LGNet: Location-Guided Network for Road Extraction from Satellite Images

Jingtao Hu, Junyu Gao, *Member, IEEE,* Yuan Yuan, *Senior Member, IEEE,* Jocelyn Chanussot, *Fellow, IEEE,* and Qi Wang, *Senior Member, IEEE*

***Abstract*—Road connectivity is vital in road extraction for accurate vehicle navigation. However, the segmentation-based methods fail to model the connectivity resulting in broken road segments. Therefore, we propose a Location-Guided Network (LGNet) for promoting connectivity performance in a very effective and efficient way. Specifically, an auxiliary Road Location Prediction (RLP) task is designed to obtain global road connectivity information, which improves the performance of road segmentation. The RLP can predict the location coordinates of the whole roads with row anchors and column anchors. By aggregating the global location context to the segmentation branch with a location-guided decoder (LG-Decoder), the features can finally capture the connectivity of each road segment. Overall, LGNet has the following advantages: 1) The proposed RLP and LCG can plug into any encoder-decoder network and achieve an impressive performance. 2) High computational efficiency. In comparison with the multi-branch method, our proposed LGNet requires about $6\times$ fewer GFLOPs. 3) The superior road connectivity performance. A series of experiments are conducted on two road extraction data sets (SpaceNet and DeepGlobe), confirming the effectiveness of the LGNet.**

***Index Terms*—Road extraction, auxiliary task, road location prediction, location-guided decoder.**

I. Introduction

ROAD extraction, a fundamental problem in remote sensing, aims to reconstruct the road network in a timely and accurate manner from satellite images. It has many applications, such as road map updating [1], vehicle navigation [2], and urban planning [3], [4]. In particular, current road extraction methods based on segmentation have made significant progress. Most of these methods [5]–[10] consider the road extraction as a pixel-wise classification task. However, due to the complexity and diversity of the road network, the pixel-wise segmentation methods are limited to ensuring connectivity. The limited ability to connect road segments has a very negative impact on topological accuracy.

This work was supported by the National Natural Science Foundation of China under Grant U21B2041, 61825603, National Key R&D Program of China 2020YFB2103902.

Jingtao Hu is with the School of Computer Science, and with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P. R. China. (e-mail: jthu@mail.nwpu.edu.cn).

Jocelyn Chanussot is with the Universit Grenoble Alpes, INRIA, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, and also with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China. (e-mail:jocelyn.chanussot@grenoble-inp.fr).

Junyu Gao, Yuan Yuan and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China. (e-mail: gjy3035@gmail.com, y.yuan1.ieee@gmail.com, crabwq@gmail.com) (*Corresponding author: Qi Wang.*)

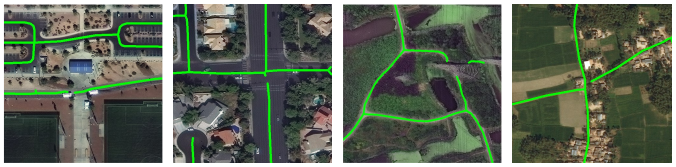


Fig. 1. Topological errors from the pix-wise segmentation methods, including incomplete junction, interrupt road segments, and faulty connections.

A large receptive field is required to extract complex road structure features effectively [11]. D-LinkNet [12] introduces multiple dilated convolutions to fuse the features from different receptive field and achieved good road segmentation. Deng *et al.* [13] integrates the strip-pooling strategy [14] into D-LinkNet, which captures the long-range context to adjust the road topology. Although the above methods attain better segmentation results, due to the shadows and occlusion, there are still topological errors, e.g., incomplete junction, interrupt road segment, and faulty connections as shown in Fig. 1.

In order to make up for the connectivity ability, some works have been proposed to introduce the road centerline task to benefit the road segmentation task [15]–[17]. Specifically, Cheng *et al.* [18] proposes a cascaded convolutional neural network (CNN) that includes a road detection network and a road centerline extraction network to capture the topology of the road. Further, Yang *et al.* [19] designs the recurrent CNN U-Net with the joint network to perform pixel-wise classification and road centerline extraction simultaneously. However, the road centerline task is still a pixel-wise classification task, which cannot explicitly increases the topological accuracy.

To obtain better road connectivity, some multi-task methods are proposed to predict the road orientation and pixel-wise binary segmentation, simultaneously. Batra *et al.* [20] focuses on classifying the road orientation angles using a multi-branch model. DiresNet learns to infer the local road directions via an asymmetric residual segmentation network [21]. The road orientation prediction task introduces additional supervised information that effectively improves road connectivity. However, these orientation-based tasks aggregate the context information in a homogeneous manner, which does not meet the requirement that road segments need different contextual dependencies. Most of the above methods perform feature fusion at the end of models, resulting in insufficient information fusion.

In this paper, we strive to achieve road connectivity while providing sufficient information aggregation across tasks. An auxiliary task, Road Location Prediction (RLP), is proposed

to learn the road connectivity. Specifically, each road segment is considered a road instance and obtained its horizontal and vertical coordinates by road anchors. Then, the abscissa or ordinate of each road instance is predicted by the RLP. Without loss of generality, two branches are used to predict the location coordinates so that the coordinate values can remain unique. To distinguish each road segment, a new road formulation is defined to parse all the roads. Meanwhile, we use the whole feature maps to infer the connectivity, which gets the global receptive field.

To effectively leverage location context information, we propose the location-guided decoder (LG-Decoder). Firstly, the feature maps from RLP are input into LCA to aggregate the global location context information. Then, we feed the output into each LG-Decoder block, which includes a location context guidance (LCG) module to achieve a better fusion of the location information and the features from the previous layer. Finally, road connectivity can be improved from the location context information. The labels of the RLP are directly extracted from the segmentation masks, and outputs are convolutional feature maps. For the above reasons, the location-guided decoder can be conveniently plugged into any encoder-decoder based road segmentation network.

To summarize, our main contributions are the following:

(1) We propose a multi-task architecture that combines road segmentation and road location prediction, effectively learning road connectivity and improving road segmentation performance.

(2) We propose the road location-guided decoder, which can aggregate the road location contextual information into the segmentation network. It can be plugged into various encoder-decoder-based segmentation networks conveniently.

(3) Extensive experiments on the SpaceNet and DeepGlobe data sets prove that LGNet is competitive and more efficient than the state-of-the-art methods.

The remaining parts of the paper proceed as follows: The section II examines the related work of the road extraction. Next, we describe the overall architecture of the proposed LGNet in section III. Section IV analyses the ablation study and the comparison experiments. Finally, the conclusion and future work are present in Section V

II. Related Work

A. Segmentation-based Road Extraction

Recently, CNN-based methods are the most widely used for road extraction due to the powerful feature representation and generalization ability [22], [23]. Here, we briefly review the encoder-decoder based methods for road segmentation. FCN [24] is the first full convolutional neural network, which uses skip connection to fuse feature maps from different levels. U-Net [25] combines multi-level features with a U-shape encoder-decoder architecture and achieves a better segmentation performance with less training data. For more efficiency, LinkNet [26] combines downsampling and deconvolution in the decoder to achieve the trade-off of accuracy and efficiency. Most road segmentation methods are improved based on the above baseline methods. CasNet [18] uses two CNNs for road segmentation and centerline extraction, respectively. In the centerline extraction network, the input features are from the decoder of road segmentation, which bridge the two tasks together. To combine with U-Net and residual units in ResNet, deep residual U-Net (ResUNet) [27] outperforms U-Net in road segmentation with fewer parameters. Similar to the ResUNet, the dense block [28] and skip connections are combined with the U-Net [29], strengthening the fusion of multi-level features. To increase the reconstruction ability of the decoder, [30] integrates multiple parallel upsampling structures to the decoder layers, extracting better multi-scale features. Instead of the convolutional units in U-Net, [19] proposes the recurrent units, which use multiple summation operations to preserve detailed spatial information. In [16], the encoder of U-Net is integrated with multi-scale features, which take advantage of the spatial information to improve the feature extraction. Dilated convolutions are also applied to enlarge the receptive field size for high-resolution satellite images [12], [31].

