ResAt-UNet: A U-Shaped Network Using ResNet and Attention Module for Image Segmentation of Urban Buildings

Zhiyong Fan, Yu Liu, Min Xia, *Member, IEEE*, Jianmin Hou, Fei Yan, and Qiang Zang

***Abstract*—Architectural image segmentation refers to the extraction of architectural objects from remote sensing images. At present, most neural networks ignore the relationship between feature information, and there are problems such as model overfitting and gradient explosion. Thus, this article proposes an improved UNet based on ResNet34 and Attention Module (ResAt-UNet) to solve the related problems. The algorithm adds a two-layer residual structure (BasicBlock) and a regional enhancement attention mechanism (Space Enhancement Area Enhancement, SEAE) to the original framework of UNet, which enhances the network depth, improves the fitting performance, and extracts small objects more accurately. The experimental results show that the network has achieved MIOU of 78.81% in the Massachusetts dataset, and the newly developed model outperforms UNet in both quantitative and qualitative aspects.**

***Index Terms*—Attention mechanism, high-resolution remote sensing, images segmentation, neural network.**

I. Introduction

REMOTE sensing satellite image is an efficient means of obtaining geospatial information and data. Compared with traditional aerial photography, it has unique advantages. Remote sensing satellites have a large monitoring area and can transmit, process, and dynamically monitor data in real time. The most important thing about research on remote sensing images of urban areas is how to effectively segment and extract architectural objects in the images. The use of remote sensing satellite technology for urban image segmentation has become an important means of planning cities and studying urban areas. How to obtain building information accurately and dynamically has arisen the interest of researchers. However, due to the wide variety of buildings together with complex and changeable image backgrounds, building image segmentation has always been a thorny problem.

Since the 21st century, image segmentation has always been widely concerned as a research hotspot and many segmentation methods have been proposed. These methods can generally be divided into two categories, traditional methods based on space and features and semantic segmentation methods based on deep learning. Traditional image segmentation methods are usually based on the feature domain and spatial domain and use the prior information of the segmentation target in the image, such as shape, brightness, texture, etc. to obtain the segmentation results. Kalyankar [1] used five different thresholding algorithms to segment remote sensing satellite images and compared their segmentation effects with each other. The best performing method was the histogram and edge maximization thresholding method. Adams et al. [2] developed a supervised learning method that generated label information for plant image segmentation, which outperformed traditional thresholding methods. Pratondo et al. [3] proposed a combination of a region-based dynamic contour model and machine learning for medical ultrasound image segmentation and the newly proposed model achieved higher segmentation accuracy than Chan-Vese [4]. Liow, Pavlidis [5] used edge detection to locate building boundaries and then adopted target area growth to determine the location and area of buildings. Avudaiammal et al. [6] combined morphological, spectral, shape, and geometric characteristic information with support vector machine (SVM) to classify remote sensing images into buildings and nonbuildings using the morphological building index. Yang et al. [7] developed a new method for local spectral angle thresholding, which segmented buildings better than the global thresholding method. In fact, due to the continuous development of remote sensing technology, the image resolution has become higher and the background area has become more complex, but the effect of traditional segmentation is not ideal.

The past decade has witnessed the continuous development of neural network, which is also used in cloud segmentation [8], [9], [10], [11], [12], [9], power transmission system [13], and building segmentation [14], [15], [16], [17], [18], [19], etc. In 2015, Long et al. [20] proposed the fully convolutional network (FCN), which was the first network successfully used for image segmentation. As a result, the researchers developed a series of FCN. Ronneberger et al. [21] developed a symmetric network

Manuscript received 31 July 2022; revised 23 October 2022; accepted 13 January 2023. Date of publication 23 January 2023; date of current version 21 February 2023. This work was supported by the National Natural Science Foundation of PR China under Grant 42075130. *(Corresponding author: Zhiyong Fan.)*

Zhiyong Fan, Min Xia, Jianmin Hou, Fei Yan, and Qiang Zang are with the College of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China, and also with the Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: zhiyongfan1981@163.com; xiamin@nuist.edu.cn; jmhou@nuist.edu.cn; fyan@nuist.edu.cn; zangq@nuist.edu.cn).

Yu Liu is with the College of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: liuyu_nuist@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3238720

with encoder and decoder, UNet, which has been widely used in medical image segmentation, but it may lead to an imbalance between decoding ability and encoding ability. Badrinarayanan et al. [22] proposed SegNet using the max pooling index in the up-sampling stage. Zhao et al. [23] proposed the pyramid scene parsing network (PSPNet), which used a pyramid pooling module to extract contextual information and stacked the contextual information with the extracted feature information to complete image segmentation, urban buildings are obscured by shadows on high-resolution remote sensing images, and the network's extraction of object information is incomplete. Sun et al. [16] proposed a Multi-Attention UNet based on the UNet network, adding a residual encoder with an attention mechanism and a self-attention mechanism. However, due to the complex background of the ground features and irregular boundaries, this method will produce some misclassifications and omissions during extraction. Therefore, accurate segmentation of images was achieved in datasets of WHDLD and DLRSD remote sensing images. Chen et al. [17] proposed a UNet architecture with a self-attention mechanism bias module, which strengthened the weight of the target area in the down-sampling part and used the bias parameter to enhance the model reshaping ability in the up-sampling part. They achieved good segmentation effects on the WHU and Massachusetts urban image datasets. Although the existing convolutional neural networks can achieve good results in the field of architectural image segmentation, most of them ignore the correlation of feature information between different channels. Soni et al. [18] proposed two-scale input-based architecture: Dual-scale CNN (Du-CNN). A Difference of Normals approach is used to isolate 3-D buildings from other objects in densely built-up areas. It shows that most building extraction results have a Precision > 0.9 and favorable Recall and F-score values [19]. Although the above algorithm applied to urban building segmentation has achieved good segmentation results, there are still some shortcomings. With the deepening of the network, the model is prone to over-fitting. In addition, the network still has insufficient extraction of building details.

In order to solve the above problems, this article combines the Space Enhancement Area Enhancement (SEAE) and the two-layer residual module (BasicBlock) to propose a ResNet and Attention Module based on UNet (ResAt-UNet).

1) An encoder-decoder network structure ResAt-UNet is designed and implemented based on UNet that enables end-to-end training, using ResNet34, a fused attention mechanism, as a feature extraction network with enhanced feature extraction capability. Additionally, adding ResNet blocks in the network solves the overfitting problem while increasing network depth.
2) The up-sampling layer of UNet reduces the optimization parameters of the network; the shallow feature map is spliced with the deep feature map, which can facilitate information fusion, reduce the numbers of network parameters, and mitigate the over-fitting phenomenon.
3) The improved CBAM (SEAE) is added to the intermediate connection layer to capture information in the channel and spatial domain and enhances feature representation.
4) Improving loss function DiceLoss and BceLoss instead of the cross-entropy loss function and facilitates information fusion.
5) A series of model comparison and ablation experiments are conducted to verify the effectiveness of the proposed model, which has a better segmentation score compared with other models.

II. Methods

This article uses the UNet proposed by Ronneberger et al. [21] in 2015 as the basic model architecture. The UNet network has a typical encoder-decoder structure. Specifically, after continuous convolution and down-sampling of the image in the encoding stage, the obtained feature map has a small scale but it contains high-dimensional semantic feature information. Then in the decoding stage, the network restores the feature map to the original size through continuous convolution and up-sampling, and finally obtains the segmentation results of the image. The middle part of the network connects the feature map of the encoding and decoding stages through the Concat layer and combines the context information to obtain the final prediction results by continuous up-sampling and feature fusion. We replace the down-sampling part of the UNet network with the residual network ResNet, and add several attention modules to the intermediate connection layer to obtain the improved network model: ResAt-UNet.

A. ResAt-UNet

Based on the original UNet network structure, the residual module BasicBlock is used to transform the encoder part and add the attention mechanism SEAE in the middle Concat layer. Specifically, the down-sampling part of UNet is modified by using the first five parts of ResNet34. The down-sampling operation uses 3×3 convolution instead of maximum pooling to reduce the loss caused by the pooling operation. Meanwhile, UNet retains jump connection part, adds the attention module AM in the middle connection layer stage and enhances the feature of the last three down-sampling output feature maps in the space and channel dimensions in the decoding stage. Then the corresponding part of the up-sampling is merged to complete the information fusion. The ResAt-UNet network structure diagram is shown in Fig. 1. The specific mechanism of the residual module BasicBlock and the attention mechanism SEAE are introduced in the following chapters.

Fig. 1 shows the detailed structure of the network model in this article. Yellow, blue, pink, red, green, and gray represent 1×1 convolution, 3×3 convolution, 7×7 convolution, maximum pooling, up-sampling, replication, and connection separately. Black and purple represent the residual modules of internal Conv1: Stride 1 and 2, which are introduced in Chapter 2.2, and SEAE represents the attention module, which is introduced in Chapter 2.3. In the encoding stage, we conduct feature extraction through four groups of 1×1 convolution, maximum pooling, and residual convolution operations. The convolution kernel uses the size of 3×3 and the maximum pooling uses the size of 3×3, which can compress the size of the feature map to 1/16.

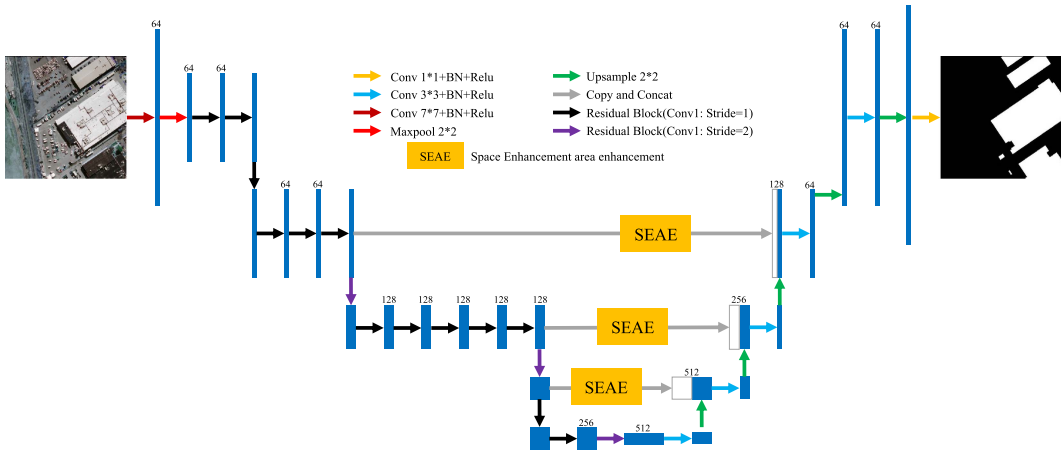


Fig. 1. ResAt-UNet network structure.

