*Sensors and Materials*, Vol. 38, No. 7 (2026) 3713–3720
MYU Tokyo



S & M 4530

# Intelligent Extraction of Urban Vegetation Information Based on Segment Anything Model and Residual Neural Network Model

Yuan Zhuang,[1] Yi Zhang,[1*] Shuang Wu,[1,3] Haizhuo Sun,[2]
Yanfeng Xie,[1] Siyang Yin,[1] and Chunyang Cui[1]

[1]Beijing Institute of Surveying and Mapping, Beijing 100045, China
[2]Beijing Revenue Technology Co. Ltd., Beijing 100080, China
[3]Beijing Key Laboratory of Urban Spatial Information Engineering, Beijing 100038, China

(Received August 6, 2024; accepted June 3, 2026)

***Keywords:*** high-resolution remote sensing image, deep learning, urban vegetation information, SAM and ResNet model, extraction

To effectively support urban ecological environment assessment and gross ecosystem product (GEP) calculation, we used Beijing-2 and Beijing-3 satellite remote sensing images with a resolution of 0.8 m as data sources and a self-made urban vegetation sample dataset, and proposed a model combining the Segment Anything Model (SAM) and residual neural network (ResNet) for vegetation extraction. This method effectively combined the segmentation ability of SAM with the classification and extraction ability of the ResNet model. The model training results showed that the *MIoU*, *MPA*, and *ACC* of this method reached 77.27, 84.74, and 85.08%, respectively, which were slightly better than those of the ResNet model, and it can accurately segment and extract the microvegetation pattern in a complex urban environment.

## 1. Introduction

China has attached considerable importance to the construction of ecological civilization. With the continuous advancement of urbanization, vegetation resources and the ecological environment have exerted a certain influence on urban development. Urban vegetation plays a crucial role in the living environment and urban ecology. The acquisition of high-precision vegetation information not only provides data support for urban environmental evaluation and gross ecosystem product accounting, but also benefits the development and utilization of vegetation resources.[1–3] Therefore, the study of efficient, accurate, and intelligent methods for extracting urban vegetation information is of great value to scientific researchers and applications.

With the development of space optical remote sensing technology, satellite remote sensing technology had clear advantages in the continuous observation of large-scale information.[4] The traditional methods for extracting vegetation information from remote sensing images mainly include visual interpretation, pixel-based methods, and object-oriented classification

*Corresponding author: e-mail: 419992776@qq.com
https://doi.org/10.18494/SAM5284

ISSN 0914-4935 

methods.[5–7] Visual interpretation results are affected by human interference and have low efficiency and high cost. Although pixel-based methods and object-oriented classification methods have higher efficiency and learning ability than artificial visual interpretation, they are difficult to fully learn and extract information owing to the large and complex information content of high-resolution remote-sensing images. The above methods mainly suffer from the incomplete extraction of remote-sensing image information and misclassification of information with similar characteristics.

With the development of deep learning theory, convolutional neural networks (CNNs) came into wide use in remote sensing image classification because of their powerful learning ability and ability to automatically extract features from massive samples. The proposal of a series of deep learning models, such as FCN, U-Net, and ResNet, provides more accurate schemes for the intelligent extraction of urban vegetation information.[8–10] Research has revealed that urban vegetation information extraction based on a depth learning method has the advantages of high precision, high efficiency, and low cost as well as a good effect on the large-scale urban vegetation information extraction.[11] Segment Anything Model (SAM) released by Meta in April 2023 is a general model for processing image segmentation tasks.. It has good zero-shot learning ability in many segmentation tasks. This model is widely used in remote sensing image segmentation.[12,13] The ResNet model is a structure based on the CNN. While it has a conventional network structure, an identity mapping layer is further added, which can well overcome the network "degradation" problem caused by the increase in the number of layers and depth.[14] We combined the SAM and ResNet model to obtain a more efficient, accurate, and intelligent model to extract urban vegetation information, as well as to provide data support for gross ecosystem product accounting and ecological quality evaluation.

## 2. Materials and Methods

### 2.1 SAM

SAM is a general model of "dividing everything" developed by Meta, and has been used to process image segmentation tasks. SAM is combined with a CNN and transformer architecture to process images in a hierarchical and multiscale manner. The idea of prompt engineering was introduced to realize segmentation based on points, boxes, masks, and even freeform text. In addition, SAM is trained on millions of images and more than one billion masks, and can return valid segmentation masks for any prompt.

