ISSN 2616-5775 Vol. 5, Issue 14: 113-125, DOI: 10.25236/AJCIS.2022.051417

Building Image Segmentation Method with Multi-Attention Mechanism

Xinyang Tian[1], Mingkun Xu[1,*]

[1]School of Computer Science (National Pilot Software Engineer School), Beijing University of Posts and Telecommunications, Beijing, China
[*]Corresponding author

Abstract: *In order to solve the problem of inaccurate edge segmentation and loss of small buildings caused by UNet which is difficult to take into account both global features and local features, CSUNet is proposed based on coordinate attention and self-attention. The CSUNet fuses the coordinate attention in the encoder, designs a Double-channel Skip Connection Transformer (DSCT) model in the skip connection, and designs a feature fusion module (FFM) based on CBAM channel attention to fuse the output of the skip connection with the upsampling result of the decoder. The model is tested on the instance dataset of typical Chinese cities and the WHU East Asia satellite dataset. On the instance dataset of typical Chinese cities, PA reaches 0.9390 and IoU reaches 0.8227, on the WHU East Asia satellite dataset, PA reaches 0.9847 and IoU reaches 0.8332. Compared with UNet, all indicators are improved. Visually, CSUNet can more accurately extract building details such as edges and corners, and can extract the location and contour of small buildings. Experiments show that CSUNet can improve the performance of building feature extraction.*

Keywords: *building image segmentation, UNet model, coordinate attention, double-channel skip connection transformer, feature fusion*

1. Introduction

In recent years, with the development of remote sensing technology, more and more remote sensing satellites have been launched, and the quality of high-resolution remote sensing images has been increasing. Based on the development of deep learning technology and remote sensing images has caused many concerns in the field of remote sensing [1]. The accurate automatic extraction of buildings from high-resolution remote sensing images has extremely important significance in urban planning, map data update, emergency response, etc. [2]. Traditional building segmentation methods are mostly based on shape, edge, texture, etc., and their effects are not satisfactory. With the development of deep learning, many deep learning algorithms have been proposed. At present, using these algorithms and high-resolution remote sensing images to recognize building characteristics has become a hot research direction in computer vision, digital image processing, etc. Remote sensing images have complex land features. How to quickly and efficiently identify targets from massive remote sensing image data is still a complex and difficult task. The extraction of building edge features has become a bottleneck problem in remote sensing image building feature extraction. There is an urgent need for a more accurate building feature extraction algorithm. In order to further solve the problems of noise interference and unclear edge segmentation faced by building segmentation, this paper proposes a feasible model improvement scheme based on the classical UNet architecture, and conducts building feature extraction experiments to prove that the improved model can effectively improve the accuracy of building segmentation.

With the advent of Convolutional Neural Networks (CNN), neural network methods based on CNN have been widely used in tasks such as semantic segmentation, object detection, and text recognition. In 2015, the UNet [3] network was proposed, which is based on the full convolutional encoding-decoding structure and strengthens the fusion of shallow and deep networks through skip connections. It has good semantic segmentation effects. The Resnet [4] uses shortcut connections to solve the problem of network degradation and deep network gradient disappearance and gradient explosion. Chen et al. proposed DeepLab V3 [5], which proposed cascading convolution with different spatial rates to extract multi-scale context, and the architecture protected the details of the image when the image was restored from the target feature to the same resolution image. Zhou et al. proposed the UNet++ [6] network, which combines long and short connections on the basis of U-Net and integrates them by means of feature superposition, integrating the semantic gap between the features of the encoder and decoder. The literature [7] proposes the parallel path neural network (MAP-Net), which preserves spatial information and extracts high-level semantic features of multiple scales through parallel multi-path learning to improve feature extraction accuracy. The DANet [8] based on the dual attention mechanism captures the global feature dependencies of the spatial and channel dimensions, and has good recognition effects on image details.

In 2017, the Google research team proposed the Transformer [9] based on Self-Attention, which has made great advantages in natural language processing. In 2020, the proposal of DERT [10] and ViT [11] for object detection and image classification models respectively has enabled the application of Transformer in the field of computer vision. However, due to the large number of parameters and high computational cost of the Transformer model, many scholars have begun to introduce successful prior knowledge from CNNs into the Transformer, including locality, hierarchy, multi-scale, residual connection, and inductive bias, etc. The Swin Transformer [12] proposed by Microsoft Research Asia uses a moving window and restricts the use of self-attention within the window, achieving good results.The CSWin Transformer [13] model replaces the W-MSA and SW-MSA modules in Swin Transformer with a cross-local window, enabling the computation of local attention across windows. Cao et al. proposed the Swin-Unet [14] network, which is a UNet network with a Transformer structure that learns local and global features. Oktay et al. proposed the Attention UNet [15], which connects an Attention Gate structure at the end of the UNet's skip connections to enable the attention mechanism on feature maps. Gao et al. proposed the UTNet, which uses a multi-headed self-attention module structure and combines it with relative position encoding to reduce model complexity. However, in the task of building extraction, the extraction of building edge details and the recognition of small buildings are still challenging problems. Therefore, this paper proposes a CSUNet model based on the UNet model and fused with the attention mechanism for building segmentation. The main contributions of this paper are as follows: introducing coordinate attention (CA) [16] module in the encoding stage to better segment building edge detail information. Designing a Double-channel Skip-Connection Transformer (DSCT) module based on Transformer to improve the ability to extract global features of images. Designing a feature fusion module based on CBAM channel attention module to improve the fusion of transformer feature output and the features of the upsampling stage.

