# Toward Accurate and Efficient Road Extraction by Leveraging the Characteristics of Road Shapes

Changwei Wang, Rongtao Xu, Shibiao Xu, *Member, IEEE*, Weiliang Meng, *Member, IEEE*, Ruisheng Wang, *Senior Member, IEEE*, Jiguang Zhang, *Member, IEEE*, and Xiaopeng Zhang, *Member, IEEE*

***Abstract*— Automatically extracting roads from very high-resolution (VHR) remote sensing images is of great importance in a wide range of remote sensing applications. However, complex shapes of roads (i.e., long, geometrically deformed, and thin) always affected the extraction accuracy, which is one of the challenges of road extraction. Based on the insight into road shape characteristics, we propose a novel road shape-aware network (RSANet) to achieve efficient and accurate road extraction. First, we introduce the efficient strip transformer module (ESTM) to efficiently capture the global context to model the long-distance dependence required by long roads. Second, we design a geometric deformation estimation module (GDEM) to adaptively extract the context from the shape deformation caused by shooting roads from different perspectives. Third, we provide a simple but effective road edge focal loss (REF loss) to make the network focus on optimizing the pixels around the road to alleviate the unbalanced distribution of foreground and background pixels caused by the roads being too thin. Finally, we conduct extensive evaluations on public datasets to verify the effectiveness of RSANet and each of the proposed components. Experiments validate that our RSANet outperforms state-of-the-art methods for road extraction in remote sensing images.**

***Index Terms*— Efficient and accurate road extraction, efficient strip transformer module (ESTM), geometric deformation estimation module (GDEM), road edge focal loss (REF loss), road shape-aware network (RSANet).**

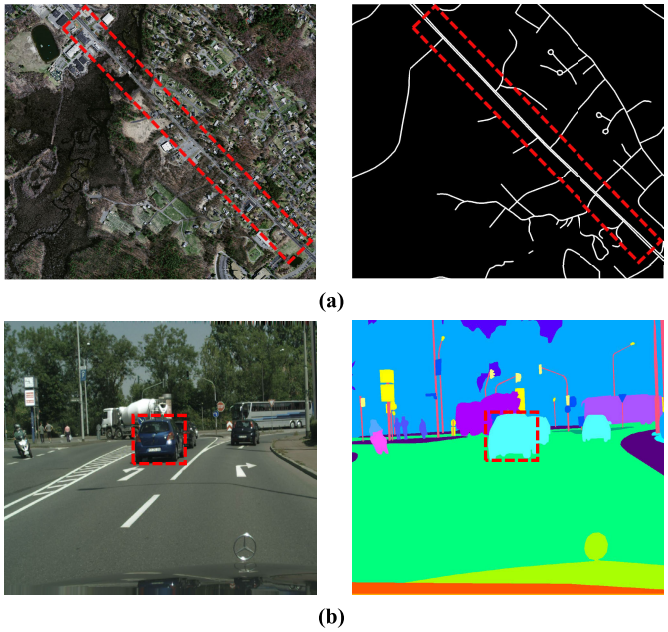



Fig. 1. Examples of (a) road extraction task versus (b) common natural object segmentation task. Obviously, there are significant differences (i.e., long, geometrically deformed, and thin) in shape between roads and common natural segmented objects.

## I. Introduction

ACCURATE road extraction from remote sensing images as an open research topic has been widely applied in urban planning [1], vehicle navigation [2], geographic information system update [3], autonomous driving [4], and other fields. Unfortunately, both the accuracy and efficiency performance of traditional road extraction have been difficult to catch up with the rapid development of optical remote sensing and automatic driving technology.

Since the flourishing and excellent performance of convolutional neural networks (CNNs), recent studies primarily resort to CNNs for tackling the above difficulties, which widely adopt the convolutional layers with square kernels as the core building blocks of CNNs [5], [6], [7], [8], [9], [10], [11], [12]. Although classic CNNs have been validated to be suitable for most common objects with bulk shape, they might not be optimal for roads with some special inherent shape properties [13], [14], [15], [16]. For better understanding, Fig. 1 shows the difference in shape between road extraction and common objects segmentation. Different from buildings, trees, vehicles, and other elements in remote sensing images of urban scenes, roads usually follow three typical shapes that bring challenges for existing road extraction methods to accurately capture.

Manuscript received 25 October 2022; revised 20 February 2023; accepted 1 June 2023. Date of publication 9 June 2023; date of current version 21 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant U21A20515, Grant 62271074, Grant U2003109, Grant 62171321, Grant 62071157, Grant 62162044, and Grant 61971418; in part by the Open Research Fund of Key Laboratory of Space Utilization, Chinese Academy of Sciences, under Grant LSU-KFJJ-2021-05; and in part by the Open Projects Program of National Laboratory of Pattern Recognition. *(Changwei Wang and Rongtao Xu contributed equally to this work.) (Corresponding authors: Shibiao Xu; Weiliang Meng.)*

Changwei Wang, Rongtao Xu, Weiliang Meng, Jiguang Zhang, and Xiaopeng Zhang are with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: weiliang.meng@ia.ac.cn).

Shibiao Xu is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: shibiaoxu@bupt.edu.cn).

Ruisheng Wang is with the Department of Geomatics Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada.

Digital Object Identifier 10.1109/TGRS.2023.3284478

First, the shape of the road is very long and distributed in strip, which means that roads often traverse the entire image. Therefore, better road extraction may need to perceive the global context but not just the local information. However, as reported in [17], pure CNN-based architectures are not suitable for capturing long-range dependencies due to inherent locality [18]. Although some CNN-based road extraction methods [8] gain a wider range of contextual information by increasing the receptive field, they still cannot enjoy the full global context. Recent NLLinkNet [9] and HMRT [19] take advantage of the global context by introducing nonlocal block [17] or transformer block [20], while these blocks need to generate enormous affinity matrix to measure the relationships with the complexity of $O(N^2)$, where $N$ is the pixel number of input feature maps. Therefore, it brings a lot of extra computation for the remote sensing images with high resolutions.

Second, remote sensing images are usually taken from different perspectives, making the shapes of road often be geometrically deformed [i.e., rotation and scale transformation in Fig. 1(a)] rather than standard [i.e., horizontal to the image like common semantic segmentation in Fig. 1(b)]. Since the convolutional layers have no rotation invariance [21], the existing CNN-based road extraction methods are difficult to deal with the complex geometric deformation of roads in practical situations. Although some approaches [21], [22], [23], [24], [25] propose instance-level geometric deformation modeling solutions and make progress in image identification, remote sensing object detection, optical character recognition, and other fields, they cannot be directly extended to the pixelwise road extraction task. To solve this task, a potential option for modeling local shape is the deformable convolution (DCN) [26], which has a high degree of freedom (DOF) for shape offset estimation. However, directly applying the DCN cannot fully utilize the prior knowledge of road shape and geometric transformation and will take a risk of overparameterizing the local shape because the road has a clear strip shape and the geometric deformation can be easily modeled using geometric transformations in computer vision.

Third, the shape of the road is thin, which causes the road to occupy only a small proportion of the entire image pixels. On the contrary, a large number of background pixels that have nothing to do with foreground objects (roads) account for a large proportion. This trend of an unbalanced distribution of pixels brings impediments to road extraction learning. Therefore, it is necessary to make the network pay more attention to the learning of pixels around the road.

Inspired by these observations, we propose a novel network structure and loss function from the unique shape characteristics of roads to achieve accurate and efficient road extraction. First, to solve the problem of long road shape in the first challenge, the efficient strip transformer module (ESTM) is introduced to capture global context information and adapt to the strip distribution of roads. Compared with the original visual transformer component [20], the proposed ESTM is more efficient and conforms to the characteristics of the road shape better. Second, we design the geometric deformation estimation module (GDEM) to alleviate the problem of road shape geometric deformation in the second challenge. Our GDEM flexibly captures the context information around the deformed road by modeling complex geometric transformations. Furthermore, our GDEM utilizes the inherent shape prior of the road to estimate the deformation estimation with fewer degrees of freedom to achieve accurate and efficient road-related context extraction. Third, we further provide the road edge focal loss (REF loss) to deal with the problem of the thin road shape mentioned in the third challenge. Specifically, our REF loss can make the network pay more attention to the pixels around the road in the optimization process, thereby alleviating the problem of the unbalanced distribution of foreground and background pixels. In general, the philosophy behind our improvements is to find better context extraction strategies and effective supervision that conform to the unique shape of the roads. For clarity, we summarize the four main technical contributions of our work as follows.

1) An ESTM that can capture the global nonlocal context is introduced, which conforms to the distribution characteristics of roads and has high computational efficiency.
2) A GDEM is designed to express complex deformation such as rotation and scaling in remote sensing images for extracting the context around the road more adaptively and efficiently.
3) A road edge focus loss function is provided for additional supplementary supervision on road around pixels to alleviate the problem of uneven distribution of foreground and background pixels in remote sensing images.
4) Based on the above innovations, a road shape-aware network (RSANet) is proposed to achieve more efficient and accurate road extraction, and thorough experimental results validate the effectiveness of each component for the road extraction task. Our RSANet achieves state-of-the-art results on multiple public road extraction datasets, which fully demonstrates its reliable performance.

## II. Related Works

Most existing approaches treat the road extraction as a pixelwise classification task where each pixel is classified into specific categories and can be summarized as a special binary segmentation task. In general, road extraction methods can be roughly categorized into the conventional and deep learning methods.

### A. Conventional Methods for Road Extraction

In the early days, some semiautomatic methods, such as the snakes model method [27], the dynamic planning method [28], and the template matching method [29], were proposed to extract the road semiautomatically with the help of humans selecting the initial point or direction of the road. Later, the other automatic road extraction methods based on classic image processing techniques were given, such as spectral-based methods [30], level set-based methods [31], Hough transform-based methods [32], mathematical morphology-based methods [33], edge feature extraction-based methods [34], and super-pixel-based methods [35]. In addition,

another methods [36] regard road extraction as a bottom-up pixel clustering task. More recently, many machine learning methods for road extraction have been proposed. Song and Civco [37] suggested a two-step technique based on support vector machine (SVM) [38] and shape index to detect road from remote sensing images, in which the SVM model was applied to roughly classify the pixels into a road group and nonroad group first, and then, the road group was further refined by a segmentation algorithm to produce the road regions. Zhang and Couloigner [39] comprehensively used k-means, fuzzy logic classifier, and shape descriptors of angular texture signature multiple technologies to extract the road area. Yuan et al. [40] developed locally excitatory and globally inhibitory oscillator networks for road detection, and the road extraction is divided into three stages, including the segmentation, the medial axis points selection, and the road grouping. Das et al. [41] used probabilistic SVM and salient features to extract roads from high-resolution multispectral satellite images, while Wegner et al. [42] designed a road extraction method based on a higher order conditional random field (CRF) model. However, all the conventional approaches are usually inefficient and cannot deal well with challenging scenes occluded by trees and buildings.

