to this sophisticated design, the spatial contexts can be better exploited to alleviate the problems of occlusion and noisy image data of cars.
2) Second, the multi-task learning scheme adopted in our framework differs from the existing methods in that we can solve the road detection and centerline extraction tasks at the same time. As the proposed RCNN-UNet model can exploit the domain knowledge learnt from the road detection task and the centerline extraction task simultaneously, more stable and better detection performance is achieved.
3) Third, extensive experiments were carried out based on real-world remote sensing image datasets, and our experimental results demonstrate that the proposed RCNN-UNet model outperforms other state-of-the-art methods in both road detection and road centerline extraction tasks.

The rest of the paper is organized as follows. Section II briefly reviews the related studies. In Section III, we first highlight the architecture, and the basis unit of the proposed RCNN-UNet model, and then we discuss the loss functions of the proposed deep network. Section VI reports our experimental results. Finally, we offer concluding remarks and discuss the future directions of our research work.

II. Related Work

A. Conventional Methods for Road Extraction

In this subsection, we review the classical methods for road detection and centerline extraction, respectively.

Road Detection. Most of the existing road detection methods are based on pixel-level segmentation or classification approach. For example, Song and Civco [30] developed a two-phase model for road detection. First, the Support Vector Machine (SVM) was applied to classify the pixels into road group (pixels with road features) and non-road group (pixels without road features). Second, the road group was further refined by a segmentation algorithm to produce the road regions. Zhang and Couloigner [31] proposed an integrated approach, which first segmented the given images into clusters, and then identified the road clusters by a fuzzy logic classifier that used angular texture signature features. In 2011, Yuan et al. [5] developed the locally excitatory and globally inhibitory oscillator networks (LEGION) and adopted a three-stage method, namely a segmentation stage, a medial axis points selection stage, and a road grouping stage, for road detection. Similarly, an multistage road detector framework was proposed by Das et al. [32]. In their proposed framework, two salient features of road, namely distinct spectral satellite contrast and locally linear trajectory, were exploited for road detection. Wegner et al. [33] developed a road detector that was underpinned by a higher-order Conditional Random Field (CRF) model. In the model, the road prior was represented by the higher-order cliques that connected sets of super-pixels along straight line segments, which could be nicely integrated into the CRF classifier for road detection.

Centerline Extraction. Identifying road centerline directly from an image is very difficult. Hence, the task is often conducted by first detecting the road, and then extracting the corresponding centerline. For example, an integrated road centerline extraction method was proposed by Gamba et al. [34]. For their proposed method, road elements were first extracted by an adaptive direction filtering procedure, and then the centerline was identified through these road elements. Huang and Zhang [9] exploited the multiscale structure features and SVM to identify the roads and extract the corresponding centerlines by using a morphological thinning algorithm. In [7], Unsalan et al. designed an automated system for centerline extraction. Their system was composed of three key modules, namely probabilistic road center detector, road shape extractor, and graph-theory-based road network builder. In [6], Miap et al. first found potential road segments with shape and spectral features, and then they applied the multivariate adaptive regression splines for centerline identification. Similarly, Shi et al. [12] obtained road segments based on an adaptive neighborhood method. Subsequently, kernel smoothing regression was applied to centerline identification. In 2014, Hu et al. [35] proposed a three-step method for road centerline extraction. First, based on multiple features, an adaptive mean-shift algorithm was applied to detect the center points of roads. Second, stick tensor voting was employed to enhance the salient linear features. Finally, the road centerlines were identified by a weighted Hough transform method. Sironi et al. [14] reformulated the centerline extraction task as a regression problem, and they developed a scale-space regressor to deal with the regression problem. Finally, based on the regression results, the centerlines were extracted by using the non-maximum suppression algorithm. In spite of the promis-

ing detection performance achieved by the aforementioned methods, these methods tended to produce discontinuities and topological mistakes. To overcome such a problem, Sironi et al. proposed the manifold approach for more spatially and geometrically consistent centerline extraction [36].

In summary, most of classical methods for road identification have two deficiencies: (i) most of these methods are based on the multi-stage approach, where the errors generated in the preceding stage will affect the detection task at the later stage. Accordingly, an end-to-end road extraction framework is desirable and worth further investigation; (ii) road detection and centerline extraction are often treated as independent tasks. However, these two tasks should be considered holistically such that the road detection result can be leveraged to improve the performance of centerline extraction, which results in more accurate overall road region identification.

B. Deep Learning Methods for Road Identification

As noted above, road identification is essentially an image segmentation task. Hence, we first briefly review deep learning based image segmentation algorithms, and then we discuss a few pioneering work on road identification.

Image Segmentation Methods. After winning the competition in ILSVRC 2012, convolution neural network (CNN) opens the door for deep learning-based approach for intelligent image analysis [15]. Deep CNN framework has also been widely adopted for image segmentation. For example, in [17], Long et al. proposed a fully connected convolutional network, called FCN, to deal with the semantic image segmentation problem. FCN achieved better performance than conventional segmentation methods. To solve the biomedical imaging partition problem, Ronneberger et al. [18] proposed a sophisticated symmetric "U" type FCN network, termed U-Net which comprised 18 layers. The first 8 layers were treated as a CNN-based encoder, accounting for summarizing features of the input image into small feature maps, while the last 8 layers were taken as a CNN-based decoder that aimed to produce the pixel-level segmentation based on the encoded data result. In between the encoder and the decoder, a 2-layer bridge was embedded. Due to the special "U" type, U-Net often achieves better performance than FCN. Badrinarayanan et al. proposed a new deep network model (SegNet) [19], which preserved symmetric encoder and decoder structure that is similar to U-Net. Different from the U-Net, SegNet memorized the indices of max pooling operators in the encoder layers, and then applied them to simplify the upsampling procedure in the decoder layers. Recently, Li et al. proposed a novel hybrid densely connected U-Net model, named H-DenseUNet [21]. The model replaced the CNN layers of U-Net by the densely connected layers. By doing so, richer spatial contextual information was expected to be preserved.

Road Identification Methods. Recently, some pioneering work that explores deep learning techniques for road identification has been conducted. For instance, inspired by the deep residual network, Zhang el al. [20] extended U-Net by introducing the short-cut connections between the CNN layers. The model was named as ResUNet and applied to road detection. In 2017, Cheng et al. developed a cascaded deep CNN approach to road identification [27]. Their method was based on a two-stage deep learning approach, where the first stage dealt with the road detection task, and the second stage focused on centerline extraction by utilizing the domain knowledge learnt from the first stage. Mattyus et al. proposed an approach called DeepRoadMapper [37]. For their proposed method, a CNN-based segmentation algorithm was applied to produce a coarse road extraction result. Then, a road network graph was constructed by using the binary thresholding and the morphological thinining methods. DeepRoadMapper can achieve promising performance for the images without occlusion, ambiguous topology or complex topology. However, such ideal case is seldom obtained in reality. Mosinska el al. [38] proposed a new loss function, called topology-aware loss, for delineation of curvilinear structures. Instead of pixel-wise loss function, e.g., crossentropy, the new loss function takes the topology structure information into account. Equipped with the topology-aware loss, U-Net can achieve better road extraction performance. Mattyus and Urtasun proposed an Matching Adversarial Network (MatAN) [39] to tackle the image segmentation problem, which could be applied to the road extraction task. In [40], Bastani et al. developed a CNN-based iterative graph construction method called Roadtracer, which could directly identify road networks from aerial images without complicated post-processing. The method started with a seeding point known to be on the road, and iteratively walked to new points which were potential road pixels within a fixed window. During the iterative learning process, a CNN-based decision function was built to decide to walk or stop, and the angle to walk. By doing so, a road network graph could be constructed. However, this method has two drawbacks: (1) it tends to fail at the road junction positions; (2) high curvature roads and long-straight roads are quite likely to be ignored. The reason is that the road graph construction is based a sequential decisions made by the CNN. Once the CNN makes a mistake, the Roadtracer will produce some wrong road segments or fail to identify some road segments labeled as the ground-truths.