2. Related Work

2.1. UNet Model

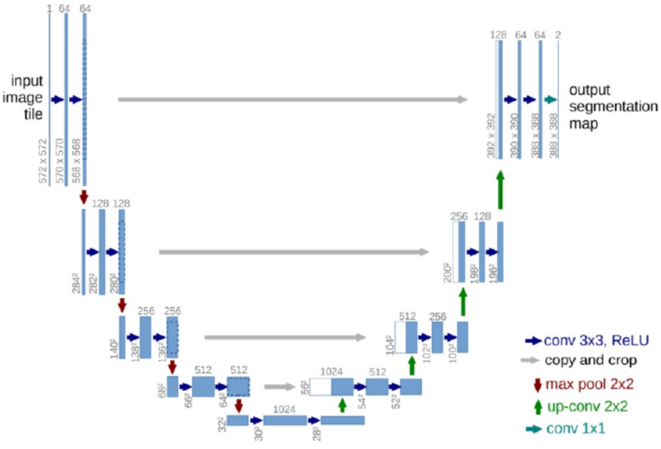


Figure 1: UNet network model structure.

UNet, as an excellent image segmentation network, is shown in Figure 1. By implementing a encoding-decoding structure, it extracts and recovers multi-scale features and designs skip connections to combine features at different scales for extraction and analysis. In the downsampling stage, the feature map passes through a convolution module and is activated by ReLU, then a 2x2 maximum pooling operation is performed to reduce the size of the feature map to half of its original size, increasing the feature depth layer by layer. In the upsampling process, the feature map from the downsampling stage is added to the upsampling process, fusing high-dimensional and low-dimensional features, complementing network details, reducing detail loss, and ultimately restoring the original image size. The deep network can obtain image detail information, and the shallow network can obtain deep information, which allows the network to better understand the image. The symmetrical structure of the encoding and decoding layers not only enhances the effect of the skip connections, but also makes the input and output sizes the same. The skip connections concatenate the features of the encoding sub-network with the features of the decoding sub-network, which can prevent the network from becoming too deep and causing gradient explosion or gradient disappearance, and at the same time can help the encoding-decoding process to recover the full spatial resolution.

2.2. Coordinate Attention

Coordinate attention not only considers the relationship between channels, but also considers the position information of the feature space. Its structure is shown in Figure 2, which consists of two steps: coordinate information embedding and coordinate attention generation.

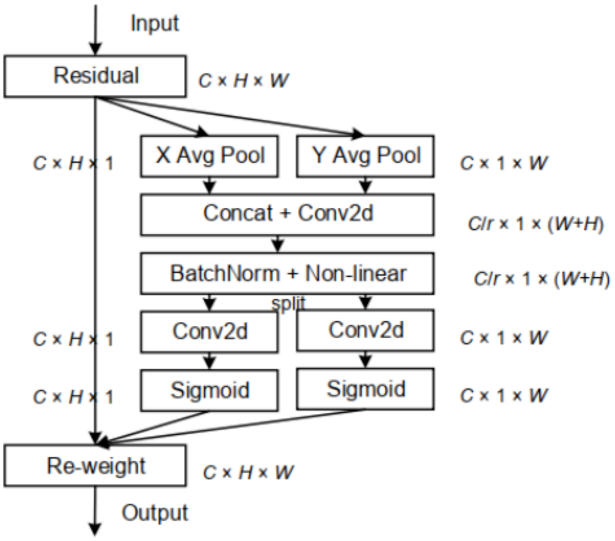


Figure 2: Coordinate attention structure.

In the coordinate information embedding part: first, the feature map is passed through channel attention, and the channel attention is averaged in the horizontal and vertical directions to obtain two directional feature maps, corresponding to X Avg Pool and Y Avg Pool in Figure 2. The input feature map is passed through pooling kernal of size (H, 1) and (1, W) respectively along the horizontal and vertical directions to perform average pooling encoding, so as to decompose the channel attention into two one-dimensional feature encodings. The output of the c-th channel of height h is obtained as formula (1).

$$z_c^h(h) = \frac{1}{W}\sum_{0\le i\le W} x_c(h,i) \qquad (1)$$

The output of the c-th channel of width w can be represented as formula (2).

$$z_c^w(w) = \frac{1}{H}\sum_{0\le j\le H} x_c(j,w) \qquad (2)$$

Where W is the width of the feature map, and H is the height of the feature map. One spatial direction is used to capture long-range dependencies, and the other spatial direction is used to preserve accurate position information, enhancing the ability of feature representation and enabling the network to more accurately locate the target of interest. The module is able to capture spatial long-range dependencies with accurate position information.

Coordinate attention generation: the coordinate information is embedded into two partially generated feature maps based on channel concatenation, then a 1 x 1 convolution is used to generate an intermediate feature map, which is split into two independent feature maps along the spatial dimension, and then each feature map undergoes a 1 x 1 convolution to change the number of channels to the original input number of channels. After Sigmoid, the horizontal and vertical spatial direction attention weights are obtained. Finally, the original input feature map is multiplied by this weight to obtain the coordinate attention feature map.

2.3. Transformer

Transformer is a encoder-decoder model based on Self-Attention, which is good at establishing long distance dependency and has become the mainstream model in natural language processing field. It consists of multiple encoding and decoding units, as shown in Figure 3.