### *B. Deep Learning Methods for Road Extraction*

In the last decade, semantic segmentation has achieved remarkable progress, driven by the rapid evolution of convolutional networks [43] (CNNs) such as fully convolutional network (FCN) [44], which extended the original CNNs architecture to enable dense pixelwise prediction. Then, UNet [45] used the symmetric encoder–decoder structure and skipped connections to restore the spatial details and location information, which is one of the classic practices of FCN architecture. Most other existing deep learning-based road extraction works are derived from UNet and achieve significantly better performance. Zhang et al. [6] extended UNet by introducing the short-cut connections from ResNet [46] and built a method called ResUNet to extract roads. Cheng et al. [5] proposed a cascaded deep CNNs method based on UNet for road identification, which dealt with the task of road detection first and then used the learned domain knowledge to extract the centerline of the road. On the other hand, Máttyus and Urtasun [47] proposed the matching adversarial network for addressing the image segmentation, which has also been applied to the road extraction task, while Bastani et al. [48] designed an iterative graph construction method with CNNs architecture called roadtracer, which could directly extract roads from remote sensing images. These general CNNs structures have reached competitive results, but the limited receptive field caused by the inherent locality (inductive bias) of CNNs cannot handle the occlusion issue of roads by cars and trees well.

Therefore, some subsequent works alleviate this problem by enhancing the perception of the context around the road. Zhou et al. [7] designed a road extraction method named DLinkNet by introducing multiscale dilated convolution to original LinkNet [49] and won the first place in the CVPR 2018 DeepGlobe competition. Yang et al. [50] proposed RCNN-UNet by introducing recurrent convolutional to make better use of spatial context for road extraction. Tao et al. [8] proposed a spatial information inference network (SIINet) to make full use of road-specific context information by 3-D convolutional recurrent neural network (RNN). Ding and Bruzzone [51] presented a direction-aware residual network (DiResNet) to improve the understanding of the context around the road by introducing road direction supervision. Wang et al. [9] combined LinkNet with nonlocal block [17] in deeper layers to grasp condensed context information. Inspired by road shapes and connections in graph networks, CoANet [52] proposes a connected attention network to jointly learn segmentation and pairwise dependencies.

Different from the above methods, we advocate improving existing CNNs architectures based on the shape characteristics of roads to achieve accurate and efficient road extraction. First, we model long-range context dependencies by proposing an efficient ESTM, which has a larger perceptual range than SIINet [8] and higher operating efficiency than NLLinkNet [9]. Second, we propose GDEM to model the complex deformation of roads, which has more diverse deformation capabilities than the strip convolution module in CoANet [52] to adapt to complex geometric transformations. In addition, we also design the REF loss to alleviate the problem of unbalanced distribution of road and background pixels.

### *C. Nonlocal Context Awareness*

It is introduced into the road extraction task to compensate for the dependence on global context information due to the long-span shape characteristics of roads. SIINet [8] is a spatial information inference structure (SIIS) based on RNN and 3-D convolution to learn global spatial context, which alleviates the occlusion problem in road extraction. GAMSNet [53] captures global information by expanding the receptive field range through multiscale residual learning. More recently, with the rapid development of visual transformer [20], some road extraction methods, such as NLLinkNet [9], GCBNet [54], and HMRT [19], use a self-attentive mechanism to model global information. However, these self-attention-based methods require computational consumption of $O(N^2)$ complexity, which will seriously impair the computational efficiency of high-resolution remote sensing image processing. In contrast, we propose ESTM to achieve efficient global context capture by introducing an axial self-attention mechanism.

### *D. Geometric Deformation Estimation*

The geometric deformation (e.g., rotation and scale) of the objects in the remote sensing images has always attracted the attention of researchers, and a series of object detection methods [22], [23], [24], [25] for rotating objects in remote sensing images is proposed. In these methods, they propose rotatable region proposal boxes by modeling the shape characteristic of the objects and extract features in the specific boxes for object recognition. Experimental results show that this feature extraction method based on object shape is significantly better than extracting features from the ordinary horizontal boxes, which

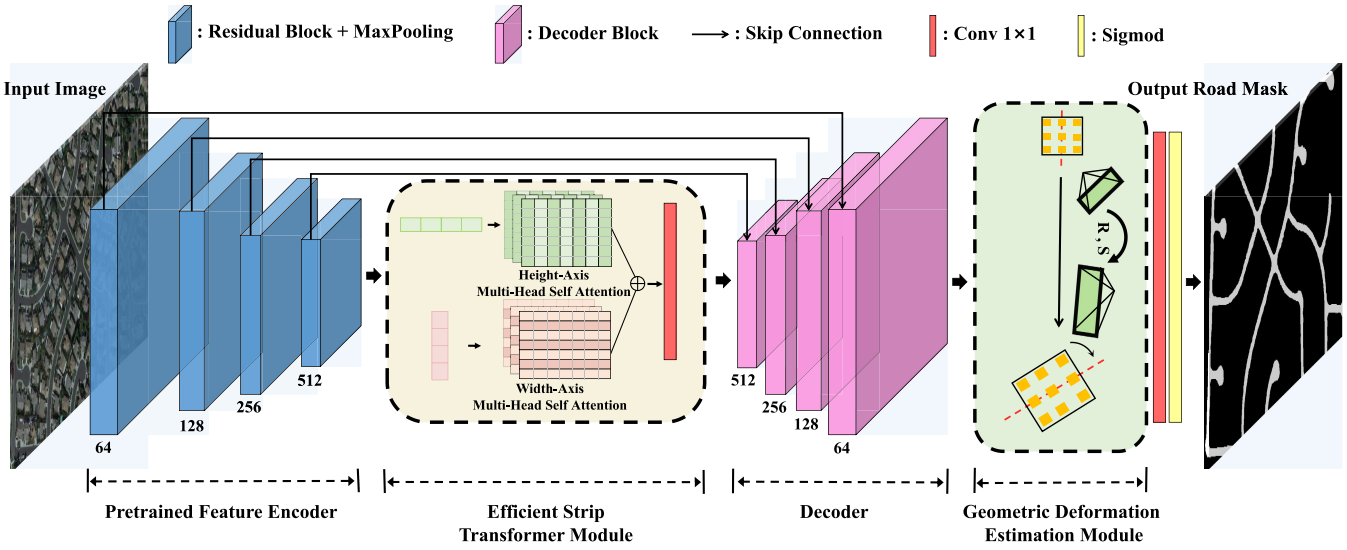



Fig. 2. Overview of our RSANet. The network architecture of our RSANet is composed of pretrained feature encoder, ESTM, decoder, and GDEM. The details of residual block, decoder block, and axial self-attention layer are shown in Fig. 3.

demonstrates that capturing the around context based on the object shape is of great significance to semantic understanding for network learning. Meanwhile, object shape modeling is also crucial to image recognition, and representative works include SAC [55] and STN [21]. These methods are all object- or image-level geometric deformation modeling, whereas our method is a pixel-level dense geometric deformation estimation developed for road extraction. The closest method to our method is DCN [26], which estimates local shape via tunable grid sampling locations. Compared to DCN, our proposed GDEM adopts the road shape and deformation priors and has a lower number of parameters to reduce the risk of overparameterization. In addition, CoANet [52] also proposes a strip convolution module inspired by the geometric deformation characteristics of road shapes. which only considers the context extraction of fixed four directions. In contrast, our proposed GDEM can adaptively extract contextual information in arbitrary directions to cope with complex road geometry deformations.

## III. Proposed Methods

In this article, we design an RSANet to better capture road shape-specific context information, which consists of three components, and each component will be deeply discussed and analyzed later. We first describe the overall network architecture of our RSANet in Section III-A, and then, we introduce ESTM to efficiently obtain global context awareness in Section III-B and design GDEM to model the shape deformation of the road to more accurately extract the surrounding context in Section III-C. Besides, we provide the road edge focus loss in Section III-D to alleviate the unbalanced distribution of road pixels in remote sensing images. Finally, we show how to train our RSANet in Section III-E.

### A. Overall Network Architecture

The overall network of our RSANet is shown in Fig. 2. In the architecture, the encoder utilizes the pretrained ResNet-34 [46], which retains the first four residual blocks to extract feature without the average-pooling layer and the fully connected layers. Pretrained residual blocks add the shortcut mechanism to avoid the gradient vanishing and accelerate the network convergence, as shown in Fig. 3.

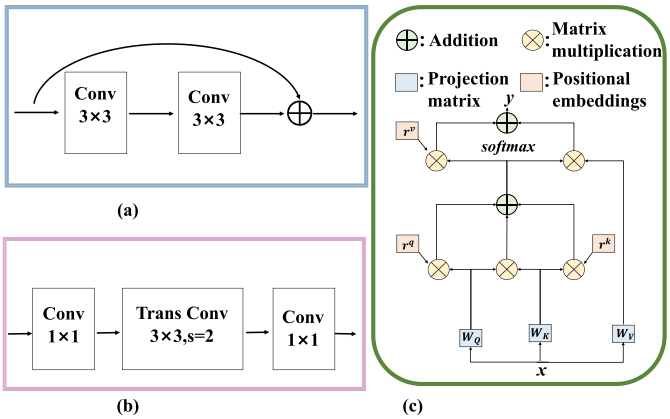



Fig. 3. Details of (a) residual block, (b) decoder block, and (c) axial self-attention layer. Note that a batch normalization layer and a Relu activation function follow each convolution layer (Conv). Each residual block is followed by a MaxPooling for downsampling.

In the encoder of our RSANet, we use the max-pooling layer to implement the downsampling operation to extract high-level semantic information, and each max-pooling layer reduces the resolution of the feature maps by half. To restore the resolution, we use the decoder block shown in Fig. 3(b) to build the decoder pipeline, and each transposed convolutional layer (Trans Conv) doubles the resolution of the feature map. Similar to UNet [45], we use skip connections from encoder to decoder at different feature scales to recover the details and position information of the feature maps. To enable our RSANet efficiently and appropriately enjoy the global context information to cope with the long distance caused by the long road shape, we introduce the ESTM in Section III-B. In addition, we also design the GDEM before the prediction head of our RSANet to model road deformation to better extract the context around the road in Section III-C.

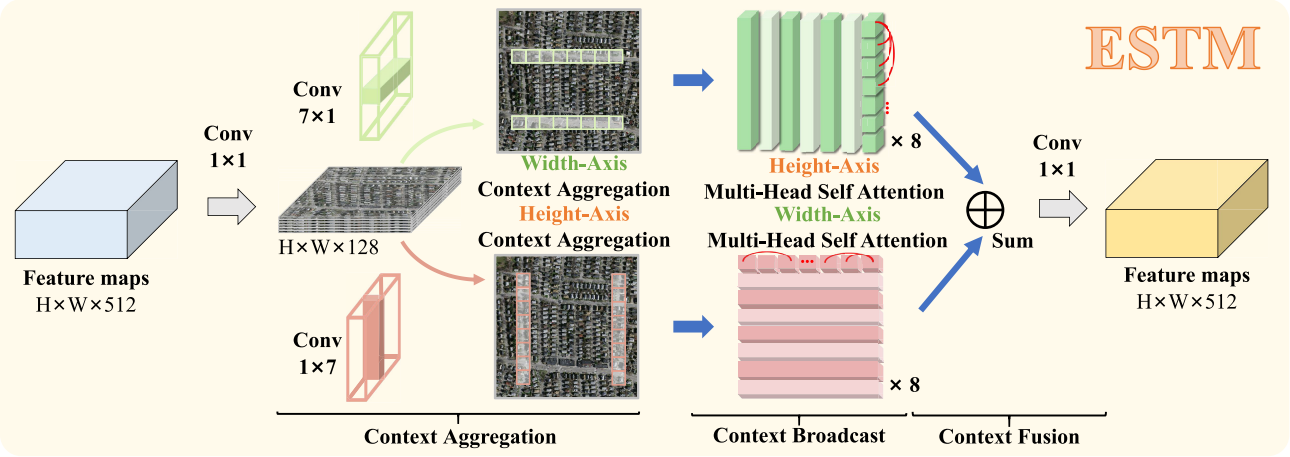



Fig. 4. Our ESTM.

