CoANet: Connectivity Attention Network for Road Extraction From Satellite Imagery

Jie Mei, Rou-Jing Li, Wang Gao, and Ming-Ming Cheng, *Senior Member, IEEE*

***Abstract*— Extracting roads from satellite imagery is a promising approach to update the dynamic changes of road networks efficiently and timely. However, it is challenging due to the occlusions caused by other objects and the complex traffic environment, the pixel-based methods often generate fragmented roads and fail to predict topological correctness. In this paper, motivated by the road shapes and connections in the graph network, we propose a connectivity attention network (CoANet) to jointly learn the segmentation and pair-wise dependencies. Since the strip convolution is more aligned with the shape of roads, which are long-span, narrow, and distributed continuously. We develop a strip convolution module (SCM) that leverages four strip convolutions to capture long-range context information from different directions and avoid interference from irrelevant regions. Besides, considering the occlusions in road regions caused by buildings and trees, a connectivity attention module (CoA) is proposed to explore the relationship between neighboring pixels. The CoA module incorporates the graphical information and enables the connectivity of roads are better preserved. Extensive experiments on the popular benchmarks (SpaceNet and DeepGlobe datasets) demonstrate that our proposed CoANet establishes new state-of-the-art results. The source code will be made publicly available at: https://mmcheng.net/coanet/.**

***Index Terms*— Road extraction, satellite imagery, connectivity attention, strip convolution, topological connectivity.**

I. Introduction

CREATING road maps is a basic and essential step in numerous application domains, such as autonomous driving, urban planning, vehicle navigation, and geographic information updating. The existing map collection methods adopted by several mapping companies are usually time-consuming, like extraction from LIDAR point clouds, aggregation of GPS trajectories, or manual road labeling. These methods are unsuitable for large-scale areas and insufficient to update the dynamic changes of road networks in a rapidly changing environment [1], [2]. Satellite imagery not only represents the geometric characteristics of the roads but also provides images from multiple periods and even real-time. To accelerate the update process, extracting road networks from satellite imagery [3]–[6] has become a promising approach.

Traditional studies focus on algorithms that utilize hand-designed features and define certain criteria to extract roads from satellite imagery [7]–[11]. These methods are usually inefficient when processing satellite imagery of large regions. With the development of deep learning, convolutional neural networks (CNNs), especially networks with fully-convolutional network (FCN) [12] architecture, have been proposed and proven to be effective in image semantic segmentation [13]–[18]. Several works have applied CNNs with encoder-decoder architecture to road segmentation tasks [19]–[22], which often obtain good segmentation results. However, extracting road from satellite imagery is challenging due to: (a) occlusions by buildings and trees, (b) complex urban environments and traffic, (c) similarities between the road and other objects. These difficulties lead to fragmentation of the road segmentation, and the above methods can not guarantee the connectivity of roads. Recently, [1], [23]–[25] introduce methods using segmentation followed by post-processing steps to refine the missing connections, where the shortest path algorithm is usually used as post-processing and it can not apply to the intricate road environments. In order to directly obtain the road with better connectivity, [5] adopts an iterative search process to automatically extract the road network, [26] utilizes U-Net combined with multiple loss functions to iteratively refine the road delineation. Besides, Liu *et al.* [27] integrate multi-level features including road surfaces, edges, and centerlines to improve road prediction. However, these methods are time-consuming and usually require complicated steps to train.

In this paper, we propose a connectivity attention network (CoANet) for road extraction from satellite imagery. We first introduce an encoder-decoder architecture network to learn the feature of roads, where the Atrous Spatial Pyramid Pooling module (ASPP) is adopted to increase the receptive field of feature points and capture multi-scale features. Since the roads are long-span, narrow, and distributed continuously, the strip convolutions are more aligned with the shapes of roads. We take advantage of it and develop a strip convolution module (SCM), which is placed in the decoder network. As shown in Fig. 1 (a), the SCM leverages four

Manuscript received December 2, 2020; revised May 7, 2021 and September 6, 2021; accepted September 15, 2021. Date of publication October 7, 2021; date of current version October 13, 2021. This work was supported in part by the New Generation of AI major project under Grant 2018AAA0100400, in part by NSFC under Grant 61922046, and in part by Tianjin Natural Science Foundation under Grant 17JCJQJC43700. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jocelyn Chanussot. *(Corresponding author: Ming-Ming Cheng.)*

Jie Mei and Ming-Ming Cheng are with the TKLNDST, College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: meijie@mail.nankai.edu.cn; cmm@nankai.edu.cn).

Rou-Jing Li is with the State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China (e-mail: lirj@mail.bnu.edu.cn).

Wang Gao is with the Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing 100191, China (e-mail: gaowang_fly@163.com).

Digital Object Identifier 10.1109/TIP.2021.3117076

Fig. 1. Motivation of the *strip convolution* module and the *connectivity attention* module that we proposed. From left to right: satellite imagery, ground truth. (a) Strip convolutions with four shapes are used to capture linear features of roads. (b) Connectivity of one pixel with neighboring pixels is predicted to capture local pair-wise dependencies and ensure road topological correctness.

strip convolutions with horizontal, vertical, left diagonal, and right diagonal to capture long-range context information from four different directions. Besides, it prevents irrelevant regions from interfering with feature learning. To alleviate occlusions in road regions caused by buildings and trees, we propose a connectivity attention module (CoA) to explore the relationship between neighboring pixels. As illustrated in Fig. 1 (b), the connectivity of a given pixel with eight neighboring pixels is predicted, which enables the topological correctness of roads. Extensive experiments on popular benchmarks in terms of pixel-based and graph-based metrics demonstrate the superiority of our CoANet compared with several state-of-the-art methods.

Our contributions are summarized as follows.

- We propose a connectivity attention network that jointly learns the segmentation and relationship between neighboring pixels to improve the connectivity of road, which achieves significant improvements over other methods on widely-used road datasets.
- We develop a strip convolution module that leverages four strip convolutions with different directions to capture long-range context information and avoid interference from irrelevant regions.
- We design a connectivity attention module, which boosts the road connectivity by exploiting dependencies between pair-wise neighboring pixels and incorporating the graphical information.

The remaining of this paper is organized as follows. Sec. II summarizes the related works of road extraction and multi-task learning. In Sec. III, we introduce the details of our proposed CoANet. In Sec. IV, datasets, evaluation metrics, and implementation details are provided, extensive experiments are conducted to evaluate the performance of our method for road extraction from satellite imagery. Conclusion and discussion are presented in Sec. V.

II. Related Work

A. Road Segmentation

Extracting road networks from satellite imagery has been attempted by numerous studies. Traditional road extraction methods usually utilize hand-designed features and define certain criteria to match [28]–[30]. He *et al.* [11] present a color-based road detection algorithm by combining the results of boundaries estimation on the gray-level image and road-area extraction on the color image. Zhang *et al.* [31] introduce a number of descriptors of angular texture and identify the road segments using a fuzzy logic classifier. Laptev *et al.* [8] conduct road extraction based on multi-scale road detection, which is combined with geometry-constrained edge extraction utilizing snakes. [10] extracts road from remote sensing images using a Gibbs point process framework. And [9] develops junction-points processes to recover line-networks in both aerial and retinal images. Wegner *et al.* [32] propose a higher-order conditional random field (CRF) model for road network extraction. However, these approaches are usually inefficient and unsuitable for large-scale areas.

With the development of deep learning, CNNs with encoder-decoder architecture [12], [14], [16], [33]–[37] have been proposed and proven to be effective in semantic segmentation. Some studies [20], [22], [38]–[40] formulate the road extraction as a segmentation problem using CNN-based models. Mnih *et al.* [19] detect roads by using a neural network implemented on a graphics process. Cheng *et al.* [20] propose a cascaded end-to-end CNN (CasNet) to simultaneously process the road segmentation and centerline extraction tasks from very high resolution (VHR) remote sensing images. Panboonyuen *et al.* [21] present a DCNN framework for road segmentation, then landscape metric is proposed to reduce misclassified road pixels and CRF is adopted to sharpen the extracted roads. U-Net [13] and LinkNet [15] are the well-admired encoder-decoder structures for semantic segmentation, their variants are also proposed to learn thin and elongated road features. Zhang *et al.* [22] propose a semantic segmentation neural network called ResUnet, which is combined with residual learning and U-Net for road area extraction. In the CVPR DeepGlobe 2018 Road Extraction Challenge [41], Zhou *et al.* [39] propose a D-LinkNet, which is built with LinkNet [15] structure and added dilated convolution layers in the center part. These methods usually obtain good road segmentation results, but they can not guarantee the connectivity of roads.

