

# Improved Mask R-CNN for Disturbed Area Extraction in Construction Projects from High-Resolution Satellite Imagery

Yuanling Zhao

Department of Soil and Water Conservation  
Changjiang River Scientific Research  
Institute  
Wuhan, China  
zhaoyuanling@mail.crsri.cn

Dongbing Cheng

Department of Soil and Water Conservation  
Changjiang River Scientific Research  
Institute  
Wuhan, China  
chengdb@mail.crsri.cn

Shengyu Shen

Department of Soil and Water Conservation  
Changjiang River Scientific Research  
Institute  
Wuhan, China  
shenshengyu@mail.crsri.cn

Daoming Cai

Department of Soil and Water  
Conservation  
Changjiang River Scientific  
Research Institute  
Wuhan, China  
cdm1528@126.com

Xujun Lyu\*

College of Engineering  
Huazhong Agricultural  
University  
Wuhan, China  
\* Corresponding author:  
lyuxujun@mail.hzau.edu.cn

**Abstract**—This study introduces an enhanced Mask R-CNN model for automatically extracting disturbed areas in construction projects from satellite imagery, aiding soil and water conservation authorities in monitoring tasks. The model optimizes the feature extraction network structure and loss function to bolster object recognition and segmentation accuracy. By integrating a channel attention mechanism, SENet, the model's capacity to detect target objects is augmented. The loss function optimization results in more precise edge detection. Utilizing the high-resolution satellite data from GF-1, the effectiveness of the proposed method has been successfully demonstrated in the study area of Liuyang City, Hunan Province, China. The proposed model exhibits substantial improvement over the traditional Mask R-CNN, delivering accurate segmentation and precise identification of disturbed areas in construction projects.

**Keywords**—mask R-CNN, soil and water conservation, disturbed area, construction projects

## I. INTRODUCTION

The implementation of construction projects acts as a catalyst for economic growth; however, it concurrently generates inevitable disturbances to the earth's surface, leading to vegetation degradation and intensified soil erosion. In the absence of timely and effective monitoring and preventative measures, these processes contribute to soil and water losses. As urbanization accelerates, the increasing quantity and extensive scope of construction projects present novel challenges to the supervision of soil and water conservation efforts. Traditional regulatory methodologies and instruments have become insufficient to

address the evolving demands of contemporary monitoring practices.

To tackle this issue, numerous researchers have employed remote sensing technology and deep learning techniques to enhance the efficiency of supervision and management. [1] proposes a convolutional neural network (CNN) for the automatic identification of disturbed areas, which demonstrates the potential to improve work efficiency and reduce costs. [2] utilizes an object-oriented direct comparison method on high-resolution remote sensing data to extract information on construction projects. This approach effectively captures the distribution information on construction projects in the study area. [3] concentrates on image enhancement for construction projects, comparing five different fusion methods based on GF-1 satellite images. The principal components and Gram-Schmidt fusion algorithms were found to be superior for disturbance information extraction, resulting in higher extraction accuracies in various regions. [4] introduces a technical process for monitoring soil erosion caused by construction projects using remote sensing images. This process effectively interprets the activity status of construction projects, determines the compliance of perturbation conditions, and improves the efficiency of soil-water conservation supervision.

The extraction of disturbed areas in construction projects can be regarded as a task that encompasses both object detection and semantic segmentation. Mask R-CNN [5] exemplifies a compact and versatile framework for general object instance segmentation. This framework not only identifies targets within images but also yields high-quality segmentation outcomes for each distinct object. [6] introduces the Oil Well Site extraction Mask R-CNN, a modified version of the original Mask R-CNN, devised for precise oil well site extraction from multi-sensor remote sensing images. [7] concentrates on the automatic recognition of ancient loess landslides utilizing deep learning object detection methods and Google Earth images.

This research was funded in part by HBSLKY202306, 202124ZDKT31 and NSFC Grant No. 51905205.

Researchers assessed three object detection algorithms for this task, with Mask R-CNN achieving the highest accuracy. [8] presents an innovative methodology for instance segmentation in multi-channel satellite imagery, offering enhanced performance in applications such as Center Pivot Irrigation System detection and demonstrating adaptability to other remote sensing and medical imaging challenges. In [9], a region-based image retrieval system is proposed based on the local color and texture features of image subregions. [10] developed an image processing pipeline in the quantifying acne task and obtained high accuracy for validation analysis using small data sets.

Addressing the challenge of automatically extracting disturbed regions in construction projects from satellite imagery, this study introduces an improved Mask R-CNN model. By optimizing the feature extraction network structure and loss function, the model aims to improve the accuracy of object recognition and segmentation. The objective is to accurately identify and extract disturbed areas in construction projects, thereby assisting soil and

water conservation regulatory authorities in rapidly responding to monitoring tasks.

## II. METHODOLOGY

### A. Mask R-CNN

Building upon the foundation of Faster R-CNN, Mask R-CNN retains the image feature extraction component and the Region Proposal Network (RPN) component. Concurrently, it incorporates a Fully Convolutional Network (FCN) branch into the classification and bounding box regression network section to achieve pixel-level segmentation of target objects within the image to be detected.

Fig. 1 depicts the network structure of Mask R-CNN. Initially, a ResNet residual network combined with a Feature Pyramid Network (FPN) extracts features from the input image, generating multi-scale feature maps. Subsequently, these feature maps are input into the RPN to produce candidate regions. The candidate regions and feature maps are then fed into the ROI Align to obtain prediction boxes. Ultimately, through fully connected layers (FC), the prediction boxes undergo classification and regression, while a FCN generates high-quality instance segmentation masks for the detected objects.

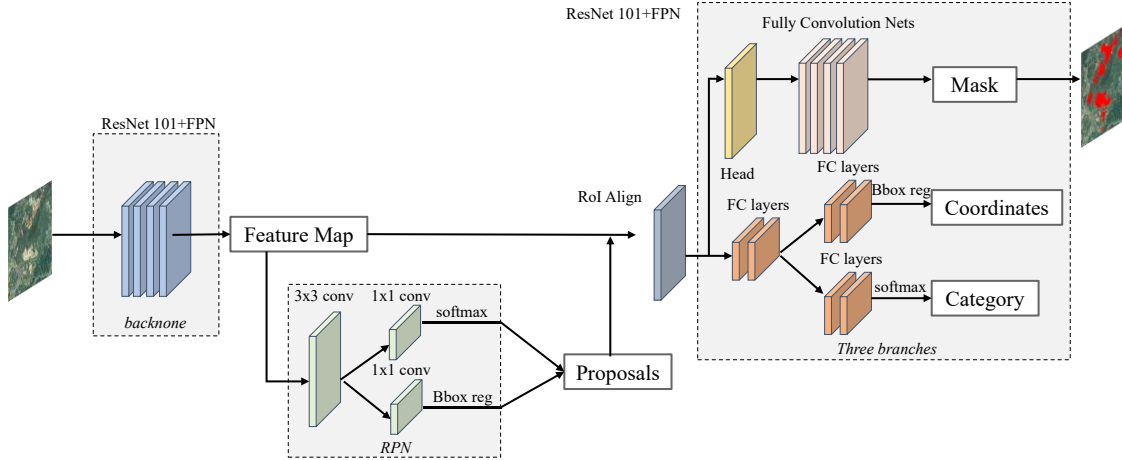


Fig. 1. Channel Attention Mechanism Structure

### B. SENet

In order to suppress complex backgrounds in remote sensing images and enhance the model's ability to detect target objects effectively, this study adopts the idea of embedding attention models into the feature extraction network. Specifically, the channel attention mechanism module, SENet [11], is embedded into the ResNet101 structure to construct the feature extraction network for Mask R-CNN. By employing the channel attention module, crucial channel features and semantic information related to the target are preserved within the feature channel domain, while suppressing other irrelevant information.

SENet consists of three components: Squeeze, Excitation, and Scale. First, the input feature map  $X$  is processed through a series of convolution operations to obtain feature map  $U$ . The feature map  $U$  undergoes Global Average Pooling to compress the features, resulting in a real-number sequence of length  $C$ , known as the squeeze operation. This allows the information

from the network's global receptive field to be utilized by lower layers. The calculation is shown in (1):

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j) \quad (1)$$

where,  $H$  and  $W$  represent the height and width of the feature map, respectively,  $u_c$  denotes the  $c$ -th channel of the feature map,  $u_c(i, j)$  represents the pixel in the  $i$ -th row and  $j$ -th column of the  $c$ -th channel, and  $z_c$  is the output of the squeeze operation.

Subsequently, the global features obtained from the squeeze operation are first passed through a fully connected layer to reduce the dimensions from  $C$  to  $1/C$ . The ReLU activation function is then applied, followed by another fully connected layer to restore the original dimension  $C$ . The Sigmoid activation function is used to obtain the weight coefficients for each channel. This series of processes is referred to as excitation. The formula is as follows:

$$s_c = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \quad (2)$$

where,  $W_1$  and  $W_2$  represent fully connected operations,  $z$  is the output of the squeeze operation,  $\delta$  denotes the ReLU activation function,  $\sigma$  represents the Sigmoid activation function, and  $s_c$  is the output of the excitation operation.

The final step is the scale operation, in which the weight coefficients obtained from the excitation operation are multiplied with the feature map  $u_c$  to recalibrate the importance of features and subsequently update the feature map. The computational formula is as follows:

$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c * u_c \quad (3)$$

where,  $s_c$  represents the weight of the  $c$ -th channel of the feature map, and  $\tilde{x}_c$  is the output of the scale operation.

In remote sensing object detection, to enhance the specificity of network computations, directly employing the SENet attention mechanism module within each residual structure of ResNet101 could lead to an increase in computational complexity and the over-processing or even loss of detailed information, resulting in suboptimal training outcomes. Therefore, this study places the SENet attention module after each group of convolutional layers with residual structures in ResNet101, specifically following the C1-C5 convolutional layers, as illustrated in Fig. 2.

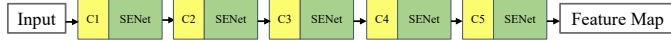


Fig. 2. The ResNet101 Network Augmented with SENet

### C. Loss Function Optimization

The loss function  $L$  of Mask R-CNN consists of three components: classification loss  $L_{cls}$ , detection loss  $L_{box}$ , and segmentation loss  $L_{mask}$ . The formula is as follows:

$$L = L_{cls} + L_{box} + L_{mask} \quad (4)$$

$L_{cls}$  employs the Softmax loss function to calculate the classification probability of the target object. The formula is as follows:

$$L_{cls} = \text{Softmax} = -\log\left(\frac{e^{f_{vi}}}{\sum_j e^{f_j}}\right) \quad (5)$$

where,  $f$  represents the score vector,  $vi$  denotes the label of sample  $i$ , and  $f_j$  represents the  $j$ -th element in the classification score vector  $f$ .

$L_{box}$  employs the  $\text{Smooth}_{L1}$  function to calculate the bounding box loss. The formula is as follows:

$$L_{box} = \text{Smooth}_{L1} = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & x < -1 \text{ or } x > 1 \end{cases} \quad (6)$$

where,  $x$  represents the input value.

$L_{mask}$  employs the Binary Cross-Entropy loss function:

$$L_{mask} = -\sum_y y \log(1 - \hat{y}) + (1 - y) \log(1 - \hat{y}) \quad (7)$$

where,  $y$  represents the expected output, and  $\hat{y}$  denotes the actual output.

The original Mask R-CNN ignores boundary information during prediction, leading to inaccurate edge detection in experiments and affecting the accuracy of the Mask. To address this issue, [12] adds an IoU Boundary loss to  $L_{mask}$ , thereby  $L_{boundary}$  optimizing the loss function for the mask portion. First, the boundary pixels of the Mask are extracted to reduce the influence of non-boundary pixels on the loss function. Then, the overlap between the true Mask boundary and the predicted Mask boundary is calculated.

The formula for  $L_{boundary}$  is as follows:

$$L_{boundary} = 1 - \frac{2|C_f \cap \hat{C}_j|}{|C_j| + |\hat{C}_j|} \quad (8)$$

where,  $|C_j|$  represents the sum of the true Mask boundary pixel intensities, and  $|\hat{C}_j|$  denotes the sum of the predicted Mask boundary pixel intensities.

The optimized loss function for the Mask portion ( $L_{mask-boundary}$ ) is as follows:

$$L_{mask-boundary} = L_{mask} + L_{boundary} = 1 - \frac{2|C_f \cap \hat{C}_j|}{|C_j| + |\hat{C}_j|} - \sum_y y \log(1 - \hat{y}) + (1 - y) \log(1 - \hat{y}) \quad (9)$$










## III. EXPERIMENT DATA, SETTINGS AND METRICS

### A. Experimental Data

The study area in this paper is Liuyang City, located in the northern section of the Luoxiao Mountains in the eastern part of Hunan Province, China. As shown in Fig. 3, the geographical coordinates are between 27°51'-28°34' north latitude and 113°10'-114°15' east longitude. The area stretches 125.8 km east to west, and 80.9 km north to south, covering a total area of 5007.75 km<sup>2</sup>. Utilizing satellite remote sensing imagery for the automatic identification of disturbed areas in construction projects can effectively assist the work of soil and water conservation regulatory agencies.

In this study, Gaofen-1 (GF-1) satellite data serves as the source of high-spatial-resolution satellite remote sensing imagery. The satellite is equipped with two 2m-resolution panchromatic and 8m-resolution multispectral cameras. We selected GF-1 satellite data encompassing Liuyang City, captured at similar times within 2020. Following preprocessing procedures such as radiometric correction, orthorectification, image fusion, image mosaicking, and cropping, we ultimately obtained a 2m-resolution satellite image covering the entirety of Liuyang City.

TABLE I. THREE PARTIAL EXTRACTION RESULTS

Number	Original Image	Label Image	Extraction Results
1			
2			
3			

Expert interpreters were assigned to delineate the construction project disturbed areas within the remote sensing imagery and corresponding vector extents. Given the heterogeneity of disturbance features and varying shapes and sizes across the selected images, including strip-like samples, the remote sensing imagery was subsequently cropped into 512 x 512-pixel samples, each encompassing labeled information. Ultimately, the acquired dataset comprised 8942 sample data entries, representing a single category, specifically the construction project disturbed areas.

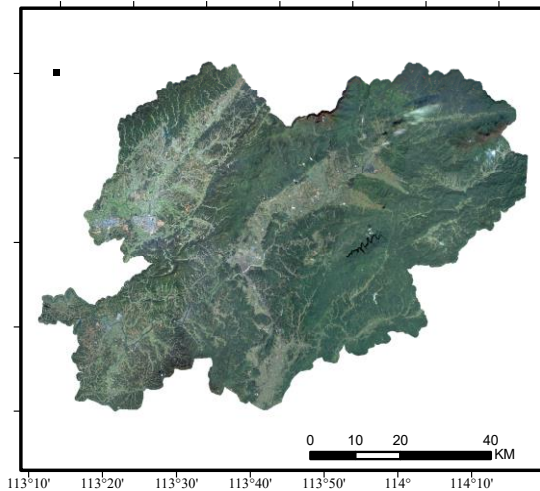


Fig. 3. GF-1 Satellite Imagery of The Study Area

### B. Experimental Platform and Settings

The experimental environment utilized the Windows 10 operating system, operating under the scientific computing integrated Python distribution Anaconda. The Keras and

TensorFlow open-source frameworks were employed, along with the CUDA-GPU acceleration scheme. The hardware specifications included an Intel Core i5-10200H CPU, NVIDIA GeForce GTX 1650 Ti GPU, 16 GB of memory, and a 250 GB solid-state drive.

Owing to the constraints imposed by the hardware, the present study utilized the hold-out method for partitioning the dataset. A total of 1000 samples were selected from the available 8942 samples to form the training and validation sets. The test set comprised samples from the remaining data, which had not been involved in training or validation processes. Apart from the dataset, other factors such as the learning rate and number of iterations were found to influence the training outcomes. The experimental setup entailed 10 epochs, with each epoch consisting of 1000 iterations. The initial learning rate was set to 0.001, the weight decay coefficient was 0.005, the momentum factor was 0.9, and the gradient clipping coefficient was 5.

### C. Evaluation Metrics

In order to reasonably evaluate the segmentation results, we employed two metrics: mean pixel accuracy (MPA) and mean intersection over union (MIoU). MPA represents the proportion of correctly predicted pixels for each class to the total number of pixels, and then computes the average value, reflecting the accuracy of the segmentation model. MIoU calculates the average value of the intersection over union ratio between the predicted results and the ground truth masks for each class. The calculation formulas are as follows:

$$\begin{cases} MPA = \frac{1}{n+1} \sum_{i=0}^n \frac{P_{ii}}{\sum_{j=0}^n P_{ij}} \\ MIoU = \frac{1}{n+1} \sum_{i=0}^n \frac{P_{ii}}{\sum_{j=0}^n P_{ij} + \sum_{j=0}^n P_{ji} - P_{ii}} \end{cases} \quad (10)$$

where,  $n$  signifies the total number of predicted categories;  $P_{ii}$  represents the count of pixels that were initially part of category  $i$  and are accurately predicted as category  $i$ ;  $P_{ij}$  corresponds to the number of pixels originally in category  $i$  but predicted as category  $j$ ; and  $P_{ji}$  denotes the count of pixels initially in category  $j$  but predicted as category  $i$ .

#### IV. RESULTS

By setting the output confidence threshold above 0.7 for target identification, Table I presents three partial segmented and recognized results. The red area represents a disturbed region in construction projects. It can be observed that the extraction results correspond well with the disturbed areas, exhibiting precise segmentation. The model accurately identifies the detailed information within the images, and the edge predictions appear relatively distinct. Due to the high degree of land use mixing in the study area, there exists a certain degree of omission in target detection as well.

To evaluate the effectiveness of the proposed improvement method in this study, two metrics, MPA and MIoU, are employed. The control group consists of the traditional Mask R-CNN, while the experimental group includes the Mask R-CNN with only SENet, the Mask R-CNN with only loss function optimization, and the Mask R-CNN with both SENet and loss function optimization. The test results are presented in Table II.

TABLE II. EVALUATION METRICS RESULTS

Method	MPA	MIoU
Mask R-CNN	83.64%	71.35%
Mask R-CNN + SENet	84.96%	72.61%
Mask R-CNN + Optimized Loss Function	84.93%	72.58%
Mask R-CNN + SENet + Optimized Loss Function	86.52%	74.17%

As evidenced by Table II, the improvement measures proposed in this study exhibit significant effects when compared to the conventional Mask R-CNN, regardless of whether a single improvement measure is adopted or both are combined.

We compared the accuracy of the proposed model with the classical semantic segmentation model on the same experimental data, which are SegNet, PSPNet and U-Net. Table III shows the comparison results, and it can be seen that the improved Mask R-CNN model has significant improvements in accuracy compared to other classical models.

TABLE III. COMPARATIVE ANALYSIS RESULTS

Method	MPA	MIoU
SegNet	82.95%	71.03%
PSPNet	83.71%	71.98%
U-Net	83.14%	71.26%
Improved Mask R-CNN	86.52%	74.17%

#### V. CONCLUSIONS

This research has demonstrated the efficacy of the improved Mask R-CNN model in automatically extracting disturbed areas from satellite imagery, paving the way for more effective soil and water conservation monitoring in construction projects. The optimizations implemented in the feature extraction network structure and loss function have proven to enhance the model's performance significantly. Future research could explore further refinements to the model and its potential applications in other domains, thereby contributing to the sustainable development and environmental preservation in the context of urbanization and infrastructure expansion.

#### REFERENCES

- [1] P. Jin, J. Huang, X. Jiang, Q. Kang, S. Yang, L. Lin, P. Yang, Z. Luo, L. Li, X. Kou, and B. Liu, "Automatic recognition and classification of construction projects' disturbed patches based on deep learning," *Science of Soil and Water Conservation*, vol. 20, no. 6, pp. 116-125, Jun, 2022.
- [2] R. Kang, M. Shi, Y. Zhao, Z. Luo, X. Wang, and E. Liu, "Extraction of Distribution Information on Production and Construction Projects in Construction Period Based on Multi-temporal GF-1 Images," *Bulletin of Soil and Water Conservation*, vol. 36, no. 3, pp. 253-257, Mar, 2016.
- [3] E. Liu, Z. Luo, X. Zhang, S. Qu, I. He, C. Zhu, and Y. Zhao, "Comparison of Fusion Algorithms for GF-1 Data from Extracted of Distribution Information on Production and Construction Projects," *Journal of Soil and Water Conservation*, vol. 32, no. 3, pp. 358-363, Mar, 2018.
- [4] T. Qiu, F. C. Lu, J. J. Zhang, and R. Wang, "Monitoring of soil erosion caused by construction projects using remote sensing images," *IOP Conference Series-Earth and Environmental Science*, 2021.
- [5] K. He, G. Gkioxari, P. Dollár, R. Girshick, and I. J. Leec, "Mask R-CNN," *IEEE International Conference on Computer Vision*, pp. 2980-2988, 2017.
- [6] H. He, H. Xu, Y. Zhang, K. Gao, H. Li, L. Ma, and J. Li, "Mask R-CNN based automated identification and extraction of oil well sites," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, Aug, 2022.
- [7] Y. Ju, Q. Xu, S. Jin, W. Li, Y. Su, X. Dong, and Q. Guo, "Loess Landslide Detection Using Object Detection Algorithms in Northwest China," *Remote Sensing*, vol. 14, no. 5, Mar, 2022.
- [8] O. F. d. Carvalho, O. A. de Carvalho Junior, A. O. d. Albuquerque, P. P. d. Bem, C. R. Silva, P. H. G. Ferreira, R. d. S. d. Moura, R. A. T. Gomes, R. F. Guimaraes, and D. L. Borges, "Instance Segmentation for Large, Multi-Channel Remote Sensing Imagery Using Mask-RCNN and a Mosaicking Approach," *Remote Sensing*, vol. 13, no. 1, Jan, 2021.
- [9] E. R. Vimina and K. Poulou Jacob, "Content Based Image Retrieval Using Low Level Features of Automatically Extracted Regions of Interest," *Journal of Image and Graphics*, Vol. 1, No. 1, pp. 7-11, March, 2013.
- [10] C. Zhang, G. Huang, K. Yao, M. Leach, J. Sun, K. Huang, X. Zhou, and L. Yuan, "A Comparison of Applying Image Processing and Deep Learning in Acne Region Extraction," *Journal of Image and Graphics*, Vol. 10, No. 4, pp. 166-171, December, 2022.
- [11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132-7141, 2018.
- [12] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pp. 6609-6617, 2017.