Article

Semantic Segmentation Method for Remote Sensing Images Based on an Improved TransDeepLab Model

Wang jin Xin , Wang man Man * , Qin zi Long

Posted Date: 6 November 2024

doi: 10.20944/preprints202411.0379.v1

Keywords: Deep learnin; High-resolution remote sensing image; Semantic segmentation; Feature extraction

Article

Semantic Segmentation Method for Remote Sensing Images Based on an Improved TransDeepLab Model

Wang Jinxin [1], Wang Manman [1,*] and Qin Zilong [2]

[1] School of Geoscience & Technology, Zhengzhou University, Zhengzhou 450001, China;
[2] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China;
* Correspondence: 13298116096@163.com

Abstract: Due to the various types of land cover and large spectral differences in remote sensing images, their high-quality semantic segmentation still faces severe challenges. This study proposes a new improved model based on the TransDeepLab segmentation method. The model introduces the GAM attention mechanism in the encoding stage, which can effectively reduce the information dispersion and enlarge the global interaction features; in the decoding stage, a multi-level linear upsampling strategy is designed to gradually amplify the multi-scale features extracted from the encoder and integrate them with the low-scale features to improve the segmentation effect of different shapes and sizes. The model makes full use of the multi-level semantic information and small target detail information in high-resolution remote sensing images, which can effectively improve the segmentation accuracy of target objects. Using open-source LoveDA large remote sensing image data sets for the validation experiment, the results show that compared to the original model, its Mean Intersection Over Union (MIOU) increased by 2.68 %, Average Pixel Accuracy (aACC), and Mean Pixel Accuracy (mACC) by 3.41% and 4.65%. Compared to other mainstream models, the model also achieved a better segmentation effect.

Keywords: deep learning; high-resolution remote sensing image; semantic segmentation; feature extraction

1. Introduction

In recent years, accurate classification and interpretation of remote sensing images has become a research frontier in the field of geoinformation science. Early remote sensing image segmentation methods used segmented images based on shallow feature semantic information such as image color, texture and gradient. They can be divided into edge detection-based segmentation methods [1], region-based segmentation methods [2], threshold-based segmentation methods [3], and segmentation methods combined with specific theories [4]. However, early remote sensing image semantic segmentation methods only performed well when extracting low-level semantic information; and had low segmentation accuracy and poor accuracy for target extraction tasks in complex environments, which cannot meet the requirements of remote sensing image interpretation applications.

Since Donald Herbst [5,6] and others proposed the backpropagation algorithm in the 1960s, the field of intelligent learning algorithms has kicked off. Subsequently, Many excellent machine learning algorithms have been applied to the semantic segmentation tasks of remote sensing images. However, these methods are not suitable for high-resolution remote sensing images with widely different features and complex spectral texture information. With the improvement of image processing requirements in real life, new technologies with stronger feature extraction and generalization capabilities are urgently needed.

With improvements in computer hardware and software; and the increasing demand for image processing in practical work, deep learning technology has gradually penetrated all aspects of life. For example, it is widely used in the fields of security [7], handwritten digit recognition [8], human action recognition [9], financial transactions [10], remote image processing [11–15], and others [16–20]. With the continuous development of deep learning methods in the field of image processing, numerous network models with strong feature-learning abilities have emerged. These models can automatically learn spatial features and topological relationships, better capture multi-level information in remote sensing images through the hierarchical feature extraction of deep neural networks, and obtain more accurate image segmentation results. For example, the FCN [21] algorithm can replace the full-connection layer with a convolution layer at the prediction output end, which can realize end-to-end training prediction regardless of the size of the input image. However, the FCN used high-level features of spatial information as the basis for pixel classification, which led to the neglect of low-level features with rich semantic information; Thus, the FCN performed poorly in processing multiple-images, and its segmentation was rough. Villa proposed the PSPNet [22] algorithm to overcome the shortcomings of FCNS. PSPNet uses a pyramid-pool module that can integrate the context information of each region to obtain better global information. Since then, several researchers have proposed effective network frameworks. The encoder-decoder structural model exhibits excellent performance. For example, Ronneberger et al. [23] proposed a U-net network structure that connects high-level features generated by the decoder with low-level features generated by the corresponding encoder through a skip connection, which can better realize the semantic segmentation task of remote sensing images. RefineNet [24] and SegNet [25] use similar network structures. This network structure uses an encoder decoder structure; the encoder is used to extract feature information from remote images, and the decoder restores the extracted features. The decoder compensates for lost information in the encoding process, fuses low-level features, and improves the segmentation accuracy of the network.

However, these models have insufficient power to aggregate contextual information, which results in poor prediction results. Subsequently, He et al. [26] proposed a variant model based on DeeplabV3Plus, which adopted MobileNetV2 as the backbone network and introduced a dual attention mechanism to extract bare land; the performance of this model was better than that of the model architecture based on Xception as the backbone network. However, owing to the locality of convolution operations, obtaining global convolution context information directly remains challenging. Inspired by the global modeling capability of a Swin transformer, He et al. [27] proposed a new remote sensing image semantic segmentation framework and constructed a dual encoder structure of a Swin transformer and a CNN, which can improve the segmentation accuracy of small-scale ground objects. Similar to the Swin transformer principle, Su et al. [28], based on the backbone of a Swin transformer, proposed a pure efficient transformer with mlphead to accelerate the inference speed; and proposed an explicit and implicit edge enhancement method to deal with the target edge problem. In 2022, based on the concept of the DeeplabV3Plus network structure and Swin transformer, Reza et al. [29] proposed a DeeplabV3+ image segmentation model, Transdeeplab, based on unconvolutional transformation. Initially used for medical image segmentation, the network leverages a layered swin-transformer with shift windows to extend DeepLabv3 and model the Spatial Pyramid Pool (ASPP) module. The Swin pyramid module of the TransDeepLab model can capture multi-scale information by using different window sizes, and then integrate the obtained multi-scale context information into the decoder module by using the leap-above context attention mechanism,which can capture more comprehensive feature information of the target object. However, the accuracy of feature extraction from complex remote-sensing images is low.

This study conducted research based on the TransDeepLab model framework, further analyzeds the existing problems of the network, and proposeds a new deep-learning network architecture. First, the Convolutional Block Attention Module (CBAM) [30] attention mechanism in the coding stage of the TransDeepLab model is replaced by a new attention mechanism, Global Attention Module (GAM) [31–34], which can enlarge the global dimensional interaction features while reducing information dispersion, thereby obtaining more abundant feature information. Second, in the decoding path, the extracted multi-scale features are up-sampled several times and connected to the low-scale features from the encoder, to refine the feature representation. Compared to many existing network models, the improved model proposed in this study achieveds better precision results. This study provides a new technical reference for large-scale land cover classification task resource surveys.

