Hindawi
Journal of Sensors
Volume 2024, Article ID 5575787, 26 pages
https://doi.org/10.1155/2024/5575787

Research Article

A Deep Learning Method for Building Extraction from Remote Sensing Images by Fuzing Local and Global Features

Yitong Wang,[1] Shumin Wang,[1] Jing Yuan,[2] Aixia Dou,[1] and Ziying Gu[1]

[1]*Institute of Earthquake Forecasting, CEA, Beijing 100036, China*
[2]*School of Information Engineering, Institute of Disaster Prevention, Langfang 065201, China*

Correspondence should be addressed to Shumin Wang; wmcnu@163.com

Received 25 July 2023; Revised 17 December 2023; Accepted 18 January 2024; Published 10 February 2024

Academic Editor: Kathiravan Srinivasan

Copyright © 2024 Yitong Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

As important disaster-bearing bodies, buildings are the focus of attention in seismic disaster risk assessment and emergency rescue. It is of great practical significance to extract buildings quickly and accurately with complex textures and variable scales and shapes from high-resolution remote sensing images. We proposed an improved TransUnet model based on multiscale grouped convolution and attention named MATUnet to retain more local detail features and enhance the representation ability of global features, while reducing the network parameters. We designed the multiscale grouped convolutional feature extraction module with attention (GAM) to enhance the representation of detailed features. The convolutional positional encoding module (PEG) was added to redetermine the number of transformer, it solved the problem of local feature information loss and the difficulty of convergence of the network. The channel attention module (CAM) of the decoder enhanced the salient information of the features and solved the problem of information redundancy after feature fusion. We experimented through MATUnet on the WHU building dataset and Massachusetts dataset. MATUnet achieved the best IOU results of 92.14% and 83.22%, respectively, and achieved better than the other generalized and state-of-the-art networks under the same conditions. We also have achieved good segmentation results on the GF2 Xichang building dataset.

1. Introduction

Building extraction based on high-resolution remote sensing images provides important technical support for earthquake disaster risk assessment and postdisaster emergency response. The development of high-resolution earth observation technology has led to more diverse and complex acquired remote sensing image data [1], presenting both opportunities and challenges for rapid and accurate building extraction. Meanwhile, the excellent performance of deep learning networks in image feature extraction and nonlinear function fitting has received extensive attention from scholars. With the great advantage of fully convolutional networks (FCN) [2] in the field of image segmentation, semantic segmentation methods based on convolutional neural networks (CNN) started to be proposed continuously. Subsequently, encoder–decoder structures have gradually been widely used in the field of segmentation, Ronneberger et al. [3] designed the Unet network, and Badrinarayanan et al. [4] designed the Segnet model, both of which improved the extraction accuracy of the model through the encoding–decoding structure, and brought new inspirations for the framework of semantic segmentation network. To improve the accuracy of deep learning methods in remote sensing building extraction, some studies by Xu et al. [5–8] have made a lot of improvements to the above-mentioned network, mainly including three strategies, i.e., achieving a larger receptive field by multiscale feature Extraction methods [9], enriching feature information through multibranching structure [10, 11], and reinforcing salient features through attention mechanisms [8, 12]. Sun et al. [13] utilized a multiscale attention approach based on Unet to recognize buildings with complex scales. Che et al. [14] proposed multiattention feature fusion HRNet [15], which preserves more detailed features based on the multibranch structure for accurate semantic segmentation. MSRF-Net [16] used different scale convolutional kernels with multiple branches in the encoder and decoder to extract features on the different scales to preserve multiscale contextual information. Yu et al. [17]

adopted ConvNeXt [18] to extract multiscale abstract features, and presented the attention module to selectively focus on some important information, improving accuracy in building extraction tasks. Shi et al. [19] employed channel–spatial attention to the fuzed features of the encoder and decoder for achieving discriminative and attentive features. The above methods improve the extraction accuracy of buildings by combining different strategies. However, due to the inherent limitations of the convolutional kernel [20, 21], the model receives limitations in capturing contextual dependencies, resulting in suboptimal semantic segmentation results.

Given the exceptional capacity of the transformer structure [22] to capture contextual features, VIT [23] was the pioneer in applying it to computer vision tasks. By creating a pure transformer with a series of image chunks as input, it achieved outstanding results in image classification tasks. The swin transformer [24] introduced a feature pyramid structure to address the low-output resolution of transformer models like VIT. This innovation not only boosted performance in semantic segmentation tasks but also decreased computational requirements. Zheng et al. [25] introduced a network known as SETR, which transforms the output of the transformer from vectors into an image. This was the first attempt to apply a transformer in the field of semantic segmentation. Yuan and Xu [26] proposed a multiscale adaptive network based on the swin transformer. This network effectively integrates the multilevel feature maps of the swin transformer to capture multiscale information, thereby enhancing the accuracy of semantic segmentation [27]. However, there are few pure transformer networks for building extraction, this is mainly because although transformers have excellent capabilities in extracting global information, they are not effective in extracting local detailed information [28]. The transformer structure lacks translation invariance and local correlation for convolution operation [23], which can neglect local information [29] and result in the loss of building detail features [30]. Therefore, some scholars have combined transformers with CNN to improve the models' feature extraction performance. Chen et al. [31] connected it to the Unet structure based on ResNet [32] and proposed the TransUnet network, addressing the issues of traditional convolution networks' inability to model the relationship of global features [33, 34], and achieved good results in the field of semantic segmentation. But, TransUnet network still has some problems to be improved in remote sensing building extraction. First, in the encoder, the traditional ResNet network has deeper layers, which may bring feature redundancy [35, 36]. The feature fusion of the decoding process does not consider the correlation between the features of different channels [37, 38], and these problems can lead to useful feature information not being effectively utilized. Second, the concatenation of convolution and transformer is operated only by linear interpolation, which can also result in the loss of feature information [39]. Meanwhile, the large computational volume of the transformer structure makes us think about the scope of application of the number of transformer layers in the field of remote sensing building extraction [40].

With the aim of further improving the extraction accuracy of remote sensing buildings, we proposed an improved TransUnet model, MATUnet, based on multiscale grouped convolution and attention mechanism in this paper. Different from TransUnet, we first designed a multiscale grouped convolutional feature extraction module with attention in the encoder part to capture richer feature information through grouped convolution with multiple branches in the shallow and middle layers, and utilized attention to enhance the global context information of the features on each convolutional branch in the deep layer. Second, deep separable convolution was utilized to implicitly construct the position information within a sequence of image blocks, contributing to the expeditious convergence of the transformer model [29, 41, 42]. In the decoder, the channel attention module (CAM) was employed to enhance the cascade feature fusion from the encoder, reinforcing the critical information of the features in the channel dimension [43]. Our MATUnet network was compared with other classical models and current state-of-the-art building extraction networks on WHU building dataset [44] and Massachusetts building dataset [45] to validate the advantage of model accuracy. We also conducted experiments on GF2 Xichang dataset to validate the effectiveness of the MATUnet model in the practical applications.

Overall, the contributions of our paper are mainly in the following areas:

(1) We proposed an improved TransUnet for building semantic segmentation based on multiscale grouped convolution and attention. Grouped convolution, depth-separable convolution, and attention methods enhance shallow feature representation and strengthen the global information representation of the deeper features, while the use of channel attention at the decoder strengthens the representation of feature-critical information, which improves the network extraction accuracy relative to TransUnet and the convergence speed.

(2) We proposed a multiscale grouped convolutional feature extraction module with attention in the encoder part to capture richer feature information through grouped convolution with multiple branches in the shallow and middle layers, and utilized attention to enhance the global context information of the features on each convolutional branch in the deep layer.

(3) We utilized depth-separable convolution to implicitly encode the position information of the transformer to accelerate network convergence, while revisiting the number of layers of the transformer to ensure the efficiency of the global information extraction of the model while reducing the computational complexity of the model. Meanwhile, we added a channel attention module to the decoder so that the encoder and decoder features are fuzed for channel-dimensional attention enhancement, which significantly improves the key information between channels.

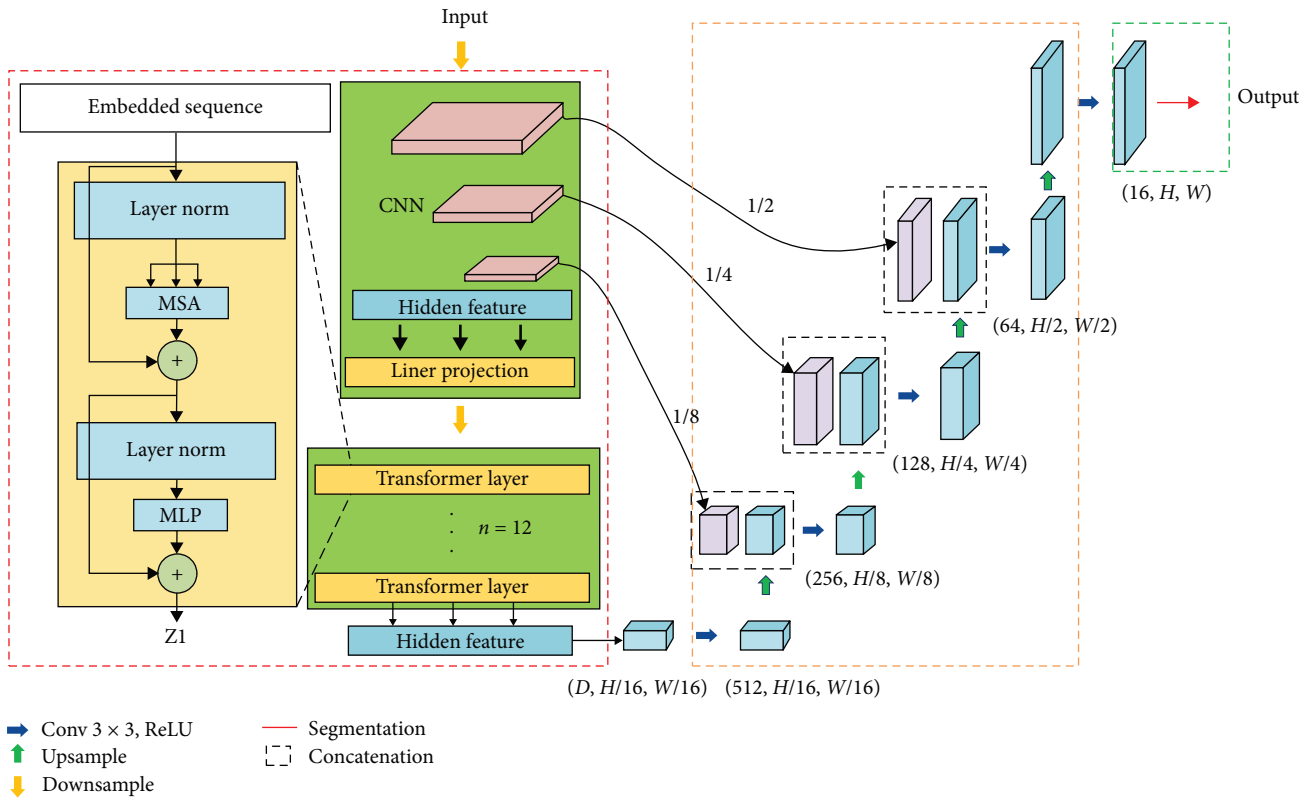


Figure 1: TransUnet network structure [31].

(4) We have achieved more significant results than the current state-of-the-art methods on two publicly available datasets, and we have also verified the effectiveness of the present model in practical application by applying the model in the GF2 image building dataset in Xichang City, Sichuan Province, China.

The remainder of this paper was organized as follows: Section 2 presented the related work. Section 3 described the specifics of the methodology of this paper. The experimental setup and the detailed analysis of results were shown in Section 4. Section 5 described the analyses of the selection of the different modules and covariates in the ablation experiments. Finally, Section 6 concluded this paper.

2. Related Works

In this section, we first presented the structure of the traditional TransUnet model, and the principle of grouped convolution to facilitate the understanding of readers for our proposed method.

2.1. TransUnet Model Overview. The TransUnet network model (Figure 1) uses an encoder that combines CNN and transformer networks, consisting of three main components: an encoder module based on CNN and transformer, a decoder module based on skip connections, and a feature extraction module.

(1) The encoder module based on CNN and transformer tandem (the red rectangular box in Figure 1). In the encoder, the original image is fed into the ResNet backbone network to obtain shallow and deep features of the buildings. The extracted shallow features are fuzed with the cascaded features sampled on the encoder, while the deep features are linearly interpolated and embedded through the image blocks as input to the transformer. The TransUnet network has a 12-layer transformer module, which collects global contextual information about the features by acquiring correlations between image blocks through the transformer's multi-head attention mechanism.

(2) The decoder module based on skip connections (the orange rectangular box in Figure 1). The deep features are combined with the shallow features of the same scale extracted by the CNN through upsampling. This prevents the loss of local building features that may occur solely from upsampling during image recovery, while also serving the purpose of decoding deep features and maintaining low-medium features.

(3) The feature extraction module (the green rectangular box in Figure 1) consists of a convolutional layer with a size of 3×3 convolutional kernel. This layer aims to maintain consistency between the feature map and the actual building labels.

2.2. The Grouped Convolution. The grouped convolution [45] (Figure 2) borrows the idea of the dot product between the input vector and the weights in a neuron. For an input

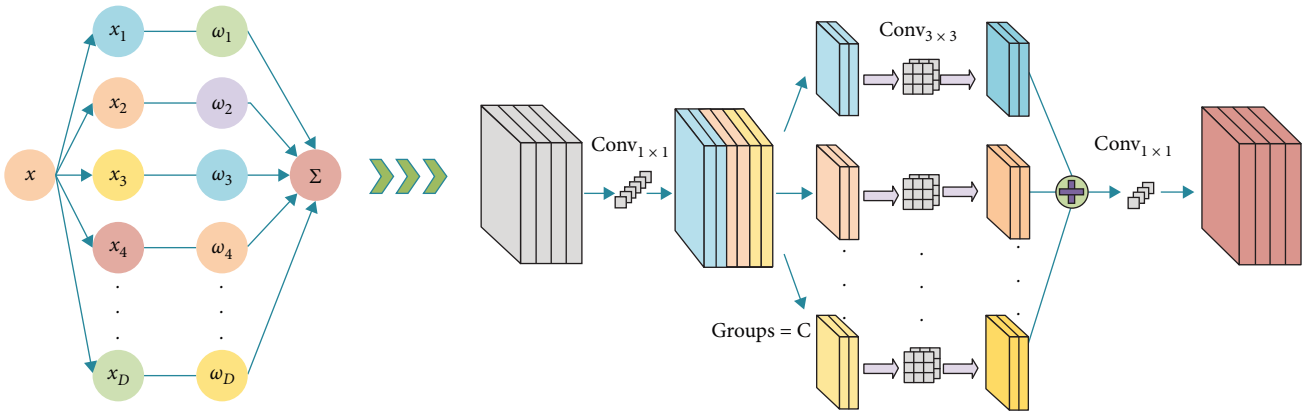


Figure 2: The structure of group convolution.

vector with a channel size of D: $x = [x_1, x_2, \ldots, x_D]$, the output obtained through the transformation of the neuron can be represented as $x = [x_1, x_2, \ldots, x_D]$. Grouped convolution considers the dot product operation in the neuron as three stages: "split-transform-aggregate." In other words, the input vector x is divided into multiple low-dimensional vectors x_i, then transformed by the corresponding number of weights ω_i in the neuron, and finally aggregates the transformed low-dimensional features, which is $\sum_{i=1}^{D} \omega_i x_i$. Applying this idea to neural networks, for input features x, there exists a function $F(x)$ that projects x onto a low-dimensional subspace with C channels, performs transformations, and finally aggregates the results, which is $\sum_{i=1}^{c} F_i(x)$. Each convolution operation divides the network into C groups, with the number of channels in each subnetwork changed from din to din/C. At this point, $F(x)$ can be considered a function with multiple convolution operations, i.e., $F(x) = \{\text{conv}_{1\times1}, \text{conv}_{3\times3}, \text{conv}_{1\times1}\}$, and finally aggregates them through concatenation to obtain the final feature output, $\sum_{i=1}^{c} F_i(x)$.