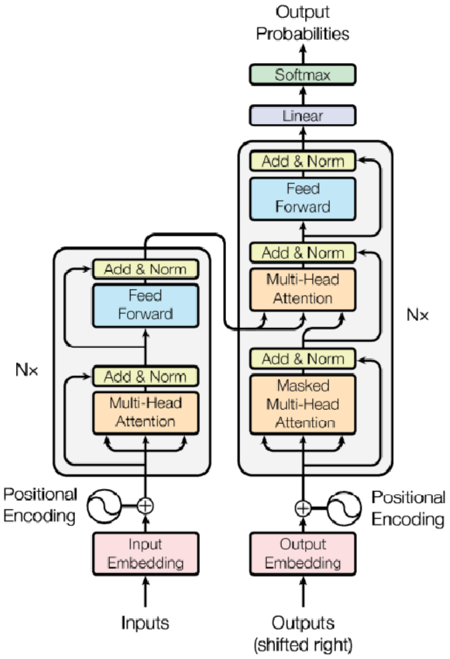


Figure 3: Transformer structure.

Transformer was proposed by Vaswani et al. to address the parallel computation problem of recurrent neural networks in natural language processing. Inspired by the great success of Transformer in the natural language processing field, in recent years, some pioneering work has begun to study how to apply Transformer to the field of computer vision and has achieved significant results. Currently, visual Transformer is still a research hotspot [17]. Dosovitskiy et al. proposed the Vision Transformer (ViT) model, which introduced the Transformer model into the field of computer vision for the first time. The Transformer structure in the ViT network is shown in the Figure 4.

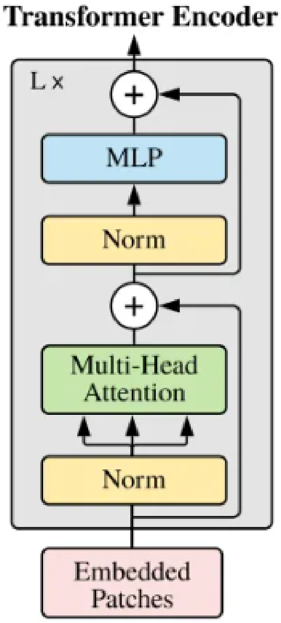
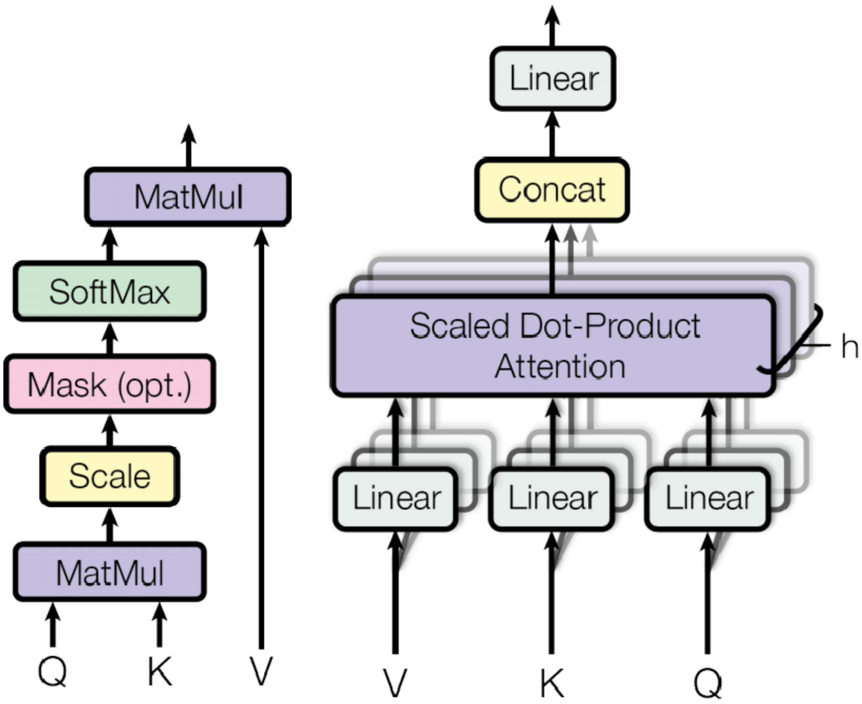


Figure 4: Vit encoder structure.

The self-attention computation process is shown in Figure 5 (a). Three matrices Wq, Wk, and Wv are defined, and all input vectors are linearly transformed three times using these matrices, and all input vectors are derived into three new vectors qt, kt, and vt. All qt vectors are concatenated into a matrix, which is denoted as the query matrix Q, all kt vectors are concatenated into a matrix, which is denoted as the key matrix K, and all vt vectors are concatenated into a matrix, which is denoted as the value matrix V. The query vector q1 of the first vector is multiplied by the key matrix K to obtain the attention weight of the first vector. The obtained value is passed through Softmax, and after obtaining the weight, the weight is multiplied by the corresponding value vector vt of the sequence, and these weighted value vectors are summed to obtain the output of the first vector. The transformation into matrix operation is used to calculate the attention feature map through dot-product attention, that is, the self-attention calculation formula is obtained,

$$X_{attention} = Self - Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \qquad (3)$$

Defining multiple sets of Q, K, and V so that they focus on different contexts can obtain the multi-head attention, where the structure of the multi-head attention is shown in Figure 5 (b).



(a) self-attention. (b) multi-head attention.

Figure 5: Attention mechanism.