Though many deep learning approaches have been developed for image segmentation, and some of them have been extended for road identification, they still suffer from two deficiencies: (i) existing methods have not fully exploited the spatial contexts of road images; classical CNN-based unit may fail to handle the occlusions of cars and the noises caused by the surrounding trees or car shadows; (ii) most of the existing methods do not simultaneously solve the road detection and centerline extraction problems such that domain knowledge learnt from each task can be utilized to bootstrap the performance of another task. In this paper, we aim to tackle the two aforementioned problems by designing a novel recurrent convolution network U-Net to enhance the performance of road identification.

III. THE PROPOSED METHODS

Our proposed methods comprise three key elements, namely the recurrent neural network (RCNN) unit, the RCNN-UNet

architecture, and the multi-task learning scheme to facilitate the road detection task and the centerline extraction task simultaneously. In this section, we will illustrate the computational details of each of these key elements.

A. RCNN unit

As discussed in related work B, the CNN model outperforms the conventional methods for road identification. CNN employs a spatial kernel function and moves it gradually to transform the visual features into feature maps. By doing so, the spatial context is transformed to higher level concepts. However, this may lead to a loss of the detailed spatial characteristics, which are important for addressing the noises and occlusions issues in road identification. Different from CNN, the RCNN layers can better exploit the spatial context, and hence lead to a richer visual feature abstraction [41]. Specially, the structure of a RCNN block example is depicted as in Fig. 1(b). As can be seen, given an image or feature map as an input I, the RCNN block first utilizes a 3×3 convolution layer to generate the first-level feature maps. After a batch normalization (BN) layer, and a ReLU activation layer, the same 3×3 convolution operation is applied to produce new feature maps. The second-level outputs $X(t-1)$ are then obtained by summing these new feature maps with the first-level feature maps. Through similar operations (i.e., a BN layer, a ReLu layer, and the same 3×3 convolution layer), other new features are produced, which will be added up with the first-level feature maps to produce the third-level outputs $X(t)$. Comparing the RCNN block with a plain CNN unit in Fig. 1(a), we can observe the advantages of the RCNN block. Without the two summation operators in Fig. 1(b), the RCNN block is in fact the same as the classical CNN unit as shown in Fig. 1(a). By using the two summation operators, RCNN can preserve the low-level detailed spatial characteristics during the multiple level feature transformation process. Moreover, the convolution layers applied in the first, second, and third levels share the same parameters. Hence, the parameter size of a RCNN unit is basically similar to that of a CNN unit. Also, due to the reuse of the same convolution layer, the block is named as recurrent convolutional neural network unit, i.e., RCNN unit. Formally, each RCNN unit can be expressed by the following transformation process:

$$\begin{aligned} X(t) &= F(f[B(X(t-1))]) + F(I), \\ X(t-1) &= F(f[B(F(I))] + F(I). \end{aligned} \tag{1}$$

where t is the time step, I and $X(t-1)$ denote the raw input and the second-level outputs, $F(\cdot)$ is the convolutional operation, $f(\cdot)$ is the activation function, $B(\cdot)$ denotes the batch normalization function (BN), and $X(t)$ represents the third-level outputs. We note in the example of Fig. 1(b) a three-layer convolution RCNN unit is built. For real-world applications, the number of convolution layers can be determined by users.

To validate the superiority of RCNN unit over the plain CNN unit, we visualize some randomly chosen feature maps in each layer by the two methods. The results are shown as Fig. 2. We can see from the Figure that the RCNN unit indeed

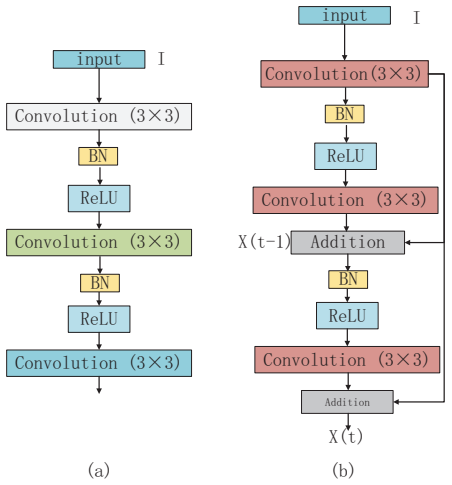


Fig. 1. Building blocks of neural networks. (a) The plain neural network in U-Net, and (b) the recurrent convolutional neural network unit used in the proposed RCNN-UNet. We note the in the figure the same color blocks denote that they share the same weights.

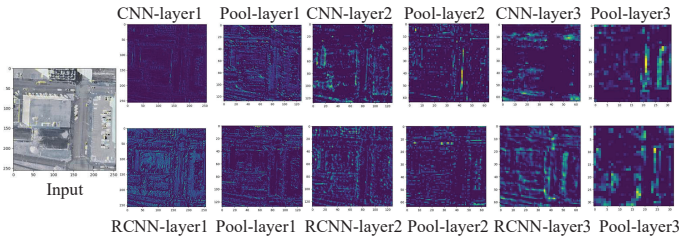


Fig. 2. The visualization of RCNN unit and plain CNN unit.

delivers better features, which consist of more details and more consistent road regions with the input image, especially in layer1 and layer3. This is attributed to the special recurrent architecture in the RCNN unit.

B. Deep RCNN-UNet Architecture

In this subsection, we illustrate how to build a RCNN-UNet model by using the RCNN unit described above. As shown in Fig. 3, we build a 7-level architecture which similar to the U-Net model; it is also a "U" type network. The proposed architecture consists of three main elements, namely an encoder, a bridge, and a decoder. The encoder accounts for abstracting the input image into low-resolution feature maps, which represent some high-level concepts. The bridge conveys the abstraction signals to the decoder. By transforming the signals, the decoder produces the road regions.

The three elements of the proposed architecture are all built upon the RCNN units, as shown as the red blocks in Fig. 3. It can be observed that either the encoder part or the decoder part is composed of three RCNN units, which are connected via the bridge. The RCNN units in the encoder are connected by max pooling operators, while the ones in the decoder are connected by the upsampling operators. As denoted by three horizontal arrows, the output of each RCNN unit is directly concatenated with feature maps of the corresponding layer in the decoder. And for clarity, we summarize the details of the