TABLE I
ARCHITECTURES FOR RESAT-UNET

Layer name	Output size	Layer
Conv1	128*128	7*7 Conv, Stride=2
Maxpool	64*64	2*2 Maxpool.Stride=2
Conv2_x	64*64	$\begin{bmatrix}3*3\ \text{Conv} & ,\text{Stride}=1\\ 3*3\ \text{Conv} & ,\text{Stride}=1\end{bmatrix}*3$
Conv3_x	32*32	$\begin{bmatrix}3*3\ \text{Conv} & ,\text{Stride}=1\\ 3*3\ \text{Conv} & ,\text{Stride}=1\end{bmatrix}*3$ $\begin{bmatrix}3*3\ \text{Conv} & ,\text{Stride}=2\\ 3*3\ \text{Conv} & ,\text{Stride}=1\end{bmatrix}*1$
Conv4_x	16.*16.	$\begin{bmatrix}3*3\ \text{Conv} & ,\text{Stride}=1\\ 3*3\ \text{Conv} & ,\text{Stride}=1\end{bmatrix}*5$ $\begin{bmatrix}3*3\ \text{Conv} & ,\text{Stride}=2\\ 3*3\ \text{Conv} & ,\text{Stride}=1\end{bmatrix}*1$
Conv5_x	8*8	$\begin{bmatrix}3*3\ \text{Conv} & ,\text{Stride}=1\\ 3*3\ \text{Conv} & ,\text{Stride}=1\end{bmatrix}*2$ $\begin{bmatrix}3*3\ \text{Conv} & ,\text{Stride}=2\\ 3*3\ \text{Conv} & ,\text{Stride}=1\end{bmatrix}*1$
Up Layer1	16.*16.	2*2 Upsample 3*3 Conv, Stride=1
Up Layer2	32*32	2*2 Upsample 3*3 Conv, Stride=1
Up Layer3	64*64	2*2 Upsample 3*3 Conv, Stride=1
Up Layer4	128*128	2*2 Upsample 3*3 Conv, Stride=1
Up Layer5	256*256	2*2 Upsample 3*3 Conv, Stride=1
Segmentation Layer	256*256	1*1 Conv, Stride=1

In the decoding stage, four sets of 2 × 2 up-sampling and 3 × 3 convolution operations are used to recover the size of the feature map. In the first three up-samplings, the up-sampling feature map completes information fusion in the middle layer and the down-sampling corresponding feature map is enhanced by the attention module feature SEAE. Finally, the prediction image with the same size as the input image is obtained by using the up-sampling and 1 × 1 convolution and the image segmentation is completed. The specific structure parameters of the network are shown in Table I, where Conv represents convolution, Maxpool represents maximum sampling, Upsample represents up-sampling, Stride represents step size and the structure name of each layer and the size of the output feature map of the network are annotated.

B. Residual Module Basic Block

He et al. [24] proposed the Deep Residual Network in 2015, in which the residual means the difference between the observed value and the estimated value. Assuming that the network input is x and the expected mapping is $H(x)$, the network mapping can be converted to the residual of the network and the corresponding relationship can be represented as

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

where x is the characteristic mapping of the upper layer network. $F(x)$ is the residual of this layer network and $H(x)$ is the observation value of this layer network. The relationship among the three is

$$H(x) = F(x) + x. \tag{2}$$

Although $H(x)$ and $F(x) + x$ have the same effect, $F(x)$ is much simpler to optimize than $H(x)$. Assuming that in layer L, the relationships between different layers can be represented as

$$X_{L+1} = X_L + F(X_L) \tag{3}$$

$$X_L = X_1 + \sum_{i=1}^{L-1} F(X_i). \tag{4}$$

Assuming that the network reaches a certain depth, the network model has reached the best and the network loss error has reached the lowest. If the network depth continues to increase, it may lead to network degradation. In order to make the network in the optimal state, we can use the residual network, so that the residual $F(x)$ value is 0.

In the k-layer network, the gradient of loss function loss with respect to x_k can be expressed as

$$\frac{\partial loss}{\partial x_k} = \frac{\partial loss}{\partial x_L} + \frac{\partial loss}{\partial x_L} * \frac{\partial}{\partial x_k} \sum_{i=1}^{L-1} F(x_i). \tag{5}$$

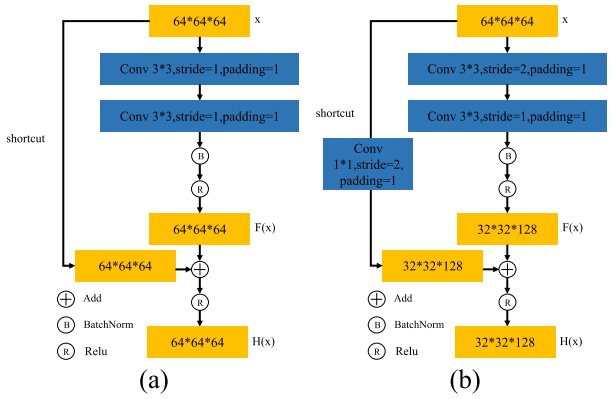


Fig. 2. Two forms of basicblock. (a) Non-dimensionality Reduction (Conv1: Stride = 1). (b) Dimensionality reduction (conv1: Stride = 2).

From (5), it can be seen that $\frac{\partial}{\partial x_k}\sum_{i=1}^{L-1} F(x_i)$ cannot always be -1 throughout the training process, and the network does not have the problem of gradient disappearance. The residual module can greatly reduce the weight of the network and the back propagation is more flexible and fast.

ResNet residual network is mainly composed of residual blocks [24], which can solve the problem of network degradation. When residual blocks are introduced, the network is built deeply and the network segmentation effect becomes better. ResNet proposed by He et al. [24] has two residual structures, namely, the two-layer residual module (BasicBlock) and the three-layer residual module (BottleNeck). For networks with fewer layers such as ResNet18 and ResNet34, BasicBlock composed of two 3×3 convolutions is often used. The schematic diagram is shown in Fig. 2. The network input is x, the expected mapping is $H(x)$ and the residual mapping of the network is $F(x)$. Conv1 and Conv2 represent two different convolution layers. The specific parameters have been annotated in detail in the graph as BN represents BatchNorm, Relu represents Relu function activation and shortcut is shortcut channel.

The black modules in Fig. 1 correspond to (a) in Fig. 2(a) and the purple modules correspond to (b). Because dimension reduction operation is not needed in (a), the convolution layer is not added to the shortcut and the output result can be directly obtained with input plus residual. The size of the input feature map is assumed to be $64 \times 64 \times 64$. In order to reduce the size and the amount of data, the first convolution step is set to 2. Since input and residual need to be added at the end of the block, we put a convolution layer with the same step of 2 in the shortcut to unify the input and output dimensions.

The ResNet34 structure to be used in this article is shown in Fig. 3, which is a 34-layer convolutional neural network. The parameters of each layer are shown in Table II. Conv represents convolution and /2 represents the down-sampling step of 2. The residual block of the dotted box corresponds to Fig. 2(b), which can reduce the dimension of the input while the residual block of the solid box corresponds to Fig. 2(a). In this article, the down-sampling part of UNet is modified. Because UNet has been down-sampled for four times, only the first six parts of ResNet34 are kept, including Conv1, Maxpool, Conv2_x, Conv3_x, Conv4_x, and Conv5_x, as shown in Fig. 2. After removing the full connection layer, it can correspond to the original network structure. At this time, the size of the input image satisfies the multiple of 32. The retained modules are shown in Fig. 4.

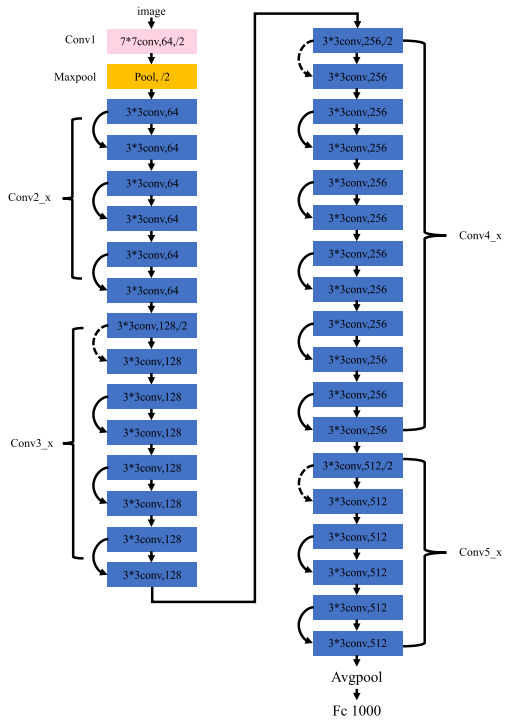


Fig. 3. ResNet34 network structure.

TABLE II
RESNET34 STRUCTURE PARAMETERS

Layer name	Outputsize	Layer
Conv1	128×128	7×7 Conv, Stride=2
Maxpool	64×64	2×2 Maxpool.Stride=2
Conv2_x	64×64	$\begin{bmatrix} 3\times3 \text{ Conv} & \text{, Stride} = 1 \\ 3\times3 \text{ Conv} & \text{, Stride} = 1 \end{bmatrix} \times 3$
Conv3_x	32×32	$\begin{bmatrix} 3\times3 \text{ Conv} & \text{, Stride} = 1 \\ 3\times3 \text{ Conv} & \text{, Stride} = 1 \end{bmatrix} \times 3$ $\begin{bmatrix} 3\times3 \text{ Conv} & \text{, Stride} = 2 \\ 3\times3 \text{ Conv} & \text{, Stride} = 1 \end{bmatrix} \times 1$
Conv4_x	16×16	$\begin{bmatrix} 3\times3 \text{ Conv} & \text{, Stride} = 1 \\ 3\times3 \text{ Conv} & \text{, Stride} = 1 \end{bmatrix} \times 5$ $\begin{bmatrix} 3\times3 \text{ Conv} & \text{, Stride} = 2 \\ 3\times3 \text{ Conv} & \text{, Stride} = 1 \end{bmatrix} \times 1$
Conv5_x	8×8	$\begin{bmatrix} 3\times3 \text{ Conv} & \text{, Stride} = 1 \\ 3\times3 \text{ Conv} & \text{, Stride} = 1 \end{bmatrix} \times 2$ $\begin{bmatrix} 3\times3 \text{ Conv} & \text{, Stride} = 2 \\ 3\times3 \text{ Conv} & \text{, Stride} = 1 \end{bmatrix} \times 1$
FullyConnected	1×1	Average pool, softmax

C. Space Enhancement Area Enhancement

Attention mechanisms are commonly used in the fields of natural image processing, knowledge graphs, and language processing. In 2018, Woo et al. [25] proposed a simple and efficient Convolutional Block Attention Module (CBAM), which can be seamlessly connected to the CNN architecture to complete the enhancement of feature map channels and spatial dimensions.

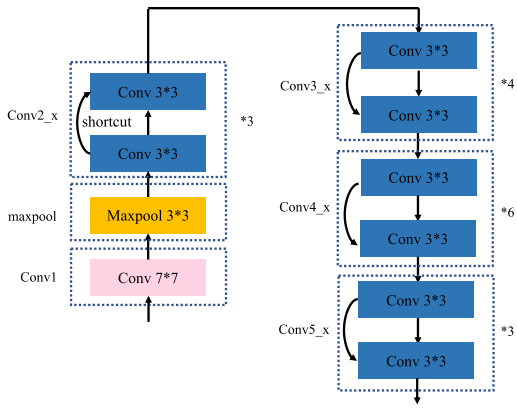


Fig. 4. Preserved ResNet34 structure.

In the same year, Park et al. [26] proposed the Bottleneck Attention Module suitable for deep neural networks, which completed the reinforcement of the feature map in the down-sampling stage. In 2020, Nanjun et al. [27] noticed the correlation between the intermediate features of the neural network, introduced the attention mechanism into the field of remote sensing image processing and inserted the newly developed hybrid first and second order attention (HFSA) into the UNet model. On this basis, the attention mechanism has been widely used in computer vision.