3. Method of This Paper

3.1. Main Network

As shown in Figure 6, the overall structure of the network in this paper is based on the UNet framework, which mainly includes an encoder, a double-channel skip-connection transformer (DSCT) module, a feature fusion module (FFM), and a decoder. The input image size of the model is 224 x 224 x 3. Based on the UNet backbone network, there are 4 levels in the encoding stage, and each level performs 2 convolution operations, and down conv is used to sample down to the next level. In the encoding stage, the coordinate attention mechanism is introduced, and in the skip connection stage, the Transformer is introduced.

In the encoding stage, the second convolution of the first and second levels are replaced with a coordinate attention module. The image first goes through a convolution module to extract the initial features of the image, and then through the coordinate attention module to capture the spatial long-range dependencies and position information of the building, in order to better learn the details of the building edges. As the depth of feature extraction increases, the details of the learned building features will continue to improve.

The skip connection part is replaced with the DSCT module, which can perform a Self-Attention-based Transformer operation on the feature map along the channel direction to better capture global semantic information. The output of the first and third levels is input to the DSCT1 module, and the output of the second and fourth levels is input to the DSCT2 module, which realizes the fusion of high-dimensional and low-dimensional feature information, enabling the network to better understand the image.

The information captured by DSCT and the upsampling result of the decoding stage are input to the feature fusion module FFM, which reduces the semantic gap between the Transformer and the decoding stage feature map. The result of the feature fusion is concatenated with the decoding upsampling feature map based on the channel and performs the decoding stage convolution operation. The decoding stage is also divided into four levels, each of which includes two convolution operations and a 2x2 upsampling operation, ultimately achieving the restoration of the original image size.

The UNet model has a convolution kernel size of 3 x 3 for convolution operations. When extracting features, the receptive field is limited and can only produce local perception. However, the Transformer in the improved model's DSCT module, based on the multi-head self-attention mechanism, can establish a long-distance pixel dependence and has strong global information extraction capabilities. By using the Transformer operation to extract long-range information and fuse global and local features, the network can better extract building edge details, bringing a significant improvement in building segmentation accuracy.

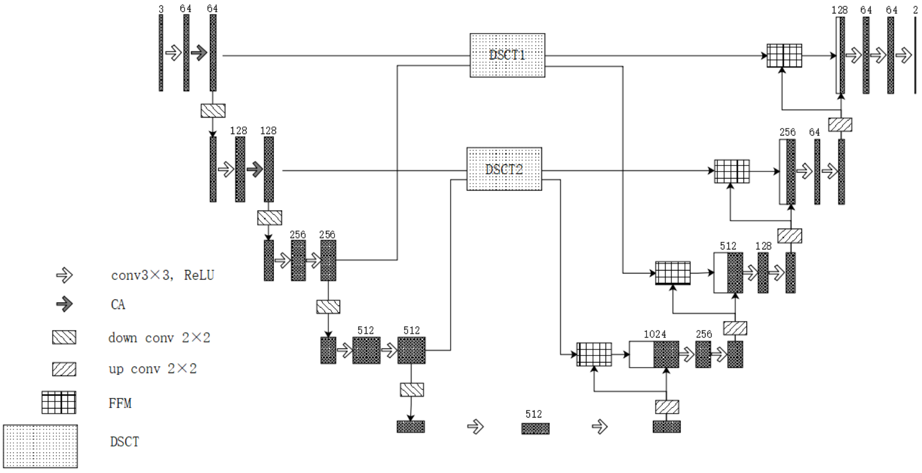


Figure 6: CSUNet network structure.

3.2. DSCT

In order to overcome the inherent limitations of the receptive field of UNet network convolution operations and further enhance the model's long-distance dependency capabilities, this paper proposes the DSCT module, as shown in Figure 7.

3.2.1. DSCT Module Framework

This module combines the Transformer model and introduces the Transformer in the skip connection stage. By using the self-attention mechanism to expand the receptive field of feature extraction, the output of the first and third levels of the encoding stage is input into DSCT1, as shown in Figure 7, to perform the self-attention mechanism-related operation and obtain the two outputs d1, d3 of the DSCT1 module. Similarly, the output of the second and fourth levels of the encoding stage is input into DSCT2, and the corresponding outputs d2, d4 are obtained. The outputs of the two DSCT modules are input into the decoding stage, where they are fused with the features of the decoding stage and connected to participate in the image recovery process of the decoding stage.

The DSCT module has two inputs, which are the outputs of two layers in the encoding stage. First, the two input features are transformed into vector sequences by splitting them into patches, and the ratio of the sizes of the two input features and the patch size remains the same. Positional encoding is added, which provides the position information of each pixel. The dimension of the positional embedding is the same as the dimension of the vector sequence after the input feature is split into patches. The position value of the input vector is added to the position embedding. The resulting sequence is concatenated and fed into a layer normalization module. The layer normalization module normalizes the hidden layer of the neural network to a standard normal distribution, which speeds up the training and convergence. The sequence is then fed into an improved multi-headed attention module and finally into a residual structure of multiple perceptron layers. The module produces two outputs, which are fed into the next attention module for multiple iterations.

