S & M 3722

Remote Sensing Image Recognition of Dust Cover Net Construction Waste: A Method Combining Convolutional Block Attention Module and U-Net

Shangwei Lv,[1*] Xiaoyu Liu,[2] and Yifei Cao[3]

[1]WuHan University, 129 Luoyu Road, Wuhan 430079, China
[2]Beijing Urban Construction Survey, Design and Research Institute Co., Ltd.,
No. 15, Yongyuan Road, Daxing District, Beijing 102616, China
[3]Beijing Institute of Surveying and Mapping, No. 15, Yangfangdian Road, Haidian District, Beijing 100038, China

(Received June 10, 2024; accepted July 16, 2024)

Keywords: construction waste, attention mechanism, U-net, CBAM, semantic segmentation

With the acceleration of urban development, the annual production of urban construction waste has been increasing yearly, which brings considerable challenges for urban supervision and management, and how to quickly and accurately identify construction waste is of great practical significance. In this paper, we propose a remote sensing image dust cover net construction waste recognition algorithm based on the improved U-network model to realize construction waste target recognition. The algorithm first prepares a dust cover net construction waste identification dataset using Google high-resolution remote sensing imagery as the database. Second, VGG16 is adopted as the backbone network of the U-Net model to improve the feature expression ability of the model. Finally, the Convolution Block Attention Module (CBAM) is embedded into the U-Net network to construct the CBAM-U-Net model to enhance the information extraction accuracy of high-resolution remote sensing images. With the remote sensing image encompassing Daxing District in Beijing as an example, the results show that the proposed algorithm can automatically and efficiently recognize the dust cover net construction waste with 95.51% recognition accuracy and 95.08% MIou, which puts forward a new idea for the supervision of construction waste.

1. Introduction

In recent years, with the rapid development of industrialization and urbanization, there has been a rapid increase of urban construction waste. According to statistics, the average annual production of construction waste in countries around the world has exceeded 2 billion tons, accounting for more than half of all waste production. The identification and disposal of construction waste is gradually becoming a serious problem for countries around the globe. Construction waste mainly includes five categories: engineering soil, engineering mud, demolition waste, and decorative waste.(1) The efficient and accurate monitoring of construction

[*]Corresponding author: e-mail: lswliesmars@whu.edu.cn

https://doi.org/10.18494/SAM5182

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

waste in the city can not only provide data support for improving the regulatory capacity of construction waste and subsequent construction waste treatment, but also be of great significance for urban governance and the improvement of the living environment for residents.

There are currently several methods for monitoring construction waste: (1) Field research—this method is costly and has low statistical efficiency; (2) Regional regulatory laws—urban streets are divided into regions to monitor construction waste, which is difficult to implement; (3) Visual interpretation—manually determining construction waste based on the knowledge and experience of professional technicians, depending on their expertise; (4) Machine learning decoding methods—by combining aerial and satellite data, traditional machine learning methods are used to extract construction waste, but most of these algorithms can only be applied to a single scene.

Remote sensing technology is characterized by real-time, high-resolution, and rich acquisition of information, and has been widely used in urban planning, disaster management, climate analysis, and other fields.(2) Currently, multiple types of remote sensing images are freely available to scholars, such as Landsat data, MODIS data, Sentinel-2 data, and Google images. Remote sensing imagery can provide a new way to monitor construction waste quickly, objectively, and dynamically.

With the rapid development of artificial intelligence and computer technology, deep learning methods are widely used in the fields of computer vision, natural language processing, and audio recognition.(3) They can learn expressions from underlying features to high-level semantic features based on the regularity of the samples. As a key task in the field of deep learning, semantic segmentation has become an important tool for information extraction from remote sensing images. Combining remote sensing images and deep learning semantic segmentation algorithms can realize different types of objects in the images to achieve pixel-level labeling classification. Convolutional neural networks (CNNs) dominate semantic segmentation owing to their feature extraction and representation capabilities. The common CNN model structure consists of an input layer, a convolutional layer, a pooling layer, a fully connected layer, and an output layer.(4) However, research on construction waste is limited, and no large, publicly available dust cover net construction waste dataset exists.

In response to the above research status, the research topics in this paper are as follows.

(1) Construction of a dust cover net construction waste dataset with high-resolution remote sensing imagery

Google image is used as a data source to construct a high-resolution construction waste dataset.

(2) Automated construction waste identification model

On the basis of the U-Net model, a Convolution Block Attention Module (CBAM) attention mechanism is embedded in the jump connection, and a CBAM-U-net model is innovatively proposed to enhance the feature extraction capability of the network and improve the accuracy of the information extraction of high-resolution remote sensing images.

(3) Model Comparison and Outcome Rating

The Daxing District of Beijing was used as the study area to verify the modeling effect.