2. Materials and Methods

2.1. Method Framework

The TransDeepLab split network model is a fully transformer-based DeepLabv3+ architecture that extends the encoder-decoder architecture of DeepLabv3+, and uses a series of swin-transformer modules to encode input images into highly representative spatial features. The encoder module divides the input image into 4×4 non-overlapping image blocks, each of which has a feature dimension of 4×4×3 = 48 (represented as C), and applies swin-transformer [33] blocks to encode local semantic information and global context information. In the coding stage, two stacked swin-transformer blocks were used to focus on inter-dimensional global context information and ignores the interaction between the channel and spatial attention, resulting in information loss. This study replaces the CBAM attention mechanism with the GAM attention mechanism module, which integrates and captures more small-scale features using nonlinear technology, minimizes information loss, and amplifies global interaction. The attention mechanism module uses two levels of channel-attention and spatial-attention, to capture feature information from each level of the pyramid, forming multi-scale interactions.

In contrast, in the decoding stage, the TransDeepLab model uses only bilinearity upsampling step and then combines low-level feature information with high-level semantic information. However, the feature information extracted by the encoder was richer and occupied a higher proportion of the training process. Therefore, in this study, the output of the encoder was upsampled four times to collect rich spatial feature information and then fused with small-scale features, ensuring that high and low-level features can fully used. The specific implementation process is the channel compression of low-level features by a 1×1 convolution,which can effectively reduce the influence of low-level features and highlight the role of high-level features in facilitating the training and optimization of the model. In the decoder, multiple upsampling operations retain high-level semantic information and improve the segmentation degree by supplementing low-level features; and effectively improving the segmentation effect of different shapes and sizes of ground objects. The improved TransDeepLab network architecture is shown in Figure 1.

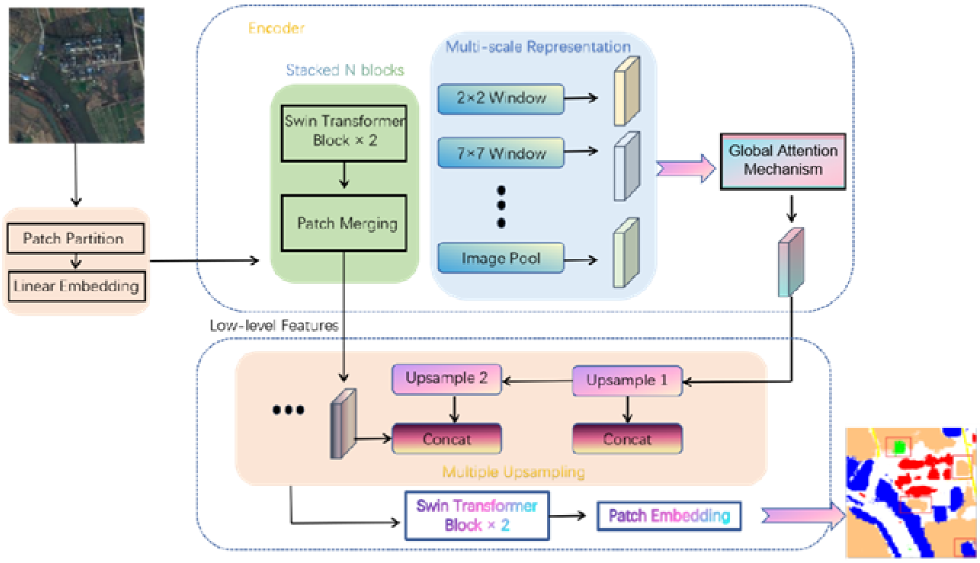


Figure 1. Overall architecture of the improved TransDeepLab network.

2.2. Attention Mechanism

The GAM attention mechanism is a global attention mechanism that is an upgraded version of CBAM. This improves the performance of deep neural networks by reducing information loss and

doi:10.20944/preprints202411.0379.v1

amplifying interactive global representations. The GAM introduces a 3D arrangement of multi-layer perceptrons for channel attention and convolution space attention submodules. The original basic model uses the CBAM attention mechanism module. Although CBAM includes channel and spatial attention operations, it has certain drawbacks. The structure of the GAM attention module is shown in Figure 2.

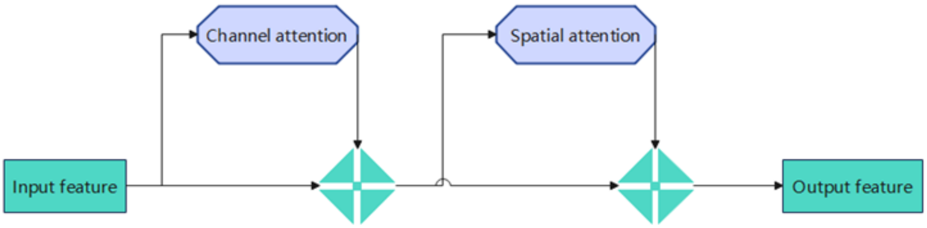


Figure 2. GAM attention structure diagram.

For the processing of channel attention, CBAM directly maximizes and averages the input feature map through Multi-Layer Perceptron (MLP) processing, which is finally activated by a sigmoid. In the spatial attention part, the feature map is a direct superposition convolution after maximum and average pooling and is then processed by a sigmoid activation function. Moreover, it causes information loss and ignores the interaction between the channel and space, thus losing cross-dimensional information.

In the attention part of the channel, GAM [35] first performs dimension transforms on the input feature map, uses a three-dimensional arrangement to retain information in three dimensions, and inputs it to the MLP, converts it to the original dimensions, and then performs sigmoid processing and outputs. Convolution processing is mainly used for spatial attention. To focus on spatial attention, the number of channels was first reduced by convolution with Convolution Kernel 7, and the number of calculations was reduced. Then, the number of channels is increased by the convolution operation with a Convolution Kernel 7 to keep the number of channels is consistent. Finally, it is outputted by the sigmoid. This amplifies the interdimensional interaction and captures the important features of the target in all three dimensions. Figure 3 shows the specific structure of the GAM.