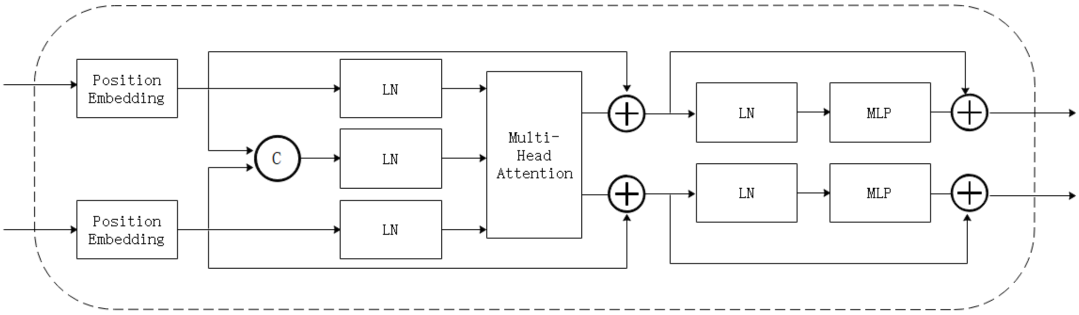


Figure 7: DSCT module structure.

3.2.2. Improved Multi-head Self-Attention Module

In order to enable the DSCT module to perform Transformer operations on two different feature layers simultaneously along the channel dimension, the multi-headed self-attention module of the Transformer operation is improved as follows. The vector sequences transformed by splitting and adding positional embedding and passing through layer normalization are denoted as E1 and E2. The improved self-attention mechanism module is shown in Figure 8.

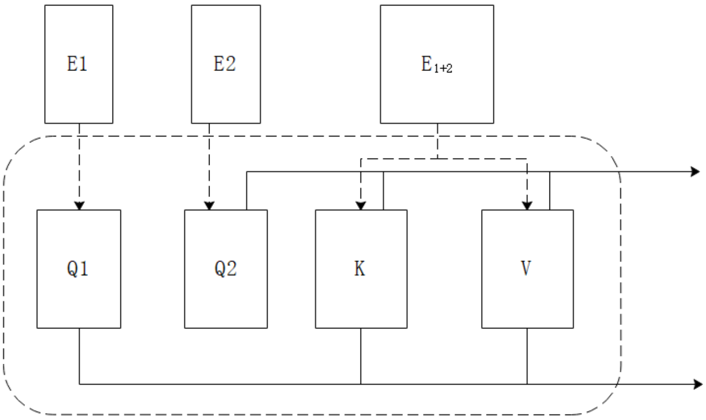


Figure 8: Improved multi-head self-attention module.

Figure 8 shows three inputs: E1, E2, and E1+2, where E1+2 is the concatenation of E1 and E2. Three matrices, Wq1, Wq2, Wk, and Wv, are defined and used to perform four linear transformations on the input vectors E1, E2, and E1+2, respectively.

$$Q_1 = Linear_{q1}(E_1) = E_1 W_{Q1}$$

$$Q_2 = Linear_{q2}(E_2) = E_2 W_{Q2} \qquad (4)$$

$$K = Linear_k(E_{1+2}) = E_{1+2} W_K$$

$$V = Linear_v(E_{1+2}) = E_{1+2} W_V$$

All input vectors produce four new vectors: q1t, q2t, kt, and vt. All q1t vectors are concatenated into a matrix called the query matrix Q1 of E1, all q2t vectors are concatenated into a matrix called the query matrix Q2 of E2, all kt vectors are concatenated into a matrix called the key matrix K, and all vt vectors are concatenated into a matrix called the value matrix V. The query vector q1 of the first vector is multiplied by the key matrix K to obtain the attention weight of the first vector, and the obtained value is passed through a Softmax function. After obtaining the weight, the weight is multiplied by the corresponding sequence value vector vt, and the weighted value vectors are summed to obtain the output of the first vector. In matrix operation, the attention feature map is calculated by dot product attention, that is, the self-attention calculation formula is obtained. By defining multiple sets of Q1, Q2, K, and V to focus on different contexts, the multi-headed attention mechanism can be obtained.

$$E_{attentioni} = Self - Attention(Q_i, K, V) = softmax(\frac{Q_i^T K}{\sqrt{\sum d_k}})V^T \tag{5}$$

$$= softmax(\frac{W_{Q_I}^T E_i^T E_{1+2} W_K}{\sqrt{d_1 + d_2}})W_V^T E_{1+2}^T$$

The improved self-attention result is added to the input and connected with a residual connection, and then passed through layer normalization and an MLP.

3.3. Feature Fusion Module

This paper proposes a feature fusion module. The design idea of the feature fusion module comes from the channel attention module in the CBAM [18], as shown in Figure 9. The channel attention module provides spatial information of the feature through MaxPool and AvgPool, and then inputs the two features into a multi-layer perceptron (MLP) with hidden layers to obtain the sum of the operation results, and then obtain the channel attention Mc through the Sigmoid operation.

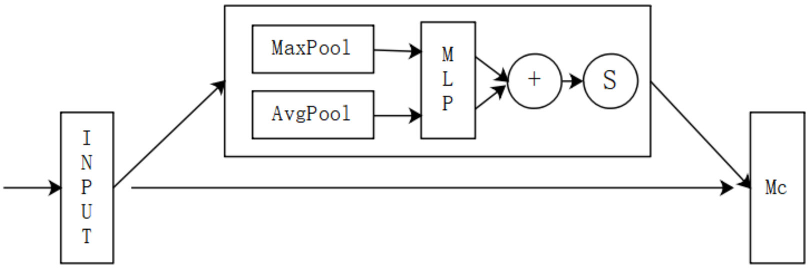


Figure 9: CBAM channel attention.

Because there is a semantic gap between the output features of the DSCT module and the features generated by upsampling on the decoding end, this feature fusion module is designed to reduce the semantic gap and better perform feature fusion, as shown in Figure 10. The module has two inputs, and the results obtained after the average pooling and MLP are added, and then the weight of the corresponding feature fusion of the module is obtained through the Sigmoid. The weight is multiplied by the input to obtain the output result. The result is concatenated with the feature of the upsampling process.