### B. Efficient Strip Transformer Module

One of the typical characteristics of road shapes is long span, continuous distribution, and often span the entire image. In this case, it is important to make full use of the global context information to capture the long-range dependencies for road extraction. However, the conventional square convolution kernel in most CNN networks cannot well capture the strip context of the road and lack the ability to model long distances, while recent nonlocal and transformer architectures will introduce additional computational complexity and time consumption. Motivated by this fact, we introduced an efficient strip conversion module (ESTM) for road extraction to efficiently capture global long-range dependencies, which can mitigate the effects of occlusions due to cars, trees, and buildings through nonlocal context supplementation. As shown in Fig. 4, our ESTM can be divided into three steps.

*1) Context Aggregation:* We first adjust the input feature maps to 128 dimensions through a $1 \times 1$ convolution layer to reduce the amount of calculation, and then, we collect the context information in the horizontal and vertical directions through two strip convolutions (i.e., $7 \times 1$ and $1 \times 7$ Conv), respectively. On the one hand, strip convolution is more suitable for the long shape features of the road to better extract the context than the normal square convolution. On the other hand, strip convolution is used to aggregate contextual information along specific directions (i.e., height and width directions).

*2) Context Broadcast:* We employ the axial self-attention mechanism [56] to broadcast the collected context information in the opposite direction to the context aggregation direction, as shown in Fig. 3(c). Specifically, the axial self-attention mechanism is only performed over the height or width axis of the feature maps; thus, the computational complexity is reduced, which is very important for the processing of remote sensing images that are usually with very high resolutions (VHRs). Besides, we use the learnable position embedding [56] to make up for the lack of position awareness of the transformer. Considering an input feature map $x \in \mathbb{R}^{H\times W\times C_{\text{in}}}$ and an output feature map $y \in \mathbb{R}^{H\times W\times C_{\text{out}}}$, where $H$ and $W$ correspond to height and width of the feature map and $C_{\text{in}}$ and $C_{\text{out}}$ are the number of input and output channels, respectively, $y$ can be derived by the following common self-attention mechanism [57]:

$$y_{i,j} = \sum_{h=1}^{H}\sum_{w=1}^{W} \text{softmax}\left(q_{i,j}^{T} k_{hw}\right) v_{hw} \tag{1}$$

where $q = W_Q x$, $k = W_K x$, and $v = W_v x$ are the query, key, and value corresponding to each input position $(i, j)$ of $x$, respectively. Here, $W_Q$, $W_K$, and $W_V$ are learnable projection matrices [20] used to encode the input $x$. As shown in (1), the self-attention mechanism calculates the affinity between arbitrary positions and therefore has an inductive bias that captures the global context. However, this quadratic complexity modeling mechanism also constrains the computational efficiency of the self-attention mechanism. On the contrary, the axial self-attention mechanism [56] is more effective and the axial self-attention calculated along the height axis can be defined as

$$y_{i,j} = \sum_{h=1}^{H} \text{softmax}\left(q_{i,j}^{T} k_{hi} + q_{i,j}^{T} r_{hi}^{q} + k_{i,j}^{T} r_{hi}^{k}\right)\left(v_{hi} + r_{hi}^{v}\right) \tag{2}$$

where $r_q, r_k, r_v \in \mathbb{R}^{H\times H}$ are the learnable position embedding for the height axis. The complexity is reduced from $O(N^2)$ in (1) to $O(N \cdot H + N \cdot W)$, where $N$ is the pixel number of input feature maps and $H$ and $W$ are the height and width of the feature maps, respectively. There are eight efficient self-attention operations in an axis multihead self-attention block, and their outputs are concatenated together. In addition, we use two branches with opposite axial directions to perform context aggregation and broadcasting in parallel to obtain diverse global context information, as shown in Fig. 4.

*3) Context Fusion:* We merge the extracted contexts of two branches by summation and adjust the feature maps to the original dimensional output by a $1 \times 1$ convolutional layer.

### C. Geometric Deformation Estimation Module

The standard convolutions can capture local context information with a square convolution kernel, which is suitable for most natural objects shot from standard angles. However, the

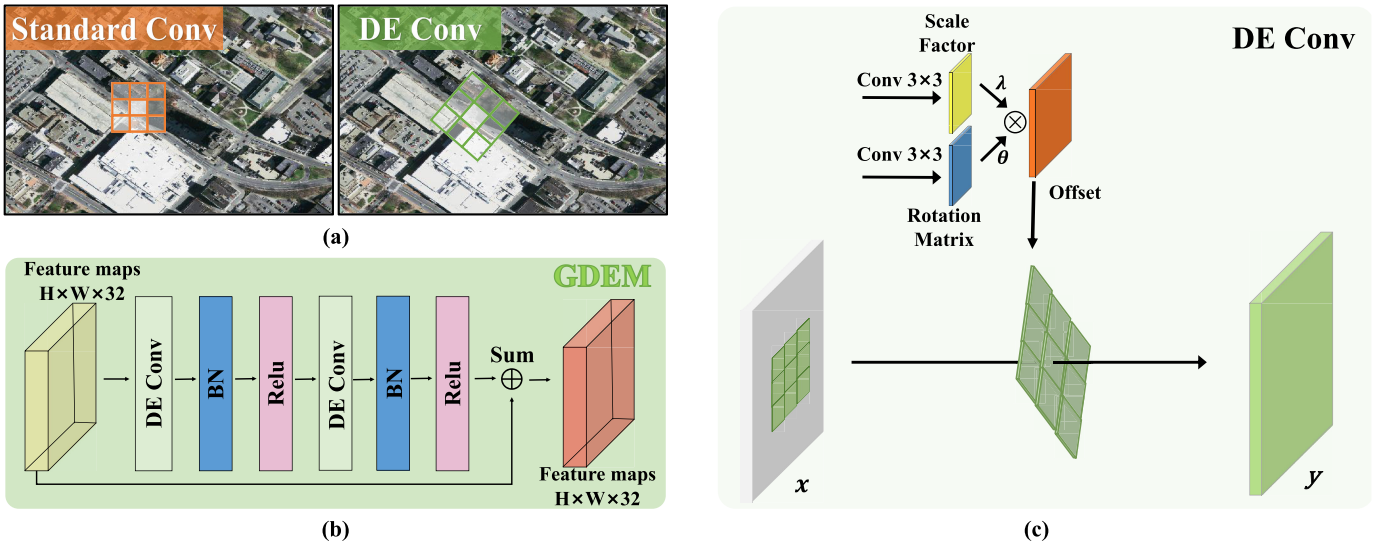



Fig. 5. (a) Illustrations on how standard Conv and our DE Conv work differently for road extraction. (b) Details of GDEM. (c) Details of deformation estimation convolutional layer (DE Conv) in (b).

shooting angle of the remote sensing image is not fixed, which causes the distribution of roads to be irregular and deformed (i.e., rotation and scale variation). Therefore, we design a GDEM to adapt to the deformation of the road for better extracting the context, as shown in Fig. 5. Our GDEM is supported by two carefully designed deformation estimation convolution layers (DE Conv), as given in Fig. 5(c). Next, we discuss the difference between the standard convolution and our DE Conv. Formally, let $\mathcal{R}$ be a regular grid that samples values over the input feature maps $x$. For a kernel of size $3 \times 3$, we have

$$\mathcal{R} = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\}. \tag{3}$$

For a standard convolution, each spatial position $p$ in the output feature maps $y$ can be written as

$$y(p) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p + p_n) \tag{4}$$

where $p_n$ denotes the locations in $\mathcal{R}$ and $w$ is the kernel weight. This equation means that the standard convolution has an inductive bias to capture the local context. In contrast, our DE Conv expands this local perception ability to adapt to the deformed shape of the road. As shown in Fig. 5(c), we use two $3 \times 3$ convolutional layers to estimate the rotation angle $\theta_p \in (0, \pi)$ and the scale factor $\lambda_p \in (0, 2]$ for each position $p$.

The following transformation matrix $\mathcal{T}_p$ can be obtained:

$$\mathcal{T}_p = \lambda_p R(\theta_p) = \lambda_p \begin{pmatrix} \cos(\theta_p) & \sin(\theta_p) \\ -\sin(\theta_p) & \cos(\theta_p) \end{pmatrix}. \tag{5}$$

For a DE Conv, each spatial position $p$ in the output feature maps $y$ can be written as

$$y(p) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p + p_n + \triangle P_n) \tag{6}$$

$$\triangle P_n = \mathcal{T}_p p_n - p_n, \quad \text{where} \quad p_n \in \mathcal{R}. \tag{7}$$

From (6) and (7), we can see that our DE Conv can adaptively sample neighborhood pixels through modeling deformation to better obtain local context information. Although our DE Conv draws inspiration from DCN [26], it is significantly different and is more suitable for road extraction task. On the one hand, the original DCN with high DOF offset parameters to sample neighborhood pixels will bring a risk of overparameterizing the local shape. Specifically, DCN needs to predict $2k^2$ offsets in each location for a $k \times k$ kernel, while high DOF parameters tend to give model training risk of overfitting and difficulty optimization. In contrast, DE Conv predicted offsets have only fixed 2-DOF for any size kernel, which greatly reduces the complexity of modeling local shape. On the other hand, our DE Conv retains the linear structure of the conventional square convolution, which can extract the linear features of the road well. In fact, the road is usually straight and therefore does not require a high DOF DCN modeling complex shape, while the linear constraints of our DE Conv can better approximate road shape and have a lower computational cost.

### D. Road Edge Focal Loss

The road in the remote sensing image is thin, making the road pixels often only occupy a small part of the image, as shown in the left part of Fig. 6. This imbalance of pixel categories between the foreground and the background will bring difficulties to the optimization of remote sensing image segmentation [58]. To deal with this problem, we provide a simple but effective REF loss to make the pixels around the road get special attention in optimization in Fig. 6. Specifically, we first use the classic logarithm (LoG) algorithm[1] [59] to detect the edge of the road in the ground-truth road mask. Since the ground-truth mask is binary, the edges of the roads are easily and automatically acquired. Given the edge pixels set $E$ and $e_i \in E, i \in N$, we can sample the corresponding

[1]We implement based on scipy.ndimage.gaussian_laplace function.

