Topology-Enhanced Urban Road Extraction via a Geographic Feature-Enhanced Network

Xingang Li, Yuebin Wang, *Member, IEEE*, Liqiang Zhang, Suhong Liu, Jie Mei, and Yang Li

***Abstract*— Urban road extraction has wide applications in public transportation systems and unmanned vehicle navigation. The high-resolution remote sensing images contain background clutter and the roads have large appearance differences and complex connectivities, which makes it a very challenging task for road extraction. In this article, we propose a novel end-to-end deep learning model for road area extraction from remote sensing images. Road features are learned from three levels, which can remove the distraction of the background and enhance feature representation. A direction-aware attention block is introduced to the deep learning model for keeping road topologies. We compare our method on public remote sensing data sets with other related methods. The experimental results show the superiority of our method in terms of road extraction and connectivity preservation.**

***Index Terms*— Convolutional neural networks(CNNs), deep learning, image segmentation, road extraction, topology relationship.**

I. Introduction

ROAD detection is one of the classical research topics in the remote sensing field. Numerous applications have benefitted from road detection, such as urban design, vehicle navigation, unmanned vehicles, and geospatial data integration. Many studies focus on algorithms to separate automatically the road from the background information [1]–[4]. There exist two categories of road detection in general: road area extraction, which attempts to obtain information on road area, and road centerline extraction, whose target is to obtain single-pixel information from the road centerline. Since the road centerline can be obtained from the road area by morphological thinning algorithms [4], road area extraction is considered our target, and road detection is mentioned later specifically for road area extraction.

Manuscript received January 16, 2020; revised March 13, 2020; accepted April 2, 2020. Date of publication May 11, 2020; date of current version November 24, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0118100 and in part by the National Natural Science Foundation of China under Grant 41861053 and Grant 41801241. *(Xingang Li, Yuebin Wang, and Jie Mei contributed equally to this work.) (Corresponding author: Suhong Liu.)*

Xingang Li, Liqiang Zhang, Suhong Liu, and Yang Li are with the Beijing Key Laboratory of Environmental Remote Sensing and Digital Cities, Beijing Normal University, Beijing 100875, China (e-mail: lixg95@126.com; zhanglq@bnu.edu.cn; liush@bnu.edu.cn; isliyang@mail.bnu.edu.cn).

Yuebin Wang is with the School of Land Science and Technology, China University of Geosciences, Beijing 100083, China, and also with the Shanxi Provincial Key Laboratory of Resources, Environment and Disaster Monitoring, Jinzhong 030600, China (e-mail: xxgcdxwyb@163.com).

Jie Mei is with the TKLNDST, College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: meijie0507@gmail.com).

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

Digital Object Identifier 10.1109/TGRS.2020.2991006

One of the most important challenges in road detection is numerous vehicles, buildings, and trees that cause road discontinuities and incompleteness in remote sensing images under complex urban traffic and geographical environments. Two resolving patterns are adopted to improve the result completeness in topological space: enhancing the feature excavation of road by more scientific network structures, such as the deeper network [4], [5] and the multitask network [6], [7], and using a two-step workflow to refine the result produced by the deep learning model [1], [8]. However, the current methods based on the first pattern are far sufficient to extract the features of the remote sensing images, nor does the two-step approach take advantage of the end-to-end benefits of deep learning.

The spatial and shape information of the roads usually can be represented by the three levels: points, polylines, and polygons. In this article, we propose a novel deep learning model for implementing road extraction on the three levels from high-resolution remote sensing images. Three kinds of blocks, i.e., pixel block, edge block, and region block, contribute to the final accuracy from different aspects. To refine further the topological relationships, we add a direction-aware attention block to the deep learning model, reformulating the segmentation as the connectivity-prediction task. The feature excavation and topology-refining processes are end-to-end trained, which benefits the parameter optimization. We compare the performance of our model with the other five methods on the SpaceNet data set [9]. The experimental results demonstrate the proposed model is superior to the others in road segmentation and topology preservation.

Our goal is to extract robust and discriminative features for road extraction and connectivity. The contributions of our method are threefold.

1) A novel deep learning framework for effectively extracting the road shape features from complex high-resolution remote sensing images is presented. The framework can effectively recognize the roads in three scales with cluttered background.
2) The shape features including point, edge, and area characteristics are automatically learned from three levels, i.e., pixels, edges, and regions through integrating the three levels into the deep learning network.
3) A direction-aware attention block is introduced to the deep learning model for keeping road topologies. The block keeps the road connectivity and further improves the road-recognition accuracy.

II. Related Works

In the early period of road detection, the mainstream approaches tended to define some criteria and traversed the whole image to select the pixels that matched the criteria, such as in geometrically constrained template matching [10] and color-based template matching [11]. The criteria-match method is limited to both specific criteria and the traversal algorithm, which is inefficient and poorly generalizable. Methods based on machine learning discern the road via presetting certain road features. Simler [12] has a support vector machine (SVM) method that detects roads using color, spectral, and geometrical features. Zhang and Couloigner [13] preseted many shape descriptors to classify the road segments obtained by the k-means clustering algorithm. However, the handcrafted feature is hard to design and not robust enough in a complex situation.

Many convolutional neural network (CNN)-based road-detection approaches have achieved the state-of-the-art results in recent years [6], [14]–[16]. Peng *et al.* [17] aggregated a multiscale context for feature learning based on the CNNs to achieve better performances in road segmentation. Lu *et al.* [18] build a multitask learning (MTL) system to detect road centerlines. Zhang *et al.* [4] proposed a method called ResUnet, which combines symmetric construction with residual networks [19] to extract the road area information and achieved the state-of-the-art results relative to the comparison methods. In the DeepGlobe Road Extraction Challenge competition [20], Zhou *et al.* [5] won the first prize with a creative feature extraction structure called D-Linknet, which appends feature excavation layers with residual blocks. Ventura *et al.* [3] proposed an algorithm that obtains a road graph of the whole urban aerial image by patch-based iteration. However, it fails to deal with some complicated situations, such as the overpasses. RoadNet, proposed by Liu *et al.* [7], successfully integrates the road regions, edges, and centerlines in a network by three CNNs that correspond to three semantic segmentation tasks: road area detection, road edge detection, and road centerline detection. The satisfactory results of RoadNet show that multiple types of features are able to enhance each other in an end-to-end network model. Many studies attempt to integrate other geographical data. Some approaches extracted road information from light detection and ranging (LiDAR) data [21]–[23]. Yuan and Cheriyadat [24] infered the road network from the noisy GPS data and guided the road area segmentation. However, both the LiDAR and GPS data are hard to acquire, and the preprocessing of these data is complicated.

A. Pixel-Level Feature

The pixel-level feature is produced by a neural network with pixel-to-pixel workflow, i.e., each road pixel has a corresponding pixel in the prediction maps, which is powerful to describe the road by accumulating spatial and spectral information over receptive fields. The method with pixel-to-pixel workflow includes the fully convolutional networks (FCNs) [25], Unet [26], Unet++ [27], D-Linknet [5], SegNet [2], and ResUnet [4]. Because of the background clutter in the remote sensing image, the results using the traditional models [28] may be largely affected by the features extracted from the background of the image.