The SAM framework includes three components: image encoder, prompt encoder, and mask decoder (as shown in Fig. 1). The image encoder is a vision transformer network pretrained by a masked autoencoder, and is used for one-time image embedding. The prompt decoder is mainly used for the prompt embedding of points, boxes, or text. The mask encoder can effectively map the image embedding, prompt embedding and output tags to masks, and calculate the effective probability of target object masks. Although SAM has good zero-shot generalization ability in many segmentation tasks and can segment arbitrary targets, some studies have shown that the segmentation ability of SAM declines significantly for images with irregular shapes, fuzzy

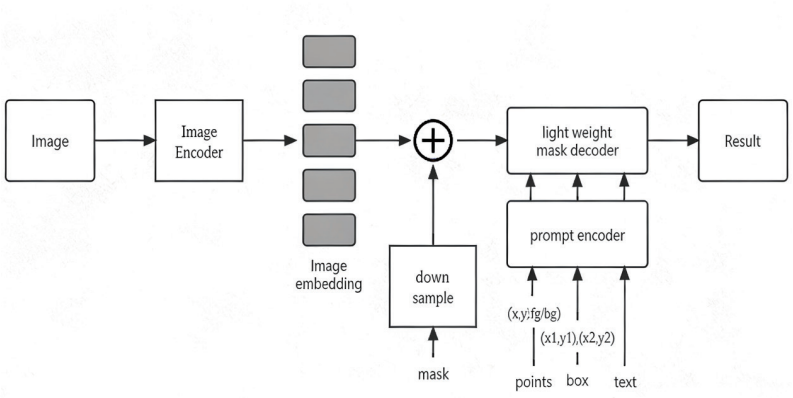


Fig. 1. SAM structure diagram.

boundaries, small size, and low contrast, mainly owing to the lack of similar image samples during SAM training.[15]

## 2.2 ResNet Model

The ResNet model is a deep neural network structure proposed by Microsoft's AI team in 2015 and is widely used in computer vision tasks such as remote sensing image classification, target detection, and semantic segmentation.[16,17] Different from traditional deep neural network structures, residual blocks were introduced in the network, where a skip connection was added between the input and output. This design made it easier for the network to learn residuals, thereby solving the problems of gradient vanishing and model degradation.

Figure 2 shows the basic residual module of the ResNet model, where $x$ and $H(x)$ are the input and output of the bottleneck residual unit, respectively; $F(x)$ is the output obtained by cascading convolution through the residual unit input, and the output of the residual unit is calculated as Eq. (1). When the learning ability of the network model reaches its optimal level, if the network continues to deepen its training, the $F(x)$ value would be trained to 0, leaving only the identity mapping $H(x) = x$ whereby the network is in its optimal state.

$$H(x) = F(x) + x \tag{1}$$

## 2.3 Accuracy evaluation indicators

In this work, mean intersection over union (*MIoU*), mean pixel accuracy (*MPA*), and accuracy (*ACC*), which are commonly used in semantic segmentation, were used as precision evaluation indicators of green vegetation information extraction results.

*MIoU*, defined as the average value of IoUP across all categories, is a metric used to evaluate the overall classification accuracy of the model. The value range of *MIoU* is 0–1, and the closer the value is to 1, the higher the accuracy of the model.

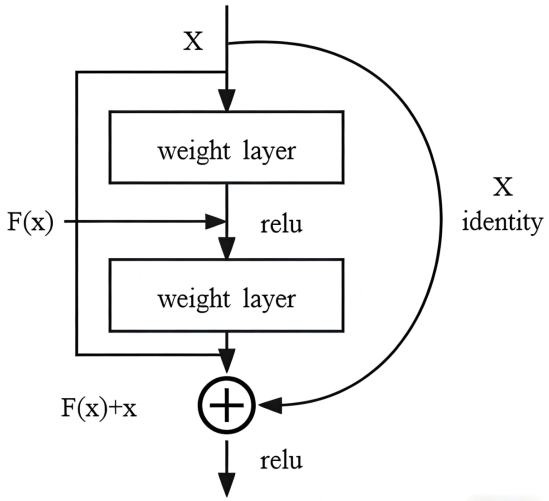


Fig. 2. Basic residual module.

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k} IoU_i \tag{2}$$

*MPA* is the average of the proportion of correctly classified pixels in each category.

$$MPA = \frac{1}{k+1}\left[\frac{TP}{(TP+FN)} + \frac{TN}{(TN+FP)}\right] \tag{3}$$

*ACC* is the ratio of correct prediction results to the total number of predicted values.

$$ACC = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{4}$$

In Eq. (4), true positive (*TP*) indicates that the prediction is true and the actual result is also true; false positive (*FP*) indicates that the prediction is true but the actual result is false; false negative (*FN*) indicates that the prediction is false but the actual result is true; true negative indicates that the prediction is false and the actual result is also false. *K* is the number of segmentation categories (excluding background).