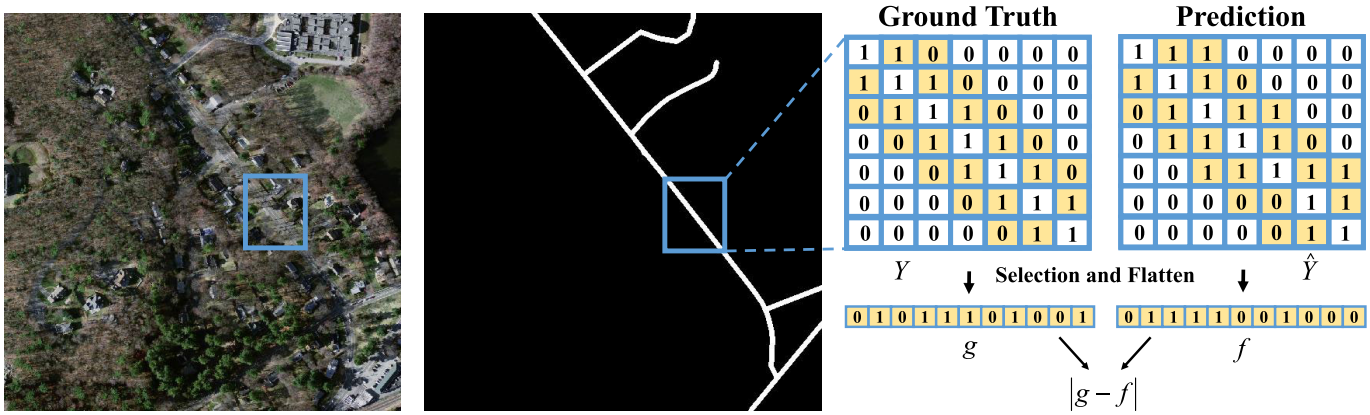



Fig. 6. Illustration of the REF loss, in which **Yellow pixels** are the road edge pixels. Our REF loss makes the network focus on the pixels around the road to alleviate the imbalance of pixel distribution by imposing additional supervision.

vectors $g$ and $f$ from the ground-truth map $Y$ and the prediction map $\hat{Y}$ as follows:

$$g = Y(e_i), \quad f = \hat{Y}(e_i). \tag{8}$$

Then, our REF loss is derived by calculating the $L1$ distance between the selected pixels

$$\mathcal{L}_{\text{REF}} = |g - f|. \tag{9}$$

### E. Training Framework

In this section, we introduce the overall loss function setting in the training phase of RSANet. In addition to the aforementioned REF loss, we also use binary cross-entropy loss (BCE loss) and Dice coefficient loss (DICE loss) [60], which are both widely used in previous road extraction works [5], [6], [7], [17], to supervise network training. Specifically, these two losses can be defined as

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=N}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \tag{10}$$

$$\mathcal{L}_{\text{DICE}} = 1 - \frac{2\sum_{i=1}^{N}(y_i \hat{y}_i)}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} \hat{y}_i^2} \tag{11}$$

where $y_i$ is the ground truth denoting road or background for a given pixel $i$ and $\hat{y}_i$ is the corresponding value in the prediction map. The total loss function of the training framework is defined as

$$\mathcal{L} = \mathcal{L}_{\text{REF}} + \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{DICE}}. \tag{12}$$

## IV. Experiments

In this section, three public datasets are introduced to evaluate and compare the performance of ours and other main stream methods. Specifically, qualitative and quantitative comparison results and ablation studies of each component are presented.

### A. Datasets

*1) Cheng Roads Dataset:* We first use the dataset from the work of [5] to conduct road extraction experiments. This dataset consists of 224 VHR images collected from Google Earth [61], and each image is manually annotated. Specifically, Google Earth takes multispectral images of 13 bands, including visible, near-infrared (NIR), and short-wave infrared bands via the sentinel-2 satellite. These images include various scenes in urban and rural areas, which are more in line with actual application requirements. In this dataset, there are at least 600 × 600 pixels in each image, and each pixel in an image represents 1.2 m in the real world. Following the setting of the previous works [5], [50], we use the same 180 images for training, 14 images for validation, and 30 images for testing.

*2) DeepGlobe Dataset:* DeepGlobe Dataset [62] contains satellite images collected from Thailand, Indonesia, and India, which covers a total land area of 2220 km$^2$ containing both urban and suburban areas, while each image has a size of 1024 × 1024 pixels and each pixel represents 0.5 m. Multispectral images are taken by the worldview-3 satellite and are in nine bands, including PAN, red, blue, green, red edge, coastal, yellow, and NIR. This original dataset contains 6226 images that are openly available with ground-truth road masks, and we further divide the accessible data into the training, validating, and testing sets. Following CoANet [52], 4696 images are selected for training, and the other 1530 images are selected for testing, with all images cropped into 512 × 512 as the input.

*3) Massachusetts Roads Dataset:* The Massachusetts Roads dataset [63] includes 1171 optic aerial images (1500 × 1500) of the state of Massachusetts, which covers an area of over 2600 km$^2$ and includes various scenes of urban, suburban, and rural regions. These images are optical remote sensing images with three bands (blue, green, and red) and each pixel represents 1 m. This dataset is divided into 1108 training images and 63 test images, and all images are cropped into 500 × 500 without overlap and resized into 512 × 512 as the input of the network.

### B. Implementation Details

We use the Adam optimizer [64] to optimize our network and set the learning rate to $5 \times 10^{-4}$ and decay to 0 with the "poly" learning rate iterative policy, where the learning rate is multiplied by $(1 - (\text{iter}/\text{maxiter}))^{\text{power}}$ with power = 3.

All training runs on a single NVIDIA Titan V GPU with 11-GB GPU memory. The batch size is set to 16 on all datasets. The maximum epoch for training on the DeepGlobe and Massachusetts Roads datasets was set to 150, while the maximum epoch for the Cheng Roads dataset was set to 300 due to the fact that the number of images in the Cheng Roads dataset is much smaller than the other two datasets. In the experiments, all methods and variants used the same dataset division and other settings. The size of the training images on all datasets is 512 × 512. During training, data augmentation, including random rotation, flipping, shift, and scaling, is applied to improve the generalization of the model and prevent overfitting.

### C. Evaluation Metrics

To evaluate the road extraction performance of our method, we use four popular metrics for evaluation: precision ($P$), recall ($R$), F1 score ($F1$), and mean intersection over union (IoU) (mIoU). For convenience, we use the following denotations: TP for true positive, TN for true negative, FP for false positive, and FN for false negative. Then, four indicators are calculated as

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{13}$$

$$R = \frac{\mathrm{TP}}{\mathrm{FP}+\mathrm{FN}} \tag{14}$$

$$F1 = \frac{2 \times P \times R}{P+R} \tag{15}$$

$$\mathrm{mIoU} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}\right). \tag{16}$$

For these four metrics, the higher the evaluation score, the better the performance.

### D. Comparison With Advanced Methods

We compared our RSANet with five state-of-the-art road extraction methods, including DeepLab V3+ [65], SIINet [8], DLinkNet [7], NLLinkNet [9], CoANet [52], and most recent SDUNet [66]. For a fair comparison, all methods are experimented with the same dataset settings. Since the datasets used are publicly available, we also show some results reported in other works that use the same dataset settings as this work.

*1) Comparisons on the Cheng Roads Dataset:* Table I reports the comparative quantitative evaluation measured in terms of $P$, $R$, $F1$, and mIoU on the Cheng Roads dataset [5]. Note that these methods have the same dataset settings as our method since the partition of the Cheng Road dataset is fixed. Our RSANet achieves a new state-of-the-art evaluation score (99.03%/96.98%/97.99%/94.21% on $P$/$R$/$F1$/mIoU) on the Cheng Roads dataset. Specifically, our method achieves the highest $P$, $F1$, and mIoU scores while achieving the second highest $R$ score.

To further verify the effectiveness of our RSANet, we also present qualitative comparisons on the Cheng Roads dataset in Fig. 7. Although all methods achieve good overall segmentation results, our RSANet has better detail segmentation results than other advanced methods. COANet can handle common straight roads well due to the introduction of fixed direction strip convolution layer but has flaws in dealing with circular roads [as in Fig. 7(d)]. In contrast, our RSANet can cope with these complex geometric deformations better with the help of GDEM that can flexibly estimate the pixel-level context extraction direction.

TABLE I
QUANTITATIVE COMPARISON OF OUR RSANET WITH SOME ADVANCED ROAD EXTRACTION METHODS ON THE CHENG ROADS DATASET, IN WHICH BOLD IS THE BEST AND UNDERLINE IS THE SECOND BEST

| Cheng Roads Dataset [5] | | | | |
|---|---|---|---|---|
| *Methods* | *P(%)* | *R(%)* | *F1(%)* | *mIoU(%)* |
| LinkNet [49] | 95.82 | 92.16 | 92.84 | 90.92 |
| ResUNet [6] | 96.15 | 93.58 | 94.84 | 91.23 |
| RCNN-UNet [50] | 97.86 | 95.12 | 96.47 | 92.89 |
| DeepLab V3+ [65] | 97.38 | 94.68 | 96.01 | 92.34 |
| DLinkNet [7] | 97.02 | 95.24 | 96.12 | 92.43 |
| SIINet [8] | 97.54 | 94.63 | 96.06 | 92.60 |
| NLLinkNet [9] | 97.72 | 96.63 | 97.17 | 92.84 |
| SDUNet [66] | 97.54 | 96.31 | 96.92 | 92.74 |
| CoANet [52] | <u>98.06</u> | **97.33** | <u>97.69</u> | <u>93.10</u> |
| RSANet (ours) | **98.96** | <u>96.81</u> | **97.87** | **93.82** |

TABLE II
QUANTITATIVE COMPARISON OF THE PROPOSED RSANET WITH SOME ADVANCED ROAD EXTRACTION METHODS ON THE DEEPGLOBE DATASET, IN WHICH BOLD IS THE BEST AND UNDERLINE IS THE SECOND BEST

| DeepGlobe Dataset [62] | | | | |
|---|---|---|---|---|
| *Methods* | *P(%)* | *R(%)* | *F1(%)* | *mIoU(%)* |
| LinkNet [49] | 71.33 | 79.81 | 75.33 | 61.34 |
| ResUNet [6] | 72.86 | 81.29 | 76.84 | 63.14 |
| RCNN-UNet [50] | 73.57 | 82.18 | 77.63 | 64.23 |
| DeepLab V3+ [65] | 73.14 | 82.10 | 77.36 | 63.23 |
| DLinkNet [7] | 73.50 | 81.38 | 77.24 | 63.36 |
| SIINet [8] | 75.42 | 83.15 | 79.09 | 64.35 |
| NLLinkNet [9] | 76.08 | 82.84 | 79.31 | 66.14 |
| SDUNet [66] | <u>78.40</u> | 74.20 | .79.40 | 66.80 |
| CoANet [52] | 77.88 | <u>84.85</u> | <u>81.22</u> | <u>68.37</u> |
| RSANet (ours) | **78.84** | **86.33** | **82.42** | **70.26** |

*2) Comparisons on the DeepGlobe Dataset:* Table II reports the comparative quantitative evaluation on the DeepGlobe dataset [62], which collects more diverse scenes and contains a large number of country roads that are constantly changing in width and shape. Besides, trees and tree shadows in the wild also pose serious occlusions, making DeepGlobe dataset to be more challenging than the other two datasets. The experiment results in Table II show that our RSANet can still handle the challenges of such complex scenarios well and achieves the evaluation scores of state-of-the-art. It is worth noting that our RSANet improves the mIoU by 1.89% compared to the suboptimal CoANet [8]. In addition, our RSANet also