B. Road Connectivity

Road connectivity is one of the most important road features, which is necessary for vehicle navigation, autonomous driving, and routing. Recently, researchers have paid more

and more attention to this characteristic and proposed several methods. Wegner *et al.* [23] first segment aerial images into superpixels, and the candidate paths with high road likelihood are connected using the shortest path algorithm. Máttyus *et al.* [1] obtain an initial segmentation of the aerial images using the model with an encoder-decoder structure. Since the segmentation results fail to predict connected roads, they then introduce a post-processing step by reasoning about missing connections with the shortest path algorithm. In these methods, road connectivity is achieved with post-processing, which is not suitable for complex environments like regions with occlusions, ambiguous road appearance, and high road density.

In order to directly obtain the road extraction results with better connectivity, Mosinska *et al.* [26] utilize U-Net combined with pixel-wise loss and topology-aware loss to iteratively refine the road delineation. Bastani *et al.* [5] propose RoadTracer that uses an iterative search process guided by a CNN-based decision function to automatically extract the road networks from aerial images. Batra *et al.* [6] propose a stacked multi-branch module to effectively utilize mutual information between segmentation and orientation learning tasks. They also develop a connectivity refinement approach to iteratively refine the topology of the predicted road networks. However, the iterative steps are time-consuming and these methods usually take a long time to train.

There are some studies that integrate multi-level road features or other geographical data to obtain connected road extraction results. RoadNet is proposed by Liu *et al.* [27] to simultaneously predict road surfaces, edges, and centerlines, where multilevel features are integrated to deal with the roads in various scenes and improve road prediction. Li *et al.* [42] develop a novel framework to effectively integrate the road shape features including point, edge, and area characteristics. Then a direction-aware attention module is introduced to further improve the road connectivity and road-recognition accuracy. The Light Detection and Ranging (LiDAR) data [43]–[45] and GPS trajectories [46]–[50] have also been used to infer road maps. [51] combines crowdsourced GPS data with aerial imagery to extract road, and experiments show that their results outperform the models using GPS data or images alone. The LiDAR and GPS data may be useful to improve road connectivity, especially in areas with occlusions. However, it is challenging to collect enough data of LiDAR and GPS covering a large region, and the preprocessing of these data is often complicated.

C. Pair-Wise Dependencies

In the task of semantic segmentation, some studies have designed modules that aggregate the contextual information to improve performance. Shen *et al.* [52] propose a joint objective to integrate segmentation features, high-order context, and boundary guidance, where a guidance CRF is adopted to further improve the segmentation performance. Dai *et al.* [53] introduce a deformable convolution and a deformable RoI pooling to enhance the transformable modeling capability of CNNs, which is a simple and efficient method to model dense spatial transformations. To capture long-range dependencies, [54] proposes the non-local block that computes contextual responses based on relationships between different positions, while it requires high computation cost. Criss-cross network [55] is developed to obtain the contextual information of all pixels on the criss-cross path, which is more efficient. A pyramid attention module is proposed in [56] to enhance saliency representations by utilizing multi-scale feature learning and an enlarged receptive field. Wang *et al.* [57] propose a pixel-wise contrastive method for semantic segmentation, which learns a well-structured pixel semantic embedding space by leveraging the global context among pixels across different images.

The idea of our connectivity attention module is also related to the methods that learn the pixel affinity in semantic segmentation. Bertasius *et al.* [58] introduce a convolutional random walk network to integrate semantic segmentation and pixel-wise affinity, which addresses the problems of poor boundary localization and spatially fragmented segments. Cheng *et al.* [59] design a locality-sensitive DeconvNet, where an affinity matrix is adopted to learn relations among neighboring pixels. AffinityNet is proposed in [60] to predict high-level semantic affinities between pairs of adjacent image coordinates. Hou *et al.* [61] develop a novel inter-region affinity knowledge distillation approach for the task of road marking segmentation. The above methods aim to recover object shape or refine outputs of semantic segmentation models, which usually establish the semantic affinity in a fixed field. However, our method improves the connectivity of roads by exploiting pair-wise affinity between multi-scale neighboring pixels. Besides, the direction and position of pixel sampling are specific, which are consistent with the distribution of most roads in the satellite imagery.

When annotating the road in satellite imagery, the human need to recognize whether a pixel belongs to road and connect road pixels considering the importance of correct road topology. Inspired by the context aggregation and pixel affinity learning in semantic segmentation, we propose a connectivity attention module to capture pair-wise dependencies among neighboring pixels. The module is able to improve the connectivity of roads, which is shown in our experiments based on pixel-based and graph-based metrics.

III. Method

Road connectivity is an important road characteristic, while segmentation based methods often produce fragmented roads. To alleviate this problem, we develop a connectivity attention network (CoANet) for road extraction from satellite imagery, as shown in Fig. 2. In CoANet, a strip convolution module (SCM) is proposed to align with the shape of the road and extract its linear feature. And we further propose a connectivity attention module (CoA) to predict the road connectivity between neighboring pixels.

A. Network Structure

1) Encoder-Decoder Architecture: In CoANet, we employ ResNet-101 [62] pre-trained on ImageNet [63] as the encoder

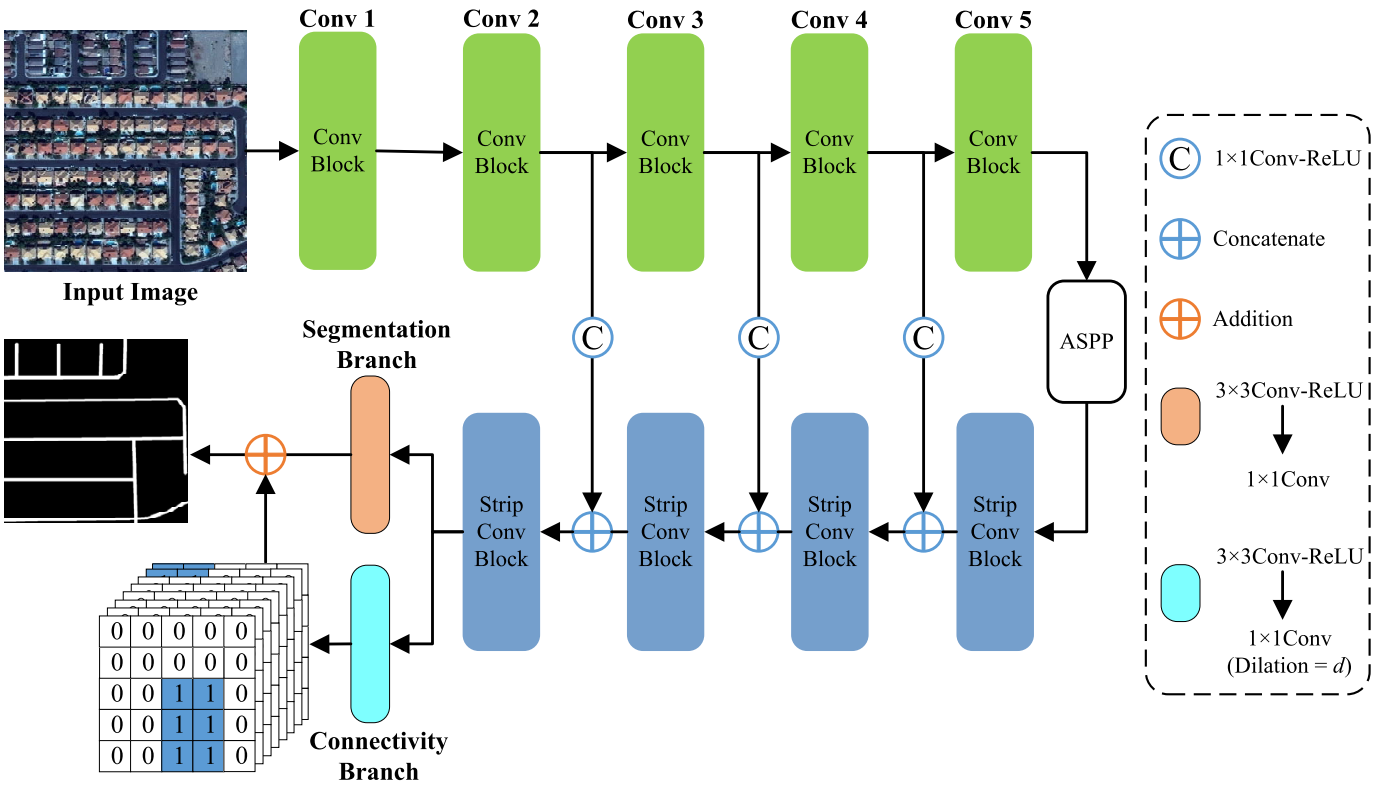


Fig. 2. Overall architecture of the proposed connectivity attention network (CoANet). The encoder module contains five convolution blocks. The ASPP is the Atrous Spatial Pyramid Pooling module, which learns multi-scale features by applying atrous convolution with multiple scales. And the decoder module includes four strip convolution blocks. d denotes the interval of a given pixel with its neighboring pixels.

because of its outstanding performance in feature learning. Since the atrous convolution is a powerful tool in controlling the filter's field-of-view and adjusting the resolution of feature maps, like [16], we apply atrous convolution with rate $r = 2$ and $r = 4$ to the last two convolution blocks in ResNet-101 for denser feature extraction.