2. Related Work

Semantic Segmentation of Remote Sensing Images: The semantic segmentation of remote sensing images initially relied on traditional image processing and machine learning methods. Common techniques include threshold-based segmentation methods, such as Otsu's method, edge detection method, region growing, and watershed algorithm. A major limitation of traditional methods is their limited segmentation ability for complex scenes, making it difficult to handle high-dimensional and multiscale features in remote sensing images. With the advancement of machine learning, supervised learning algorithms such as support vector machine (SVM) and random forest (RF) have been applied to the semantic segmentation of remote sensing images. These methods segment images by training classifiers. Although machine learning methods have improved the accuracy of segmentation to a certain extent, their performance depends on the quality of feature engineering and has a low efficiency in processing large-scale remote sensing data. In recent years, deep learning methods have developed rapidly and achieved significant success in the semantic segmentation of remote sensing images. Fully convolutional networks (FCNs) are the first results of pioneering work to apply deep learning to semantic segmentation.(5) Subsequently, U-Net with an encoder–decoder structure that performed well in medical image segmentation was proposed and quickly applied in the field of remote sensing images. The success of U-Net lies in its skip connections, which combine low-level features with high-level features, thereby improving segmentation accuracy.(6) SegNet and DeepLab series are two other deep learning models widely used in semantic segmentation. SegNet reduces computational complexity and improves efficiency by introducing maximum pooling indexes. DeepLab performs excellently in improving resolution and boundary accuracy by introducing dilated convolution and conditional random field (CRF).(7,8) In addition, PSPNet performs well in remote sensing image segmentation tasks by introducing a Pyramid Pooling Module to capture multi-scale contextual information. Recently, the Transformer architecture has also been explored in remote sensing image segmentation.(9) Vision Transformer (ViT) and SegFormer have shown potential in capturing long-range dependencies by introducing self-attention mechanisms. In addition, multimodal methods such as fusing optical images and SAR data have also been proven to improve segmentation performance. These methods improve the robustness and accuracy of segmentation by combining the data of different sensors and using their complementary information.(10,11)

U-Net: U-Net was originally proposed by Ronneberger *et al.*(6) to address the limitations of FCN in biomedical image segmentation. The key innovation of U-Net lies in its symmetrical encoder–decoder structure and skip connections. Jumping connections allow the network to combine high-resolution features in the encoder with upsampled features in the decoder, thereby preserving spatial information and achieving precise localization. The architecture of U-Net has performed well in medical image segmentation and achieved state-of-the-art results in various challenges. After the success of U-Net, many variants and improved versions were proposed. 3D U-Net extends the original 2D architecture to 3D, achieving volume segmentation of medical images.(12) Other enhanced versions, such as ResUNet-a and DenseUNet, respectively introduce residual connections and dense blocks to further improve feature propagation and network training. In addition to U-Net and its variants, other architectures have also driven the development of semantic segmentation.(13,14)

Attention: With the rapid development of deep learning, CNNs and recurrent neural networks (RNNs) have achieved significant success in Computer Vision (CV) and natural language processing (NLP) tasks. However, these models still have limitations in capturing long-range dependencies and global information. To overcome the above limitations, Bahdanau *et al.* first introduced the attention mechanism in neural machine translation.(15) The attention mechanism allows the model to dynamically focus on different parts while processing input sequences, thereby better capturing global dependencies. This mechanism significantly improves translation quality by computing context vectors.(15) Vaswani *et al.* proposed the transformer model, which incorporates the self-attention mechanism as its core component.(16) The transformer model abandons traditional RNN structures and relies entirely on self-attention mechanisms and feedforward neural networks. The self-attention mechanism can directly model the dependency relationship between any two positions in a sequence, thus having significant advantages in capturing long-range dependencies and parallel computing. The transformer model has achieved state-of-the-art performance in machine translation tasks and has become the standard architecture for many NLP tasks. Inspired by the success of the transformer model in the field of NLP, researchers have begun to apply it to CV tasks.(16) ViT divides images into a series of fixed size patches and treats them as sequential data input into the transformer model. ViT has achieved significant results in image classification tasks, demonstrating the potential of the transformer model in the field of CV. On the basis of the successful transformer and attention mechanisms, many variants and improved versions have been proposed.(10) For example, DeiT made the transformer model more efficient in image classification tasks by introducing efficient training strategies and distillation methods.(17) Swin Transformer proposed a hierarchical self-attention mechanism that enables transformers to better handle features at different scales.(18) In addition, researchers have also explored combining the attention mechanism with traditional models such as CNN and RNN to further improve performance. The success of the attention mechanism and transformer in multiple fields has sparked widespread research interest. Researchers not only continuously optimize and expand these models in NLP and CV tasks, but are also beginning to explore their applications in multimodal learning, reinforcement learning, and generative models. For example, BERT has made breakthroughs in pretraining language models, and the GPT series models have performed well in generating text. Meanwhile, the attention mechanism has also shown great potential in image generation, video understanding, and cross modal tasks.(19)

3. Methodology

On the basis of the U-net model, we used the VGG16 model as the backbone feature extraction network, embedding the channel attention mechanism and spatial attention mechanism into the U-net model to improve the feature expression ability of convolutional neural networks. By embedding CBAM attention modules in the skip connections of the encoding and decoding U-net network, a CBAM U-net model is proposed to segment dust cover net construction waste, solving the problem of poor feature extraction ability and segmentation effect of the U-net. The overall technical process of this article is shown in Fig. 1.

Introducing the attention mechanism into neural networks to improve the model effectively enhances its feature expression ability and the accuracy of semantic segmentation. The essence of the attention mechanism is to use neural networks to give more attention to useful information, ignore or suppress useless information in the scene, and obtain more global contextual information. The CBAM is a lightweight attention module that can be embedded into any backbone network to improve performance and achieve plug and play. CBAM includes two submodules, the channel attention module (CAM) and the spatial attention module (SAM), which concatenate and combine the channel dimension and spatial dimension to add attention. The CBAM structure diagram is shown in Fig. 2.