SEAE used in this article is composed of channel attention module (CAM) and spatial attention module (SAM). The specific mechanism will be introduced below. After introducing the attention module, the network can learn the connection mode between channel and space to improve the efficiency of information processing. The implementation method is shown in Fig. 5. The attention module used in this article is based on CBAM. The attention module used in this article is modified based on CBAM. First, the MLP depth in the attention module of the CBAM channel is expanded from 2 to 6 on CAM, which can learn the attention weights more profoundly and calculate more accurate attention weight values. Second, in SAM, the feature maps after spatial enhancement and the feature maps of the corresponding sizes of downsampling are added to alleviate the problem of gradient degradation caused by increasing the depth in the neural network, followed by the learning of spatial features in both mean pooling and maximum pooling channels, using concat for feature fusion of different channels, enriching the spatial information of the module, and by supplementing the results of 1×1 point convolution, the squeezing operation. It is possible to obtain richer information and form a more effective attention map, so as to better grasp the spatial information. Building image targets are fine, and the improved CBAM can enhance spatial information and facilitate the extraction of tiny targets.

The feature map $\mathrm{W_x}$ sampled under Encode phase is the same size as that sampled map $\mathrm{W_g}$ under Decode phase. First, CAM is used to enhance the channel feature of $\mathrm{W_x}$ and the enhanced feature map is $\mathrm{W_c}$. Subsequently, $\mathrm{W_c}$ and $\mathrm{W_g}$ in SAM are added and the enhancement of $\mathrm{W_c}$ space area is completed to obtain $\mathrm{W_s}$. Skip connection operation is then performed in Concat layer.

Fig. 6 shows the CAM module. First, maximum pooling and average pooling are used to compress the feature map $\mathrm{W_x}$ from Encode stage to channel feature map with 1×1 size, namely $P_{\mathrm{avg}}(\mathrm{x})$, $P_{\max}(\mathrm{x})$. New nonlinear elements are then introduced through the activation of continuous convolution and linear rectification function (Relu) in Multilayer Perceptron (MLP) [28]. The MLP structure diagram is shown in Fig. 7 and the specific operation of MLP is shown in (6) and (7)

$$\begin{cases} f_1(x) = \sigma_1(W_{1*1}(x)) \\ f_2(x) = \sigma_1(W_{1*1}(f_1(x) \\ f_3(x) = \sigma_1(W_{1\times1}(f_2(x)) \\ f_4(x) = \sigma_1(W_{1\times1}(f_3(x))) \\ f_5(x) = \sigma_1(W_{1\times1}(f_4(x))) \\ f_6(x) = \sigma_1(W_{1\times1}(f_5(x))) \end{cases} \tag{6}$$

$$F(x) = f_6(x) \tag{7}$$

where x is the input while $f_1(x)$, $f_2(x)$, $f_3(x)$, $f_4(x)$, $f_5(x)$, and $f_6(x)$ are the results of convolution and Relu activation of the first, second, third, fourth, fifth, and sixth times respectively. σ_1 is the activation function Relu and $W_{1\times1}$ is the convolution of 1×1. In (7), $F(x)$ is the final output of the feature map through MLP.

Then the results of the two channels after MLP processing are gotten and the Sigmoid function is used to make the weight value of each channel between 0 and 1. Finally, the weight matrix is multiplied by the original feature map to complete the feature enhancement on the channel. CAM can capture the relationship between spatial features and improve the network segmentation performance [29]. The whole process is shown in (8) and (9)

$$C_{\mathrm{weight}} = \sigma_2\left(F(P_{\mathrm{avg}}(\mathrm{W_x})) + F(P_{\max}(\mathrm{W_x}))\right) \tag{8}$$

$$\mathrm{W_c} = \mathrm{C_{weight}} * \mathrm{W_x}. \tag{9}$$

In (8) and (9), σ_2 is the activation function Sigmoid, F is the activation of MLP sensor, P_{avg} is the average pooling, $P_{\max}$ is the maximum pooling, $\mathrm{W_x}$ is the input feature map, $\mathrm{C_{weight}}$ is the channel weight matrix, and $\mathrm{W_c}$ is the feature map after channel feature enhancement.

Fig. 8 shows the SAM used in this article, which is an improvement based on Attention-Gate [30] Adding SAM in the middle layer helps the network to focus on a specific part of the input.

First, the $\mathrm{W_c}$ enhanced by CAM is gotten and the feature map $\mathrm{W_g}$ obtained by upsampling in the Decode stage, use 1×1 convolution to compress the two feature maps to the same size channel, and then add the two feature maps element by element. Next, through the activation function, Relu activation and 1×1 convolution compression, the feature matrix with channel 1 is obtained. Here, the max pooling and average pooling processing matrices are used to obtain two different spatial feature maps, then CAM stitchs the maps according to dimension 1, uses 1×1 convolution to compress the map channel to 1 and finally uses the sigmoid function to activate the feature and weight map, so that the value of the weight map can be standardized between 0-1 and the weight map is multiplied with the feature map $\mathrm{W_x}$

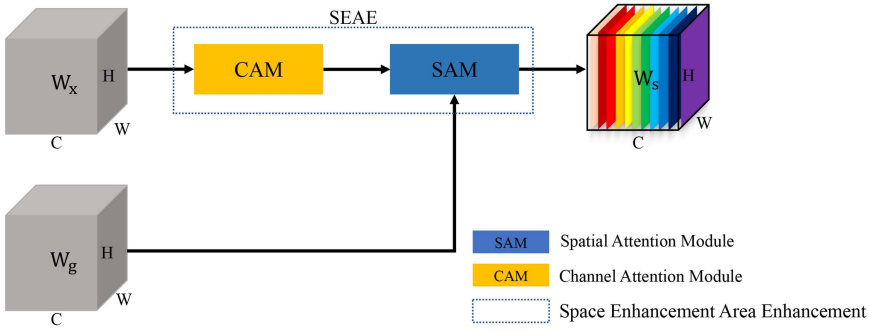


Fig. 5. SEAE schematic.

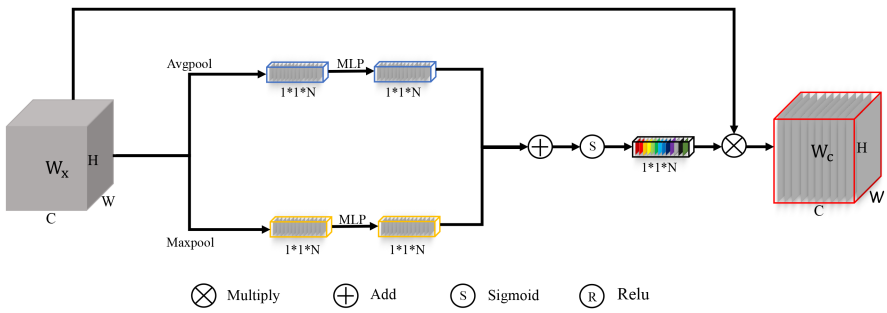


Fig. 6. Schematic of the CAM.

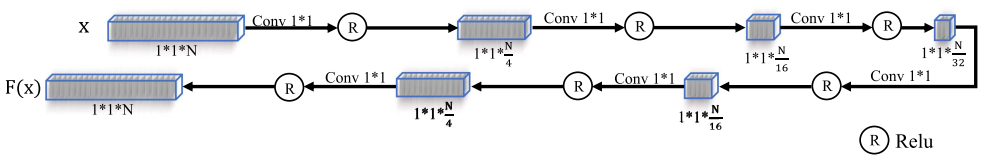


Fig. 7. MLP schematic.

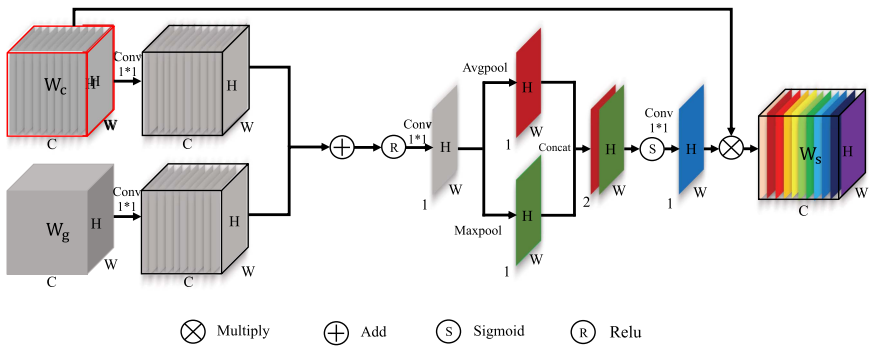


Fig. 8. Schematic of the SAM.

to perform feature re-calibration to complete the enhancement of the spatial area of the building image.

The whole process is illustrated in the following formula, where P_{avg} represents average pooling, $P_{\max}$ represents maximum pooling, σ_1 is the activation function Relu, σ_2 is the activation function Sigmoid, and $W_{1\times1}$ is the 1×1 convolution and $\mathrm{C_{at}}$ represents the splicing operation. S_{a} and S_{b} are the spatial feature maps of the upper and lower pooling channels, $\mathrm{W_x}$ and $\mathrm{W_g}$ are the input images and S_{weight} is the attention coefficient, which is obtained by addition. Although these steps

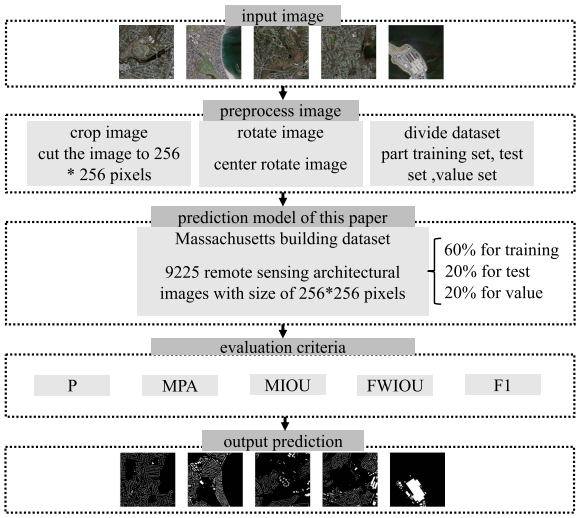


Fig. 9. Experimental flow.

will generate more computation than multiplication cost, a more ideal segmentation effect can be obtained. $\mathrm{W_S}$ is the feature map after spatial region feature enhancement

$$S_{\mathrm{a}} = P_{\mathrm{avg}}\left(W_{1*1}\left(\sigma_1\left(W_{1*1}\left(\mathrm{W_x}\right) + W_{1*1}\left(\mathrm{W_g}\right)\right)\right)\right) \tag{10}$$

$$S_{\mathrm{b}} = P_{\mathrm{max}}\left(W_{1*1}\left(\sigma_1\left(W_{1*1}\left(\mathrm{W_x}\right) + W_{1*1}\left(\mathrm{W_g}\right)\right)\right)\right) \tag{11}$$

$$S_{\mathrm{weight}} = \sigma_2\ \left(W_{1*1}\left(C_{\mathrm{at}}\left(S_{\mathrm{a}}, S_{\mathrm{b}}\right)\right)\right) \tag{12}$$

$$\mathrm{W_S} = S^*_{\mathrm{weight}}\mathrm{W_x}. \tag{13}$$

In the network jump connection stage, different numbers of SEAE modules are added to enhance the building pixels of the feature map and suppress the feature information expression of the background pixels, trying to improve the anti-noise ability of the network model.

III. Experiment

In this chapter, the above improved convolutional neural network model is used for experiments to complete the segmentation of building images. The detailed process of the algorithm is shown in Fig. 9, which mainly includes processing of image, division of test set and training set, training of the neural network and acquisition of test results.

Step 1: The original data of the Massachusetts dataset is large and the difference between the background and the building is not obvious, so it needs to be preprocessed, mainly including image cropping to reduce the image size and image rotation to expand the dataset.

