

# Remote Sensing Road Extraction by Refining Road Topology



Huiqin Gao, Yuan Yuan, and Xiangtao Zheng

**Abstract** Remote sensing road extraction is one of the research hotspots in high-resolution remote sensing images. However, many road extraction methods cannot hold the edge interference, including shadows of sheltered trees and vehicles. In this paper, a novel remote sensing road extraction (RSRE) method based on deep learning is proposed, which considers the road topology information refinement in high-resolution image. Firstly, two parallel operations, which named dilation module (DM) and message module (MM) in this paper, are embedded in the center of semantic segmentation network to tackle the issue of incoherent edges. DM containing dilated convolutions is used to capture more context information in remote sensing images. MM consisting of slice-by-slice convolutions is used to learn the spatial relations and the continuous prior of the road efficiently. Secondly, a new loss function is designed by combining dice coefficient term and binary cross-entropy term, which can leverage the effects of different loss. Finally, extensive experimental results demonstrate that the RSRE outperforms the state-of-the-art methods in two public datasets.

**Keywords** High resolution · Road extraction · Deep learning · Feature fusion

## 1 Introduction

Road extraction in high-resolution remote sensing image aims at detecting and segmenting road pixels in images. It refers to judging pixels as road or non-road, usually regarded as a binary classification problem. Road is an integral part of vehicle navigation, city planning, geographic and information updating, and so on. At present,

---

H. Gao · X. Zheng (✉)

Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China

e-mail: [zhengxiangtao@opt.cn](mailto:zhengxiangtao@opt.cn)

H. Gao

University of Chinese Academy of Sciences, Beijing 100049, China

Y. Yuan

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China

© Springer Nature Singapore Pte Ltd. 2020

L. Wang et al. (eds.), *Proceedings of the 6th China High Resolution Earth Observation Conference (CHREOC 2019)*, Lecture Notes in Electrical Engineering 657,

[https://doi.org/10.1007/978-981-15-3947-3\\_14](https://doi.org/10.1007/978-981-15-3947-3_14)

the task of road extraction mainly contains road surface detection [1] and road centerline extraction [2, 3]. The former extracts all road pixels out, while the latter only label the skeletons of roads, which used to provide directions. Some methods also extract surface and centerline of road simultaneously [4].

In the field of high-resolution remote sensing road extraction, numerous methods have been proposed in recent years. Early methods extract low-level features (*e.g.*, edge, corner, gradients) and define heuristic rules (*e.g.*, geometrical shape) to classify pixels into road or non-road. S. Hinz and A. Baumgartner combined road and context information to extract road, including radiation measurements and geometry information [5]. In [6, 7], uniform areas with shape or geometric characteristics in the image were detected firstly and then used region growth technique to generate road map. The problem of these methods is that the features and rules used are only for simple scenes, while roads in the high-resolution remote sensing images are complex and irregular.

Several methods applied deep learning to road extraction in high-resolution remote sensing image. Y. Zhang et al. used multi-source data and multi feature to improve accuracy of road extraction [8]. Y. Wei et al. designed a road structural loss function to constrain road edge [9]. G. Mátyus et al. inferred the correct road in the result of initial segmentation [10]. F. Bastani et al. designed an iterative graph construction method to output road map [11]. Recent methods adopted the idea of semantic segmentation [12–15] and take roads as foreground and non-roads as background. Z. Zhang et al. combined the strengths of residual learning and U-Net [16] to extract road [17]. Y. Xu et al. fused attention mechanisms in DenseNet to capture local and global road information simultaneously [18]. L. Zhou et al. used larger receptive field to preserve detailed information, thus obtained a better result of road extraction [19]. Although methods based on deep learning made some progress, it still had incoherent issues of road edge. The issues mainly caused by edge interference, including shadows of roadside trees or buildings and vehicles on the roads, which can be observed in high-resolution remote sensing images. To solve the issues, a novel high-resolution remote sensing road extraction (RSRE) method is proposed to refine road topology information. In addition to increasing the receptive field to keep context information, RSRE also considers the spatial relations of road pixels of an image. The spatial relations of pixels in image contribute to learning topology information of road with weak coherence in high-resolution remote sensing image. Therefore, RSRE can alleviate the incoherent issues which occurred in many existed methods based on deep learning.