B. Multi-Task Learning

Multi-task learning (MTL) is a training pattern where multiple sub-tasks are performed with a shared network simultaneously [32]. With these architectures, we gain advantages such as improved data efficiency, reduced overfitting, and fast learning [33]. There have been successful applications of multi-task learning in a wide variety of fields, including natural language processing [34]–[37] and computer vision [38]–[41].

In the road extraction field, segmentation-based methods are often trained together with centerline extraction as an auxiliary task [15]–[19]. Despite the centerline extraction task, [20] predicts the road orientations for promoting road connectivity. With feature fusion and multi-branch architecture, the orientation task can efficiently improve road connectivity. By joint segmentation and path classification, [42] reduces the number of disconnected road segments. [43] designs three sub-tasks including pixel-level, edge-level and region-level classification. The MTL framework integrates different levels of features, enhancing the topology of road segmentation. In our work, we design a new auxiliary task, road location prediction (RLP), which uses multi-task learning to correct the disconnected road segments.

III. Methodology

In this section, we give the details of the proposed LGNet for road extraction. We first describe the overview of our LGNet. Then, the auxiliary road location prediction task will be introduced. It captures the location context in horizontal and vertical directions. To improve road connectivity, we propose the location-guided decoder module for transferring the location context to the road segmentation branch. To obtain the global location context, we use the attention-based approach for location context aggregation [44]–[46]. Meanwhile, we propose to use the location context information to guide the feature map of the decoder of segmentation task for correcting the disconnect road segments. In the following section, we demonstrate how to build LGNet upon an encoder-decoder based network.

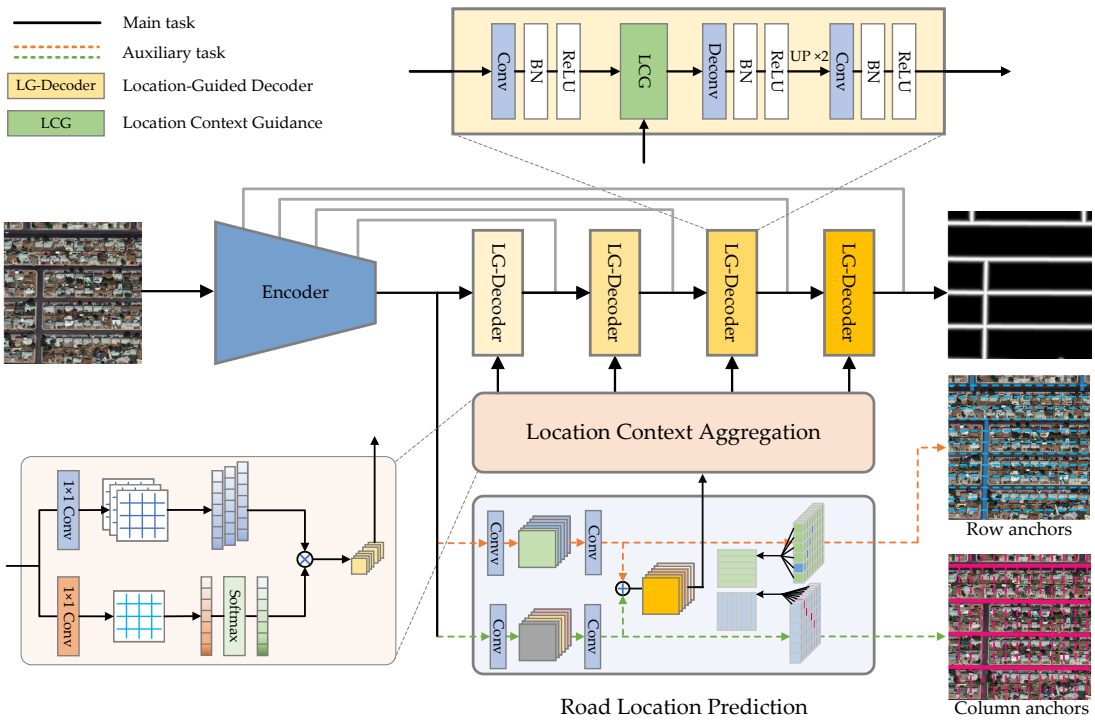


Fig. 2. The overall architecture of the proposed LGNet for road extraction.

A. Network Architecture

LGNet consists of three components, a shared encoder based on a deep convolution neural network(DCNN), two task-specific decoders including road segmentation and road location prediction, and the location-guided decoder across the two tasks. The network architecture is shown in Fig. 2. A satellite image is input DCNN to produce a feature map F with the spatial size of $H \times W$. For each branch in the RLP task, given the F, we first apply a convolutional layer with batch normalization and activation to obtain the feature maps of road locations. Then the feature maps are followed by a 1×1 convolutional layer to predict the road location results. Then, we apply a point-wise addition operation to the outputs of the two branches that can obtain the global road location information F'. To get the richer location context information, feature map F' is fed into location context aggregation (LCA). Therefore, the output feature map F'' gathers the information from all positions. In the segmentation task, the feature map F'' is fed into each decoder block through location context guidance (LCG). Finally, the fused features through the successive location-guided decoder (LG-Decoder) blocks are passed to the final layers of segmentation to get the predicted mask of road segmentation. As mentioned above, the modules, *ie.*, RLP, LCA, LCG, and LG-Decoder participate in the training and are trained simultaneously.

B. Road Instance Formulation

As mentioned in the introduction section, shadows and occlusion mainly affect road extraction performance. Due to the weak ability of learning road connectivity, the segmentation-based methods are hard to solve the above problems. A road formulation is proposed to extract individual road instances which are used to capture the global context of each road segment. Inspired by [47], we define several line anchors [48] including row anchors and column anchors to obtain locations of each road instance, as shown in Fig. 3. The road location prediction is to predict the coordinate x or y of having a road at anchors, aiming to enable independent prediction for each road instance. Due to the complex topological structure of the road, line anchors are divided into row anchors and column anchors, which are used to tackle location confusion of road instances parallel to the x-axis or y-axis.

The location coordination must be keep unique in each line anchor. The road can intuitively be divided into different categories based on its direction. However, more than one road instance may have the same orientation in a dense road scene. Thus, we sort all the road segments depending on the starting point and select N_s instances from the same directions. We can obtain the number of road types $N_t = N_d \times N_s$, in which N_d belongs to number of orientations and N_s belongs to the number of road instances with the same orientation. To facilitate the training, the N_s is set to a fixed number. Also, it is essential to determine the grid size of road location $\hat{H} \times \hat{W}$ in which $\hat{H}$ is the number of anchors and $\hat{W}$ is the length of anchors. The value of $\hat{H}$ and $\hat{W}$ is preferably larger than the road width to contain the whole road instance. Considering that the roads have different widths, we perform a detailed parameter analysis in section IV-C. The ground truth of road locations can be extracted from the road masks and does not require any extra annotation effort.

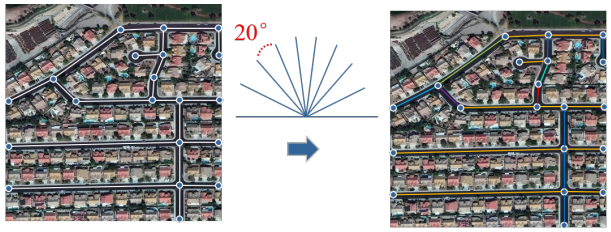


Fig. 3. Road formulation parsed the road according to the directions.