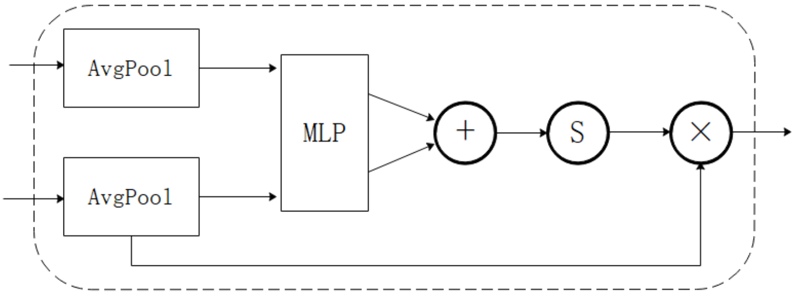


Figure 10: Feature fusion module.

4. Analysis of Experimental Results

The experimental environment of this paper is running on a 24-core CPU, 80GB memory, NVIDIA GeForce RTX 3090 GPU computing platform, and the development environment is PyTorch1.10.0, Python 3.8.10, cuda11.3.

4.1. Experiment Dataset

The experimental results of the CSUNet on the typical Chinese city building instance dataset and the WHU East Asian satellite dataset with high spatial resolution and better label quality are explored to investigate the effect of CSUNet on the remote sensing image building extraction task with different resolutions and different label quality. The Chinese typical city dataset is selected from four representative city centers in Beijing, Shanghai, Shenzhen, and Wuhan as the target area for data collection. The dataset covers a total of 120 square kilometers. The East Asian satellite dataset consists of six adjacent satellite images covering 550 km² in East Asia with a ground resolution of 2.7 m. The vector building map contains a total of 29,085 buildings. The entire image is seamlessly cut into 17,388 512x512 small blocks for training and testing. The two datasets are divided into training sets, validation sets, and test sets according to the ratio of 6:2:2, as shown in Table 1.

Table 1: Data splitting method.

Dataset	train	validation	test
the typical Chinese city building instance dataset	4549	1436	1275
the WHU East Asian satellite dataset	3135	462	441

4.2. Evaluation Index

In order to quantitatively evaluate the performance of the network, PA, mPA, mIoU, Recall, and F1 Score are selected. Precision focuses on whether there is a false alarm in the result, and recall focuses on whether there is a missed result. F1 score is an important evaluation criterion for measuring the accuracy of binary classification in the field of computer science. It takes into account both the precision and overlap of the classification results, that is, the ratio of their intersection to the union, and the ratio when they are completely overlapped is 1. TP represents True Positive, the number of pixels predicted to be positive class; FP represents False Positive, the number of pixels predicted to be positive class but negative class; FN represents False Negative, the number of pixels predicted to be negative class but positive class; TN represents True Negative, the number of pixels predicted to be negative class.

PA (Pixel Accuracy): Pixel accuracy is calculated by taking the proportion of correctly classified pixels for each class, summing them up, and then taking the average.

$$\mathrm{PA} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{6}$$

mPA (Mean Pixel Accuracy): Classaverage pixel accuracy is calculated by taking the proportion of correctly classified pixels for each class (CPA), summing them up, and then taking the average.

$$\mathrm{CPA} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{7}$$

mIoU:Mean Intersection over Union is the average of the ratios of the intersection of the model's predicted and true values for each class to their union, summed over all classes.

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP+FP+FN}} \tag{8}$$

Recall:Recall is the proportion of samples that were correctly identified as buildings in the dataset.

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{9}$$

F1:The F1 score is used to measure the Precision and Recall of a model.

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision+Recall}} \tag{10}$$

4.3. Comparative Analysis of Experimental Results

This paper presents a comparative experiment between UNet and CSUNet networks, demonstrating the effectiveness of CSUNet. The results of building extraction from high-resolution building datasets of typical Chinese city building instances and the WHU East Asia satellite dataset are shown in Figure 11 and Figure 12, respectively, which illustrate the validation set's PA, mPA, mIoU, Recall, and F1 score change curves during the training process of UNet and CSUNet networks. Quantitative analysis shows that CSUNet has improved PA, mPA, mIoU, Recall, and F1 score compared to the UNet network.

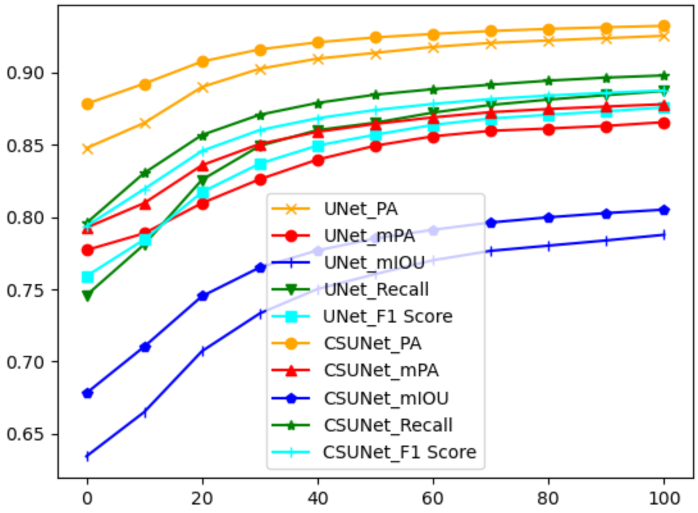


Figure 11: Experimental results of the typical Chinese city building instance dataset - Evaluation indicators.

