Road Extraction From Satellite Imagery by Road Context and Full-Stage Feature

Zhigang Yang, Daoxiang Zhou, Ying Yang, Jiapeng Zhang, and Zehua Chen

***Abstract*— Road extraction from satellite imagery is vital in a broad range of applications. However, extracting complete roads is challenging due to road occlusions caused by the surroundings. This letter proposed an improved encoder–decoder network via extracting road context and integrating full-stage features from satellite imagery, dubbed as RCFSNet. A multiscale context extraction (MSCE) module is designed to enhance inference capabilities by introducing adequate road context. Multiple full-stage feature fusion (FSFF) modules in the skip connection are devised to provide accurate road structure information, and we devise a coordinate dual-attention mechanism (CDAM) to strengthen the representation of road features. Extensive experiments are carried out on two public datasets, and as a result, our RCFSNet outperforms other state-of-the-art methods. The results indicate that the road labels extracted by our method have preferable connectivity. The source code will be available at https://github.com/CVer-Yang/RCFSNet.**

***Index Terms*— Attention mechanism, full-stage feature fusion (FSFF), road context, road extraction.**

I. Introduction

ROAD extraction from satellite imagery is of great significance in plentiful fields, such as urban planning, road monitoring, autonomous driving, and so on. Deep learning has gained remarkable performance in image processing of remote sensing, such as subpixel mapping [1], [2] [3], target segmentation [4], and change detection [5], [6]. With the boost of convolution neural networks (CNNs), road extraction methods based on CNNs have attracted increasing attention.

However, extracting accurate road labels from satellite imagery is still challenging. First, satellite imagery contains abundant ground objects, and road pixels only account for a small part of remote sensing imagery. In addition, some roads in the imagery are blocked by buildings or trees, which causes road labels predicted by the model to be disconnected.

Substantial works combined with the deep-learning model have provided high-quality road extraction results in the past years. Mnih and Hinton [7] used restricted Boltzmann machines to extract roads from aerial imagery, realizing road extraction based on deep-learning for the first time. Zhou et al. [8] combined the LinkNet with dilated convolution, which aims to provide the characteristics of high-level feature maps in different receptive fields. Zhu et al. [9] improved the integrity of the forecast road by capturing the road global context. Wang et al. [10] introduced nonlocal structure to grasp the long-distance relationship of roads. Chen et al. [11] designed a dual-branch encoder to extract multiscale image features. Lu et al. [12] captured long-distance dependencies from the spatial and channel dimensions, which enables the network to overcome the limitation of local receptive field. Zhou et al. [13] adopted a graph convolution network to obtain global information on road space and channel features, and the method has the highest accuracy in complex backgrounds compared with competitive models. Li et al. [14] combined UNet3+ [15] with channel mechanism to realign channel-wise feature maps adaptively. Lu et al. [16] provided a cascaded multitask extraction network to solve the problem of producing roads with poor connectivity.

The attention mechanism has been widely applied in various computer vision tasks. Hu et al. [17] designed an effective Squeeze-and-Excitation (SE) Network, which can preserve the key features of channels. Huang et al. [18] designed a Criss-Cross network to solve the computationally expensive problem of nonlocal structure while capturing the dependence between long-range pixels. Hou et al. [19] designed a channel attention mechanism combined with spatial information to achieve accurate target localization. Recently, various visual transformer networks have appeared in many vision tasks, showing the power of self-attention.

Capturing multiscale context is required for road extraction tasks. There is a common method to capture the long-range dependencies by using dilated convolution. However, using ordinary $N \times N$ convolution kernels to model long and narrow road features always introduces much unrelated context. Moreover, some current existing works equipped with attention mechanisms do not consider road shapes.

In this letter, we propose a novel road extraction method from satellite imagery named RCFSNet via extracting road context and integrating full-stage features. The multiscale context extraction (MSCE) module and full-stage feature fusion (FSFF) module are designed to improve the segmentation quality of the model on the road area.

The main contributions of this work are summarized as follows.

1) We propose the MSCE module to capture the long-range dependencies, which can efficiently capture road context over long distances.

Manuscript received 15 October 2022; revised 27 November 2022; accepted 8 December 2022. Date of publication 12 December 2022; date of current version 6 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62101376, in part by the Natural Science Foundation of Shanxi Province of China under Grant 201901D211078, and in part by the Shanxi Transportation and Control Technology Research Project under Grant 19-JKKJ-2. *(Corresponding author: Zehua Chen.)*

Zhigang Yang, Daoxiang Zhou, and Zehua Chen are with the College of Data Science, Taiyuan University of Technology, Taiyuan 030024, China (e-mail: chenzehua@tyut.edu.cn).