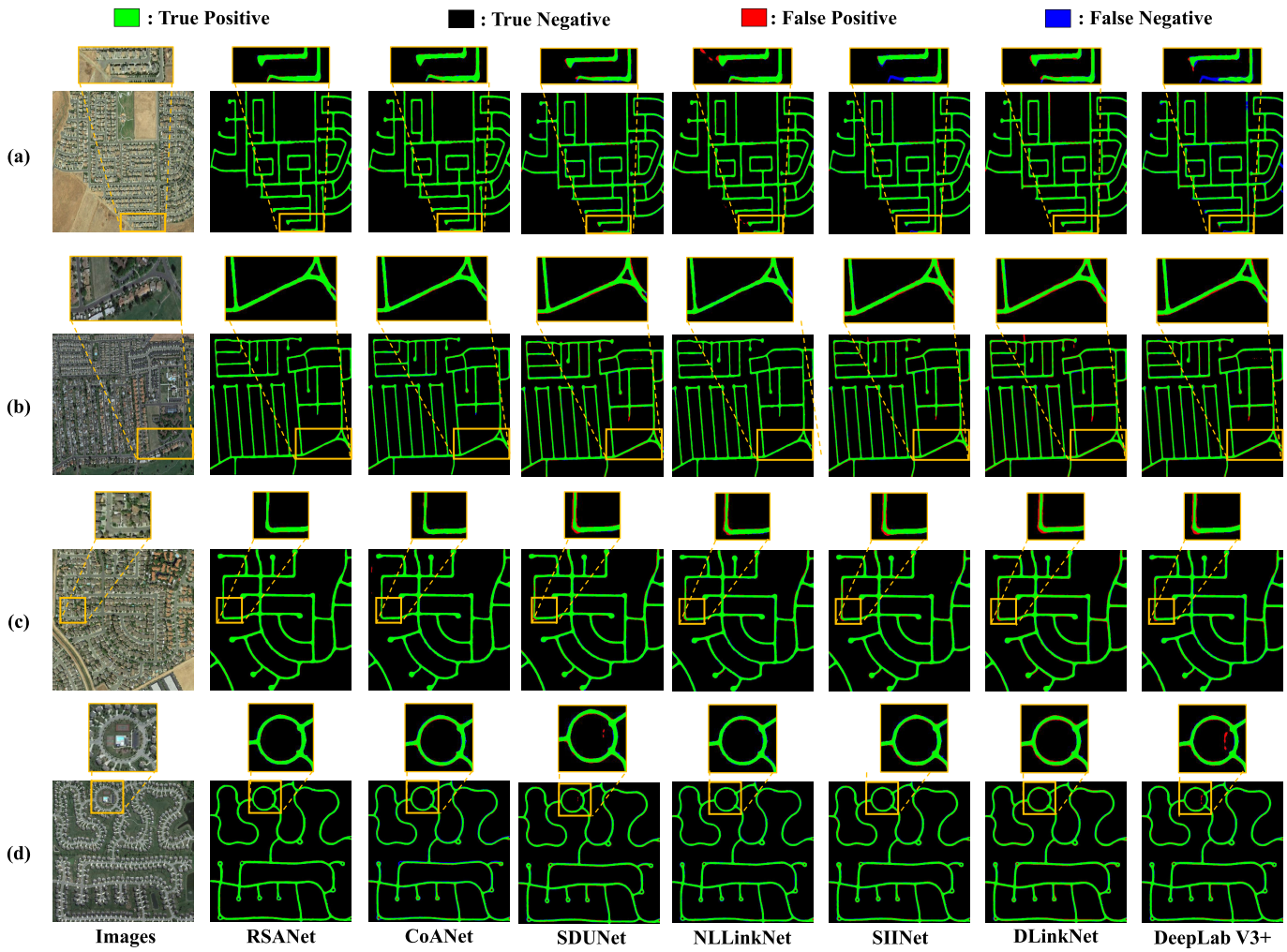



Fig. 7. Qualitative comparison of the proposed RSANet with other state-of-the-art road extraction methods on the Cheng Roads dataset.

achieves the best balance between precision $P$ and recall $R$. The corresponding recall and precision values are 86.33% and 78.84%, respectively, indicating that 86.33% of all road pixels in the reference map are detected correctly and 78.84% of the detected road pixels are also road pixels in the ground-truth map.

In addition to quantitative analysis, qualitative analysis is equally important. Specifically, we show in Fig. 8 a visual comparison of CoANet, which performs the closest to our RSANet, on some challenging samples. The results show that our method is closer to the ground truth than CoANet. More qualitative comparisons with the three state-of-the-art methods are shown in Fig. 9. Our RSANet gives a more complete road mask (green pixels in Fig. 9) and produces less false alarms (red pixels in Fig. 9), as our method can better capture both global and local context according to road shape features. Specifically, the global context awareness brought by the proposed ESTM maintains the continuity of long-span roads [e.g., Fig. 9(c) and (e)], while the local context awareness brought by the proposed GDEM can handle the complex reasoning deformations [e.g., Fig. 9(a), (b), and (d)].

*3) Comparisons on the Massachusetts Roads Dataset:* Table III reports the comparative quantitative evaluation results of our RSANet and some competitive approaches on the Massachusetts Roads dataset [63]. Road extraction for this dataset is challenging as many roads are occluded by trees, shadows, and buildings, which seriously affect the road's integrity and clarity in the remote sensing images. In this case,

TABLE III

QUANTITATIVE COMPARISON OF THE PROPOSED RSANET WITH SOME ADVANCED ROAD EXTRACTION METHODS ON THE MASSACHUSETTS DATASET, IN WHICH BOLD IS THE BEST AND UNDERLINE IS THE SECOND BEST

| Massachusetts Dataset [63] | | | | |
|---|---|---|---|---|
| *Methods* | *P(%)* | *R(%)* | *F1(%)* | *mIoU(%)* |
| LinkNet [49] | 76.56 | 71.82 | 74.11 | 69.80 |
| ResUNet [6] | 79.74 | 72.54 | 75.96 | 70.85 |
| RCNN-UNet [50] | 80.98 | 75.10 | 77.92 | 71.86 |
| DeepLab V3+ [65] | 75.14 | 72.56 | 73.83 | 69.29 |
| DLinkNet [7] | 78.11 | 73.06 | 75.50 | 71.10 |
| SIINet [8] | 85.36 | 74.13 | 79.35 | 73.86 |
| NLLinkNet [9] | 86.64 | <u>78.24</u> | <u>82.23</u> | 73.59 |
| SDUNet [66] | 81.20 | 75.70 | 78.40 | <u>74.10</u> |
| CoANet [52] | <u>87.34</u> | 77.08 | 81.89 | 74.00 |
| RSANet (ours) | **88.22** | **78.54** | **83.10** | **76.17** |

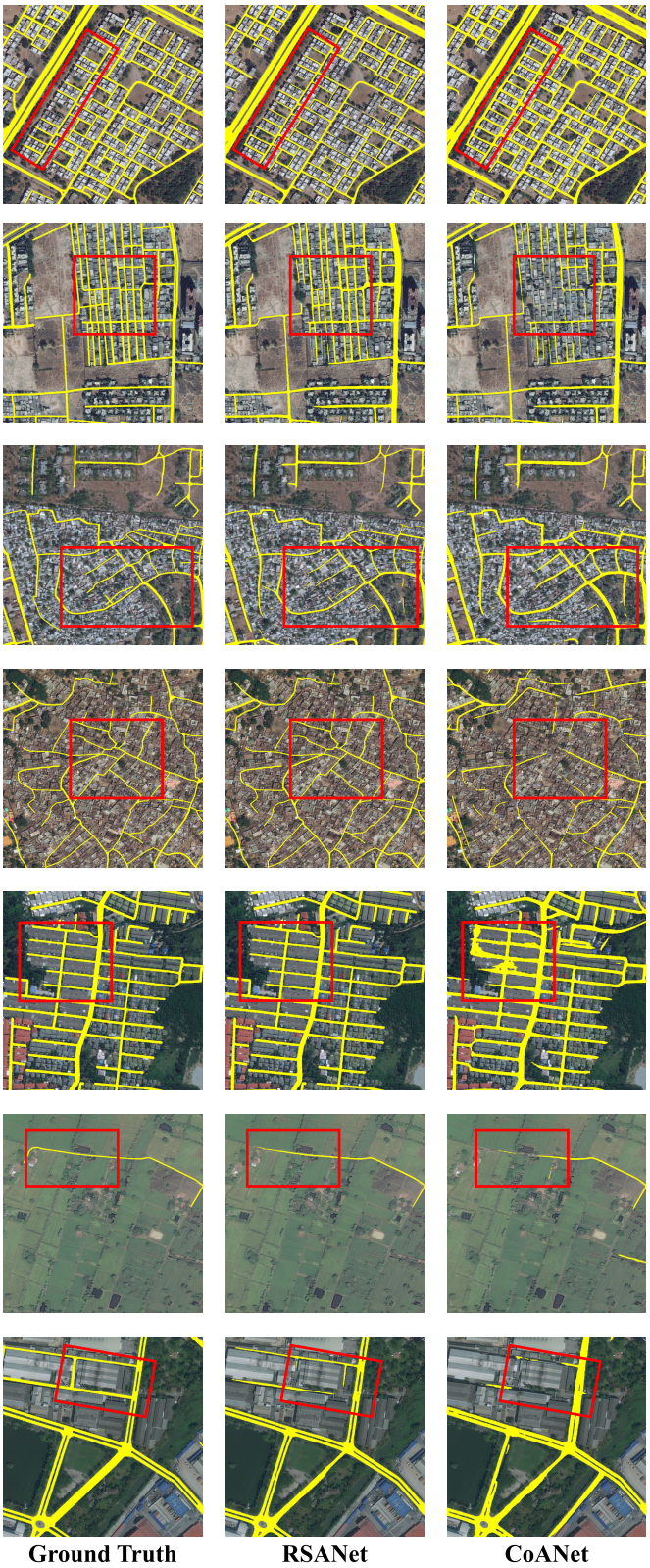



Fig. 8. Some examples of challenging road extraction on the DeepGlobe dataset. Compared with the recent CoANet [52], the segmentation results of our RSANet are closer to the ground truth.

SIINet [8], NLLinkNet [9], SDUNet [66], and CoANet [52] achieve competitive performance since their designed network architectures have better context capture ability to alleviate the challenge of occlusion. Since our RSANet designs novel modules and losses according to the shape characteristics of the road to better perceive context information, we achieve higher evaluation scores than them (i.e., 2.07% higher than SDUNet and 2.17% higher than CoANet on mIoU). Furthermore, our RSANet also achieves the highest *P*, *R*, and *F1* proving its generalization on the Massachusetts dataset. These scores are reflected in the higher accuracy of our proposed method in extracting road boundaries with fewer false detections from a cluttered background. Specifically, we present a qualitative comparative example in Fig. 10. Our RSANet has fewer false positive and false negative regions in both urban and rural examples. Specifically, in some regions where occlusions exist, the roads extracted by other methods may be disconnected, while our RSANet maintains the connection well, as shown in Fig. 10. Furthermore, due to the proposed road edge focus loss, our RSANet has higher quality edge segmentation results, as shown in Fig. 10(e).

### *E. Ablation Study*

To evaluate the effectiveness of each proposed component in our RSANet, including ESTM, GDEM, and REF loss, we performed comprehensive ablation studies. Specifically, we report the *F1* and mIoU scores of different variants of our method on the Cheng Road and DeepGlobe datasets, as described in the following. In addition, we also give qualitative comparison and analysis of ablation experiments in Fig. 11.

*1) Effectiveness of the Proposed Components:* Table IV reports the results of the ablation studies for our three proposed components. In addition, the design motivation of these three components is also given in Table IV ("*long, deform., thin*" in the second row). Motivated by the road shape features described by these three words, we carefully design the three components of RSANet.