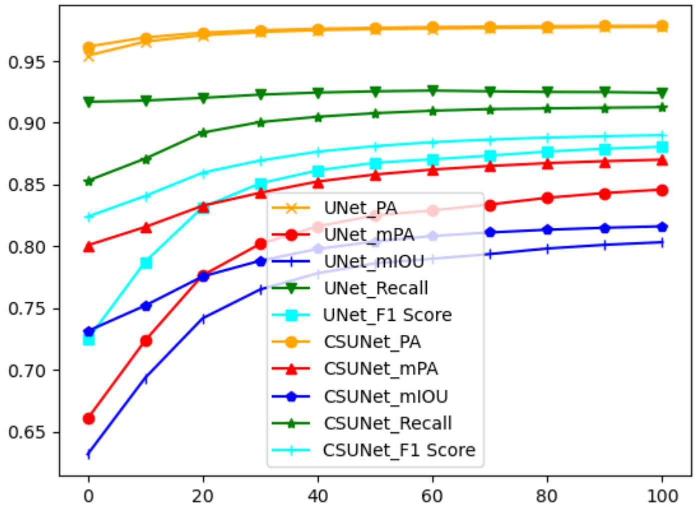


Figure 12: Experimental results of the WHU East Asian satellite dataset - Evaluation indicators.

As shown in Figure 13 and Figure 14, the extraction experiments were performed on the two datasets using CSUNet and UNet, respectively. In the results of the extraction of the typical Chinese city building instances dataset and the WHU East Asia satellite dataset, CSUNet showed more stability and higher detail retention. UNet has difficulty in extracting buildings with significant size differences, and tends to blur when extracting edge corners. For edges and corners, CSUNet accurately extracts building details such as edges and corners. It can be seen that the CSUNet result has sharper corners and clearer edges, and is closer to the labeled visual effect than UNet, proving that CSUNet can better understand the relationship between images and labels. For the extraction effect of buildings of different scales, please see the red box in the figure. CSUNet completely extracts the houses with a large difference in scale from the surrounding houses, which UNet cannot do. Although UNet extracts the location of the houses, there are fuzzy areas and it is almost impossible to confirm the number and contour of the houses. CSUNet can well observe and clearly restore these details in the result figure. Unavoidable buildings of different scales will appear in remote sensing images. CSUNet can accurately extract the location and contour of small and medium-sized buildings. These prove that CSUNet can extract the contour and detailed information of buildings under different scales and complexities, and prove the performance of the network.

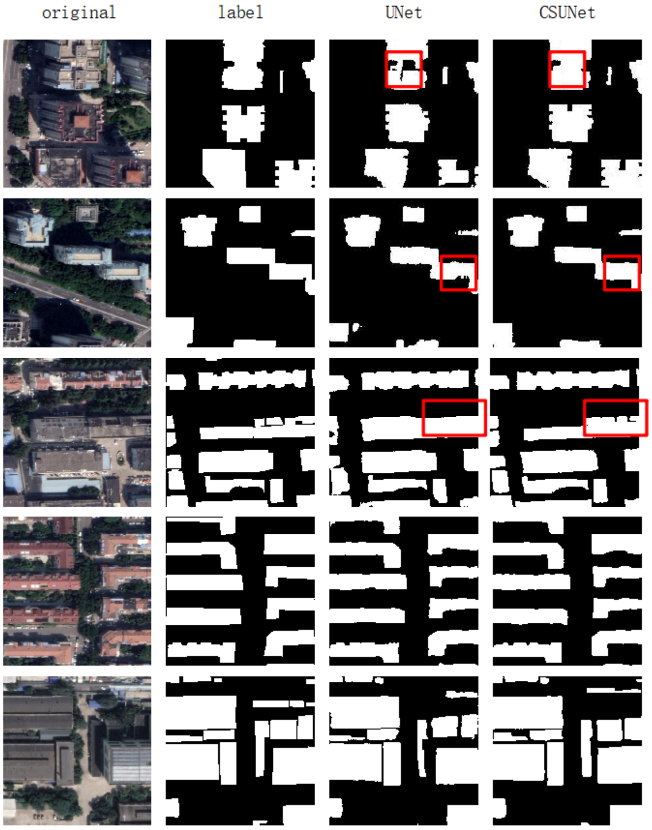


Figure 13: Experimental results of the typical Chinese city building instance dataset.

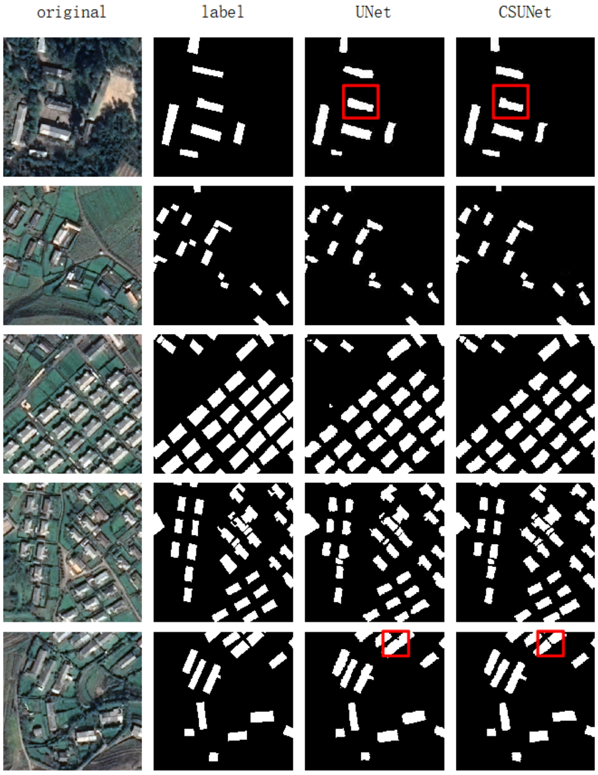


Figure 14: Experimental results of the WHU East Asian satellite dataset.