Ying Yang and Jiapeng Zhang are with Shanxi Transportation Technology Research and Development Company Ltd., Taiyuan 030024, China.

Digital Object Identifier 10.1109/LGRS.2022.3228967

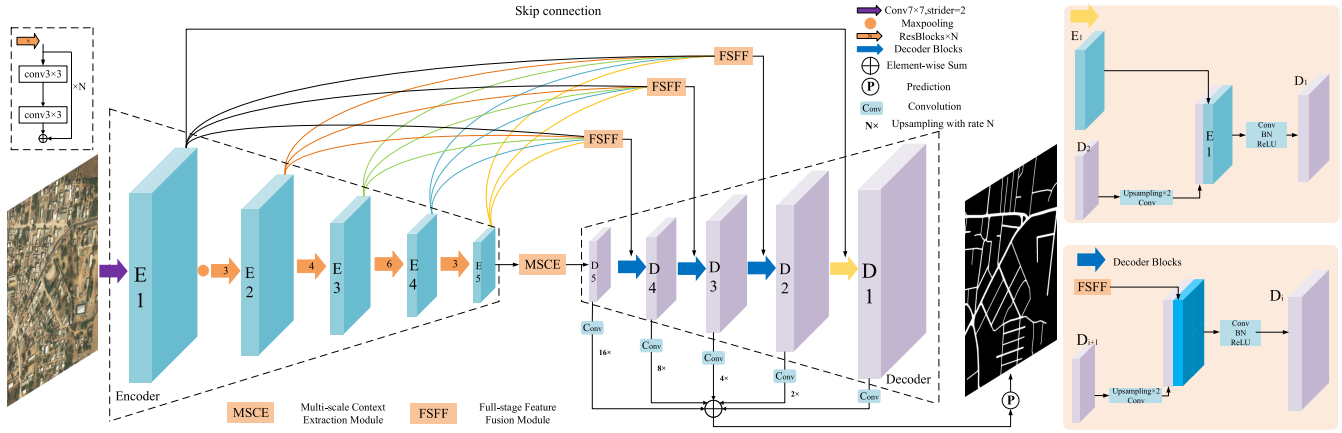

Fig. 1. Basic structure of the proposed RCFSNet.

2) Multiple FSFF in skip connection that integrates features in different stages are constructed, which can supply accurate road structure information to solve the problem of broken lines in predicted results. Meanwhile, we devise a coordinate dual attention mechanism (CDAM) in FSFF to strengthen the channel and spatial features of roads.
3) We conduct extensive experiments on two public datasets, which show RCFSNet outperforms other state-of-the-art methods.

II. Proposed Methods

In this section, we design a road extraction method via extracting road context and integrating full-stage features, called RCFSNet in Fig. 1, which is composed of four modules: 1) feature encoder; 2) MSCE; 3) FSFF; and 4) decoder.

A. Feature Encoder

The ResNet34 [20] network pretrained on the ImageNet dataset is used as the encoder for image feature extraction. The preliminary features E1 are obtained through convolution with a kernel size of 7 and stride of 2. Afterward, the feature extraction of the image is achieved through Maxpooling and multiple residual blocks in multiples of 3, 4, 6, and 3.

B. Multiscale Context Extraction Module

There are plenty of roads that are occluded by surroundings in remote-sensing images. Capturing the relationships between roads and surroundings can enrich the overall characteristics of roads, which is of great help in improving road connectivity. Unlike common objects, the road has strong connectivity and wide coverage naturally, and it often covers the entire image. To tackle the issues, we design an MSCE module to introduce rich road context.

As shown in Fig. 2, MSCE leverages three branches made of 3×3 convolution, 3×1 convolution (horizontal kernel), and 1×3 convolution (vertical kernel), with dilation rates of $\{1, 2, 4\}$, and two branches made of horizontal pooling and vertical pooling. The convolution branches extract road features of different scales, and the pooling branches can preserve the global information of the road in vertical and horizontal directions. We utilize an element-wise addition operation to fuse the feature maps at different receptive fields in the same branch. Then, concatenation and convolution operations combine the feature maps output from different branches. Finally, E5 is combined with the fusion feature maps to obtain feature map D5 with abundant road context, where p is a learnable parameter.