B. Edge-Level Feature

The edge-level feature has an advantage in discerning the boundaries between different objects. In earlier studies, the edge-detection methods, including Candy [29] and Sober [30], mainly focus on using the color and intensity information of the images. DeepContour [31] and DeepEdge [32] develop the edge-level automatic hierarchical feature with the deep neural network. Liu and Lew [33] built hierarchical supervisory convolutional networks with relaxed deep supervision to strengthen edge detection. Xie and Tu [34] propose a holistically nested-edge-detection (HED) network to extract the edge information. Based on HED, Liu *et al.* [35] proposed a deeper convolutional hierarchy network, achieving better accuracy in the field of edge detection. The categorywise edge-detection model named CASENet presented by Yu *et al.* [36] achieved fine-edge-detection performance in the Street View data set. Marmanis *et al.* [28] showed that boundary detection significantly optimizes the parameters of the classification network. Liu *et al.* [7] also proved that the accurate edge detection can refine the road-recognition accuracy. We describe the workflow as pixel to edge. The edge-level feature has more sensitivity in the high-frequency part of images, exerting positive impacts on road detection.

C. Region-Level Feature

Scene classification is an important topic in the remote sensing field. The solution has evolved from the bag of visual words [37], [38] to deep learning-related methods [39]–[41]. We take the pattern of scene classification as the pixel-to-region workflow. The region-level feature preserves the contextual information during the scene classification process, considering more information around the road when discerning the road.

D. Feature Fusion

The idea of feature fusion has been applied to various visual tasks in recent years [2], [7], [26], [27], [42], proving that the performance of the segmentation task can be improved by stacking different level features. The MTL framework also gives a guideline to integrate the different levels of features [43]–[45].

E. Topology Enhancing

There are a lot of studies on road topology structure preservation. The road topology is simulated by the point process [46]. The Markov random field is used to correct the topology of the roads [8]. Zhang *et al.* [47], [48] proposed the model based on the generative adversarial network (GAN), obtaining complete road networks by refining the imperfect road topology. DeepRoadMapper [1], reasoning the missing connections of the CNN, outputs as a shortest path problem. Bastani *et al.* [49] used the postprocessing heuristics to infer the graph connectivity. The PolyMapper [50] has the CNN outputs

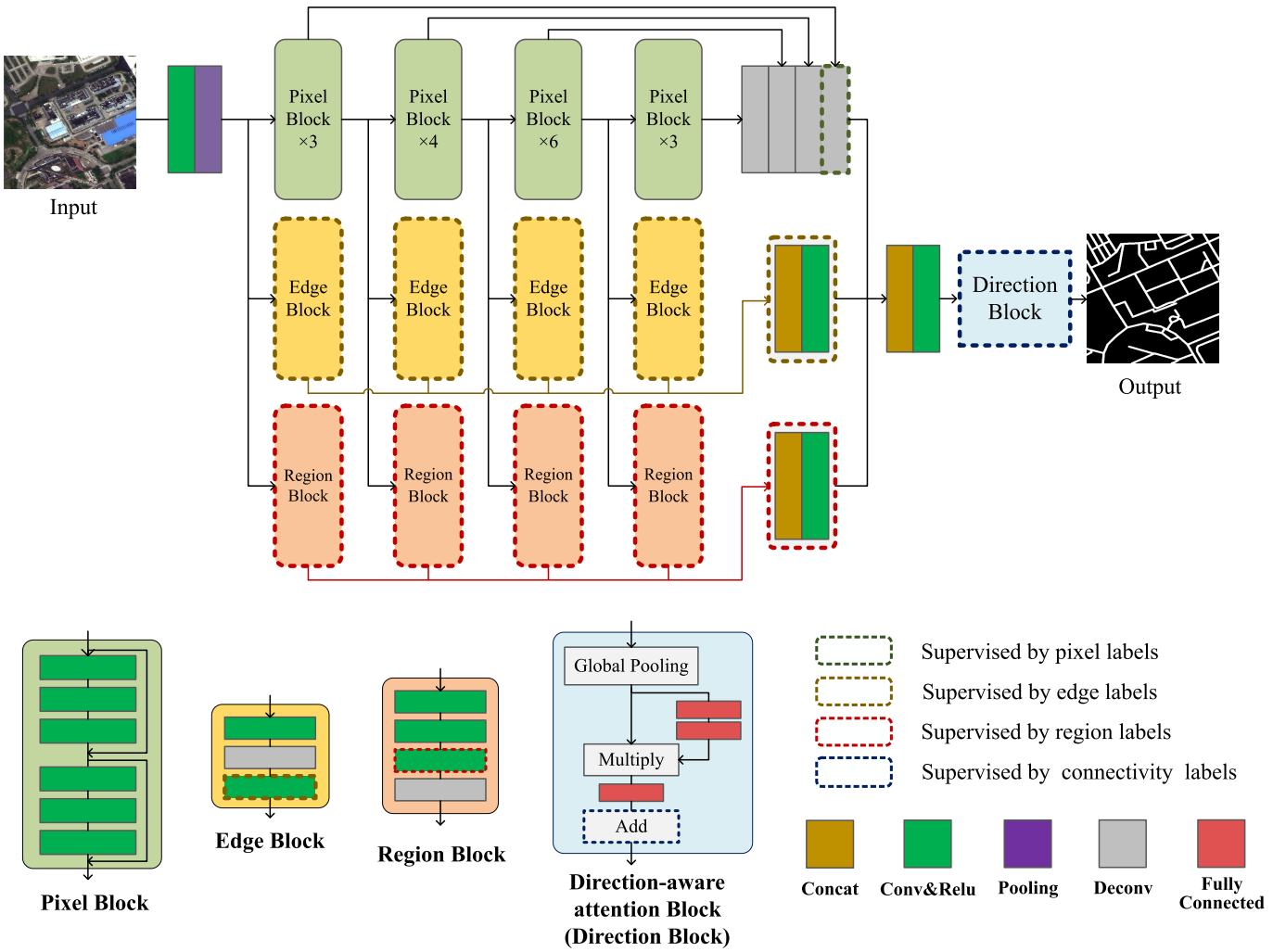


Fig. 1. Illustration of our network architecture for road detection. The pixel block, edge block, and region block are proposed to improve feature excavation. In the stage of feature refining, we have the direction-aware attention block (abbreviated as direction block) recalibrating the stacked features to model the interdependences between the road pixels.

connected sequentially by the recurrent neural networks (RNNs). RoadNet [7] manually tags the connection information in remote sensing images to train the network. Ventura *et al.* [3] used the patch-based method to solve the road topology problem. ConnNet, a novel pixel-connectivity network proposed by Kampffmeyer *et al.* [51], improved the overall semantic rationality and region smoothing by modeling the adjacent relationships between the pixels. Inspired by that, we introduce a direction-aware attention block to enhance the integrity of the topology information in the remote sensing images.

III. Proposed Method

The features are extracted by three cascaded CNN structures based on the pixel, edge, and region levels. Each level contributes distinctive features for road segmentation via MTL [43]. A direction-aware attention block is proposed to enhance the connectivity of the roads by predicting the connectivity probabilities of each road pixel with its neighboring ones, reformulating the segmentation as the connectivity-prediction task [51]. Fig. 1 shows the architecture of the method.