### 2.4 Model training environment and samples

The experimental environment was configured with an Intel Xeon 24-core processor running at 2.3 GHz, an NVIDIA GeForce RTX 3090 graphics card, 64 GB of RAM, and the Windows 10

Pro operating system. An automatic classification network was built using the Pytorch deep learning framework.

The model training sample set consisted of 17420 training sets and 4355 validation sets. The size of images and tags in the vegetation sample set was 200 pixels × 200 pixels. The initial learning rate of training settings was 0.001, and the batch size was 128. The learning rate determined the step size for parameter updates in each iteration, guiding the model to descend along gradients until it reached the optimal state. Batch size represented the number of samples used to update the weight value in one iteration. The number of iterations represented the number of times the training set samples were completed.

## 3. Study Area and Data

### 3.1 Study area

Beijing is a megacity located in the northwest of the North China Plain. It has a unique topography that is high in the northwest and low in the southeast, naturally separating the mountains from the plain. With the vigorous development of the capital's economy, the challenges and problems of very large cities have become increasingly prominent. To comprehensively build a model of a modern city with harmonious coexistence between human and nature, Beijing upheld the core concept of "the integration of city and garden, harmony between man and city" and intensively promoted the strategy of garden city construction. Through the addition of green microspaces, urban green coverage was gradually increased, which steadily and effectively alleviated and cured the various "diseases" unique to megacities. We selected the central urban area of Beijing as the research area, and conducted research and analysis of the characteristic scenes of megacities.

The sources of the main research data of this study were the domestic Beijing 2 and Beijing 3 satellite multispectral images. The images were taken in September 2023; the cloud cover in the experimental area was less than 5%. The image data included a 0.8-m-resolution panchromatic image and 3.2-m-resolution multispectral image. The 0.8-m-resolution multispectral image was synthesized through radiometric calibration, atmospheric correction, ortho-correction, image fusion, and other forms of preprocessing.

### 3.2 Urban vegetation sample set

The urban vegetation sample set was composed of a group of corresponding images and tags and was used to train and verify the model and calculate the accuracy of the model classification. In this study, the labels of the vegetation sample set were made and established by means of manual visual interpretation. First, with Beijing 2 and Beijing 3 images, ArcGIS was employed to manually interpret the vegetation in the experimental area and draw its vector boundary. Then the manually interpreted vegetation vector boundary was modified and corrected in combination with field survey and verification. Finally, the results were exported in the form of grid images as real label data, where the white area represented urban vegetation and the black area represented the background of other land features. The urban vegetation sample set is shown in Fig. 3.

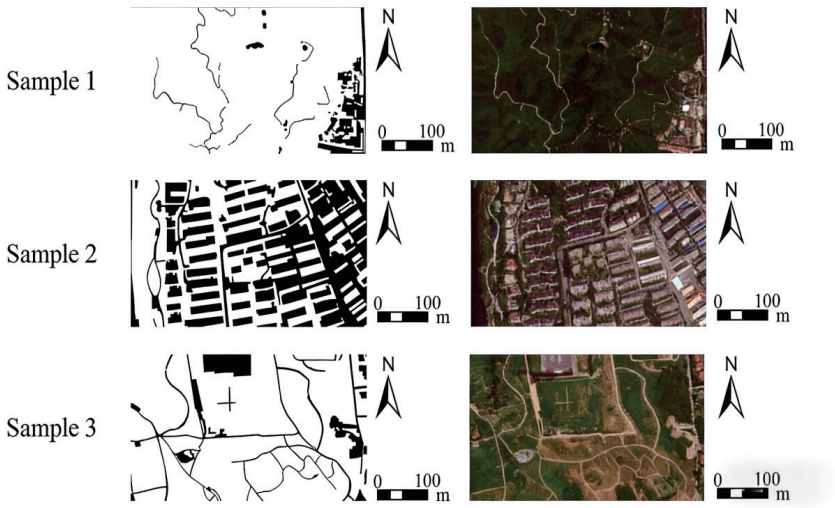


Fig. 3. (Color online) Schematic diagram of sample set.