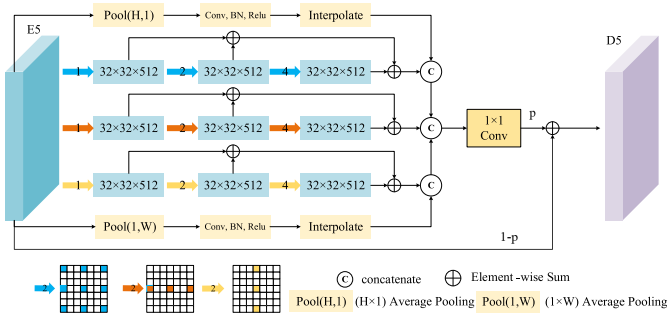

Fig. 2. Structure of MSCE.

C. FSFF Module

The encoder feature maps at different stages contain different levels of information. The low-level feature maps contain rich spatial information, which can provide the overall structure information of roads; the high-level feature maps contain accurate road semantic information, which can enhance the model's ability to distinguish the road and background. Combining feature maps at different stages can effectively supplement the decoder with sufficient road hierarchical features, making the road boundaries predicted by the model clearer. Meanwhile, we design a novel CDAM to facilitate FSFF.

The FSFF module is shown in Fig. 3, taking the E3 as an example. First, the fine-grained feature maps (E5 and E4), and the coarse-grained feature maps (E1 and E2) are adjusted to the same size as the E3, and the number of channels is adjusted to 64 through convolution. Then, the resized feature maps are transmitted by the concatenation operation. Finally, the acquired merged feature map is fed into the CDAM to strengthen road feature representation. The visualization result is illustrated in Fig. 4.

The coordinate channel attention mechanism (CCAM) is shown in Fig. 5. The fused feature map $F_I \in R^{C \times H \times W}$ is squeezed by strip pooling layers with shapes $(H, 1)$ and $(1, W)$

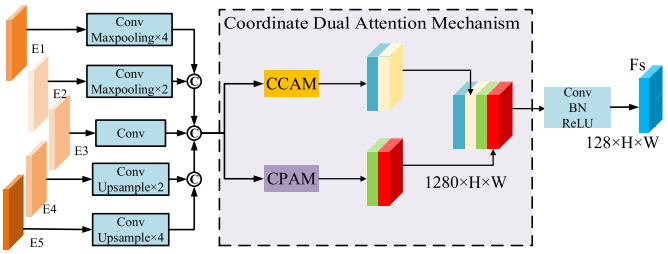

Fig. 3. Structure of FSFF.

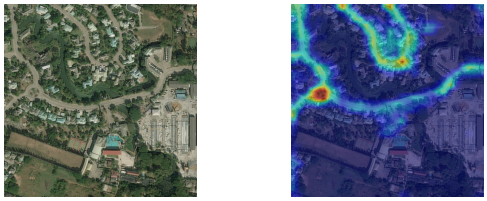

Fig. 4. We use Grad-CAM [21] as our visualization tool. It is obvious that our CDAM can precisely locate the road.

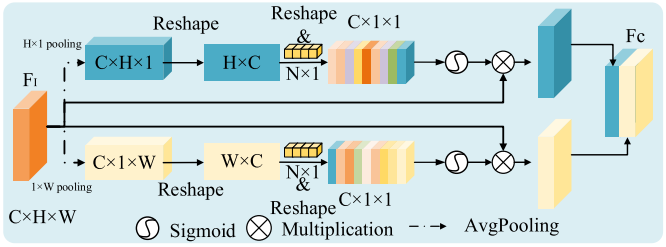

Fig. 5. Structure of CCAM.

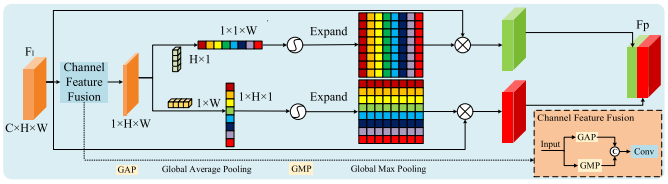

Fig. 6. Structure of CPAM.

simultaneously, and the reshape operation is used to convert the feature maps into $F_H \in R^{H\times C}$ and $F_W \in R^{W\times C}$. We adopt 1-D convolution with one filter to obtain cross-channel interaction of feature maps. The sigmoid activation is used to obtain the channel weights combined with horizontal and vertical features, respectively. The input feature map combined with the channel weight is delivered by concatenation operation to obtain the feature map $F_C \in R^{2C\times H\times W}$.