Step 2: Before model training, the training set, test set and validation set are roughly divided into a ratio of 6:2:2 and each image must be preprocessed.

Step 3: In the model training stage, the process initializes the parameters of the model, inputs the data set, trains and optimizes the network and saves the trained model.

TABLE III
Laboratory Environment

Environment	Implement parameters
CPU	Intel Core E5
GPU	NVDIA TITAN X
Python	3.8
Torch	1.10.1
Opencv-python	4.5.5.62

TABLE IV
Training Parameters

Hyperparameters	Implement
Batch_size	15
Epochs	150
Optimizer	RMSProp
Loss function	CombinedLoss
Learning rate	0.001

Step 4: After each test image is input to the prediction module, the output result should be an image with the building and the background separated. The final prediction result is obtained and the effect is evaluated.

In the following, we first introduce the experimental environment and then preprocess the dataset to illustrate the training methods and evaluation criteria. Finally, the experimental results are displayed and analyzed.

A. Experimental Environment

Hardware environment: This experiment uses a Macbook Air produced in 2022 with Ubuntu operating system. The industry generally believes that the most popular deep learning framework is pytorch [31]. This article selects this framework to build a neural network on the integrated development environment Pycharm2021, and the specific configuration is shown in Table III.

The training parameters are set as shown in Table IV. When the times of training reach 8400, the model stops.

B. Dataset

In this article, the Massachusetts data set is used to generate multiple similar images by using data enhancement methods such as flipping, translation, scaling, and cropping to increase the size of the data set. As the data set increases, the model cannot overfit all samples, so the model has to be generalized.

The dataset used in this article is Massachusetts remote sensing city image [32]. The original dataset contains 137 large images for training, 4 for validation, and 10 for testing. The original size is 1500 × 1500. After rough estimation, the image resolution is 1.5 meters and the image quality is not very high. Due to the large image size in the dataset, it is not suitable for training test and verification directly. After preprocessing, the image size is reduced to 256 × 256. The training set, test

Fig. 10. Image before preprocessing. (a) Original images. (b) Ground truth.

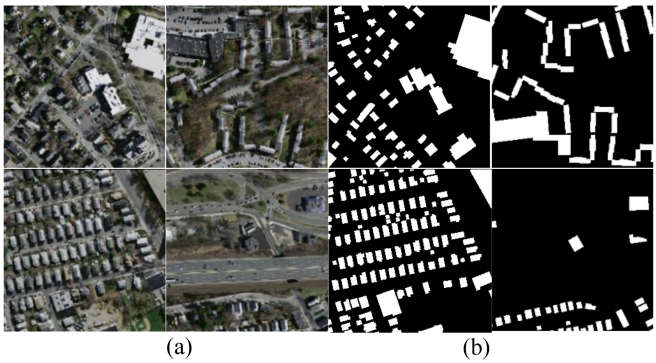


Fig. 11. Image after preprocessing. (a) Original images. (b) Ground truth.

set, and verification set are randomly divided according to the proportion, 7200 images for training, 2400 images for testing, and 2400 images for verification.

The preprocessing method is described in detail below. The original image of Massachusetts is shown in Fig. 10 and the experimental image after clipping is shown in Fig. 11. As can be seen from Fig. 11, the image quality is general, the background is complex and the dataset has many different categories, such as docks, farms, forests, hills, lakes, roads, etc. The sizes of different targets are quite different as the resolutions of cars and pedestrians are small while the targets of lakes, rivers, and forests are large. Moreover, the boundaries of different targets are fuzzy and the pixel classification between adjacent targets is difficult.

C. Data Augmentation

1) Image Cropping: For the original remote sensing image with the pixel value of 1050 × 1050, we need to cut to construct the building image dataset. In order to facilitate processing, we cut each image into 25 256 × 256 images. Fig. 12 shows the effect of cutting a single image into 25 images in proportion.

2) Image Rotation: To expand the dataset and facilitate the next training of convolutional neural networks, we can use the image rotation method for preprocessing. Image rotation can use affine transformation to preserve the original "straightness" and "parallelism" of the image. The effect of expanding the dataset by image rotation is shown in Fig. 13 and the affine transformation matrix is shown in (14). After data augment, the model performance is improved and overfitting is mitigated

$$\begin{pmatrix} \mathrm{x}' \\ \mathrm{y}' \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \mathrm{t_x} \\ 0 & 1 & \mathrm{t_y} \\ 0 & 0 & 1 \end{pmatrix} * \begin{pmatrix} \mathrm{x} \\ \mathrm{y} \\ 1 \end{pmatrix}. \tag{14}$$

In (14), x and y are the input pixels, x' and y' are the output pixels, $\mathrm{t_x}$ is the distance moving on the horizontal axis and $\mathrm{t_y}$ is the distance moving on the vertical axis.

D. Improvement of Loss Function

In this task, since our goal is to divide the input image into two categories, we combine the binary cross entropy loss function $BceLoss$ and set similarity loss function $DiceLoss$ to obtain the improved loss function. For data containing N samples, BceLoss and DiceLoss are shown in (15) and (16) while our improved loss function CombinedLoss is shown in (17)

$$\mathrm{BceLoss} = \frac{1}{\mathrm{N}} \sum_{\mathrm{n}=1}^{\mathrm{N}} - (\mathrm{x_n} * \log \mathrm{y_n} + (1 - \mathrm{x_n}) * log(1 - \mathrm{y_n})) \tag{15}$$

$$\mathrm{DiceLoss} = 1 - \frac{2 \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{x_n} * \mathrm{y_n}}{\sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{x_n} + \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{y_n}} \tag{16}$$

$$\mathrm{CombinedLoss} = \lambda * BceLoss + \mu * DiceLoss (\lambda + \mu = 1). \tag{17}$$

In the equation, x_n indicates the predicted target category and y_n indicates the actual target category.

BceLoss applies to pixel-level prediction tasks, especially targeting at learning small samples, but is vulnerable to uneven sample distribution [33]. DiceLoss considers losses more globally and tends to process large samples, which is more suitable for binary segmentation [34]. In this article, the building objects that need to be segmented in the building image segmentation task are very small and the number of objects and background samples in the image is not balanced. DiceLoss can complete learning and training without the influence of the background size. The combination of the two can introduce additional weights to BceLoss, alleviates the problem of imbalance in the number of buildings and background samples and is conducive to the training optimization and parameter updating of the network.

E. Evaluation Criteria

To comprehensively evaluate the performance of the model, we use four widely used evaluation indexes for remote sensing building image segmentation, namely Precision (P), Mean Pixel Accuracy (MPA), Mean Intersection over Union (MIOU) and

Fig. 12. Image cropping. (a) Original images. (b) Cropped images.

Fig. 13. Image rotation. (a) Original images. (b) Center rotated 90 degrees. (c) Center rotated 180 degrees.

frequency weighted over Union (FWIOU), recall (Recall), F_1-score (F_1). The representations of which are as follows:

$$P = \frac{\mathrm{BB}}{\mathrm{BB}+\mathrm{NB}} \tag{18}$$

$$\mathrm{MPA} = \left(\frac{\mathrm{BB}}{\mathrm{BB}+\mathrm{NB}} + \frac{\mathrm{NN}}{\mathrm{NN}+\mathrm{BN}}\right)/2 \tag{19}$$

$$\mathrm{MIOU} = \left(\frac{\mathrm{BB}}{\mathrm{BB}+\mathrm{BN}+\mathrm{NB}} + \frac{\mathrm{NN}}{\mathrm{NN}+\mathrm{BN}+\mathrm{NB}}\right)/2$$

$$\mathrm{FWIOU} = \frac{\mathrm{BB}+\mathrm{BN}}{\mathrm{BB}+\mathrm{BN}+\mathrm{NB}+\mathrm{NN}} * \left(\frac{\mathrm{BB}}{\mathrm{BB}+\mathrm{BN}+\mathrm{NB}}\right) \tag{20}$$

$$+ \frac{\mathrm{NN}+\mathrm{NB}}{\mathrm{BB}+\mathrm{BN}+\mathrm{NB}+\mathrm{NN}} * \left(\frac{\mathrm{NN}}{\mathrm{NN}+\mathrm{BN}+\mathrm{NB}}\right) \tag{21}$$

$$\mathrm{Precision} = \frac{\mathrm{BB}+\mathrm{NN}}{\mathrm{BB}+\mathrm{BN}+\mathrm{NB}+\mathrm{NN}} \tag{22}$$

$$\mathrm{Recall} = \frac{\mathrm{BB}}{\mathrm{BB}+\mathrm{BN}} \tag{23}$$

$$F_1 = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{24}$$

In the above formulas, BB is the number of pixels correctly predicted as the building target; BN is the number of pixels whose background target is falsely detected; NB is the number of pixels wrongly detected as building targets; NN is the number of pixels correctly detected as the background target.

F. Experimental Results

1) Parameter Experiment: In order to set the parameters of the loss function, the neural network ResAt-UNet(2SEAE) with two SEAE modules is used for image segmentation in Massachusetts. We set six intervals (λ, μ) = (0, 1), (0.2, 0.8), (0.4, 0.6), (0.6, 0.4), (0.8, 0.2), (1, 0). For experiments, where λ is the weight of BceLoss and μ is the weight of DiceLos. The segmentation results are shown in Table V.

When λ is set to 0.2, μ is set to 0.8, MPA, $MIOU$, $FWIOU$ are segmented, which have obvious advantages over other networks. Therefore, λ is set to 0.2 and μ is set to 0.8 in the $CombinedLoss$ loss function.

2) Ablation Experiment: The original meaning of ablation is surgical resection of body tissue [35]. Long et al. [36] defined ablation experiment as: on relatively complex neural networks, deleting some network structures to test network performance to understand the process of network internal structure. In the field of deep learning, the use of ablation experiments to remove some parts of the network contributes to a better understanding of the network behavior, which is very important for deep learning research and can help to study the causal relationship of the system [36]. In this chapter, in order to select the network with the optimal performance of the attention module, this article conducts four ablation experiments and compares the segmentation results of the network without adding, adding one, two, and three attention modules SEAE, namely ResAt-UNet, ResAt-UNet(1SEAE), ResAt-UNet(2SEAE), ResAt-UNet(3SEAE). In order to evaluate differently deep ResNet performance, this article compares several ResNet networks, namely UNet, ResAt-UNet(2SEAE), UNet-ResNet50(2SEAE), UNet-ResNet101(2SEAE). To evaluate the loss in each network training, the results of P, MPA, $MIOU$, $FWIOU$ are shown in Tables VI and VII. As the network is deeper, the model

TABLE V
PARAMETER EXPERIMENTAL RESULTS

ResAt-UNet(2SEAE)	*P*	*MPA*	*MIOU*	*FWIOU*	*F1*
$\lambda = 0,\ \mu = 1$	0.8334	0.5459	0.4701	0.7045	0.8144
$\lambda = 0.2, \mu = 0.8$	**0.9366**	**0.8701**	**0.7881**	**0.8852**	**0.9059**
$\lambda = 0.4, \mu = 0.6$	0.8056	0.5011	0.4144	0.6611	0.7998
$\lambda = 0.6, \mu = 0.4$	0.8159	0.5392	0.4501	0.6821	0.7801
$\lambda = 0.8, \mu = 0.2$	0.8402	0.5701	0.4679	0.7137	0.8199
$\lambda = 1, \mu = 0$	0.8009	0.6411	0.5701	0.7922	0.8012

TABLE VI
QUANTITATIVE COMPARISON OF REMOTE SENSING BUILDING IMAGE SEGMENTATION RESULTS FROM ABLATION EXPERIMENTS

Method	*P*	*MPA*	*MIOU*	*FWIOU*	*F1*
ResAt-UNet	0.9349	0.8501	0.7732	0.8798	0.8072
ResAt-UNet(1SEAE)	0.7857	00.7971	0.716.2	0.8361	0.8192
ResAt-UNet(2SEAE)	**0.9366**	**0.8701**	**0.7881**	**0.8852**	**0.9059**
ResAt-UNet(3SEAE)	0.7372	0.8014	0.7188	0.8463	0.8320

TABLE VII
COMPARISON OF DIFFERENT DEPTH NETWORK SEGMENTATION RESULTS

Method	*P*	*MPA*	*MIOU*	*FWIOU*	*F1*
UNet	0.8811	0.8102	0.7167	0.8502	0.6878
ResAt-UNet(2SEAE)	**0.9366**	**0.8701**	**0.7881**	**0.8852**	**0.9059**
UNet-ResNet50(2SEAE)	0.8944	0.7570	0.7327	0.8152	0.6800
UNet-ResNe101(2SEAE)	0.8065	0.8610	0.7114	0.8114	0.5686

performs increasingly poorly, so the ResAt-UNet(2SEAE) is selected for the next comparative experiment.