3. Methodology

To further improve the extraction accuracy of remotely sensed buildings, this paper proposed an improved TransUnet model based on multiscale group convolution and attention mechanism, MATUnet. Figure 3 represents the rough framework of MATUnet. The model is an encoder–decoder structure, which is different from TransUnet in that the encoder part consists of a multiscale grouped convolutional feature extraction module with attention (MGM) and eight transformer structures with convolutional position embedding module (PEG), and the decoder part adds the CAM. Specifically, MATUnet captures richer feature information at all four scales of the encoder through MGM and utilizes attention to enhance the global information of features in each convolutional branch. In addition, a depth-separable convolution with zero-padding in PEG is utilized to implicitly encode the position information and speed up the convergence of the transformer. In the decoder, MATUnet enhances the encoder and upsampling fusion features with CAM to strengthen the key information representation of features in each channel of the grouped convolution.

We detailed the MGM in the encoder in Section 3.1, and the PEG in Section 3.2, and validated the selection of the number of transformer layers in the subsequent ablation experiments. The CAM is introduced in Section 3.3, and the loss function in Section 3.4.

3.1. Multibranch Grouped Convolutional Feature Extraction Module. In encoder, traditional convolution brings redundancy as the number of layers increases, so we designed multibranch grouped convolutional feature extraction module (MGM) with attention to improve the ability of building feature extraction, which performed feature extraction by convolution of different branches to get more subfeatures than the traditional convolution. At the same time, to enhance the interaction of the information between subfeatures with different branches, we concatenated each subfeature with a global enhancement, which improved the representation of the salient information between subfeatures, and suppressed the irrelevant features. The whole module was shown in Figure 4.

First, to reduce the computational effort of high-resolution images on the model, 7×7 convolution kernel with a large perceptual field is employed to reduce the image resolution and preserve as many image features as possible. The input image $\boldsymbol{x} \in R^{512\times512\times3}$ is conducted convolution, and then max-pooling is adopted to obtain the building features $\boldsymbol{z}_1 \in R^{256\times256\times64}$, as shown in Equation (1).

$$z_1 = \text{ReLU}(\text{Maxpool}_{3\times3}(\text{conv}_{7\times7}(x))), \tag{1}$$

where $\text{conv}_{7\times7}(\cdot)$ denotes the 7×7 convolution, Maxpool$(\cdot)$ denotes max-pooling.

Subsequently, the features calculated through Equation (1) are input into the MGM to compute multiscale features. The number of MGM for the three scales is 3, 3, and 9. Each MGM contains a 1×1, 3×3, and 1×1 grouped convolution, respectively, with the number of groups set to 32. As shown in the red font in Figure 4, the step size of the 3×3 convolution in the first convolution module of each scale is set to 2 to reduce the feature scale. The features after the grouped convolution module fuze multiple subfeature information, which are pooled by global average and then subjected to SoftMax operation and

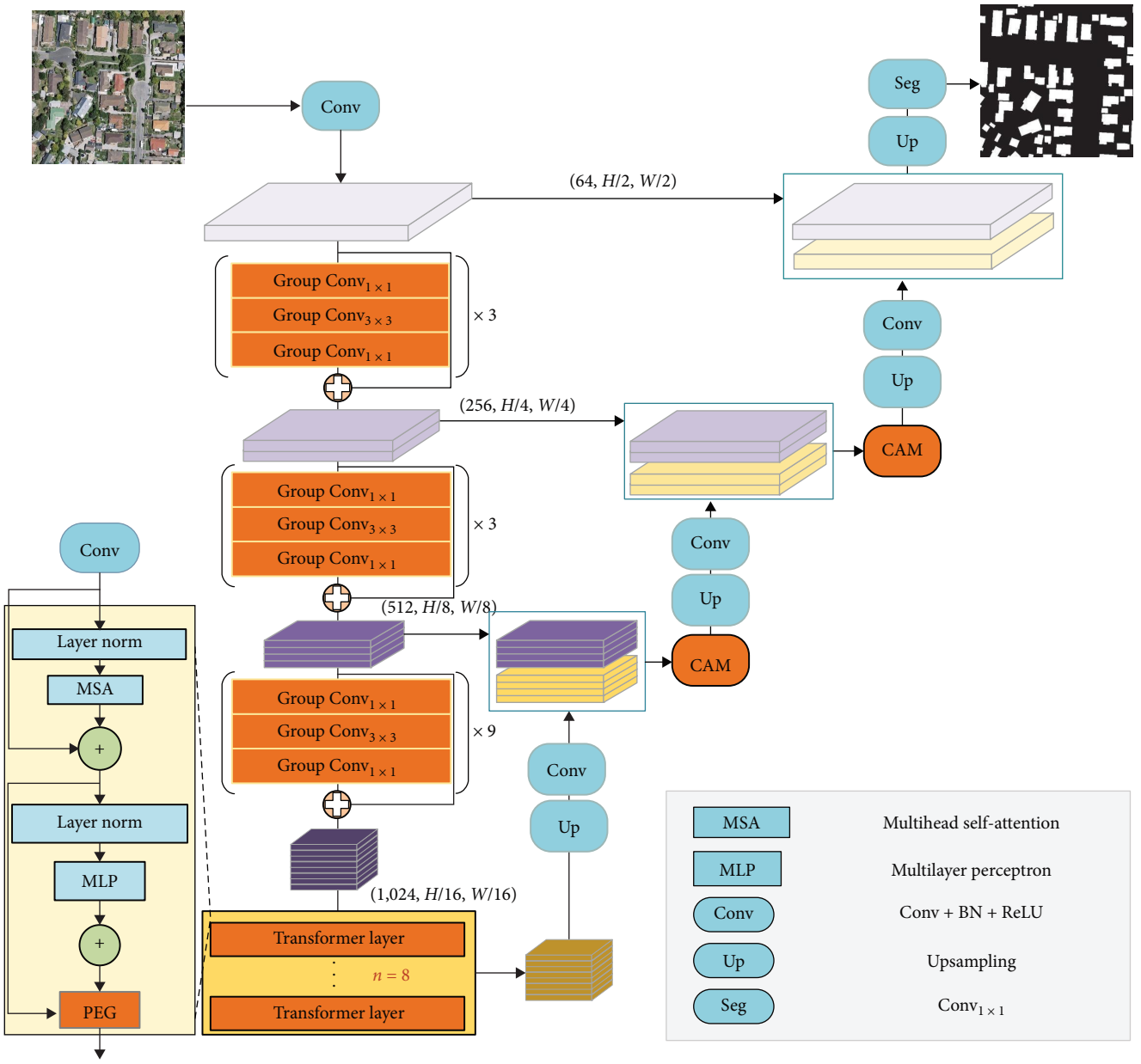


FIGURE 3: The structure of MATUnet.

multiplied with the original features to obtain the attention-enhanced features. The three scales of shallow building features are $z_2 \in R^{128\times128\times256}$, $z_3 \in R^{64\times64\times512}$, and $z_4 \in R^{32\times32\times1,024}$.

The MGM in this module (the red dashed rectangular box in Figure 4) is calculated as shown in Equations (2)–(6).

$$z_{i+1} = z \times \text{softmax}(\text{GAP}(z)), \tag{2}$$

$$z = \text{ReLU}\left(\text{conv}_{1\times1}\left(\text{concate}\left(\sum_{i=1}^{d} y_3^i + z_i\right)\right)\right), \tag{3}$$

$$y_3^i = \text{ReLU}(\text{conv}_{1\times1}^i(y_2^i))(0<i\le d), \tag{4}$$

$$y_2^i = \text{ReLU}(\text{conv}_{3\times3}^i(y_1^i))(0<i\le d), \tag{5}$$

$$y_1^i = \text{ReLU}(\text{conv}_{1\times1}^i(x'))(0<i\le d). \tag{6}$$

In Equation (2), GAP denotes global average pooling, in Equation (3), concate$(\cdot)$ denotes stitching the features $\sum_{i=1}^{d} y_3^i$ obtained by d group convolutions with the building features x, and the output of the convolution module z is obtained after nonlinear activation ReLU$(\cdot)$. Equations (3)–(5) represent the calculation of the convolution module with three layers of group convolution on features x', where $\text{conv}_{1\times1}^i(\cdot)$, $\text{conv}_{3\times3}^i(\cdot)$, and $\text{conv}_{1\times1}^i(\cdot)$, respectively, denotes the group convolution with convolution size of 1×1, 3×3, and 1×1. The i indicates the i-th group convolution. It improves the channel local correlation of building features by d parallel identical convolutions.

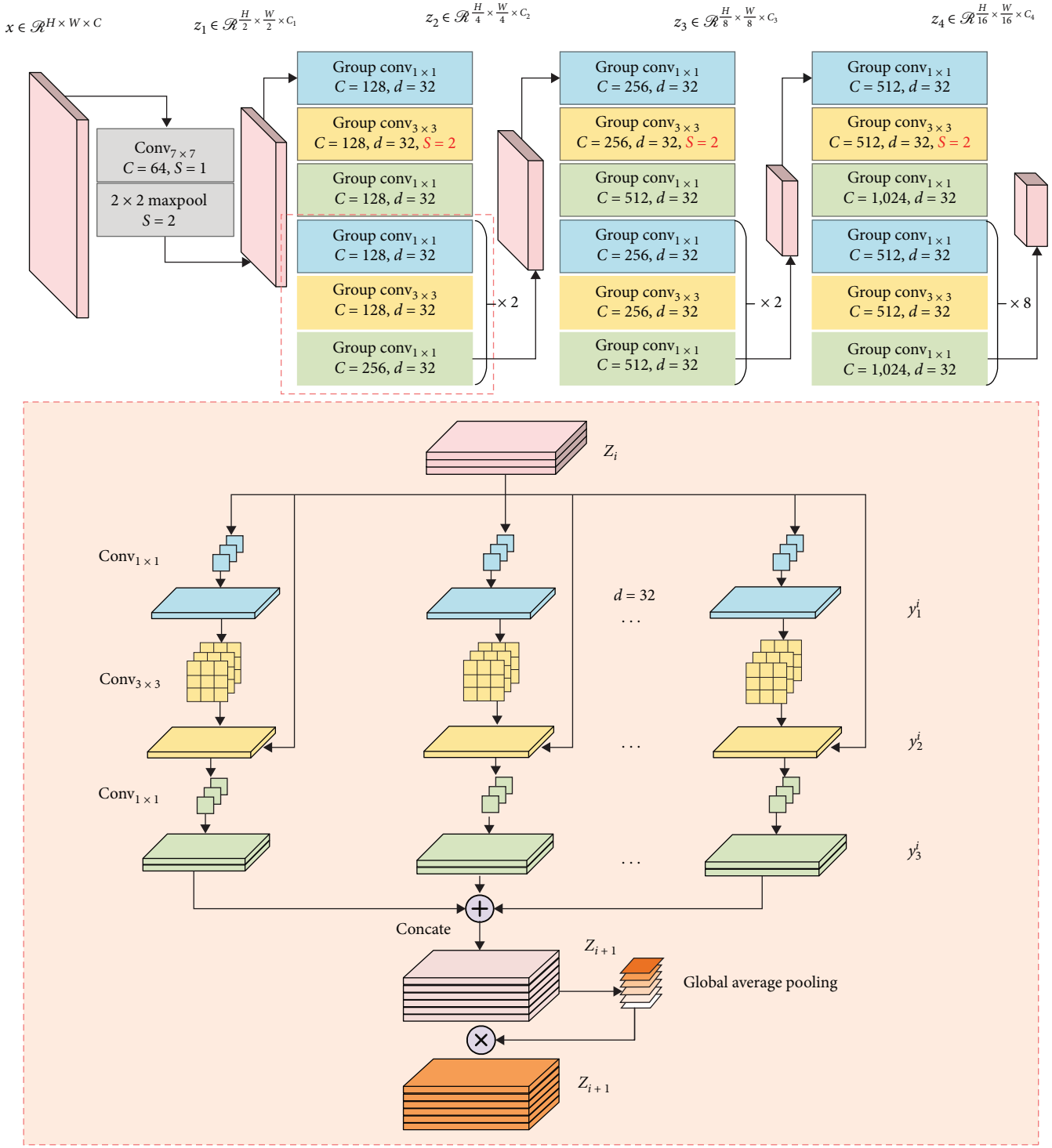


Figure 4: Structure of multiscale feature extraction module.

The improved module is shown in the red rectangular box ① in Figure 3.

3.2. Transformer Structures with PEG. In the transformer structures with PEG, depthwise separable convolution, zero-padding convolution, and the attention mechanism are employed to fuze the local and global features of the building. Transformer structure with PEG was shown in Figure 5, we will introduce the PEG module in the next process.