The CAM is designed to enhance the effect of certain channel images while weakening the effect of other channels. The input feature map F first obtains two feature maps through max pooling and average pooling, which are used to reduce the information loss caused by pooling. Next, we input both feature maps simultaneously into a weight-shared multilayer perception network (MLP) for dimensionality reduction and dimensionality enhancement operations. Then,

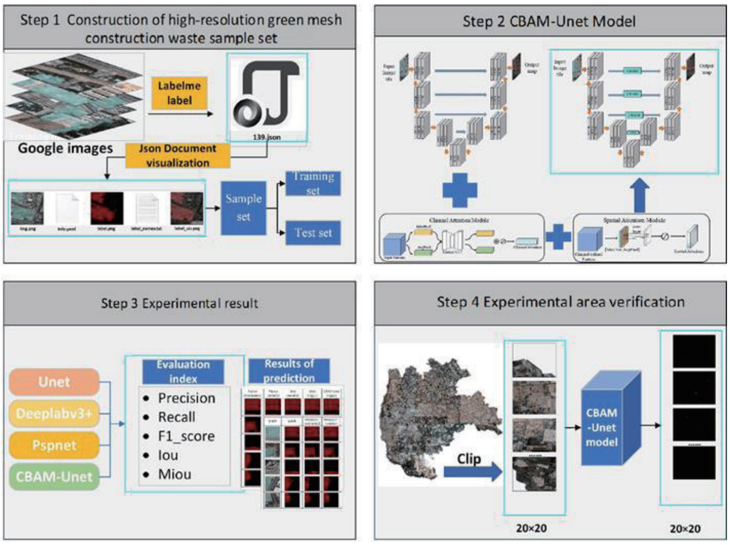


Fig. 1. (Color online) Technical flowchart.

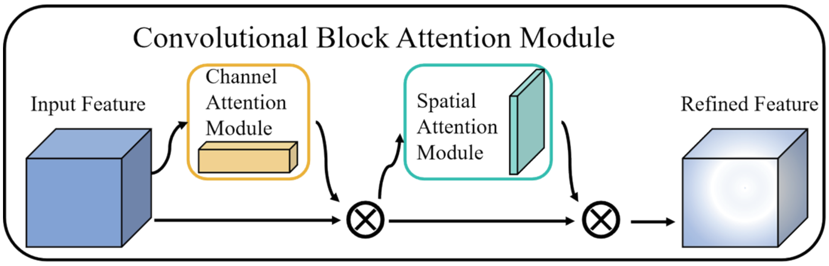


Fig. 2. (Color online) CBAM structure diagram.

the two feature maps output from MLP are added pixel by pixel and processed by the sigmoid activation function to obtain the feature map of the CAM. The feature map of the CAM can be expressed as

$$M_c\left(F\right)=\sigma\left\{MLP\left[AvgPool\left(F\right)\right]+MLP\left[\mathrm{M}axPool\left(F\right)\right]\right\}. \tag{1}$$

In the equation, $M_c(F)$ is the feature map of the channel attention module; σ is the sigmoid activation function; *MLP*( ) is a multilayer perceptron; *AvgPool*( ) and *MaxPool*( ) represent average pooling and maximum pooling, respectively. The structure diagram of the CAM model is shown in Fig. 3.

The SAM focuses on positional information in images, enhancing the correlation between similar feature channels across the entire feature map. We use average pooling and maximum pooling to obtain the feature map of the channel attention module input. Then, on the basis of channel concatenation, we use a convolution operation with a size of 7 × 7 to increase the number of channels to 1. Finally, we generate the SAM feature map through the activation function sigmoid. The SAM can be represented as

$$M_s\left(F'\right)=\sigma\left\{f\left[\mathrm{A}vgPool\left(F'\right);MaxP\mathrm{ool}\left(F'\right)\right]\right\}. \tag{2}$$

In the equation, $M_s(F')$ is the spatial attention module; f is a convolutional layer operation. The structure diagram of the SAM model is shown in Fig. 4.

In CBAM, the input feature map F is first dot multiplied with the channel attention feature map $M_c(F)$ to obtain the feature map F'. F' is used as the input feature map and multiplied with

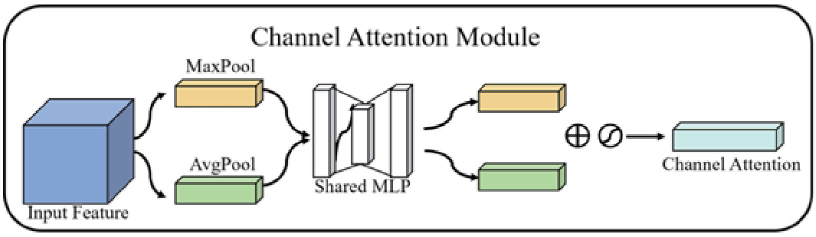


Fig. 3. (Color online) CAM structure diagram.

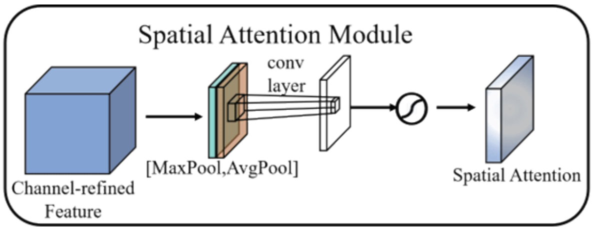


Fig. 4. (Color online) SAM structure diagram.

the spatial attention feature map $M_S(F)$ to obtain the final feature map F'' of the overall CBAM. The calculation equation is as follows.

$$F' = M_c\left(F\right)\otimes F \tag{3}$$

$$F'' = M\ \left(F'\right)\otimes F' \tag{4}$$

In this study, we added a CBAM module to the U-net model's skip connection layer to improve the network efficiency. The model in this study consists of four encoding blocks, four decoding blocks, and four layers of CBAM attention modules. The structure diagram of the CBAM U-net model is shown in Fig. 5.