4.4. Ablation Experiment of the Attention Mechanism

To further analyze the reasons for the performance improvement and verify the enhancement effect of the self-attention module on the UNet accuracy, five groups of module ablation experiments were conducted to prove the improvement of each module on the extraction accuracy of the deep learning network. In this paper, the following ablation experiments were designed on two datasets: UNet model, UNet_CA containing only CA attention module, UNet_DSCT containing only DSCT module, UNet_DSCT_FFM containing DSCT module and feature fusion module, and our network CSUNet. The experimental results on the test set of the typical Chinese city building instance are shown in Table 2.

Table 2: The typical Chinese city building instance test set experimental results.

Method	PA	mPA	mIoU	Recall	F1 Score
UNet	0.9269	0.8732	0.8042	0.8944	0.8832
UNet_CA	0.9363	0.8749	0.8102	**0.9089**	0.8905
UNet_DSCT	0.9360	0.8768	0.8103	0.9065	0.8906
UNet_DSCT_FFM	**0.9397**	0.8901	0.8223	0.9080	0.8986
CSUNet	0.9390	**0.8966**	**0.8227**	0.9014	**0.8989**

The experimental results on the WHU East Asian satellite test set are shown in Table 3.

Table 3: The WHU East Asian satellite test set experimental results.

Method	PA	mPA	mIoU	Recall	F1 Score
UNet	0.9834	0.8661	0.8194	0.9216	0.8917
UNet_CA	0.9846	0.8812	**0.8340**	0.9233	0.9015
UNet_DSCT	0.9839	0.8764	0.8264	0.9199	0.8968
UNet_DSCT_FFM	0.9841	0.8765	0.8274	0.9211	0.8974
CSUNet	**0.9847**	**0.8814**	0.8332	**0.9240**	**0.9015**

Observing the five groups of experiments, it can be seen that compared with the original U-shaped network, the evaluation indicators are improved to different degrees after the introduction of various modules proposed in this paper. Among them, the segmentation effect is the best when CA, DSCT and FFM modules are used simultaneously in the U-shaped network. Compared with the UNet network, UNet_CA with the coordinate attention module performs better on two datasets, indicating that the coordinate attention module can help improve the segmentation accuracy. Compared with the network with the added DSCT module and the UNet network, UNet_DSCT also performs better on two datasets, which means that the global information features extracted by the Transformer supplement to the decoding end can effectively improve the segmentation accuracy. For the network with the added DSCT module and the added feature fusion module, the segmentation effect can also be further improved by the feature fusion between the DSCT module and the decoding end. Comparing the results with the CSUNet results, the feature extraction network using the coordinate attention module, the DSCT module and the feature fusion module can improve the result accuracy.

5. Conclusion

This article analyzes the UNet network and attention mechanism in the current deep learning field, improves the UNet network model, and proposes a CSUNet network model library that integrates the coordinate attention mechanism and self-attention mechanism. Enhance the edge detection accuracy of building feature extraction in remote sensing images and improve the accuracy of feature extraction. The architecture embeds the CA attention mechanism module in the sampling phase of the UNet network model to change the resource allocation method, making more resources inclined to building details, and in the module training process, assigns more weight parameters to the edge details of the image, improves the performance of feature extraction, and reduces the influence of background noise. In the jumping connection phase, the DSCT module is designed to fuse the improved self-attention mechanism to expand the field of feature extraction and design the feature fusion module to fuse the output of the jumping connection DSCT module with the result of the up-sampling in the decoding phase, improving the segmentation effect.

References

[1] Xiaofei HE, Zhengrong ZOU, Chao TAO, et al. Combined Saliency with multi-convolutional neural network for high resolution remote sensing scene classification [J]. Acta Geodaetica et Cartographica

Sinica, 2016, 45(9): 1073.

[2] Wang Z, Zhou Y, Wang S, et al. House building extraction from high resolution remote sensing image based on IEU-Net[J]. J. Remote Sens, 2021.

[3] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[5] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.

[6] Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation [M]//Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, Cham, 2018: 3-11.

[7] Zhu Q, Liao C, Hu H, et al. MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery [J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 59(7): 6169-6181.

[8] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3146-3154.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

[10] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Springer, Cham, 2020: 213-229.

[11] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:2010.11929, 2020.

[12] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[13] Dong X, Bao J, Chen D, et al. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows [J]. 2021.

[14] Cao H, Wang Y, Chen J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation [J]. arXiv preprint arXiv:2105.05537, 2021.

[15] Oktay O, Schlemper J, Folgoc L L, et al. Attention u-net: Learning where to look for the pancreas [J]. arXiv preprint arXiv:1804.03999, 2018.

[16] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.

[17] Gao Y, Zhou M, Metaxas D N. UTNet: a hybrid transformer architecture for medical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021: 61-71.

[18] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module[C]// European Conference on Computer Vision. Springer, Cham, 2018.