First, the deep-level features $X \in R^{32\times32\times1,024}$ are input to 8×8 convolution, which includes zero-padding, and the channel size is reduced to $X' \in R^{14\times14\times768}$, which preserves more localized features compared to the original TransUnet through linear interpolation operations. Meanwhile, the zero

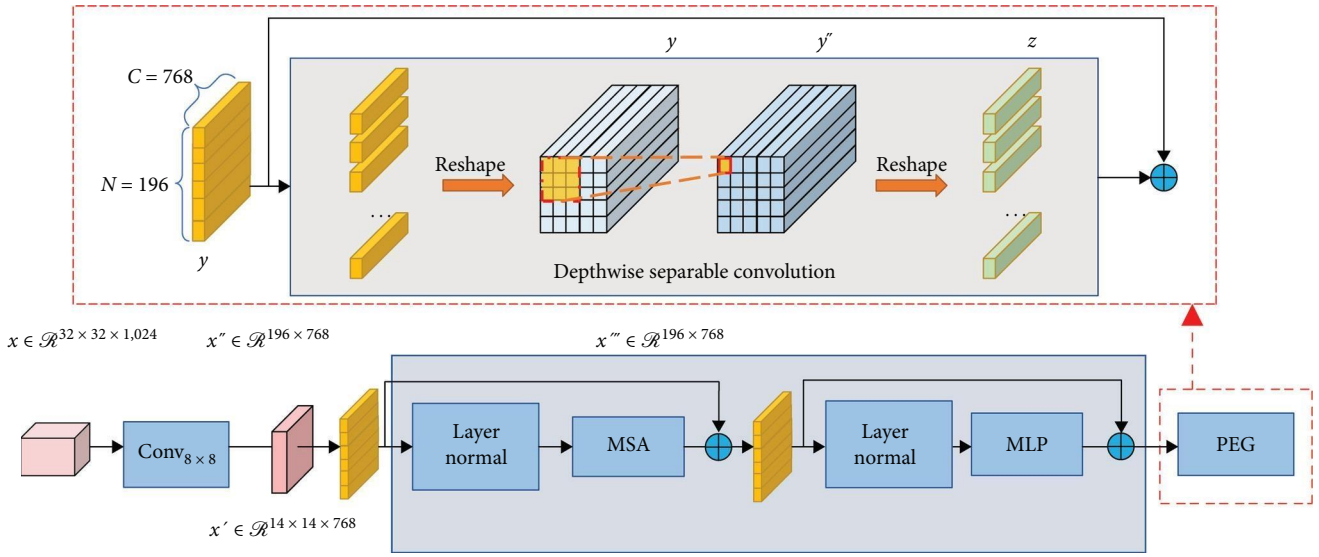


Figure 5: Transformer structure with PEG.

value of a filled position is computed in a convolution operation with other input values of nonfilled positions, thus preserving the positional information in the output. This implicit encoding of positional information helps the transformer to understand the relative relationship of different positions in the sequence [46]. Subsequently, X' is chunked without overlapping to get the image block sequence $X'' \in R^{N\times C}$ as transformer input. After the attention calculation, the features calculated by multi-head attention are spliced with X'' to get the image block sequence $X''' \in R^{N\times C}$. Subsequently, X''' is input into the MLP module for nonlinear transformation to obtain Y. And then Y is input into the PEG module, which is resized as $H \times W \times D$. Considering the excellent extraction performance of depthwise separable convolution in remote sensing semantic segmentation task [47], local semantic information interaction is performed on building features through depthwise separable convolution to obtain positional encoding information $Y'' \in R^{H\times W\times D}$. The image boundary effect and the zero-padding operation of convolution are adopted to obtain the encoding position information to achieve the purpose of strengthening the local semantic information of the building [48]. Finally, the building features Y'' with location encoding information are reshaped to the image block sequence size and added with transformer output features to obtain $Z \in R^{N\times C}$, and then Z is input into the next module. The complete calculation process is shown in Equations (7)–(12), the PEG calculation process includes Equations (7) and (8).

$$Z = \text{reshape}_2\left(Y''\right) + Y, \tag{7}$$

$$Y'' = \text{reshape}_1(Y) + \text{GN}(\text{DSW}(\text{reshape}_1(Y))), \tag{8}$$

$$Y = \text{concate}(W_2(\text{GeLU}(W_1(\text{LN}(X''')) + b_1)) + b_2), \tag{9}$$

$$X''' = \text{concate}\left(\text{MSA}\left(\text{LN}\left(X''\right)\right) + X''\right), \tag{10}$$

$$X'' = \text{split}_{1\times1}(X'), \tag{11}$$

$$X' = (\text{conv}_{8\times8}(X)). \tag{12}$$

In Equation (7), $\text{reshape}_2(\cdot)$ indicates reshaping the building features to a sequence of image blocks with size $N \times C$. In Equation (8), $\text{reshape}_1(\cdot)$ denotes reshaping the image block sequence $Y \in R^{N\times C}$ to the building features, $\text{DSW}(\cdot)$ denotes the depthwise separable convolution with a 7×7 convolution kernel and the padding 3. In Equation (9), W_1, b_1 and W_2, b_2 denote the weights of the two fully connected layers in the multilayer perception (MLP), and $\text{GeLU}(\cdot)$ [49] denotes the nonlinear activation function. In Equation (10), LN denotes the normalization of X'' in C dimensions [50], $\text{MSA}(\cdot)$ indicates the calculation of multi-head self-attention, and $\text{concate}(\cdot)$ denotes the stitching of X'' with the features computed by multi-head attention. In Equation (11), $\text{split}_{1\times1}(\cdot)$ denotes a 1×1 image block split window. In Equation (12), $\text{conv}_{8\times8}(\cdot)$ denotes the deep convolution with convolution kernel K ($K = 8 \times 8$), step size S ($S = 3$), and padding value (Padding = 1).

3.3. Channel Attention Module. Due to the differences in the information on different channels of the shallow features from the encoder and the deep features sampled on the decoder, the channel attention enhancement module is added at the skip connection to optimize the integration of the two features. The most significant feature of the feature on the channel dimension is computed using global maximum pooling, and the mean of the feature on the channel dimension is computed using global average pooling, and the two are used to aggregate the spatial information of the features by summing and sigmoid nonlinear activation to obtain the channel attention weight map, and finally elementwise multiplication operation is performed with the input features to obtain the enhanced features on the channel

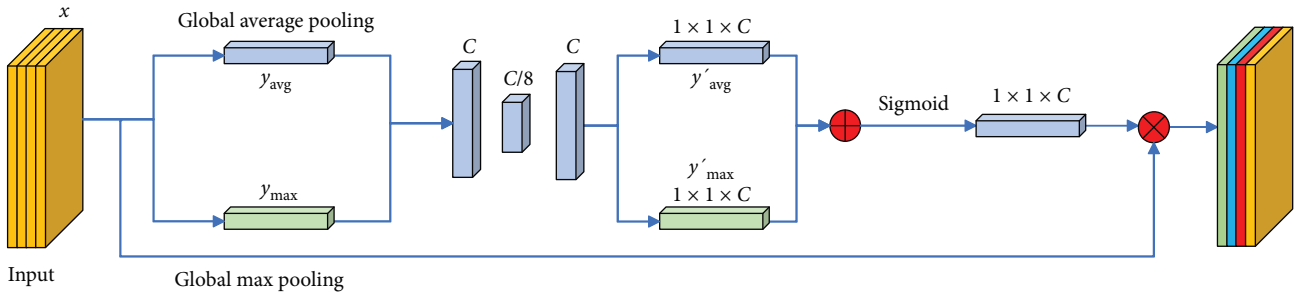


Figure 6: Channel attention enhancement module [43].

dimension. The channel attention enhancement module is shown in Figure 6.

In this module, global max pooling [44] and global average pooling [51] are performed on the building feature $x \in R^{H\times W\times C}$ in the channel dimension to obtain two $1\times 1\times C$ building features, y_{max} and y_{avg}. Subsequently, they are input to a multilayer perceptron with shared weights for semantic interaction, and then the features y'_{max} and y'_{avg} outputted from the perceptron are summed and input into nonlinear activation to obtain the channel attention weight map, and finally multiplied with the spliced features to obtain the channel-enhanced features $z_{channel}$. The complete calculation process is shown in Equations (13)–(17).

$$z_{channel} = x \times \text{Sigmoid}\left(y'_{avg} + y'_{max}\right), \tag{13}$$

$$y'_{max} = \text{MLP}_C\left(\text{ReLU}\left(\text{MLP}_{\frac{C}{r}}\left(y_{max}\right)\right)\right), \tag{14}$$

$$y'_{avg} = \text{MLP}_C\left(\text{ReLU}\left(\text{MLP}_{\frac{C}{r}}\left(y_{avg}\right)\right)\right), \tag{15}$$

$$y_{avg} = \text{GAP}(x), \tag{16}$$

$$y_{max} = \text{GMP}(x). \tag{17}$$

In Equation (13), Sigmoid$(\cdot)$ is a nonlinear activation function. In Equation (14), $\text{MLP}_{\frac{C}{r}}(\cdot)$ indicates the first layer of the perceptron, which has C/r neurons (r is the reduction rate, here $r = 8$), and ReLU$(\cdot)$ is the activation function. In Equation (15), $\text{MLP}_C(\cdot)$ is the second layer of the perceptron which has C neurons. In Equation (16), GAP$(\cdot)$ denotes global average pooling operation. In Equation (17), GMP$(\cdot)$ is global max pooling operation.

Finally, other building features after attention enhancement are up-sampled through three layers, and then the results are input to a 3×3 convolution for semantic segmentation. The predicted building extraction results are obtained. The improved module is shown in the green rectangular box in Figure 3.

3.4. Loss Function. In this paper, the loss function L_{total} which is combined cross-entropy loss function L_{ce} with Dice loss function L_D [52] was selected to optimize the predicted values in the training process. When the loss function corresponds to the smallest loss value during the training process, weight parameters ω in the network are solved, as shown in Equation (18), and the weights of L_{ce} and L_D are set to 0.5.

$$\text{argmin}(L_{total}|\omega) = \text{argmin}(0.5\times L_{ce} + 0.5\times L_D|\omega), \tag{18}$$

where L_{ce} denotes the cross-entropy loss function, L_D denotes the Dice loss function.

Cross-entropy loss function L_{ce} is defined as Equation (19).

$$L_{ce} = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i 1 - \frac{1}{N}\sum_{c=1}^{C} y_i \lg(p_i), \tag{19}$$

where C is the number of sample categories (in this paper $C = 1$), and y_i indicates the sample belongs to category or not ($y_i = 1$ or 0). p_i denotes the probability that sample i belongs to category c. L_{ce} is used to evaluate the loss incurred when classifying pixels in the image segmentation process. It measures the degree of difference between the labels and the predicted values. The smaller the function value, the more similar they are, and the better the model prediction.

Dice loss function L_D is defined as Equation (20).

$$L_D = 1 - \frac{2|x \cap y|}{|x| + |y|}, \tag{20}$$

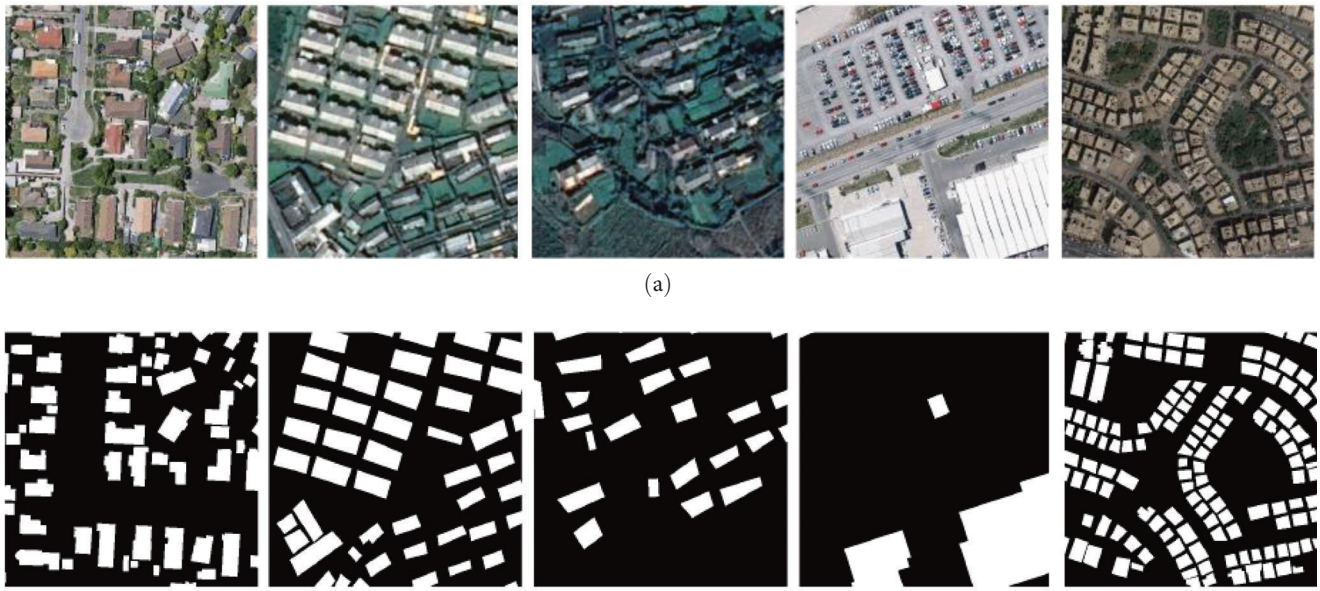
where $|x \cap y|$ is the intersection of true samples and predicted samples, and $|x| + |y|$ denotes the union of true samples with predicted samples. $|x|$ and $|y|$ indicate the element number of the samples, respectively. L_D is the loss metric used to evaluate the similarity between the predicted images and the real images.

4. Materials and Methods

In this section, we focus on the dataset we used in Section 4.1 and the preprocessing of the data in Section 4.2.

4.1. Introduction of the Experimental Dataset

4.1.1. Wuhan University Building Dataset. To verify the building extraction capability of the network model proposed in this paper, the sample dataset was produced using WHU building dataset (http://gpcv.whu.edu.cn/data/building_dataset.html) to train, validate and test the model. The building



(a)

(b)

Figure 7: Images (a) and sample labels (b) of WHU building dataset.

experiment dataset of the Wuhan University (Figure 7) is a large building dataset composed of multisource remote sensing images, mainly including aerial and satellite images, each of which is 512 × 512 pixels. Among them, there are 8,819 aerial images with 0.3 m spatial resolution, covering ground area about 450 km^2, and 17,388 satellite images (Satellite Dataset II (East Asia)) with 2.7 m spatial resolution, covering ground area about 550 km^2. The labels of the whole building dataset are divided into building and background. In this paper, 65% of the images in the dataset were randomly selected as the training set, 5% of the images were randomly selected as the validation set, and the remaining 30% of the images were the test set for training and testing the building extraction capability of the network.

4.1.2. Massachusetts Dataset. To verify the building extraction capability of the improved network in this paper, the Massachusetts building dataset (https://www.cs.toronto.edu/-vmnih/data/) was also selected for training and testing the network to further demonstrate the robustness of the network model in this paper. The dataset covers urban and suburban areas in the Boston area of the United States, such as office buildings, individual homes and garages, and other buildings. The dataset includes 151 high-resolution remote sensing images with a size of 1,500 × 1,500 pixels and 1.0 m spatial resolution, covering ground area about 340 km^2. After random cropping, an image dataset with 512 × 512 pixels for each image were generated (Figure 8). About 3,000,200 and 1,200 images were randomly selected from them as the training, evaluation, and test sets.

4.1.3. GF2 Xichang City Research Area. To verify the building extraction capability of the improved network proposed in this paper in the practical application process, the GF-2 remote sensing imagery collected in Xichang City, Liangshan Yi Autonomous Prefecture, Sichuan Province was selected, and 1 m resolution image (Figure 9) was obtained after orthorectification, image fusion, and mosaic. The images offered regions from ①–④ in Figure 9 were selected as the training images for the network model. Each red area has 3,000 × 5,000 pixels, and the image in green area is the test data with 6,500 × 10,000 pixels. After random cropping, a sample dataset with 512 × 512 pixels for each image block was obtained.