## 4. Results

### 4.1 Comparison of model extraction accuracies

We input the training sample set into the ResNet and SAM+ResNet models for model training. Then we fed the validation set images and corresponding labels into the trained models to extract urban vegetation. The extraction accuracy (Table 1) showed that compared with the ResNet model, the accuracy of the SAM+ResNet model was slightly better. *MIoU* was increased by 1.63 percentage points, MAP by 1.76 percentage points, and *ACC* by 1.39 percentage points.

### 4.2 Extraction results of urban vegetation validation sample set

Three representative images of the experiment area were selected, containing vegetation information such as forest land, grassland, and road greening, as well as nonvegetation information such as buildings, roads, and squares, with the characteristics of the complex information of megacity application scenarios. Figure 4 shows the results of three image areas in the model comparison experiment. In image 1, the ResNet model misidentifies the hardened path in a small urban park as vegetation, whereas SAM+ResNet can accurately extract it and establish the boundary, but the shadow area of buildings was mistakenly labeled in both methods. In image 2, ResNet mistakenly classified part of the water in the image as vegetation, and the green microspaces interspersed between buildings in the upper right corner of the image were not subdivided. By contrast, SAM+ResNet better handled these two types of problem, and the overall extraction effect of vegetation information had better similarity to that of the labels. Image 3 shows a complex environment on the edge of the city. Compared with the ResNet model, the SAM+ResNet model accurately segmented scattered and small areas of nonvegetation, and the vegetation identification pattern was consistent with the real value.

Although the accuracy of the SAM+ResNet model was not significantly better than that of the ResNet model, the segmentation results of the ResNet model were relatively poor for the complex

Table 1
Accuracies of ResNet and SAM+ResNet models.

| | *MIoU* (%) | *MPA* (%) | *ACC* (%) |
|---|---|---|---|
| ResNet | 75.64 | 82.98 | 83.69 |
| SAM+ResNet | 77.27 | 84.74 | 85.08 |

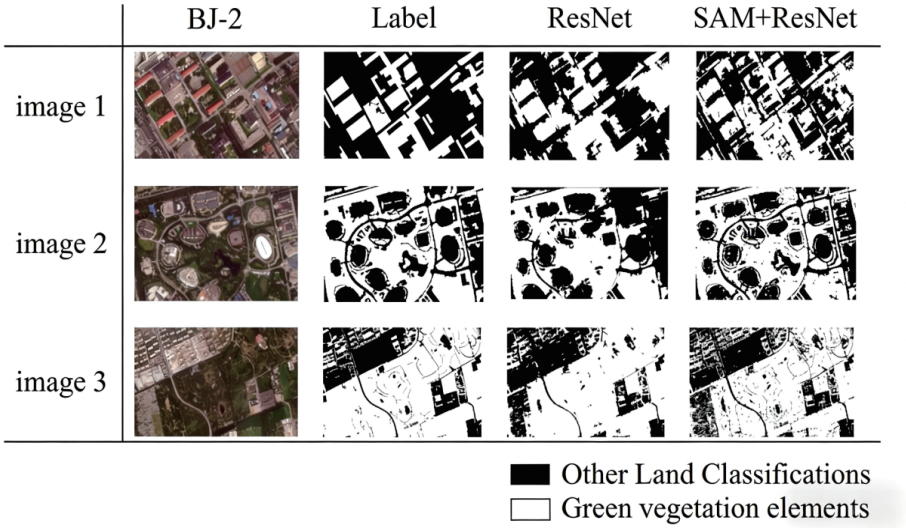


Fig. 4. (Color online) Vegetation information extraction results.

scenes of megacities, and many mini–vegetation polygons were difficult to accurately segment and extract. SAM+ResNet could effectively solve this problem and extract subtle vegetation information in building areas.

## 5. Conclusions

We combined the segmentation ability of SAM with the classification extraction ability of the ResNet model, aiming to explore an efficient and accurate method for identifying urban vegetation. We then analyzed and discussed the research results.

(1) Compared with using the ResNet model alone for urban vegetation extraction, the combined SAM+ResNet model showed minor improvements in *MIoU*, MAP, and *ACC*. Through the analysis of the classification results of specific application scenarios, we found that SAM+ResNet can accurately segment and extract small vegetation patches in complex environments. This provided strong support for verifying the practical effectiveness of ecological strategies such as green microspaces in very large cities, demonstrating the enormous potential of the SAM and ResNet model in promoting the refined management and evaluation of urban green spaces.

(2) In this study, we prepared our own vegetation sample set data that was manually visually interpreted and inevitably had some manual errors that can affect the accuracy of the model. Differences in the quality of the sample datasets can also result in variations in model accuracy. Therefore, a large number of more refined sample datasets will be needed in a later study of deep learning models.