A. Feature Learning

1) Pixel-Level Feature Learning: We improve the ResNet50 [19] model to obtain robust pixel-level feature representation. The repeated convolution and pooling operations in original ResNet50 reduces the resolution of the feature map stepwise, retaining rich and multiscale semantic information. Each convolutional layer is followed by a rectified-linear-unit (ReLU) [52] activation function to alleviate the vanishing gradient problem [53]. Then, four deconvolution layers are added, upsampling the feature map to the size of 64×64, 128×128, 256×256, and 512×512 respectively. The probability map is obtained by the softmax layer and supervised by the pixel labels (depicted by the green dashed lines in Fig. 1).

2) Edge-Level Feature Learning: The edge-level feature learning structure consists of four edge blocks. Each edge block contains two convolution layers and a deconvolution layer. Table I lists the details of the structure, which is inspired by the deeply supervised network (DSN) [54]. Four outputs of the pixel blocks, which we call the shared feature maps temporarily, are used as the input of the edge-level feature

TABLE I

EDGE-LEVEL FEATURE LEARNING PROCESS. "CONV" DENOTES THE CONVOLUTIONAL LAYER. "DECONV" AND "SOFTMAX" REPRESENT THE DECONVOLUTION OPERATION AND SOFTMAX FUNCTION, RESPECTIVELY

Layer name	Input size	Kernel	Stride	Pad	Output size
Conv1-1	256×256	3	1	Yes	256×256
Deconv1	256×256	2	-	-	512×512
Conv1-2	512×512	4	1	Yes	512×512
Softmax1	-	-	-	-	512×512
Conv2-1	128×128	3	1	Yes	128×128
Deconv2	128×128	4	-	-	512×512
Conv2-2	512×512	4	1	Yes	512×512
Softmax2	-	-	-	-	512×512
Conv3-1	64×64	3	1	Yes	64×64
Deconv3	64×64	8	-	-	512×512
Conv3-2	512×512	4	1	Yes	512×512
Softmax3	-	-	-	-	512×512
Conv4-1	32×32	3	1	Yes	32×32
Deconv4	32×32	16	-	-	512×512
Conv4-2	512×512	4	1	Yes	512×512
Softmax4	-	-	-	-	512×512

TABLE II

ENTIRE REGION-LEVEL FEATURE LEARNING PROCESS. "CONV" AND "SOFTMAX" DENOTE THE CONVOLUTIONAL LAYER AND SOFTMAX FUNCTION, RESPECTIVELY. THE PARTITION LAYER REDUCES THE IMAGE SIZE BY THE KERNEL SIZE AND STRIDE WITH NO PADDING

Layer name	Input size	Kernel	Stride	Pad	Output size
Partition layer1	256×256	8	8	No	32×32
Conv1-1	32×32	3	1	Yes	32×32
Conv1-2	32×32	3	1	Yes	32×32
Softmax1	-	-	-	-	32×32
Partition layer2	128×128	4	4	No	32×32
Conv2-1	32×32	3	1	Yes	32×32
Conv2-2	32×32	3	1	Yes	32×32
Softmax2	-	-	-	-	32×32
Partition layer3	64×64	2	2	No	32×32
Conv3-1	32×32	3	1	Yes	32×32
Conv3-2	32×32	3	1	Yes	32×32
Softmax3	-	-	-	-	32×32
Partition layer4	32×32	1	1	No	32×32
Conv4-1	32×32	3	1	Yes	32×32
Conv4-2	32×32	3	1	Yes	32×32
Softmax4	-	-	-	-	32×32

learning structure, introducing multiscale semantic information to edge level. The shared feature maps are convoluted by a kernel size of 3×3 first. The deconvolution layer upsamples the feature map to the size of 512, which is the standard size in the experiment. The second convolution layer functions feature refining in the edge level. Each output of the edge blocks is the one-dimensional feature maps and is supervised by the edge labels.

3) Region-Level Feature Learning: The region-level feature learning structure is inspired by the idea of scene classification, which is able to eliminate the effect of background clutter and noises, discerning the scene category. The whole structure consists of four region-level blocks, depicted in Table II. Each block owns three convolution layers, i.e., a partition layer and two feature-refining layers. The partition layer has the kernel size of 8×8, 4×4, 2×2, and 1×1 with the strides of 8, 4, 2, and 1, respectively. The shared feature maps (with the size of 256×256, 128×128, 64×64, and 32×32) are resized into 32×32 through the partition layer, each unit representing a local feature map with the size of 16×16. The two feature-refining layers have the kernel size of 3×3 and 1×1, fusing the hierarchical features. The final outputs are supervised by the region labels depicted by the red dotted lines in Fig. 1.

4) Feature Fusion: The obtained three-level features reflect the characteristics of the roads in the remote sensing images from different perspectives. Therefore, feature fusions can further make features of the roads more representative. The feature fusion can be summarized as follows. First, we use the deconvolution layer to expand the output in the region level to the size of 512×512. Then, the stacked outputs in the three levels are convolved by the kernel of 1×1 for feature aggregation [55], [56] and sent to the direction-aware attention block for topology enhancing.

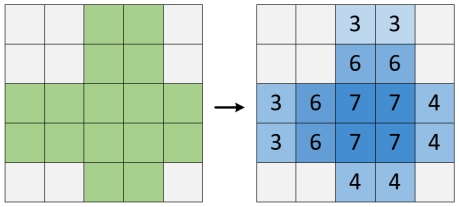


Fig. 2. Process of generating the connectivity labels. The left-hand-side cube represents a road area in simplified form, where the green area denotes the road pixels and the white area is the background. The cube on the right-hand side is the connectivity labels generated by counting the road pixels with the neighboring ones.

B. Direction-Aware Attention Block

To extract better the topological information, we introduce the direction-aware attention block to enhance the integrity of the topological information. Specifically, we predict an n-dimensional connectivity cube, where n is the number of road directions and $n = 8$ as default. The method mainly includes the following two steps.

1) Connectivity Labels: To predict the topological relationships, we first make the connectivity labels. Commonly, the segmentation result is obtained by segmenting the probability map $\sigma(y_i)$ with t. Here, $\sigma(*)$ and t denote the sigmoid nonlinearity operation and the threshold parameter, respectively. On the basis of the pixel labels, we exclude the background information before building the connectivity labels, as shown in Fig. 2. The left-hand-side cube represents a road area, where