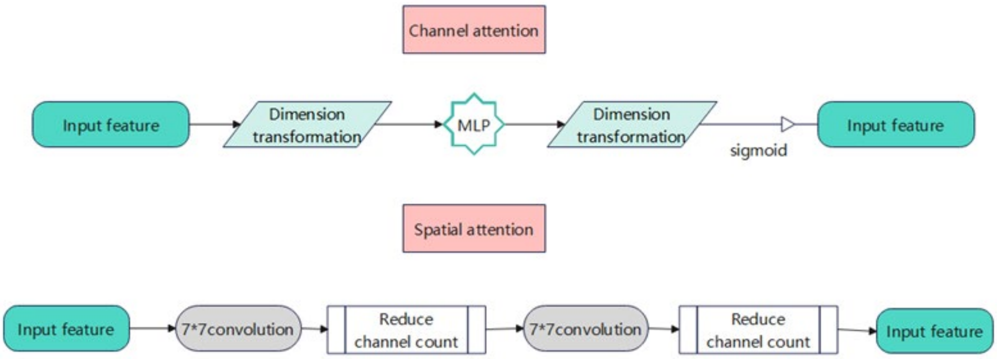


Figure 3. GAM structure detail diagram.

2.3. Decoder

The improved decoding network consists of four up-sampling blocks that are used to restore the feature map to the size of the input image. Each up-sampling block first adjusts the channel number and image size of the high-level features to the same level as the corresponding low-level features of the coding network through a transposed convolution. The high and low-level features are then concatenated, and the result is used as an input to a Channel Selection Block (CSB). After each upper sampling block, the number of channels in the feature map is halved.

To further enhance the feature representation, several upper sampling layers are added in the decoding process, and the “connect” module is introduced after each upper sampling layer. This

doi:10.20944/preprints202411.0379.v1

module connects high-level features with corresponding low-level features to form a cross-layer feature fusion, thereby retaining more details and achieving a thinning feature representation. Each upper sampling block first uses transposition convolution to adjust the channel number and spatial resolution of the high-level feature map to be the same as those of the corresponding low-level feature, and then concatenates the transposition convolution of the high-level feature with the corresponding low-level feature in the coding network to form a rich feature representation. The concatenated feature map was input into the CSB, and important features were further extracted using a nonlinear transformation. During the decoding process, several up-sampling operations are performed, each time the feature map is enlarged, and the size of the input image is gradually restored. Finally, the high-dimensional features are translated into 8-channel features using convolution kernel 1 to obtain the final predictive output. In the decoder, the adjustment and concatenation processes of the feature map are as follows:

Up-sampling process: High-level feature graphs are first up-sampled by transposing the convolution:

$$F_{up}^{l} = T\left(F_{high}^{l};\theta_T\right) \ (1)$$

Among them, F_{up}^{l} is the feature map after up-sampling on the L-layer, F_{high}^{l} is the high-level feature map of the L-layer, and $T\left(F_{high}^{l};\theta_T\right)$ represents a transposed convolution operation with parameter θ_T.

Feature splicing: After up-sampling, the feature map is spliced with the corresponding low-level feature map in the encoder:

$$F_{concat}^{l} = \left[F_{up}^{l}, F_{low}^{l}\right] \ (2)$$

where; F_{concat}^{l} is the feature map after concatenation, F_{low}^{l} is the low-level feature map of the L-layer in the encoder, and $\left[F_{up}^{l}, F_{low}^{l}\right]$ represents the concatenation operation in the channel dimension.

CSB processing: The concatenated feature map is nonlinear transformed by CSB:

$$F_{csb}^{l} = CSB(F_{concat}^{l};\theta_{csb}) \ (3)$$

where; F_{csb}^{l} is the feature map with the output of CSB, $CSB(F_{concat}^{l};\theta_{csb})$ indicates the CSB operation with parameter θ_{csb}.

Multiple up-sampling: Assuming that L layers are up-sampled, the output of the L layer can be expressed as

$$F_{out}^{l} = \{F_{csb}^{l}, \quad \mathrm{if} l < L Conv_{1\times1}\left(F_{csb}^{l};\theta_{conv}\right), \quad \mathrm{if} l = L \ (4)$$

where; $Conv_{1\times1}(F_{csb}^{l};\theta_{conv})$ represents $1*1$ convolution operation with parameter θ_{conv}.

Final output: After several upsamples, the final output feature maps are as follows:

$$F_{final} = F_{out}^{L} \ (5)$$

The F_{final} is the final feature map for the eight channels. This process can be expressed in the following recursive form:

$$F_{out}^{l} = \{CSB\left(\left[T(F_{out}^{l-1};\theta_T^l), F_{low}^{l}\right];\theta_{csb}^{l}\right), \quad \mathrm{if}\ 1 \le l < L Conv_{1\times1}(CSB([T(F_{out}^{L-1};\theta_T^L), F_{low}^{L}];\theta_{csb}^{L});\theta_{conv}), \quad \mathrm{if} l = L \ (6)$$

Where F_{out}^{l} is the initial high-level feature map.

3. Experiments and Results

3.1. Datasets

LoveDA, a land use and cover data set from Google Earth, was used to collect real urban and rural remote sensing images from Nanjing, Changzhou and Wuhan in 2016. The image data had a spatial resolution of 0.3m, including 5,987 high-resolution images and 166,768 annotated objects that were divided into urban and rural parts. This data set had three characteristics. First, multi-scale objects, the same category of objects in different scenes, have completely different geographical landscapes and increase the change of scale; second, they have extremely complex background samples. The high resolution and different complex scenes as background in the LoveDA dataset bring richer details and greater intra-class variance. Finally, there are differences in the class distribution. Urban scenes with high population densities contain many artificial objects, such as buildings and roads. Compared with rural scenes, which contain more natural elements, such as forest and water, this dataset focuses on stylistic differences in the geographical environment compared with common image segmentation datasets, and inconsistent category distribution will bring special challenges to image segmentation tasks. Compared to existing submeter high-resolution land coverage datasets, this dataset contains more pixel samples and labeled ground objects. The labels of the dataset included seven types of ground objects: buildings, roads, water, barren, forest, agriculture, and background, among which building examples were marked the most frequently. Figure 4 shows a few of the selected raw images and label maps.