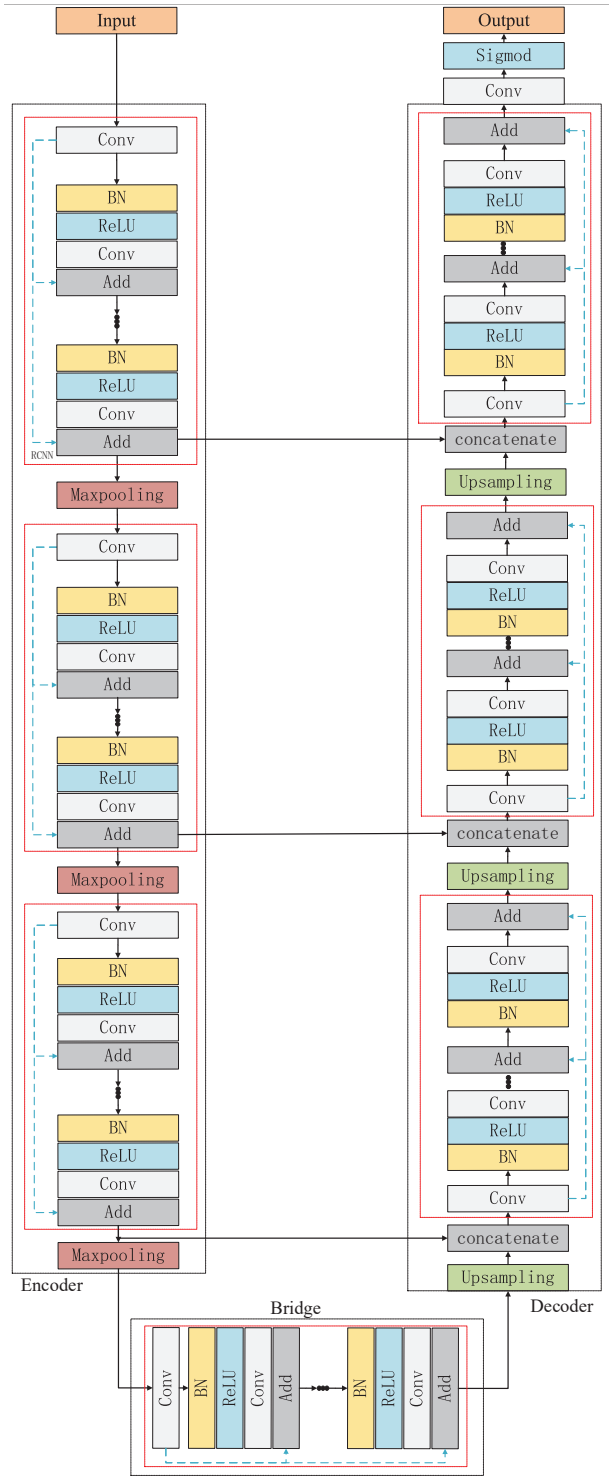


Fig. 3. The architecture of the proposed deep RCNN-UNet. Each red box represents a RCNN unit, which is composed of five convolutional layers. In each red box, we only depict two convolutional layers and the other three ones are abbreviated to save space.

parameters pertaining to the network in Table I, which include the number of layers, the size and number of filters, and the stride and the size of output in each layer. As shown in Table I, each RCNN unit of the proposed architecture is composed of five convolution layers. Note that each convolution layer is appended by a BN layer and a ReLU activation function, which are abbreviated in Table I. As a whole, there are 36 convolution layers, 3 max pooling layers, and 3 upsampling layers.

The proposed RCNN-UNet has two distinct advantages over the conventional UNet: (1) The RCNN unit will preserve more detailed spatial characteristics during the feature transformation process, which are critical for tackling the noise and occlusion issues in road identification. The advantage is because our special design in RCNN unit can enlarge the effective receptive fields to better capture the spatial context. As the RCNN unit consists of five convolutional layers, after each convolutional and addition operators, the effective receptive fields will be enlarged gradually. However, the effective receptive fields of a plain CNN unit are fixed. (2) Though there are five convolutional layers in each RCNN unit, these five layers share the same weight parameters. As a result, the recurrent architecture increases the network depth while keep the number of adjustable parameters constant by weight sharing. And the complexity of the proposed model is reduced, alleviating the problem of model overfitting. Note that the number of convolutional layers in a RCNN Unit has been empirically obtained.

C. Multi-task Learning

As noted in the introduction part, the road identification task consists of two sub-tasks, namely road detection and centerline extraction. There are two ways to tackle the road identification task. One way is to follow the existing approach of first detecting road regions with the designed RCNN-UNet, and then identify the centerlines by applying centerline extraction methods (e.g., morphological thinning algorithms) to the results obtained at the first stage. However, the two-stage approach cannot effectively utilize the domain knowledge learnt from each sub-task. A more desirable approach is to leverage a multi-task learning scheme to simultaneously train two predictors with respect to the two subtasks. Next, we will illustrate the RCNN-UNet1, RCNN-UNet2, and RCNN-UNet3 unit, where RCNN-UNet1 is the instantiation of the RCNN-UNet model discussed above, while RCNN-UNet2 and RCNN-UNet3 are two instantiations of the RCNN-UNet model by using the multi-task learning scheme.

RCNN-UNet1 Method.

As shown in Fig. 4(a), the method employs the Deep RCNN-UNet architecture discussed before for road detection. In this case, we consider the road detection problem as a binary semantic segmentation problem, where the road regions are taken as foregrounds, and the other parts are treated as backgrounds. Hence, RCNN-UNet1 is still a 36-layer network, and the last layer is a sigmoid activation function for each pixel i. The sigmoid activation function converts the corresponding feature maps into a probability value p_i, which indicates the

probability that this pixel is a road pixel. Let y_i denote the ground-truth of the (i)-th pixel, where $y_i = 1$ if it corresponds to a road part, and $y_i = 0$ otherwise. Then, the following binary cross entropy function is adopted as the loss function:

$$L_{road} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i + (1 - y_i)\log(1 - p_i) \quad (2)$$

where N is the number of pixels of the input image. Given a set of training images with the ground truth labels y, we can fit RCNN-UNet1 by minimizing the loss function with the standard stochastic gradient descent (SGD)method.

RCNN-UNet2 Method.

Different from RCNN-UNet1, RCNN-UNet2 trains two predictors at the same time, where one for the road detection task and the another one for the centerline extraction task, as shown as in Fig. 4(b). Since the two tasks are strongly correlated, we thus let them share the first 35-layer of the RCNN-UNet architecture. In the last layer, two branch of sigmoid activation functions are used for road detection and centerline prediction, respectively. For the road detection task, we leverage the same loss function L_{road} in Eq. (2). Similarly, a binary cross entropy loss function can be defined for centerline extraction as follows:

$$L_{cen} = -\frac{1}{N}\sum_{i=1}^{N} z_i \log q_i + (1 - z_i)\log(1 - q_i) \quad (3)$$

where z_i denotes the ground truth of the i-th pixel, i.e., $z_i = 1$ if the i-th pixel belongs to the centerline, and $z_i = 0$ otherwise; q_i represents the probability that the i-th pixel belongs to the centerline. Hence, the overall loss function of RCNN-UNet2 is defined as follows:

$$Loss = W_1 L_{road} + W_2 L_{cen} \quad (4)$$

where W_1 and W_2 are the weighting parameters, and we set $W_1 = W_2 = 1$ in our experiments.

RCNN-UNet3 Method.

This is our second instantiation of the multi-task learning scheme. The main idea of this method is inspired by the conventional two-stage centerline prediction method. First, we try to detect roads. Then, we extract the centerlines from the road regions detected before. As shown in Fig. 4(c), RCNN-UNet3 also trains two predictors simultaneously as that of RCNN-UNet2. The upper predictor still accounts for detecting the road regions. Then, a plain convolution layer is applied to the detected road regions to extract the centerlines. For this instantiation, we also utilize a binary cross entropy loss function as follows:

$$L_{new} = -\frac{1}{N}\sum_{i=1}^{N} m_i \log k_i + (1 - m_i)\log(1 - k_i) \quad (5)$$

where m_i denotes the ground truth of the i-th pixel, i.e., $m_i = 1$ if the i-th pixel belongs to the centerline and $m_i = 0$ otherwise; k_i represents the probability that the i-th pixel belongs to the centerline. The produced feature maps in this part are combined with the bottom feature maps (Fig. 4(c) to predict the final centerline. To sum up, the loss function of our RCNN-UNet3 is defined as follows:

$$Loss = W_1 L_{road} + W_2 L_{cen} + W_3 L_{new} \quad (6)$$

where $W_1 = W_2 = W_3 = 1$ was used in our experiments.

TABLE I
THE NETWORK STRUCTURE OF DEEP RCNN-UNET.