In this paper, RSRE focuses on road topology information refinement in high-resolution remote sensing image. Topology information refinement means the maintenance of the shape, structure, or connection of the road throughout the whole image. Based on the encoder–decoder architecture which often used in semantic segmentation network, RSRE adopts dilation module (DM) and message module (MM) between encoder and decoder to enhance connectivity of road edge. Dilated convolutions in DM can increase the receptive field to keep the detailed context information in image, while slice-by-slice convolutions in MM enable messages passing across rows and columns in image to capture spatial relations of pixels. After extracting

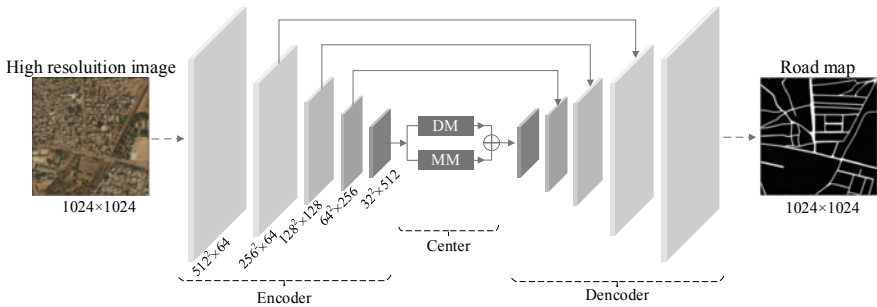
the features of image by encoder, DM and MM are used to reprocess the features between encoder and decoder. Finally, RSRE uses sigmoid layer and threshold value to output road maps. Furthermore, a new loss function is proposed to make RSRE not favor the non-road which has most of pixels in image. Experimental results show that RSRE is excellent both on DeepGlobe Road dataset [20] and Massachusetts Road dataset [21].

## 2 Method

### 2.1 RSRE Architecture

Due to the initial high-resolution remote sensing image has large size, and road always span the whole image with natural properties like topology and connectivity. Therefore, RSRE receives  $1024 \times 1024$  high-resolution image as input to reduce the loss of detail caused by cropping images and generates road map with road topology information refinement and better road connectivity recovery. As shown in Fig. 1, RSRE has an encoder–decoder structure and combines low-level detail information and high-level semantic information to extract road in high-resolution image.

There are three parts in RSRE: encoder, center, and decoder parts. Like the architecture of D-LinkNet, the encoder part extracts feature maps of input high-resolution remote sensing image and uses ResNet34 [22] pretrained on ImageNet [23] dataset. The center of RSRE fuses the results of feature reprocessing by DM and MM to keep topology information of road. The decoder part uses transposed convolution layers [24] to do up-sampling and restores the resolution from  $32 \times 32$  to  $1024 \times 1024$ . Finally, RSRE uses sigmoid layer and threshold value to output road maps. Pixels which probability of sigmoid layer output larger than 0.5 are considered roads, while others are considered as non-roads.



**Fig. 1** RSRE architecture. Symmetrical blocks represent features with the same size and channels. The below expression  $n^2 \times c$  means that the size is  $n \times n$ , and the number of channels is  $c$ . RSRE has an encoder–decoder structure. Center part is the core of RSRE, including DM and MM