The coordinate position attention mechanism (CPAM) is shown in Fig. 6. Avgpooling, maxpooling, and convolution are used in the channel dimension of the input feature map $F_I \in R^{1\times H\times W}$ to aggregate the features. Then, the convolution with kernel sizes of $(H, 1)$ and $(1, W)$ is used to extract the features in the horizontal and vertical directions. The sigmoid activation function is used to obtain the position weight of the feature map. The input feature map weighted by the position weight is sent into concatenation to get the feature map $F_P \in R^{2C\times H\times W}$.

The obtained feature maps F_C and F_P are spliced and then input into 1×1 convolution to generate the precise road structure feature map $F_S \in R^{C\times H\times W}$.

D. Feature Decoder

Each decoder takes the outputs from the FSFF and the previous decoder. Feature maps of the previous stage adjust the size and channel's number through up-sampling operation and convolution. Then it is integrated with the supplementary road structure information by concatenation and convolution, batch normalization (BN), and rectified linear unit (ReLU) to generate the feature map of this decoder stage.

In prediction, D5, D4, D3, and D2 are adjusted to the equivalent resolution and channels as D1 by using up-sampling operation and convolution. These feature maps are fused to obtain the $D_{\text{out}} \in R^{64\times 512\times 512}$. The merged feature map D_{out} is fed into up-sampling operation and 3×3 convolution to obtain the final predicted semantic segmentation result $F_{\text{pred}} \in R^{1\times 1024\times 1024}$.

III. Experiments

A. Dataset

We conducted experiments on the DeepGlobe [22] road dataset and the Massachusetts [23] road dataset. The DeepGlobe road dataset covers multiple scenarios, and it is randomly divided into 5500 pairs of images as training and 726 pairs as testing. The Massachusetts road dataset collected by Mnih contains 1171 pairs of remote sensing images. Due to some images being occluded, we selected road images and cropped them into 1024×1024 from training with 1108 pairs of images as training and 49 pairs as testing.

B. Implementation Details

The combination of Binary CrossEntropy (BCEloss) and Dice coefficient is used as our loss function, Adam is used as the optimizer, and the training batch size is set to 8 [(two images for each graphic processing unit (GPU)]. The initial learning rate is set to 2e-4. If the training loss does not decrease five times, the model will end the training. The number of convolution kernels in the CCAM is set as 5 (E4), 7 (E3), and 9 (E2). The test enhancement techniques are used in the prediction stage. All experiments are implemented on 4 NVIDIA V100 with 16 GB of memory. The precision, recall, intersection over union (IoU), and F1-score are adopted as evaluation metrics.

C. Comparative Analysis and Visualization

As it can be seen in Fig. 7, RCFSNet demonstrates manifest advantages compared with several state-of-the-art models under all thresholds, and we set the threshold as 0.3.

As shown yellow-black in Fig. 8, our RCFSNet performs well when trees and buildings block the road, and road boundary labels generated by RCFSNet have stronger brightness. The first and third rows of Fig. 8 show the scene where surroundings occlude the road. Due to the lack of context, U-Net and DBRANet cannot effectively handle the situation where the road is blocked in the image. DeepLabv3+ and DLinkNet introduce context by ordinary dilated convolution,

TABLE I

ROAD EXTRACTION RESULTS ON THE DEEPGLOBE ROAD DATASET AND THE MASSACHUSETTS ROAD DATASET

	DeepGlobe				Massachuseets			
Method	Recall	Precision	IoU	F1-score	Recall	Precision	IoU	F1-score
U-Net[24]	0.8357	0.7322	0.6362	0.7676	0.8088	0.7401	0.6315	0.7705
DeepLabV3+[25]	0.8474	0.7268	0.6373	0.7571	0.7634	0.7588	0.6129	0.7536
DLinkNet[8]	0.8307	0.7901	0.6758	0.7955	0.8016	0.7744	0.6490	0.7821
NLinkNet[10]	0.8177	0.8092	0.6858	0.8002	0.8001	0.7727	0.6467	0.7805
DBRANet[11]	0.8172	**0.8106**	0.6837	0.8001	0.8013	0.7747	0.6490	0.7821
MACU-Net[14]	0.8055	0.7967	0.6657	0.7894	**0.8141**	0.7593	0.6469	0.7827
RCFSNet	**0.8546**	0.7898	**0.6934**	**0.8101**	0.8046	**0.7825**	**0.6552**	**0.7866**