Name	Unit level	Layers	Filter	Stride	Output size
Input					256×256×3
Encoder	Level 1	Conv1	3×3 /64	1	256×256×64
		Conv2	3×3 /64	1	256×256×64
		Conv3	3×3 /64	1	256×256×64
		Conv4	3×3 /64	1	256×256×64
		Conv5	3×3 /64	1	256×256×64
		Pool1	2×2	2	128×128×64
	Level 2	Conv6	3×3 /128	1	128×128×128
		Conv7	3×3 /128	1	128×128×128
		Conv7	3×3 /128	1	128×128×128
		Conv9	3×3 /128	1	128×128×128
		Conv10	3×3 /128	1	128×128×128
		Pool2	2×2	2	64×64×128
	Level 3	Conv11	3×3 /256	1	64×64×256
		Conv12	3×3 /256	1	64×64×256
		Conv13	3×3 /256	1	64×64×256
		Conv14	3×3 /256	1	64×64×256
		Conv15	3×3 /256	1	64×64×256
		Pool3	2×2	2	32×32×256
Bridge	Level 4	Conv16	3×3 /512	1	32×32×512
		Conv17	3×3 /512	1	32×32×512
		Conv18	3×3 /512	1	32×32×512
		Conv19	3×3 /512	1	32×32×512
		Conv20	3×3 /512	1	32×32×512
Decoder	Level 5	upsampling	2×2	2	64×64×256
		Conv21	3×3 /256	1	64×64×256
		Conv22	3×3 /256	1	64×64×256
		Conv23	3×3 /256	1	64×64×256
		Conv24	3×3 /256	1	64×64×256
		Conv25	3×3 /256	1	64×64×256
	Level 6	upsampling	2×2	2	128×128×128
		Conv26	3×3 /128	1	128×128×128
		Conv27	3×3 /128	1	128×128×128
		Conv28	3×3 /128	1	128×128×128
		Conv29	3×3 /128	1	128×128×128
		Conv30	3×3 /128	1	128×128×128
	Level 7	upsampling	2×2	2	256×256×64
		Conv31	3×3 /64	1	256×256×64
		Conv32	3×3 /64	1	256×256×64
		Conv33	3×3 /64	1	256×256×64
		Conv34	3×3 /64	1	256×256×64
		Conv35	3×3 /64	1	256×256×64
Output		Conv36	1×1	1	256×256×1

IV. EXPERIMENTS

To demonstrate the effectiveness of the proposed RCNN-UNet models, we conducted extensive experiments based on two publicly available road identification datasets. As for a comparative evaluation, nine well-known methods were adopted as the baselines which include FCN [17], SegNet [19], U-Net [18], DenseUNet [21], ResUNet [20], Cascaded CNN [27], DeepRoadMapper [37], RoadCNN[1] [40], and

[1]The RoadCNN is the CNN based segmentation baseline developed in [40], which was named as Seg(ours). In this paper, we call it RoadCNN.

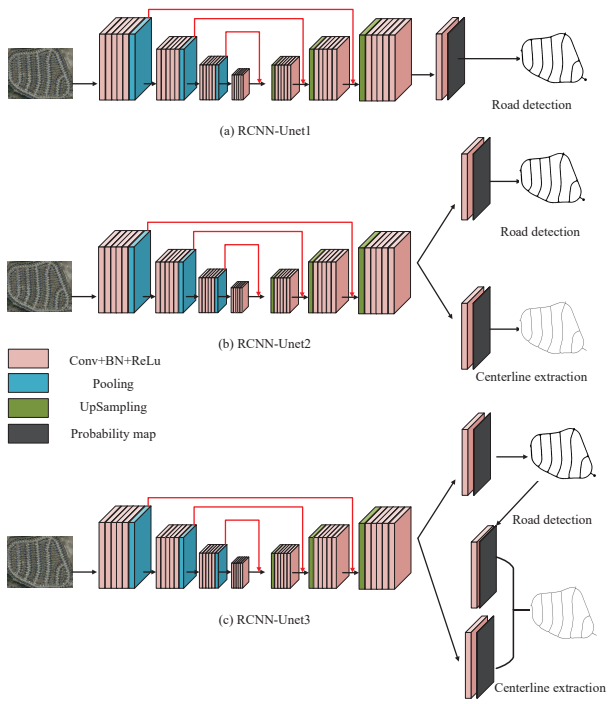


Fig. 4. Flowchart of the proposed deep RCNN-UNets. In (a), the RCNN-UNet model is only designed for road detection; In (b), the RCNN-UNet model is designed for road detection and road centerline extraction, simultaneously; In (c), the RCNN-UNet is designed based on the thinning layer for road detection and road centerline extraction.

RoadTracer [40]. The first five baselines are state-of-the-art image segmentation methods, and the reminders are state-of-the-art road identification methods.

A. Dataset

Cheng Dataset. We leverage the benchmark dataset built by Cheng et al. [27]. The dataset consists of 224 images which are all collected from Google Earth. Though the sizes of the 224 images vary, the resolution of each image is at least 600 $\times$ 600 pixels. Each pixel denotes 1.2 meters. The dataset covers different geographical regions including urban areas, sub-urban areas, and rural areas, and it captures a wide range of objects on the ground e.g., roads, rivers, various buildings, and so on. Most of the images contain complex backgrounds and occlusions caused by trees or buildings. This makes the road identification task extremely challenging. Following [27], 180 images were leveraged for training, 14 images for validation, and the reminder 30 images for testing.

Roadtracer Dataset. The Roadtracer dataset [40] is also adopted. The dataset is a large corpus of high-resolutional satellite images, which includes the road networks from the urban core of 40 cities in six countries. In each city, the center region about 24 sq km is covered. The dataset contains 300 images, and the resolution of each image is 4096×4096. Following [40], the images of 25 cities are utilized as training set and the ones of the remaining 15 cities as test set. We note that the dataset does not include the centerline ground truth.

B. Experiment Setting

Cheng Dataset. Since all the baselines and the proposed three methods were deep learning based methods, we adopted the same data augmentation scheme to enrich the training and the validation samples. Specifically, given a training image or a validation image, we first extracted 5 fixed-position patches (the 4 corner patches and the center patch), and then sampled another 15 random free position patches. The sizes of the patches were 256×256. Subsequently, the patches were flipped in horizontal and vertical directions, and then rotated with a step of 90 degree. By doing so, the number of training samples was increased from 180 to $180 \times 20 \times 2 \times 3$ (21,600), and the number of validation samples increased from 14 to $14 \times 20 \times 2 \times 3$ (1,680).

RoadTracer Dataset. For the dataset, we follow the settings in [40]. Given each 4096×4096 image, we divide it into 256 small images, whose sizes are all 256×256. As the centerline labels are not given, only the road extraction task can be tested on the dataset. The algorithms such as RCNN-UNet2 and RCNN-UNet3 cannot be applied on the dataset.

All the methods were trained based on a NVIDIA Titan 1080ti GPU. The minibatch size was 16, and the learning rate was 0.00001; the max epoch was set to 20.

C. Evaluation Metrics

Following [27], we employed four well-known metrics to evaluate the performance of the road detection task and the centerline extraction task, respectively. These metrics include completeness (COM), correctness (COR), quality (Q), and F1 score. The COM measures the portion of matched areas with respect to the ground truth reference map. The COR denotes the percentage of matched road areas (centerline areas) from among the road areas (centerline areas) detected by a computational method under our evaluation. The Q is an combined metric that takes into account both COM and COR. The F1 score is also called balanced F score, which is a harmonic average between COM and COR. Specifically, the four metrics are defined as follows.

$$\begin{aligned} COM &= \frac{TP}{TP+FN}, \\ COR &= \frac{TP}{TP+FP}, \\ Q &= \frac{TP}{TP+FN+FP}, \\ F1 &= \frac{2 \times COM \times COR}{COM+COR}. \end{aligned} \tag{7}$$

where TP denotes the true positive; FP denotes the false positive, and FN denotes the false negative. The larger an evaluation metric value is, the better performance of the method is.