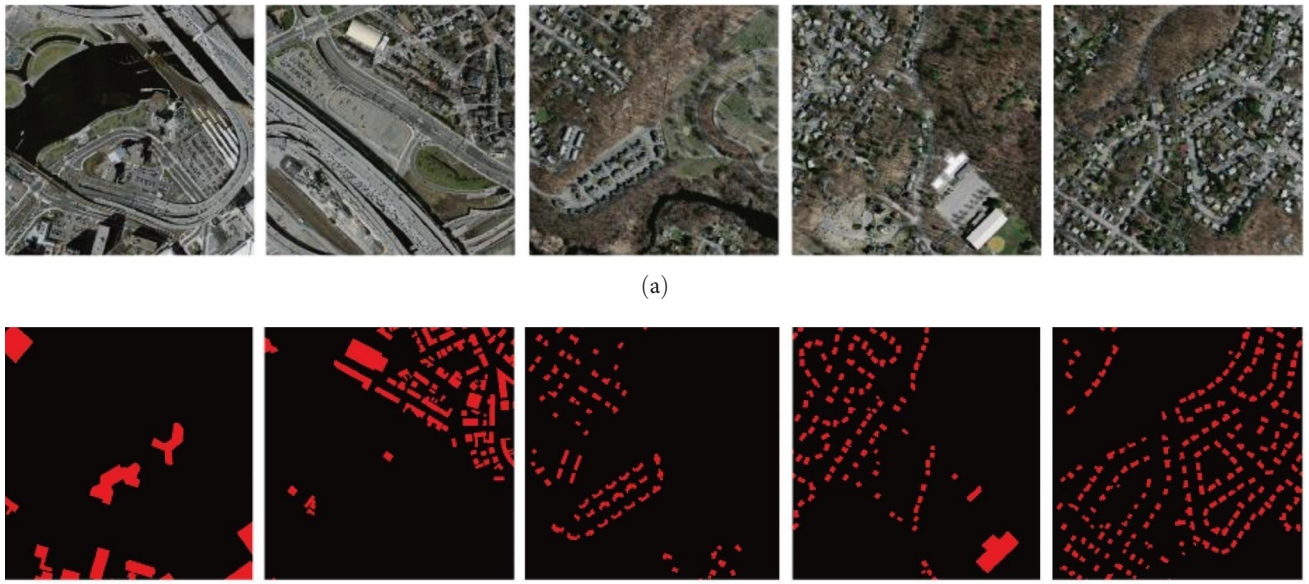
4.2. Dataset Preprocessing. Image enhancement can increase the amount of data and improve the generalization performance of the network. In this paper, data augmentation for sample datasets was carried out from the following aspects:

(1) To prevent the network model from overfitting, the sample datasets are subjected to data augmentation. The training samples in the above three datasets are rotated 90°, 180°, and 270° clockwise, flipped horizontally and flipped vertically (Figure 10).

(2) During the training process, a random value in the range of (0, 1) is randomly generated. When it is greater than 0.5, random Gaussian noise with variance in the range of (0, 2) is added. Meanwhile, random brightness transformation is performed to simulate images collected under different sunlight conditions. Data augmentation is performed through the above operations (Figure 11) to prevent overfitting of the network.

5. Experimental Results and Discussion

5.1. Network Training. The details of experimental environment and hyperparameters are as follows. We used a Windows



(a)

(b)

Figure 8: Images (a) and sample labels (b) of Massachusetts building dataset.

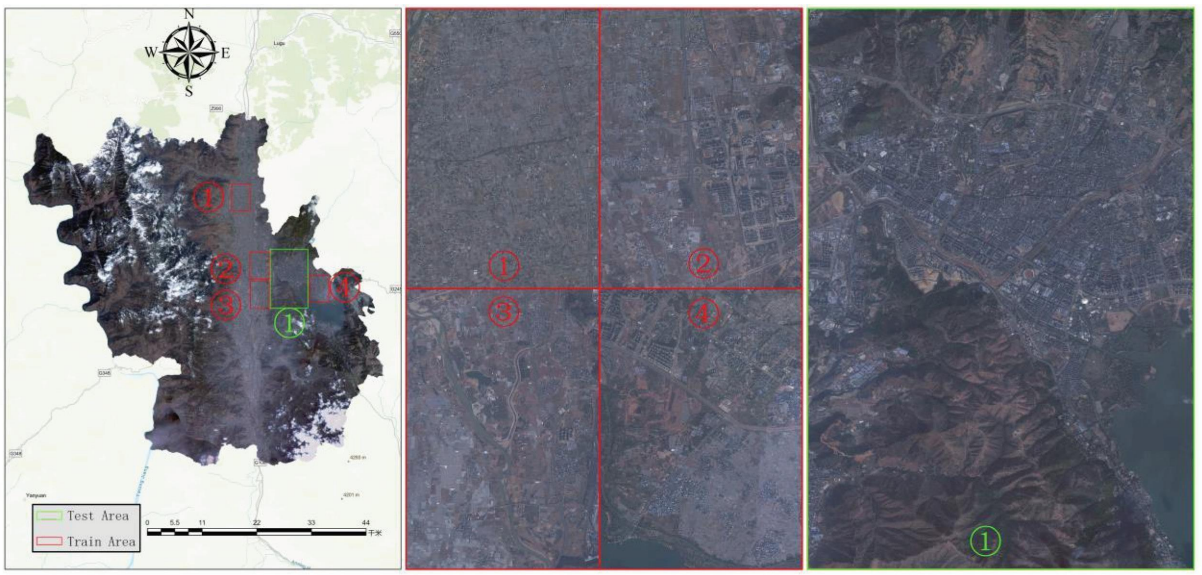


Figure 9: The samples selection on Xichang GF-2 image.

operating system with an RTX2080Ti GPU with 11 GB of video memory and a 16-core CPU, and chose the PyTorch framework to build the network. The optimizer is Adam, which is a momentum-based algorithm that uses the same learning rate for each parameter and reduces it adaptively as the network learns. We set the batch size to 2 due to computational resource constraints, epoch count is 50, initial learning rate is 0.001.

The training and validation loss curves for our network on the WHU dataset and the Massachusetts dataset are shown in Figure 12.

5.2. Precision Evaluation Index and Evaluation Strategy. Evaluation indexes are used to assess the performance strengths and weaknesses of the model in the semantic segmentation task. In this paper, after referring to relevant research results [16, 17], Accuracy (Acc), Recall (R), Precision (P), F1 score (F1), and intersection over union (IOU) are used to test the prediction ability of the network model. They are defined as follows in Equations (21)–(25).

$$\text{Accuracy} = \frac{\text{TP} + \text{FN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}, \tag{21}$$

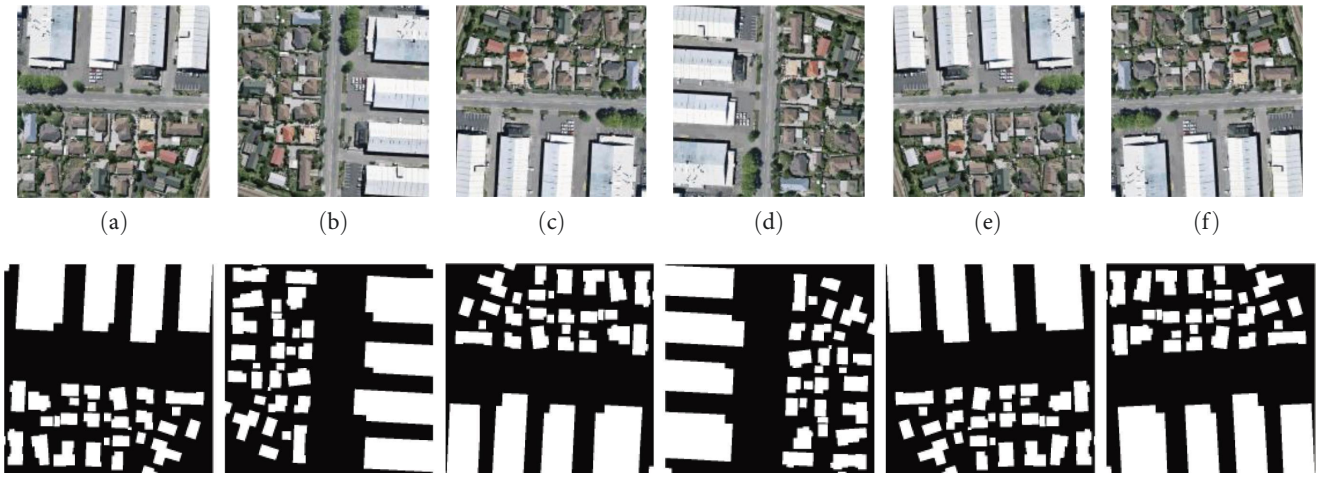


Figure 10: Image and label preprocessing. (a) Original image; (b) 90° clockwise rotate; (c) 180° clockwise rotate; (d) 270° clockwise rotate; (e) horizontal mirroring; (f) vertical mirroring; and (g) ground truth.

Figure 11: Image data enhancement: (a) original image; (b) with gaussian noise; and (c) with brightness shift.

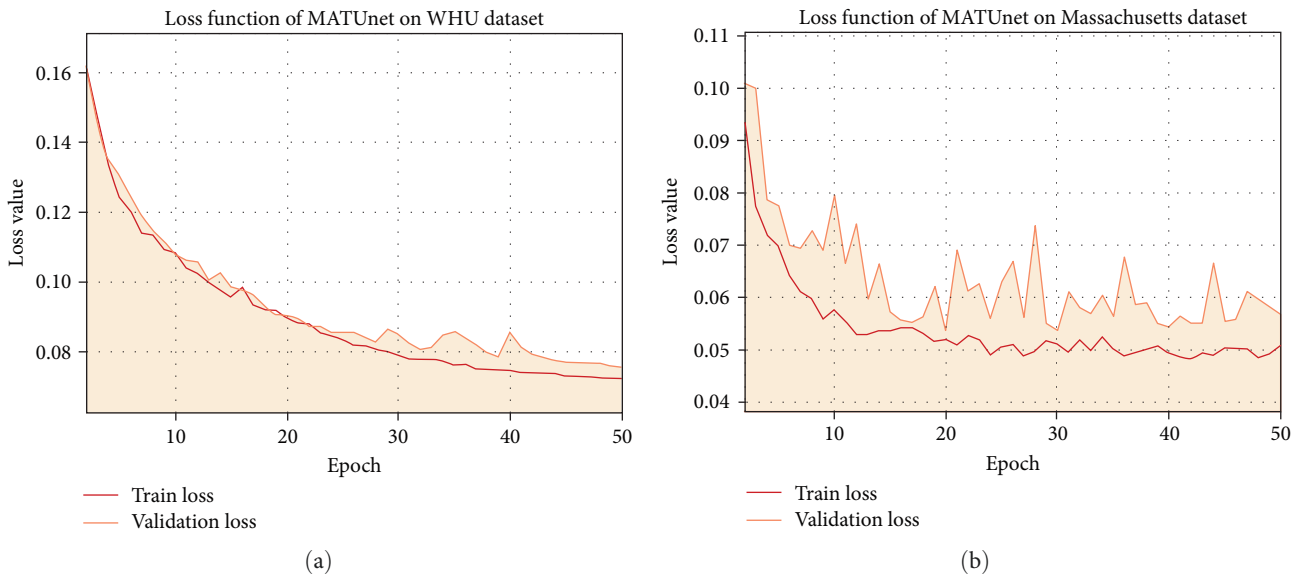


Figure 12: Loss curves of MATUnet: (a) WHU dataset and (b) Massachusetts dataset.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}, \tag{22}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{23}$$

$$\text{IOU} = \frac{\text{TP}}{(\text{TP} + \text{FP} + \text{FN})}, \tag{24}$$

$$\text{F1} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}. \tag{25}$$

TP is the number of samples labeled as building pixels while predicted as building pixels. FN is the number of samples labeled as background pixels while predicted as background pixels. FP is the number of samples labeled as background pixels whereas predicted as building pixels. TN is the number of samples labeled as building pixels whereas predicted as background pixels. Acc indicates the proportion of building pixels and background pixels that are correctly predicted to the predicted pixels and sample pixels. P is proportion of the correctly predicted building pixels to the predicted building pixels. R indicates the proportion of building pixels correctly predicted to the building sample pixels. IOU indicates the ratio of the intersection of the predicted building pixels and the building sample pixels to the union of the predicted building pixels and the building sample pixels. F1 Score is used to comprehensively evaluate the extraction results.

5.3. Building Extraction Results. To evaluate the effectiveness of the network model proposed in this paper, the classical semantic segmentation models Unet [3], Segnet [4], and the building extraction model of TransUnet were used as the baseline models for quantitative and qualitative evaluations on three different datasets. Meanwhile, to further demonstrate the advantages of our model, we also compared the evaluation indicators with the state-of-the-art building extraction methods MAP-Net [53], MSRF-Net [16], and TransFuse [54] on the two public baseline building datasets. MAP-Net uses three independent paths to combine different scale features in the encoding part. MSRF-Net is a block-level built-up area extraction framework combing densely connected dual-attention network and multiscale context, which used the designed DCDA-Net [55] for feature representation and discrimination of the image blocks. The proposed DCDA-Net is a lightweight network that combines dense connection and dual attention.

5.3.1. WHU Dataset.

(1) Quantitative Evaluation of Model Extraction Accuracy. The experiments were conducted on the WHU dataset, and the results of accuracy evaluation were obtained as shown in Table 1.

From the comparison of the indicators, it can be found that MATUnet is optimal in all metrics. *P* metric reaches 95.05%, which is an improvement of about 1.3% compared to the traditional TransUnet. IOU reaches 92.14%, which indicates that the MATUnet over TransUnet has resulted in the improved performance. In the latest method, compared with MAP-Net, the *P* metric of MATUnet improves by 1.26% and IOU improves by 2.74%. MAP-Net learns the spatial locations of multiscale features through multiple parallel paths while applying an attention-based approach to enhance the features. As reflected from the accuracy metrics, the combination of our multibranching strategy and the attention mechanism outperforms MAP-Net in terms of performance. Compared with the TransFuse model, our network improves the *P* metric by 0.88% and the IOU by 2.21%. TransFuse combines the transformer and CNN in parallel to capture global and spatially detailed features, but integrating the features extracted by both of them at a shallow level led to redundancy of extracted information, while our MATUnet fully utilizes the strengths of CNN and transformer to accurately extract local features and global features, and enhances the channel attention on the features during upsampling to improve the accuracy of the model.

As shown in Figure 13, by comparing the number of parameters and efficiency (flops) of different networks, we can find that the number of parameters of MATUnet network

Table 1: Evaluation indicators of different networks.

Method	P (%)	R (%)	F1 (%)	IOU (%)	Acc (%)
Segnet	95.01	74.82	83.71	81.88	93.20
Unet	93.71	81.93	87.42	85.43	94.51
MAP-Net	93.79	90.82	92.28	89.40	96.17
MSRF-Net ($k = 38$)	94.97	91.68	93.29	90.16	96.65
TransUnet	93.75	87.21	90.36	88.49	95.66
TransFuse-L	94.17	90.18	92.12	89.93	96.84
MATUnet	**95.05**	**92.23**	**93.62**	**92.14**	**97.06**

Note. Bold numbers indicate the best performance of each indicator.