C. Road Location Prediction Task

We introduce the road location prediction which is used to promote the connectivity representation ability of the features of CNN. RLP includes two branches to predict the road location with row anchors and column anchors. In one branch, the last feature maps of encoder F_{global} are pass through a 3×3 convolutional layer with batch normalization and LeakyReLU to encode the location information. Then, the output features are input a 1×1 convolutional layer which changes the feature maps into a half channel dimension. The feature maps from two branches are finally summed to the location context agregtion module, as shown in Fig. 2. $\mathbf{Y}_{row} \in \mathbb{R}^{(\hat{W}+1)\times\hat{H}\times N_t}$ represents the prediction of the row anchors branch, where $\mathbf{Y}_{row} = \left\{y^1_{row}, y^2_{row}, \ldots, y^{\hat{W}}_{row}, y^{\hat{W}+1}_{row}\right\}$, in which $y^{\hat{W}}_{row} \in \mathbb{R}^{\hat{H}\times N_t}$, containing the correct location index for each road instance. $\mathbf{Y}_{col} = \left\{y^1_{col}, y^2_{col}, \ldots, y^{\hat{W}}_{col}, y^{\hat{W}+1}_{col}\right\}$ represents the prediction of the column anchors branch, in which $y^{\hat{W}}_{col} \in \mathbb{R}^{\hat{H}\times N_t}$. Suppose $g^{i,j}_{row}$ is the classifier which is used for predict the road location coordinates on the i-th row anchor, j-the road type. Then, the location prediction for road instances can be described as:

$$\mathbf{Y}^{i,j}_{row} = g^{i,j}_{row}(F_{global}), \text{ s.t. } i \in [1, \hat{H}], j \in [1, N_t], \quad (1)$$

$$\mathbf{Y}^{i,j}_{col} = g^{i,j}_{col}(F_{global}), \text{ s.t. } i \in [1, \hat{H}], j \in [1, N_t], \quad (2)$$

in which $\mathbf{Y}^{i,j}_{row}$ and $\mathbf{Y}^{i,j}_{col}$ represents the probability of selecting $(\hat{W}+1)$ grid blocks for the i-th anchor, j-the road type. The maximum probabilities represent the correct predicted locations and the ground truth of road locations is $\mathbf{T}_{row} \in \mathbb{R}^{\hat{H}\times N_t}$ and $\mathbf{T}_{col} \in \mathbb{R}^{\hat{H}\times N_t}$. Then, the optimization corresponds to:

$$L_{row} = L_{FL}(\mathbf{Y}_{row}, \mathbf{T}_{row}), \quad (3)$$

$$L_{col} = L_{FL}(\mathbf{Y}_{col}, \mathbf{T}_{col}), \quad (4)$$

in which L_{FL} represents the Focal Loss [49].

D. Location-Guided Decoder

Combined with the auxiliary RLP task, the features of the encoder are properly refined by the road location information. Meanwhile, the road location information can be aggregated with a decoder to improve the final segmentation results. Thus, we propose the location-guided decoder (LG-Decoder) module to fuse road location context effectively and efficiently. Before fusion, we need the location information to be more discriminative and less noisy. Thus, the lightweight non-local module

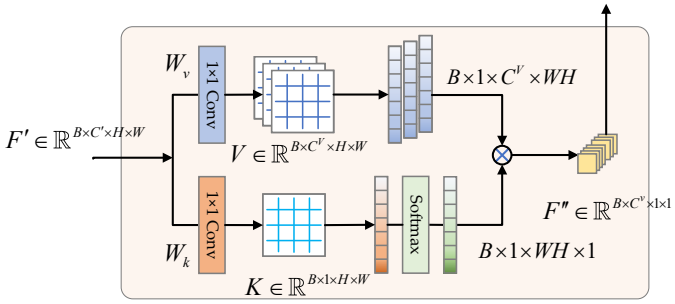


Fig. 4. Our proposed location context aggregation module.

named location context aggregation is used to aggregate road location features in different directions. Then, the enhanced location features are input to each decoder block to filter the features from the previous layer through the location context guidance module. By directly learning road locations and integrating them into segmentation networks, road connectivity can be effectively improved. In the following paragraph, we give the details about the location context aggregation module and the location context guidance module.

1) Location Context Aggregation(LCA): For aggregating road location information, we introduce a location context aggregation (LCA) module to model the location context dependencies of all the road instances using lightweight computation and memory. LCA is a simplified non-local block following the fact that variants with and without the query achieve the comparable performance [45].As shown in Fig. 4, the local location context feature map $F^{'} \in \mathbb{R}^{B\times C^{'}\times H\times W}$ first inputs two 1×1 convolutions (W_k and W_v) to generate the feature maps K and V, respectively, where $K \in \mathbb{R}^{B\times1\times H\times W}$ and $V \in \mathbb{R}^{B\times C^{V}\times H\times W}$. Then, the global location context $F^{''} \in \mathbb{R}^{B\times C^{V}\times1\times1}$ is obtained with matrix product operation. The operation defined as follow.

$$F^{''} = \sum_{i=1}^{N}(W_v \cdot F^{'}_i)\frac{\exp W_k F^{'}_j}{\sum_{j=1}^{N}\exp(W_k F^{'}_j)} \quad (5)$$

Finally, we feed the global location context F'' to each of the decoder block to strengthen the road connectivity.

2) Location Context Guidance(LCG): Take LinkNet as an example, and a 1×1 convolution is first employed to minimize the channel dimension of features to reduce computational cost within each LG-Decoder block. The features are then combined with the global location context using the location context guidance (LCG), which is shown in Fig. 5. Within the LCG, a depth-wise convolution operation conditioned on the road location representation is used to filter the input features. Specifically, two fully connected layers are applied to the location representation, followed by reshape operation to produce filters. The input features are then processed with a 3×3 depth-wise convolution and a 1×1 convolution to produce F_1. Moreover, LCG also learns to aggregate the global location context to the filtered features. Specifically, $F^{''}$ which represents the global location information is passed through a convolution layer to reduced the dimension in the LG-Decoder. Then, the location feature is passed to two 1×1 convolutional

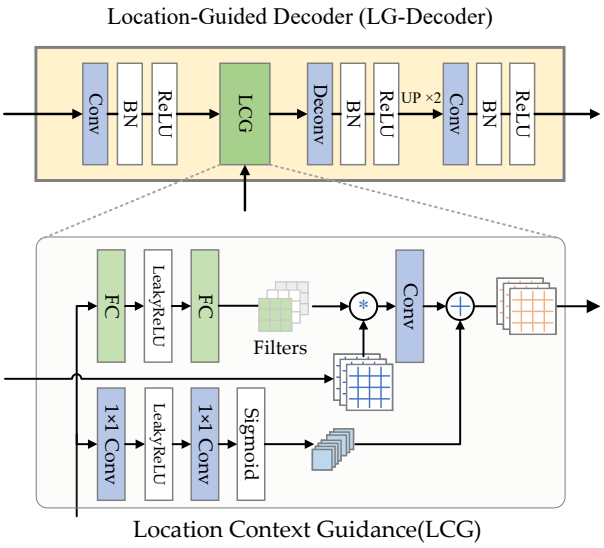


Fig. 5. Our proposed location context guidance module.

layers to generate channel-wise weighted features F_2. Finally, we can get the output feature which is the sum of F_1 and F_2 . The LCG can be embedded into other encoder-decoder based architectures before the upsampling operation in the decoder.

E. Loss Function

We use the binary cross-entropy (BCE), and Dice loss [50] to train the road segmentation task with direct addition:

$$L_{seg} = L_{BCE} + L_{Dice}, \tag{6}$$

The final learning objective function is shown as follow:

$$Loss = L_{seg} + \lambda L_{row} + \beta L_{col}, \tag{7}$$

where λ and β are scalar weights that control the influence of auxiliary losses.

IV. Experiments and Analyses

To verify the effectiveness of the LGNet, experiments are carried out on the two public datasets, SpaceNet [51] and DeepGlobe [52]. Experimental results demonstrate that LGNet achieves the state-of-the-art road connectivity performance on Spacenet. In the following subsections, we first introduce the datasets, the evaluation metrics, and training details, then we perform a series of ablation experiments on Spacenet and DeepGlobe data sets. Finally, we report our results on the above two datasets.

A. Datasets and Evaluation Metrics

1) SpaceNet: The SpaceNet data set [51] provides 2,780 images from four different cities with pixel resolution of 1300×1300 and the ground sample distance (GSD) is 0.3m/pixel. We randomly split the data set with the percentage of 8:2 for training and validation. Following [20], we crop each image into 650×650 patches with the overlapping of 215 pixels for training and without any overlapping window for validation. After this process, the data sets are split into 35,584 train images and 2,224 validation images, respectively.