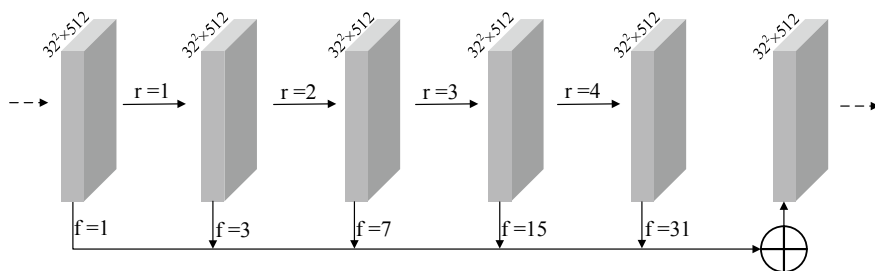
## 2.2 DM and MM

The center part can refine the topology structure of road in the high-resolution remote sensing image. The high-dimensional hidden layer features are selected as the input of center part because of rich information. This part is composed of two parallel operations: DM and MM. DM increases receptive field without reducing the image resolution through dilated convolution layers [25] with series and parallel connections [25]. As shown in Fig. 2, DM stacks result of each dilation rate, which contributes to capturing multi-scale context.

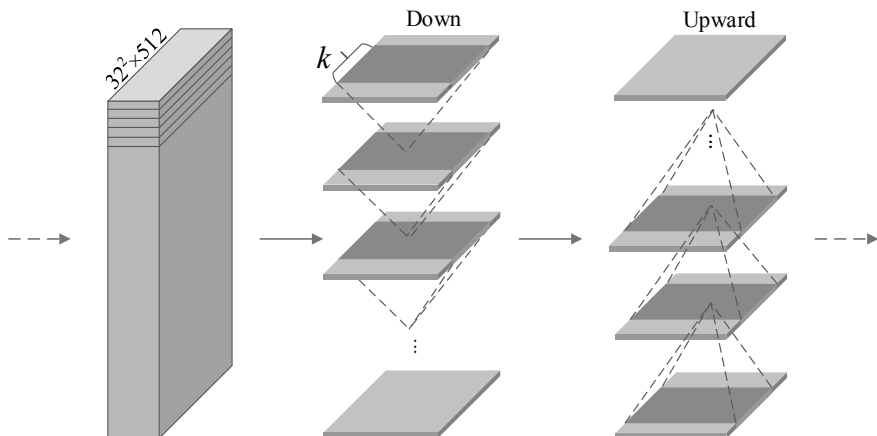
Although DM contributes to obtain multi-scale context information by increasing receptive field, it has an issue of lacking correlation information between pixels. Due to the values of dilated convolution is obtained from the pixels of mutually independent rows and columns, and these pixels lack correlation on each other. The issue causes loss of local information might be not relevant and thus bring poor continuity but is critical to the road, which has long-distance continuous, and strong spatial relation while weak appearance clue in the high-resolution remote sensing image.

In order to solve the potential issue of DM, MM uses the module of spatial CNN [26] in the field of computer vision to enhance spatial relations of pixels in high-resolution remote sensing image. Though slice-by-slice convolutions within feature maps, it can better propagate spatial information of pixels on rows and columns and thus can effectively preserve the topology information of road with long thin structure in high-resolution image.

As shown in Fig. 3, MM also applies on high-dimensional hidden layer features. The height, width, and the number of channel of the input feature are 32, 32, 512. MM has four directions to slice: upward, down, left, and right. Only the down and upward directions are shown in figure, and the left and right directions are similar. In each direction, the feature is sliced along height (upward and down) or width (left and right) of feature. The first slice goes through convolution and rectified linear unit (ReLU) and then adds next slice to formulate a new slice. The new slice repeats the same processing sequentially until the last slice is updated and thus obtains a new



**Fig. 2** DM architecture. DM contains dilated convolution with series and parallel connections. The expression  $n^2 \times c$  on the feature block means that the size is  $n \times n$ , and the number of channels is  $c$ . The parameters  $r$  represents the dilation rate, and  $f$  means the receptive field



**Fig. 3** MM Architecture. MM contains slice-by-slice convolutions within features. Expression  $n^2 \times c$  on the feature means that the size is  $n \times n$ , and the number of channels is  $c$ . Parameter  $k$  is kernel width used in MM

feature of size  $32 \times 32 \times 512$ . Weights of the slice-by-slice convolutions are shared with the same direction and are initialized randomly different with spatial CNN.