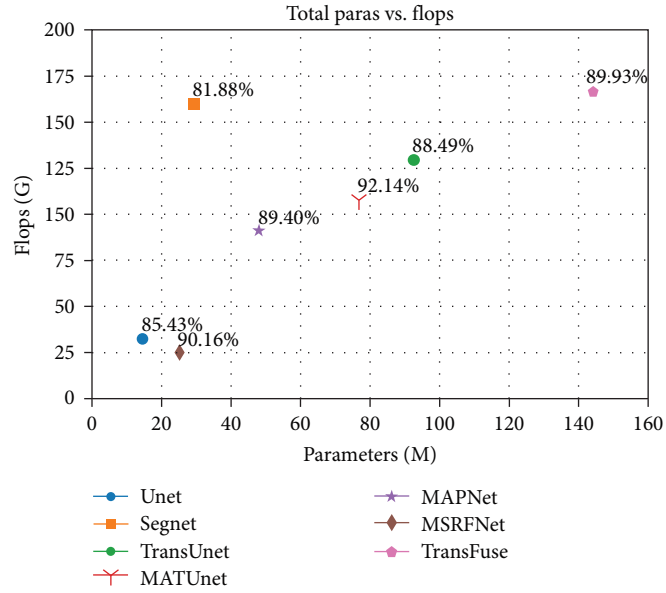


Figure 13: Comparison of parameters (M) and flops (G) of different network models (the marked percentage is IOU of the model).

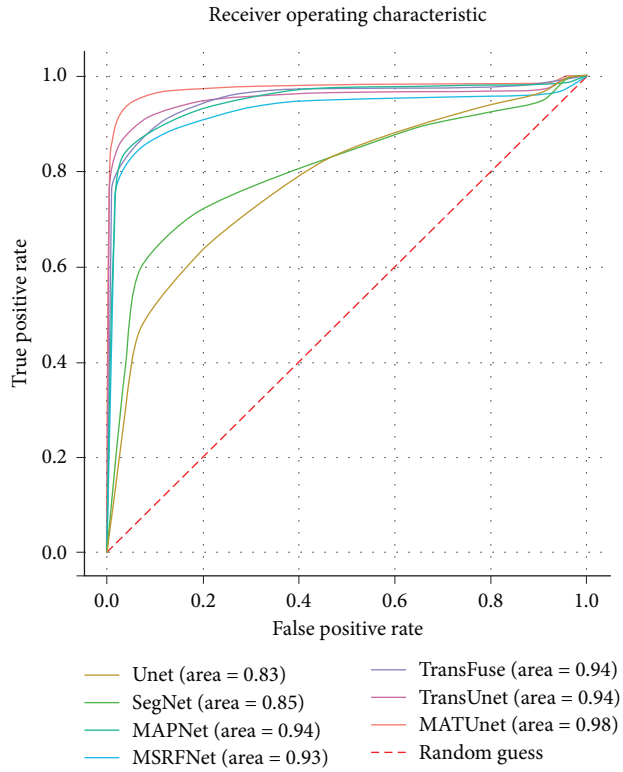


Figure 14: ROC on WHU dataset.

is larger compared with those of Unet and Segnet, but smaller than that of TransUnet, and the IOU of MATUnet is better than that of TransUnet.

Meanwhile, we plotted the receiver operating characteristic (ROC) curve of the model on the WHU dataset to judge the performance of the model. As shown in Figure 14, The ROC curve is obtained by changing the threshold of classification, which in turn yields a series of points, and then the obtained points are plotted as a curve according to the threshold from small to large. The horizontal coordinate of the curve represents the true positive rate (TPR), i.e., Recall, and the vertical coordinate represents the false positive rate (FPR). The area encircled by the curve is called the area under the curve (AUC), and a larger area of AUC indicates better performance.

(2) Qualitative Analysis of Model Extraction Results. To qualitatively compare the results of the classical models, the building recognition results were visualized and compared. The prediction result is overlaid with the labeled image, where white pixels represent the building that are correctly predicted by the network, red pixels represent the wrongly extracted building, and blue pixels represent the unextracted building, as shown from Figures 15–17.

By comparing the yellow boxes in the first row in Figure 15, we can find that the buildings of the images extracted from MATUnet network are more complete, the buildings have a lower miss detection rate in the small buildings extraction results. By comparing the yellow boxes in the second row, we can find that MATUnet is able to distinguish building pixels with high similarity and maintains the integrity of buildings. Meanwhile, MATUnet has a lower false alarm rate compared with TransUnet.

From the extracted results shown in Figure 16, we can find that MATUnet has better extraction results for large buildings, with more complete building boundaries and no combination phenomenon that occurs in other networks for building prediction. By comparing the yellow box area in the first row, we can find that Unet wrongly extracts nonbuilding objects, and Segnet and TransUnet also wrongly extract some nonbuilding objects. Compared to the other networks, MATUnet reduces this phenomenon and accurately distinguishes between building and nonbuilding objects. By comparing the yellow box area in the second row, we can find that for buildings surrounded by forests, MATUnet has achieved complete and accurate extraction of the area not covered by forests, which is a segmentation advantage different from the other general networks.

We can see that in the building extraction results of WHU dataset (Figure 17) with 0.45 m image resolution, MATUnet still obtains good extraction results and can keep the integrity of the buildings relative to the other network extraction results. By comparing the yellow area in the first row of Figures 17(c) and 17(f), we can find that Unet has missed some building pixels when buildings and backgrounds are similar in the extraction process. Segnet and TransUnet have extracted relatively few building pixels, whereas MATUnet can extract the complete buildings compared to other networks. By comparing the yellow area in the second row of Figure 17(d)–17(f), we can find that TransUnet incorrectly extracts the nonbuilding objects, whereas Segnet and MATUnet have correctly extracted buildings. By comparing the yellow area in the third row, we can find that those are relatively dense and large buildings. Although MATUnet can extract relatively complete buildings compared with other networks, it does not distinguish the buildings when the buildings are close together. It is due to the distance between buildings is too short, resulting in the incorrect extraction of some pixels.

5.3.2. Massachusetts Dataset. To further validate the building extraction capability of the network model, the Massachusetts dataset is also used in experiment. The Massachusetts dataset is adopted the same enhancement method in this paper.

(1) Quantitative Evaluation of Model Extraction Accuracy. By analyzing the prediction accuracy results of each network in the Massachusetts dataset (Table 2), we can see that in this dataset, the metrics of MATUnet are significantly better than those of the other classical network models, which further suggests that MATUnet has a better building extraction performance. In the latest network, we can find that the P metrics and IOU metrics of TransFuse and TransUnet are lower than some convolution-based networks, which due to the smaller number of Massachusetts dataset, it is more difficult for TransFuse and TransUnet with multiple transformer structures to converge, which leads to the two models have a lower accuracy on this dataset. Whereas our network reconsiders the number of transformer layers after adding convolutional positional coding, which reduces

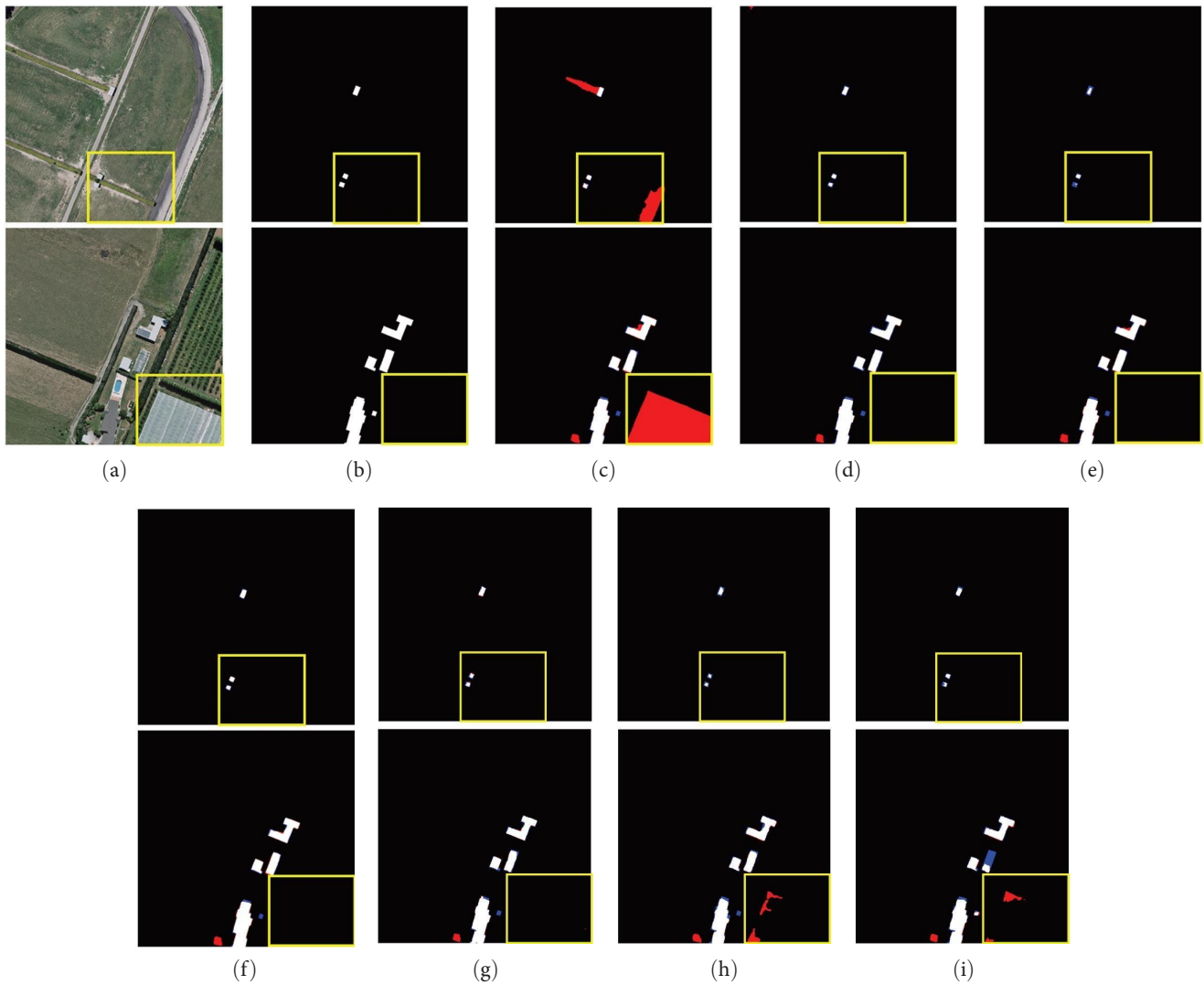


Figure 15: Extraction results of small buildings (0.3 m resolution): (a) original image; (b) ground truth; (c) Unet; (d) Segnet; (e) TransUnet; (f) MATUnet; (g) MAPNet; (h) MSRFNet; and (i) TransFuse.

the number of references while making the network easier to converge, and therefore has higher accuracy.

Similarly, we plotted the ROC curve of the model on the Massachusetts dataset (Figure 18) to determine the performance of the model, and the AUC area of our network in Figure 18 indicates that our method has a good performance.

(2) Qualitative Analysis of Model Extraction Results. As can be seen in Figure 19, MATUnet outperforms other network models in building extraction on this dataset. From the first row and fourth row images, we find that MATUnet can extract more building pixels correctly in white color and less wrongly and missed building pixels in red and blue color, respectively. From the fifth-row images, we can find that Unet, Segnet, and TransUnet have poor ability in the extraction of large buildings, whereas MATUnet shows excellent extraction performance and keeps the integrity and accuracy of buildings. From the third-row images, we can find that for buildings with complex shapes, although MATUnet extracts better integrity of buildings compared to the other networks, there are still incorrect extractions due to the presence of shadows between the buildings, and the network model is not able to distinguish the boundaries of buildings due to the small spacing.

5.3.3. Generalization Ability Assessment. To verify the feature extraction effect of the model proposed in this paper in scene transferring applications, the prediction of buildings is performed for the GF-2 Xichang study area. The images are cropped to 512×512 nonoverlapping image blocks. The 512×512 resulting map is obtained after network model prediction, and then the resulting map is merged into a raster map with geospatial location information using Python and GDAL open-source library. We chose three networks, Unet, Segnet, and TransUnet, which have been widely applied in the practical scenarios, to compare with our MATUnet. All four networks were tested for prediction in the designated areas, respectively, and the results are displayed as shown in Figure 20.

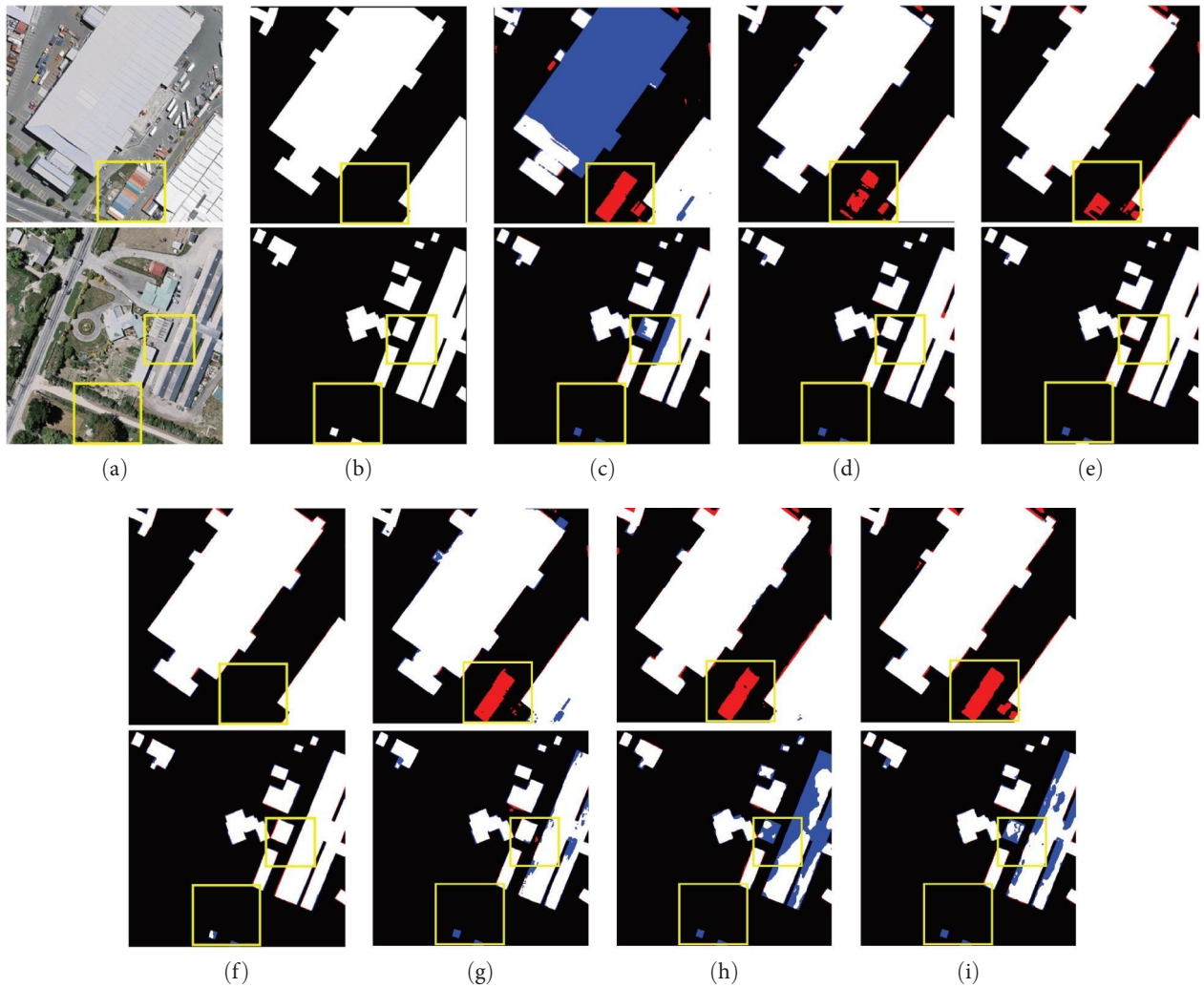


Figure 16: Extraction results of large buildings (0.3 m resolution): (a) original image; (b) ground truth; (c) Unet; (d) Segnet; (e) TransUnet; (f) MATUnet; (g) MAPNet; (h) MSRFNet; and (i) TransFuse.