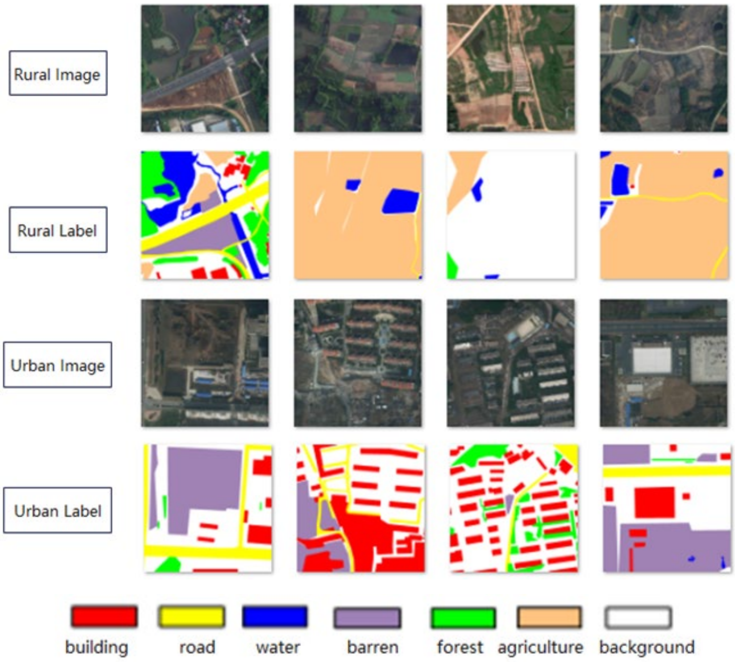


Figure 4. Training sample and label example.

3.2. Implementation Details

In this study, a deep learning network model was built on an NVIDIA 3090 (90G video memory) GPU combined with a PyTorch framework, and the model effect was verified. The initial learning rate was set to 0.001 during the experiment, which stabilized the model during the training process. The momentum was 0.9, which helped the model accelerate and converge in the right direction during training. The batch size was set to 16 to reduce the risk of overfitting to a certain extent. A conventional cross-entropy loss function was selected to calculate the loss value, and a stochastic gradient descent learning rate strategy was used for the weight optimization iteration. The total number of iterations of the model training was set to 200, and the model was stored once every five epochs to avoid storing the model too frequently. Simultaneously, the training process was guaranteed to have sufficient checkpoints to monitor the training progress and performance, and the model with the best performance on the verification set was saved during training for verification and testing experiments.

3.3. Evaluation Metric

In this experiment, we selected three evaluation metrics to measure network performance: mean intersection over union (MIOU), Average Pixel Accuracy (aACC), and mean Pixel Accuracy (mACC). The IOU is mainly used to measure the degree of coincidence between the segmentation and the real results. The higher the IOU, the closer the segmentation result is to the real result, and the better the segmentation performance. MIOU is the mean of all categories of IOU and represents the average of the intersection and union ratio of the two sets of predicted and true values of all categories of ground objects. Acc is mainly used to measure the number of correctly classified pixels in the segmentation results; aACC is the proportion of correctly classified pixels in the total number of pixels and mACC represents the average rate of pixel classification accuracy of all categories, that is, the number of correctly classified pixels in each category and the proportion of pixels in the category [36–38]. The formulas for calculating each indicator are as follows:

$$aACC = \frac{\sum_{i=0}^{k} pii}{\sum_{i=0}^{k}\sum_{j=0}^{k} pij} \quad (7)$$

$$mACC = \frac{1}{k+1}\sum_{i=0}^{k}\frac{pii}{\sum_{j=0}^{k} pij} \quad (8)$$

$$mIOU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{pii}{\sum_{j=0}^{k} pij + \sum_{j=0}^{k} pij - pii} \quad (9)$$

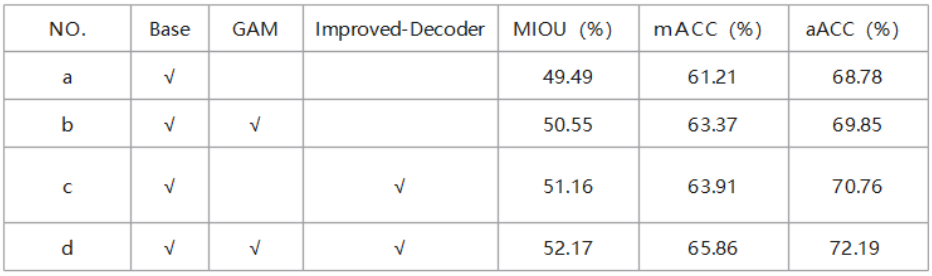
where, K represents the total number of classes, Pii represents the number of pixels class in class i, $\sum_{j=0}^{k} pij$ represents the number of pixels belonging to class i but predicted to belong to another class, $\sum_{j=0}^{k} pji$ represents the number of pixels belonging to another class and predicted to belong to other classes, represented by i.

3.4. Experiment and Result Analysis

3.4.1. Ablation Experiment

To verify the effectiveness of the improved modules, this study added the improved modules and process experiments individually to verify the effect of the improved decoder and the newly proposed GAM attention mechanism module on model performance. It can be seen from the experimental results in Table 1 that each improvement modules improved the model performance. First, multiple up-samplings using the improved decoder can improve the utilization of different receptive fields and extract more complete object contours by focusing on more local information. Compared to the original model, the MIOU increased from 49.49 % to 51.16 %, mACC and aACC also increase from 61.21 % and 68.78 % to 63.91 % and 70.19 %, respectively. Second, after replacing the CBAM attention mechanism of the original model with the GAM module, the model can improve the aggregation of feature information on the channel attention and convolution space attention submodules and reduce information loss by amplifying global information interaction. Compared with the original model, the MIOU, mACC and aACC increased from 49.49 %, 61.21 % and 68.78 % to 50.55 %, 63.37 % and 69.85 %, respectively. Finally, the overall performance of the model was improved by adding an improved module. The model combined with all the improved modules showed the highest segmentation accuracy, improving the MIOU, aACC, and mACC by 2.68 %, 3.41 % and 4.65 %, respectively, compared with the base model. The experimental accuracies before and after improvement are listed in Table 1.

Table 1. Comparison of ablation experimental evaluation results.



NO.	Base	GAM	Improved-Decoder	MIOU (%)	mACC (%)	aACC (%)
a	√			49.49	61.21	68.78
b	√	√		50.55	63.37	69.85
c	√		√	51.16	63.91	70.76
d	√	√	√	52.17	65.86	72.19

To demonstrate the validity of the improved model directly, we compared the segmentation results of the improved model with those of the original model and the classical network model Deeplabv3+ in the loveDA datasets, and then showed the detailed classification of various types of ground items, as shown in Table 2.