TABLE II

ABLATION RESULTS ON THE DEEPGLOBE ROAD DATASET AND THE MASSACHUSETTS ROAD DATASET

	Components			DeepGlobe				Massachuseets			
Method	Baselline	MSCE	FSFF	Recall	Precision	IoU	F1-score	Recall	Precision	IoU	F1-score
Model1	✓			0.8141	0.7916	0.6712	0.7901	0.8040	0.7689	0.6473	0.7808
Model2	✓		✓	0.8360	**0.7986**	0.6898	0.8048	0.8043	0.7786	0.6536	0.7857
Model3	✓	✓		**0.8593**	0.7778	0.6857	0.8047	**0.8063**	0.7747	0.6515	0.7841
RCFSNet	✓	✓	✓	0.8546	0.7898	**0.6934**	**0.8101**	0.8046	**0.7825**	**0.6552**	**0.7866**

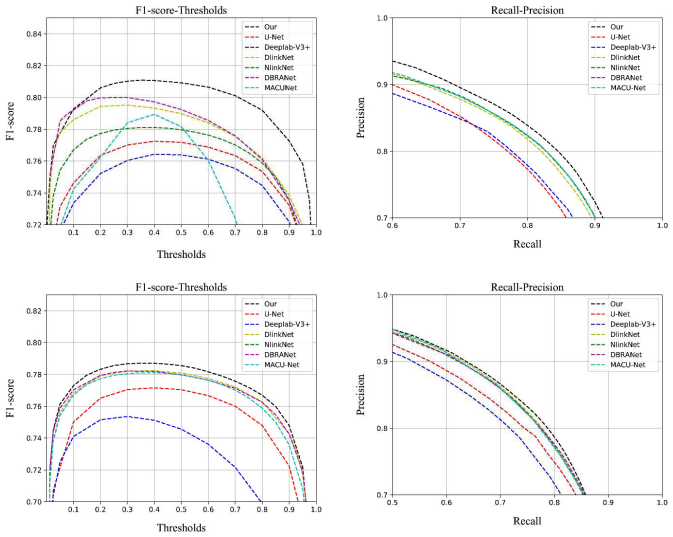

Fig. 7. Accuracy curves in different thresholds on the (first row) DeepGlobe road dataset and the (second row) Massachusetts road dataset.

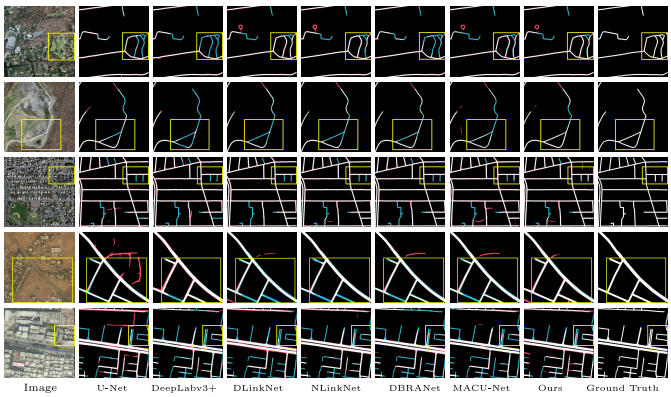

Fig. 8. We cropped some representative regions on the DeepGlobe road dataset and the Massachusetts road dataset. The red and blue represent false positives (FPs) and false negatives (FNs), respectively.

so partially occluded roads can be extracted. Labels predicted by our RCFSNet are complete because our MSCE integrates multiscale road context and global information to establish the long-range dependencies between roads, which effectively enhances the model's inference about occluded areas. Since FSFF provides a range of road structure information for the decoder, the proposed RCFSNet has perfect performance at road boundaries. The labels predicted by our model contain fewer blue lines (blue represents the unrecognized roads). The experimental results in Fig. 8 verify that RCFSNet can solve the problem of road occlusion and provide complete road extraction results.

From Table I, we can conclude that our proposed RCFSNet achieves the best recall (0.8646), IoU (0.6911), and F1-score (0.8088) on the DeepGlobe road dataset. Compared with NLinkNet, the RCFSNet gets an improvement of 3.69% for recall, which shows our model can extract the complete road network structure. Compared with DLinkNet, our IoU is improved by 1.76%, demonstrating that the road labels extracted by RCFSNet are more in line with ground truth. We also achieve the best precision (0.7825), IoU (0.6552), and F1-score (0.7866) on the Massachusetts road dataset. It proves that our approach has excellent robustness in road extraction from satellite images.