4. Experimental Results and Analysis

4.1 Experimental data

Deep learning models require large-scale samples. Considering that there is currently no dataset for semantic segmentation of construction waste, we analyzed the characteristics of construction waste on high-resolution remote sensing images and created a dust cover net construction waste sample set.

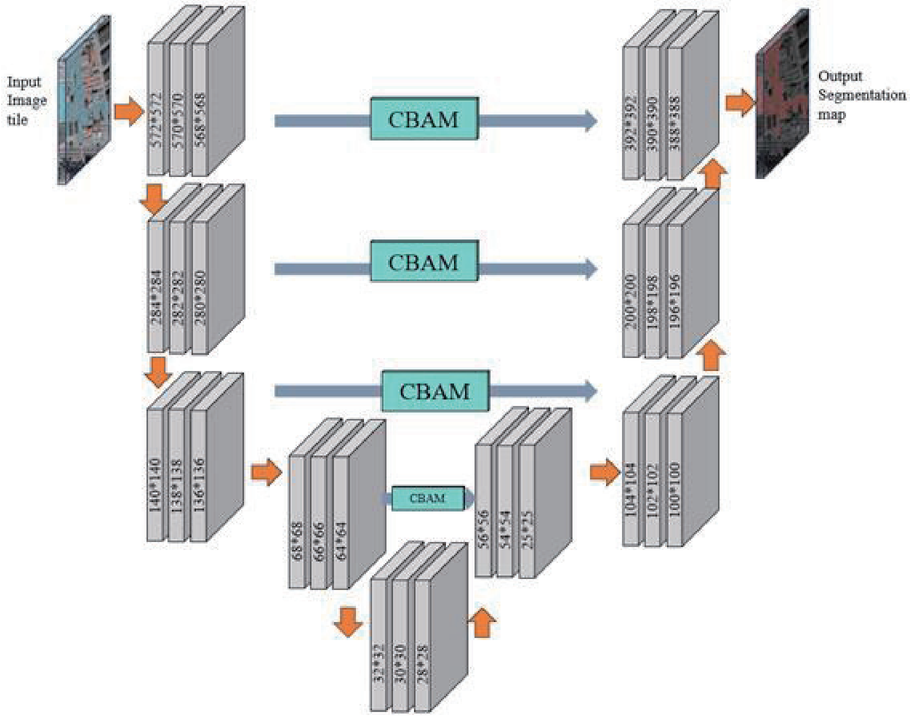


Fig. 5. (Color online) Structure diagram of CBAM U-net model.

We collected construction waste samples from different cities in China in different years on Google images, with a spatial resolution of 0.3 m. There are a total of 1000 images in the red, green, and blue bands. We cropped each image in batches to a size of 512 × 512 pixels. Each image contains varying amounts of construction waste with a dust cover net. The data at this resolution can clearly display the form of construction waste, which is helpful for feature extraction in network models.

The construction waste with a dust cover net is mainly green in the remote sensing image, and gray, brown, white, and brown spots appear below the dust cover net. The shape of construction waste has no specific pattern, is complex and diverse, and mainly manifests as piles concentrated in one area, disorderly, and uneven. This article is based on the inconsistent expression of texture, color, and shape characteristics of different types of construction waste. For dust cover net construction waste, the Labelme software is used to visually interpret and manually annotate the dust cover net construction waste dataset, which is divided into two categories: Green-CDW and background. The construction waste images with a dust cover net are true positive samples, while the other features are background negative samples, with a true value of 1 and a background value of 0. Original and label images are shown in Fig. 6. The ratio

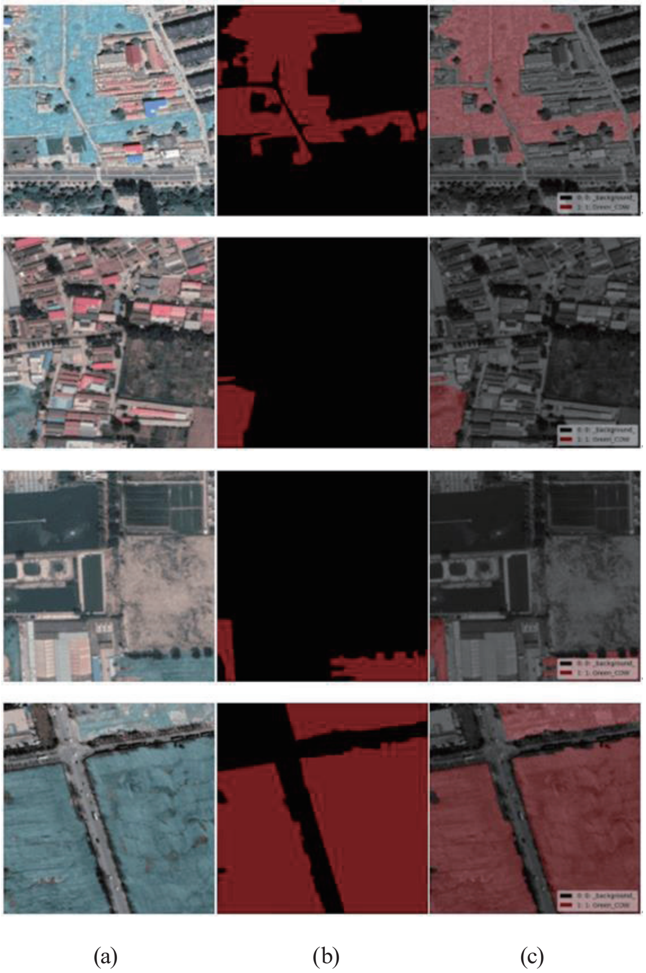


Fig. 6. (Color online) Annotation image of construction waste sample set: (a) Original image, (b) label image, and (c) visualized image.

of training and testing sets for the dust cover net construction waste dataset is 8:2, with 800 training sets and 200 testing sets.