To effectively learn features at multiple scales, the Atrous Spatial Pyramid Pooling module (ASPP) in [18] is adopted. Since the roads are narrow, complex, and long-span, the use of ASPP will increase the receptive field of feature points and improve the connectivity of roads. The decoder module contains four strip convolution blocks for up-sampling the feature maps to an appropriate size and extracting linear features of the roads. Each strip convolution block contains four strip convolutions with different directions to capture long-range context information, including horizontal, vertical, left diagonal, and right diagonal. Besides, each output feature map of strip convolution blocks is adjusted by a 1×1 convolution, which is then concatenated with the corresponding feature map of convolution blocks in the encoder.

2) Loss Function: There are two branches after the decoder module: the segmentation branch and the connectivity branch. The connectivity branch corresponds to the connectivity attention module we developed, where the connectivity of a given pixel with eight neighboring pixels is predicted to incorporate the graphical information and guarantee the topological correctness of roads. As for the segmentation branch, it contains a 3×3 convolution and a 1×1 convolution for reducing the number of channels to one. The loss function of the segmentation branch is defined as:

$$L_{seg} = L_{BCE} + \alpha(1 - L_{Dice}), \tag{1}$$

where L_{BCE} is binary cross entropy, and L_{Dice} is the Dice coefficient, which are defined as:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)], \tag{2}$$

$$L_{Dice} = \frac{2\sum_{i=1}^{N}(y_i \hat{y}_i)}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} \hat{y}_i^2}, \tag{3}$$

where α is a constant. N indicates the number of elements in a $H \times W$ slice, y_i is the ground truth denoting road or background for a given pixel in position i, and $\hat{y}_i$ is the corresponding predicted probability of the segmentation branch.

B. Strip Convolution Module

The convolutions in most CNN architectures often have square kernels and learn the feature map within square windows, which is suitable for most natural objects with bulk shape. However, the roads are long-span, narrow, and distributed continuously. Taking advantage of square convolution can not capture the linear features of roads well, and it would inevitably incorporate irrelevant information from neighboring pixels. The strip convolution is more aligned with the shape

of roads, which utilizes a long kernel shape along one spatial direction to capture long-range dependencies in road regions. Besides, it captures local context along the other spatial direction and prevents irrelevant regions from interfering the feature learning.

Motivated by the above fact and the 1D transpose convolution in [51], we propose a novel *strip convolution module* (SCM). As shown in Fig. 3, SCM leverages four strip convolutions with horizontal, vertical, left diagonal, and right diagonal to capture long-range context information from four different directions. Let $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ denote the input tensor for the SCM, Where H, W, and C represent the height, width, and the number of channels. In the strip convolution block, $\mathbf{X}$ is fed into four parallel pathways after a 1×1 convolution, each of which contains a strip convolution with one shape. Then the output feature maps of four strip convolutions are concatenated, which is followed by an up-sampling operation and a 1×1 convolution to obtain the output of the strip convolution block.

Let $\mathbf{w} \in \mathbb{R}^{2k+1}$ be the strip convolution filter with size $2k+1$, $\boldsymbol{D} = (D_h, D_w)$ is the direction of filter $\mathbf{w}$, and $\mathbf{Z}_{\boldsymbol{D}} \in \mathbb{R}^{H\times W\times C'}$ denotes the result of strip convolution. The strip convolution can be defined as:

$$\begin{aligned}\mathbf{Z}_{\boldsymbol{D}}[i,j] &= (\mathbf{X} * \mathbf{w})_{\boldsymbol{D}}[i,j] \\ &= \sum_{l=-k}^{k} x[i + D_h l, j + D_w l] \cdot w[k-l], \quad (4)\end{aligned}$$

where $\mathbf{X} * \mathbf{w}$ denotes the convolution operation. $\boldsymbol{D}$ is the direction vector of the strip convolution, which is $(0,1)$, $(1,0)$, $(1,1)$, and $(-1,1)$ for convolutions of horizontal, vertical, left diagonal, and right diagonal, respectively. For filter $\mathbf{w}$, we set $k = 4$ to make each strip convolution have 9 parameters, which is the same as a 3×3 convolution filter.

In the above SCM, each position in the output feature map is allowed to establish relationships with multiple positions from four directions in the input feature map. The four directions we chose are aligned with the distributions of most roads in the satellite imagery and are relatively easy to implement.

C. Connectivity Attention Module

Extracting roads from satellite imagery is challenging due to the occlusions caused by buildings and trees, which would interfere with road connectivity. To alleviate this problem, we develop a *connectivity attention* module (CoA) to effectively predict road connectivity between neighboring pixels and disentangle the background regions. The CoA is able to explore the relationship between pairs of pixels, which is seamlessly combined with the feature learning process. This module enables our model to integrate information that is usually learned in the graphical models and leads to better connectivity of roads.

Taking advantage of the binary ground truth mask, we first generate the ground truth connectivity cube $O \in \mathbb{R}^{H\times W\times C_o}$, where C_o denotes the number of sampled neighboring pixels for a given pixel and we set $C_o = 8$. In the connectivity cube, $O_{i,j,c}$ indicates the connectivity of a pixel with the neighboring pixel at a specific position, where i, j denote the spatial position of the pixel and c denotes the position of its neighboring pixel. $O_{i,j,c} = 1$ if the two pixels are connected, which means both of them are road pixels. And the background pixels are not connected to reduce the irrelevant noise. For the neighboring pixels, we select the pixels with an interval of $d = 1$ from the given pixel. As shown in Fig. 4 (a), the pixels at positions $C1 - C8$ are chosen as neighboring pixels. By checking if each pixel with its neighboring pixel at the specific position is connected and concatenating the connectivity masks at positions $C1 - C8$, we can obtain the ground truth connectivity cube O.

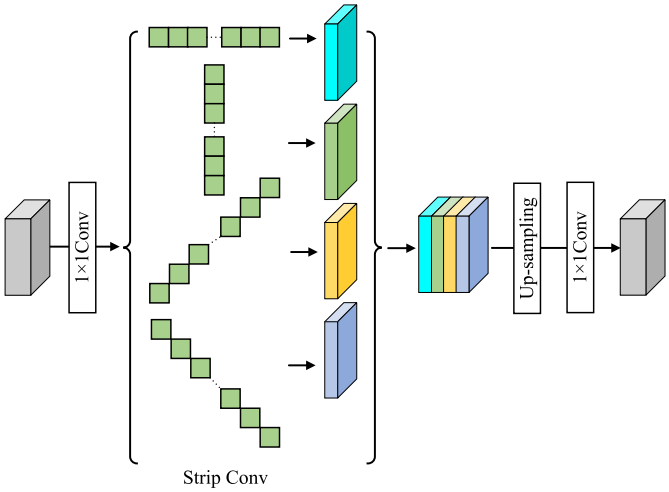


Fig. 3. The strip convolution block. The strip convolutions contain four different shapes: horizontal, vertical, left diagonal, and right diagonal.

In our CoA module, as shown in Fig. 4 (b), the input tensor is fed into a 3×3 convolution, which is followed by a 3×3 atrous convolution with rate $r = d$. The atrous convolution is used to increase the receptive field and learn the relations between neighboring pixels. Then the Squeeze-Excitation (SE) block in [64] is adopted to fully exploit the connectivity, which re-calibrates the predicted connectivity cube by using the channel attention mechanism. The SE block contains two fully connected layers and a sigmoid function. The input feature map is fed into this block after a global average pooling and we can obtain a vector ranging (0, 1), where each factor is multiplied by the corresponding channel in the input feature map. The final output of the CoA module is a $H \times W \times C_o$ connectivity cube to predict the connectivity between neighboring pixels.