Table 2. Precision comparison of transdeeplab and its improved model.

model	IOU and Acc scores for each class (%)							MIOU/%	mACC/%	aACC/%	Speed (FPS)
	building	road	water	barren	forest	agriculture	background				
Deeplabv3+	42.97	50.88	52.02	74.36	10.40	44.21	42.97	42.41	54.13	66.37	64.6
	54.17	48.40	76.91	16.94	40.21	57.90	84.40				
transdeeplab	43.04	51.34	50.93	74.77	10.03	42.19	43.03	49.49	61.21	68.78	27.3
	75.11	61.46	78.14	26.75	48.85	50.59	87.58				
transdeeplab +GAM	43.06	52.74	52.78	73.08	10.33	43.05	43.06	50.55	63.37	70.76	24.4
	79.15	63.80	80.47	28.27	51.73	56.79	83.38				
transdeeplab +Improved-Decoder	42.85	52.58	52.82	74.51	11.42	44.42	42.85	51.16	63.91	70.76	15.4
	77.08	63.40	84.96	29.06	50.64	58.62	83.65				
Textual method	59.59	54.57	67.46	24.69	38.63	49.98	54.81	52.17	65.86	72.19	14.5
	79.94	65.88	83.54	25.09	59.55	68.48	78.53				

*Note: The first line of each model is the IOU value, and the second line is the ACC score.

As can be seen from the data in the table, the segmentation IOU and ACC scores of the proposed method for each class were higher than those of the basic unimproved TransDeepLab model. Only the barren IOU score was 25.09 %, which was lower than that of the basic model 74.77 %; however, the overall MIOU score was 52.17 %. The overall segmentation result is still closer to the real situation than that of the basic model, which may be due to the fuzzy boundary of the bare land category and the difference in the geographical environment of the image itself; thus, the accuracy decreases. In addition, by comparing the Deeplabv3+ model with the mainstream segmentation network model, the scores of all kinds of IOU and ACC in the proposed method are much higher than those of the external network, except the barren, and the overall MIOU score is approximately 10 % higher than that of the Deeplabv3+ algorithm. The conclusion can be obtained from the results of the study that the improved method proposed has greatly raised segmentation precision and accuracy compared with the original and external segmentation methods. However, the running speed is slow, which may be the reason for the increase in the parameters, and further improvement and optimization are required.

Improvements in the prediction effect can be observed directly from the prediction results of each model for high-resolution images. The specific experimental results are shown in Figure 5, where the obvious improvement areas are highlighted in red boxes.

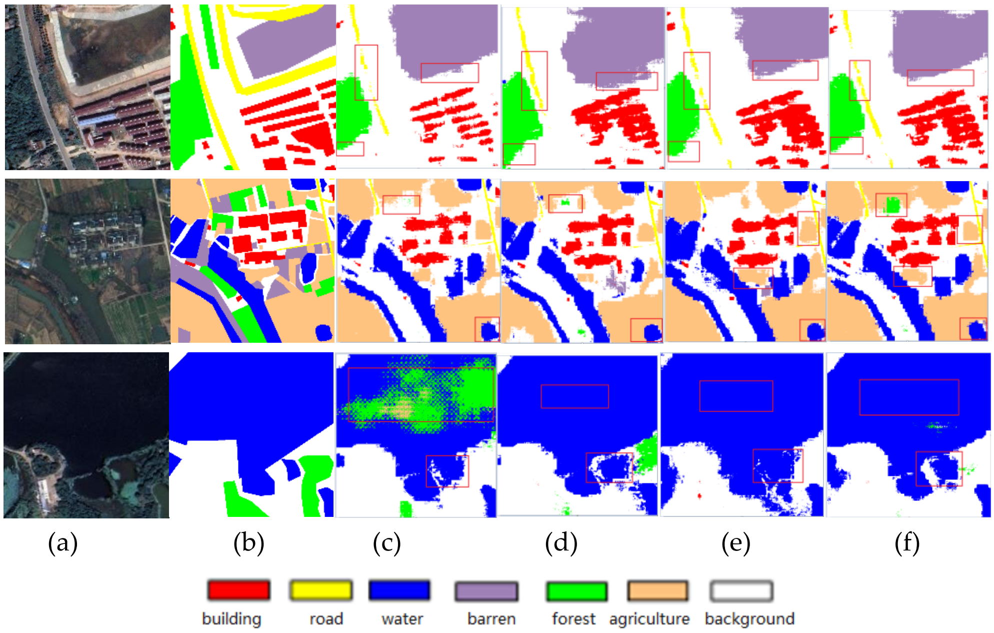


Figure 5. Land use classification prediction map of the ablation experiment. (a) Input image, (b) Ground Truth, (c) base, (d) base+GAM, (e) base+Improved-Decoder, (f) Textual method.

As shown in Figure 5, the basic model misclassified a large body of water into the forest, and the extraction effect of the forest was particularly poor, with many misextractions and omissions. However, after the addition of the GAM module, the model's identification accuracy for each class was greatly improved, and there was no misjudgment, especially when the extraction effect of the water body was relatively obvious. The recognition of small objects with a poor recognition effect in the original model was improved, and the recognition rate was higher than that of similar objects in the original model. This is because, after the addition of the GAM module, through the channel attention mechanism and spatial attention mechanism, the model can focus on the interaction between feature maps in the channel and space and; capture more comprehensive target information. Thus, the difference between ground objects of each class can be more clearly identified, the accuracy of the model can be increased, and various ground objects can have richer multi-scale features to express their semantic information. Therefore, we can successfully increase the accuracy of the model by using its advanced features to distinguish between classifications.

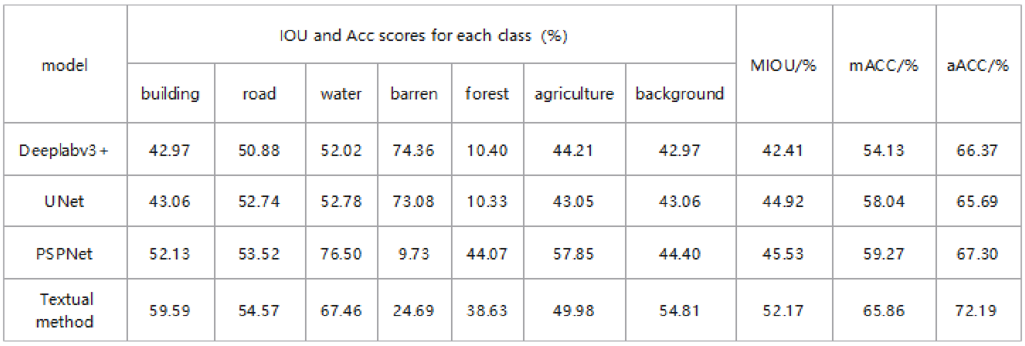
In addition, it can be seen from the figure adding a connection to the decoder of the basic model also improves the recognition effect compared with the original model. Owing to the loss of spatial information, the segmentation map generated by the basic model is fuzzy; in particular, the object contour extraction is not clear. However, after adding connection to the decoder and performing quartic up-sampling, the geometric and complex contour information of the ground object can be better preserved. Compared with the original model, the MIOU, mACC and aACC values of the improved decoder model increased by 1.67 %, 2.7 % and 1.98 %, respectively. This is because the improved decoder pays more attention to the extraction of low-level feature information; and inhibits error information in the feature fusion stage of the decoder, which enhances the useful feature expression. Compared with the original model, the method proposed in this study significantly reduces the confusion rate between classes, provides more accurate recognition between classes, and can accurately identify the detailed information of each class feature.