### 2.3 Loss Function

Although roads are distributed in the entire high-resolution image, the imbalance in road pixels and non-road pixels in an image has the potential to tilt the training result toward non-road class with more pixels. Therefore, a new loss function is used in this paper, which uses dice coefficient item (1) and binary cross-entropy (2) simultaneously.

$$l_{\text{dice}} = 1 - \frac{\sum_{n=1}^N y_n f_w(x_n) + m}{\sum_{n=1}^N y_n + f_w(x_n) + m}. \quad (1)$$

$$l_{\text{bce}} = -\frac{1}{N} \sum_{n=1}^N [y_n \cdot \log f_w(x_n) + (1 - y_n) \cdot \log(1 - f_w(x_n))]. \quad (2)$$

$x_n$  means  $n$ th high-resolution remote sensing image, where  $n = 1, 2, 3, \dots, N$  and  $N$  is the mini-batch size.  $y_n$  indicates the ground truth (GT) of image  $x_n$ , and GT is a binary map. Expression  $f_w(x_n)$  denotes the output of RSRE, where  $w$  represents the weight of RSRE that needs to be optimized. Parameter  $m$  is an adjustable parameter between numerator and denominator in  $l_{\text{dice}}$ .

Loss  $l_{\text{dice}}$  can be regarded as the degree of similarity of road contours between GT and predicted  $f_w(x_n)$ . Loss  $l_{\text{bce}}$  is often used in the training of semantic segmentation

network. However, in the case of extremely unbalanced data, the cross-entropy loss is much smaller than the dice loss after multiple iterations, and the effect of  $l_{\text{bce}}$  will be lost. Road extraction can be regarded as a task of pixel-level recognition, only pixels of road can view as positive sample. Clearly, there is great imbalance in road pixels and non-roads pixels. To release the question of imbalance, RSRE combines  $l_{\text{dice}}$  and  $l_{\text{bce}}$  with  $\lambda$  and adjusts  $m$  to control the effects of different losses on training results. Thus, the final loss function is the form in (3):

$$\text{loss}_w = l_{\text{dice}} + \lambda l_{\text{bce}}. \quad (3)$$

Parameter  $w$  is the weight of RSRE that needs to be updated.  $\lambda$  is a constant coefficient by manual setting. Though minimizing the loss function, the optimal  $w$  is obtained gradually. RSRE chooses Adam as optimizer to optimize the loss function.

### 3 Experiment

#### 3.1 Datasets

The method is tested on two big datasets. The first is DeepGlobe Road dataset. The resolution of each image is  $1024 \times 1024$ . Image scenes include urban, rural, wilderness, seaside, tropical rainforest, and others. On account of only training images have labels, for the convenience of measuring the accuracy of road extraction, the experiment divides the labeled 6226 training image into 4358 for training and 1868 for test.

The second dataset is Massachusetts Road dataset. The size is  $1500 \times 1500$  with a resolution of 1.2 m per pixel. In the original training data, there are some images that do not match labels; because in the work of the original thesis, the dataset is used to study the effectiveness of robustness. In this paper, first delete the mismatch image and GT pairs, then crop to size  $1024 \times 1024$  from center, and train RSRE on the rest of 737 images and test in 49 images.

#### 3.2 Implementation Details

In the experiments, PyTorch [27] is used as the deep learning framework. In the training phase, the min-batch size is 16 and uses 2 GPUs. The learning rate was initially set to  $2e-4$  and reduced by a factor of 0.1 in every 20 epochs. RSRE adopts data augmentation to avoid the problem of over-fitting without cross-validation, including flip, vertical flip, diagonal flip, color jittering, image shifting, and scaling. In the predicting phase, each image has operations of horizontal flip, vertical flip, and

**Table 1** Results of RSRE compared with other methods in two datasets, and the best values of precision ( $P$ ), recall ( $R$ ), and  $F_1$ -score have been highlighted in bold

Method	DeepGlobe Road dataset			Massachusetts Road dataset		
	$P$	$R$	$F_1$ -score	$P$	$R$	$F_1$ -score
U-Net	0.730	0.768	0.748	0.613	0.760	0.679
LinkNet	0.753	0.734	0.743	0.681	0.750	0.714
D-LinkNet	0.776	<b>0.829</b>	0.802	0.737	<b>0.816</b>	0.774
RSRE	<b>0.793</b>	0.821	<b>0.807</b>	<b>0.755</b>	0.810	<b>0.781</b>

diagonal flip. Hence, each image is predicted 8 times and averaged the probability of each prediction.