The connectivity branch in our proposed CoANet consists of two connectivity attention modules, one of which is the module described above with $d = 1$ and the other is the module with $d = 3$. As for the CoA module with $d = 3$, the interval of a given pixel with its neighboring pixels is set to 3 and the rate of 3×3 atrous convolution in the CoA module is set to $r = 3$. The two CoA modules with different settings are adopted to capture multi-scale connectivity information and improve the connectivity of predicted roads. We will provide more analysis in the experiments on the performance of our approach with different configurations of the CoA module.

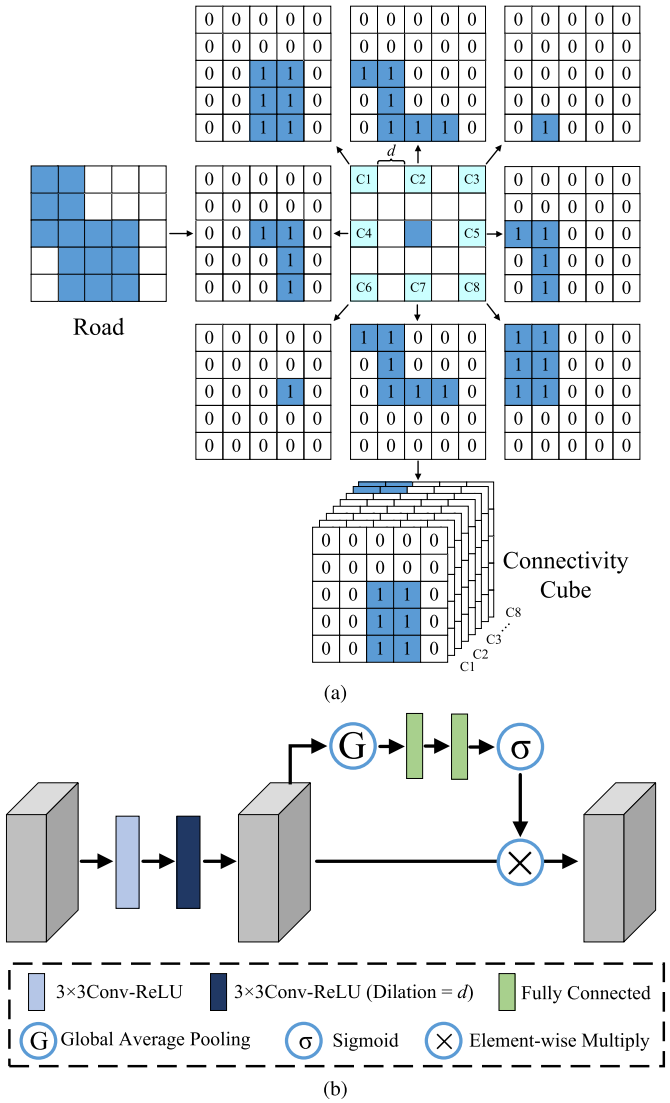


Fig. 4. (a) Illustration of how the connectivity cube is generated, the interval between sampled pixels $d = 1$. In the road image, the white pixels denote the background and the blue pixels are road. In the connectivity cube, all pixels have binary values, where 1 denotes the pixel is connected with a neighboring pixel in one direction and 0 is for not-connected pixels. (b) The connectivity attention module (CoA).

The loss function of the connectivity branch is defined as:

$$L_{con} = L_{d1} + \beta L_{d3}, \tag{5}$$

$$L_{d1} = -\frac{1}{C_o \times N}\sum_{c=1}^{C_o}\sum_{i=1}^{N}[y_i^c \cdot \log(\hat{y_i^c}) + (1-y_i^c)\cdot \log(1-\hat{y_i^c})], \tag{6}$$

where β is a constant. C_o represents the number of sampled neighboring pixels for a given pixel, and N is the number of elements in a $H \times W$ slice. y_i^c is the ground truth denoting connectivity or non-connectivity for a given pixel in position i with its neighboring pixel in position c, and $\hat{y_i^c}$ is the corresponding predicted connectivity of the connectivity branch. The loss function L_{d3} is the same as L_{d1}.

The overall loss function can be defined as:

$$L_{CoANet} = L_{seg} + \lambda L_{con}, \tag{7}$$

where λ is a constant.

IV. Experiments

In this section, we introduce the two datasets and evaluation metrics used in our experiments. Besides, detailed evaluation results in both quantitative and qualitative are presented.

A. Datasets

Two datasets are used in our experiments to evaluate the performance of the proposed method.

1) SpaceNet [65]: This dataset provides 30cm/pixel imagery with a pixel resolution of 1300 × 1300 from four different cities: Paris, Las Vegas, Shanghai, and Khartoum. The annotations of road are provided in the form of line-string that indicating the centerline of road. The dataset consists of 2,780 images, which are split into 2,213 images for training and 567 images for testing following [66]. We augment the training dataset by creating crops of 650 × 650.

2) DeepGlobe [41]: This dataset includes 50cm/pixel imagery with a pixel resolution of 1024 × 1024. The images are collected from three different regions: Thailand, Indonesia, and India. It provides pixel-level annotation, including road and background classes. The dataset contains 6,226 images, following [66], we split it into 4,696 images for training and 1,530 for testing. The training dataset is augmented by creating crops of 512 × 512.

B. Evaluation Metrics

1) Pixel-Based Metrics: To evaluate the performance of our method for road segmentation, we use $F1$-score and Intersection over Union (IoU) metrics. Since the annotations of the SpaceNet dataset are provided in the form of line-string, we obtain the ground truth for road segmentation by rasterizing the line-string with constant width. The buffer of road centerline is set to 3 meters (10 pixels) in our experiments.

2) Graph-Based Metric: The Average Path Length Similarity (APLS) [65] is used in our experiments to evaluate topological correctness and connectivity of roads. The APLS metric measures the differences in optimal path lengths between all pair of nodes in the ground truth graph G and the proposed graph $\hat{G}$, which is defined as:

$$APLS = 1 - \frac{1}{n}\sum min\left\{1, \frac{\left|L(a,b) - L(\hat{a},\hat{b})\right|}{L(a,b)}\right\} \tag{8}$$

where $\hat{a}, \hat{b}$ are the nodes in the predicted graph $\hat{G}$ nearest the location of nodes a, b in the ground truth graph G, respectively. $L(\hat{a},\hat{b})$ and $L(a,b)$ denote the path length between the corresponding nodes in graphs $\hat{G}$ and G, respectively. n is the number of unique paths.

C. Implementation Details

In our CoANet, the Stochastic Gradient Descent (SGD) optimizer is used with a batch size of 16. The momentum and weight decay coefficients are set to 0.9 and 5×10^{-4}, respectively. The learning rate is initially set to 0.01 and we adopt the 'poly' policy to gradually reduce the learning rate,

TABLE I

QUANTITATIVE COMPARISON OF OUR PROPOSED COANET WITH SOME STATE-OF-THE-ART ROAD EXTRACTION METHODS ON THE SPACENET DATASET (%). COANET-UB DENOTES THE UPPER BOUND FOR OUR COANET WITH THE GROUND TRUTH OF CONNECTIVITY BRANCH

	F1	IoU	APLS
DeepRoadMapper [1] $_{ICCV17}$	71.47	55.61	46.76
Topology Loss [26] $_{CVPR18}$	58.44	41.29	39.08
LinkNet34 [15] $_{VCIP17}$	73.96	58.68	63.12
D-LinkNet [39] $_{CVPRW18}$	69.77	53.57	50.20
RoadCNN [5] $_{CVPR18}$	73.74	58.40	59.39
ImprovedConnectivity [6] $_{CVPR19}$	75.91	61.17	62.81
VecRoad [69] $_{CVPR20}$	63.63	46.65	61.64
CoANet	**76.91**	**62.48**	**65.53**
CoANet-UB	**85.54**	**74.73**	**76.98**

TABLE II

QUANTITATIVE COMPARISON OF OUR PROPOSED COANET WITH SOME STATE-OF-THE-ART ROAD EXTRACTION METHODS ON THE DEEPGLOBE DATASET (%). COANET-UB DENOTES THE UPPER BOUND FOR OUR COANET WITH THE GROUND TRUTH OF CONNECTIVITY BRANCH

	F1	IoU	APLS
DeepRoadMapper [1] $_{ICCV17}$	78.04	63.98	58.85
Topology Loss [26] $_{CVPR18}$	56.07	38.95	46.99
LinkNet34 [15] $_{VCIP17}$	79.65	66.18	72.93
D-LinkNet [39] $_{CVPRW18}$	77.49	63.26	71.81
RoadCNN [5] $_{CVPR18}$	79.08	65.40	71.15
ImprovedConnectivity [6] $_{CVPR19}$	79.93	66.58	71.69
CoANet	**81.22**	**68.37**	**73.48**
CoANet-UB	**89.25**	**80.58**	**85.14**

where the learning rate is multiplied by $(1 - \frac{iter}{maxiter})^{power}$ with $power = 3$. And our method is performed using the machine learning framework PyTorch [67], the experiments are implemented on 4 NVIDIA RTX TITAN GPUs with 24GB memory. Both PyTorch [67] and Jittor [68] versions of the source code will be made publicly available. During training, the data augmentation including random rotation, horizontal flipping, rescaling, and gaussian blurring are applied to improve the generalization of the model. Finally, the images are cropped to a fixed size of 512×512 for both datasets.