As there exist deviations between manually labeled road map and the real road region, it is better to set up a tolerance threshold ρ. Specifically, if the areas in the predicted results are in a range of ρ pixels, they are considered as the matching areas. In this paper, we follow [4] to set the tolerance threshold to $\rho = 2$.

D. Comparative Evaluation on Road Detection

Cheng Dataset. To compare the performance of road detection, all the methods were evaluated based on the test samples for road detection. For a qualitative comparison, we show the results produced by all methods based on one example images depicted in Fig. 5. The quantitative comparisons are reported in Table II.

We can see from Fig. 5 that the proposed RCNN-UNet models generally perform better than the baseline methods. For FCN, we find that it misses road regions in many places i.e., the false negative (red) part is large, especially in the intersection and corner parts of roads, or the occlusion parts caused by trees and buildings. This can be attributed to two reasons: (i) in the FCN architecture, the segmentation images is upsampled from very low-resolution feature maps by only one deconvolutional operator, which is too coarse; (ii) FCN utilizes the plain CNN unit and it fails to take into account the rich spatial context. As for the baselines of U-Net type, namely SegNet, U-Net, DenseUNet, RoadCNN, and ResUNet, we observe that they outperform FCN. The reason is that the symmetric "U"-type architecture effectively constructs the segmentation image from low-resolution feature maps. Specifically, there are two key differences between the construction processes of FCN and those of the U-Net type baselines. The first key difference is that multiple upsampling layers are utilized by the U-Net type baselines, and the another difference is that multi-scale low-resolution feature maps produced in the encoding layers are fused into the upsampling results of the corresponding decoding layers of the U-Net type baselines. For cascaded CNN, apart from the centerline extraction, it is also a "U" type architecture, with the only difference in that multi-scale low-resolution feature maps are not used. We can observe that it also outperforms FCN. Compared to all the baseline methods, the proposed three RCNN-UNet methods achieve more promising performance. Though most of the baseline methods also utilize the U-Net like architecture, the basis unit i.e., RCNN unit of the proposed methods can better capture the spatial context than conventional CNN layer. This is the reason why the proposed RCNN-UNet methods outperform other baselines.

Table II presents the comparative quantitative evaluation measured in terms of COM, COR, Q, and F1. In the table, the best performance is denoted in bold, and the second best is marked with underlines. We observe from the table that the proposed three methods outperform baselines in terms of all the four metrics. None of the baseline methods can beat other baselines under all the four metrics. Among the three proposed methods, the RCNNUNet2 achieves more promising and robust performance because its result is either the best or the second best in terms of the three metrics. Furthermore, we find that RCNN-UNet1 performs better for the evaluation tasks in terms of COR; RCNN-UNet3 tends to yield better performance in terms of COM, whereas RCNN-UNet2 is more likely to achieve better performance in terms of Q. As RCNN-UNet1 was merely trained by using the road regions, it tended to achieve a higher precision (COR). Both RCNN-UNet2 and RCNN-UNet3 adopt the multi-task learning scheme, which detect road and extract centerlines simultaneously. As the centerlines extracted can help identify the road regions (areas around centerlines are quite likely to be roads), a better recall (COM) is achieved by these methods, especially by RCNN-UNet3. In addition, RCNN-UNet3 is better than other methods in terms of all the evaluation metrics. As DeepRoadMapper and RoadTracer both need the latitude and longitude information of each node as input, which is not available in Cheng dataset, their results are missing in Figs. 5 and Table II.

RoadTracer Dataset. Here we report the results of road extraction on RoadTracer dataset. Fig. 6 visualizes the results of different methods. We can see from the figure that RoadTracer performs the worst, which misses many ground-truth road segments and causes a few wrong detections. The conventional CNN based methods FCN and DeepRoadMapper deliver better results than RoadTracer. The approaches based on the U-Net architecture, namely U-Net, DenseUNet, SegNet, ResUNet, RoadCNN and RCNN-UNet outperform FCN and DeepRoadMapper. Among all the methods, the proposed RCNN-UNet yields the best results. This is attributed to our designed RCNN unit, which can better capture the spatial characteristics for road extraction. Table III summarizes the results evaluated by COM, COR, Q and F1 metrics. We can see from the table that RCNN-UNet1 performs the best, which is consistent with our visual judgement in the qualitative evaluation. We note that the proposed RCNN-UNet2 and RCNN-UNet3 cannot be applied on the dataset.

E. Comparison of Road Centerline Extraction

Cheng Dataset. In this subsection, we report the results of a comparative evaluation based on the centerline extraction task. For FCN, SegNet, U-Net, DenseUnet, ResUNet, and RCNN-UNet1, we applied the morphological thinning algorithm [42] to the road detection results so as to bootstrap centerline extraction. Fig. 7 visualizes the centerlines identified by different methods. From Fig. 7, we can see that FCN and U-Net produce spurs in the intersection points between roads. Moreover, a few discontinuities and false positive parts are observed in their identified centerlines. Though spurs are rarely seen, discontinuities and false positive parts still appear in the results produced by ResUNet. As for RCNN-UNet, several spurs appear at the interaction points. Although without spurs and discontinuities, Cascaded CNN tends to make mistakes at small road areas and intersection points. The centerlines identified by the proposed RCNN-UNet2 and RCNN-UNet3 are the most consistent results with real centerlines. The reason is that the two methods can solve road detection and centerline extraction tasks simultaneously under a multi-task learning scheme, whereas the prediction of one sub-task can bootstrap the performance of solving another sub-task.

Table IV summarizes the results evaluated by COM, COR, Q and F1 metrics. We can see from the table that RCNN-UNet2 and RCNN-UNet3 perform the best, which is consistent with our visual judgement in the qualitative evaluation. Among all the baseline methods, cascaded CNN is the best one. RCNN-UNet2 and RCNN-UNet3 outperform cascaded CNN by 1.89%, 1.44%, 0.023% and 1.95% in terms of COM, COR,

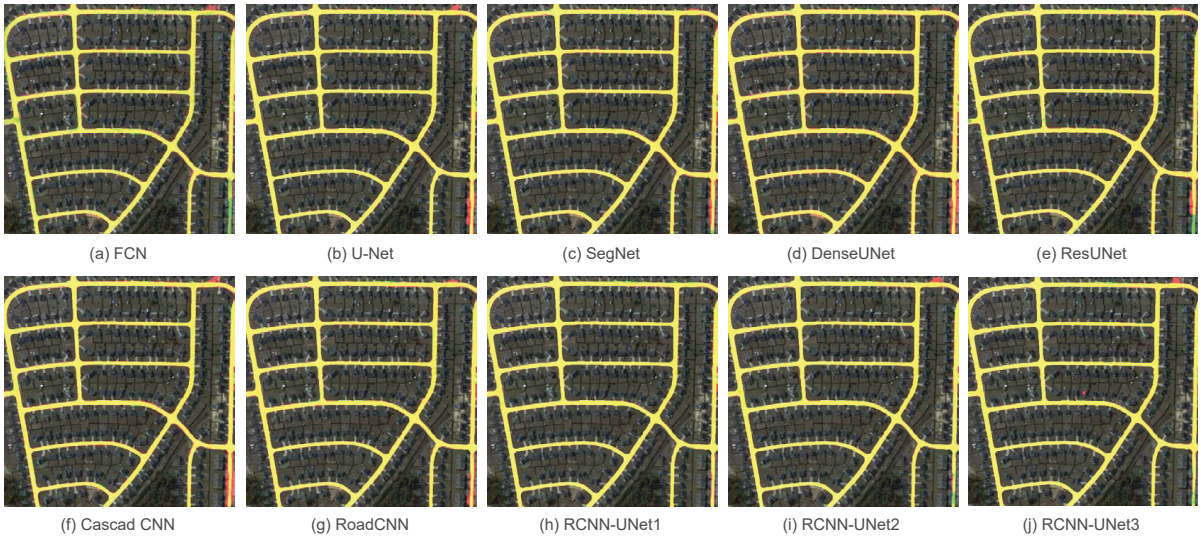


Fig. 5. Qualitative comparison of road detection results produced by different methods based on one example image from Cheng dataset. The yellow color denotes the true positive parts; the red color denotes the false positive parts; and the green color represents the false negative parts.