2) DeepGlobe: The DeepGlobe data set [52] includes 6,226 images with the spatial resolution of 1024×1024. The GSD is 0.5m/pixel. Same as the Spacenet, the images are randomly separated into train and validation sets with the percentage of 8:2. We create the crops of size 512×512, yielding 44,838 and 4,976 images for training and validation.

3) Evaluation Metrics: In our experiments, we use both connectivity-based and pixel-based metrics for evaluation.

Conectivity-Based Metric: TLTS [53] statics the same length of the shortest path between two points randomly selected in the ground-truth and estimated road networks. It records percentages of correct, too long, too short, and infeasible paths. In experiments, we report the results with the relative length difference within 5%. Too long and too short indicate the missing links and hallucinated connections, respectively. Average Path Length Similarity (APLS) [51] is defined as the average relative length difference. The difference of shortest paths are compared between corresponding points sampled from the ground truth and predicted road network graphs. The definition of APLS is as follow:

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

where N is the total number of paths. $L(a,b)$ represents the length of path(a,b) sampled from the ground truth and $L(a',b')$ is from the predicted road network graphs where (a',b') denotes the location in predicted graph closest to the ground truth node (a,b).

Pixel-Based Metric: We use Completeness (COM), Correctness (COR), and Quality (Q) which are widely used in road extraction, where the definition has been relaxed within a distance of 5 pixels [54]. The road intersection over union (Road IoU) is also used to evaluate the performance with a constant width. Specifically, the metrics are defined as follows:

$$COM = \frac{TP}{TP + FN}, \tag{9}$$

$$COR = \frac{TP}{TP + FP}, \tag{10}$$

$$Q = \frac{2 \times COM \times COR}{COM + COR}, \tag{11}$$

B. Experiments Setting

Dataset Preprocessing: Firstly, the road line strings are extracted from the binary masks and smoothed with Ramer-Douglas-Peucker (RDP) algorithm. We can obtain both the road locations and orientations from the line strings. Then, the correct vertical coordinates (with row anchors) or the horizontal coordinates (with column anchors) can be obtained. On the SpaceNet data set, similar to [20], we create the centerline of the road from the line strings and use distance transform with a Gaussian kernel along the centerline. The binary masks are created with a threshold of 0.76. On the DeepGlobe data set, we use its own binary masks. The N_d, N_s, λ, and β is empirically set to 9, 4, 4, and 4.

Training Details: We randomly crop the images into 256 × 256 for training. Data augmentation methods including random

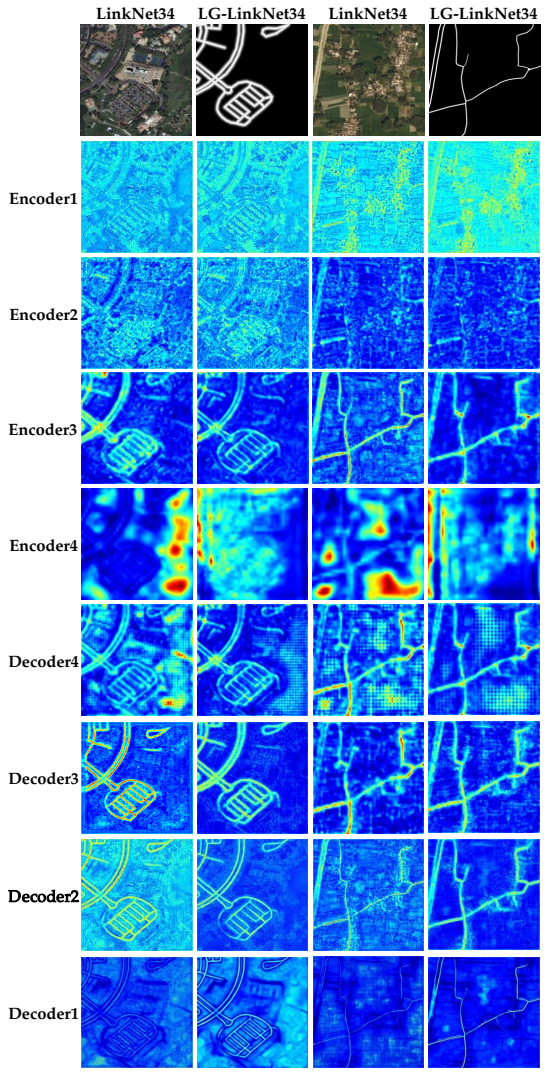


Fig. 6. Visualization of the feature maps on LinkNet34 and LG-LinkNet34.

horizontal flip, mirroring and 90/180/270 degree rotation are utilized. Our models are trained with the batch size of 32 in total 120 epochs and optimized by stochastic gradient descent (SGD). The momentum is 0.9 and weight decay is set to 0.0005. The initial learning rate is 1×10^{-2} dropped with 10 at epochs 60, 90, 110. We perform simple post-process to remove small segments, fill small holes, and use RDP to smooth the final graph. All these settings keep the same with the multi-branch [20] for fair comparison.

C. Ablation study

We conduct extensive ablation experiments on the validation set of SpaceNet and DeepGlobe data sets to analysis the affect of different settings for LGNet.

Effect of the road location prediction We first investigate the performance of our proposed road location prediction (RLP). The experiments are conducted on LinkNet34 (with ResNet34 as the backbone), and we evaluate the Road IoU and APLS metrics on the SpaceNet and DeepGlobe validation set. As shown in Table I, the RLP can improve the Road IoU from 62.94 % to 63.02% for SpaceNet. For DeepGlobe, it also has a 0.25% increase. By incorporating the road location

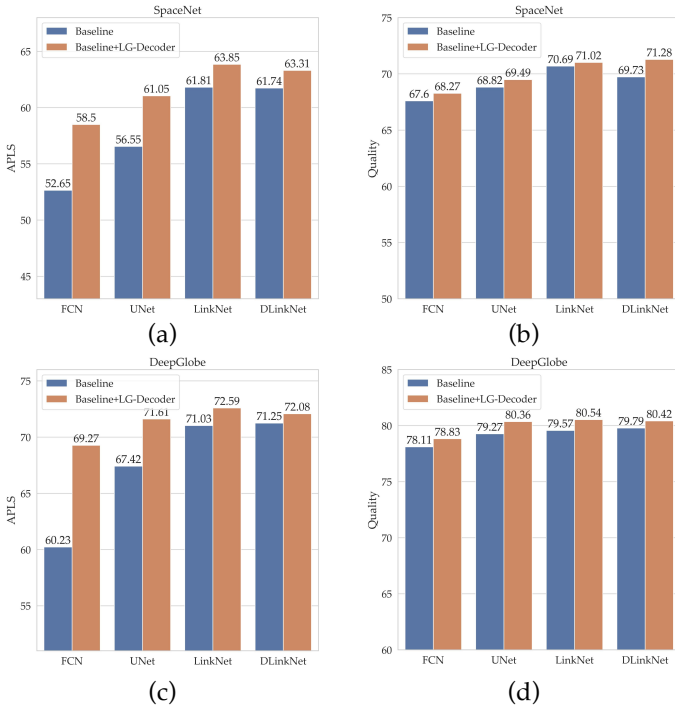


Fig. 7. Ablation results of different encoder-decoder architectures on SpaceNet and DeepGlobe data set. (a) and (b) are on SpaceNet. (c) and (d) are on DeepGlobe.

TABLE I
COMPARISON OF JOINT LEARNING MODULES WITH ROAD LOCATION PREDICTION (RLP) AND LOCATION-GUIDED DECODER (LG-DECODER) INCLUDING LCG AND LCA FOR ROAD SEGMENTATION.

RLP	LCA	LCG	Spacenet		DeepGlobe	
			Road IoU	APLS	Road IoU	APLS
×	×	×	62.94	61.77	67.99	71.03
✓	×	×	63.02	62.08	68.24	72.04
✓	×	✓	63.12	63.02	68.12	72.22
✓	✓	×	63.03	60.01	68.08	71.41
✓	✓	✓	63.15	63.85	68.29	72.69

prediction, the results show that APLS is improved for both datasets by 0.31% and 1.01%, respectively. The results show that RLP improves the feature generalization to refine road segmentation.