In the inference phase, we utilize the predictions of connectivity branch to enhance the results of road extraction. There are eight channels in the output of the connectivity attention module, where the prediction of each channel can be regarded as a sub-problem of the segmentation task. For example, given the predicted connectivity cube O, if $\sigma(O_{i,j,c}) > t$, the pixel in location (i, j) is connected to its neighboring pixel in position c and both of them are road pixels. $\sigma()$ is a sigmoid nonlinearity function and t is a threshold value. We estimate each channel of O and sum it up along the channel dimension, then we can get a one-dimensional road mask. It is added to the output of the segmentation branch to get the final road extraction results.

D. Comparison With State-of-the-Art Methods

In this section, we compare the performance of our CoANet with several state-of-the-art road extraction methods on the SpaceNet and DeepGlobe datasets, including DeepRoadMapper [1], Topology Loss [26], LinkNet34 [15], D-LinkNet [39], RoadCNN [5], ImprovedConnectivity [6], and VecRoad [69]. In these methods, DeepRoadMapper [1] and ImprovedConnectivity [6] utilize post-processing steps to improve the connectivity of the road, while the other methods directly generate the extraction results.

1) Experiments on the SpaceNet Dataset: The quantitative experiment results on the SpaceNet dataset are listed in Table I. It is noted that our CoANet outperforms other methods in terms of pixel-based and graph-based metrics. For example, CoANet obtains an F1 score of 76.91% and an IoU score of 62.48%, which are better than ImprovedConnectivity [6] by 1.00% and 1.31%, respectively. As for the APLS, which is used to evaluate the topological correctness of roads, our method improves the second-best method LinkNet34 [15] by 2.41%. Since our proposed CoANet integrates the segmentation and relationship between neighboring pixels into one framework, it can extract roads with higher accuracy from satellite imagery. What's more, the use of connectivity attention module can further improve the topological connectivity of the road. We also compare a graph-based method VecRoad [69], which introduces an iterative graph exploration model with flexible steps. Our method obtains improvements of 15.83% on the IoU score and 3.89% on the APLS score. The graph-based methods usually guarantee the connectivity of the extracted roads, but there may be a large number of missing roads. Therefore our CoANet also has advantages over the graph-based methods. Besides, we devise an upper bound to our proposed method CoANet by using the ground truth of connectivity branch during inference. The upper bound results indicate that there is still large room for the connectivity branch to improve, and one possibility is to use a larger pixel interval in the future work.

SpaceNet [65] is a dataset where the remote sensing images are mainly collected from the urban areas, which contains a variety of road types in the city, such as motorway, residential, and highway, *etc*. There are also occlusions by buildings, shadows of buildings, and trees. Our method achieves the best results on pixel-based and graph-based metrics, which shows that the CoANet is able to deal with the complex urban traffic environment and extract roads with better connectivity.

2) Experiments on the DeepGlobe Dataset: Table II shows the quantitative results of our proposed CoANet compared with previous state-of-the-art methods on the DeepGlobe dataset. Since the VecRoad [69] needs the ground truth of road line string but DeepGlobe only has the pixel-level annotations, we do not show the experiment results of VecRoad. Our CoANet achieves an IoU score of 68.37%, which is better than other methods and improves the second-best method ImprovedConnectivity [6] by 1.79%. Besides, CoANet also obtains the best APLS score, which outperforms the LinkNet34 [15] by 0.55% and D-LinkNet [39] by 1.67%.

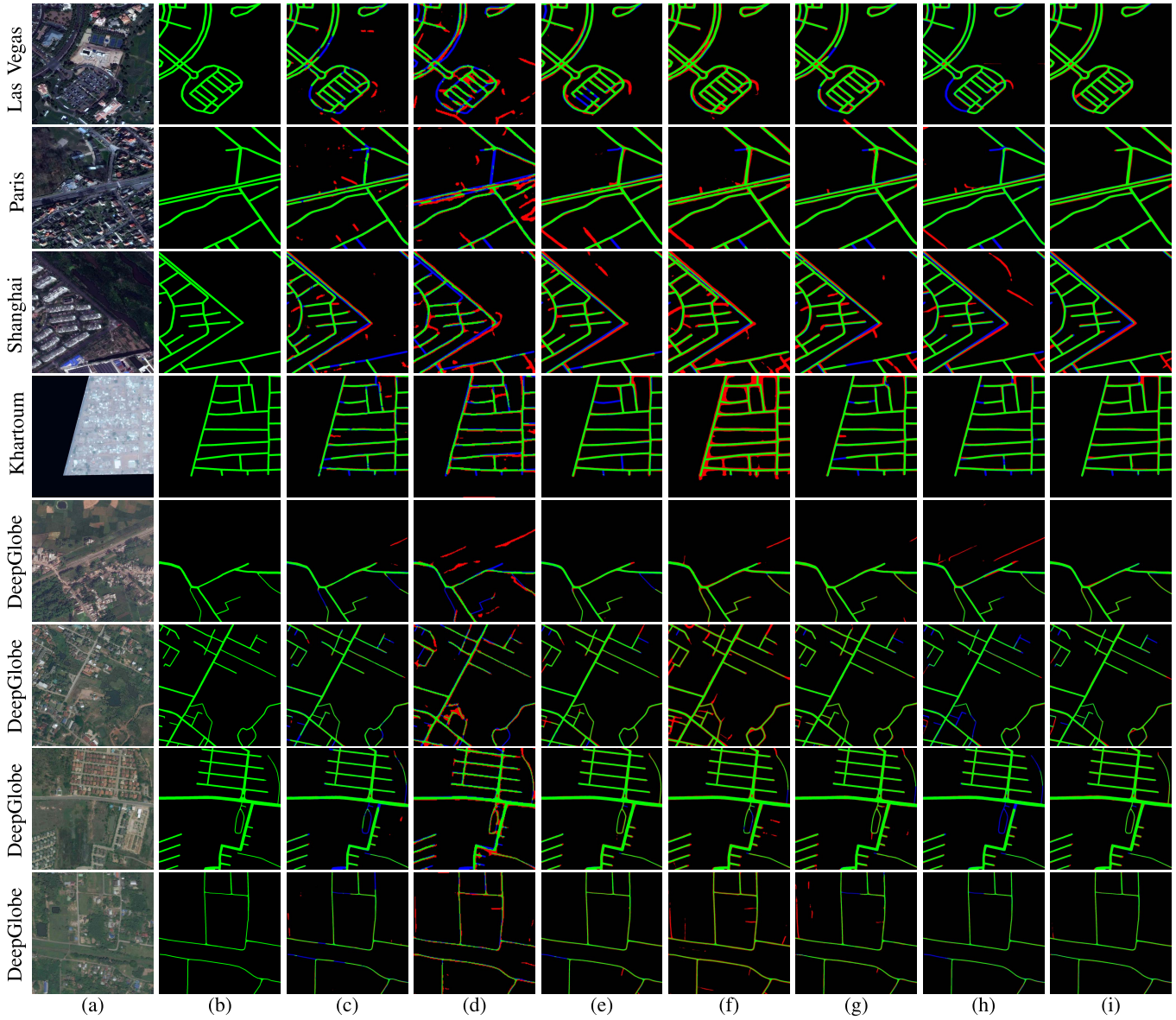


Fig. 5. Qualitative comparison of our CoANet with other state-of-the-art methods. Green: true positive, red: false positive, blue: false negative. The first to the fourth rows show the comparison results with different cities of SpaceNet [65], including Las Vegas, Paris, Shanghai, and Khartoum. The fifth to the eighth rows show the comparison results of DeepGlobe [41]. (a) Satellite imagery. (b) Ground truth. (c)-(i) Road extraction results of DeepRoadMapper [1], Topology Loss [26], LinkNet34 [15], D-LinkNet [39], RoadCNN [5], ImprovedConnectivity [6], and our CoANet. .

The remote sensing images in DeepGlobe [41] are mainly gathered from rural areas. It contains a large number of country roads, and the width of roads is constantly changing, which means that a road may have different widths. And there are severe occlusions caused by trees and shadows of trees. The experiment results show that our CoANet is also effective for the roads in rural areas. Combined with the results on the SpaceNet [65] dataset, it shows that our proposed method is robust to different road types in different regions.