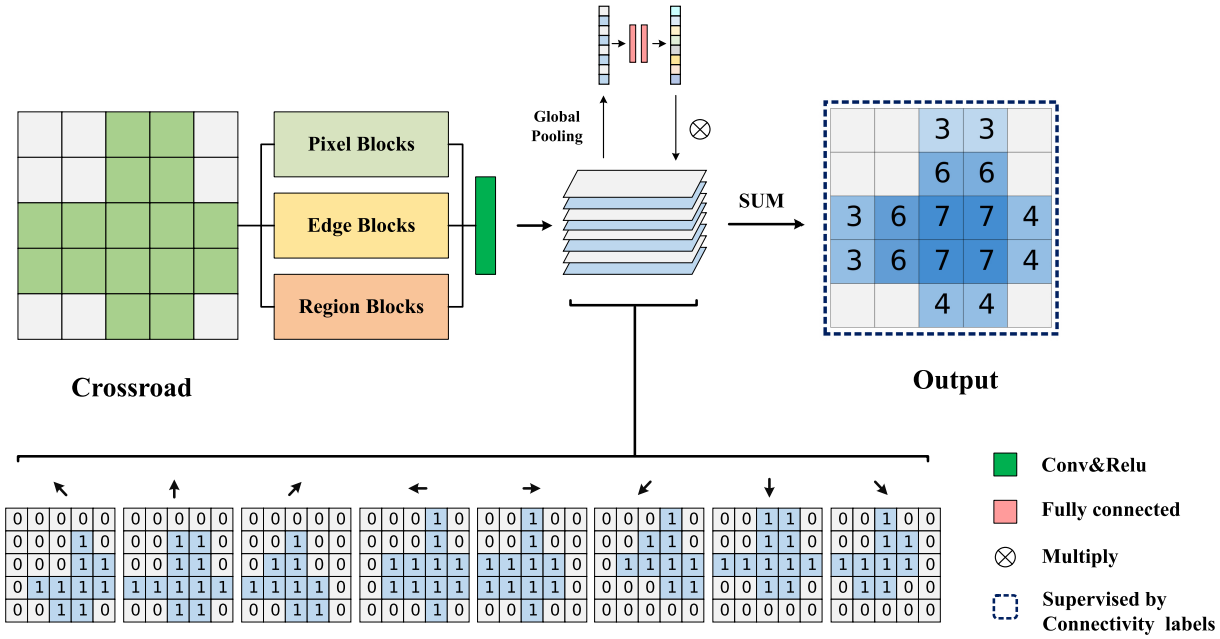


Fig. 3. Description of the proposed direction-aware attention block. A simplified crossroad is represented by the green grid. The direction block produces an eight-dimensional feature map, each dimension representing the number of neighboring road pixels in a direction. The channelwise attention structure makes the network pay more attention to the information of various directions. Finally, the output is summed up and supervised by the connectivity labels.

the green area denotes the road pixels and the white area is the background. The cube on the right-hand side is the connectivity labels generated by counting the road pixels with the neighboring ones.

2) Channel Attention Mechanism: After the three levels of feature learning, we convolute the features to C channels by the kernel of 1×1. In our approach depicted in Fig. 3, $C = 8$ denotes the eight directions. Each channel represents the adjacency information in the corresponding direction. We describe the input feature maps as $\mathbf{f} = [\mathbf{f_1}, \mathbf{f_2}, \ldots, \mathbf{f_c}]$, where $\mathbf{f_i} \in \mathbb{R}^{W \times H}$ is the ith slice of f. After the average pooling operation, the feature of each channel is obtained. Then, we use two fully connected layers and a sigmoid activation function to recalibrate the direction features. The final output is a vector $\mathbf{v} \in \mathbb{R}^C$, and each element of the vector represents the weight of the corresponding channel. Then, the weight vector is applied to the unified feature map

$$\mathbf{x} = \mathbf{F}(\mathbf{f}, \mathbf{v}) \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^{W \times H}$ refers to the channelwise weighted feature maps and $\mathbf{F}$ denotes a channelwise multiplication operation between the feature maps and the weight vector $\mathbf{v}$. The final output, obtained by the sigmoid function on $\mathbf{x}$, is the probability of each pixel belonging to either the road or the nonroad.

C. Optimization Method

The supervision of the three-level feature learning is performed by optimizing the balanced cross-entropy loss. Since the pixels belonging to the roads only occupy a small part in the remote sensing images, the positive and negative ratios are unbalanced. Thus, the native cross-entropy loss function should pay more attention to the negative sample features. The balanced cross-entropy loss [34] weights the loss according to the number of corresponding pixels

$$L_f = -\beta \sum_{y \in y_+} \log \Pr\left(P_j = 1 \middle| \mathbf{X}, \mathbf{W}, \mathbf{w}\right) - (1-\beta) \sum_{y \in y_-} \log \Pr\left(P_j = 0 \middle| \mathbf{X}, \mathbf{W}, \mathbf{w}\right) \tag{2}$$

where $\beta = |y_-|/|y|$ represents the class balancing weight, which refers to the ratio of the number of negative samples to the total number of pixels. $|y|$, $|y_+|$, and $|y_-|$ refer to the total number of pixels, number of positive pixels, and number of negative pixels in image $\mathbf{X}$, respectively. The class balancing weight, $\Pr(\cdot) \in [0,1]$, represents the probability of a pixel belonging to a certain class, which is computed using the softmax function on the feature maps.

In our topology-enhanced process, the problem to be solved is approximately equivalent to a regression problem, since each pixel needs to be assigned a value representing the number of neighboring road pixels ranging from zero to eight. We address this issue by optimizing the mean squared error (MSE)

$$L_{\text{topo}} = \frac{1}{n} \sum_{i=1}^{n} w_i \left(y^i_{\text{pred}} - y^i_{\text{GT}}\right)^2 \tag{3}$$

where n denotes the total number of elements in the connectivity cube and y^i_* indicates the connectivity value in position i. The supervised process using the connectivity ground truth is indicated in Figs. 1 and 3 with a blue dotted box.

Our method comprises a deep learning network with multiple outputs at different levels. The network architecture resembles that of a deeply supervised frame [54], which has been proven to improve the convergence and generalization of the

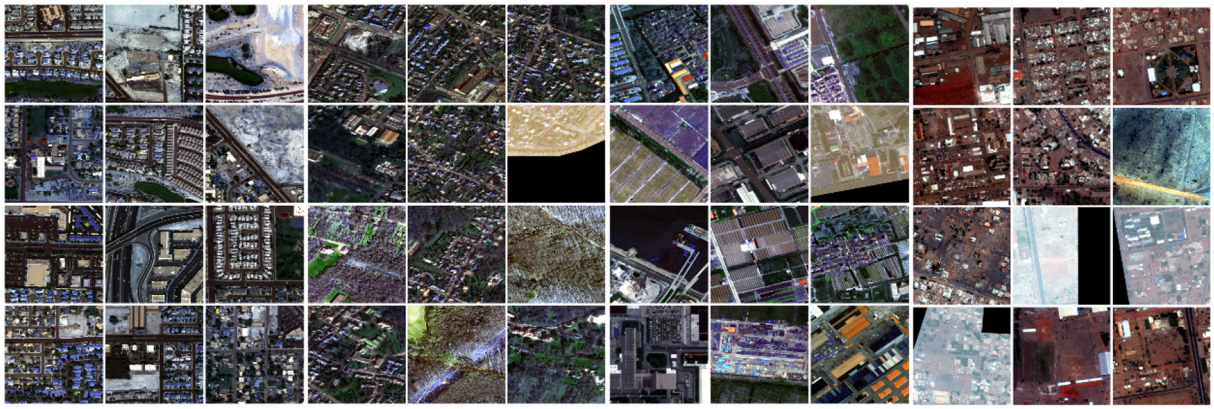


Fig. 4. Overview of the data set from several typical urban areas in Las Vegas, Paris, Shanghai, and Khartoum gathered from SpaceNet [9].