TABLE II
COMPARATIVE QUANTITATIVE EVALUATION AMONG DIFFERENT METHODS FOR ROAD DETECTION ON CHENG DATASET. IT SHOULD BE NOTED THAT THE RESULTS ARE THE AVERAGE PERFORMANCE OF ALL IMAGES IN TEST SET. A HIGHER VALUE INDICATES A BETTER PERFORMANCE.

Methods Name	COM	COR	Q	F1
FCN	0.7620±0.0390	0.8534±0.0407	0.8665±0.0658	0.8051±0.0392
U-Net	0.9516±0.0062	0.9093±0.0038	0.9105±0.0034	0.9300±0.0017
SegNet	0.9174±0.0081	0.9407±0.0091	0.8665±0.0150	0.9289±0.0085
DenseUNet	0.8751±0.0275	0.8917±0.0121	0.7937±0.0301	0.8833±0.0186
ResUNet	0.9324±0.0142	0.9524±0.0211	0.8934±0.0311	0.9423±0.0167
RoadCNN	0.9577±0.0659	0.9447±0.1329	0.8905±0.1061	0.9612±0.1302
Cascaded CNN	0.9377±0.0030	0.9247±0.0042	0.8705±0.0067	0.9312±0.0035
RCNN-UNet1	0.9627±0.0048	0.9683±0.0085	0.9336±0.0044	0.9655±0.0079
RCNN-UNet2	0.9679±0.0240	0.9638±0.0081	0.9339±0.0070	0.9658±0.0099
RCNN-UNet3	**0.9772**±0.0095	**0.9688**±0.0012	**0.9362**±0.0092	**0.9730**±0.0047

TABLE III
QUANTITATIVE COMPARATIVE EVALUATION AMONG DIFFERENT METHODS FOR ROAD DETECTION ON ROADTRACER DATASET. IT SHOULD BE NOTED THAT THE RESULTS ARE THE AVERAGE PERFORMANCE OF ALL IMAGES IN TEST SET. A HIGHER VALUE INDICATES A BETTER PERFORMANCE.

Methods Name	COM	COR	Q	F1
FCN	0.6397±0.0643	0.3113±0.1127	0.2629±0.0958	0.4057±0.1349
U-Net	0.6575±0.0624	0.3775±0.1283	0.3144±0.1025	0.4693±0.1259
SegNet	0.6548±0.0545	0.4359±0.1254	0.2721±0.0958	0.5147±0.1109
DenseUnet	0.6397±0.0709	0.3113±0.1305	0.2629±0.1049	0.4057±0.1388
ResUNet	0.6525±0.0710	0.3368±0.1359	0.2840±0.1111	0.4308±0.1451
Cascaded CNN	0.6583±0.0043	0.3861±0.0177	0.3205±0.0113	0.4757±0.0170
DeepRoadMapper	0.5461±0.0963	0.3663 ±0.1576	0.2757±0.1095	0.4212±0.1382
RoadCNN	0.6329±0.0659	0.4121±0.1329	0.3326±0.1061	0.4991±0.1302
RoadTracer	0.3664±0.0457	0.3259±0.0012	0.2063 ±0.0012	0.3414±0.0023
RCNN-UNet1	**0.6576**±0.0588	**0.4936**±0.1307	**0.3931**±0.1018	**0.5590**±0.1162

Q and F1. We note that as DeepRoadMapper and RoadTracer cannot detect centerlines, we do not compare with them here.

V. CONCLUSIONS

In this paper, we propose a novel RCNN-UNet deep learning model for road identification. The proposed model is essentially a "U" type image segmentation deep learning architecture. To alleviate the problems of noises, occlusions, and complex backgrounds often captured in road images, we successfully design a recurrent convolutional neural network (RCNN) unit, and incorporate it into the U-Net architecture. The distinct advantage of the proposed unit is that it can preserve the detailed low-level spatial characteristics. Since the prediction results pertaining to road detection and centerline extraction have the potential to bootstrap the prediction performance on one another, two multi-task learning schemes are developed. Extensive experiments were conducted based on real-world benchmark datasets, and our experimental results demonstrate that the proposed RCNN-UNet methods outperform state-of-the-art techniques for both the road detection and the centerline extraction tasks. The proposed RCNN-UNet is a

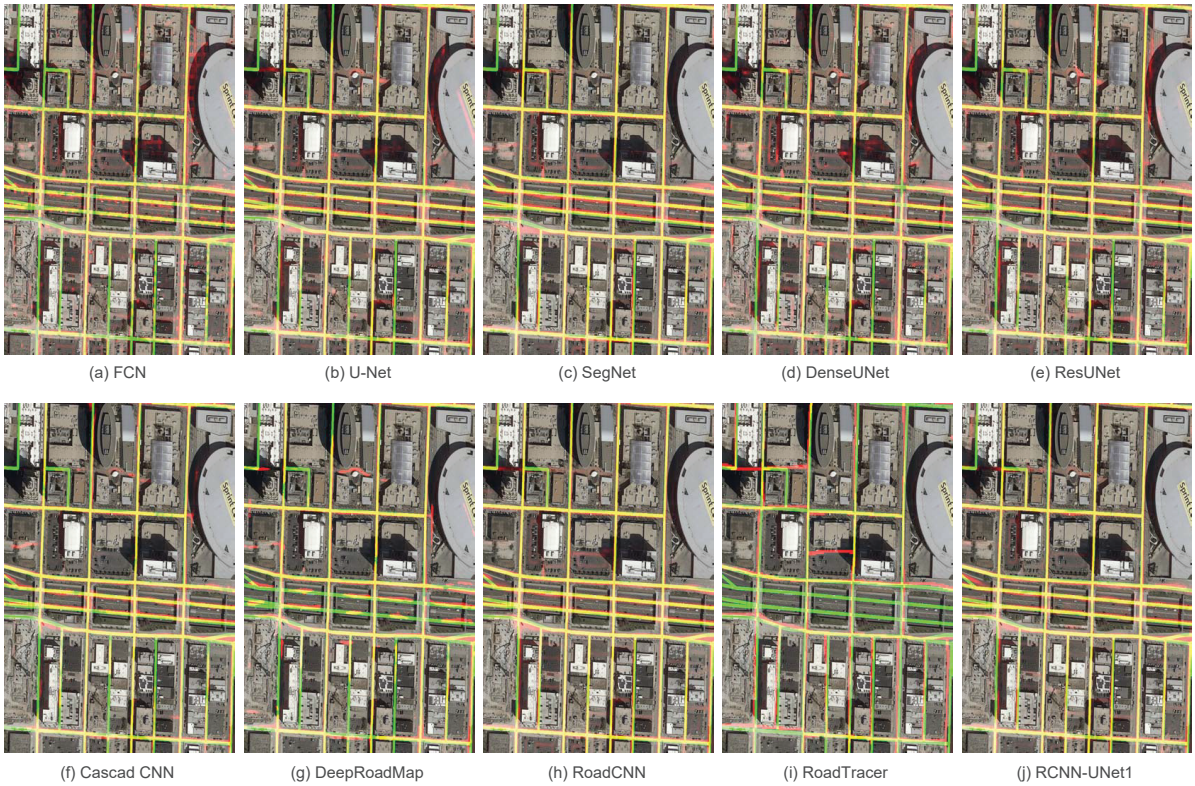


Fig. 6. Qualitative comparison of road detection results achieved by different methods based on one example image from RoadTracer dataset. The yellow color denotes the true positive parts; the red color denotes the false positive parts; and the green color represents the false negative parts.