3.4.2. Contrast Experiment

To further evaluate the effectiveness of the proposed method, we compared the improved TransDeepLab model with several mainstream image segmentation models, including Deeplabv3+ [39], Unet [40], and PSPNet [41]. To ensure experimental fairness and data validity, all networks were trained and tested using the same software and hardware. A comparison of the specific accuracy and prediction results is presented below.

For the LoveDA dataset, the MIOU value of the proposed method exceeded DeepLabv3Plus by 9.76 percentage points and outperformed PSPNet by 6.65 percentage points. This stems from the GAM's focus on global information interaction, which helps capture irregularly shaped features more comprehensively. For the classification of many rule instance images, such as buildings, roads, and other categories, the proposed method was significantly better than the DeepLabv3Plus and UNet methods. In addition, among the seven categories shown in Table 3, compared with the mainstream PSPNet method, the proposed method improved the accuracy of the extraction of buildings and roads to varying degrees. However, the segmentation accuracy of lakes, forests and farmland was slightly lower than that of the PSPNet network, which may be due to the dense connection used in the decoder. Although specific upsampling models can be learned, some details are lost, and there is a certain degree of segmentation ambiguity for targets with complex edge details.

Table 3. Precision comparison between the proposed method and other models.



model	IOU and Acc scores for each class (%)							MIOU/%	mACC/%	aACC/%
	building	road	water	barren	forest	agriculture	background			
Deeplabv3+	42.97	50.88	52.02	74.36	10.40	44.21	42.97	42.41	54.13	66.37
UNet	43.06	52.74	52.78	73.08	10.33	43.05	43.06	44.92	58.04	65.69
PSPNet	52.13	53.52	76.50	9.73	44.07	57.85	44.40	45.53	59.27	67.30
Textual method	59.59	54.57	67.46	24.69	38.63	49.98	54.81	52.17	65.86	72.19

According to the experimental prediction results in Figure 6, the popular mainstream segmentation methods have certain effects on object segmentation and can roughly identify various objects, but the misclassification rate is high and there is more confusion among different types. Meanwhile, missing or misclassification is often observed in the segmentation of buildings, and the segmentation among similar types is not sufficiently accurate; the object boundaries are not sufficiently fine and there is more noise. Compared with previous models, the proposed method has a higher classification accuracy and can more accurately distinguish each ground object category. Meanwhile, the segmentation results are more in line with geological logic, and there are no disadvantages to other models with more noise. Therefore, edge extraction between different types of objects is more accurate. Overall, the comparative analysis shows, that the segmentation map of the improved model is the closest to the label map, and the segmentation of elongated targets such as buildings, roads and lakes is more accurate. Experiments show that the improved method proposed in this study is effective, and that the improved network has better segmentation performance than the general segmentation network.

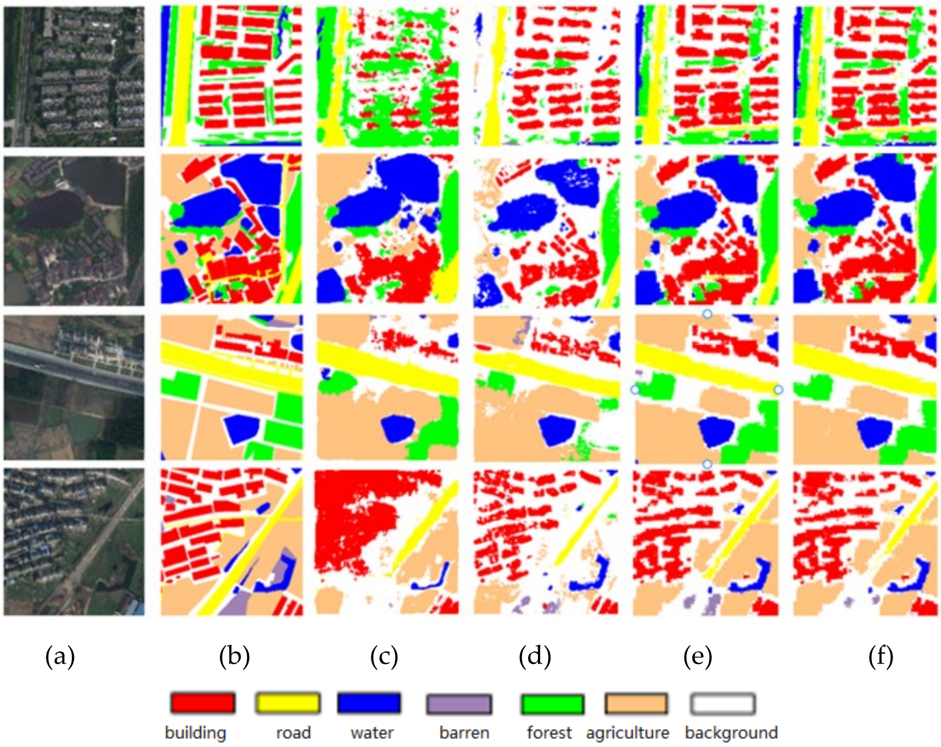


Figure 6. Comparison of the results of different models. (a) Real image, (b) label, (c) Deeplabv3+, (d) unet, (e) PSPNet, (f) Textual method.