(3) Although some progress was made, this study also had certain limitations, In particular, when dealing with vegetation under building shadows in remote sensing images, it was still difficult for our model to accurately identify and extract such vegetation. This challenge highlighted the importance of future research in enhancing the model's ability to detect vegetation under complex lighting conditions. We look forward to further improving and perfecting our model by introducing more advanced shadow detection algorithms, optimizing feature extraction mechanisms, and integrating multisource data to overcome the current limitations.

In summary, the SAM+ResNet model provides valuable data support and a decision-making basis for urban ecological environment assessment and construction owing to its accuracy, transferability, and universality in extracting urban vegetation information. It not only effectively responds to diverse urban scenarios, but also opens up new paths for the intelligent and refined monitoring and management of urban vegetation.

## Acknowledgments

The publication was supported by the ministry-province cooperation project with the grant number 2024ZRBSHZ085.

## References

1 L. Chen and F.F. Sun: Ecological Science. **42** (2023) 169. https://doi.org/10.14108/j.cnki.1008-8873.2023.05.020
2 X. J. Mou, X. H. Wang, X. Zhang, and Z. X. Zhu: Res. Soil Water Conserv. **27** (2020) 265. https://doi.org/10.13869/j.cnki.rswc.2020.01.037
3 J. X. Gao, H. W. Wan, Y. C. Wang, P. Hou, P. R. Shi, and C. X. Sun: Nat. Remote Sensing Bull. **12** (2023) 2860. https://doi.org/10.11834/jrs.20221186
4 L. Q. Yang, K. Jia, S. L. Liang, X. Q. Wei, Y. J. Yao, and X. T. Zhang: Remote Sens. **9** (2017) 857. https://doi.org/10.3390/rs9080857
5 K. Zhou, Y. Q. Yang, Y. N. Zhang, R. Miao, Y. Yang, and L. Liu: Sci. Technol. Eng. **32** (2021) 13603. https://doi.org/10.3969/j.issn.1671-1815.2021.24.001
6 D. D. Zhang and X. M. Wang: For. Grassland Resour. Res. **3** (2021) 108. https://doi.org/10.13466/j.cnki.lyzygl.2021.03.017
7 C. F. Li, J. Y. Yin, and J. J. Zhao: Sci. Surv. Mapp. **36** (2011) 112. https://doi.org/10.16251/j.cnki.1009-2307.2011.05.004
8 H. Y. Ma, T. Y. Zhang, Q. L. Dai, F. Dai, and L. G. Wang: J. Southwest Forestry University (Natural Sciences) **3** (2019) 117. https://doi.org/10.11929/j.swfu.201903111
9 N. Lin, J. He, B. Wang, F. F. Tang, J. Y. Zhou, and J. Gou: Geo-Inf. Sci. **8** (2023) 1717. https://doi.org/10.12082/dqxxkx.2023.220866
10 L. Shi: Central South University of Forestry and Technology (2022). https://doi.org/10.27662/d.cnki.gznlc.2022.000120
11 G. Q. Men, G. J. He, and G. Z. Wang: Forests **12** (2021) 1441. https://doi.org/10.3390/f12111441
12 S. Li, Q. F. Chu, R. J. Yuan, Y. He, K. S. Zhang, and J. M. Li: Geomatics Spatial Inf. Technol. **S1** (2024) 1. https://doi.org/10.11834/jig.240540
13 L. B. Zhang: Railway Invest. Surv. **3** (2024) 21. https://doi.org/10.19630/j.cnki.tdkc.202310180002
14 F. He, T. Liu, and D. Tao: IEEE Trans. Neural Networks Learn. Syst. **12** (2020) 5349. https://doi.org/10.1109/TNNLS.2020.2966319
15 Y. H. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. R. Zhou, R. S. Chen, J. X. Yu, J. Q. Chen, C. Y. S. J. Liu, H. Z. Chi, X. D. Hu, K. J. Yue, L. Li, V. Grau, D. P. Fan, F. J. Dong, and N. Dong: Med. Image Anal. **92** (2024) 1361. https://doi.org/10.1016/j.media.2023.103061
16 J. F. Zhu, M. H. Zhang, C. Ding, J. H. Luo, and Y. Gu: Comput. Digital Eng. **2** (2023) 479.
17 C. Q. Shen, K. Wang, and W. J. Wang: Geospatial Inf. **6** (2023) 21.