3) Qualitative Comparison: The qualitative comparison results of our CoANet and other methods are illustrated in Fig. 5, which show four examples from different cities of SpaceNet [65] and four examples of DeepGlobe [41]. It is noted that the extracted roads of our method are consistent with those of ground truth and exist very few false positive pixels. In some regions where occlusions exist, the roads extracted by other methods may be disconnected, while our CoANet maintains the connectivity very well. For example, in the results of Las Vegas (the first row in Fig. 5), there is a parking lot in the lower-left corner of the image, where is parked a lot of vehicles and planted several trees. The roads extracted by DeepRoadMapper [1], Topology Loss [26], LinkNet34 [15], and ImprovedConnectivity [6] have many defects and fail to preserve the connectivity, but the result obtained by CoANet is consistent with the ground truth. In the results from DeepGlobe (eighth row in Fig. 5), where the roads are from rural areas and exist severe occlusions caused by trees. The roads extracted by other approaches have poor connectivity compared with our method. These visualized

TABLE III

ABLATION STUDY (%) FOR THE PROPOSED STRIP CONVOLUTION MODULE (SCM) AND CONNECTIVITY ATTENTION MODULE (COA). THE BASELINE IS THE SEGMENTATION MODEL BASED ON RESNET-101 (NO. 1). WE ADD THE SCM MODULE AND THE COA MODULE TO SHOW THE EFFECTIVENESS OF THEM (NO. 2 AND NO. 3). NO. 4 IS THE COMPLETE VERSION OF OUR PROPOSED COANET

No.	SCM	CoA	SpaceNet		DeepGlobe	
			IoU	APLS	IoU	APLS
1			59.57	58.69	63.09	69.13
2	✔		61.84	63.93	63.89	70.09
3		✔	61.10	61.27	64.32	70.45
4	✔	✔	**62.48**	**65.53**	**68.37**	**73.48**

results verify the superiority of our CoANet in the task of road extraction from satellite imagery.

E. Ablation Study

1) Effectiveness of Our Proposed Modules: In the CoANet model, SCM and CoA modules are proposed to capture the long-range relations in road regions and explore the dependencies between neighboring pixels, respectively. To validate the effectiveness of the two modules, we conduct experiments with different configurations, as shown in Table III. The baseline (No. 1) is the FCN model based on ResNet-101, when we add our proposed SCM and CoA to the baseline, the performance in terms of the IoU score increase from 59.57% to 61.84% and 61.10% on the SpaceNet dataset. As for the APLS score on the SpaceNet, adding SCM and CoA improve the baseline by 5.24% and 2.58%. After combining the SCM and CoA, we achieve a 2.91% improvement on the IoU score and a 6.84% improvement on the APLS score. The experiment performance on the DeepGlobe dataset is also improved by adding the two modules. These results verify that our proposed two modules are effective for road extraction.

The qualitative results of our method under different settings are illustrated in Fig. 6. There are some broken road segments in the results of the baseline, especially in the regions occluded by trees. After adding the CoA module, most broken segments in the baseline are connected, while the edges of the roads are coarse and there are some discrete points identified as roads. The SCM module helps improve the connectivity and the roads extracted are smoother. However, since the SCM is developed to capture long-range dependencies, some areas that are other classes may be recognized as roads and it makes the road longer than it actually is. It is noted that there are some disadvantages if we use the two modules separately. The SCM module connects the broken road segments where the neighboring pixels in the CoA module can not reach, and the CoA module can prevent the irrelevant noise from the background. Therefore, our CoANet obtains the best road extraction results by combining the two modules.

To verify the effectiveness of our connectivity attention module on the task of road extraction, we also compare it with the module that learns the pixel affinity. In [60], AffinityNet

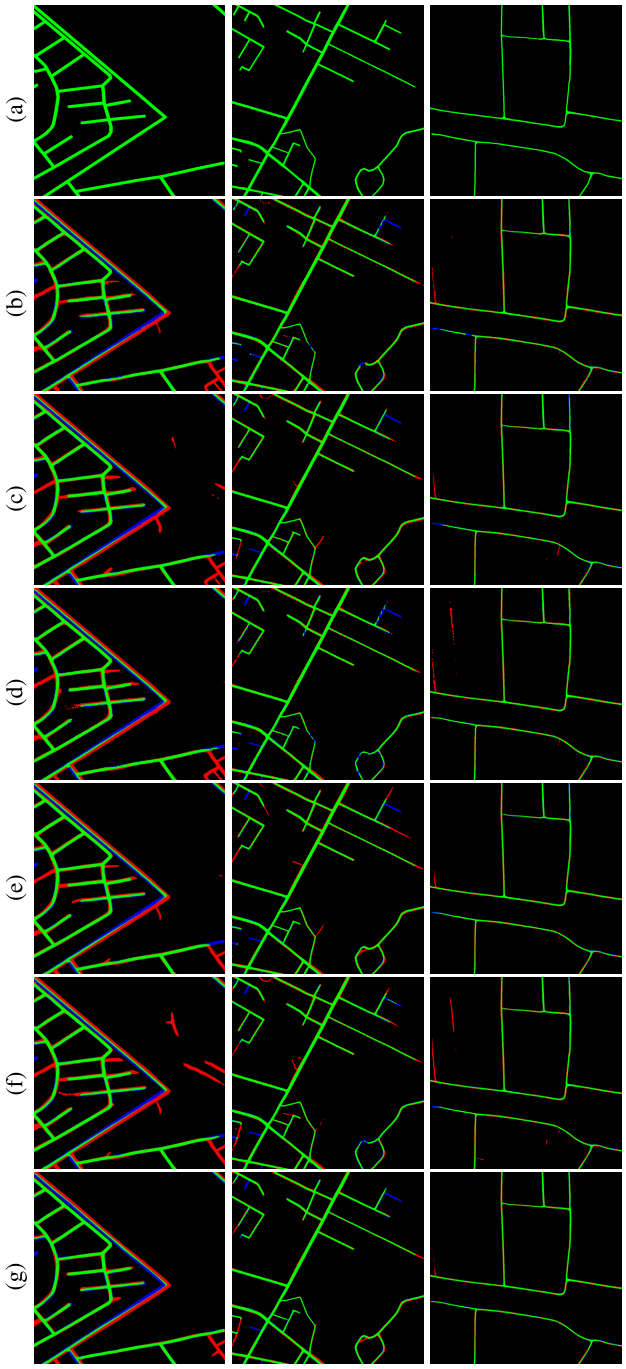


Fig. 6. Visual results of the proposed method under different model configurations. Green: true positive, red: false positive, blue: false negative. The first to the third columns show three samples from SpaceNet [65] and DeepGlobe [41]. (a) Ground truth. (b) Baseline. (c) Baseline + SCM. (d) Baseline + CoA. (e) CoANet contains only one CoA module with $d = 1$. (f) CoANet contains three CoA modules. (g) CoANet. .

is developed to learn class-agnostic semantic affinity between a pair of adjacent coordinates on an image, which is similar to our CoA module. We replace the proposed CoA module with the AffinityNet module in [60], and the experimental results are shown in Table IV. The performance of our CoANet is better than CoANet-Affinity by 5.94% on the IoU score and 3.94% on the APLS score for the DeepGlobe dataset. The AffinityNet in [60] assigns the affinity label of two

TABLE IV

ABLATION STUDY (%) FOR THE PROPOSED CONNECTIVITY ATTENTION MODULE (COA). COANET-AFFINITY DENOTES THE COANET THAT THE PROPOSED CONNECTIVITY ATTENTION MODULE IS REPLACED WITH THE AFFINITY MODULE IN [60]

	SpaceNet		DeepGlobe	
	IoU	APLS	IoU	APLS
CoANet-Affinity	61.93	62.67	62.43	69.54
CoANet	**62.48**	**65.53**	**68.37**	**73.48**

TABLE V

ABLATION STUDY (%) FOR THE PROPOSED CONNECTIVITY ATTENTION MODULE (COA) WITH DIFFERENT CONFIGURATIONS. $d1$, $d3$, AND $d5$ DENOTE THE COA MODULES WITH $d = 1$, $d = 3$, AND $d = 5$, RESPECTIVELY. NO. 1 IS THE PROPOSED COANET THAT CONTAINS ONLY ONE COA MODULE WITH $d = 1$. NO. 2 REPRESENTS THE COMPLETE VERSION OF OUR METHOD. AND NO. 3 DENOTES THE CONNECTIVITY BRANCH CONSISTS OF THREE COA MODULES

				SpaceNet		DeepGlobe	
No.	$d1$	$d3$	$d5$	IoU	APLS	IoU	APLS
1	✔			62.04	64.87	64.39	70.50
2	✔	✔		**62.48**	**65.53**	**68.37**	**73.48**
3	✔	✔	✔	62.09	65.13	64.89	70.83

adjacent coordinates to 1 if their classes are the same, and the coordinate pairs are sampled within a small radius. However, our CoA module explores the connectivity between neighboring road pixels and disentangle the background regions. And the neighboring pixels are sampled from eight specific directions around, where the directions are consistent with the distribution of most roads in the satellite imagery. There are two different settings for the interval between sampled pixels in our CoA module and it can capture multi-scale connectivity information, while the sampling radius in AffinityNet [60] is fixed. These advantages make our CoA module achieve better performance on the task of road extraction.

