RMNET: A Residual and Multi-scale Feature Fusion Network For High-resolution Image Semantic Segmentation

ZiRui Shen
College Of Information
Science And Technology
Nanjing Forestry University
Nanjing, China
szr@njfu.edu.cn

Xin Li
College Of Information
Science And Technology
Nanjing Forestry University
Nanjing, China
lixin99@njfu.edu.cn

Sheng Xu
College Of Information
Science And Technology
Nanjing Forestry University
Nanjing, China
xusheng@njfu.edu.cn

Abstract—High-resolution remote sensing images have high clarity and provide significant support for urban planning, resource management, environmental monitoring, and disaster warning. Semantic segmentation accurately helps extract the boundaries of objects, thereby increasing the application value of scene understanding. Traditional encoder-decoder architecture networks lack multi-scale information fusion and fail to capture precise multi-scale semantic information, when segmenting targets at different scales. Additionally, these semantic segmentation networks have inadequate handling of class-imbalanced data, resulting in unsatisfactory classification results and final segmentation effect. This paper proposes a semantic segmentation network based on residual blocks and multi-scale feature fusion. Building upon the U-Net network, we design residual modules and multi-scale feature fusion modules to extract information-rich feature maps. Then, the multi-scale feature fusion module is used to interpolate and upsample the obtained feature maps, which are then concatenated with feature maps at the same layer, resulting in a novel fusion feature map. In experiments, the performance of the proposed model surpasses U-Net with improvements reaching 6.06% for MIoU. The introduced network identifies complex land features including dense distribution of objects, small objects, large differences in object characteristics and complex background effectively preserves and restores feature information by incorporating the multi-scale feature fusion module, achieving higher precision segmentation results and providing rich multi-scale and spatial information.

Index Terms—deep learning, high-resolution, semantic segmentation, residual block, multi-scale feature fusion

I. Introduction

Semantic segmentation of high-resolution remote sensing images plays a crucial role in various fields such as land resource investigation, natural disaster monitoring, and national security. It aims to label each pixel in an image with a corresponding semantic class, dividing the image into meaningful semantic regions. Unlike traditional image segmentation, semantic segmentation requires classifying each pixel into classes with semantic meanings, such as pedestrians, vehicles, roads, and buildings. Semantic segmentation of remote sensing images enables the efficient extraction of category and geometric information from different scenes.

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62102184), in part by the Natural Science Foundation of Jiangsu Province (Grant No. BK20200784), and in part by the Graduate Research and Innovation Projects of Jiangsu Province (Grant No. SJCX23_0320) (Corresponding author: Sheng Xu).

However, due to the characteristics of remote sensing data, semantic segmentation poses greater challenges than natural images. The challenges arise from the large data volume and high computational complexity associated with high-resolution, wide-range, and multi-spectral remote sensing images. Furthermore, remote sensing images capture diverse and complex targets, including buildings, roads, vegetation, etc., making segmentation diverse and complex.

Image segmentation methods can be divided into traditional image segmentation and deep learning-based image segmentation. Traditional methods are usually based on heuristic rules or statistical models, which have relatively low model complexity. Deep learning methods typically employ deep neural network models with higher model complexity and parameters. This enables deep learning methods to have stronger expressive power and flexibility, making them suitable for more complex image segmentation tasks.

High-resolution image segmentation in deep learning networks aims to divide images into different regions or objects with semantic or structural significance. This task is of great importance for applications such as image understanding, scene analysis, object detection and image processing [2] [3] [4]. In high-resolution image segmentation, the goal is to assign each pixel to its corresponding semantic category or region, thus achieving a fine-grained analysis of the image. This segmentation result provides more detailed and accurate image information, making subsequent analysis and processing tasks more precise and efficient. To achieve high-resolution image segmentation, researchers have used various techniques and methods [5]. By using deep neural networks such as Convolutional

Neural Networks (CNN) and its variants, it is possible to learn feature representations and semantic information, thereby achieving pixel-level segmentation.

In semantic segmentation models, downsampling is employed in the encoder to extract features, gradually reducing the size and channel number of feature maps, leading to information loss. Thus, efficiently extracting target features, preserving them during the extraction and reconstruction processes, and minimizing information loss are crucial for achieving semantic segmentation. The main work of this paper is as follows:

The proposed remote sensing image semantic segmentation network is based on residual blocks and multi-scale feature fusion to improve segmentation performance.