As can be seen in Figure 20, MATUnet extracts more building areas and fewer wrongly extracted buildings compared with Unet, Segnet, and TransUnet, which shows that MATUnet has more excellent extraction performance. From the local images of prediction results (Figure 20(b)), we can find that MATUnet can identify more complete and regular building boundaries compared with the other networks. To a certain extent, it shows that MATUnet has better generalization ability than the other general models.

5.3.4. Analysis of Ablation Studies to Model Structure. To explore the impact of the three modules improved in this paper on the feature extraction performance of the model, the following ablation experiments were conducted on the WHU dataset by means of control variables:

(1) Effect of MGM on the Network. To verify the ability of MGM, we designed two networks. The first one is to replace the convolution module in TransUnet with MGM, named TransUnet + MGM, this network is to verify whether there is any improvement in the performance of TransUnet after adding MGM. The second network is to replace the MGM in MATUnet with the standard convolutional module, named MATUnet-Group, which is to verify whether there is any performance degradation of the network after removing the MGM. Our prediction results for MATUnet, MATUnet-MGM, TransUnet, and TransUnet + MGM are shown in Table 3.

By comparing the prediction accuracies of TransUnet and TransUnet + MGM, we can find that after adding MGM, the *P*-accuracy of TransUnet + MGM is 2.37% higher than that of conventional TransUnet, and the IOU is higher than 1.21%. By comparing the prediction accuracies of MATUnet and MATUnet-MGM (without MGM), we can find that after removing MGM, the *P*-accuracy of MATUnet-MGM is 0.08% lower than that of MATUnet, and the IOU is 1.67% lower than MATUnet, which shows the advantage of MGM in improving the model accuracy. Meanwhile, by comparing the number of parameters of the network model between MATUnet-MGM and MATUnet, we can find that MGM does not make the parameters of the network not increase significantly, which proves the effectiveness of MGM.

The feature maps of the standard convolutional layer in TransUnet (corresponding to three blocks) and the MGM in

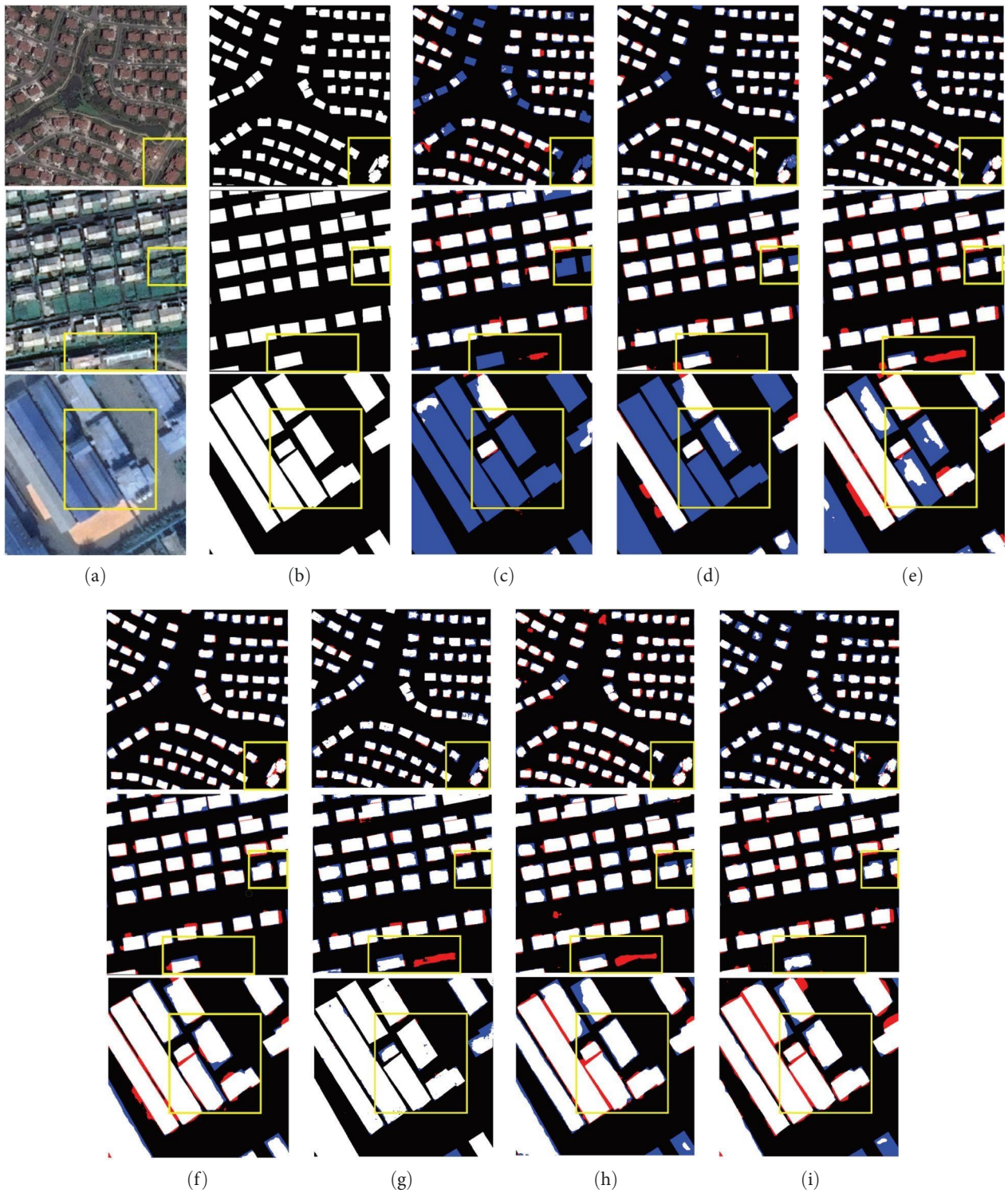


Figure 17: Extraction results of dense buildings (0.45 m resolution): (a) original image; (b) ground truth; (c) Unet; (d) Segnet; (e) TransUnet; (f) MATUnet; (g) MAPNet; (h) MSRFNet; and (i) TransFuse.

MATUnet were visualized. Specifically, we output the feature maps after 1×1 convolution, 3×3 convolution, and 1×1 convolution in the standard convolutional had and MGM, and the results are shown in Figure 21(a)–21(f), where Figure 21(a)–21(c) are the outputs of the standard convolutional layer at three scales and Figure 21(d)–21(f) are the outputs of the MGM. The color in the figure represents the feature value. The brighter the color, the higher the feature value.

In Figure 21(a), 21(b), 21(d), and 21(e) all have nine convolutional feature maps, Figures 21(c) and 21(f) have 27 convolutional feature maps. By comparing the feature maps, we

Table 2: Evaluation indicators of different networks.

Method	P (%)	R (%)	F1 (%)	IOU (%)	Acc (%)
Unet	82.28	69.31	75.24	74.37	90.16
Segnet	78.82	65.04	71.27	72.13	90.24
MAP-Net	85.94	79.80	82.76	70.59	—
MSRF-Net ($k = 32$)	86.53	80.88	83.61	71.84	93.09
TransUnet	80.80	71.95	76.11	74.96	90.26
TransFuse-L	82.54	78.97	80.48	72.02	92.76
MATUnet	**89.17**	**81.22**	**85.01**	**83.22**	**93.82**

Note. Bold numbers indicate the best performance of each indicator, and "–" indicates the network did not use this indicator in the original article.

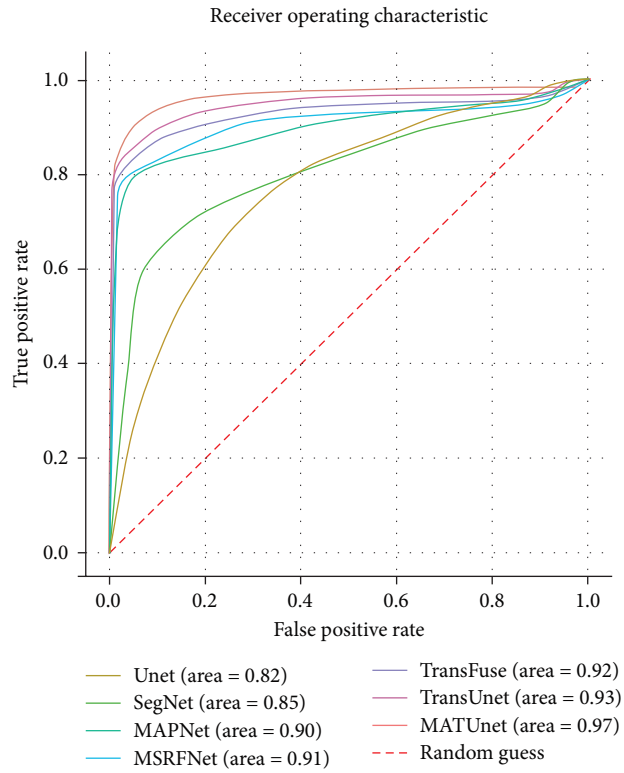


Figure 18: ROC on Massachusetts dataset.

can find that the semantic information (e.g., texture, etc.) obtained from the network containing the standard convolution gradually decreases as the network layer gets deeper, and the representation of features tends to be categorized, with the semantic information becoming more abstract. However, the features output from MGM end up with abstract semantic information while retaining some localized detailed semantic information, which helps the model to recover the detailed information of the features by sampling on the encoder. This indicates to some extent that our MGM has a better ability to capture architectural object information than the standard convolution, which helps to improve the accuracy of the building feature extraction.

In contrast to increasing the depth and width of the network, group convolution improves the channel local correlation of building features by increasing the number of groups, which improves the building feature extraction capability of the network without significantly increasing the number of parameters.

(2) Effect of Transformer with PEG on the Network. To verify the impact of transformer with PEG on the model feature extraction performance, this paper conducts experiments from two aspects.

First, the number of transformer layers in the transformer with PEG is discussed and analyzed in this paper. In previous studies, some scholars [33] explored the influence of transformer layers on the performance of network feature extraction in the field of remote sensing semantic segmentation. We also set different layers of transformer structure for building extraction from remote sensing images to explore the feature extraction effect of different transformer layers in the remote sensing semantic segmentation task. The

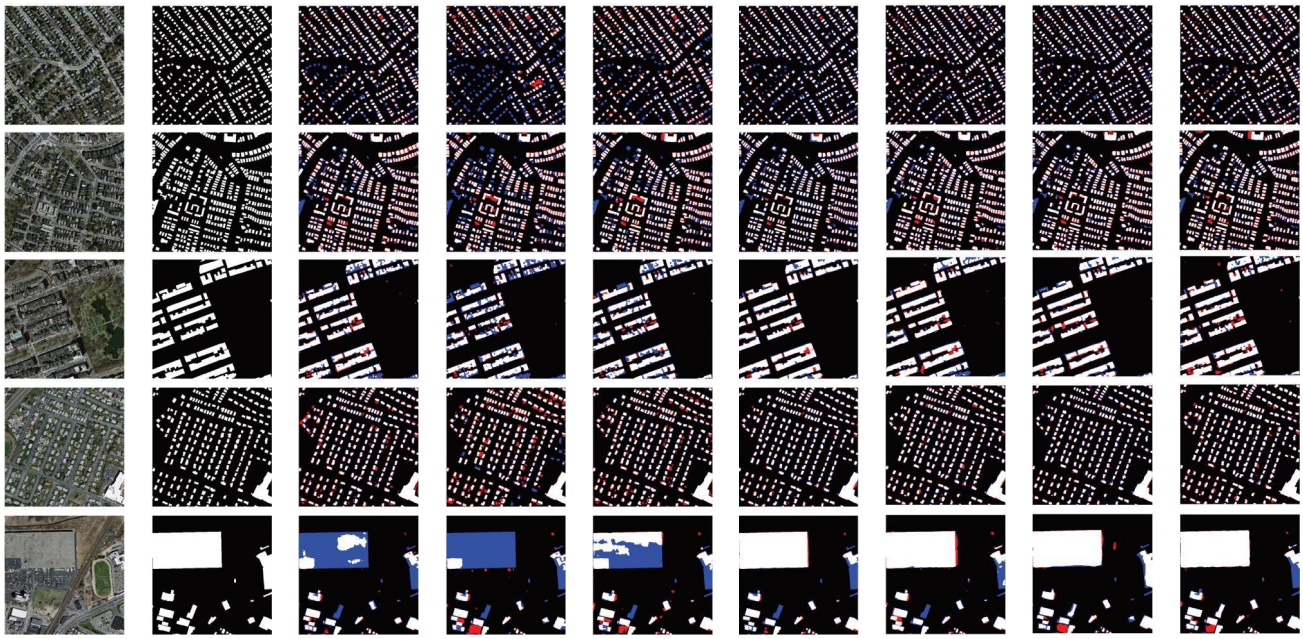


Figure 19: Building extraction results of different networks: (a) original image; (b) ground truth; (c) Unet; (d) Segnet; (e) TransUnet; (f) MATUnet; (g) MAPNet; (h) MSRFNet; and (i) TransFuse.