4.2 Test results and analysis

The experimental research area is located in Daxing District, southern Beijing, with a total area of 1036.33 square kilometers ranging from 39°26‘–39°50’N north latitude and 116°13’–116°43’E. The image size is 190732 × 145046 pixels. The verification area is shown in Fig. 7.

We divided a TIF image into 20 rows and 20 columns, with a total of 400 small images. The overlap between adjacent images is 0, and then the model predicts it. After the prediction is completed, the predicted results are concatenated. By taking two dense construction waste sites as examples, the original image and predicted image are as shown in Fig. 8.

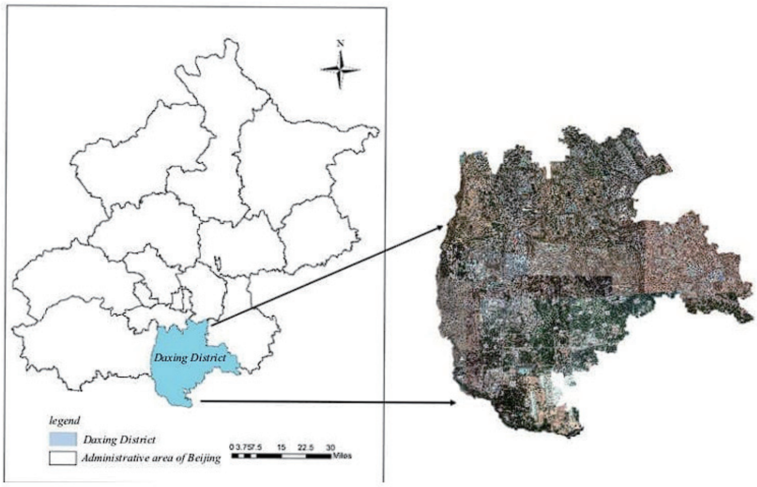
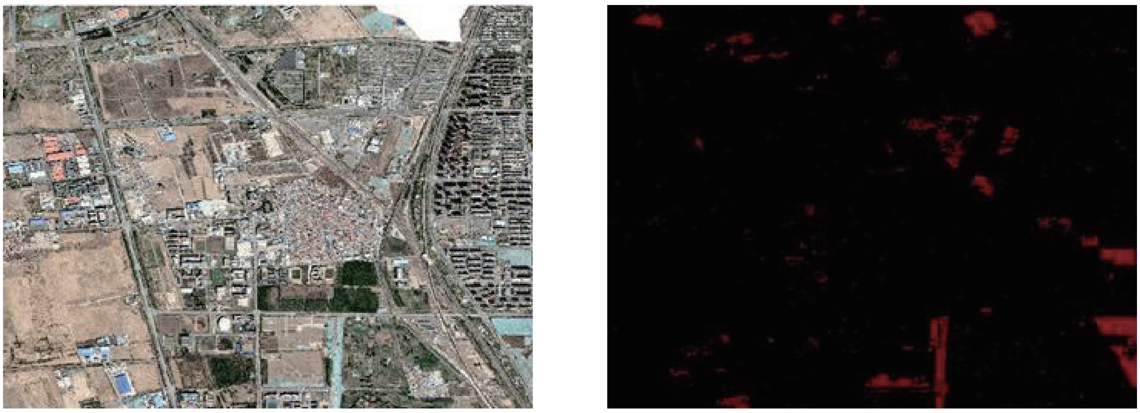


Fig. 7. (Color online) Schematic diagram of research area.



(a) (b)

Fig. 8. (Color online) CBAM U-net model for predicting construction waste in Daxing District.