4. Discussion

Current popular semantic segmentation methods fail to make full use of the rich contextual semantic information in high-resolution remote sensing images, and the segmentation accuracy of small-scale surface objects is low because of the imbalance of surface object categories and large scale differences in high-resolution remote sensing images. In this study, an improved network architecture based on the transformer model was proposed. The GAM attention mechanism was introduced in the coding stage, and a multi-level linear up-sampling strategy was added in the decoding stage. The experiment is verified by using open source data set, and a good segmentation effect was obtained. The main conclusions are as follows:

1) Compared the CBAM mechanism used in the base model, the GAM global attention mechanism further optimizes feature representation. By enhancing the global interactive representation and reducing information loss, the model can consider a larger receptive field and extract more scaled features.

2) The multi-level linear up-sampling strategy added in the decoding stage can gradually amplify the multi-scale features extracted from the encoder, collect rich spatial feature information, and fuse with low-scale features. This cross-layer connection method preserves the semantic information of high-level features, refines the feature representation, and improves the granularity resolution and accuracy of the segmentation results.

3) The experimental results show that the above improvements make the model achieve higher accuracy and robustness in the semantic segmentation of high-resolution remote sensing images, especially in the segmentation of complex scenes and multi-scale targets, and the advantages are more obvious. In the open source LoveDA dataset, the mIOU, aACC, and mACC of the improved model increased by 2.68 %, 3.41 % and 4.65 %, respectly, compared with the basic transformer model before the improvement, and were superior to other commonly used semantic segmentation models in multiple evaluation indicators. These include Deeplabv3+, Unet and PSPNet.

However, this model still has certain limitations. First, the number of parameters in the network is too large, which increases the training cost and reduces the training speed. Secondly, there are missing or incorrect classifications of similar ground objects. Lightweight Networks, such as MobileNet and ShuffleNet, can be considered for subsequent research to reduce the scale of the model and speed up the inference. Simultaneously, techniques such as Model Pruning and Quantization can be explored to reduce the computational load and memory footprint of the model, thereby increasing segmentation speed.

Author Contributions:: Conceptualization, M.W.; methodology, M.W. and Z.Q.; validation, M.W. and Z.Q.; formal analysis, M.W. and Z.Q.; investigation, M.W.; writing—original draft preparation, M.W. and J.W.; writing—review and editing, M.W. , Z.Q. and J.W.; visualization, M.W.; supervision, J.W. and Z.Q.; All authors have read and agreed to the published version of the manuscript.

Funding: This study received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code and data can be downloaded below: https://github.com/JunJue-Wang/LoveDA

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dharampal, V. M. Methods of image edge detection: A review. J. Electr. Electron. Syst,2015, 4(2), 2332-0796. DOI:10.4172/2332-0796.1000150
2. Feng, J., Qing, G., Huizhen, H., Na, L., Yan-Wen, G., & Dao-Xu, C. Survey on content-based image segmentation methods. Journal of software, 2017, 28(1), 160-183. DOI:10.13328/j.cnki.jos.005136
3. Abdullah, S. L. S., & Jamil, N. Segmentation of natural images using an improved thresholding-based technique. Procedia Engineering, 2012, 41, 938-944. DOI:10.1016/j.proeng.2012.07.266

4. Hossain, M. D., & Chen, D. Segmentation for Object-Based Image Analysis (OBIA): A review of algorithms and challenges from remote sensing perspective. ISPRS Journal of Photogrammetry and Remote Sensing, 2019, 150, 115-134. DOI:10.1016/j.isprsjprs.2019.02.009
5. Rumelhart, D. E., Hinton, G. E., Williams, R. J. Learning representations by back-propagating errors[J]. nature, 1986, 323(6088): 533-536. DOI:10.1038/323533a0
6. Yu, X., Efe, M. O., & Kaynak, O. A general backpropagation algorithm for feedforward neural networks learning. IEEE transactions on neural networks, 2002, 13(1), 251-254. DOI:10.1109/72.977323
7. Tang, J., Han, P., & Liu, D. Adhesive Handwritten Digit Recognition Algorithm Based on Improved Convolutional Neural Network. In 2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS)2020, (pp. 388-392). IEEE. DOI:10.1109/icaiis49377.2020.9194797
8. Lee, S.; Xiong, W.; Bai, Z. Human action recognition based on supervised class-specific dictionary learning with deep convolutional neural network features. Comput. Mater. Contin. 2020, 63, 243–262. DOI:10.32604/cmc.2020.06898
9. Sezer, O.; Ozbayoglu, A. Financial trading model with stock bar chart image time series with deep convolutional neural networks. Intell. Autom. Soft Comput. 2020, 26, 323–334. DOI:10.31209/2018.100000065
10. Qian, J.; Xia, M.; Zhang, Y.; Liu, J.; Xu, Y. TCDNet: Trilateral Change Detection Network for Google Earth Image. Remote Sens. 2020, 12, 2669. DOI:10.3390/rs12172669
11. **a, M., Tian, N., Zhang, Y., Xu, Y., & Zhang, X. Dilated multi-scale cascade forest for satellite image classification. International Journal of Remote Sensing, 2020, 41(20), 7779-7800. DOI:10.1080/01431161.2020.1763511
12. Weng, L.; Xu, Y.; Xia, M.; Zhang, Y.; Liu, J.; Xu, Y. Water Areas Segmentation from Remote Sensing Images Using a Separable Residual SegNet Network. Int. J. Geo-Inf. 2020, 9, 256. DOI:10.3390/ijgi9040256
13. Xia, M.; Cui, Y.; Zhang, Y.; Liu, J.; Xu, Y. DAU-Net: A Novel Water Areas Segmentation Structure for Remote Sensing Image. Int. J. Remote Sens. 2021, 42, 2594–2621. DOI:10.1080/01431161.2020.1856964
14. Yan, Z., Yan, M., Sun, H., Fu, K., Hong, J., Sun, J., ... & Sun, X. Cloud and cloud shadow detection using multilevel feature fused segmentation network. IEEE Geoscience and Remote Sensing Letters, 2018, 15(10), 1600-1604. DOI:10.1109/LGRS.2018.2846802
15. Chen, B.; Xia, M.; Huang, J. MFANet: A Multi-Level Feature Aggregation Network for Semantic Segmentation of Land Cover. Remote Sens. 2021, 13, 731. DOI:10.3390/rs13040731
16. Xia, M.; Liu, W.; Wang, K.; Song, W.; Chen, C.; Li, Y. Non-intrusive load disaggregation based on composite deep long short-term memory network. Expert Syst. Appl. 2020, 160, 113669. DOI:10.1016/j.eswa.2020.113669
17. Lee, S.; Ahn, Y.; Kim, H. Predicting concrete compressive strength using deep convolutional neural network based on image characteristics. Comput. Mater. Contin. 2020, 65, 1–17. DOI:10.32604/cmc.2020.011104
18. Janarthanan, A.; Kumar, D. Localization based evolutionary routing (lober) for efficient aggregation in wireless multimedia sensor networks. Comput. Mater. Contin. 2019, 60, 895–912. DOI:10.32604/cmc.2019.06805
19. Yang, W.; Li, J.; Peng, W.; Deng, A. A rub-impact recognition method based on improved convolutional neural network. Comput. Mater. Contin. 2020, 63, 283–299. DOI:10.32604/cmc.2020.07511
20. Fang, W.; Zhang, W.; Zhao, Q.; Ji, X.; Chen, W. Comprehensive analysis of secure data aggregation scheme for industrial wireless sensor network. Comput. Mater. Contin. 2019, 61, 583–599. DOI:10.32604/cmc.2019.05237
21. Villa, M., Dardenne, G., Nasan, M., Letissier, H., Hamitouche, C., & Stindel, E. FCN-based approach for the automatic segmentation of bone surfaces in ultrasound images. International journal of computer assisted radiology and surgery, 2018, 13, 1707-1716. DOI:10.1007/s11548-018-1856-x
22. Zhu, X., Cheng, Z., Wang, S., Chen, X., & Lu, G. Coronary angiography image segmentation based on PSPNet. Computer Methods and Programs in Biomedicine, 2021, 200, 105897. DOI:10.1016/j.cmpb.2020.105897
23. Singh, N. J., & Nongmeikapam, K. (2023). Semantic segmentation of satellite images using deep-unet. Arabian Journal for Science and Engineering, 48(2), 1193-1205. DOI:10.1007/s13369-022-06734-4
24. Mao, Y., Ren, W., Li, X., Yang, Z., & Cao, W. Sep-RefineNet: A Deinterleaving Method for Radar Signals Based on Semantic Segmentation. Applied Sciences, 2023, 13(4), 2726. DOI:10.3390/app13042726
25. Nanfack, G., Elhassouny, A., & Thami, R. O. H. Squeeze-SegNet: a new fast deep convolutional neural network for semantic segmentation. In Tenth International Conference on Machine Vision (ICMV 2017) ,2018.04, (Vol. 10696, pp. 703-710). SPIE. DOI:10.1117/12.2309497
26. Wang, X., Hu, Z., Shi, S., Hou, M., Xu, L., & Zhang, X. A deep learning method for optimizing semantic segmentation accuracy of remote sensing images based on improved UNet. Scientific reports, 2023, 13(1), 7600. DOI:10.1038/s41598-023-34379-2