### 3.3 Result and Analysis

To assess the effectiveness of RSRE in road extraction with high-resolution remote sensing image, the precision ( $P$ ), recall ( $R$ ) [28], and  $F_1$ -score are introduced as follows:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, F_1 = 2 \times \frac{P \times R}{P + R}. \quad (4)$$

where TP, FP, TN, FN represent the number of true positives, false positives, false negatives, and false negatives.  $P$  means the percentage of all road pixels in image that is predicted correctly.  $R$  represents the proportion of all pixel predicted correctly that detected as road.  $F_1$ -score is an evaluation metric for the harmonic mean between  $P$  and  $R$ .

RSRE is compared with U-Net, LinkNet, D-LinkNet in two datasets and evaluated  $P$ ,  $R$ ,  $F_1$ -score of road in Table 1 The architecture of U-Net and LinkNet is modified to fit the high-resolution remote sensing image input of  $1024 \times 1024$ , and they use cross-entropy as loss function. As baseline of RSRE, D-LinkNet only uses DM, but has no MM and constrained loss. The best result of RSRE is obtained when  $m = 0.5$ ,  $\lambda = 0.01$ . The results show that after the center feature fusion and loss constrained, the  $P$  and  $F_1$ -score of road extraction have increased. The  $R$  drops a little, because precision and recall are a pair of contradictory measures, when precision is high, the recall is often low.

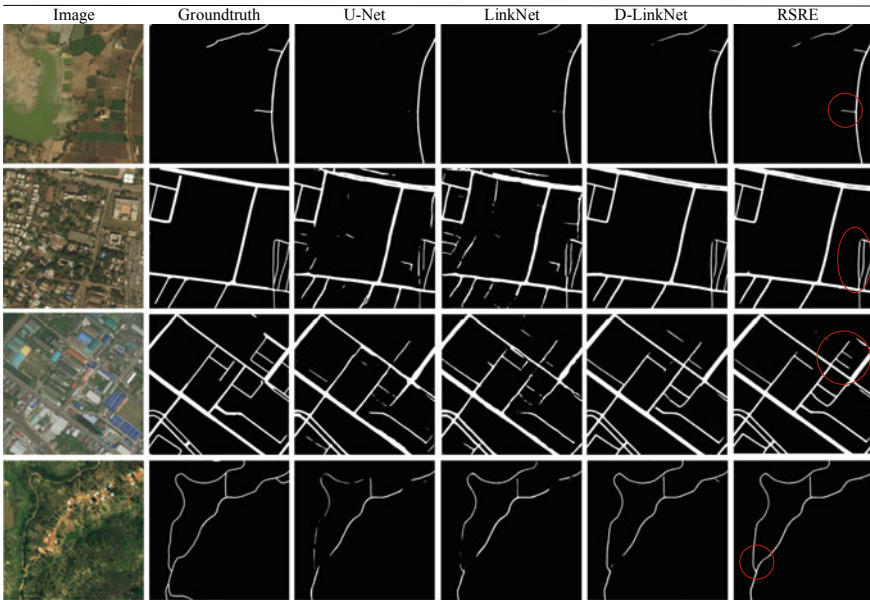
The effect of different weight combinations tested in Massachusetts Road dataset is shown in Table 2. As  $\lambda$  decreases, the influence of  $l_{bce}$  decreases gradually, and both  $P$  and  $F_1$ -score improve. It shows that the constraint binary cross-entropy can alleviate the poor performance caused by sample imbalance.  $m$  has the best adjustment effect on  $l_{dice}$  at 0.5. Therefore, RSRE gets the best result at  $m = 0.5$ ,  $\lambda = 0.01$ .