D. Ablation Study

In this section, the ablation study proves the effectiveness of the MSCE and FSFF. Table II shows that we use the ResUnet as a baseline and delete MSCE and FSFF separately.

TABLE III
ABLATION RESULTS OF CDAM ON THE DEEPGLOBE ROAD DATASET AND CBAM REPRESENT THE CONVOLUTIONAL BLOCK ATTENTION MODULE

Methods	Recall	Precision	IoU	F1-score
FSFF without CDAM	0.8318	**0.8106**	0.6905	0.8054
FSFF with CDAM	0.8546	0.7898	**0.6934**	**0.8101**
FSFF with CBAM[26]	**0.8564**	0.7839	0.6901	0.8074
FSFF with SE[17]	0.8537	0.7626	0.6690	0.7890
FSFF with CA[19]	0.8393	0.7958	0.6886	0.8046

Table II shows the quantitative results. After removing MSCE and FSFF, the IoU, and recall of the model dropped significantly, which indicates the necessity of our work. After removing MSCE, the performance in terms of the recall declined from 85.46% to 83.60%, showing that road context is useful for road extraction. Without FSFF, the model's precision and IoU dropped by 1.2% and 0.77%, respectively, which confirms that the road structure information plays an outstanding role in improving the accuracy of road labels. RCFSNet achieves the best IoU and F1, proving that the model can provide a complete road network structure with competitive accuracy.

CDAM is proposed to strengthen the channel and spatial features of roads. As shown in Table III, after deleting the CDAM, the IoU and F1-score of the model drop 0.29% and 0.47%, respectively, which indicates that CDAM improves the model's road extraction results. We also replaced the CDAM with other attention mechanisms. Compared with the model without CDAM, the recall is significantly improved, and the combination of FSFF and CDAM gets the best IoU and F1, proving that CDAM can aggregate road characteristics effectively.

IV. CONCLUSION

This letter proposes a novel method named RCFSNet for road extraction from satellite imagery, which captures the road context and fuses full-stage features effectively.

1) Given the shape of roads, a multiscale road context extract module is designed to capture the long-range dependencies in road regions. The introduced road context can significantly improve the completeness of the prediction results.
2) The FSFF is proposed to integrate full-stage features for supplying precious structure information, and the CDAM is designed to strengthen road features.
3) Extensive experiments on two public datasets indicate the superiority of our proposed RCFSNet compared with several state-of-the-art methods. Detailed ablation experiments are conducted to corroborate the necessity of MSCE, FSFF, and CDAM.

REFERENCES

[1] D. He, Y. Zhong, X. Wang, and L. Zhang, "Deep convolutional neural network framework for subpixel mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9518–9539, Nov. 2021.

[2] Q. Zhu, W. Deng, Z. Zheng, Y. Zhong, and D. Li, "A spectral–spatial-dependent global learning framework for insufficient and imbalanced hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 11709–11723, Nov. 2022.

[3] Y. Li et al., "MFVNet: Deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation," *Sci. China Inform.*, 2022.

[4] 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.

[5] M. Hu, C. Wu, L. Zhang, and B. Du, "Hyperspectral anomaly change detection based on autoencoder," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3750–3762, 2021.

[6] Q. Zhu et al., "Land-use/land-cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 63–78, Feb. 2022.

[7] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. ECCV*, 2010, pp. 210–223.

[8] 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.

[9] 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.

[10] 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.

[11] S.-B. Chen, Y.-X. Ji, J. Tang, B. Luo, W.-Q. Wang, and K. Lv, "DBRANet: Road extraction by dual-branch encoder and regional attention decoder," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[12] 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.

[13] G. Zhou, W. Chen, Q. Gui, X. Li, and L. Wang, "Split depth-wise separable graph-convolution network for road extraction in complex environments from high-resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[14] R. Li, C. Duan, S. Zheng, C. Zhang, and P. M. Atkinson, "MACU-Net for semantic segmentation of fine-resolution remotely sensed images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[15] H. Huang et al., "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1055–1059.

[16] X. Lu et al., "Cascaded multi-task road extraction network for road surface, centerline, and edge extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

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

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

[19] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13713–13722.

[20] 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.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 2017, pp. 618–626, Dec. 2017.

[22] 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.

[23] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Graduate Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2013.

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

[25] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, *Encoder–Decoder With Atrous Separable Convolution for Semantic Image Segmentation*. Cham, Switzerland: Springer, 2018.

[26] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, 2018, pp. 3–19.