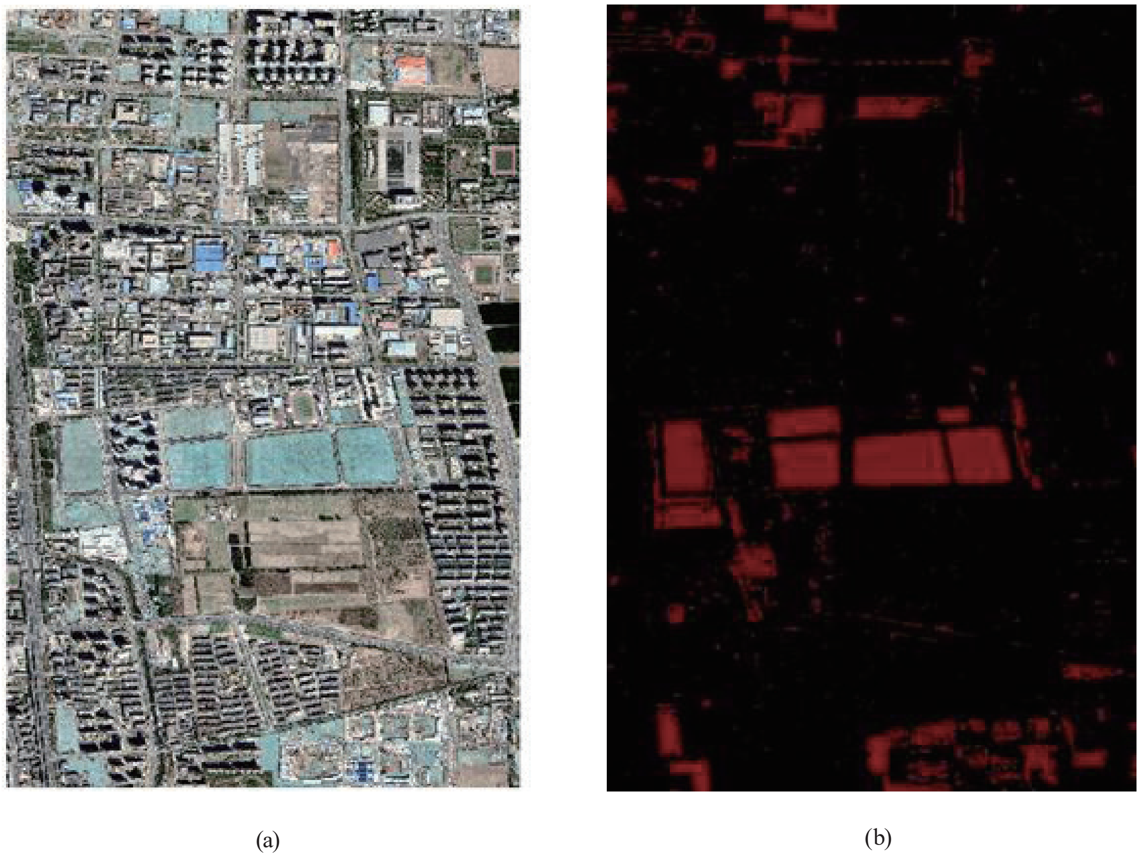


Fig. 8. (Color online) (continued) CBAM U-net model for predicting construction waste in Daxing District.

4.3 Comparison and analysis of algorithm performance

The accuracy evaluation in this study is visualized using a confusion matrix, and the final result obtained is a binary classification, namely, construction waste and background nonconstruction waste, as shown in Table 1. Among them, *TP* represents the correctly predicted number of pixels correctly predicted as dust cover net construction waste; *FP* represents the number of pixels of the background predicted to be dust cover net construction waste; *FN* represents the number of pixels of construction waste predicted to be the background; *TN* represents the number of pixels correctly predicted as the background.

The quantitative evaluation indicators used in this study include precision, recall, $F1\_score$, pixel accuracy, and mean intersection over union. Accuracy is an indicator that represents the accuracy of the model's detection results, that is, the proportion of correctly predicted dust cover net construction waste pixels to the total predicted dust cover net construction waste pixels. The calculation equation is

$$Precision=\frac{TP}{TP+FP}. \tag{5}$$

Table 1
Dust cover net construction waste semantic segmentation confusion matrix.

	Construction waste	Background
Construction waste	*TP*	*FN*
Background	*FP*	*TN*

The *recall* rate represents the proportion of dust cover net construction waste pixels in the dataset that the model detects correctly and is calculated as

$$Recall = \frac{TP}{TP+FN}. \tag{6}$$

The *F*1*_score* is a comprehensive indicator that considers both accuracy and recall, and can comprehensively reflect the true performance of the algorithm. It is calculated as

$$F1\_score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{7}$$

The *IOU* represents the overlap rate between the segmentation result and the true value range as the ratio of intersection and union. *MIOU* represents the average ratio of intersection area to merged area, and the higher the value, the better the segmentation effect. When the *MIOU* value is 1, it indicates that the segmentation result completely overlaps with the true value; When the *MIOU* value is 0, it means there is no overlap. *MIOU* refers to the average handover ratio of k categories divided by the number of IOU categories. It is calculated as

$$MIOU = \frac{\square}{k+1}\sum_{i=0}^{k}\frac{}{TP+FN+FP}. \tag{8}$$

To further validate the effectiveness of the CBAM_U-net network model in the extraction of construction waste from high-resolution images in the dust cover net, we compare the extraction results with the current three typical semantic segmentation models, Pspnet, U-net, and Deeplabv3+. For the sake of fairness and objectivity in comparison, all the models were trained with the same experimental environment and parameter configuration, using the same dataset for training and testing. The backbone feature network adopted a transfer learning strategy, and the pretraining weights on the Imagenet dataset were used as the initial weights. The evaluation indicators for dust cover net construction waste in each model are shown in Table 2.

The experimental results show that the CBAM U-net model has a better overall performance in identifying construction waste on the dust cover net than other models. The *MIOU* value in the recognition results of the CBAM U-net model is 95.08%, which is 9.61% higher than that of the Deeplabv3+(xception) model. The *MIOU* value of construction waste with a dust cover net is 16.23% higher. The precision value in the CBAM U-net model's dust cover net construction waste identification results is 95.51%, which is 3.11% higher than that in the Pspnet (Mobilenetv2)

Table 2
Dust cover net construction waste semantic segmentation confusion matrix.