Fig. 14 is the contrast map of the visual effect of remote sensing building image segmentation in the ablation experimental network and four groups of different scene segmentation images are selected. In the image, the white part stands for the building and the black part is the background. The existence of roads and vegetation interferes with image segmentation [29].

The first image in Fig. 14(a) is remote sensing images of urban suburbs. The first image in Fig. 14(b) is the ResAt-UNet segmentation result, which is rough; the first image in Fig. 14(c) is the result of ResAt-UNet(1SEAE) segmentation, which can segment some large buildings but cannot completely separate them from the background; and the first image in Fig. 14(d) is the ResAt-UNet(2SEAE) segmentation result, which can describe the building contour as the segmentation is more accurate and there is no false detection. The first image in Fig. 14(e) is the segmentation result of ResAt-UNet(3SEAE). The extracted targets are relatively complete, but the buildings and roads cannot be well separated and there is a situation where some backgrounds are mistakenly regarded as building targets.

The second image in Fig. 14(a) is the remote sensing image of university town areas. The second image in Fig. 14(b) is the ResAt-UNet segmentation result, which does not segment small targets; the second image in Fig. 14(c) is the result of ResAt-UNet(1SEAE) segmentation, which can separate the building and the background, but there are missing and false detection; the second image in Fig. 14(d) is the segmentation result of ResAt-UNet(2SEAE), which segments most building targets completely and accurately and there are fewer missed buildings; and the second image in Fig. 14(e) is the result of ResAt-UNet(3SEAE) segmentation, which can segment most buildings, but the contour of the building target is rough and there is a problem of false detection.

The third image in Fig. 14(a) is the remote sensing image of residential areas. The third image in Fig. 14(b) is the ResAt-UNet segmentation result, which misses a small number of targets; the third image in Fig. 14(c) is the ResAt-UNet(1SEAE) segmentation result, which has more missed targets and larger false detection area. The third image in Fig. 14(d): ResAt-UNet(2SEAE) and the third image in Fig. 14(e): ResAt-UNet(3SEAE) can segment the vast majority of building targets whereas ResAt-UNet(2SEAE) for building target extraction is more detailed.

The fourth image in Fig. 14(a) is the remote sensing image of hilly areas. The fourth image in Fig. 14(b) is the ResAt-UNet segmentation result, which misses individual targets. The fourth image in Fig. 14(c) is the ResAt-UNet(1SEAE) segmentation result, which mistakenly regards land and road as building targets and missed some buildings. The fourth image in Fig. 14(e) is the result of ResAt-UNet(3SEAE) segmentation, which can extract most of the building targets, but there is a false detection and a small amount of rock is segmented into buildings.

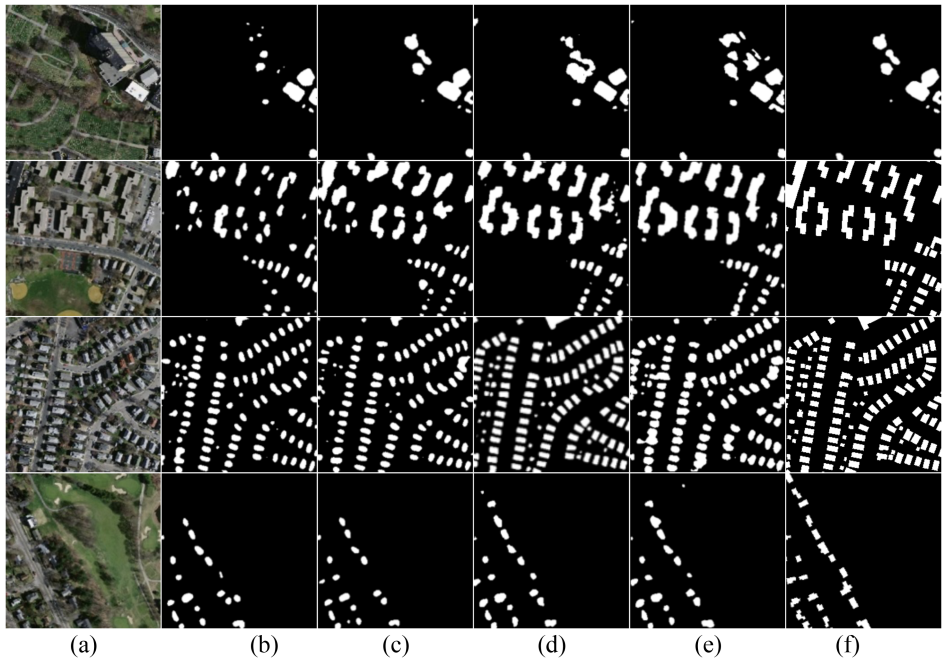


Fig. 14. Visual effect diagram of building segmentation in the ablation experiment. (a) Original images. (b) ResAt-Unet. (c) ResAt-UNet(1SEAE). (d) ResAt-UNet(2SEAE). (e) ResAt-UNet(3SEAE). (f) Grouth truth.

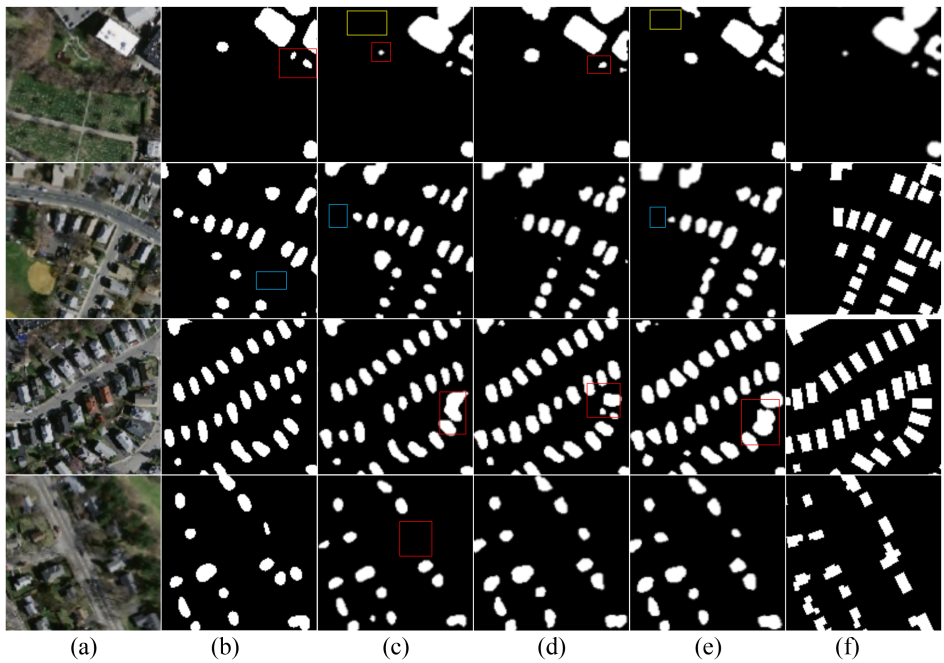


Fig. 15. The visual effect diagram of the detail part of the building segmentation in the ablation experiment. (a) Original images. (b) ResAt-Unet. (c) ResAt-UNet(1SEAE). (d) ResAt-UNet(2SEAE). (e) ResAt-UNet(3SEAE). (f) Grouth truth.

The fourth image in Fig. 14(d) is the result of ResAt-UNet(2SEAE) segmentation. The segmentation effect is obviously improved and missed detection and false detection are rare, which can better segment the building from the background.

Fig. 15 shows the comparison of visual effect details of remote sensing building image segmentation in the ablation experiment, and the parts with poor segmentation effect are selected for analysis and comparison. The yellow frame in the following diagram indicates that the area is missing and the building is mistaken as the background. The red frame indicates that the segmentation effect is poor. The blue frame indicates that the area is missing, and the background is mistaken as the building.

The first image in Fig. 15(a) is the remote sensing image of some suburban areas. The first image in Fig. 15(b) is the

ResAt-UNet segmentation result, which is rough; the first image in Fig. 15(c) is the ResAt-UNet(1SEAE) segmentation result, which misses many building targets and the segmentation effect is not accurate enough; the first image in Fig. 15(d) is ResAt-UNet(2SEAE) segmentation results, of which the segmentation is more accurate and missing and wrong detection problems rarely appear; and the first image in Fig. 15(e) is the result of ResAt-UNet(3SEAE) network segmentation, which separates most of the targets, but there are also many inaccurate segmentation regions.

The second image in Fig. 15(a) is the remote sensing image of the university town area. The second image in Fig. 15(b) is the ResAt-UNet segmentation result, which misses a target; the second image in Fig. 15(c) is the segmentation result of ResAt-UNet(1SEAE), which misdetects many building targets and does not process the edge of building targets accurately. The second image in Fig. 15(d) is the segmentation result of ResAt-UNet(2SEAE), which is more accurate in segmentation and accurate in the extraction of building feature information. Missed detection and false detection almost do not appear and all building targets are segmented. The second image in Fig. 15(e) is the result of ResAt-UNet(3SEAE) network segmentation, which separates most of the targets, but there are errors.

The third image in Fig. 15(a) is the remote sensing image of residential areas. The third image in Fig. 15(b) is the segmentation result of ResAt-UNet. The third image in Fig. 15(c) is the segmentation result of ResAt-UNet(1SEAE). It misdetects many building targets and the processing of building target edge is not accurate enough, so the segmentation effect is not ideal. The third image in Fig. 15(d) is ResAt-UNet(2SEAE) segmentation result, the segmentation being more accurate and the processing of details being in place. The third image in Fig. 15(e) is the result of ResAt-UNet(3SEAE) network segmentation, which detects most of the building targets, but the edge information of small targets is not accurate enough and the building contour is not clear enough.

The fourth image in Fig. 15(a) is the remote sensing image of hilly areas. The fourth image in Fig. 15(b) is the ResAt-UNet segmentation result. The fourth image in Fig. 15(c) is the ResAt-UNet(1SEAE) segmentation result with missed detections of many building targets and nondetections of some targets. The fourth image in Fig. 15(d) is the ResAt-UNet(2SEAE) segmentation result, which segments all the targets without missed detection and false detection. The fourth image in Fig. 15(e) is the result of ResAt-UNet(3SEAE) network segmentation, which has a false detection and regards some rocks as building segmentation.

Combining the subjective observations of the performance experiments in Figs. 14 and 15 with the objective actual data, ResAt-UNet(2SEAE) is more accurate in the extraction of small building targets, more perfect in the processing of image detail information, and better to strip the building from the background, so it is selected for the following comparative experiments.