**Table 2** Results of RSRN with different combinations of  $m, \lambda$  in Massachusetts Road dataset

$m, \lambda$	$P$	$R$	$F_1$ -score
0.0, 1.00	0.739	<b>0.815</b>	0.777
0.0, 0.50	0.741	0.816	0.776
<b>0.5, 0.01</b>	<b>0.755</b>	0.810	<b>0.781</b>
1.0, 0.01	0.753	0.808	0.780

In order to observe ability of RSRE proposed in this paper, some typical images tested in DeepGlobe Road dataset are shown in Fig. 4. The test results of RSRE at different background including river, building roof, urban, and rural. The road maps of comparison demonstrate that RSRE can maintain road topology information effectively in the presence of disturbances. The results in D-LinkNet are superior to U-Net and LinkNet because of the use of DM, which proves that it is necessary to increase the receptive field. Though river in image has approximate linear edge, building roof may block part of the road edge or pavement and thus destroy the continuity of the road. The connectivity of roads in a red circle of figure can be well kept through RSRE than other methods.

Results of RSRE compared with other methods in Massachusetts Road dataset are shown in Fig. 5. Compare with the results of U-Net, LinkNet, and D-LinkNet, RSRE can detect most of the roads correctly with the road topology information refinement. Though increase the receptive field and enhance spatial relations by adding DM and



**Fig. 4** Example results of RSRE and other methods tested in DeepGlobe Road dataset. From top to bottom, the contained background is river, building roof, urban, and rural



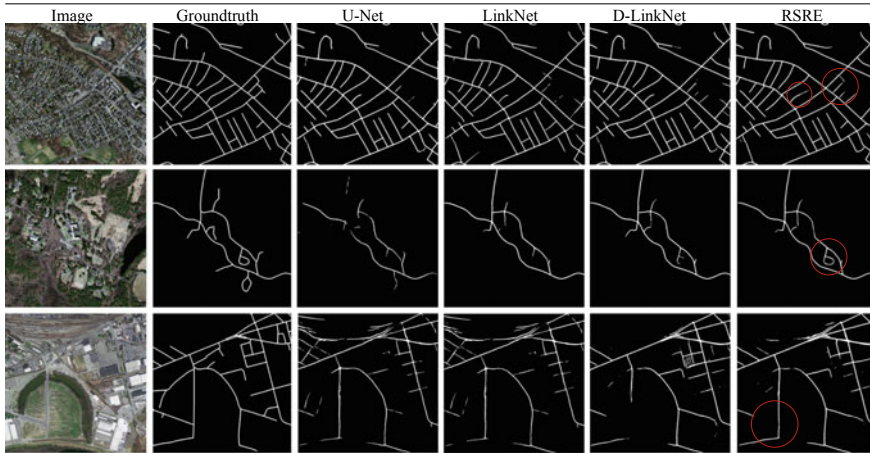


Fig. 5 Example results of RSRE and other methods tested in Massachusetts Road dataset

MM in the center of deep learning network, RSRE keeps rich information of road to refine topology in high-resolution remote sensing image. Therefore, in the phase of generating road map, the context and spatial relations information of road obtained in RSRE alleviate effectively the incoherent issues caused by shadows of sheltered trees and vehicles.

### 4 Summary

In this paper, RSRE extracts road from high-resolution remote sensing image, which pays more attention to refine road topology information. By the feature fusion processing of DM and MM, it can refine topology information of road and thus effectively preserve the continuity of long thin structure of road. The new combined loss function can solve the imbalance of road and non-road pixels. The results in two datasets show that RSRE could alleviate the discontinuity and incoherence of road which come from edge interference. It also has a good performance in different backgrounds. However, the test results of all methods are not very accurate for very short and thin sections of road, and RSRE has the wrong recognition in a heavily sheltered area, Therefore, the future work will focus on how to solve the above issues.