Model	Backbone feature network	*Precision* (%)		*Recall* (%)		*F*1_*score* (%)		*IOU* (%)		*MIOU* (%)
		Green_ CDW	Background	Green_ CDW	Background	Green_ CDW	Background	Green_ CDW	Background	
Deeplabv3+	mobilenetv2	94.62	98.63	92.15	99.08	93.36	98.85	87.56	97.74	92.65
	xception	95.25	96.31	78.21	99.32	85.89	97.79	75.27	96.67	85.47
Pspnet	mobilenetv2	92.40	98.24	89.86	98.71	91.11	98.47	83.68	96.99	90.33
	resnet50	93.93	98.90	93.73	98.94	93.83	98.92	88.37	97.87	93.12
U-net	resnet50	94.55	98.89	93.65	99.06	94.10	98.97	88.85	97.97	93.41
	vgg16	94.09	99.01	94.34	98.97	94.21	98.99	89.07	98.00	93.53
CBAM-U-net	vgg16	95.51	99.33	95.61	99.32	95.56	99.32	91.5	98.66	95.08

model. The *recall* value in the CBAM U-net model's dust cover net construction waste identification results is 95.61%, which is 17.40% higher than that in the Deeplabv3+(xception) model. The *F*1_*score* in the CBAM U-net model's dust cover net construction waste identification results is 95.56%, which is 9.67% higher than that in the Deeplabv3+(xception) model. The predicted results of construction waste are shown in Fig. 9.

To visually compare and analyze the prediction and recognition results of Google Image Dust cover net construction waste using various models, a visual display is shown in the following figure. CBAM U-net is more accurate in dividing the boundaries of objects than the other two methods, and has stronger recognition ability for dust cover nets and construction waste.

5. Discussion and Conclusions

In this paper, we proposed a remote sensing image dust cover net construction waste identification algorithm, CBAM U-net, based on an improved U-net model, to address the difficulty of identifying dust cover net construction waste in remote sensing images. The CBAM U-net algorithm is based on the U-net semantic segmentation model, using VGG16 as the backbone for feature extraction and adding a four-layer CBAM attention mechanism between downsampling and upsampling. Starting from both channel and spatial aspects, the algorithm improves its ability to learn features of dust cover net construction waste and to process detailed information. At the same time, to confirm the effectiveness of the algorithm, we prepared a dust cover net construction waste identification dataset based on Google high-resolution remote sensing images. A total of 1000 high-resolution dust cover net construction waste segmentation datasets were created nationwide through manual interpretation of the dataset. In the test set, the CBAM U-net model achieved better segmentation results than the other models. The *MIOU* value in the model recognition result is 95.08%, and the *F*1_*score* of the dust cover net construction waste recognition result is 95.56%. Compared with the U-net, Pspnet, and Deeplabv3+ models, the recognition effect of the CBAM U-net model is better.

Although the CBAM U-net model proposed in this article has achieved good recognition results on Google high-resolution remote sensing images, it uses Google images with a resolution of 0.30 m and high data quality. However, owing to the effect of seasons, weather, and the

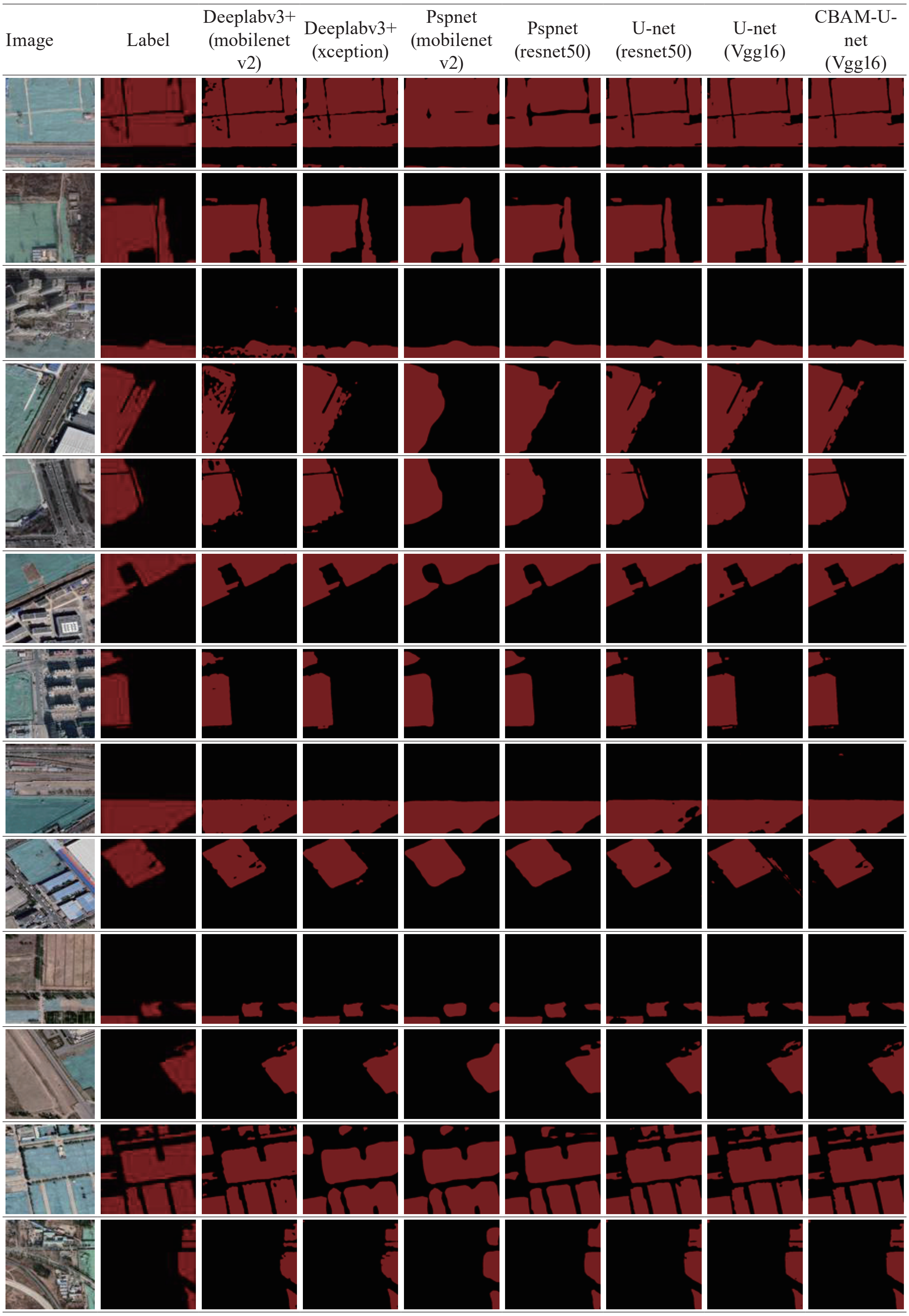


Fig. 9. (Color online) Prediction results of construction waste for each model.