*a) Quantitative analysis:* Specifically, our ESTM is used to model long-distance dependencies to cope with long roads shape, GDEM is employed to flexibly capture local context according to road deformation to adapt to complex road deformations, and the REF loss is designed to alleviate the uneven class distribution caused by thin road shape. The baseline in Table IV is obtained by removing all proposed components and using only Dice and BCE loss for training the model. Then, we can get significant performance gains by adding each component individually to the baseline, in which GDEM provides the largest improvement. The highest evaluation scores are achieved after we sequentially add components to the baseline. Besides, our proposed components improve the mIoU of Cheng road from 89.89% to 93.82% (+3.93%) and the mIoU of DeepGlobe from 61.52% to 70.26% (+8.74%). These results validate that our proposed components can provide independent and consistent performance improvements on different road extraction datasets. The proposed SETM contributes the highest magnitude of performance improvement on both datasets, validating the importance of improving long-range context awareness for the road extraction task. Note that GDEM achieves higher gains on the DeepGlobe dataset than on the Cheng Road dataset, which may be due to more complex road deformations on the DeepGlobe dataset. Evidently, the consistent performance improvement after applying

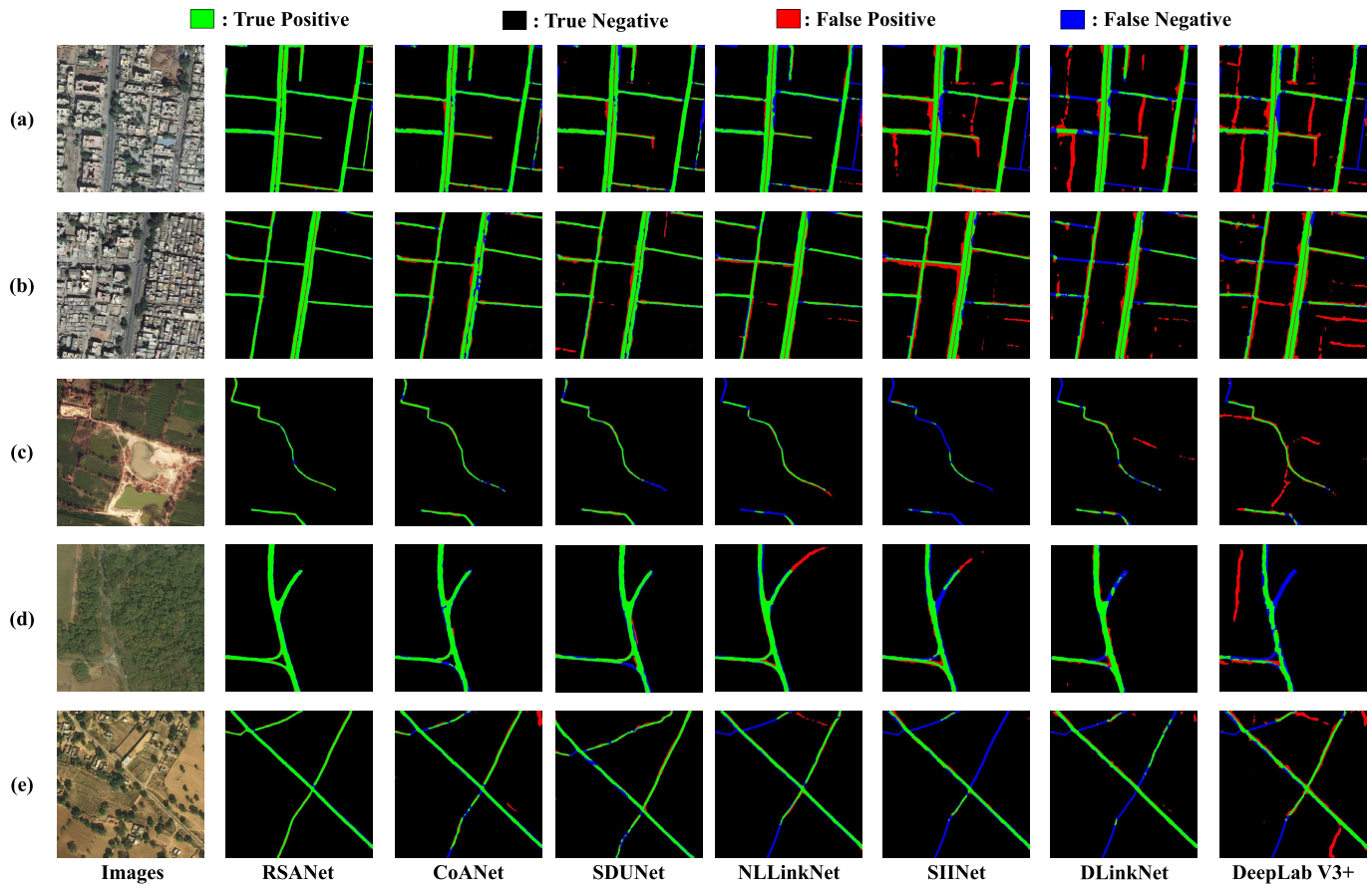



Fig. 9. Qualitative comparison of proposed RSANet with other state-of-the-art road extraction methods on the DeepGlobe dataset.

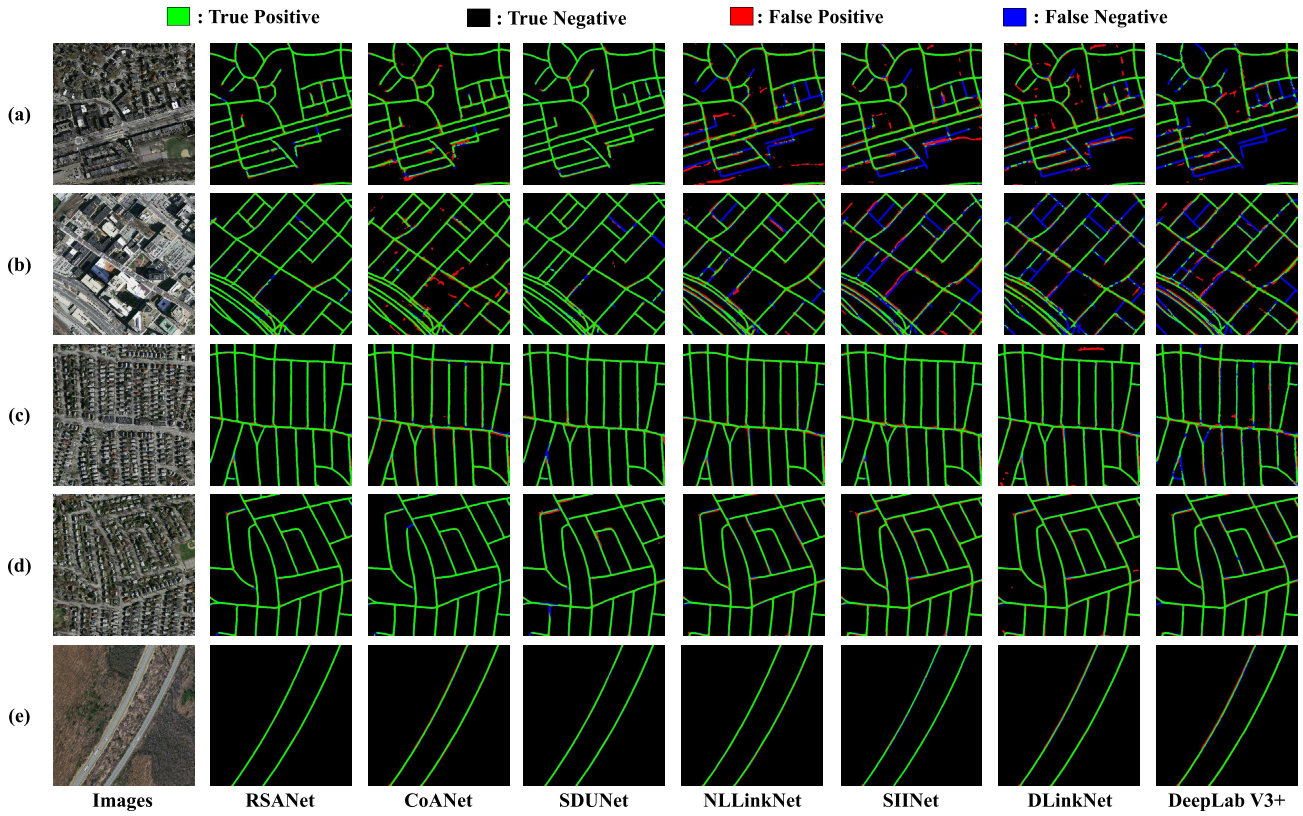



Fig. 10. Qualitative comparison of the proposed RSANet with other state-of-the-art road extraction methods on the Massachusetts Roads dataset.

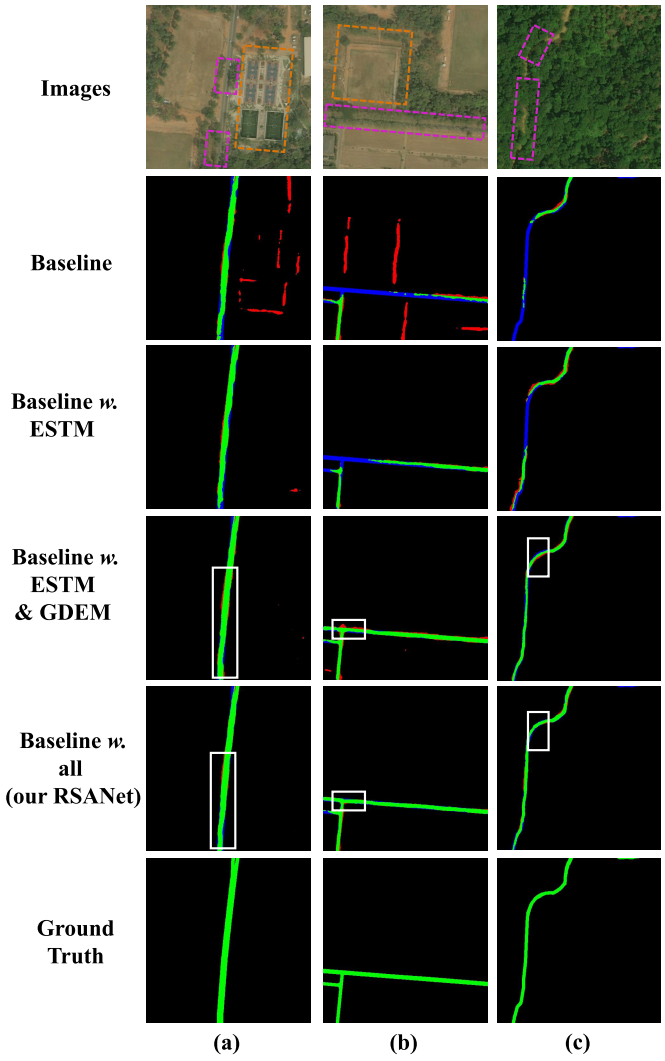



Fig. 11. Qualitative comparison of ablation experiments on challenge examples of DeepGlobe dataset. Green pixels mean true positive samples. Black pixel means true negative samples. Red pixels mean false positive samples. Blue pixels mean false negative samples. Pink boxes indicate occlusion areas. Orange boxes indicate interference areas.