27. He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., & Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60, 1-15. DOI:10.1109/TGRS.2022.3144165
28. Xu, Z., Zhang, W., Zhang, T., Yang, Z., & Li, J. Efficient transformer for remote sensing image segmentation. Remote Sensing,2021, 13(18), 3585. DOI:10.3390/rs13183585
29. Azad, R., Heidari, M., Shariatnia, M., Aghdam, E. K., Karimijafarbigloo, S., Adeli, E., & Merhof, D. Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation. Cham: Springer Nature Switzerland, In International Workshop on PRedictive Intelligence In Medicine, 2022.09, (pp. 91-102). DOI:10.1007/978-3-031-16919-9_9
30. He, C., Liu, Y., Wang, D., Liu, S., Yu, L., & Ren, Y. Automatic extraction of bare soil land from high-resolution remote sensing images based on semantic segmentation with deep learning. Remote Sensing, 2023, 15(6), 1646. DOI:10.3390/rs15061646
31. Zheng, Z., Zhang, X., **ao, P., & Li, Z. Integrating gate and attention modules for high-resolution image semantic segmentation. Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14, 4530-4546.
32. DOI:10.1109/JSTARS.2021.3071353
33. Zhao, D., Zhu, C., Qi, J., Qi, X., Su, Z., & Shi, Z. Synergistic attention for ship instance segmentation in SAR images. Remote Sensing, 2021, 13(21), 4384. DOI:10.3390/rs13214384
34. Wu, H., Huang, Z., Zheng, W., Bai, X., Sun, L., & Pu, M. SSGAM-Net: A Hybrid Semi-Supervised and Supervised Network for Robust Semantic Segmentation Based on Drone LiDAR Data. Remote Sensing, 2023, 16(1), 92. DOI:10.3390/rs16010092
35. Liu, Y., Zhu, Q., Cao, F., Chen, J., & Lu, G. High-resolution remote sensing image segmentation framework based on attention mechanism and adaptive weighting. ISPRS International Journal of Geo-Information, 2021, 10(4), 241. DOI:10.3390/ijgi10040241
36. Ma, B., & Chang, C. Y. Semantic segmentation of high-resolution remote sensing images using multiscale skip connection network. IEEE Sensors Journal, 2021, 22(4), 3745-3755. DOI:10.1109/JSEN.2021.3139629
37. He, C., Liu, Y., Wang, D., Liu, S., Yu, L., & Ren, Y. Automatic extraction of bare soil land from high-resolution remote sensing images based on semantic segmentation with deep learning. Remote Sensing, 2023, 15(6), 1646. DOI:10.3390/rs15061646
38. Chen, M., Yang, B., Wang, F., Guo, Y., & Duan, T. Identification of open-pit mines and surrounding vegetation on high-resolution satellite images based on improved bilateral segmentation network semantic segmentation model. Journal of Applied Remote Sensing, 2023, 17(4), 044518-044518. DOI:10.1117/1.JRS.17.044518
39. Lin, X., Cheng, Y., Chen, G., Chen, W., Chen, R., Gao, D., ... & Wu, Y. Semantic Segmentation of China's Coastal Wetlands Based on Sentinel-2 and Segformer. Remote Sensing, 2023, 15(15), 3714. DOI:10.3390/rs15153714
40. Wang, Z., Wang, J., Yang, K., Wang, L., Su, F., & Chen, X. Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3+. Computers & Geosciences, 2022, 158, 104969.
41. DOI:10.1016/j.cageo.2021.104969
42. Sun, Y., Bi, F., Gao, Y., Chen, L., & Feng, S. A multi-attention UNet for semantic segmentation in remote sensing images. Symmetry,2022, 14(5), 906. DOI:10.3390/sym14050906
43. Yuan, X., Chen, Z., Chen, N., & Gong, J. Land cover classification based on the PSPNet and superpixel segmentation methods with high spatial resolution multispectral remote sensing imagery. Journal of Applied Remote Sensing, 2021, 15(3), 034511-034511.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.