2) Effect of Different Configurations for CoA: As described in Sec. III-C, the connectivity branch in our proposed CoANet contains two CoA modules, one of which is $d = 1$ and another is $d = 3$. Here we analyze the effect of different configurations for the CoA module in the connectivity branch, the results are listed in Table V. We define that CoANet-*d1* denotes our proposed CoANet contains only one CoA module with $d = 1$, CoANet-*d5* represents that our CoANet consists of three CoA modules with $d = 1$, $d = 3$, and $d = 5$. CoANet-*d1* obtains 62.04% in terms of the IoU score and 64.87% in terms of the APLS score on SpaceNet [65]. After adding a CoA module with $d = 3$, the performance is improved to 62.48% on the IoU score and 65.53 on the APLS score. However, if the CoANet consists of three CoA modules (No. 3 in Table V), its performance is higher than CoANet-*d1* (No. 1) but lower than our CoANet that with two CoA modules (No. 2). The reason can be analyzed from the visualization results. As shown in Fig. 6, there are still some broken road segments in the results of CoANet-*d1*. Besides, since the intervals of a given pixel with its neighboring pixels are bigger, the road extraction results of CoANet-*d5* contain more background areas and are much longer. Therefore we select the configuration with two CoA modules, it obtains results that are more consistent with the ground truth.

TABLE VI

ABLATION STUDY (%) FOR THE PROPOSED STRIP CONVOLUTION MODULE (SCM) WITH DIFFERENT CONFIGURATIONS. 'H', 'V', 'L', AND 'R' DENOTE THE FOUR STRIP CONVOLUTIONS WITH DIFFERENT SHAPES: HORIZONTAL, VERTICAL, LEFT DIAGONAL, AND RIGHT DIAGONAL, RESPECTIVELY. SPECIALLY, NO. 1 IS THE SCM THAT CONTAINS TWO STRIP CONVOLUTIONS: HORIZONTAL AND VERTICAL. NO. 4 REPRESENTS THE COMPLETE VERSION OF OUR METHOD. NO. 5 DENOTES THAT THE SCM CONTAINS TWO CONVOLUTIONS FOR EACH SHAPE AND HAS A TOTAL OF EIGHT CONVOLUTIONS

		SpaceNet		DeepGlobe	
No.	SCM	IoU	APLS	IoU	APLS
1	H&V	61.85	64.58	64.57	70.65
2	H&V&L	62.01	65.01	64.69	70.92
3	H&V&R	61.93	64.85	64.97	70.74
4	H&V&L&R	**62.48**	**65.53**	**68.37**	**73.48**
5	2×H&V&L&R	61.98	65.41	65.19	71.03

3) Effect of Different Configurations for SCM: As shown in Table VI, we analyze the influence of different configurations for SCM in the decoder. When SCM contains horizontal and vertical strip convolutions, it obtains 61.85% on the IoU score and 64.58% on the APLS score for the SpaceNet dataset. The performance is improved by adding a type of strip convolution, like No. 2 and No. 3 in Table VI. After adding left diagonal and right diagonal strip convolutions, our CoANet achieves 62.48% on the IoU score and 65.53% on the APLS score. The above experimental results are in line with our expectations. With four strip convolutions of different shapes, our method can capture local context along with multiple spatial directions. It is noted that the four directions are consistent with most roads in satellite images and are relatively easy to implement. We also report results of the configuration that SCM contains two convolutions for each shape and has a total of eight convolutions. As listed in No. 5, it harms the improvement of performance and increases the computational cost due to the additional strip convolutions. Thus, we apply four strip convolutions in the SCM.

F. Discussion

1) Analysis of Failure Cases: As mentioned in the above experiments, our proposed CoANet achieves new state-of-the-art performance on two public datasets SpaceNet [65] and DeepGlobe [41]. However, there are still some failure cases for our approach. As illustrated in Fig. 7, we show three samples from the two datasets. For Fig. 7 (a), there are a tunnel and an overpass in the middle of the satellite image. Besides, there are server occlusions in road regions caused by trees and buildings, as shown in Fig. 7 (b) and (c). Since the occluded

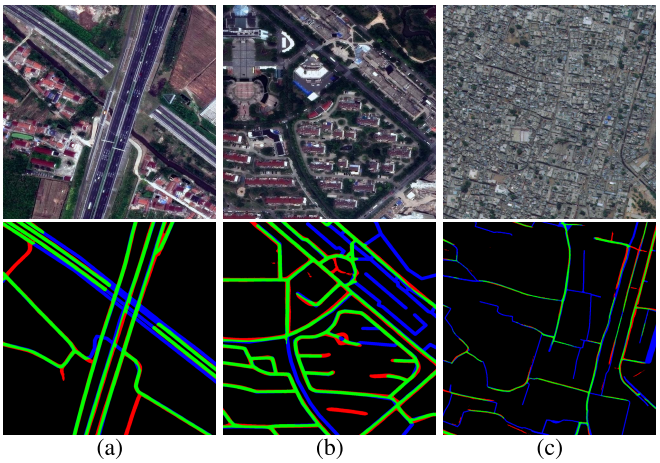


Fig. 7. Visualization of some cases that our CoANet fails. Green: true positive, red: false positive, blue: false negative. The first row is satellite imagery, the second row is road extraction results of our CoANet. (a)-(b) Three samples from SpaceNet [65] and DeepGlobe [41]. .

TABLE VII

RUNNING TIME OF OUR PROPOSED COANET AND SOME STATE-OF-THE-ART ROAD EXTRACTION METHODS UNDER THE SAME CONDITIONS. WE LIST THE TRAINING TIME FOR IMAGES OF AN IDENTICAL BATCH SIZE AND INFERENCE TIME FOR ONE IMAGE (S)

	Training	Inference
DeepRoadMapper [1] $_{ICCV17}$	0.579	0.121
Topology Loss [26] $_{CVPR18}$	0.168	0.049
LinkNet34 [15] $_{VCIP17}$	0.201	0.066
D-LinkNet [39] $_{CVPRW18}$	0.218	0.071
RoadCNN [5] $_{CVPR18}$	0.156	0.028
ImprovedConnectivity [6] $_{CVPR19}$	0.361	0.074
CoANet	0.146	0.022

areas are very large and the roads in these areas may not be visible in the satellite images, our CoANet fails to generate roads and preserve road connectivity in these areas. In the future, we will consider adding other information to extract the roads in these challenging areas, such as the GPS trajectories of pedestrians and cars. What's more, the road network in satellite imagery can be regarded as a graph with edges and nodes. We can take advantage of the graph convolutional network to extract the road, which may be effective for roads in occluded areas.

2) Analysis of Running Time: As listed in Table VII, we analyze the running time of our CoANet and several state-of-the-art road extraction methods on the SpaceNet [65] dataset. The comparison experiments of all methods are executed on a workstation with 4 NVIDIA RTX TITAN GPUs. For fair comparison, we report the training time for images of an identical batch size and inference time for one image with the size of 512×512. It is noted that our CoANet achieves the fastest training time and inference time. In addition, the methods that utilize post-processing steps, such as DeepRoadMapper [1] and ImprovedConnectivity [6], require more time for training and inference. With better performance and faster execution, our proposed CoANet is more suitable for extracting roads from satellite imagery.