TABLE IV
ABLATION STUDY ON THE PROPOSED COMPONENTS. HERE, ✔MEANS THAT THIS COMPONENT IS APPLIED

| *Variants* | | | Cheng Road | | DeepGlobe | |
|---|---|---|---|---|---|---|
| *long* ESTM | *deform.* GDEM | *thin* REF loss | *F1(%)* | *mIoU(%)* | *F1(%)* | *mIoU(%)* |
| | | | 94.02 | 89.89 | 77.15 | 61.52 |
| ✔ | | | 96.24 | 92.17 | 79.83 | 67.82 |
| | ✔ | | 95.62 | 91.05 | 79.64 | 67.74 |
| | | ✔ | 95.10 | 90.21 | 78.51 | 63.86 |
| ✔ | ✔ | | 97.31 | 93.22 | 81.93 | 69.45 |
| ✔ | ✔ | ✔ | **97.87** | **93.82** | **82.42** | **70.26** |

REF loss on RSANet suggests that increasing attention to road edges can further improve segmentation accuracy.

*b) Qualitative analysis:* Although the evaluation on average pixel-level numerical metrics in Table IV can reflect the overall quality of road extraction, it cannot show how these variants perform in some challenging scenes that often arise in real-world applications. Therefore, we also present a qualitative comparison of different variants in ablation studies on some challenging examples, as shown in Fig. 11. In Fig. 11(a), the part of the road is obscured by trees and there is a large area of road-like building interference. In this case, the baseline has a large number of false alarms (red pixels) and missing road pixels (blue pixels). The false alarms are significantly reduced when our ESTM with global context capture capability is applied. In addition, our GDEM reduces road loss pixels through adaptive shape local context awareness. When the REF loss is also applied, the road edges become smoother, which further improves the road extraction performance. In Fig. 11(b), the road is almost completely covered by plants, yet our method still achieves an exciting performance. The road in Fig. 11(c) has a curved shape distribution, in addition to being occluded by trees. The results show that our RSANet can still cope with this challenging example well after adding all proposed components, especially applying our GDEM that can adaptively model road geometric deformations.

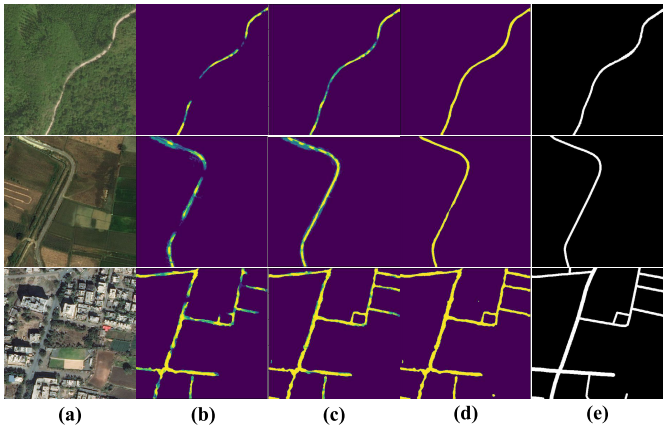



Fig. 12. Visualization of the feature maps (mean value) of the last decoder. (a) Input images. (b) *w/o* ESTM. (c) *w/o* GDEM. (d) RSANet. (e) Ground truths.

*c) Visualization analysis:* To investigate the contribution of the proposed modules to the feature learning capability, we visualize the mean value of the output feature maps of the last decoder convolution layer in Fig. 12. Specifically, Fig. 12(b) and (c) represents the feature maps when ESTM and GDEM are not connected, respectively, while Fig. 12(d) shows the feature maps when they are both working. It can be seen that when ESTM is removed, the continuity of road features is significantly affected, especially in the case of tree/vehicle shading. After removing GDEM, the activation value becomes lower in some places on the road where the deformation or occlusion exists. This demonstrates that the proposed ESTM and GDEM can capture road-related context even under challenging situations where visual features are not reliable.

*2) Ablation Study on Our GDEM:* In our GDEM, we decompose the deformation of road shapes in remote sensing images into rotation and scale transformations. Table V shows the effects of rotation and scale estimation in our GDEM. Here DOF denotes the DOF of the deformation estimation, and a higher DOF means a higher computational

TABLE V
ABLATION STUDY ON GDEM. HERE, ✔MEANS THAT THIS COMPONENT IS APPLIED

| *DoF* | Variants | | Cheng Road | | DeepGlobe | |
|---|---|---|---|---|---|---|
| | *Rotation* | *Scale* | *F1(%)* | *mIoU(%)* | *F1(%)* | *mIoU(%)* |
| 0 | | | 96.52 | 92.29 | 80.73 | 68.37 |
| 1 | ✔ | | 97.26 | 92.98 | 82.20 | 69.39 |
| 1 | | ✔ | 96.68 | 92.44 | 82.27 | 69.45 |
| 2 | ✔ | ✔ | **97.87** | **93.82** | **82.42** | **70.26** |
| $2k^2$ | *w.* DCN [26] | | 96.71 | 92.56 | 81.31 | 69.04 |

TABLE VI
ABLATION STUDY ON OUR ESTM. HERE, ✔MEANS THAT THIS COMPONENT IS APPLIED

| *Variants* | | Cheng Road | | DeepGlobe | |
|---|---|---|---|---|---|
| *height axis* | *width axis* | *F1(%)* | *mIoU(%)* | *F1(%)* | *mIoU(%)* |
| | | 95.98 | 91.45 | 79.98 | 67.90 |
| ✔ | | 96.75 | 92.98 | 81.10 | 68.23 |
| | ✔ | 97.04 | 93.12 | 81.21 | 68.44 |
| ✔ | ✔ | **97.87** | **93.82** | **82.42** | **70.26** |
| *w.* Non-local Block [9] | | 96.88 | 91.74 | 81.07 | 69.33 |

TABLE VII
EVALUATION OF RESOLUTION ROBUSTNESS ON THE CHENG ROAD DATASET. "RESOLUTION" REFERS TO THE MINIMUM RESOLUTION OF THE TEST IMAGES

| Resolution | Multiplier | Cheng *mIoU(%)* RSANet | Cheng *mIoU(%)* CoANet [52] |
|---|---|---|---|
| 2400 × 2400 | ×4 | 94.07 | 93.58 |
| 1200 × 1200 | ×2 | 94.12 | 93.43 |
| 600 × 600 | ×1 | 93.82 | 93.10 |
| 300 × 300 | $\times\frac{1}{2}$ | 87.24 | 86.88 |
| 150 × 150 | $\times\frac{1}{4}$ | 64.05 | 60.74 |

complexity with more parameters to be learned. For the original DCN, the sampling points of each convolution kernel need to predict two offsets in the horizontal and vertical directions. Therefore, the original DCN without any geometric constraints requires a DOF of $2k^2$, where $k$ is the kernel size of the convolution layer. Taking $3 \times 3$ convolution as an example, our proposed DE Conv reduces the DOF from 18 to 2. As shown in Table V, the estimation of the rotation angle provides more performance contribution than the scale. In addition, simultaneously estimating rotation and scale from road shape can capture local context information more flexibly and achieve the higher metrics scores, especially on the DeepGlobe Dataset. Here, *replace w.* DCN in Table V means to replace DE Conv in our GDEM with DCN [26], while the results show that the metric scores of proposed DE Conv are significantly higher than the original DCN due to the deformation estimation capability more suitable for road shape characteristics.

*3) Ablation Study on Our ESTM:* In this section, we explore the influence of the two branches of the height axis and width axis in the ESTM, as shown in Table VI. As described in Table VI, these two branches bring close performance improvements. Note that higher metrics scores are achieved when they are used together, as these two branches are complementary and combined to provide more comprehensive global context information. Here, *w.* nonlocal block in Table VI means to replace ESTM with nonlocal block proposed in [17]. Although nonlocal block also has the ability to model long-range dependencies, it needs higher complexity than our ESTM [$O(N^2)$ versus $O(N \cdot H + N \cdot W)$]. Furthermore, on the more challenging DeepGlobe dataset, our ESTM achieves a performance advantage with large margin. This may be due to the sparse distribution of roads in the DeepGlobe dataset, where a large number of nonroad pixels affect the pixel-to-pixel affinity modeling in the nonlocal block. In contrast, our ESTM decomposes long-range distance modeling into three steps: first aggregating local context information through strip convolutions that conform to the shape of the road distribution, then broadcasting the context along opposite directions, and finally fusing the complementary contexts of the two branches. Our ESTM significantly reduces the computational complexity of nonlocal interpixel modeling and can avoid the influence of a large number of road-independent background pixels.

*4) Ablation Study on Resolution Robustness:* Table VII reports the resolution robustness study of our method and the most competitive CoANet [52] on the Cheng Road dataset [5]. The models are tested with inputs of different resolutions to verify resolution robustness. Specifically, we use bilinear interpolation to adjust the resolution of test images to obtain the corresponding output masks and then scale the masks to the original resolution for comparison with the ground-truth masks. The results show that the performance of the segmentation model does not degrade or even improves when the image resolution is increased. However, the performance improvement is limited, which may be caused by the fact that the interpolated image does not contain detailed information compared to the true high-resolution image. Interestingly, the segmentation performance decreases significantly when the resolution of the test image decreases. This means that there is still room for improvement in the road extraction methods to deal with low resolution or particularly small roads. To the best of our knowledge, there are almost no existing methods to

TABLE VIII
COMPARISON ON RUNNING CONSUMPTION, IN WHICH BOLD IS THE BEST AND UNDERLINE IS THE SECOND BEST

| *Methods* | FLOPs | Parameters | FPS | DeepGlobe *mIoU(%)* |
|---|---|---|---|---|
| Baseline | 22.72G | 10.65M | 70 | 61.52 |
| + ESTM | 23.79G | 13.14M | **66** | 67.82 |
| + GDEM | 24.28G | 13.85M | 57 | 69.45 |
| + REF loss (our RSANet) | <u>24.28G</u> | <u>13.85M</u> | 57 | **70.26** |
| SIINet $_{ISPRS'19}$ [8] | **23.96G** | **11.36M** | <u>62</u> | 64.35 |
| NLLinkNet $_{TGRSL'21}$ [9] | 31.41G | 21.82M | 48 | 66.14 |
| CoANet $_{TIP'21}$ [52] | 47.42G | 59.15M | 42 | <u>68.37</u> |
| SDUNet $_{PR'22}$ [66] | 53.28G | 80.24M | 31 | 66.80 |

explore this problem, and we plan to improve the performance of the challenging low-resolution (small roads) segmentation in future.

### F. Comparison on Running Efficiency

We report the floating point operations (FLOPs), parameters, and FPS of some variants and recent methods in Table VIII. Specifically, experiments are performed on a single Titan V GPU and use a single 512 × 512 image as the input to these models. In addition to sequentially applying the variants driven by our proposed components, we also report four competitive approaches SIINet [8], NLLinkNet [9], SDUNet [66], and CoANet [52]. From Table VIII, we can see that our ESTM causes the most computation and space consumption, while our REF loss does not cause consumption during the inference phase and only works in the training phase. Although visual transformers tend to bring more resource consumption, our ESTM is only deployed at low resolution and employs the efficient axial self-attention mechanism to reduce the cost. The computation of our RSANet is significantly less than recent NLLinkNet and CoANet, whose performance is closest to our method. Although the consumption of SIINet is comparable to our RSANet, its performance is lower than RSANet (70.26% versus 64.35%). As a result, our RSANet achieves the best tradeoff of speed and accuracy for efficient and accurate road extraction and shows potential for practical applications.