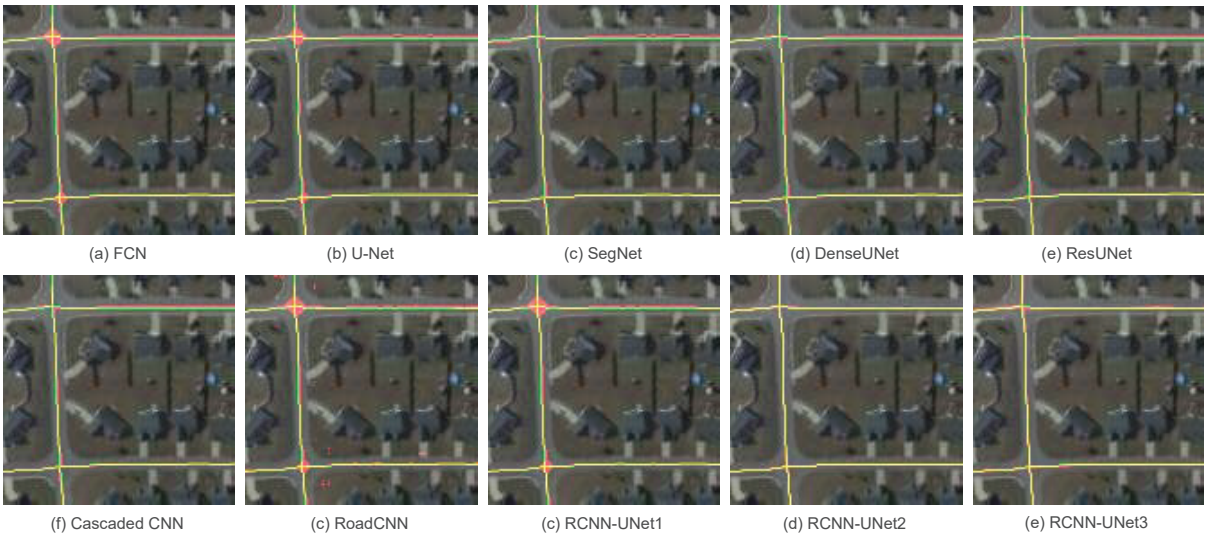


Fig. 7. Qualitative comparison of road centerline extraction results achieved by different methods based on one example image from Cheng dataset. The yellow color denotes the true positive parts; the red color denotes the false positive parts; and the green color represents the false negative parts.

TABLE IV
QUANTITATIVE COMPARATIVE EVALUATION AMONG DIFFERENT METHODS FOR ROAD CENTERLINE EXTRACTION ON CHENG DATASET. IT SHOULD BE NOTED THAT THE RESULTS ARE THE AVERAGE PERFORMANCE OF ALL IMAGES IN TEST SET. A HIGHER VALUE INDICATES A BETTER PERFORMANCE.

Methods Name	COM	COR	Q	F1
FCN	0.7427±0.0390	0.9098±0.0407	0.8516±0.0658	0.8178±0.0392
U-Net	0.9534±0.0050	0.9285±0.0038	0.9335±0.0034	0.9656±0.0014
SegNet	0.9641±0.0071	0.9671±0.0165	0.9560±0.0152	0.9408±0.0079
DenseUnet	0.9230±0.0094	0.9686±0.0460	0.8865±0.0522	0.9453±0.0190
ResUNet	0.9262±0.0142	0.9641±0.0211	0.9005±0.0311	0.9448±0.0167
Cascaded CNN	0.9549±0.0030	0.9790±0.0042	0.9362±0.0067	0.9668±0.0035
RoadCNN	0.9687±0.0650	0.9561±0.1219	0.9141±0.1150	0.9624±0.1103
RCNN-UNet1	0.9738±0.0017	0.9800±0.0067	0.9583±0.0044	0.9769±0.0048
RCNN-UNet2	**0.9871**±0.0034	0.9856±0.0076	0.9732±0.0070	0.9863±0.0055
RCNN-UNet3	0.9794±0.0095	**0.9959**±0.0012	**0.9744**±0.0092	**0.9876**±0.0047

supervised multi-task learning model for road detection, which needs sufficient training samples. In the future, it would be interesting to study a semi-supervised model that can detect road and centerlines by using a smaller amount of training samples.

ACKNOWLEDGMENT

This work was supported by the National Key R&D Program of China, 2018YFB0504900, 2018YFB0504905 and the Shenzhen Science and Technology Program under Grant JCYJ20170811160212033, Grant JCYJ20160330163900579, Grant JCYJ20180507183823045, and Grant JCYJ20170413105929681. Lau's work was supported in part by the Research Grants Council of the Hong Kong SAR (Projects: CityU 11502115 and CityU 11525716), the National Natural Science Foundation of China (NSFC) Basic Research Program (Project 71671155), and CityU Shenzhen Research Institute. X. Huang was supported in part by Natural Science Foundation of Jiangxi Province under Grant No.20181BAB202024 and in part by the Education Department of Jiangxi Province under Grant GJJ170413.

REFERENCES

[1] Q. Li, L. Chen, M. Li, S.-L. Shaw, and A. Nüchter, "A sensor-fusion drivable-region and lane-detection system for autonomous vehicle navigation in challenging road scenarios," IEEE Transactions on Vehicular Technology, vol. 63, no. 2, pp. 540–555, 2014.

[2] B. Du, M. Zhang, L. Zhang, R. Hu, and D. Tao, "Pltd: Patch-based low-rank tensor decomposition for hyperspectral images," IEEE Transactions on Multimedia, vol. 19, no. 1, pp. 67–79, 2017.

[3] R. Bonnefon, P. Dhérété, and J. Desachy, "Geographic information system updating using remote sensing images," Pattern Recognition Letters, vol. 23, no. 9, pp. 1073–1083, 2002.

[4] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in European Conference on Computer Vision. Springer, 2010, pp. 210–223.

[5] J. Yuan, D. Wang, B. Wu, L. Yan, and R. Li, "Legion-based automatic road extraction from satellite imagery," IEEE transactions on geoscience and remote sensing, vol. 49, no. 11, pp. 4528–4538, 2011.

[6] Z. Miao, W. Shi, H. Zhang, and X. Wang, "Road centerline extraction from high-resolution imagery based on shape features and multivariate adaptive regression splines," IEEE geoscience and remote sensing letters, vol. 10, no. 3, pp. 583–587, 2013.

[7] C. Unsalan and B. Sirmacek, "Road network detection using probabilistic and graph theoretical methods," IEEE Transactions on Geoscience and Remote Sensing, vol. 50, no. 11, pp. 4441–4453, 2012.

[8] R. Alshehhi and P. R. Marpu, "Hierarchical graph-based segmentation for extracting road networks from high-resolution satellite images," ISPRS journal of photogrammetry and remote sensing, vol. 126, pp. 245–260, 2017.

[9] X. Huang and L. Zhang, "Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines," International Journal of Remote Sensing, vol. 30, no. 8, pp. 1977–1987, 2009.

[10] D. Chaudhuri, N. Kushwaha, and A. Samal, "Semi-automated road detection from high resolution satellite images by directional morphological enhancement and segmentation techniques," IEEE journal of selected topics in applied earth observations and remote sensing, vol. 5, no. 5, pp. 1538–1544, 2012.