V. CONCLUSION

In this paper, we propose a connectivity attention network (CoANet) for road extraction from satellite imagery, which jointly learns the segmentation and pair-wise dependencies. We first introduce an encoder-decoder architecture network to learn the feature of roads. Motivated by the shape of roads, which are long-span, narrow, and distributed continuously. We propose a strip convolution module (SCM) since it is more aligned with the shape of roads. The SCM leverages four strip convolutions to capture long-range context information from four different directions, which prevents irrelevant regions from interfering the feature learning. What's more, to alleviate the occlusions in road regions caused by the buildings and trees, a connectivity attention module (CoA) is developed to explore the relationship between neighboring pixels. The connectivity of a given pixel with eight neighboring pixels is predicted, which incorporates the graphical information and enables the connectivity of roads are better preserved. Extensive experiments on popular benchmarks (SpaceNet and DeepGlobe datasets) demonstrate the superiority of our proposed CoANet compared with several state-of-the-art methods. We also perform ablation experiments to show the effectiveness of the SCM and CoA modules, and provide insights into the choices of different configurations for the CoA module. In the future, we aim to exploit more tasks that the connectivity attention module can be used, *e.g.,* salient segmentation and semantic segmentation, since the prediction of connectivity cube can be regarded as a series of subproblems of segmentation task. Besides, we can treat the road network in satellite imagery as a graph and take advantage of the graph convolutional network to extract the road.

REFERENCES

[1] G. Máttyus, 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] A. V. Etten, "City-scale road extraction from satellite imagery v2: Road speeds and travel times," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1775–1784.

[3] R. Alshehhi and P. R. Marpu, "Hierarchical graph-based segmentation for extracting road networks from high-resolution satellite images," *ISPRS J. Photogramm. Remote Sens.*, vol. 126, pp. 245–260, Apr. 2017.

[4] B. Liu, H. Wu, Y. Wang, and W. Liu, "Main road extraction from ZY-3 grayscale imagery based on directional mathematical morphology and VGI prior knowledge in urban areas," *PLoS ONE*, vol. 10, no. 9, Sep. 2015, Art. no. e0138071.

[5] 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.

[6] A. Batra, S. Singh, G. Pang, S. Basu, C. V. Jawahar, and M. Paluri, "Improved road connectivity by joint learning of orientation and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10385–10393.

[7] M. Barzohar and D. B. Cooper, "Automatic finding of main roads in aerial images by using geometric-stochastic models and estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 7, pp. 707–721, Jul. 1996.

[8] I. Laptev, H. Mayer, T. Lindeberg, W. Eckstein, C. Steger, and A. Baumgartner, "Automatic extraction of roads from aerial images based on scale space and snakes," *Mach. Vis. Appl.*, vol. 12, no. 1, pp. 23–31, 2000.

[9] D. Chai, W. Forstner, and F. Lafarge, "Recovering line-networks in images by junction-point processes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1894–1901.

[10] 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.

[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] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image. Comput. Comput. Assist. Interv.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[14] 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.

[15] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[16] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[18] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[19] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 210–223.

[20] 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.

[21] 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.

[22] 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.

[23] J. D. Wegner, J. A. Montoya-Zegarra, and K. Schindler, "Road networks as collections of minimum cost paths," *ISPRS J. Photogramm. Remote Sens.*, vol. 108, pp. 128–137, Oct. 2015.

[24] V. Mnih and G. E. Hinton, "Learning to label aerial images from noisy data," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 567–574.

[25] D. Costea and M. Leordeanu, "Aerial image geolocalization from recognition and matching of roads and intersections," 2016, *arXiv:1605.08323*. [Online]. Available: http://arxiv.org/abs/1605.08323

[26] A. Mosinska, P. Marquez-Neila, M. Koziński, and P. Fua, "Beyond the pixel-wise loss for topology-aware delineation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3136–3145.

[27] 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. 2018.

[28] H. Kong, J.-Y. Audibert, and J. Ponce, "General road detection from a single image," *IEEE Trans. Image Process.*, vol. 19, no. 8, pp. 2211–2220, Aug. 2010.

[29] W. Song, J. M. Keller, T. L. Haithcoat, and C. H. Davis, "Automated geospatial conflation of vector road maps to high resolution imagery," *IEEE Trans. Image Process.*, vol. 18, no. 2, pp. 388–400, Feb. 2008.

[30] M. Amo, F. Martínez, and M. Torre, "Road extraction from aerial images using a region competition algorithm," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1192–1201, May 2006.

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

[32] J. D. Wegner, J. A. Montoya-Zegarra, and K. Schindler, "A higher-order CRF model for road network extraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1698–1705.

[33] S. H. Gao, M. M. Cheng, and K. Zhao, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[34] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.

[35] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 11–19.

[36] 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 2018.

[37] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5229–5238.

[38] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. Dalla Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 139–149, Aug. 2017.

[39] 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.

[40] G. Mattyus and R. Urtasun, "Matching adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8024–8032.

[41] 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. 172–179.

[42] X. Li, Y. Wang, L. Zhang, S. Liu, J. Mei, and Y. Li, "Topology-enhanced urban road extraction via a geographic feature-enhanced network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8819–8830, Dec. 2020.

[43] 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.

[44] 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.

[45] 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.

[46] J. Hu, A. Razdan, J. C. Femiani, M. Cui, and P. Wonka, "Road network extraction and intersection detection from aerial images by tracking road footprints," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4144–4157, Dec. 2007.

[47] J. Biagioni and J. Eriksson, "Map inference in the face of noise and disparity," in *Proc. 20th Int. Conf. Adv. Geographic Inf. Syst. (SIGSPATIAL)*, 2012, pp. 79–88.

[48] Z. Shan, H. Wu, W. Sun, and B. Zheng, "COBWEB: A robust map update system using GPS trajectories," in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput. (UbiComp)*, 2015, pp. 927–937.

[49] R. Stanojevic, S. Abbar, S. Thirumuruganathan, S. Chawla, F. Filali, and A. Aleimat, "Kharita: Robust map inference using graph spanners," 2017, *arXiv:1702.06025*. [Online]. Available: http://arxiv.org/abs/1702.06025

[50] 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.

[51] T. Sun, Z. Di, P. Che, C. Liu, and Y. Wang, "Leveraging crowdsourced GPS data for road extraction from aerial imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7509–7518.

[52] F. Shen, R. Gan, S. Yan, and G. Zeng, "Semantic segmentation via structured patch prediction, context CRF and guidance CRF," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1953–1961.

[53] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[54] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.

[55] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[56] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1448–1457.

[57] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," 2021, *arXiv:2101.11939*. [Online]. Available: http://arxiv.org/abs/2101.11939

[58] G. Bertasius, L. Torresani, S. X. Yu, and J. Shi, "Convolutional random walk networks for semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 858–866.

[59] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3029–3037.

[60] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4981–4990.

[61] Y. Hou, Z. Ma, C. Liu, T.-W. Hui, and C. C. Loy, "Inter-region affinity distillation for road marking segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12486–12495.

[62] 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.

[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[64] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[65] 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

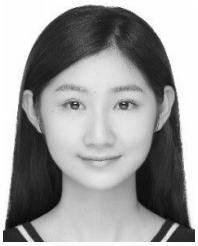
[66] S. Singh *et al.*, "Self-supervised feature learning for semantic segmentation of overhead imagery," in *Proc. Brit. Mach. Vis. Conf.*, 2018, vol. 1, no. 2, p. 4.

[67] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inform. Process. Syst.*, 2019, pp. 8026–8037.

[68] S.-M. Hu, D. Liang, G.-Y. Yang, G.-W. Yang, and W.-Y. Zhou, "Jittor: A novel deep learning framework with meta-operators and unified graph execution," *Sci. China Inf. Sci.*, vol. 63, no. 12, pp. 1–21, Dec. 2020.

[69] Y.-Q. Tan, S.-H. Gao, X.-Y. Li, M.-M. Cheng, and B. Ren, "VecRoad: Point-based iterative graph exploration for road graphs extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8910–8918.

Jie Mei is currently pursuing the Ph.D. degree with the College of Computer Science, Nankai University. He is supervised via Prof. Ming-Ming Cheng. His research interests include computer vision, machine learning, and remote sensing image processing.



Rou-Jing Li is currently pursuing the master's degree with the Faculty of Geographical Science, Beijing Normal University, Beijing, China. Her research interests include remote sensing image processing, machine learning, and spatial analysis.

Wang Gao received the master's degree from the Third Research Institute, China Aerospace Science and Industry Corporation, in 2017. He is currently with the Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing, China. His research interests include computer vision, scene matching, and visual navigation.

Ming-Ming Cheng (Senior Member, IEEE) received the Ph.D. degree from Tsinghua University in 2012. He was a Research Fellow with Prof. Philip Torr at Oxford. He is currently a Professor at Nankai University, leading the Media Computing Laboratory. His research interests include computer graphics, computer vision, and image processing. He received research awards, including the ACM China Rising Star Award, the IBM Global SUR Award, and the CCF-Intel Young Faculty Researcher Program. He is on the Editorial Board of IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP).