ISBN: 978-981-18-1791-5
2021 the 11th International Workshop on Computer Science and Engineering (WCSE 2021)
doi: 10.18178/wcse.2021.06.012

LSDNet: Boosting Change Detection of High-resolution Remote Sensing Images by Combining Convolution-Involution and Ensemble Coordinate Attention

Yifan Liu [1,2,3], Jingdong Liu [1,2,3], Asif Raza [1,2,3], Zeng Li [1,2,3], Hong Huo [1,2,3 +], Tao Fang [1,2,3]

[1] Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China

[2] Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China

[3] Shanghai Engineering Research Center of Intelligent Control and Management, Shanghai 200240, China.

Abstract. Change detection has always been a crucial task in remote sensing fields, and there have already been great efforts made on it for decades. However, as high resolution (HR) remote sensing images generally contain abundant ground details, it still faces a huge challenge for their change detection, especially from the aspects of change detection accuracy and speed. Concerning this issue, a novel lightweight Siamese deep network (LSDNet) is proposed, and it combines Convolution-Involution Module (CIM) and Ensemble Coordinate Attention Module (ECAM) for boosting the change detection of HR remote sensing images. CIM summarizes the context of ground objects and reweights the importance of different positions, while ECAM aggregates multiple levels of semantic features and pays different attention to different spatial information. The experiments on CNZ data set have shown that the proposed LSDNet performs better than state-of-the-art (SOTA) change detection methods, especially it improves the accuracy by 1.92% and reduces the amount of model parameters by 32.89% compared to SNUNet-CD which has the best performance currently.

Keywords: artificial intelligence, change detection, Ensemble Coordinate Attention, high-resolution remote sensing images

1. Introduction

Change detection of remote sensing images is the process of identifying differences in the state of a ground object by observing it at different times[1], which plays a significant role in effective land management, resources survey, and damage assessment. Nowadays, there are many change detection methods based on deep neural networks, and they may roughly be divided into two categories, binary change detection and multi-type change detection. Binary change detection only aims to identify whether a ground object has changed or not, while multi-type change detection needs not only identify the change, but also determine the change type, namely, from which land cover type to which land cover type.

There are different deep networks utilized for change detection. U-Net[2] is a benchmark model for the first time. Siamese network is gradually being used and becomes the mainstream for change detection[3][4][5][6]. Siamese NestedUNet Concate (SNC)[3], which combinates DenseNet[6] and UNet++[7], narrows down the loss of information transmission by densely connected mechanism. Based on SNC[3], SNUNet-CD[3] is obtained by introducing channel attention module[8]. DSIFN[9] is composed of two sub-networks, one is called difference discrimination and the other is called shared deep feature extraction. Besides different networks, a great deal of efforts have been made on deep feature extraction, such as multiscale features extracted by pyramid network[10], coordinate attention[11] and so on.

Although above mentioned methods have achieved relatively good change detection performance, there still lies two key issues, down-samplings that bring the loss of accurate spatial position information and a large number of parameters that cost much computing power. Inspired by SNC[3] and RedNet[12], a novel lightweight Siamese deep network (LSDNet) is proposed, and it aims to tackle the typical two issues involving change detection of HR remote sensing images, namely, accuracy and speed.

+ Corresponding author. Tel.: +86-21-34204758.
E-mail address: huohong@sjtu.edu.cn.

The contributions of the article are mainly as follows:

- Inspired by convolution unit[13] and involution layer[12], CIM is designed for summarizing the context of ground objects and reweighting the importance of different positions in the spatial domain.
- ECAM is proposed to aggregate multiple levels of semantic features and learn the long-range dependencies for improving the overall change detection accuracy.
- It is the first time that convolution-involution and ensemble coordinate attention are introduced and combined for change detection of HR remote sensing images.

The rest of this paper is organized as follows. Section 2 details the structure of proposed LSDNet. Section 3 is the comparative analysis of experiments. Finally, the conclusion is drawn in Section 4.

2. Network Architecture

The architecture of LSDNet is shown in Fig.1. Its backbone is a Siamese network with shared weights, also called the encoder, and it ultimately has four outputs of the same size which are at different semantic levels and spatial positions. The outputs of shallower sub-decoders have more precise localization, while the outputs of deeper sub-decoders have richer semantics. Coordinating the outputs of LSDNet can obtain more accurate representations.

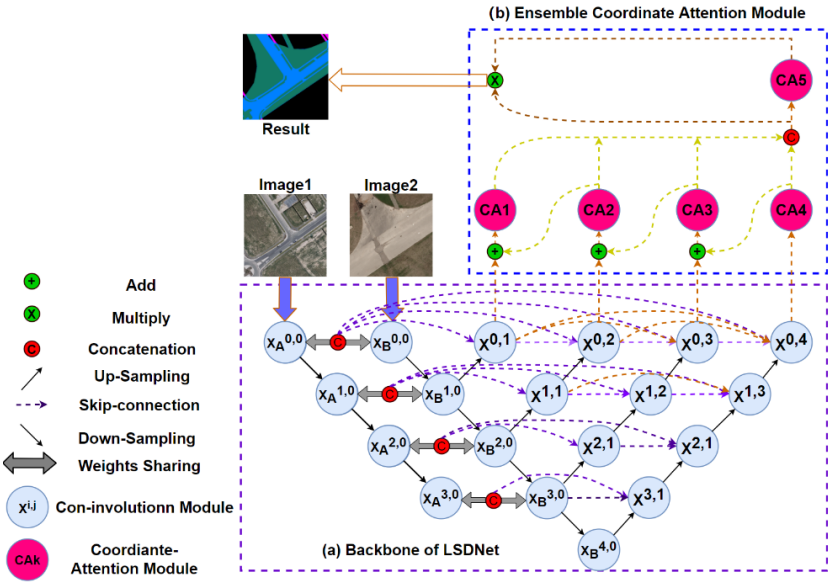

Fig. 1: The architecture of LSDNet. (a) the proposed backbone. (b) ECAM.

LSDNet mainly includes Convolution-Involution Module (CIM) and Ensemble Coordinate Attention Module (ECAM). Each $X^{i,j}$ denotes a CIM and each CAk denotes a CAM[6]. Firstly, two-temporal images are input into each branch of this Siamese network, respectively. In this way, their features are extracted by same convolution filters, which is beneficial for detecting the changed ground objects. Then, the feature maps extracted by two branches separately are merged for obtaining the complete information accurately. To maintain features of high-resolution remote sensing images, the dense skip connection mechanism[6] is introduced into LSDNet, as the doted arrows shown in Fig.1(a), purple doted arrows indicate connections between encoders and sub-decoders, and rose gold doted arrows indicate connections between sub-decoders and sub-decoders).

The CIM has a residual unit structure, as shown in Fig. 2(b). It totally includes 6 layers. The first layer is 2-Dimension convolution layer (Conv2D) which is responsible for adjusting the number of input channels of the feature map to match the number of output channels of the feature map. After that, there are a BatchNorm (BN) layer and a Rectified Linear Unit (ReLU) activation layer. Then, an involution layer[12] and a BN layer are applied. The outputs of Conv2d and the second BN layer will be added and input to the last ReLU activation layer.

Let $x^{i,j}$ denote the output of $X^{i,j}$:

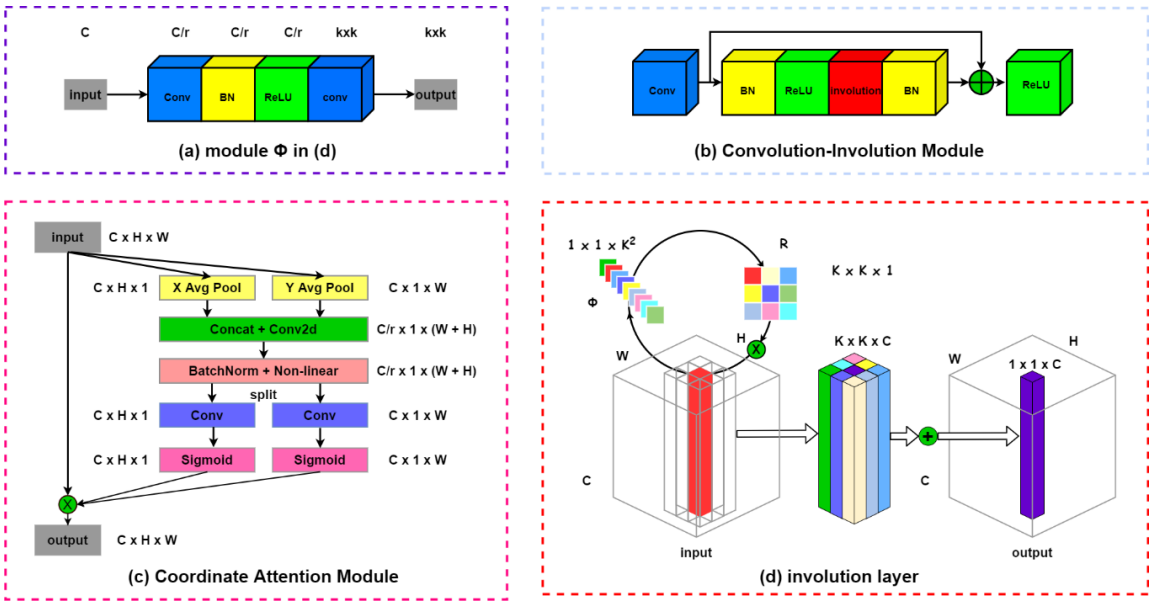

Fig. 2: The architecture of some modules. (a) Module Φ in (d). (b) CIM. (c) CAM[11]. (d) involution layer[12].

$$x^{i,j} = \begin{cases} P(H(x^{i-j,j})), j = 0 \\ H([x_A^{i,0}, x_B^{i,0}, U(x^{i+1,j-1})]), j = 1 \\ H([x_A^{i,0}, x_B^{i,0}, [x^{i,j}]_{k=1}^{j-1}, U(x^{i+1,j-1})]), j > 1 \end{cases} \tag{1}$$

where $H(\cdot)$ denotes the operation of a CIM, and $P(\cdot)$ denotes a 2x2 max pooling operation for down-sampling, which is indicated by the downward arrows in Fig. 1(a). $U(\cdot)$ denotes the up-sampling using transpose convolution, which is indicated by the upward arrows. [·] denotes that the features are concatenated and fused in the channel dimension. If j=0, the encoder will extract features and then downsample by max pooling; If j>0, the features in the encoder are directly transmitted to the decoder by the dense connected mechanism.

The involution layer[12] is the key layer of the CIM and can adaptively allocate the weights over different positions and prioritize the most informative visual elements in the spatial domain. As shown in Fig. 2(d), the involution kernel size K equals 3, which is a replacement for a Conv2D with a convolution kernel size of 3. The involution kernel is generated by module Φ in Fig. 2(a). After that, a rearrangement "R" is implemented from channel dimension to spatial dimension. Then the involution kernel is expanded to C dimension on the channel dimension and multiplied by the original K×K spatial neighbourhood. Finally, the output features are aggregated within the K×K spatial neighborhood.

To coordinate channel attention generation and spatial information embedding, CAM[11] encodes both long-range dependencies and channel relationships. The structure of CAM[11] is shown in Fig.2(c), "X Avg Pool" and "Y Avg Pool" refer to 1D horizontal global pooling and 1D vertical global pooling, respectively. It not only reweights the importance of different channels, but also considers encoding the spatial information. By disassembling the channel attention into horizontal and vertical parts and inputting these two parts to a tensor at the same time for combining them, it can further focus more on the spatial location of the object of interest for better detection. ECAM includes five CAMs. Four of them ensemble the four outputs of backbone of LSDNet, and the other one is used to further aggregate the outputs of these four CAMs, as shown in Fig.1(b). It can be formulated as follows:

$$x_,^{0,4} = C(x^{0,4}) \tag{2}$$

$$x_,^{0,3} = C(x_,^{0,4} + x^{0,3}) \tag{3}$$

$$x_,^{0,2} = C(x_,^{0,3} + x^{0,2}) \tag{4}$$

$$x_,^{0,1} = C(x_,^{0,2} + x^{0,1}) \tag{5}$$

$$X_{concate} = [x,^{0,1}, x,^{0,2}, x,^{0,3}, x,^{0,4}] \tag{6}$$

$$Y = h(X_{concate} \otimes C(X_{concate})) \tag{7}$$

where $C(\cdot)$ denotes a CAM[11], $x,^{i,j}$ denotes the output of $C(\cdot)$, $[\cdot]$ denotes the concatenation of features in the channel dimension, $\otimes$ denotes element-wise product, $h(\cdot)$ denotes a 1×1 convolution layer to generate $Z \times H \times W$ change map Y ("Z" indicates the number of change types).

3. Experiments

3.1. Data Set

To evaluate the proposed method, a series of experiments are conducted on one bi-temporal high resolution (BHR) data set. The data set covers the city of Christchurch, New Zealand, called CNZ data set[17]. The image size of the data set is 30k * 30k with RGB three channels. The images are taken on February 24, 2011 (called CNZ-1) and April 10, 2014 (called CNZ-2). The ground resolution is 0.3m. It has 5 unchanged categories and 19 changed categories[18]. Fig.3 shows one pair of images of CNZ data set.

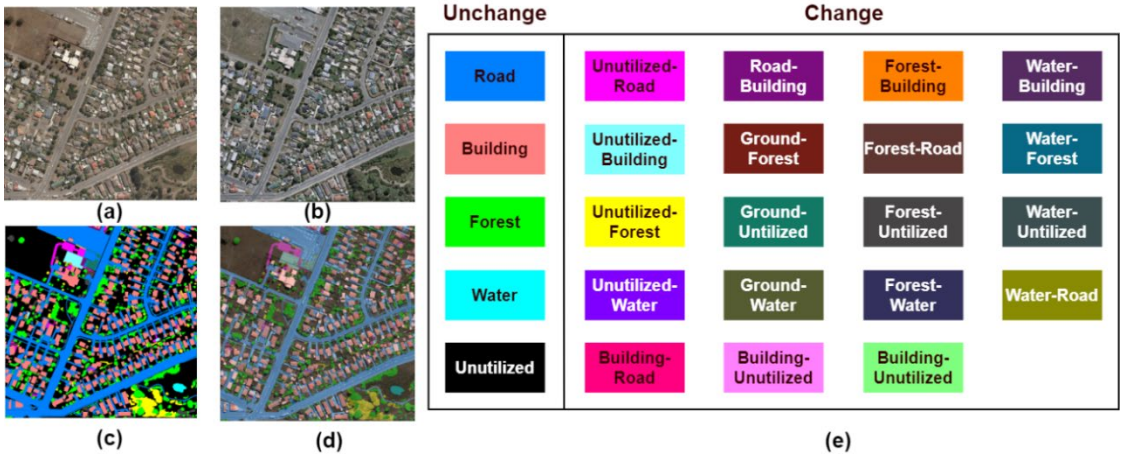

Fig. 3: The data set of (a) CNZ-1[17][18], (b) CNZ-2[17][18]. (c) the ground truth. (d) the mask visualization result. (e) the label of different change types.

To evaluate the performance of the proposed LSDNet, three evaluation indicators are used: overall accuracy (OA), model parameters and floating point of operations (FLOPs). The larger OA and the less model parameters and FLOPs are, the better prediction result is. OA is expressed as follow:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

where, TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative respectively.

3.2. Implementation Details

All experiments are powered by 2×GTX 2080Ti under PyTorch framework. During training, the batch size is set to 4, and Adam is applied as an optimizer. The learning rate adjustment strategy is cosine annealing with warmup. The warmup stage contains 10 epochs, and the learning rate of each epoch increases by 1e-4. The learning rate starts to decrease from 1e-3 to 0 in the annealing stage, which lasts 90 epochs to make the network converge. The weights of each convolution layer are initialized by the KaiMing normalization. The loss function is cross-entropy loss, which is shown in formulate (9).

$$loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\, log(p_{ik}) \tag{9}$$

"N" denotes the number of pixel class; y_{ik} indicates the variable. If the category is the same as the category of sample "i", it is 1, otherwise it is 0; p_{ik} stands for the probability that the observed sample "i" belongs to category "k".

3.3. Comparison and Analysis

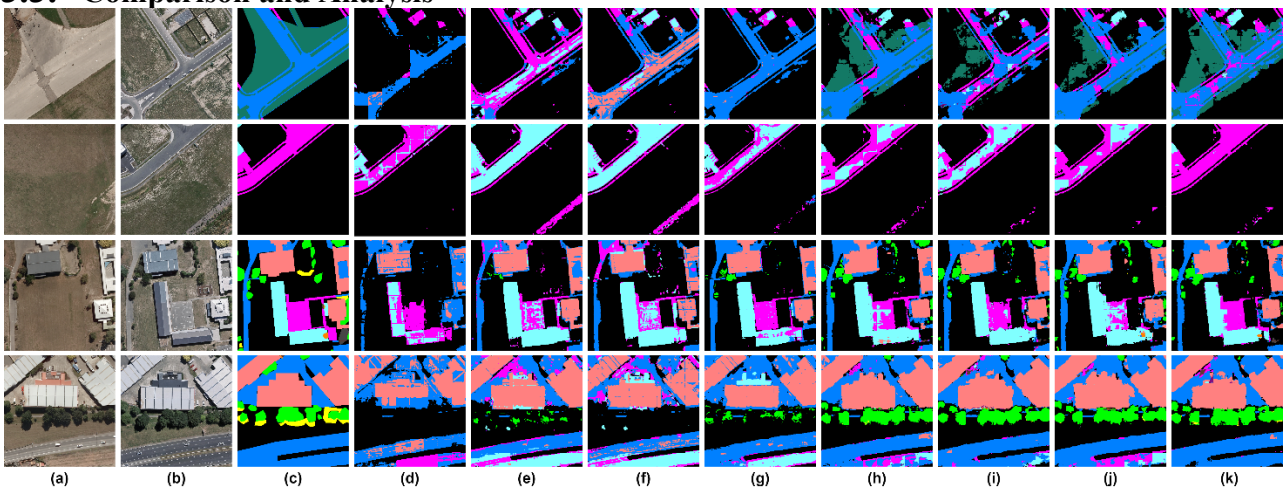

Fig. 4: Visualization results: (a) and (b) are original bi-temporal images; (c) the ground truth; (d) the result of DSIFN[9]; (e) the result of FC-Siam-conc[4]; (f) the result of FC-Siam-diff[4]; (g) the result of FC-EF[4]; (h) the result of SNUNet-CD[3] (8 channels); (i) the result of SNC[3]; (j) the result of LSDNet (without ECAM); (k) the result of LSDNet.

Several change detection methods are selected for comparison. FC-Siam-conc[4], FC-Siam-diff[4], and FC-EF[4] are the mainstream methods for change detection. DSIFN[9] is composed of two sub-networks. SNC[3] combinates DenseNet[6] and UNet++[7]. SNUNet-CD[3] is obtained by introducing channel attention module[8] to SNC[3].

As shown in Table 1, LSDNet, which has 0.51 M parameters and 0.38 G FLOPs, achieves the highest OA and lowest model parameters. Compared with SNUNet-CD[3], which has the best performance currently, LSDNet has improved overall accuracy by 1.92%, and the amount of model parameters has been reduced by 32.89%, and FLOPs has been reduced by 13.64%. CIM is the main reason for improving overall accuracy and reducing the amount of model parameters and FLOPs. Comparing LSDNet (without ECAM) with SNC[3], which is the baseline method, CIMs bring 2.15% OA improvement and reduce 0.25 M model parameters. Although the ECAM increases FLOPs by 5.6%, the OA has increased by 0.66%. As is shown in Fig. 4, comparing "(j)" and "(k)", it could be found that ECAM can integrate different levels of semantic features and greatly capture long-range interactions spatially, which also boosts the performance of change detection.