Effect of the location-guided decoder The key of our method is the effective fusion of road location information into the segmentation branch, resulting in improved road connectivity. Table I shows that LinkNet with location context guidance(LCG) outperforms the baseline on both the SpaceNet data set (62.08% vs 63.12%) and the DeepGlobe data set (72.04% vs 72.22%) in APLS. Using the location context aggregation (LCA) module independently resulted in a significant decrease in connectivity, but when used in combination with LCG, connectivity was further improved. This indicates that the LCG module is a more effective way to leveraging the road location information compared to simply adding it to the segmentation network. By adding the location-guided decoder (LG-Decoder) which uses both LCA and LCG, the APLS improved significantly by 1.77% for the SpaceNet data set and Road IoU also improved by 0.13%. The combination

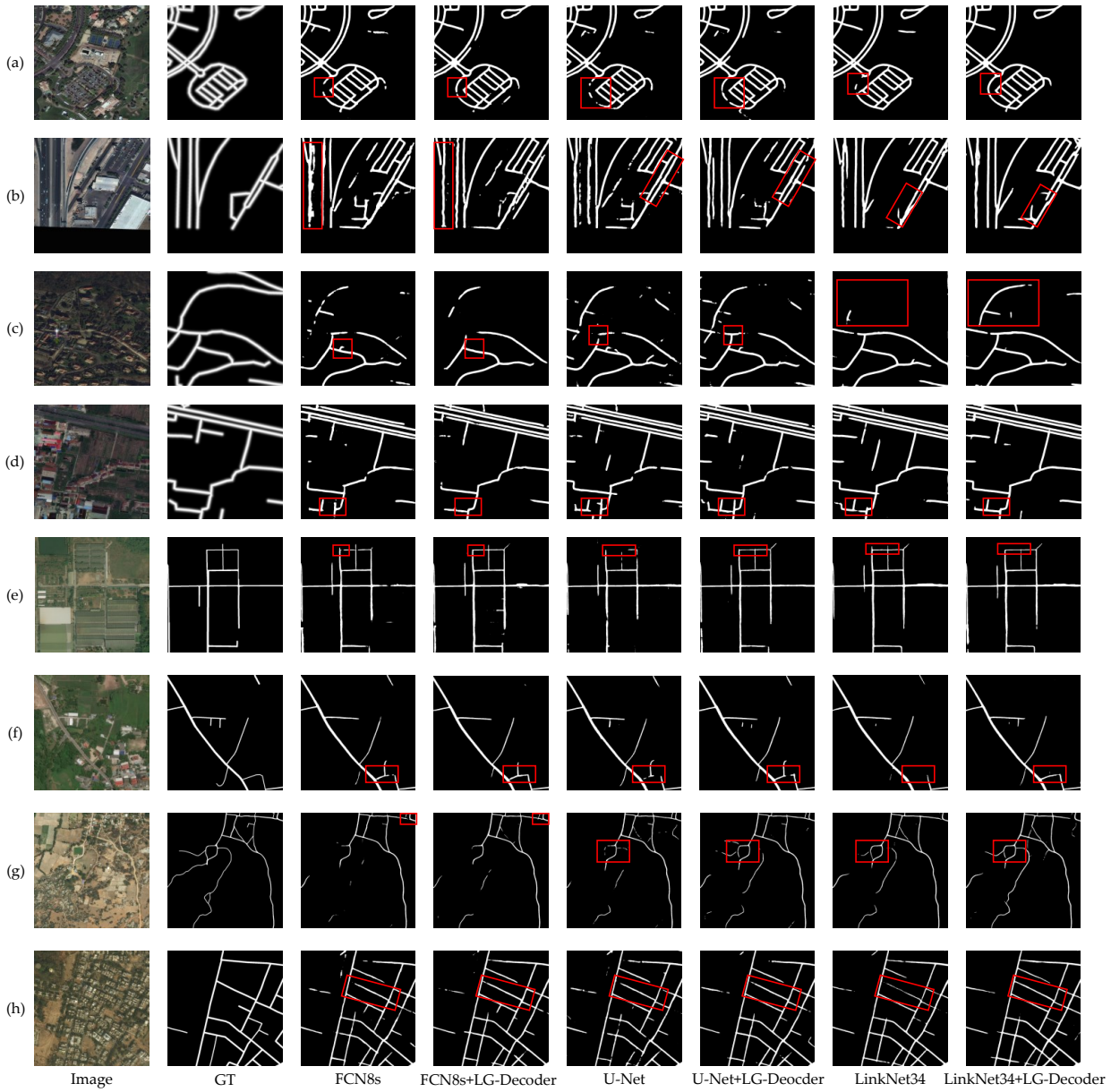


Fig. 8. Visulization of the segmentation results with or without LG-Decoder (ablation study). (a)–(d) are selected from the SpaceNet data set. (e)–(h) are selected from the DeepGlobe data set.

of the LG-Decoder on the DeepGlobe data set also shows a 0.65% improvement in the APLS measure. These results consistently demonstrate the effectiveness of LGNet.

To better understand the LGNet, we visualize the feature maps of different layers of the baseline LinkNet34 and our LG-LinkNet34. As shown in Fig. 6, we can see that the low-level features include more discriminable road edge features. Our method focuses more on the road foreground than the baseline in the high-level feature map. It is evident that LG-LinkNet34 can significantly enhance the road connectivity information and ease the background noise, thus undoubtedly strengthening the road connectivity of the model.

Effect of various encoder-decoder architecture To further verify the effectiveness of our method, we apply the location-guided module to other architectures, which include FCN, UNet, LinkNet, and D-LinkNet. Fig. 7 presents the road segmentation and connectivity accuracy on SpaceNet and DeepGLobe. The results identify the generality of our method. With the LG-Decoder, four networks (FCN, U-Net, LinkNet34, and D-LinkNet34) outperform the accuracy of its corresponding baseline. Especially for road connectivity in DeepGlobe datasets, FCN has a significant improvement with a 9.03% increase. UNet, LinkNet, and D-LinkNet have improved by 4.19%, 1.56%, and 0.83%, respectively.

Fig. 8 shows the quantitative results on the FCN, UNet, and LinkNet combined with the LG-Decoder. As we can see from the visualization results, the LG-Decoder effectively promotes road connectivity. In complex road scenes, LGNet can obtain smoother results that effectively improve the performance of road segmentation. The above results further verify the generalization performance of our method, which can be applied in different encoder-decoder based road segmentation

TABLE II
RESULTS OF THE APLS AND ROAD IOU WITH DIFFERENT GRIDS.

datasets	metrics	16×16	16×32	16×64	32×16	32×32	32×64	64×16	64×32	64×64
SpaceNet [51]	APLS	61.82	63.92	63.85	62.18	**64.05**	63.09	62.27	63.03	63.29
	Road IoU	63.07	63.13	63.15	**63.20**	63.10	62.98	63.11	63.04	62.96
DeepGlobe [52]	APLS	71.26	72.33	**72.69**	71.71	72.17	71.66	72.72	71.26	71.51
	Road IoU	68.16	68.02	**68.29**	67.70	68.10	67.93	68.21	67.97	68.02

TABLE III
THE RESULTS OF COMPARATIVE EXPERIMENTS ON SPACENET DATA SET FOR ROAD EXTRACTION.

Model	Connectivity-based Metric					Pixel-based Metric			
	TLTS				APLS	CCQ			Road IoU
	correct	too long	too short	infeasible↓		COR	COM	Q	
FCN [24]	57.34	4.87	3.96	33.83	52.65	81.09	80.25	67.60	60.71
UNet [25]	60.02	5.17	4.02	30.80	56.55	79.88	83.26	68.82	61.68
LinkNet34 [26]	65.99	4.45	4.51	25.04	61.77	82.83	82.84	70.69	62.94
D-LinkNet34 [12]	64.77	5.08	4.48	25.67	61.74	81.47	82.86	69.73	62.44
DeepRoadMapper [55]	51.77	5.48	3.60	39.15	46.01	76.00	79.81	63.74	55.56
RoadCNN [56]	60.16	4.56	3.90	31.38	57.36	82.51	80.91	69.07	60.83
ResUNet [27]	42.28	5.94	3.38	48.40	37.15	73.04	79.38	61.39	56.79
Diresnet [21]	56.55	5.03	5.27	33.15	47.17	66.01	64.1	48.19	45.64
CoANet [57]	66.23	4.28	4.45	25.04	62.73	82.24	83.36	70.64	61.30
GCBNet [58]	62.73	4.40	4.19	28.68	62.04	83.97	82.36	71.18	61.40
Multi-Branch [20]	66.76	4.59	4.55	24.10	63.46	**84.34**	82.94	**71.87**	**63.35**
LG-DLinkNet34(ours)	67.34	4.52	5.06	23.08	63.31	83.39	83.07	71.28	63.24
LG-LinkNet34(ours)	**68.12**	4.54	4.93	**22.41**	**63.85**	82.73	**83.38**	71.02	63.15

networks with an improvement.

Effect of the different grid size In the road location prediction task, there are two parameters, $\hat{H}$ and $\hat{W}$, which represent the spatial size of the location feature map. In Table. II, we show how this parameters impact the APLS and Road IoU on SpaceNet and DeepGlobe datasets. We find that the Road IoU metric is insensitive to the parameters, and its fluctuation range does not exceed 0.6%. The APLS has the best result in the grid size of 32×32 for SpaceNet, but the result on DeepGlobe is not. Thus, for the trade-off, we selected the grid size of 16×64.

D. Comparisons with state-of-the-arts

1) SpaceNet Data Set: In this section, we conduct experiments on the SpaceNet data set. The compared methods include the baseline FCN [24], the U-Net [25], and the LinkNet34 [26]. We also compare with the sate-of-the-art road segmentation methods include D-LinkNet [12], DeepRoadMapper (DRM) [55], RoadCNN [56], CoANet [57], GCBNet [58], Multi-branch [20], and DiresNet [21]. The last two networks are the multi-task architecture, which is used to compare our proposed auxiliary task to verify the advantages. Table III presents the comparative quantitative results measured with pixel-based metric and connectivity-based metric. The best performance is denoted in bold, and the second-best is marked with underlines in the table. As shown in Table III, our method outperforms baseline in both pixel-based metrics and connectivity-based metrics. Compared with the two multi-task networks, our method LG-DLinkNet34 and LG-LinkNet34 are essentially the best or second-best results on the connectivity-based metric. LG-LinkNet34 achieves the state-of-the-art performance of 63.85% in the APLS metric. Multi-branch achieves the second best performance with stacked hourglass network [59] through the multi-task fusion strategies of simple feature addition. In contrast, our LG-LinkNet34 adopts a depth-wise convolution to fuse the multi-task features more effectively and efficiently and achieve the best results.

Fig. 9 shows the visualization results of road segmentation results. The top four rows show the segmentation results on the SpaceNet Datasets. And the bottom four rows are on the DeepGlobe. Compared with the RoadCNN, DRM, and ResUNet, we can see that our method has less noise. For the first row of Fig. 9, there is an irregularly curved road with occlusion in the top right of the image. Our method effectively predicts the continuous road, but other methods do not. As shown in the Fig. 9 (c) and (d), the prediction of CoANet has more wrong road segments than our method. In the last three rows, there is a situation where the road has shadows and is mostly occluded with trees. As is evident, our method shows the best road connectivity results against the shadow and occlusion situation, connecting the separated road segment efficiently.

2) DeepGlobe Data Set: In this section, we report the results of the road extraction on DeepGlobe data set. The DeepGlobe data set has more rural scene which includes some narrower

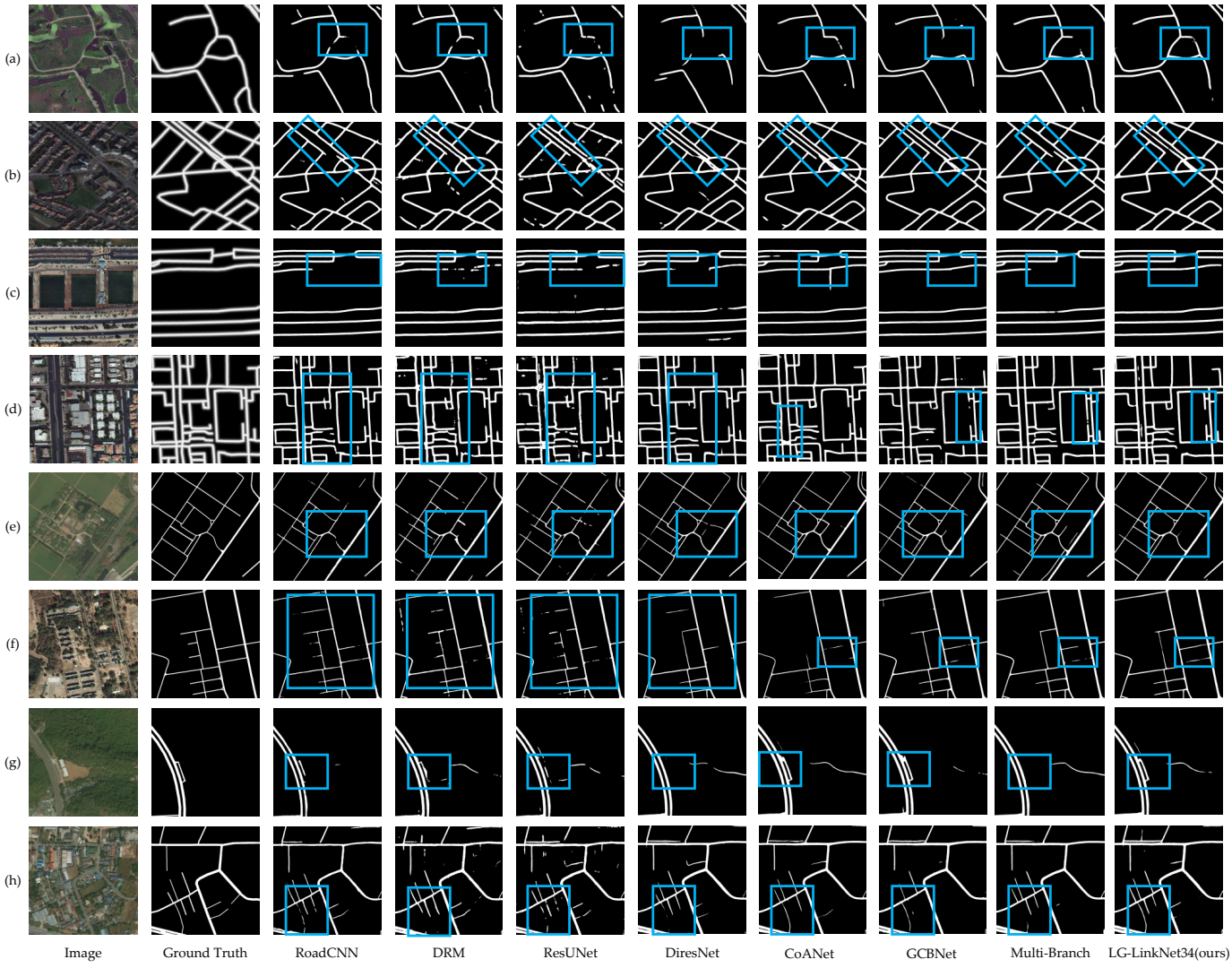
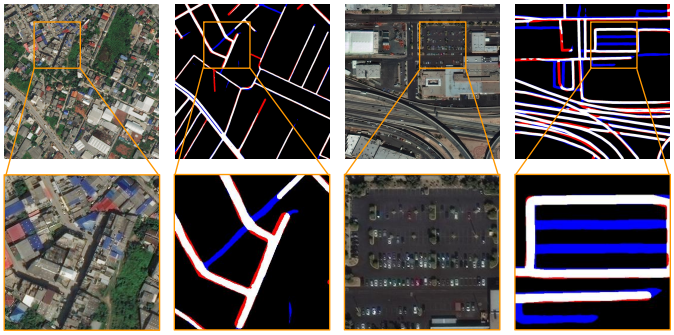


Fig. 9. Visualization of the segmentation results output by our proposed LGNet and other baseline methods for the comparison. (a)–(d) are selected from the SpaceNet data set. (e)–(h) are from the DeepGlobe data set.