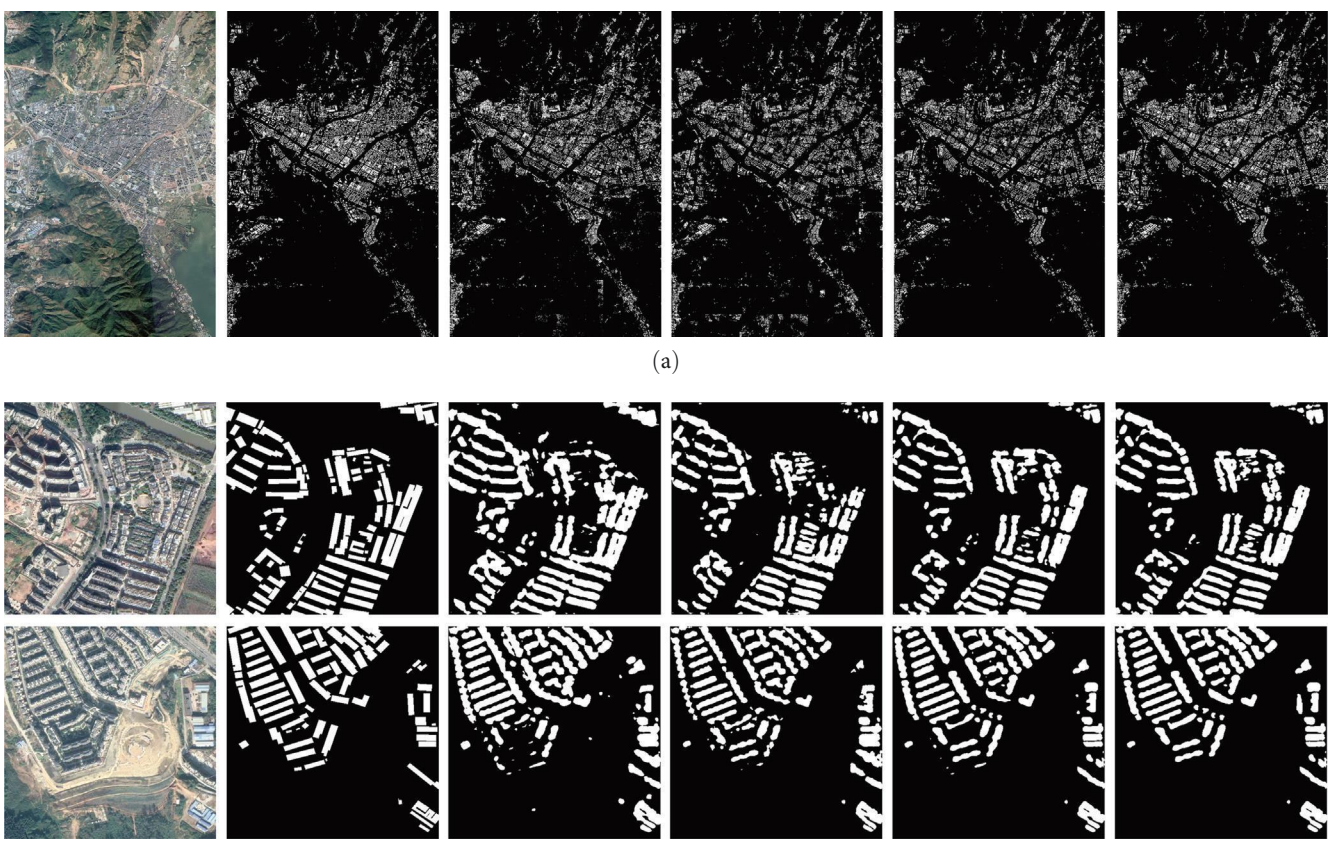
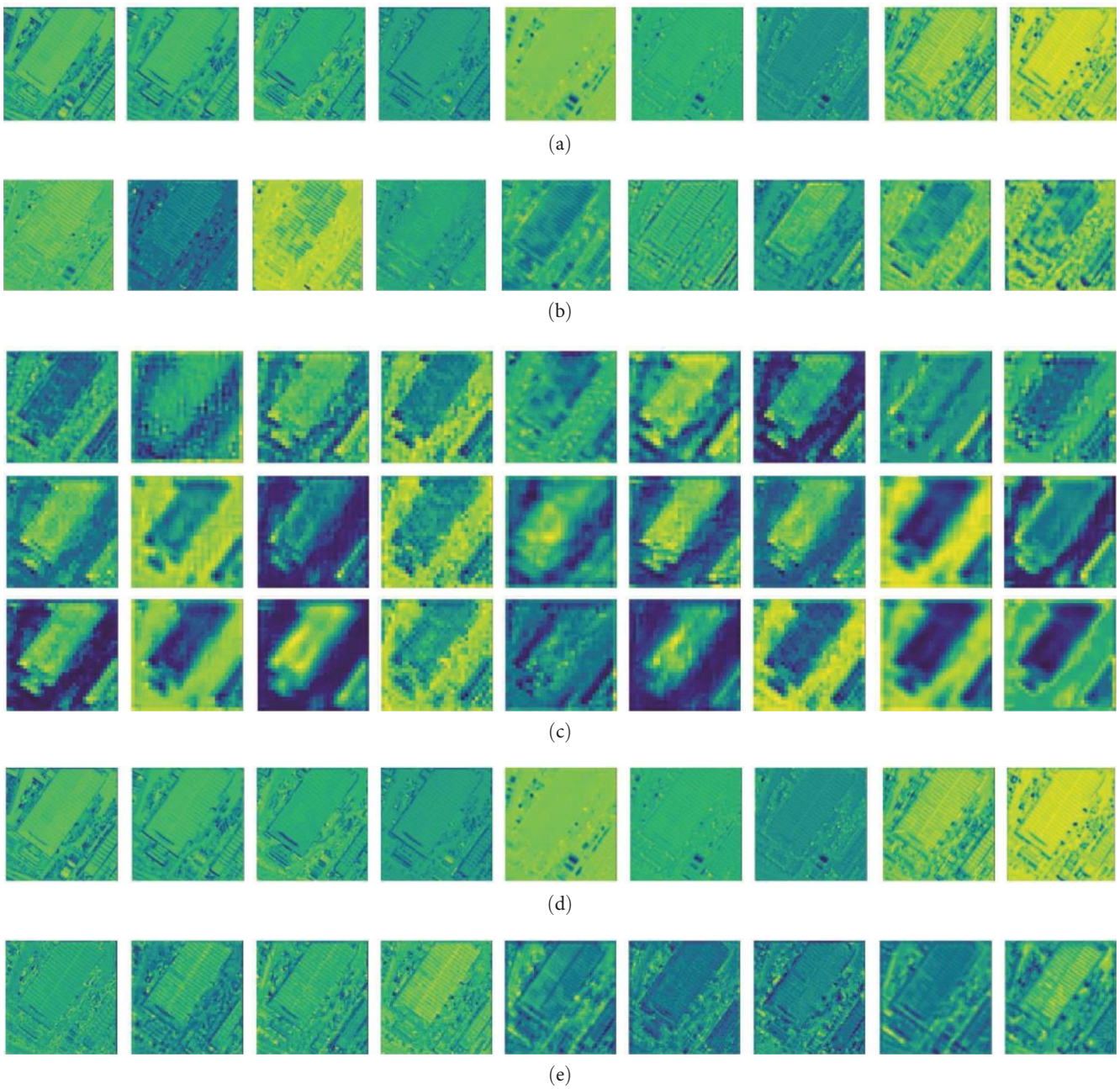


Figure 20: Building extraction results of different networks: (a) overall building extraction results and (b) local extraction results. In sequence, the columns represent the original image, label, Unet, Segnet, TransUnet, and our network, MATUnet.

Table 3: Evaluation indicators of different networks.

Method	P (%)	R (%)	F1 (%)	IOU (%)	Acc (%)	Paras (M)
MATUnet	95.05	92.23	93.62	92.14	97.06	64.98
MATUnet-MGM	94.97	90.56	92.71	91.16	96.70	64.98
TransUnet	93.75	87.21	90.36	88.49	95.66	93.23
TransUnet + MGM	96.12	87.38	91.57	89.78	96.12	93.23



(a)

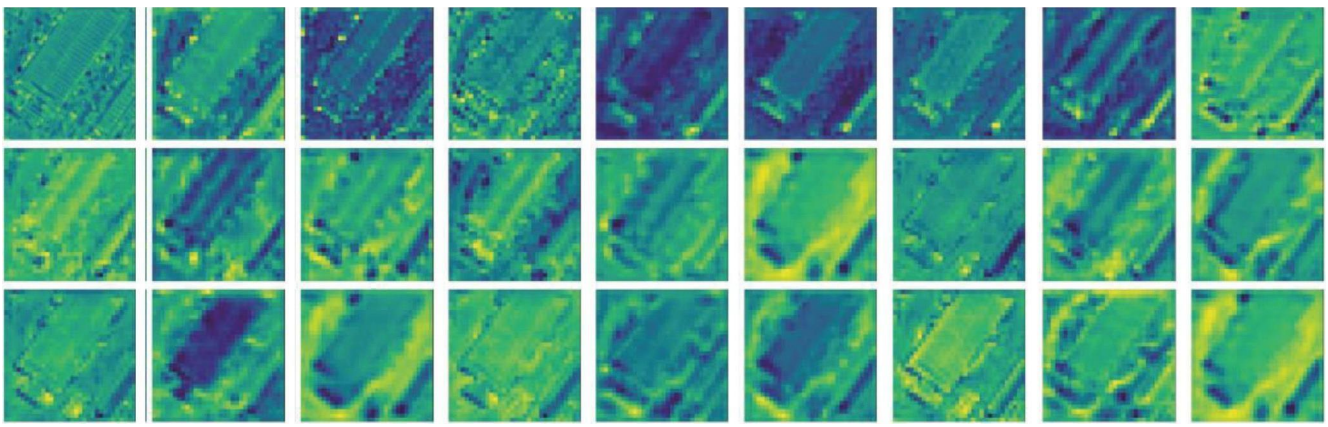
(b)

(c)

(d)

(e)

Figure 21: Continued.



(f)

Figure 21: Feature visualization of standard convolution features ((a) Block 1, (b) Block 2, and (c) Block 3) and MGM features ((d) Block 1, (e) Block 2, and (f) Block 3).

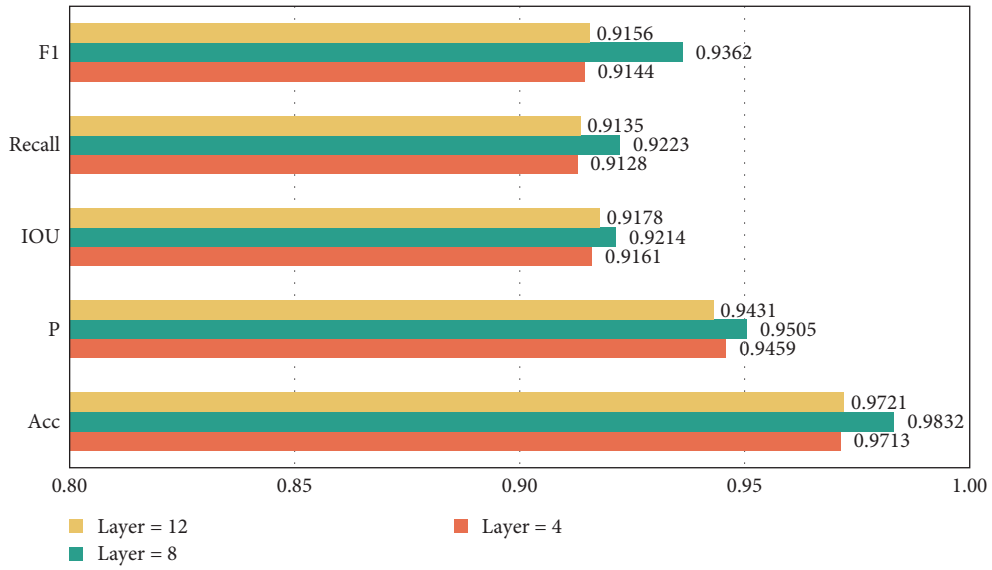


Figure 22: Comparison of transformer structure with different layers.

numbers of transformer groups in different networks in this paper are designed to be 4, 8 (8 layers for MATUnet), and 12, respectively, which are named MATUnet-4, MATUnet, and MATUnet-12. The above networks were trained and predicted in the WHU building dataset, and the segmentation accuracy results of each network were obtained as shown in Figure 22. MATUnet network has the highest extraction accuracy. The extraction accuracy of MATUnet-12 network is higher than that of MATUnet-4. MATUnet has reduced the number of layers compared with MATUnet-12, and the feature extraction accuracy has been further improved. Comparing the numbers of parameters and flops of three network models in Figure 23, we can see that although the number of parameters of MATUnet-12 is larger than that of MATUnet, the accuracy index of MATUnet-12 is not better than that of MATUnet. Therefore, the model of 8-layer transformer structure is selected as MATUnet.

Second, we conducted a comparative experiment on the position encoding methods in the transformer structure through designing two networks, respectively. One network is TransUnet + PEG, and the other network is MATUnet-PEG. TransUnet + PEG is to add the PEG module to TransUnet and delete the original position encoding method. By comparing the TransUnet and TransUnet + PEG networks, we can show that adding the PEG module will improve the accuracy of the network. MATUnet-PEG is to delete the PEG module based on our MATUnet and adopts the position encoding method of the traditional TransUnet. By comparing MATUnet and MATUnet-PEG networks, we can show that the PEG module contributes to the high performance of MATUnet, and the PEG module is necessary. We conduct the experiments on the WHU dataset using TransUnet, TransUnet + PEG, MATUnet, MATUnet-PEG to compare the indicators. The extraction accuracies of four networks are shown in Table 4.

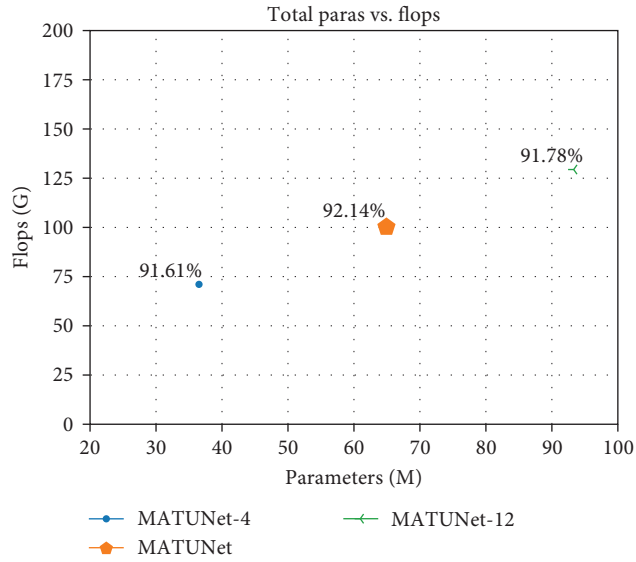


Figure 23: Comparison of parameters (M) and flops (G) of different network models (the marked percentage is IOU of the model).

Table 4: The impact of location encoding on the networks.