3) Comparative Experiment: To test the effectiveness of the improved network in this article, namely, adding the residual module and attention module ResAt-UNet(2SEAE), MLP [28], SVM [37], FCN8 [20], Bilateral Segmentation Network (BisNet) [38], Dual Attention Network (DANet) [39], SegNet [22], PSPNet-101 [23], DRNet [40], Deep Feature Aggregation Network (DFA-Net) [41], Spatial residual inception convolutional neural network (SRI-Net) [17], DeepLabV3+ [42], UNet [21], Du-CNN [18], MA-FCN [45], BRRNet [44] and ResAt-UNet network with only residual modules are selected for building image segmentation in this article. They are also compared with the improved network ResAt-UNet(2SEAE) segmentation effect to test the improved network segmentation effect.

Table VIII shows a comparison between quantitative results of the improved network and other classical building segmentation networks. The results of image quantitative indicators obtained by traditional network MLP, SVM, FCN8s, SegNet, BisNet, and DANet segmentation are not ideal. The results obtained by PSPNet, DRNet, and SRI-Net have been improved, but the performances are slightly worse than that of UNet, DeepLabV3+, DFA-Net, and ResAt-UNet(2SEAE) segmentation results. ResAt-UNet(2SEAE) has obtained the best P, MPA, MIOU and FWIOU. Compared with UNet, it increases by 0.0554, 0.0513, 0.0639, and 0.0422 respectively, which proves the accuracy of this algorithm as well as the reliability and effectiveness of network improvement.

Fig. 16 is the visual effect comparison chart of eight remote sensing building image segmentation methods with good quantitative indicators, and eight groups of segmentation images of different scenes are selected. In the image, the white part represents the building and the black part is the background. The existence of roads and vegetation interferes with image segmentation.

The first image in Fig. 16(a) is the remote sensing image of dense residential areas. The first image in Fig. 16(b) is the MLP segmentation result. The segmentation effect is not good and the building contour cannot be described. The first image of Fig. 16(c) is the FCN8s segmentation result and the building contour cannot be described; the first image of Fig. 16(d) is the BisNet segmentation result; the first image of Fig. 16(e) is the PSPNet-101 segmentation result; and the first image of Fig. 16(f) is the UNet segmentation result, which is better in large building areas. The first image of Fig. 16(g) is the ResAt-UNet segmentation result. Most of the residential buildings are segmented. The first image in Fig. 16(h) is the segmentation result of ResAt-UNet(2SEAE), which can well describe the building and the segmentation is relatively accurate without false detection.

The second image in Fig. 16(a) is the remote sensing image of residential areas. The second image in Fig. 16(b) is the MLP segmentation result, and the second image in Fig. 16(c) is the FCN8s segmentation result, which cannot describe the building contour. The second image in Fig. 16(d) is the BisNet segmentation result, the second image in Fig. 16(f) is the UNet segmentation result, and the second image in Fig. 16(g) is the ResAt-UNet segmentation result, which performs well in the building area. The second image in Fig. 16(e) is the PSPNet-101 segmentation result, which missed the large building. The second image in Fig. 16(h) is the result of ResAt-UNet(2SEAE) segmentation. The vast majority of building targets are segmented completely and accurately, and there are almost no missed targets.

TABLE VIII
QUANTITATIVE COMPARISON OF THE SEGMENTATION RESULTS OF REMOTE SENSING BUILDING IMAGES IN CONTRASTIVE EXPERIMENTS

Method	*P*	*MPA*	*MIOU*	*FWIOU*	*F1*
MLP	0.7299	0.5667	0.3989	0.5799	0.5052
SVM	0.7406	0.5904	0.4877	0.5409	0.5245
FCN8s	0.7427	0.5537	0.3709	0.5077	0.4990
SegNet	0.7336	0.5537	0.3962	0.6639	0.6939
BisNet	0.7413	0.5709	0.4596	0.6649	0.5132
DANet	0.7734	0.6906	0.6210	0.7156	0.6199
PSPNet-101	0.8659	0.7799	0.7089	0.8166	0.6579
DRNet	0.8932	0.7831	0.7097	0.7999	0.6351
SRI-Net	0.8902	0.7899	0.7011	0.8012	0.6811
UNet	0.8811	0.8102	0.7167	0.8502	0.6878
DeepLabV3+	0.8999	0.8071	0.7309	0.8419	0.7089
DFA-Net	0.9006	0.8189	0.7309	0.8521	0.7204
Du-CNN	0.8337	0.8643	0.7350	0.8474	0.7651
MA-FCN	0. 9042	0.8692	0.7389	0.7625	0.8498
BRRNet	0.9002	0.8249	0.7289	0.8422	0.8432
ResAt-UNet	0.9349	0.8501	0.7732	0.8798	0.8072
ResAt-UNet(2SEAE)	**0.9366**	**0.8701**	**0.7881**	**0.8852**	**0.9059**

TABLE IX
GENERALIZATION EXPERIMENT RESULTS

Method	*MP*	*MPA*	*MIOU*	*MRecall*
FCN8s	0.6846	0.6811	0.5617	0.6238
UNet	0.6970	0.7024	0.6043	0.6516
SegNet	0.6732	0.7071	0.5373	0.6890
PSPNet	0.6632	0.7128	0.5581	0.6889
DeepLabv3+	0.6998	0.7296	0.5936	0.7011
DA-UNet	0.7404	0.7012	0.5926	0.7279
DFA-Net	0.7327	**0.7580**	0.6029	0.7263
ResAt-UNet(2SEAE)	**0.7421**	0.7011	**0.6156**	**0.7289**

The third image in Fig. 16(a) is the remote sensing image of dense residential areas. The third image in Fig. 16(b) is the MLP segmentation result, and the third image in Fig. 16(c) is the FCN8s segmentation result, which separates some large buildings. The third image in Fig. 16(d) is the BisNet segmentation result, the third image in Fig. 16(e) is the PSPNet-101 segmentation result, the third image in Fig. 16(f) is the UNet segmentation result, and the third image in Fig. 16(g) is the ResAt-UNet segmentation result. These four methods perform well in the building area, but the individual small buildings are missed. The third image of Fig. 16(h) is the result of ResAt-UNet(2SEAE) segmentation, which is more detailed for extraction of the building target.

The fourth image in Fig. 16(a) is the remote sensing image of industrial areas. The fourth image of Fig. 16(b) is the MLP segmentation result, which is not ideal, and the building segmentation is incomplete. The fourth image of Fig. 16(c) is the FCN8s segmentation result, the fourth image of Fig. 16(d) is the BisNet segmentation result, the fourth image of Fig. 16(e) is the PSPNet-101 segmentation result, and the fourth image of Fig. 16(f) is the UNet segmentation result. These four methods are incomplete for the segmentation of large warehouse targets, and the fourth image of Fig. 16(g) is the ResAt-UNet segmentation result. The fourth image in Fig. 16(h) is the ResAt-UNet(2SEAE) segmentation result, which is more accurate.

The fifth image in Fig. 16(a) is the remote sensing image of dense residential areas. The fifth image of Fig. 16(b) is the MLP segmentation result, and the segmentation effect is not good. The fifth image of Fig. 16(c) is the FCN8s segmentation result, which cannot describe the building contour. The fifth image of Fig. 16(d) is the BisNet segmentation result; the fifth image of Fig. 16(e) is the PSPNet-101 segmentation result; the fifth image of Fig. 16(f) is the UNet segmentation result; and the fifth image of Fig. 16(g) is the ResAt-UNet segmentation result, which segments most of the residential buildings. The fifth image in Fig. 16(h) is the result of ResAt-UNet(2SEAE)

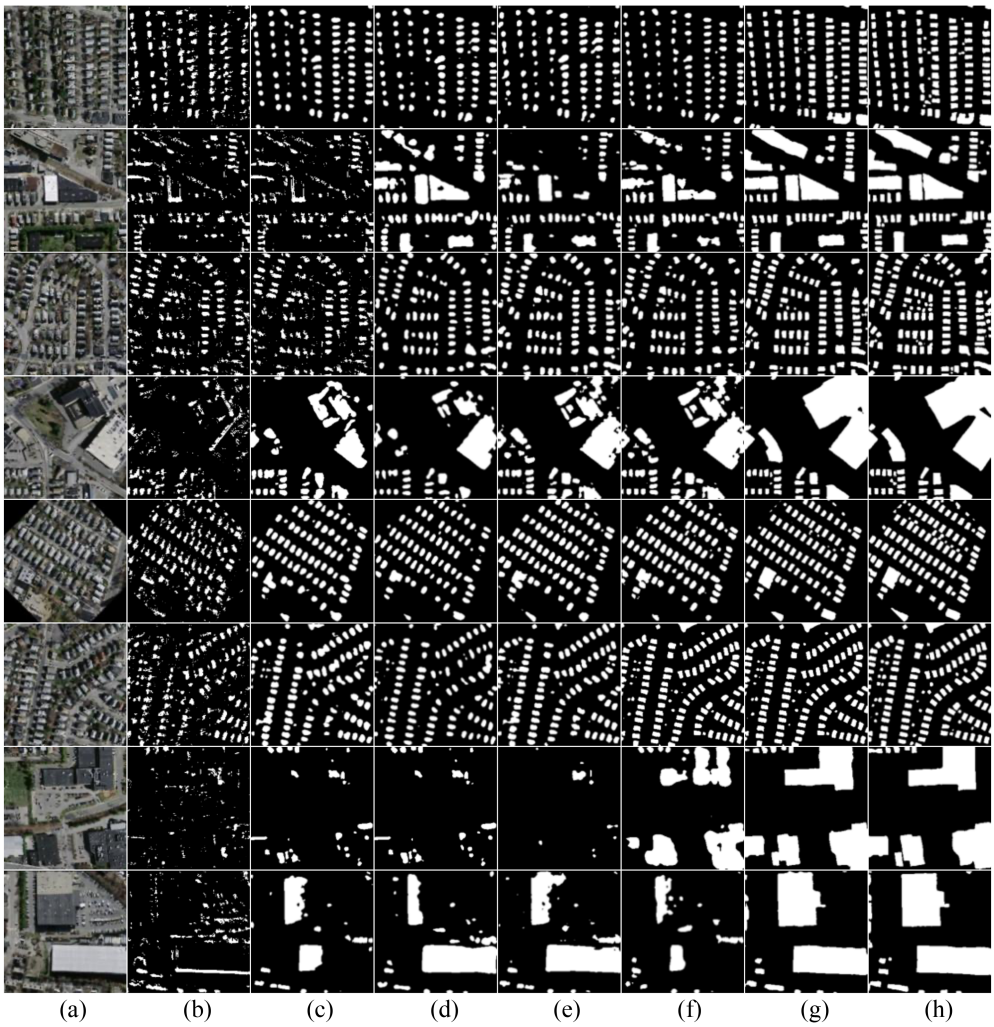


Fig. 16. Visual effect diagram of building segmentation in contrast experiment. (a) Original images. (b) MLP. (c) FCN8s. (d) BisNet. (e) PSPNet-101. (f) Unet. (g) ResAt-Unet. (h) ResAt-UNet(2SEAE).

segmentation, which completely separates small objects and can better separate buildings and backgrounds.