## References

1. Mnih V, Hinton GE (2010) Learning to detect roads in high-resolution aerial images. In: Proceedings of the 11th European conference on computer vision, pp 210–223
2. Gao L, Shi W, Miao Z, Lv Z (2018) Method based on edge constraint and fast marching for road centerline extraction from very high-resolution remote sensing images. *Remote Sens* 10:900
3. Zhang Z, Zhang X, Sun Y, Zhang P (2018) Road centerline extraction from very-high-resolution aerial image and LiDAR data based on road connectivity. *Remote Sens.* 10:1284
4. Cheng G, Wang Y, Xu S, Wang H, Xiang S, Pan C (2017) Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. *IEEE Trans Geosci Remote Sens* 55:3322–3337
5. Hinz S, Baumgartner A (2003) Automatic extraction of urban road networks from multi-view aerial imager. *ISPRS-J Photogramm Remote Sens* 58:83–98
6. Hu J, Razdan A, Femiani JC, Cui M, Wonka P (2007) Road network extraction and intersection detection from aerial images by tracking road footprints. *IEEE Trans Geosci Remote Sens* 45:4144–4157
7. Song M, Civco D (2004) Road extraction using SVM and image segmentation. *Photogramm Eng Remote Sens* 70:1365–1371
8. Zhang Y, Xia W, Zhang Y, Sun S, Sang L (2018) Road extraction from multi-source high-resolution remote sensing image using convolutional neural network. In: Proceedings of IEEE 2018 international conference on audio, language and image processing (ICALIP), pp 201–204
9. Wei Y, Wang Z, Xu M (2017) Road structure refined CNN for road extraction in aerial image. *IEEE Geosci Remote Sens Lett* 14:709–713
10. Mátyus G, Luo W, Urtasun R (2017) Deeproadmapper: extracting road topology from aerial images. In: Proceedings of IEEE international conference on computer vision, pp 3438–3446
11. Bastani F, He S, Abbar S, Alizadeh M, Balakrishnan H, Chawla S, Madden S, DeWitt D (2018) RoadTracer: automatic extraction of road networks from aerial images. In: IEEE conference on computer vision and pattern recognition, pp 4720–4728
12. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Proceedings of international conference on medical imaging and computer-assisted intervention, pp 234–241
13. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell* 39:2481–2495
14. Chaurasia A, Culurciello E (2017) Linknet: exploiting encoder representations for efficient semantic segmentation. In: Proceedings of IEEE visual communications and image processing, pp 1–4
15. Chen L, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision, pp 801–818
16. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of international conference on medical imaging and computer-assisted intervention, pp 234–241
17. Zhang Z, Liu Q, Wang Y (2018) Road extraction by deep residual u-net. *IEEE Geosci Remote Sens Lett* 15:749–753
18. Xu Y, Xie Z, Feng Y, Chen Z (2018) Road extraction from high-resolution remote sensing imagery using deep learning. *Remote Sens* 10:1461
19. Zhou L, Zhang C, Wu M (2018) D-Linknet: linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In: IEEE conference on computer vision and pattern recognition. Workshops, pp 182–186
20. Demir I, Koperski K, Lindenbaum D, Pang G, Huang J, Basu S, Hughes F, Tuia D, Raskar R (2018) Deepglobe 2018: a challenge to parse the earth through satellite images. In: IEEE/CVF conference on computer vision and pattern recognition. Workshops, pp 172–179
21. Mnih V (2013) Machine learning for aerial image labeling. University of Toronto, Canada

22. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 770–778
23. Deng J, Dong W, Socher R, Li L, Li K, Li F (2009) Imagenet: a large-scale hierarchical image database. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 248–255
24. Zeiler MD, Taylor GW, Fergus R (2011) Adaptive deconvolutional networks for mid and high level feature learning. In: Proceedings of the IEEE conference on computer vision, pp 2018–2025
25. Yu F, Koltun V (2016) Multi-scale context aggregation by dilated convolutions. In: Proceedings of international conference on learning representations
26. Pan X, Shi J, Luo P (2018) Spatial as deep: spatial CNN for traffic scene understanding. In: Proceedings of AAAI conference on artificial intelligence, pp 7276–7283
27. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A (2017) Automatic differentiation in PyTorch
28. Martin DR, Fowlkes CC, Malik J (2004) Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans Pattern Anal Mach Intell* 26:530–549