TABLE III
NUMBER OF TRAINING, VALIDATING, AND TESTING SAMPLES FOR EACH CITY

	Las Vegas	Paris	Shanghai	Khartoum	Total
Train	619	163	704	173	1659
Valid	108	28	124	30	290
Test	245	60	200	63	568
Total	972	251	1028	266	2517

deep networks [34], [35]. We minimize the following objective function via an adaptive moment-estimation algorithm (Adam) [57]:

$$L = w_t L_{\text{topo}} + \sum_{m=1}^{M} w_m L_f \tag{4}$$

where M refers to the total number of outputs supervised by the three-level labels. w_t and w_m are the weights, and their values are determined according to the loss during the training process. After many epoch traversals, each loss function in the network is well maintained between 0 and 1, so both w_t and w_m are set to 1.0 by default. This default setting may cause the parameters not in the best optimization state, and we will study this issue in future work.

IV. EXPERIMENTS

In this section, we evaluate the performance of the proposed method.

A. Data Set

The Road data set [9] in SpaceNet, as shown in Fig. 4, is a publicly available benchmark for the high-resolution remote sensing image. The images from four cities, i.e., Las Vegas, Paris, Shanghai, and Khartoum, collected from the WorldView-3 satellite are used to compare the methods. Each image is in the size of 3000 × 3000 pixels, with the spatial resolution of 30 cm, containing roads, various buildings, vehicles, rivers, and vegetation in urban regions. Table III lists the number of training, validating, and testing samples for each city.

(a) (b) (c) (d)

Fig. 5. Visualization of the road labels with pixel level, edge level, and region level. The columns display the training sample images from (a) Las Vegas, (b) Paris, (c) Shanghai, and (d) Khartoum.

The data set gives only the centerline labels in a vector format. To explore better the performance of different approaches, we annotate the pixel labels with a width of 3 m to ensure the label covers the majority of the road region. We outline the pixel labels with 2 pixels as the edge pixels. The region labels contain 32 × 32 units, each unit representing a local image with a size of 16 × 16. The label of each unit depends on the proportion of road pixels. We preset the proportion as 8% to retain more road information. Fig. 5 shows some aerial images with the corresponding labels.

B. Evaluation Methods

The intersection-over-union (IoU) metric mainly measures the overlap between the predictions and the labels. The F1-score is mainly used to evaluate the image segmentation accuracy. Van Etten *et al.* [9] suggested that the two metrics may not always reflect the road characteristics, because they fail to consider the connectivity relationships among the roads.

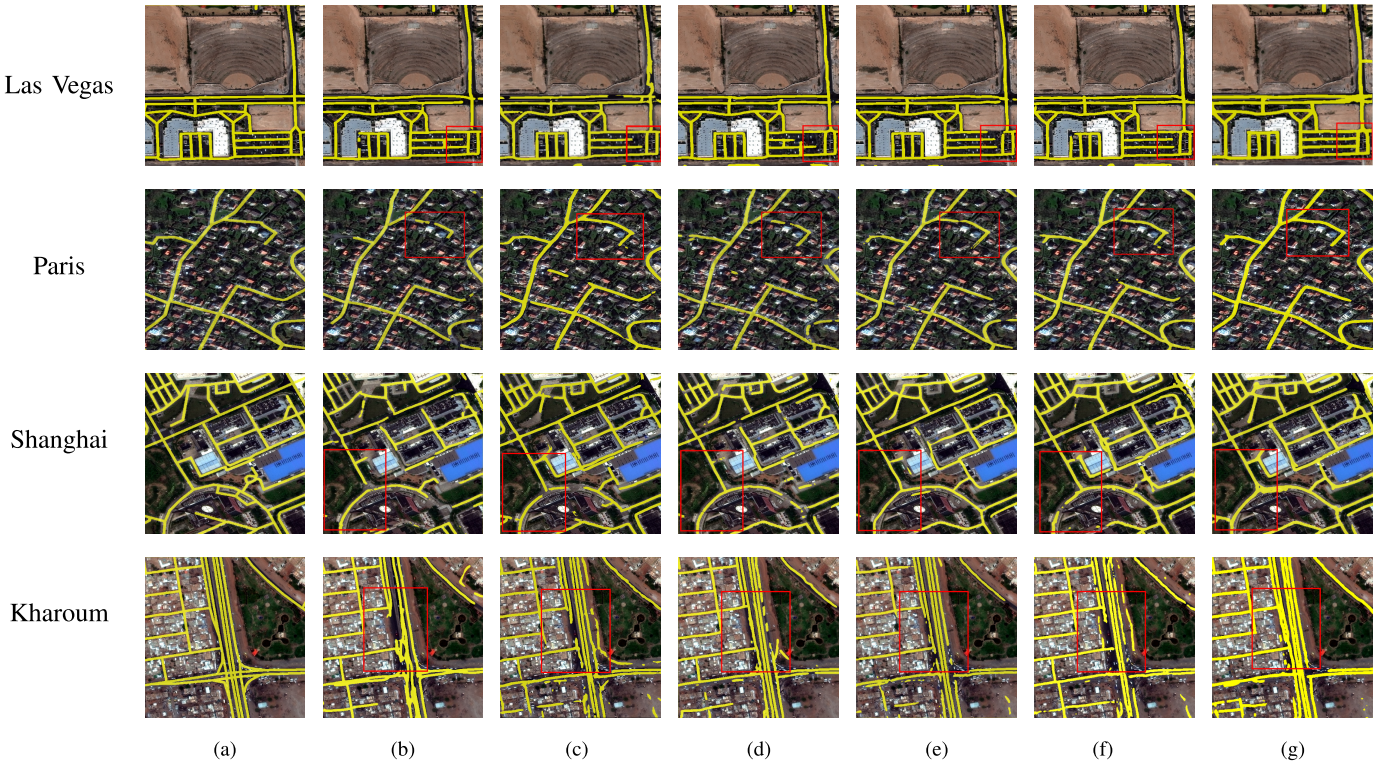


Fig. 6. Visual comparison of the urban road extraction results with different comparison algorithms. The figure is organized in four rows and seven columns. The prediction results of (a) labels, (b) U-net, (c) U-net++, (d) Segnet, (e) ResUnet, (f) D-Linknet, and (g) our method are arranged in a row.

TABLE IV
QUANTITATIVE COMPARISON OF THE APLS INDICATOR (%)

Method	Cities				Average
	Las Vegas	Paris	Shanghai	Khartoum	
U-Net	0.7586	0.5980	0.5213	0.4021	0.5625
U-Net++	0.7621	0.5893	0.5143	0.4774	0.5855
SegNet	0.7849	0.5849	0.5399	0.4118	0.5919
Res-Unet	0.7894	0.6003	0.5318	0.5142	0.6107
D-LinkNet	0.7692	0.5985	0.4898	0.4719	0.5791
Our Method	**0.8104**	**0.6081**	**0.5504**	**0.5335**	**0.6252**

TABLE V
QUANTITATIVE COMPARISON OF THE IOU INDICATOR (%)

Method	Cities				Average
	Las Vegas	Paris	Shanghai	Khartoum	
U-Net	0.6143	0.5242	0.5113	0.4671	0.5292
U-Net++	0.6275	0.5336	0.5237	0.5188	0.5559
SegNet	0.6247	0.5259	0.5003	0.4993	0.5376
Res-Unet	0.6348	0.5476	0.5218	0.5198	0.5585
D-LinkNet	0.6267	0.5511	0.5024	0.4996	0.5442
Our Method	**0.6416**	**0.5722**	**0.5248**	**0.531**	**0.5667**

To evaluate the performance of road semantic segmentation and connectivity completeness, we adopt the following metrics to compare our method with others.