The sixth image in Fig. 16(a) is the remote sensing image of commercial areas. The sixth image of Fig. 16(b) is the MLP segmentation result, which misses many targets. The sixth image of Fig. 16(c) is the FCN8s segmentation result, and the sixth image of Fig. 16(d) is the BisNet segmentation result. These two methods miss some small targets. The sixth image of Fig. 16(e) is the PSPNet-101 segmentation result; the sixth image of Fig. 16(f) is the UNet segmentation result; the sixth image of Fig. 16(g) is the ResAt-UNet segmentation result; and the sixth image in Fig. 16(h) is the ResAt-UNet(2SEAE) segmentation result. These four methods can complete the accurate segmentation of small targets.

The seventh image in Fig. 16(a) is the remote sensing image of industrial areas. The seventh image in Fig. 16(b) is the MLP segmentation result, and the seventh image in Fig. 16(c) is the FCN8s segmentation result, which does not extract the warehouse target. The seventh image of Fig. 16(f) is the result of UNet segmentation; the seventh image of Fig. 16(g) is the result of ResAt-UNet segmentation; and the seventh image of Fig. 16(h) is the result of ResAt-UNet(2SEAE) segmentation. These three methods can accurately segment large and small targets, and ResAt-UNet(2SEAE) has the best segmentation effect.

The eighth image in Fig. 16(a) is the remote sensing image of industrial areas. The eighth image in Fig. 16(b) is the MLP segmentation result, and its segmentation effect is poor. The eighth image in Fig. 16(c) is the FCN8s segmentation result; the eighth image in Fig. 16(d) is the BisNet segmentation result; the eighth image in Fig. 16(e) is the PSPNet-101 segmentation result; and the eighth image in Fig. 16(f) is the UNet segmentation result. The segmentation of large targets by the above methods is incomplete. The eighth image in Fig. 16(g) is the ResAt-UNet segmentation result. The eighth image in Fig. 16(h) is the result of ResAt-UNet(2SEAE) segmentation. The two large warehouse targets are completely segmented without missed detection and false detection.

Fig. 17 compares the detail parts of visual effects of remote sensing building image segmentation in the contrast experiment,

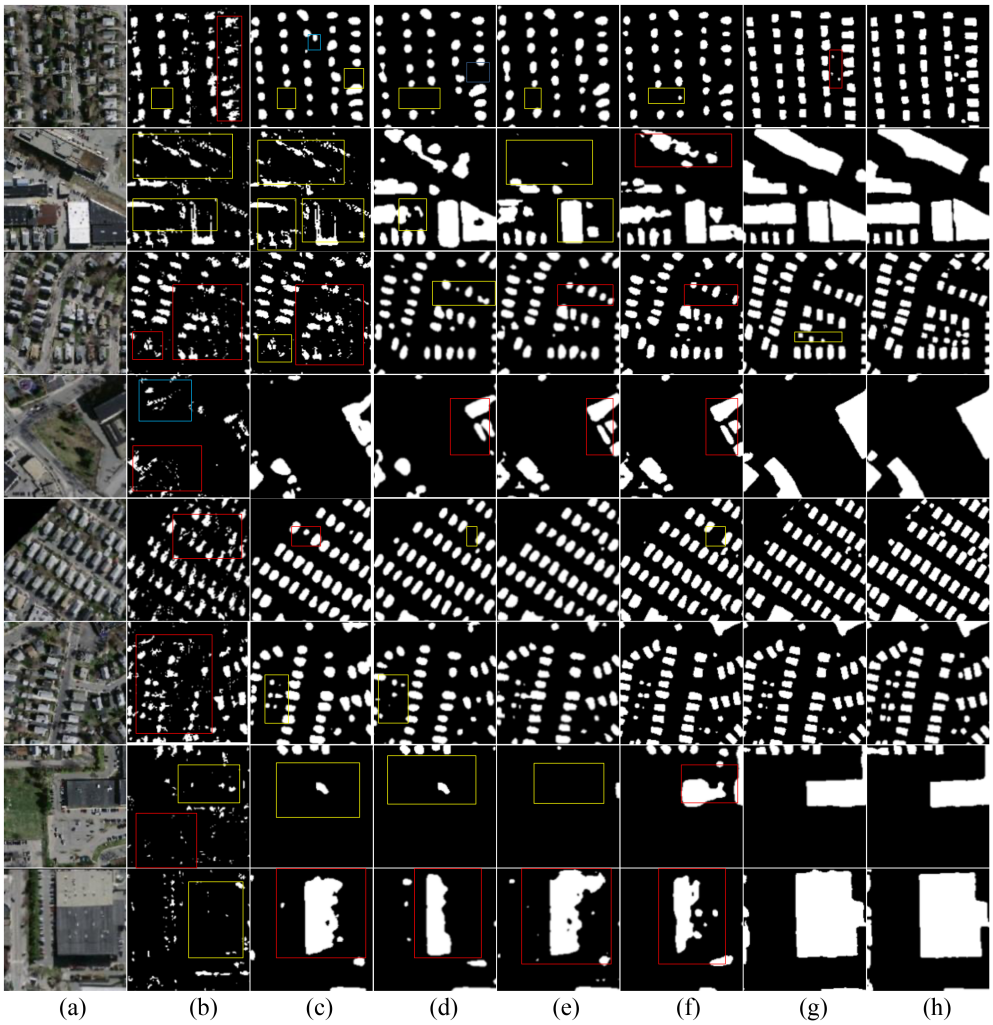


Fig. 17. Visual effect diagram of the detail part of the building segmentation in the comparison experiment. (a) Original images. (b) MLP. (c) FCN8s. (d) BisNet. (e) PSPNet-101. (f) UNet. (g) ResAt-Unet. (h) ResAt-UNet(2SEAE).

and choose to intercept the parts with poorer segmentation effect for analysis and comparison. The yellow frame in the following diagram indicates that the area is missing, and the building is mistaken as the background. The red frame indicates that the segmentation effect is poor. The blue frame indicates that the area is missing, and the background is mistaken as the building.

The first image in Fig. 17(a) is the remote sensing image of dense residential areas. The first image in Fig. 17(b) is the MLP segmentation result; the first image in Fig. 17(c) is the FCN8s segmentation result; the first image in Fig. 17(d) is the BisNet segmentation result; the first image in Fig. 17(e) is the PSPNet-101 segmentation result; the first image in Fig. 17(f) is the UNet segmentation result, which misses many buildings; the first image in Fig. 17(g) is ResAt-UNet segmentation result, which segments most residential buildings; the first image of Fig. 17(h) is the result of ResAt-UNet(2SEAE) segmentation, which detects all targets.

The second image in Fig. 17(a) is the remote sensing image of residential areas. The second image of Fig. 17(b) is the MLP segmentation result; the second image of Fig. 17(c) is the FCN8s segmentation result; the second image of Fig. 17(d) is the BisNet segmentation result; the second image of Fig. 17(f) is the UNet segmentation result, which misses the large targets; the second image of Fig. 17(g) is the ResAt-UNet segmentation result; the second image of Fig. 17(h) is the ResAt-UNet(2SEAE) segmentation result, which detects all the targets.

The third image in Fig. 17(a) is the remote sensing image of dense residential areas. The third image of Fig. 17(b) is the MLP segmentation result, and the segmentation result is poor. The third image of Fig. 17(c) is the FCN8s segmentation result; the third image of Fig. 17(d) is the BisNet segmentation result; the third image of Fig. 17(e) is the PSPNet-101 segmentation result; the third image of Fig. 17(f) is the UNet segmentation result; the third image of Fig. 17(g) is the ResAt-UNet segmentation result. These five methods miss some small buildings. The third image in Fig. 17(h) is the segmentation result of ResAt-UNet (2SEAE), which separates all building targets and backgrounds.

The fourth image in Fig. 17(a) is the remote sensing image of industrial areas. The fourth image of Fig. 17(b) is the

MLP segmentation result, the segmentation result of which is not ideal and the building segmentation is incomplete. The fourth image of Fig. 17(c) is the FCN8s segmentation result; the first image of Fig. 17(d) is the BisNet segmentation result; the fourth image of Fig. 17(e) is the PSPNet-101 segmentation result; and the first image of Fig. 17(f) is the UNet segmentation result. These four methods are incomplete for the segmentation of large warehouse targets. The fourth image of Fig. 17(g) is the ResAt-UNet segmentation result. The fourth image of Fig. 17(h) is the ResAt-UNet(2SEAE) segmentation result, which is more accurate.

The fifth image in Fig. 17(a) is the remote sensing image of dense residential areas. The fifth image of Fig. 17(b) is the MLP segmentation result, and its segmentation is relatively rough. The fifth image of Fig. 17(c) is the FCN8s segmentation result; the fifth image of Fig. 17(d) is the BisNet segmentation result; the fifth image of Fig. 17(e) is the PSPNet-101 segmentation result; the fifth image of Fig. 17(f) is the UNet segmentation result; and the fifth image of Fig. 17(g) is the ResAt-UNet segmentation result. The above methods segment most of the residential buildings and miss out on very small buildings. The fifth image of Fig. 17(h) is the ResAt-UNet(2SEAE) segmentation result, which is accurate and complete.

The sixth image in Fig. 17(a) is the remote sensing image of commercial areas. The sixth image in Fig. 17(b) is the MLP segmentation results, the segmentation effect of which is not ideal. The sixth image in Fig. 17(c) is the FCN8s segmentation result. The sixth image in Fig. 17(d) is the BisNet segmentation result, which misses individual buildings. The sixth image in Fig. 17(e) is the PSPNet-101 segmentation result; the sixth image in Fig. 17(f) is the UNet segmentation result; the sixth image in Fig. 17(g) is the ResAt-UNet segmentation result; and the sixth image in Fig. 17(h) is ResAt-UNet(2SEAE) segmentation results. These four methods split all targets.

The seventh image in Fig. 17(a) is the remote sensing image of industrial areas. The seventh image in Fig. 17(b) is the MLP segmentation result; the seventh image in Fig. 17(c) is the FCN8s segmentation result; the seventh image in Fig. 17(d) is the BisNet segmentation result; and the seventh image in Fig. 17(f) is the UNet segmentation result. The above methods are not ideal for the detection and extraction of large targets. The seventh image in Fig. 17(g) is the ResAt-UNet segmentation result, and the seventh image in Fig. 17(h) is the ResAt-UNet(2SEAE) segmentation result. These two methods can detect two large targets.

The eighth image in Fig. 17(a) is the remote sensing image of industrial areas. The eighth image of Fig. 17(b) is the MLP segmentation result, and its segmentation effect is poor. The eighth image of Fig. 17(c) is the FCN8s segmentation result; the eighth image of Fig. 17(d) is the BisNet segmentation result; the eighth image of Fig. 17(e) is the PSPNet-101 segmentation result; and the eighth image of Fig. 17(f) is the UNet segmentation result. The above methods have a poor extraction effect on large buildings. The eighth image of Fig. 17(g) is the ResAt-UNet segmentation result and the eighth image in Fig. 17(h) is the segmentation result of ResAt-UNet (2SEAE). These two methods can segment large buildings and small targets. Information,

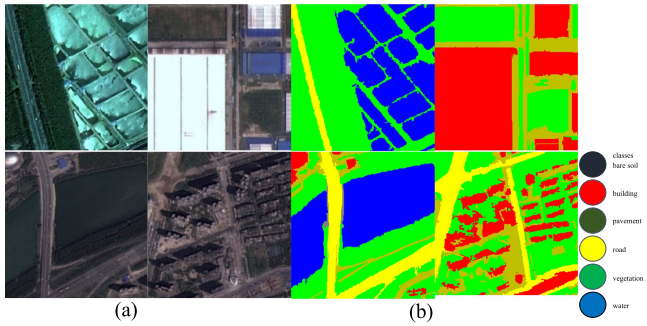


Fig. 18. WHDLD datasets. (a) Original images. (b) Ground truth.

attention mechanism, and residual module are added. In the encoding stage, the residual module of ResNet34 network is used for down-sampling. Meanwhile, SEAE is added to the intermediate connection layer, which makes a full use of the correlation of intermediate features and improves the accuracy of the network segmentation image.