[11] W. Shi, Z. Miao, Q. Wang, and H. Zhang, "Spectral–spatial classification and shape features for urban road centerline extraction," IEEE Geoscience and Remote Sensing Letters, vol. 11, no. 4, pp. 788–792, 2014.

[12] W. Shi, Z. Miao, and J. Debayle, "An integrated method for urban main-road centerline extraction from optical remotely sensed imagery," IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 6, pp. 3359–3372, 2014.

[13] G. Cheng, F. Zhu, S. Xiang, and C. Pan, "Road centerline extraction via semisupervised segmentation and multidirection nonmaximum suppression," IEEE Geoscience and Remote Sensing Letters, vol. 13, no. 4, pp. 545–549, 2016.

[14] A. Sironi, V. Lepetit, and P. Fua, "Multiscale centerline detection by learning a scale-space distance transform," in Conference on Computer Vision and Pattern Recognition (CVPR), no. EPFL-CONF-198153, 2014.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 12, pp. 2481–2495, 2017.

[20] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," IEEE Geoscience and Remote Sensing Letters, 2018.

[21] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P. A. Heng, "H-denseunet: Hybrid densely connected unet for liver and liver tumor segmentation from ct volumes," arXiv preprint arXiv:1709.07330, 2017.

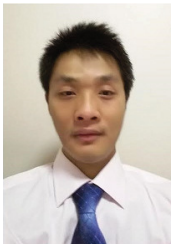
[22] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, vol. 1, no. 2, 2017, p. 3.

[23] V. Mnih and G. E. Hinton, "Learning to label aerial images from noisy data," in Proceedings of the 29th International conference on machine learning (ICML-12), 2012, pp. 567–574.

[24] Q. Zhang, Y. Wang, Q. Liu, X. Liu, and W. Wang, "Cnn based suburban building detection using monocular high resolution google earth images," in Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International. IEEE, 2016, pp. 661–664.

[25] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," IEEE Geoscience and Remote Sensing Magazine, vol. 4, no. 2, pp. 22–40, 2016.
[26] X. Yang, Y. Ye, X. Li, R. Y. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," IEEE Transactions on Geoscience and Remote Sensing, 2018.
[27] G. Cheng, Y. Wang, S. Xu, H. Wang, S. Xiang, and C. Pan, "Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network," IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 6, pp. 3322–3337, 2017.
[28] S. Saito, T. Yamashita, and Y. Aoki, "Multiple object extraction from aerial imagery with convolutional neural networks," Electronic Imaging, vol. 2016, no. 10, pp. 1–9, 2016.
[29] J. M. Alvarez, T. Gevers, Y. LeCun, and A. M. Lopez, "Road scene segmentation from a single image," in European Conference on Computer Vision. Springer, 2012, pp. 376–389.
[30] M. Song and D. Civco, "Road extraction using svm and image segmentation," Photogrammetric Engineering & Remote Sensing, vol. 70, no. 12, pp. 1365–1371, 2004.
[31] Q. Zhang and I. Couloigner, "Benefit of the angular texture signature for the separation of parking lots and roads on high resolution multi-spectral imagery," Pattern recognition letters, vol. 27, no. 9, pp. 937–946, 2006.
[32] S. Das, T. Mirnalinee, and K. Varghese, "Use of salient features for the design of a multistage framework to extract roads from high-resolution multispectral satellite images," IEEE transactions on Geoscience and Remote sensing, vol. 49, no. 10, pp. 3906–3931, 2011.
[33] J. D. Wegner, J. A. Montoya-Zegarra, and K. Schindler, "A higher-order crf model for road network extraction," in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013, pp. 1698–1705.
[34] P. Gamba, F. Dell'Acqua, and G. Lisini, "Improving urban road extraction in high-resolution images exploiting directional filtering, perceptual grouping, and simple topological concepts," IEEE Geoscience and Remote Sensing Letters, vol. 3, no. 3, pp. 387–391, 2006.
[35] X. Hu, Y. Li, J. Shan, J. Zhang, and Y. Zhang, "Road centerline extraction in complex urban scenes from lidar data based on multiple features," IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 11, pp. 7448–7456, 2014.
[36] A. Sironi, V. Lepetit, and P. Fua, "Projection onto the manifold of elongated structures for accurate extraction," in International Conference on Computer Vision, no. EPFL-CONF-211536, 2015.
[37] G. Mattyus, W. Luo, and R. Urtasun, "Deeproadmapper: Extracting road topology from aerial images," in The IEEE International Conference on Computer Vision (ICCV), Oct 2017.
[38] A. Mosinska, P. Márquez-Neila, M. Kozinski, and P. Fua, "Beyond the pixel-wise loss for topology-aware delineation," CoRR, vol. abs/1712.02190, 2017. [Online]. Available: http://arxiv.org/abs/1712.02190
[39] G. Mttyus and R. Urtasun, "Matching adversarial networks," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
[40] F. Bastani, S. He, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, S. Madden, and D. DeWitt, "Roadtracer: Automatic extraction of road networks from aerial images," in Computer Vision and Pattern Recognition (CVPR), 2018.
[41] P. H. O. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in ICML, 2014.
[42] T. Zhang and C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," Communications of the ACM, vol. 27, no. 3, pp. 236–239, 1984.

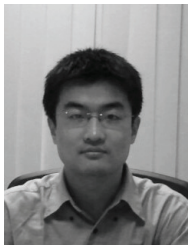
Xiaofei Yang Xiaofei Yang received the B.Sc. from Suihua University in 2007 and 2011, and received M.Sc. degrees from Harbin Institute of Technology in 2011 and 2013, respectively. Currently, he is a Ph.D. candidate in Shenzhen Graduate School, Harbin Institute of Technology. His research interests are in the areas of semi-supervised learning, deep learning, remote sensing, transfer learning and graph mining.



Xutao Li Xutao Li is now an Associate Professor in the Shenzhen Graduate School, Harbin Institute of Technology. He received the Ph.D. and Master degrees in Computer Science from Harbin Institute of Technology in 2013 and 2009, and the Bachelor from Lanzhou University of Technology in 2007. His research interests include data mining, machine learning, graph mining and social network analysis, especially tensor based learning and mining algorithms.

Yunming Ye Yunming Ye received the Ph.D. in Computer Science from Shanghai Jiao Tong University. He is now a professor in the Shenzhen Graduate School, Harbin Institute of Technology. His research interests include data mining, text mining, and ensemble learning algorithms.

Raymond Y. K. Lau Raymond Y. K. Lau is an Associate Professor in the Department of Information Systems at City University of Hong Kong. He is the author of two hundred refereed international journals and conference papers. His research work has been published in renowned journals such as MIS Quarterly, INFORMS Journal on Computing, ACM Transactions on Information Systems, IEEE Transactions on Knowledge and Data Engineering, IEEE Internet Computing, Journal of MIS, Decision Support Systems, etc. His research interests include Big Data Analytics, Social Media Analytics, FinTech, and AI for Business. He is a senior member of the IEEE and the ACM, respectively.



Xiaofeng Zhang Xiaofeng Zhang received the MSc degree from Harbin Institute of Technology in 1999, and the Ph.D. degree from Hong Kong Baptist University in 2008. He has worked in R&D center of Peking University Founder Group and E-business Technology Institute of Hong Kong University. He is now an associate professor at department of computer science of Harbin Institute of Technology Shenzhen Graduate School. His research interests include data mining, machine learning and graph mining.

Xiaohui Huang Xiaohui Huang received the B.Eng. and masters degrees from Jiangxi Normal University, Nanchang, China, in 2005 and 2008, respectively, and the Ph.D. degree from the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China, in 2014. Since 2015, he has been with the School of Information Engineering Department, East China Jiaotong University, Nanchang, China, where he is currently a lecturer of computer science. His current research interests include clustering, social media analysis, and deep learning.