1) Average Path Length Similarity (APLS) Metric: The APLS metric [9] mainly measures the similarity between the ground truth and the recognition results in terms of the logical topology of the roads. The metric is calculated by the following equation:

$$M_{\text{apls}} = 1 - \frac{1}{N}\sum_{k} \min\left\{1, \frac{\left|L(a,b) - L\left(a',b'\right)\right|}{L(a,b)}\right\}. \quad (5)$$

Here, N is the number of unique paths, $L(\cdot)$ is the length between the two nodes, and k denotes all the possible source and target nodes in each specific graph.

2) IoU: The IoU is the ratio between the intersection and union parts of the segmentation results and ground truth. The IoU is rewritten as follows:

$$\text{IoU} = \frac{\text{GT} \cap \text{DR}}{\text{GT} \cup \text{DR}} \quad (6)$$

where GT and DR denote the road pixels in the ground truth and prediction results, respectively.

3) Confusion Matrix: Three metrics are also used in this experiment: the precision (P), recall (R), and F-score (F_{score}) metrics. The precision denotes the proportion of the correctly predicted pixels in all the positive cases. The recall represents the proportion of the true-positive samples in the prediction results. The F_{score} is an overall metric that combines the precision and recall metrics. They are defined as follows:

$$P = \text{TP}/(\text{TP} + \text{FP}) \quad (7)$$

$$R = \text{TP}/(\text{TP} + \text{FN}) \quad (8)$$

$$F_{\text{score}} = 2PR/(P + R) \quad (9)$$

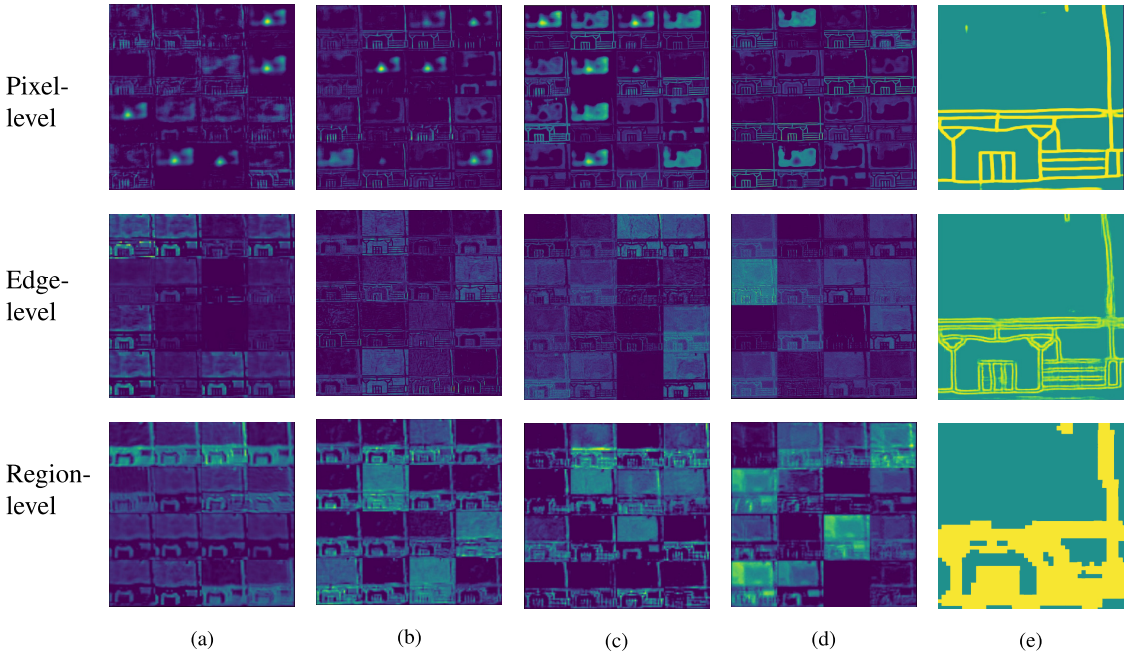


Fig. 7. Visualization of the convolutional layers and the final result of the pixel-, edge-, and region-level blocks. It is obvious that the edge-level features preserve more high-frequency information, while the region-level features reflect the global information. (a) Block 1. (b) Block 2. (c) Block 3. (d) Block 4. (e) Result.

Fig. 8. Output of the direction block. (Left subfigures) Labels of road area. (Middle subfigures) Output of the direction block displayed in the color stripe, which contains eight values from 0 to 8, where pixels in zeros represent the background information and nonzero values indicate the road pixel. (Right subfigures) Threshold output of the direction block.

where TP, FP, and FP denote the *true-positive*, *false-positive*, and *false-negative* counts, respectively.

C. Implementation Details

The proposed model is implemented using Keras and optimized through the Adam algorithm on an NVIDIA GeForce GTX 1080 Ti GPU with 11 GB of onboard memory. We randomly select 15% of the data set in each city from the training data for validation. We use ImageNet [58] to pretrain the model to help improving the convergence of the model [58]. To speed up computation, the images are resized to 512×512 pixels and input to the model. We train the model

TABLE VI
QUANTITATIVE COMPARISON OF THE PRECISION, RECALL, AND F1-SCORE METRICS

Methods	Las Vegas			Paris			Shanghai			Khartoum		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
U-Net	0.7402	0.7834	0.7612	0.6028	0.7076	0.6510	0.5815	0.6895	0.6309	0.5769	0.7224	0.6420
U-Net++	0.7528	0.7979	0.7747	0.6730	0.7629	0.7151	0.6465	0.7339	0.6874	0.6691	0.6978	0.6832
SegNet	0.7365	**0.8014**	0.7676	0.6125	0.7647	0.6802	0.5759	**0.7530**	0.6526	0.6137	0.7285	0.6662
Res-Unet	0.7585	0.7983	0.7779	**0.6612**	0.7523	0.7038	0.6412	0.6678	0.6542	0.6412	**0.7386**	0.6864
D-LinkNet	0.7856	0.7356	0.7598	0.6471	0.7817	0.7081	0.627	0.6569	0.6416	0.6687	0.6941	0.6663
Our Method	**0.826**	0.7498	**0.7861**	0.6397	**0.7946**	**0.7088**	**0.6882**	0.6885	**0.6883**	**0.6739**	0.7076	**0.6903**

with a minibatch size of 3 due to the graphics memory limit, and the learning rate is initially set to $1e^{-4}$ and reduced by a factor of 0.05 in every ten epochs.

D. Comparisons of the Road Detection

We compare our method with five state-of-the-art road segmentation methods: U-Net [26], U-Net++ [27], SegNet [2], ResUnet [4], and D-LinkNet [5]. The performances of the six approaches are listed in Tables IV–VI. It is noted that our method performs better than the other four approaches in most of the indicators in terms of both segmentation and topology.

In Fig. 6, we display the detection results of our proposed approach and the other five approaches. Compared with the five approaches, our method has the ability to extract the subtle roads and keep the connectivity better. Especially, when there are occlusions or shadows in the image, our method can successfully segment the roads, while other methods fail to obtain the roads in the occluded region.

E. Discussion

In this section, we further evaluate the effectiveness of the proposed feature learning structure and the direction-aware attention block.

1) Effectiveness of the Feature Learning Structure: We choose the remote sensing images of Las Vegas as an example to validate the effectiveness of the multiple feature learning at the three levels. Feature visualization is an effective way to explain the method. The visualization of each block in Fig. 1 is displayed in Fig. 7. The last column shows the final output obtained by convolving the last layer to the one-dimensional feature map by the kernel of 1×1 at each level, denoting that the network has the ability to learn the features of pixel level, edge level, and region level. We discover that the edge-level features (the second row in Fig. 7) preserve more high-frequency information that is reflected in the sharp edges of the feature maps, while the region-level features (the third row in Fig. 7) gather more global information that removes the distraction of the background.

We compare our method to the ones with different combinations among the pixel blocks, edge blocks, region blocks, and direction blocks. As shown in Table VII, our method significantly outperforms the compared ones, which demonstrate that each part of our model contributes to the final results. Thus, we conclude that the proposed feature extraction structure consisting of pixel level, edge level, and region level can focus on the interesting regions and learn the features of roads from various geographical features.

TABLE VII
PERFORMANCE COMPARISON OF ROAD-DETECTION METHOD WITH DIFFERENT CONFIGURATIONS OF THE PROPOSED NETWORK

	IoU	F1-score	APLS
Pixel only	0.6154	0.7683	0.7819
Pixel + Edge	0.6312	0.7758	0.7753
Pixel + Region	0.6255	0.7696	0.8032
Pixel + Edge+ Region (no Direction Block)	0.6375	0.7786	0.8077
Pixel + Edge + Region	**0.6416**	**0.7861**	**0.8104**

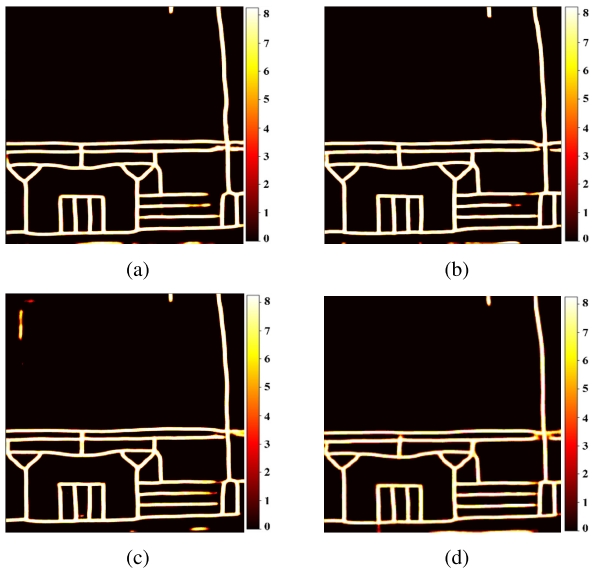


Fig. 9. Outputs of direction block with different configurations. We can find that the method is able to represent the road information more strongly with the addition of edge-level and region-level feature excavation structures. (a) Pixel only. (b) Pixel + Edge. (c) Pixel + Region. (d) Pixel + Edge + Region.

2) Effectiveness of the Direction-Aware Attention Block: To validate the effectiveness of the proposed direction-aware attention block, we compare our model with the one in which the direction block is replaced with a widely used fully connected layer. As shown in Table VII, the method with the direction block performs better in both the evaluation

indicators. We visualize the output of the direction block, as shown in Fig. 8. The middle subfigure depicts the output of the direction block displayed in the color stripe, which contains eight values from 0 to 8, where pixels in zeros represent the background information and nonzero values indicate the number of neighboring road pixels. The larger pixel value denotes higher connectivity probability. It is noted that the road detection is more accurate after fusing the edge-level and region-level features. We find that some fragile road sections are successfully connected as shown in Fig. 8, denoting the direction block that contributes to the topological relationship. Compared with other topology reconstruction models, we adopt the direction block to enhance the topology relationship, which is more concise in understanding and more holistic in parameter optimizing.

To compare the performance of the direction block based on different configurations, we display the visualization results in Fig. 9 and the quantitative results in Table VII. We find that the ability of the network to detect roads becomes stronger with the addition of the feature excavation modules and the direction-aware attention block.

V. Conclusion

In this article, we propose a novel end-to-end CNN-based network that combines the pixel-, edge-, and region-level road geographical features that are commonly recognized in high-resolution remote sensing images to recognize the urban roads as well as the direction-aware attention block to enhance the topological relationship.

In the feature learning stage, the encoder part is shared among the pixel, edge, and region levels so that the feature maps are able to integrate the three levels to describe the road simultaneously. Each level is learned by the cascaded CNN-based blocks. In the topology-enhanced stage, we reformulate the problem of binary segmentation into connectivity prediction and introduce the attention mechanism to recalibrate adaptively the features. The whole process is holistically trained in an end-to-end manner. The experimental results prove that our method performs better in segmentation and topology compared with the other state-of-the-art road-detection approaches, especially in the connectivity and integrity of the road areas. There exist some unsolved problems. The addition of edge level and region level introduces more parameters to the whole network, although we adjust the network structure to alleviate the problem. The MSE loss function just computes the numerical regression, which ignores the geographical features of the actual expression in the direction-aware attention block. Roads apparently broken in the images cannot be connected in prediction. Future work will involve the solution of the abovementioned problems and exploit this framework in more geography-related tasks.

References

[1] G. Mattyus, W. Luo, and R. Urtasun, "DeepRoadMapper: Extracting road topology from aerial images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3438–3446.

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

[3] C. Ventura, J. Pont-Tuset, S. Caelles, K.-K. Maninis, and L. Van Gool, "Iterative deep learning for road topology extraction," 2018, *arXiv:1808.09814*. [Online]. Available: http://arxiv.org/abs/1808.09814

[4] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[5] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182–186.

[6] 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 Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3322–3337, Jun. 2017.

[7] Y. Liu, J. Yao, X. Lu, M. Xia, X. Wang, and Y. Liu, "RoadNet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2043–2056, Apr. 2019.

[8] G. Mattyus, S. Wang, S. Fidler, and R. Urtasun, "Enhancing road maps by parsing aerial images around the world," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1689–1697.

[9] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "SpaceNet: A remote sensing dataset and challenge series," 2018, *arXiv:1807.01232*. [Online]. Available: http://arxiv.org/abs/1807.01232

[10] G. Vosselman and J. De Knecht, "Road tracing by profile matching and Kaiman filtering," in *Automatic Extraction of Man-Made Objects From Aerial and Space Images*. Springer, 1995, pp. 265–274.

[11] Y. He, H. Wang, and B. Zhang, "Color-based road detection in urban traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 4, pp. 309–318, Dec. 2004.

[12] C. Simler, "An improved road and building detector on VHR images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2011, pp. 507–510.

[13] 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 Recognit. Lett.*, vol. 27, no. 9, pp. 937–946, Jul. 2006.

[14] G. Cheng, Y. Wang, F. Zhu, and C. Pan, "Road extraction via adaptive graph cuts with multiple features," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 3962–3966.

[15] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Road segmentation of remotely-sensed images using deep convolutional neural networks with landscape metrics and conditional random fields," *Remote Sens.*, vol. 9, no. 7, p. 680, 2017.

[16] R. Liu *et al.*, "Multiscale road centerlines extraction from high-resolution aerial imagery," *Neurocomputing*, vol. 329, pp. 384–396, Feb. 2019.

[17] B. Peng, Y. Li, L. He, K. Fan, and L. Tong, "Road segmentation of UAV RS image using adversarial network with multi-scale context aggregation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 6935–6938.

[18] X. Lu *et al.*, "Multi-scale and multi-task deep learning framework for automatic road extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9362–9377, Nov. 2019.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[20] I. Demir *et al.*, "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 17–172.

[21] Y.-W. Choi, Y.-W. Jang, H.-J. Lee, and G.-S. Cho, "Three-dimensional LiDAR data classifying to extract road point in urban area," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 725–729, Oct. 2008.

[22] M. Yadav, A. K. Singh, and B. Lohani, "Extraction of road surface from mobile LiDAR data of complex road environment," *Int. J. Remote Sens.*, vol. 38, no. 16, pp. 4655–4682, Aug. 2017.

[23] J. Liang, N. Homayounfar, W.-C. Ma, S. Wang, and R. Urtasun, "Convolutional recurrent network for road boundary extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9512–9521.

[24] J. Yuan and A. M. Cheriyadat, "Image feature based GPS trace filtering for road network generation and road segmentation," *Mach. Vis. Appl.*, vol. 27, no. 1, pp. 1–12, Jan. 2016.

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

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2015, pp. 234–241.

[27] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested U-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.

[28] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.

[29] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[30] W. Gao, X. Zhang, L. Yang, and H. Liu, "An improved Sobel edge detection," in *Proc. 3rd Int. Conf. Comput. Sci. Inf. Technol.*, vol. 5, Jul. 2010, pp. 67–71.

[31] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang, "DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3982–3991.

[32] G. Bertasius, J. Shi, and L. Torresani, "DeepEdge: A multiscale bifurcated deep network for top-down contour detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4380–4389.

[33] Y. Liu and M. S. Lew, "Learning relaxed deep supervision for better edge detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 231–240.

[34] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1395–1403.

[35] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jul. 2017, pp. 3000–3009.

[36] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam, "CASENet: Deep category-aware semantic edge detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5964–5973.

[37] C.-F. Tsai, "Bag-of-words representation in image annotation: A review," *ISRN Artif. Intell.*, vol. 2012, pp. 1–19, Nov. 2012.

[38] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2010, pp. 143–156.

[39] K. Nogueira, W. O. Miranda, and J. A. D. Santos, "Improving spatial feature representation from aerial scenes by using convolutional networks," in *Proc. 28th SIBGRAPI Conf. Graph., Patterns Images*, Aug. 2015, pp. 289–296.

[40] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[42] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018.

[43] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*. [Online]. Available: http://arxiv.org/abs/1706.05098

[44] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris, "Fully-adaptive feature sharing in multitask networks with applications in person attribute classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5334–5343.

[45] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multitask learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3994–4003.

[46] R. Stoica, X. Descombes, and J. Zerubia, "A Gibbs point process for road extraction from remotely sensed images," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 121–136, May 2004.

[47] Y. Zhang, Z. Xiong, Y. Zang, C. Wang, J. Li, and X. Li, "Topology-aware road network extraction via multi-supervised generative adversarial networks," *Remote Sens.*, vol. 11, no. 9, p. 1017, 2019.

[48] Y. Zhang, X. Li, and Q. Zhang, "Road topology refinement via a multi-conditional generative adversarial network," *Sensors*, vol. 19, no. 5, p. 1162, 2019.

[49] F. Bastani *et al.*, "RoadTracer: Automatic extraction of road networks from aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4720–4728.

[50] Z. Li, J. D. Wegner, and A. Lucchi, "Topological map extraction from overhead images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1715–1724.

[51] M. Kampffmeyer, N. Dong, X. Liang, Y. Zhang, and E. P. Xing, "ConnNet: A long-range relation-aware pixel-connectivity network for salient segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2518–2529, May 2019.

[52] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*. [Online]. Available: http://arxiv.org/abs/1505.00853

[53] K. Hara, D. Saito, and H. Shouno, "Analysis of function of rectified linear unit used in deep learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–8.

[54] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. Artif. Intell. Statist.*, Feb. 2015, pp. 562–570.

[55] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[57] A. Gulli and S. Pal, *Deep Learning With Keras*. Birmingham, U.K.: Packt, 2017.

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

Xingang Li is pursuing the master's degree with the Faculty of Geographical Science, Beijing Normal University, Beijing, China.

His research interests include machine learning and remote sensing imagery processing and application.

Yuebin Wang (Member, IEEE) received the Ph.D. degree from the School of Geography, Beijing Normal University, Beijing, China, in 2016.

He was a Post-Doctoral Researcher with the School of Mathematical Sciences, Beijing Normal University. He is an Associate Professor with the School of Land Science and Technology, China University of Geosciences, Beijing. His research interests include remote sensing imagery processing and 3-D urban modeling.

Liqiang Zhang received the Ph.D. degree in geoinformatics from the Institute of Remote Sensing Applications, Chinese Academy of Science, Beijing, China, in 2004.

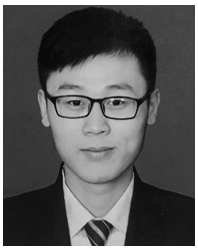
He is a Professor with the Faculty of Geographical Science, Beijing Normal University, Beijing. His research interests include remote sensing image processing, 3-D urban reconstruction, and spatial object recognition.

Suhong Liu received the B.S. degree in computer science from Southwest Jiaotong University, Chengdu, China, in 1988, the M.S. degree in geophysical well-logging from Jianghan Petroleum University, Jingzhou, China, in 1991, and the Ph.D. degree in cartography and remote sensing from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 1999.

She is a Professor with the Faculty of Geographical Science, Beijing Normal University, Beijing. Her research interests include spatiotemporal analysis of remotely sensed data and retrieval of land biophysical parameters from satellite data.

Jie Mei is pursuing the Ph.D. degree with the College of Computer Science, Nankai University, Tianjin, China.

His research interests include computer vision, machine learning, and remote sensing image processing.



Yang Li is pursuing the master's degree with the Faculty of Geographical Science, Beijing Normal University, Beijing, China.

His research interests include deep learning and remote sensing image processing.