TABLE IV
THE RESULTS OF COMPARATIVE EXPERIMENTS ON DEEPGLOBE DATA SET FOR ROAD EXTRACTION.

Model	Connectivity-based Metric					Pixel-based Metric				FLOPs
	TLTS				APLS	CCQ			Road IoU	
	correct	too long	too short	infeasible↓		COR	COM	Q		
FCN8s [24]	59.83	4.04	3.59	32.53	60.23	89.12	86.35	78.11	66.72	25.54G
UNet [25]	59.61	4.01	3.27	33.12	67.42	90.53	86.44	79.27	67.71	40.13G
LinkNet34 [26]	67.37	3.67	3.95	25.00	71.03	90.18	87.12	79.57	67.99	6.85G
D-LinkNet34 [12]	66.63	3.81	3.71	25.83	71.25	90.88	86.73	79.79	68.11	7.44G
RoadDeepMapper [55]	50.64	3.71	3.28	42.37	60.23	89.22	83.04	75.47	61.97	-
RoadCNN [56]	49.55	3.38	2.99	44.08	58.86	**93.17**	77.83	73.62	60.76	-
ResUNet [27]	42.07	3.28	3.09	51.48	54.63	88.90	82.32	74.65	63.78	80.98G
CoANet [57]	65.69	2.74	2.96	28.60	68.93	88.55	86.45	77.76	65.78	69.31G
GCBNet [58]	66.87	2.93	2.56	27.57	71.38	90.61	87.70	80.40	68.07	8.43G
Diresnet [21]	65.46	3.58	3.55	27.46	71.55	91.99	87.28	**81.11**	**69.09**	19.00G
Multi-Branch [20]	**72.58**	3.66	4.01	**19.62**	**74.46**	89.52	**88.26**	80.01	68.19	41.55G
LG-DLinkNet34(ours)	68.18	3.53	3.86	24.44	72.08	90.72	87.63	80.42	68.19	7.69G
LG-LinkNet34(ours)	67.99	3.58	3.81	24.62	72.69	91.17	87.35	80.54	68.29	**7.09G**



(a) (b)

Fig. 10. Visualization of failure examples of our proposed LGNet. White represents True Positive, Blue represents False Positive, and Red represents False Negative.

TABLE V
COMPARISON OF COMPUTATION COST AND MEMORY COST IN TERMS OF PARAM. AND FLOPS, RESPECTIVELY.

method	#param.	FLOPs	APLS	
			SpaceNet	DeepGlobe
FCN8s [24]	18.64M	25.54G	52.65	60.23
UNet [25]	17.27M	40.13G	56.55	67.42
LinkNet [26]	21.64M	6.85G	61.81	71.03
D-LinkNet [12]	31.08M	7.44G	61.74	71.25
ResUNet [27]	13.04M	80.98G	54.63	37.15
DiresNet [21]	21.56M	19.00G	47.17	71.55
CoANet [57]	59.15M	69.31G	62.73	68.93
GCBNet [58]	31.23M	8.43G	62.04	71.38
Multi-branch [20]	29.00M	41.55G	63.46	74.46
LG-LinkNet34(ours)	**27.66M**	**7.09G**	63.85	72.69

curved roads than the SpaceNet data set that do have not enough road context information. The insufficient road context reduces the ability of the model to handle with shadows and occlusions. Our method models the road as multiple line segments and ignores some short-length road segments due to the fixed grid size in the location prediction task resulting in the reduction of performance in road connectivity.

As shown in Fig. 9, the bottom four rows visualize the results of the compared methods. The DRM is the worst that misses some road segments and brings additional background noise. Although our method also gets some wrong road segments, it retains better connectivity. For example, in some scenes where the road is occluded by the trees (Fig. 9 (e)), the road extractions by other methods may be disconnected or wrong-connected, while our LGNet maintains the better road connectivity. As shown in Table IV, we can see LG-LinkNet34 performs the second-best connectivity, which is similar to the visualization results in the qualitative evaluation. The DiresNet achieves the best performance in the pixel-based metric, while our method is also the second-best. In the connectivity-based metric, the infeasible path of the proposed LG-LinkNet34 is much smaller than the Diresnet. It can show that our methods predict more correct connected roads which is more effective in reality application.

E. Analysis of Road Extraction Failure

As mentioned in the above experiments, our proposed LGNet achieves the best connectivity performance on SpaceNet and the second-best on DeepGlobe. As shown in Fig. 10, our method still has road extraction errors with hallucinated and missing connections. In Fig. 10 (a), the incorrect connection of road segments is caused by occlusion and shadows created by tall buildings. In these areas, the roads may not be visible in satellite images, leading to inaccurate mapping. Fig. 10 (b) illustrates this issue, which is also compounded by incomplete labeling of parking areas. Fig. 10 (b) shows a stacked road scene, where our LGNet model predicts most of the roads accurately. However, some roads are still incorrectly connected due to the complex nature of the scene. By comparing the shortest path lengths of ground-truth and predicted road graphs, we can quantify the accuracy of a predicted road network. Tables III and IV demonstrate that incorporating LG-Decoder into LinkNet and D-LinkNet results in a slight increase in the "too short" metric on the SpaceNet data set, and a slight decrease on the DeepGlobe data set. Meanwhile, the "correct" metric increased for both datasets, indicating that our method can effectively improve road connectivity without introducing excessive connectivity errors. In future work, we will introduce multi-modal information including depth and trajectories of buses which can get the local spatial relationship and information compensation to predict the continuous road more effectively.

F. Complexity of LGNet

For fairly evolution, the memory and computation cost are compared in Table V. The FLOPs are calculated based on the spatial size of 256×256. Our LG-LinkNet34 has a 6.02M increase in the total parameters compared with the baseline LinkNet34. It also has 0.24 GFLOPs[1] improvement in the computation cost. Despite the baseline models, LG-LinkNet34 has fewer parameters and the least FLOPs which indicate that it is a lightweight and effective model for road extraction. LG-LinkNet34 is almost $6\times$ fewer than multi-branch in terms of FLOPs while it has competitive performance in road extraction on both SpaceNet and DeepGlobe data sets. Our method achieves the best trade-off between performance and complexity.

V. CONCLUSION

In this paper, we have presented LGNet combining the auxiliary road location prediction for the road extraction task, which captures the road connectivity through the continuous coordination prediction. To feed the connectivity information to the segmentation decoder, we introduce the location-guided decoder, which aggregates the global location context and filters the road features by the road location information. The experiments demonstrate that LGNet promotes the road connectivity of the network at less computation cost. In the future, we will explore semi-supervised road extraction to further reduce the training computation cost and improve the efficiency of learning road connectivity.

[1]We use the flops-counter to calculate GFLOPs https://github.com/sovrasov/flops-counter.pytorch.

REFERENCES

[1] C. Boucher and J.-C. Noyer, "Automatic detection of topological changes for digital road map updating," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 11, pp. 3094–3102, 2012.

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

[3] Y. Yuan, C. Wang, and Z. Jiang, "Proxy-based deep learning framework for spectral–spatial hyperspectral image classification: Efficient and robust," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[4] A. G.-O. Yeh, T. Zhong, and Y. Yue, "Angle difference method for vehicle navigation in multilevel road networks with a three-dimensional transport gis database," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 140–152, 2017.

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

[6] 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, 2017.

[7] J. Yuan, D. Wang, B. Wu, L. Yan, and R. Li, "Legion-based automatic road extraction from satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4528–4538, 2011.