characteristics of the construction waste itself, some images still have unclear edges of construction waste, which leads to inaccurate labeling and affects model feature learning, resulting in errors between the model prediction results and the real data. In later research, other types of data can be used to analyze the robustness of the model. Furthermore, we conducted an experiment on the binary classification recognition of construction waste on dust cover net, which can be used to identify various construction waste by type in the future. At the same time, owing to the complexity of remote sensing image scenes and the variability of target scales in practice, we improved the encoder–decoder network model by adding attention mechanisms. In the future, resource allocation can be further optimized, and how to improve the computational efficiency and segmentation accuracy of the model can be considered adding as few parameters as much as possible.

References

1 P. Wang, Y. Liu, Q. Sun, Y. Bai, and C. Li: Sustainability **14** (2022) 12286. https://doi.org/10.3390/su141912286

2 Y. Li, Q. Hou, Z. Zheng, M. M. Cheng, J. Yang, and X. Li: Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV, 2023) 16794–16805.

3 X. Li, M. Li, P. Yan, G. Li, Y. Jiang, H. Luo, and S. Yin: Int. J. Netw. Dyn. Intell. **2** (2023) 93. https://doi.org/10.53941/ijndi0201006

4 A. W. Salehi, S. Khan, G. Gupta, B. I. Alabduallah, A. Almjally, H. Alsolai, T. Siddiqui, and A. Mellit: Sustainability **15** (2023) 5930. https://doi.org/10.3390/su15075930

5 J. Long, E. Shelhamer, and T. Darrell: IEEE Trans. Pattern Anal. Mach. Intell. **39** (2017) 640. https://doi.org/10.1109/TPAMI.2016.2572683.

6 O. Ronneberger, P. Fischer, and T. Brox: 18th Int. Conf. Munich, Germany, October 5–9 (MICCAI, 2015) 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

7 V. Badrinarayanan, A. Kendall, R. Cipolla: IEEE Trans. Pattern Anal. Mach. Intell. **39** (2017) 2481. https://doi.org/10.1109/tpami.2016.2644615

8 L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille: arXiv (2014) https://doi.org/10.48550/arXiv.1412.7062

9 H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia: Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition. (IEEE, 2017) 2881–2890.

10 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, N. Houlsby: arXiv (2020) https://doi.org/10.48550/arXiv.2010.11929

11 E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo: Adv. Neural Inf. Process. Syst. **34** (2021) 12077.

12 Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger: 19th Int. Conf. Athens, Greece, October 17–21, (MICCAI, 2016) 424–432.

13 F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu: ISPRS J. Photogramm. Remote Sens. **162** (2020) 94.

14 X. Li, H. Chen, X. Qi, Q. Dou, C. W. Fu, and P. A. Heng: IEEE Trans. Med. Imaging **37** (2018) 2663.

15 D. Bahdanau, K. Cho, Y. Bengio: arXiv (2014) https://doi.org/10.48550/arXiv.1409.0473

16 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, I. Polosukhin: Adv. Neural Inf. Process. Syst. **30** (2017).

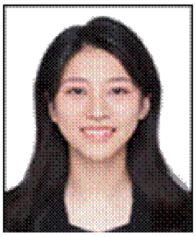
17 H. Touvron, M. Cord, and M. Douze: International Conference on Machine Learning (PMLR, 2021) 10347–10357.

18 Z. Liu, Y. Lin, Y. Cao: Proc. IEEE/CVF Int. Conf. Computer Vision (IEEE/CVF, 2021) 10012–10022.

19 J. Devlin, M. W. Chang, K. Lee, and K. Toutanova: arXiv (2018) https://doi.org/10.48550/arXiv.1810.04805

About the Authors

Shangwei Lv received his B.E. degree in geographic information science from Zhengzhou University, Zhengzhou, China, in 2018, and his master’s degree from Wuhan University in 2021. Since 2021, he has been a front-end engineer at Meituan, Shanghai. His current research interests include WebGIS platform development and spatiotemporal data analysis.
(lswliesmars@whu.edu.cn)



Xiaoyu Liu received her master's degree in surveying and mapping engineering from Beijing University of Civil Engineering and Architecture, Beijing, China, in 2023. Since 2023, she has been working as a project engineer at Beijing Urban Construction Survey, Design and Research Institute Co., Ltd. Her current research interests include three-dimensional modeling and image recognition.(18600385831@163.com)

Yifei Cao received his B.E. degree in surveying and mapping engineering from Beijing University of Civil Engineering and Architecture, Beijing, China, in 2019 and his master’s degree from Beijing University of Civil Engineering and Architecture in 2022. Since 2022, he has been an assistant engineer at Beijing Institute of Surveying and Mapping, China. His current research interests include target detection and tracking in remote sensing.
(1023362448@qq.com)