Table 1: Performance of Comparison on CNZ data set

Method	Params (M)	FLOPs (G)	Overall Accuracy (%)
DSIFN[9]	50.44	10.29	78.09
FC-Siam-conc[4]	1.55	0.69	71.58
FC-Siam-diff[4]	1.35	0.62	70.05
FC-EF[4]	1.35	0.47	80.02
SNUNet-CD[3]	0.76	0.44	83.66
SNC[3]	0.76	0.44	82.77
LSDNet (without ECAM)	0.51	**0.36**	84.92
LSDNet	**0.51**	0.38	**85.58**

4. Conclusion

In this letter, a novel lightweight Siamese deep network for HR remote sensing images is proposed, called LSDNet. It combines two essential modules for boosting the change detection: CIM and ECAM. CIM summarizes the context of ground objects and reweights the importance of different positions, while ECAM aggregates multiple levels of semantic features and pays different attention to different spatial information. The experiments on CNZ data set have shown that the proposed LSDNet has higher accuracy and speed with less parameters comparing to the state-of-the-art (SOTA) change detection methods.

5. Acknowledgement

This study is supported by the National Key Research and Development Program of China (No.2018YFB0505000), the National Science and Technology Major Project (21-Y20A06-9001-17/18), and the Science Fund for Creative Research Groups of the National Natural Science Foundation of China (No. 61221003).

6. References

[1] A. Singh, “Review Article: Digital change detection techniques using remotely-sensed data,” international Journal of Remote Sensing, vol. 10, no. 6, pp. 989–1003, 1989.

[2] 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.

[3] S Fang, K Li, J Shao, et al. SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images[J]. IEEE Geoscience and Remote Sensing Letters, 2021, PP(99):1-5.

[4] R. Caye Daudt, B. Le Saux, and A. Boulch, “Fully convolutional Siamese networks for change detection,” in Proc. 25th IEEE Int. Conf. Image Process. (ICIP), Oct. 2018, pp. 4063–4067.

[5] F. Rahman, B. Vasu, J V. Cor, et al., “Siamese network with multi-level features for patchbased change detection in satellite imagery,” in Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP), Nov. 2018, pp. 958–962.

[6] G Huang, Z Liu, V Laurens, et al. Densely Connected Convolutional Networks[J]. IEEE Computer Society, 2016.

[7] Z Zhou, M Siddiquee, N Tajbakhsh, et al. UNet++: A Nested U-Net Architecture for Medical Image Segmentation[J]. 4th Deep Learning in Medical Image Analysis (DLMIA) Workshop, 2018.

[8] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 3–19.

[9] A, Cz, Y Peng, E Dt, et al. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 166:183-200.

[10] H. Chen and Z. Shi, “A spatial-temporal attention-based method and a new dataset for remote sensing image change detection,” Remote Sens., vol. 12, no. 10, p. 1662, 2020.

[11] Q Hou, D Zhou, J Feng. Coordinate attention for efficient mobile network design[J]. arXiv preprint arXiv:2103.02907, 2021.

[12] D Li, J Hu, C Wang, et al. Involution: Inverting the Inherence of Convolution for Visual Recognition[J]. 2021.

[13] D Peng, Y Zhang, H Guan. End-to-end change detection for high resolution satellite images using improved UNet++[J]. Remote Sensing, 2019, 11(11): 1382.

[14] J Chen, Z Yuan, J Peng, et al. DASNet: Dual attentive fully convolutional siamese networks for change detection of high-resolution satellite images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, PP(99).

[15] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.

[16] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.

[17] S Ji, S, Wei, M Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 57(1): 574-586.

[18] C Fu, T Bao, L Lv, et al. Multi-task Learning for Bi-temporal Remote Sensing Scene Parsing via Patch-pixel Representation[C]// ICMLC 2020: 2020 12th International Conference on Machine Learning and Computing. 2020.