4) Generalization Experiment: To verify the robustness and generalization ability of the ResAt-UNet(2SEAE), we use the model to carry out multilabel segmentation on The WHDLD dataset [45]. The WHDLD dataset is an open-source dataset for remote sensing image segmentation, published by Wuhan University. The image is $256 \times 256 \times 3$ and categories are divided into 6 classes containing bare soil, buildings, pavement, roads, vehicles, and water, with a total of 4940 images. We randomly divided the training sets and validation sets at a ratio of 0.8:0.2 for each category, among which 3952 images were used for training and 988 were used for testing. The images and labels of the WHDLD datasets are shown in Fig. 18. We use four widely used evaluation indexes: MPA, MP, MIOU, MRecall, the representations of which are as follows:

$$\mathrm{MP} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{25}$$

$$\mathrm{MPA} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \tag{26}$$

$$\mathrm{MIOU} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{27}$$

$$\mathrm{MRecall} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{28}$$

where TP stands for true positives, FP stands for false positives, TN stands for true negatives, and FN stands for false negatives, k stands for numbers of categories.

As shown in Table IX, ResAt-UNet(2SEAE) has a better degree of misclassification, though DFA-Net achieves better MPA values. Overall, ResAt-UNet(2SEAE) has better segmentation results. In contrast, ResAt-UNet(2SEAE) achieves finer segmentation results by focusing on different dimensions and enhancing the semantic representation of features between different categories, proving the robustness of the model. In contrast,

ResAt-UNet(2SEAE) not only enhances the feature extraction ability through residuals and attention in the encoder but also enhances the feature enhancement ability through SAM and CAM in the feature fusion stage, which makes ResAt-UNet(2SEAE) achieve the highest score on both datasets.

IV. Summary

The research of remote sensing image in urban areas is becoming increasingly important. Efficient and accurate building image segmentation algorithm has gradually attracted people's attention. The combination of deep learning and remote sensing image segmentation has become an inevitable trend. This article mainly introduces a new segmentation algorithm based on the UNet network: ResAt-UNet(2SEAE). Because the amount of urban remote sensing image data used in this article is small, and UNet requires few training samples, which meets the task requirements of this article, so it is selected as the basic framework. In addition, in view of the problem that the down-sampling of UNet is easy to lose context and detail information, attention mechanism and residual module are added. In the encoding stage, the residual module of ResNet34 network is used for down-sampling. Meanwhile, SEAE is added to the intermediate connection layer, which makes a full use of the correlation of intermediate features and improves the accuracy of the network segmentation image [43].

References

[1] S. Al-Amri and N. V. Kalyankar, "Image segmentation by using threshold techniques," 2010, *arXiv:1005.4020*.

[2] J. Adams et al., "Plant segmentation by supervised machine learning methods," *Plant Phenome J.*, vol. 3, no. 1, 2020, Art. no. e20001.

[3] A. Pratondo, C. K. Chui, and S. H. Ong, "Integrating machine learning with region-based active contour models in medical image segmentation," *J. Vis. Commun. Image Representation*, vol. 43, pp. 1–9, 2017.

[4] P. Getreuer, "Chan-vese segmentation," *Image Process. Line*, vol. 2, pp. 214–224, 2012.

[5] Y. T. Liow and T. Pavlidis, "Use of shadows for extracting buildings in aerial images," *Comput. Vis. Graph. Image Process.*, vol. 49, pp. 242–277, 1990.

[6] R. Avudaiammal et al., "Extraction of buildings in urban area for surface area assessment from satellite imagery based on morphological building index using SVM classifier," *J. Indian Soc. Remote Sens.*, vol. 48, no. 9, pp. 1325–1344, 2020.

[7] J. Yang, Y. He, and J. Caspersen, "Region merging using local spectral angle thresholds: A more accurate method for hybrid segmentation of remote sensing images," *Remote Sens. Environ.*, vol. 190, pp. 137–148, 2017.

[8] C. Lu, M. Xia, M. Qian, and B. Chen, "Dual-branch network for cloud and cloud shadow segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5410012.

[9] S. Miao et al., "Cloud/shadow segmentation based on multi-level feature enhanced network for remote sensing imagery," *Int. J. Remote Sens.*, pp. 1–21, 2022, doi: 10.1080/01431161.2021.2014077.

[10] C. Lu, M. Xia, and H. Lin, "Multi-scale strip pooling feature aggregation network for cloud and cloud shadow segmentation," *Neural Comput. Appl.*, vol. 34, pp. 6149–6162, 2022.

[11] M. Xia, Y. Qu, and H. Lin, "PADANet: Parallel asymmetric double attention network for clouds and its shadow detection," *J. Appl. Remote Sens.*, vol. 15, no. 4, 2021, Art. no. 046512.

[12] Y. Qu, M. Xia, and Y. Zhang, "Strip pooling channel spatial attention network for the segmentation of cloud and cloud shadow," *Comput. Geosci.*, vol. 157, 2021, Art. no. 104940.

[13] Z. Wang, M. Xia, M. Lu, L. Pan, and J. Liu, "Parameter identification in power transmission systems based on graph convolution network," *IEEE Trans. Power Del.*, vol. 37, no. 4, pp. 3155–3163, Aug. 2022.

[14] B. Chen et al., "MANet: A multilevel aggregation network for semantic segmentation of high-resolution remote sensing images," *Int. J. Remote Sens.*, vol. 43, no. 15/16, pp. 1–21, 2022, doi: 10.1080/01431161.2022.2073795.

[15] J. Gao et al., "MLNet: Multichannel feature fusion lozenge network for land segmentation," *J. Appl. Remote Sens.*, vol. 16, no. 1, 2022, Art. no. 016513.

[16] Y. Sun et al., "A multi-attention UNet for semantic segmentation in remote sensing images," *Symmetry*, vol. 14, 2022, Art. no. 906.

[17] Z. Chen et al., " Self-attention in reconstruction bias U-Net for semantic segmentation of building rooftops in optical remote sensing images," *Remote Sens.*, vol. 13, no. 13, 2021, Art. no. 2524.

[18] A. Soni, R. Koner, and V. G. K. Villuri, "Fusion of dual-scale convolution neural network for urban building footprints," *Ain Shams Eng. J.*, vol. 13, no. 3, 2022, Art. no. 101622.

[19] P. Zhang et al., "3D urban buildings extraction based on airborne LiDAR and photogrammetric point cloud fusion according to U-Net deep learning model segmentation," *IEEE Access*, vol. 10, pp. 20889–20897, 2022.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[21] O. Ronneberger, P. Fischer, and T. Brox, "Unet: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Berlin, Germany: Springer-Verlag, 2015, pp. 234–241.

[22] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[23] H. Zhao et al., "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[24] K. He et al., "Deep residual learning," *Image Recognit.*, 2015.

[25] S. Woo et al., *CBAM: Convolutional Block Attention Module*. Berlin, Germany: Springer-Verlag, 2018.

[26] J. Park et al., "Bam: Bottleneck attention module," 2018, *arXiv: 1807.06514*.

[27] H. Nanjun, F. Leyuan, and P. Antonio, "Hybrid first and second order attention Unet for building segmentation in remote sensing images," *Sci. China Inf. Sci.*, vol. 63, no. 4, pp. 69–80, 2020.

[28] I. O. Tolstikhin et al., "Mlp-mixer: An all-mlp architecture for vision," *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24261–24272.

[29] E. Thomas et al., "Multi-res-attention UNet: A CNN model for the segmentation of focal cortical dysplasia lesions from magnetic resonance images," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1724–1734, May 2021.

[30] O. Oktay et al., "Attention u-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[31] J. Pan et al., "Salgan: Visual saliency prediction with generative adversarial networks," 2017, *arXiv:1701.01081*.

[32] V. Mnih, "Machine learning for aerial image labeling," Accessed: May 03, 2021. [Online]. Available: https://www.cs.toronto.edu/∼{}vmnih/data/

[33] W. C. Tu et al., "Learning superpixels with segmentation-aware affinity loss," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 568–576.

[34] X. Li et al., "Dice loss for data-imbalanced NLP tasks," 2019, *arXiv: 1911.02855*.

[35] I. Joo et al., "CT-monitored minimal ablative margin control in single-session microwave ablation of liver tumors: An effective strategy for local tumor control," *Eur. Radiol.*, vol. 32, pp. 6327–6335, 2022.

[36] T. Walsh et al., "Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia," *Science*, vol. 320, no. 5875, pp. 539–543, 2008.

[37] C. Schuldt, I. Laptev, and B. Caputo, " Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit.*, vol. 3, 2004, pp. 32–36.

[38] P. Boonma and J. Suzuki, "BiSNET: A biologically-inspired middleware architecture for self-managing wireless sensor network," *Comput. Netw.*, vol. 51, no. 16, pp. 4599–4616, 2007.

[39] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[40] M. Chen et al., "DR-Net: An improved network for building extraction from high resolution remote sensing image," *Remote Sens.*, vol. 13, 2021, Art. no. 294.

[41] W. Wang and H. Zhu, "Main aortic segmentation from CTA with deep feature aggregation network," in *Proc. IEEE Int. Conf. Imag. Syst. Techn.*, 2018, pp. 1–5.

[42] P. Liu et al., "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 830.
[43] Y. Sun et al., "A multi-attention UNet for semantic segmentation in remote sensing images," *Symmetry*, vol. 14, no. 5, 2022, Art. no. 906.
[44] Z. Shao et al., "BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images," *Remote Sens.*, vol. 12, no. 6, 2020, Art. no. 1050.
[45] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.

Zhiyong Fan received the M.Sc. degree in system analysis and integration from Nanjing University of Information Science and Technology, Nanjing, China, in 2005, and the Ph.D. degree in pattern recognition and intelligent systems from Nanjing University of Science and Technology, Nanjing, China, in 2016.

He is currently a Lecturer with Nanjing University of Information Science and Technology. His current research interests include machine learning and its application.

Yu Liu received the B.S. degree in cybernetics control engineering from College of Automation, Nanjing University of Information and Technology, Nanjing, P.R. China, in 2022.

He is mainly engaged in image processing, computer vision. His research interests include semanic segmentation and target detection.

Min Xia (Member, IEEE) received the Ph.D. degree in cybernetics control engineering from Donghua University, Shanghai, China, in 2009.

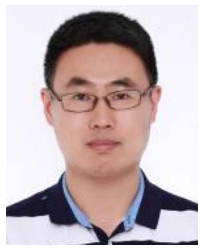
He is currently a Professor with Nanjing University of Information Science and Technology. He is currently the Deputy Director of Jiangsu Key Laboratory of Big Data analysis technology. His research interests include machine learning theory and its application, graph structure data analysis.

Jianmin Hou received the Ph.D. degree in management science and engineering (economic and industrial management) from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2014.

She is currently an Associate Professor with Nanjing University of Information Science and Technology. Her current research interests include integrated energy system design and optimization and artificial intelligence.

Fei Yan received the Ph.D. degree in circuit and system from the University of Chinese Academy of Sciences, Nanjing, P.R. China, in 2014.

He is mainly engaged in image processing, information display technology. His research interests cover 3-D measurement and imaging, and embedded application.



Qiang Zang received the Ph.D. degree in control theory and control engineering from Southeast University, Nanjing, China, in 2009.

He is currently an Associate Professor with Nanjing University of Information Science and Technology. His research interests include complex systems control and applications.