## V. CONCLUSION

In this article, we propose an RSANet for efficient and accurate road extraction from remote sensing images. According to the shape characteristics of roads, three corresponding components are carefully designed to better capture effective context learning for complex road extraction. First, we introduce the ESTM to efficiently capture the global context to model the long-distance dependence required by long roads. Then, we design GDEM to adaptively model road deformation for better local feature extraction. Besides, we provide a special REF loss to guide the network focus on the optimization of the pixels around the road to mitigate optimization difficulty caused by the thin roads. Finally, experiments are carried out on three challenging datasets and validate that our RSANet can achieve the state-of-the-art performance. Specifically, our RSANet achieves an mIoU score of 70.26% on the DeepGlobe dataset with a competitive operating efficiency (57 FPS) and a size of parameters (13.82M).

## REFERENCES


[1] Z. Sun, J. Wu, J. Yang, Y. Huang, C. Li, and D. Li, "Path planning for GEO-UAV bistatic SAR using constrained adaptive multiobjective differential evolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 11, pp. 6444–6457, Nov. 2016.

[2] W. J. Emery, D. Baldwin, and D. Matthews, "Maximum cross correlation automatic satellite image navigation and attitude corrections for open-ocean image navigation," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 1, pp. 33–42, Jan. 2003.

[3] D. Mckeown, "The role of artificial intelligence in the integration of remotely sensed data with geographic information systems," *IEEE Trans. Geosci. Remote Sens.*, vol. GE-25, no. 3, pp. 330–348, May 1987.

[4] R. Miyamoto et al., "Vision-based road-following using results of semantic segmentation for autonomous navigation," in *Proc. IEEE 9th Int. Conf. Consum. Electron. (ICCE-Berlin)*, Sep. 2019, pp. 174–179.

[5] 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.

[6] 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.

[7] 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.

[8] C. Tao, J. Qi, Y. Li, H. Wang, and H. Li, "Spatial information inference net: Road extraction using road-specific contextual information," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 155–166, Dec. 2019.

[9] Y. Wang, J. Seo, and T. Jeon, "NL-LinkNet: Toward lighter but more accurate road extraction with nonlocal operations," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[10] Y. Wang et al., "DDU-Net: Dual-decoder-U-Net for road extraction using high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4412612.

[11] A. Abdollahi, B. Pradhan, and A. Alamri, "SC-RoadDeepNet: A new shape and connectivity-preserving road extraction deep learning-based network from remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5617815.

[12] Y. Xu, H. Chen, C. Du, and J. Li, "MSACon: Mining spatial attention-based contextual information for road extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604317.

[13] B. Southall and C. J. Taylor, "Stochastic road shape estimation," in *Proc. 8th IEEE Int. Conf. Comput. Vision. (ICCV)*, Jul. 2001, pp. 205–212.

[14] H. Wang, Y. Hou, and M. Ren, "A shape-aware road detection method for aerial images," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 4, Apr. 2017, Art. no. 1750009.

[15] L. Jiao, Y. Liu, and H. Li, "Characterizing land-use classes in remote sensing imagery by shape metrics," *ISPRS J. Photogramm. Remote Sens.*, vol. 72, pp. 46–55, Aug. 2012.

[16] K. Wang and D. Ming, "Road extraction from high-resolution remote sensing images based on spectral and shape features," *Proc. SPIE*, vol. 7495, Oct. 2009, Art. no. 74953R.

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

[18] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, and L. Sagun, "ConViT: Improving vision transformers with soft convolutional inductive biases," 2021, *arXiv:2103.10697*.

[19] Z. Sun, W. Zhou, C. Ding, and M. Xia, "Multi-resolution transformer network for building and road segmentation of remote sensing image," *ISPRS Int. J. Geo-Inf.*, vol. 11, no. 3, p. 165, Feb. 2022.

[20] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[21] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. NIPS*, vol. 28. Red Hook, NY, USA: Curran Associates, 2015, pp. 2017–2025.

[22] J. Ding, N. Xue, Y. Long, G. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2844–2853.

[23] Y. Xu et al., "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.

[24] X. Yang et al., "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8231–8240.

[25] Z. Liu, H. Wang, L. Weng, and Y. Yang, "Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 8, pp. 1074–1078, Aug. 2016.

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

[27] J. Youn and J. S. Bethel, "Adaptive snakes for urban road extraction," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 35, no. 3, pp. 465–470, 2004.

[28] A. Dal Poz and G. Do Vale, "Dynamic programming approach for semi-automated road extraction from medium-and high-resolution images," *ISPrS Arch.*, vol. 34, no. 3, p. W8, 2003.

[29] X. Lin, R. Zhang, and J. Shen, "A template-matching based approach for extraction of roads from very high-resolution remotely sensed imagery," *Int. J. Image Data Fusion*, vol. 3, no. 2, pp. 149–168, Jun. 2012.

[30] W. Shi, Z. Miao, Q. Wang, and H. Zhang, "Spectral–spatial classification and shape features for urban road centerline extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 4, pp. 788–792, Apr. 2014.

[31] M. Rajeswari, K. S. Gurumurthy, L. P. Reddy, S. N. Omkar, and J. Senthilnath, "Automatic road extraction based on normalized cuts and level set methods," *Int. J. Comput. Appl.*, vol. 18, no. 7, pp. 10–16, Mar. 2011.

[32] D. Herumurti, K. Uchimura, G. Koutaki, and T. Uemura, "Urban road extraction based on Hough transform and region growing," in *Proc. 19th Korea-Japan Joint Workshop Frontiers Comput. Vis.*, Jan. 2013, pp. 220–224.

[33] H. Youquan, Q. Hanxing, W. Jian, Z. Wei, and X. Jianfang, "Studying of road crack image detection method based on the mathematical morphology," in *Proc. 4th Int. Congr. Image Signal Process.*, vol. 2, Oct. 2011, pp. 967–969.

[34] N. L. Gavankar and S. K. Ghosh, "Automatic building footprint extraction from high-resolution satellite image using mathematical morphology," *Eur. J. Remote Sens.*, vol. 51, no. 1, pp. 182–193, Jan. 2018.

[35] S. Guiming and S. Jidong, "Remote sensing image edge-detection based on improved Canny operator," in *Proc. 8th IEEE Int. Conf. Commun. Softw. Netw. (ICCSN)*, Jun. 2016, pp. 652–656.

[36] X. Huang and L. Zhang, "Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines," *Int. J. Remote Sens.*, vol. 30, no. 8, pp. 1977–1987, Apr. 2009.

[37] M. Song and D. Civco, "Road extraction using SVM and image segmentation," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 12, pp. 1365–1371, Dec. 2004.

[38] V. K. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in linear SVM: A review," *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 803–855, 2019.

[39] Q. Zhang and I. Couloigner, "Benefit of the angular texture signature for the separation of parking lots and roads on high resolution multi-spectral imagery," *Pattern Recognit. Lett.*, vol. 27, no. 9, pp. 937–946, Jul. 2006.

[40] 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, Nov. 2011.

[41] S. Das, T. T. Mirnalinee, and K. Varghese, "Use of salient features for the design of a multistage framework to extract roads from high-resolution multispectral satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3906–3931, Oct. 2011.

[42] 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.

[43] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The Handbook of Brain Theory and Neural Networks*, vol. 3361, no. 10. New York, NY, USA: ACM Digital Library, 1995, p. 1995.

[44] 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.

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

[46] 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.

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

[48] 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.

[49] 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.

[50] 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, Sep. 2019.

[51] 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. 10243–10254, Dec. 2021.

[52] J. Mei, R. Li, W. Gao, and M. Cheng, "CoANet: Connectivity attention network for road extraction from satellite imagery," *IEEE Trans. Image Process.*, vol. 30, pp. 8540–8552, 2021.

[53] 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, May 2021.

[54] Q. Zhu et al., "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, May 2021.

[55] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan, "Scale-adaptive convolutions for scene parsing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2050–2058.

[56] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 108–126.

[57] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[58] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4095–4104.

[59] M. Basu, "Gaussian-based edge-detection methods—A survey," *IEEE Trans. Syst., Man Cybern., C, Appl. Rev.*, vol. 32, no. 3, pp. 252–260, Aug. 2002.

[60] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[61] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.*, vol. 202, pp. 18–27, Dec. 2017.

[62] 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. 172–181.

[63] V. Mnih, *Machine Learning for Aerial Image Labeling*. Toronto, ON, Canada: Univ. Toronto, 2013.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[65] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[66] M. Yang, Y. Yuan, and G. Liu, "SDUNet: Road extraction via spatial enhanced and densely connected UNet," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108549.

**Changwei Wang** received the B.S. degree in software engineering from Tiangong University, Tianjin, China, in July 2019. He is currently pursuing the Ph.D. degree in pattern recognition and intelligent systems with the Institute of Automation, Chinese Academy of Sciences, Beijing, China.

His research interests include computer vision and image processing.

**Rongtao Xu** received the B.S. degree in information and computing science from the Huazhong University of Science and Technology, Wuhan, China, in July 2019. He is currently pursuing the Ph.D. degree in pattern recognition and intelligent systems with the Institute of Automation, Chinese Academy of Sciences, Beijing, China.

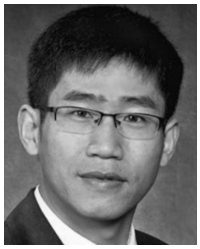
His research interests include computer vision and image processing.

**Shibiao Xu** (Member, IEEE) received the B.S. degree in information engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2009, and the Ph.D. degree in computer science from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2014.

He is currently a Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. His research interests include computer vision and image-based 3-D scene reconstruction and understanding.

**Weiliang Meng** (Member, IEEE) received the Ph.D. degree in computer application from the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China, in 2010.

He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His main research interests include artificial intelligence, computer vision, 3-D scene analysis, 3-D geometry processing, and computer graphics.



**Ruisheng Wang** (Senior Member, IEEE) received the B.Eng. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1993, the M.Sc.E. degree in geomatics engineering from the University of New Brunswick, Fredericton, NB, Canada, in 2004, and the Ph.D. degree in electrical and computer engineering from McGill University, Montreal, QC, Canada, in 2011.

In 2012, he joined the Department of Geomatics Engineering, University of Calgary, Calgary, AB, Canada, where he is currently a Professor. Prior to that, he worked as an Industrial Researcher at HERE (formerly NAVTEQ), Chicago, IL, USA, in 2008, where his primary research focus was on mobile light detection and ranging (LiDAR) data processing for next-generation map making and navigation. His research interests include geomatics and computer vision, especially point cloud processing.

**Jiguang Zhang** (Member, IEEE) received the M.S. and Ph.D. degrees in computer science from Mysore University, Mysore, India, in 2018.

He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include 3-D reconstruction, scene analysis, robotic perception, and image processing.

**Xiaopeng Zhang** (Member, IEEE) received the B.S. and M.S. degrees in mathematics from Northwest University, Xi'an, China, in 1984 and 1987, respectively, and the Ph.D. degree in computer science from the Institute of Software, Chinese Academy of Sciences, Beijing, China, in 1999.

He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His main research interests are computer graphics and computer vision.