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

[9] Z. Miao, W. Shi, H. Zhang, and X. Wang, "Road centerline extraction from high-resolution imagery based on shape features and multivariate adaptive regression splines," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 583–587, 2013.

[10] X. Lu, Y. Zhong, Z. Zheng, and L. Zhang, "Gamsnet: Globally aware road detection network with multi-scale residual learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 340–352, 2021.

[11] Z. Xiong, Y. Yuan, N. Guo, and Q. Wang, "Variational context-deformable convnets for indoor scene parsing," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2020, pp. 3991–4001.

[12] 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 Comput. Soc. Conf. Comput. Vis. Pattern Recogn. Workshops(CVPRW)*, 2018, pp. 192–1924.

[13] Y. Deng, J. Yang, C. Liang, and Y. Jing, "Spd-linknet: Upgraded d-linknet with strip pooling for road extraction," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.(IGARSS)*, 2021.

[14] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip Pooling: Rethinking spatial pooling for scene parsing," in *CVPR*, 2020.

[15] 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, 2019.

[16] X. Lu, Y. Zhong, Z. Zheng, Y. Liu, J. Zhao, A. Ma, and J. Yang, "Multi-scale and multi-task deep learning framework for automatic road extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9362–9377, 2019.

[17] Y. Wei, K. Zhang, and S. Ji, "Simultaneous road surface and centerline extraction from large-scale remote sensing images using cnn-based segmentation and tracing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8919–8931, 2020.

[18] 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, 2017.

[19] X. Yang, X. Li, Y. Ye, R. Y. K. Lau, X. Zhang, and X. Huang, "Road detection and centerline extraction via deep recurrent convolutional neural network u-net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7209–7220, 2019.

[20] A. Batra, S. Singh, G. Pang, S. Basu, C. Jawahar, and M. Paluri, "Improved road connectivity by joint learning of orientation and segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2019, pp. 10 377–10 385.

[21] L. Ding and L. Bruzzone, "Diresnet: Direction-aware residual network for road extraction in vhr remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10 243–10 254, 2021.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2017.

[23] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "PSANet: Point-wise spatial attention network for scene parsing," in *ECCV*, 2018.

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

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

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

[27] 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, 2018.

[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2017, pp. 2261–2269.

[29] J. Xin, X. Zhang, Z. Zhang, and W. Fang, "Road extraction of high-resolution remote sensing images derived from denseunet," *Remote Sens.*, vol. 11, no. 21, p. 2499, 2019.

[30] Q. Wu, F. Luo, P. Wu, B. Wang, H. Yang, and Y. Wu, "Automatic road extraction from high-resolution remote sensing images using a method based on densely connected spatial feature-enhanced pyramid," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 3–17, 2021.

[31] Z. Zhang and Y. Wang, "Jointnet: A common neural network for road and building extraction," *Remote Sens.*, vol. 11, no. 6, p. 696, 2019.

[32] Y. Zhang and Q. Yang, "An overview of multi-task learning," *National Science Review*, vol. 5, no. 1, pp. 30–43, 2018.

[33] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv preprint arXiv:2009.09796*, 2020.

[34] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *International Conference on Learning Representations*, 2019.

[35] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 160–167.

[36] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," *arXiv preprint arXiv:1901.11504*, 2019.

[37] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "The natural language decathlon: Multitask learning as question answering," *arXiv preprint arXiv:1806.08730*, 2018.

[38] Y. Gao, J. Ma, M. Zhao, W. Liu, and A. L. Yuille, "Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2019, pp. 3200–3209.

[39] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2018, pp. 7482–7491.

[40] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2018, pp. 675–684.

[41] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris, "Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2017, pp. 1131–1140.

[42] A. Mosinska, M. Koziński, and P. Fua, "Joint segmentation and path classification of curvilinear structures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1515–1521, 2020.

[43] 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, 2020.

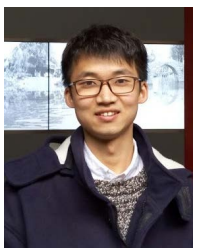
[44] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2018, pp. 7794–7803.

[45] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," *arXiv preprint arXiv:1904.11492*, 2019.

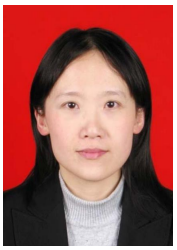
[46] L. Wang, Y. Wang, X. Dong, Q. Xu, J. Yang, W. An, and Y. Guo, "Unsupervised degradation representation learning for blind super-resolution," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2021, pp. 10 576–10 585.

[47] Z. Qin, H. Wang, and X. Li, "Ultra fast structure-aware deep lane detection," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*. Springer, 2020, pp. 276–291.
[48] J. Hu, Q. Wang, and X. Li, "Road extraction from satellite image via auxiliary road location prediction," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.(IGARSS)*, 2021, pp. 2182–2185.
[49] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020.
[50] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. - Int. Conf. 3D Vis.(3DV)*, 2016, pp. 565–571.
[51] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "Spacenet: A remote sensing dataset and challenge series," *arXiv preprint arXiv:1807.01232*, 2018.
[52] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. Workshops(CVPRW)*, 2018, pp. 172–17 209.
[53] J. D. Wegner, J. A. Montoya-Zegarra, and K. Schindler, "A higher-order crf model for road network extraction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2013, pp. 1698–1705.
[54] C. Wiedemann, C. Heipke, H. Mayer, and O. Jamet, "Empirical evaluation of automatically extracted road axes," *Empirical evaluation techniques in computer vision*, vol. 12, pp. 172–187, 1998.
[55] G. Máttyus, W. Luo, and R. Urtasun, "Deeproadmapper: Extracting road topology from aerial images," in *Proc. IEEE Int. Conf. Comput. Vision. (ICCV)*, 2017, pp. 3458–3466.
[56] F. Bastani, S. He, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, S. Madden, and D. DeWitt, "Roadtracer: Automatic extraction of road networks from aerial images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2018, pp. 4720–4728.
[57] J. Mei, R.-J. Li, W. Gao, and M.-M. Cheng, "Coanet: Connectivity attention network for road extraction from satellite imagery," *IEEE Trans. Image Process.*, vol. 30, pp. 8540–8552, 2021.
[58] "A global context-aware and batch-independent network for road extraction from vhr satellite imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 353–365, 2021.
[59] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*. Springer, 2016, pp. 483–499.

Jingtao Hu is currently pursuing the Ph.D. degree with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing, computer vison and machine learning.



Junyu Gao received the B.E. degree and the Ph.D. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China, in 2015 and 2021 respectively. He is currently a researcher with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.



Yuan Yuan (M'05-SM'09) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.

Jocelyn Chanussot (M'04-SM'04-F'12) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Universit de Savoie, Annecy, France, in 1998. Since 1999, he has been with Grenoble INP, where he is a Professor of signal and image processing. He has been a Visiting Scholar at Stanford University, Stanford, CA, USA, KTH Royal Institute of Technology, Stockholm, Sweden, and National University of Singapore, Singapore. Since 2013, he is an Adjunct Professor of the University of Iceland, Reykjavk, Iceland. In 2015-2017, he was a Visiting Professor at the University of California, Los Angeles (UCLA), Los Angeles, CA, USA. He holds the AXA Chair in remote sensing and is an Adjunct Professor at the Chinese Academy of Sciences, Aerospace Information Research Institute, Beijing, China. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence.

Dr. Chanussot is the founding President of IEEE Geoscience and Remote Sensing French chapter (2007-2010) which received the 2010 IEEE GRSS Chapter Excellence Award. He has received multiple outstanding paper awards. He was the Vice-President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia (2017-2019). He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing (WHISPERS). He was the Chair (2009-2011) and Cochair of the GRS Data Fusion Technical Committee (2005-2008). He was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society (2006-2008) and the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing (2009). He is an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the PROCEEDINGS OF THE IEEE. He was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (2011-2015). In 2014 he served as a Guest Editor for the IEEE Signal Processing Magazine. He is a member of the Institut Universitaire de France (2012-2017) and a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters, 2018-2019).

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, machine learning, pattern recognition and remote sensing.