The proposed decoder part of the model incorporates a multi-scale feature fusion module to better capture multi-scale semantic information. By combining bilinear interpolation and deconvolution operations during the upsampling process, we address the issue of target boundary detail processing insufficiency, thereby preserving image smoothness and detail feature information.

We conduct experimental analysis using the WHDLD dense labeling dataset to validate the accuracy of the proposed model [1]. Results demonstrate that the proposed residual blocks enables better identification of complex land targets, achieving higher accuracy. Moreover, the incorporation of multi-scale feature fusion modules effectively preserves and restores feature information, providing richer multi-scale and spatial information as substantial support for prediction.

II. The Method

A. Model Architecture

Traditional CNN semantic segmentation models suffer from the problems of feature information loss and blurring during experimentation. This is mainly due to two reasons. Firstly, remote sensing images contain abundant geographical information, and the quality of the images is unstable due to factors such as shadows. This affects the accuracy and robustness of the semantic segmentation task. Secondly, the characteristics of upsampling and downsampling lead to the loss of feature information, thereby affecting the segmentation accuracy.

In order to achieve semantic segmentation of remote sensing images, this section proposes a semantic segmentation network based on residual blocks and multi-scale feature fusion called RMNet, as shown in Fig . 1. RMNet aims to introduce residual blocks and increase the depth of the network within a reasonable range to enhance feature representation capability. In the decoding part, a combination of bilinear interpolation and deconvolution is used for upsampling, and the output is concatenated with the same-level feature information through the concat operation to complete the decoding and obtain detailed features.

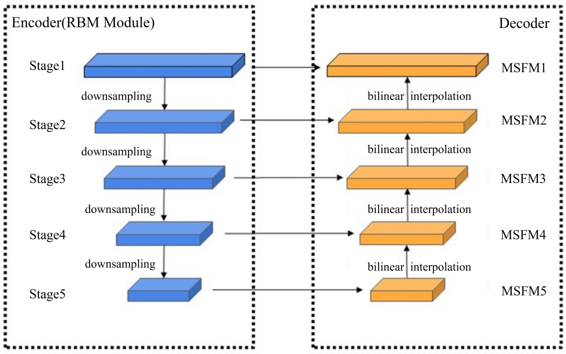

Fig. 1. RMNet network structure.

B. Encoder

Due to the high-resolution and complex nature of objects in remote sensing images, our work employs ResNet50 as the feature extraction network for better feature extraction. The input image size is illustrated in the schematic diagram of the feature extraction network (Fig. 2).

The structure consists of a convolutional layer and four residual blocks. Each residual block is composed of important modules or bottleneck modules, which are concatenated optimally. The basic module uses two 3x3 convolutions and a shortcut connection to implement the residual structure. The bottleneck module reduces the number of channels with a 1x1 convolution, extracts features with a 3x3 convolution, and increases the number of channels with another 1x1 convolution. A 1x1 convolution is used in the shortcut connection to adjust the number of channels. The basic module is suitable for low-dimensional input data, while the bottleneck module uses more channels and fewer convolutional kernels to reduce the number of parameters and computations, making it suitable for high-dimensional input data. Therefore, in Stage 2, three basic modules are used for shallow feature extraction, while Stage 3, 4, and 5 respectively use 4, 6, and 3 bottleneck modules for deep feature extraction. The residual structure solves the problems of gradient vanishing or degradation in traditional deep networks by introducing skip connections. It accelerates the training process, improves the network's expressive power, and makes deep networks easier to train and optimize.

C. Decoder

To generate pixel-level results in semantic segmentation, feature maps need to be upsampled to restore low-resolution feature maps to the size of the original image. Deep feature maps contain more semantic information that can be used for feature recognition and classification. Therefore, to enhance segmentation accuracy for semantic segmentation tasks, it is necessary to consider both the low-level features and semantic information of the image.

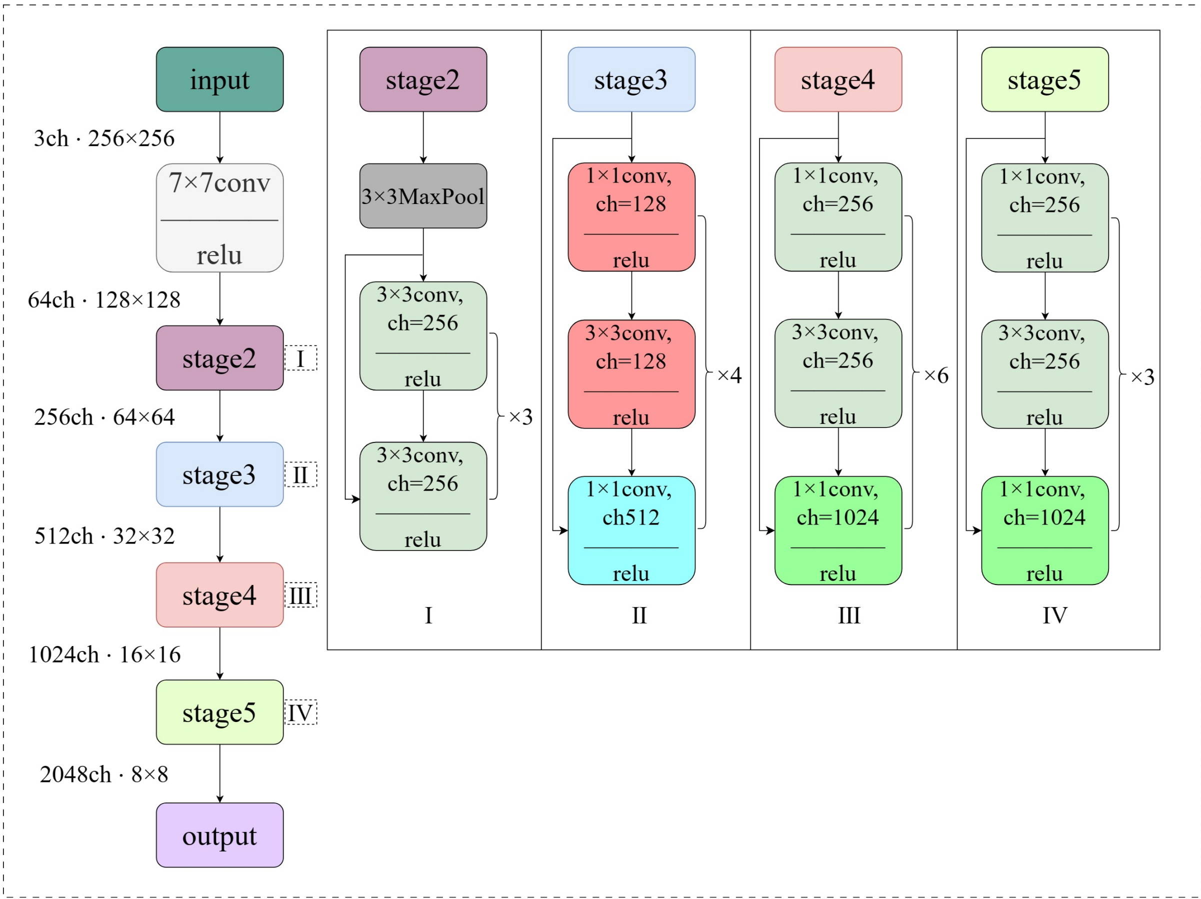

Fig. 2. Feature extraction network structure.

This necessitates the fusion of features from different levels.

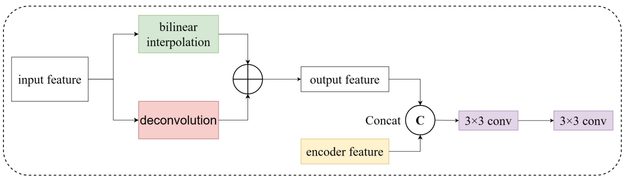

Fig. 3. MSFM module structure schematic.

In this section, we propose the MSFM module, as shown in Fig. 3. Firstly, the feature maps are subjected to deconvolution to map from low-level features to high-level features. To preserve the integrity of the low-level features, a bilinear interpolation method is utilized. Our method helps retain the detailed information of the original image. Next, the obtained feature maps are combined to yield a comprehensive and rich set of features. Our resulting feature information is concatenated with the feature information from the same layer encoder, creating a more enriched feature layer. The fusion of low-level features with semantic information provides the network with a more diverse range of channel information, which supports accurate predictions. Finally, our concatenated feature map undergoes two additional convolution operations to refine the feature maps and extract more valuable information, thus improving the accuracy of semantic segmentation.

D. Loss Function

Since the majority of pixels are typically background, there is an imbalance between foreground and background. To address this issue, we adopt the DiceLoss function. This function focuses more on foreground region exploration to ensure a low number of false negatives, which can lead to loss saturation. To mitigate this saturation issue, we combine the FocalLoss with the DiceLoss. The final loss function is:

$$L = DiceLoss + \lambda \cdot FocalLoss, \tag{1}$$

where λ represents the coefficient of FocalLoss in the whole loss function. This paper sets λ to 0.9.

The expression of DiceLoss is:

$$DiceLoss = 1 - \frac{2\,|X \cap Y| + S}{|X| + |Y| + S}, \tag{2}$$

where |X| and |Y| represent number of predicted and true values respectively, while $|X \cap Y|$ represents the intersection of x and y. The parameter S is introduced in the formula with a value of 10^{-5}, it helps prevent overfitting.

The expression of FocalLoss is:

$$FocalLoss = -\alpha_t (1 - p_t)^{\gamma} \log(p_t), \quad (3)$$

where p_t represents the probability of correctly classifying the predicted sample, while both α_t and γ are adjustable factors. This study sets α_t and γ to 2 and 0.25 respectively.

III. Experiment

A. WHDLD Dataset and Evaluation

WHDLD is a collection of 4940 RGB images in 256 × 256, captured by the Gaofen 1 and ZY-3 Satellites, specifically focusing on the Wuhan urban area. It contains six types of remotely sensed feature types, extracted from UC Merced and released by Wuhan University in 2018. The targets are divided into bare, building, pavement, road, vegetable and water. In order to verify the effectiveness of the method, we compare and analyze RMNet with common semantic segmentation models. In which all network models are trained with an epoch of 100, and save the optimal weight file obtained after the training is completed, the prediction was performed on the test set, and the accuracy evaluation of the comparison experiments is shown in Table I:

TABLE I
Experimental results of different model on WHDLD

Model	OA	AA	K	mIoU	F_1
SegNet [6]	80.229	63.787	71.403	52.940	66.529
U-Net [7]	81.830	67.724	74.422	55.706	68.567
Tiramisu [8]	82.188	70.712	74.903	58.167	71.276
FGC [9]	82.975	68.855	75.927	57.368	70.274
MSFCN [10]	84.168	72.081	77.558	60.366	73.031
CNet+RBM	82.216	70.934	75.004	56.957	69.618
CNet+MSFM	82.934	71.307	76.288	58.019	71.530
RMNET	84.395	73.320	78.895	61.763	73.852

The performance of different model on WHDLD is shown in Tables I and II. OA and AA represent overall accuracy and average accuracy respectively, while K and F_1 represent Kappa coefficient and F_1 score respectively. RMNet demonstrates improvements in all F_1 score performance metrics, all six models perform better in segmenting road and water, but encounter challenges in segmenting sidewalks and vegetation. The analysis of predicted images in the test set suggests that shadows and other effects may cause sidewalks and vegetation to appear similar to road and water in color performance. Fig. 4 presents the results of three models.

The overall segmentation effect is more accurate and the edges of the features are relatively smoother and more accurate. In comparison, MSFCN reduces miss classification in larger regions compared with the U-Net model, and overall, its performance is relatively satisfactory. However, there are still some miss classification issues in certain detail regions. The classic U-Net misses classification results and its overall effect is weak.

B. Ablation Experiment

This section discusses the residual module (RBM) and the multiscale feature fusion module (MSFM). The effects of each module on the segmentation results are explored and analyzed through ablation experiments conducted using the control variable approach. The model without the introduction of both modules is referred to as a normal convolutional network (CNet).

The effectiveness of the network for segmentation of detail regions and its vulnerability to segmentation leakage were observed after removing the MSFM module, as shown in Fig. 4. This suggests that the integration of low-level features and the fusion of low-level features with semantic information are essential for achieving effective semantic segmentation. The removal of the RBM module leads to rougher segmentation results, increased leakage, and incorrect segmentations, indicating the model's sub-optimal performance.

The evaluation indexes have decreased to different degrees when moving out of each module, as shown in Tables I and II. Specifically, when the MSFM module is removed, there is a slight decrease in each indicator. This is mainly due to the loss of detail feature information after removing the MSFM board. As a contrast, the MSFM module preserves low-level features of the image. These features are then combined with the feature information of the same layer through the concat operation, resulting in a richer feature layer. This fusion of low-level features with semantic information provides strong support for the network's accurate prediction. As for RBM module, it utilizes residual blocks and cross-layer connections that enable direct transfer of information between different layers. This effectively reduces overfitting and improve the generalization ability. Furthermore, RBM facilitates faster training speed and higher accuracy can be achieved.

IV. Conclusion

This work proposes a semantic segmentation network for remote sensing images based on residual blocks and multi-scale feature fusion. The network architecture consists of an encoder stage to perform deep feature extraction and a decoder stage enhances image segmentation accuracy and robustness by integrating low-level and high-level feature information. The proposed semantic segmentation network has several advantages. Firstly, it significantly improves segmentation accuracy and the ability to locate and segment the target accurately. Secondly, it reduces false detections, thereby filtering background interference and improving the reliability of semantic segmentation.

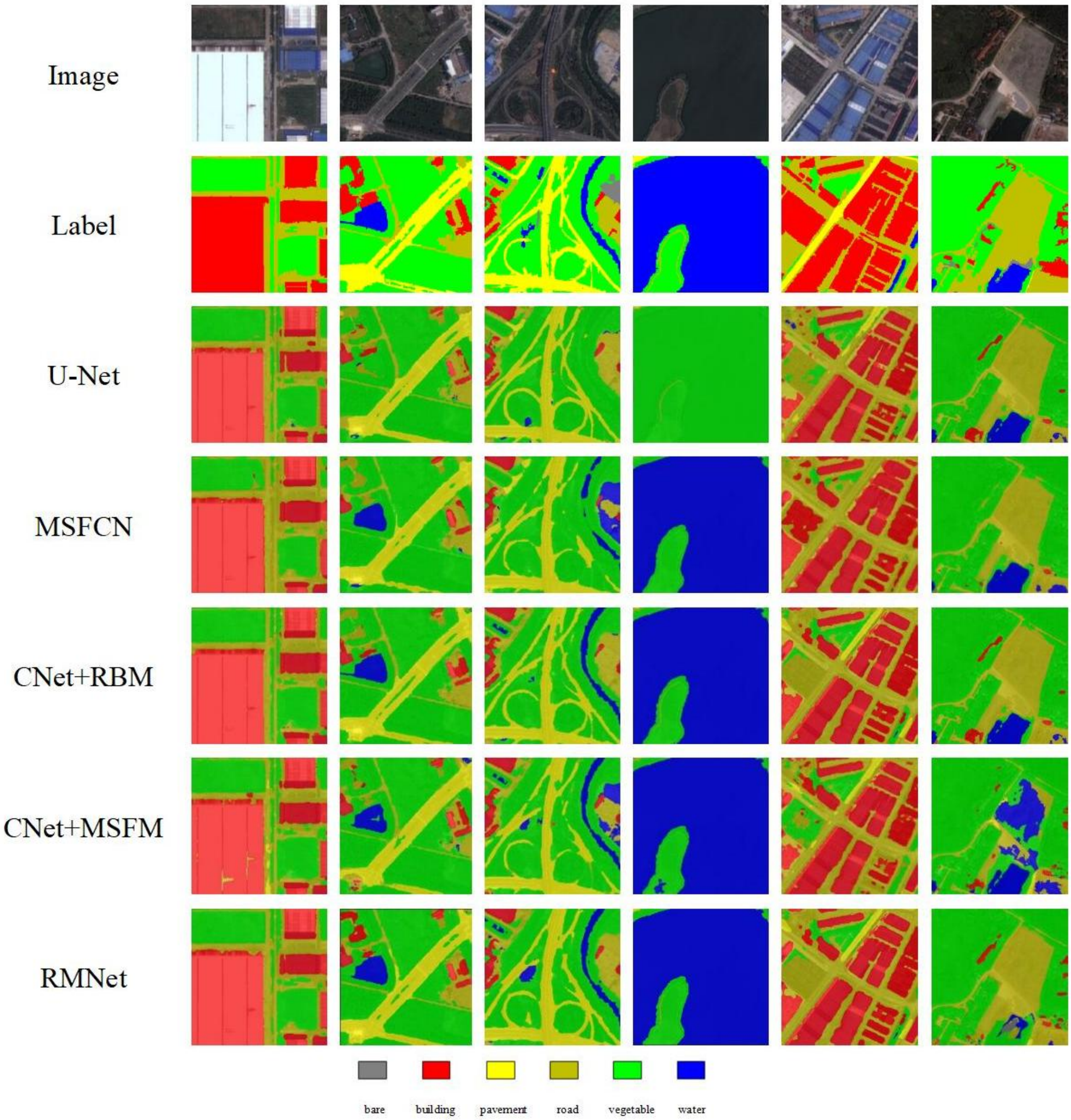

Fig. 4. Prediction results of different methods and ablation experiment.

TABLE II
F_1 score index results for each category of the experiment

Model	bare	building	pavement	road	vegetable	water
SegNet [6]	47.682	63.253	51.466	54.649	86.473	95.649
U-Net [7]	43.097	70.752	52.609	58.668	89.185	97.089
Tiramisu [8]	50.313	68.918	53.576	70.047	88.206	96.598
FGC [9]	50.282	72.642	53.842	57.931	89.651	97.294
MSFCN [10]	52.178	74.499	55.177	68.797	90.024	97.511
CNet+RBM	49.010	69.271	53.760	60.143	88.522	97.003
CNet+MSFM	49.694	73.380	54.048	65.312	89.502	97.244
RMNet	51.242	78.147	55.829	69.434	90.506	97.956

Additionally, the network maintains a fast-processing speed while achieving high accuracy, enhancing its real-time performance. Ablation experiments are conducted on the residual block and multi-scale feature fusion module, demonstrating their contribution to the segmentation task.

Future research should consider breaking through the limitations of CNN, for example, the initial stage of the network using CNN can only utilize local information due to the limited size of the convolutional kernel, thus resulting in a lack of comprehensive understanding of the input image, and subsequently impact the distinguishability of the features extracted by the encoder at the end.

References

[1] Z. Shao, W. Zhou, X. Deng, M. Zhang, Q. Cheng, Multilabel Remote Sensing Image Retrieval Based on Fully Convolutional Network, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13(1):318-328.

[2] H. Wu, Z. Gui, and Z. Yang, Geospatial Big Data for Urban Planning and Urban Management, Geo-Spatial Information Science, 2020, 23 (4): 273–274.

[3] P. Bharati, A. Pramanik, Deep learning techniques—R-CNN to mask R-CNN: a survey, Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019, 2020: 657-668.

[4] C. Yang, Q. Zhan, S. Gao, H. Liu, Characterizing the Spatial and Temporal Variation of the Land Surface Temperature Hotspots in Wuhan from a Local Scale, Geo-spatial Information Science , 2020, 23 (4): 327–340.

[5] V. S. F. Garnot, L. Landrieu, S. Giordano, C. Nesrine, Satellite image time series classification with pixel-set encoders and temporal self-attention, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 12325-12334.

[6] V. Badrinarayanan, A. Kendall, R.Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.

[7] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer International Publishing, 2015: 234-241.

[8] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017: 11-19.

[9] S. Ji, Z. Zhang, C. Zhang, S. Wei, M. Lu, and Y. Duan, "Learning discriminative spatiotemporal features for precise crop classification from multi-temporal satellite images," International Journal of Remote Sensing, 2020, 41(8) :3162-3174.

[10] R. Li, S. Zheng, C. Duan, L. Wang, C. Zhang, Land cover classification from remote sensing images based on multi-scale fully convolutional network, Geo-spatial information science, 2022, 25(2): 278-294.