Method	P (%)	R (%)	F1 (%)	IOU (%)	Acc (%)
TransUnet	93.75	87.21	90.36	88.49	95.66
TransUnet + PEG	95.03	96.67	90.66	88.84	95.83
MATUnet-PEG	94.61	91.62	93.08	91.52	96.82
MATUnet	95.05	92.23	93.62	92.14	97.06

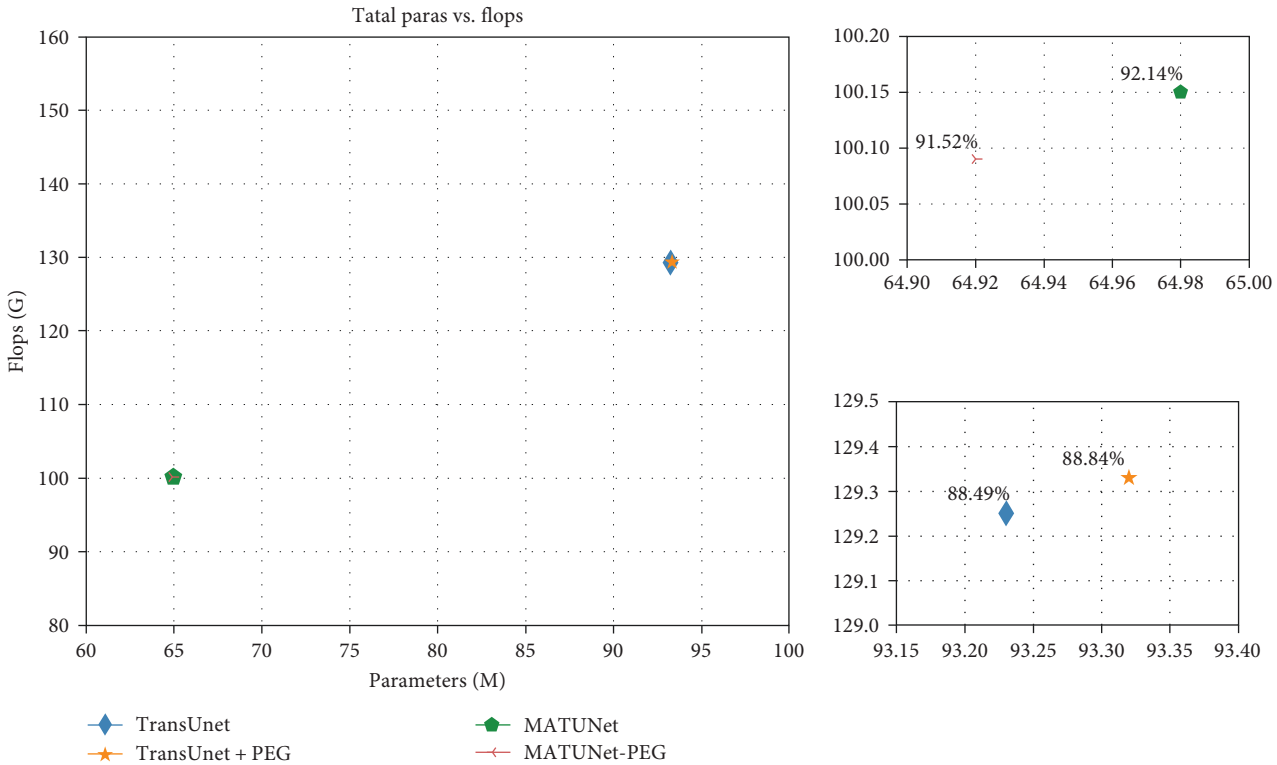


Figure 24: Comparison of parameters (M) and flops (G) of different network models (the marked percentage is IOU of the model).

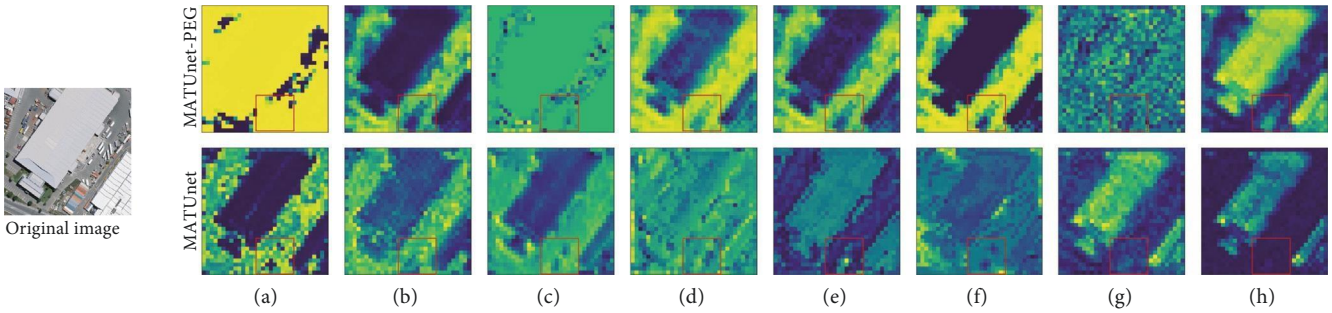


Figure 25: Features visualization output from multi-headed attention modules under two position coding methods. (a)–(h): Encoder 1–Encoder 8.

Table 5: The impact of CAM on network performance.

Method	P (%)	R (%)	F1 (%)	IOU (%)	Acc (%)
TransUnet	93.75	87.21	90.36	88.49	95.66
TransUnet + CAM	93.88	91.06	92.45	90.78	96.53
MATUnet	95.05	92.23	93.62	92.14	97.06
MATUnet-CAM	94.91	91.75	93.30	91.77	96.92

In Table 4, comparing the extraction accuracies of TransUnet and TransUnet + PEG, we can find that TransUnet + PEG have higher accuracy indicators than those of TransUnet. P has improved by 1.28% compared to the corresponding indicator of TransUnet. It indicates that the addition of the PEG module can help the network model to extract more accurate building pixels. Similarly, by comparing the evaluation indicators of the MATUnet and MATUnet-PEG networks, we can find that the MATUnet with the PEG module has a certain improvement in feature extraction evaluation indicators compared with the MATUnet-PEG without the PEG module. Comparing the numbers of parameters and Flops of four network models in Figure 24, we can see that the number of parameters of MATUnet is increasing, and the validation indicators of the model are also better.

At the same time, extracted features are visualized after the multi-head attention layer calculation in the MATUnet and MATUnet-PEG networks. The highlighted color in the feature map denotes a network with high-corresponding values, while the dark black color denotes a network with low-response values. We can find the changes in features extracted before and after improvement. The results are shown in Figure 25.

From Figure 25, we can find that MATUnet can extract clearer building edge features than MATUnet-PEG. MATUnet makes the difference between the features of building and features of the background more obvious. In the final output feature map (h), MATUnet can determine nonbuilding objects as background, whereas MATUnet-PEG misidentifies nonbuildings as buildings. The above results show that PEG module not only provides an implicit encoding method through convolution calculation, but also compensates for the local semantic loss due to interpolation. The PEG module has the multi-head attention mechanism to fuze local and global semantic information of image, which improves the extraction accuracy of building features.

(3) Effect of the CAM on the Network. To verify the effect of CAM on feature extraction performance, we designed two networks. One is TransUnet with CAM, which is named TransUnet + CAM, and the other is MATUnet without CAM, which is named MATUnet-CAM. By comparing the accuracy indicators of TransUnet and TransUnet + CAM, we explore whether the CAM module is effective in improving network performance. Meanwhile, we hope that through the comparison of MATUnet and MATUnet-CAM networks, we could find that the contribution of CAM module to the high performance of MATUnet. And the CAM module is necessary. Similarly, the above networks were used for the building extraction experiment on the WHU dataset, and the results of feature extraction accuracy are shown in Table 5.

From Table 5, we find that the evaluation indicators of TransUnet + CAM are higher than those of TransUnet after adding channel attention. Similarly, the feature evaluation indicators of MATUnet with CAM module are improved compared with MATUnet-CAM which does not include the CAM module. Comparing the numbers of parameters and flops of four network models in Figure 26, we can see that the number of parameters of MATUnet has increased, and the validation indicators of the model are also better. To a certain extent, it shows the effectiveness of the channel attention enhancement module in improving the ability of network feature extraction.

At the same time, to intuitively reflect the effect of the channel attention enhancement module, the output features of convolutional layer before and after adding the CAM are visualized. The Grad-Cam heat map of the features in this layer before and after enhancement was obtained by calculating the product of the gradient of the backward propagation and the feature map of the network in this layer. The calculated results are shown in Figure 27. The color in the

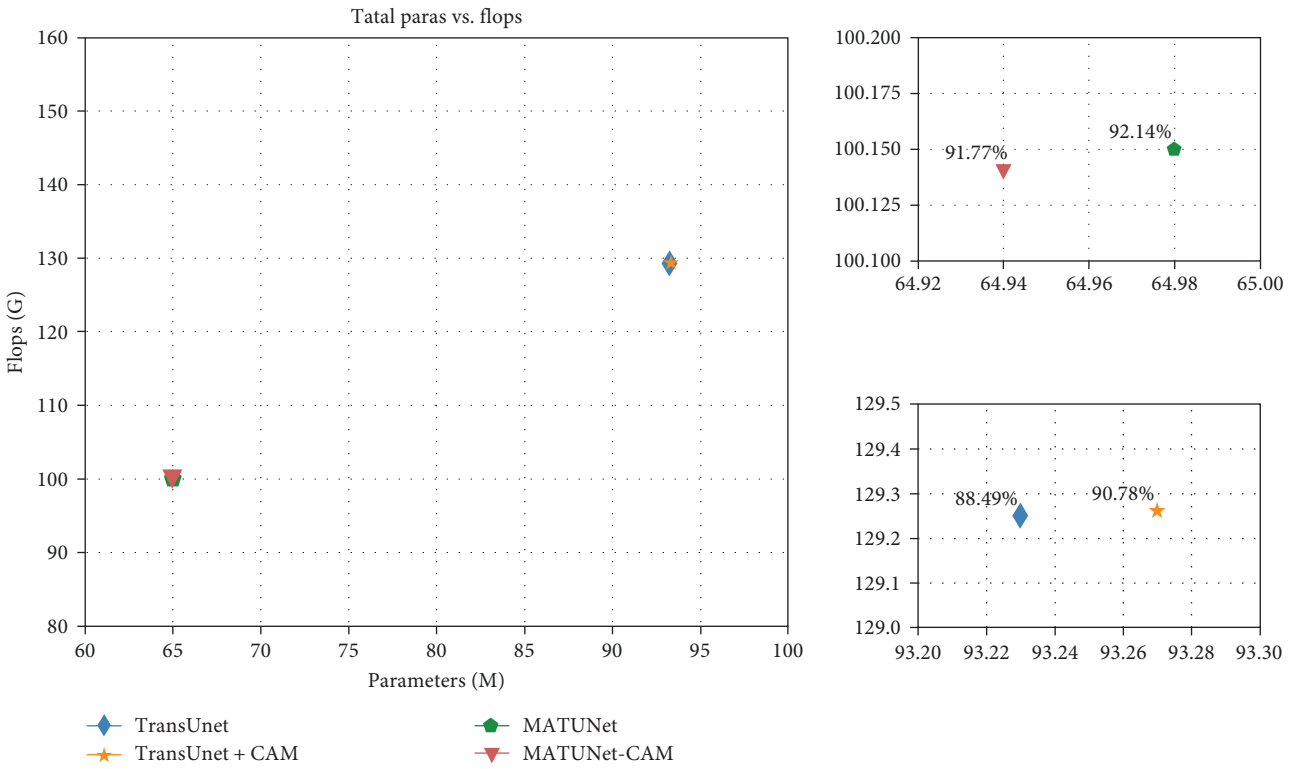


Figure 26: Comparison of parameters (M) and flops (G) of different network models (the marked percentage is IOU of the model).

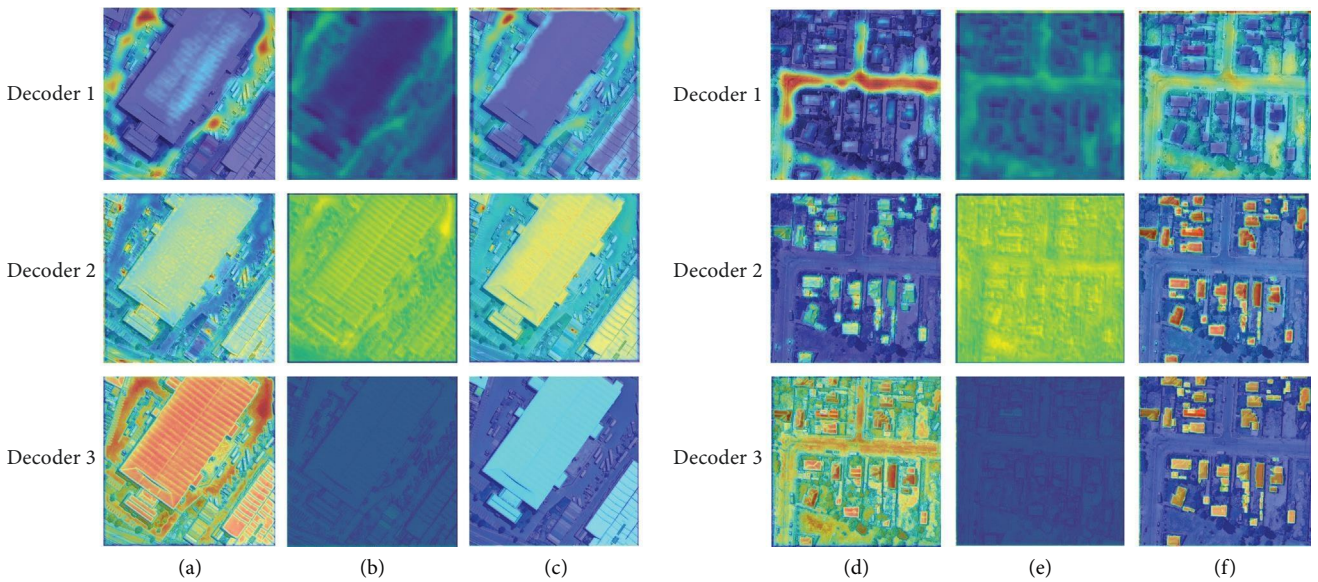


Figure 27: Effect of feature enhancement module on model segmentation: (a) heat map of large buildings before enhancement; (b) feature map of large buildings after enhancement; (c) heat map of large buildings after enhancement; (d) heat map of dense buildings before enhancement; (e) feature map of dense buildings after enhancement; and (f) heat map of dense buildings after enhancement.

figure represents the correlation of the network to the area. The brighter the color, the higher the correlation.

As can be seen in Figure 27, after calculation by Decoder 1, the heat map shows that the background information has a high value, which means that after adding the CAM, the attention of the network pays more attention to the background information at this layer, which shows that the enhanced image background features are more obvious. After calculation by Decoder 2 and Decoder 3, the attention of the network pays more attention to building information at these layers, and the building features are more obvious, which indicates that the network has higher attention to building features at these layers. We can find from the above analysis that after adding the channel attention module, the

ability of the network to extract building features has been greatly improved.

6. Conclusion

In this paper, we propose an improved TransUnet model, MATUnet, based on multiscale grouped convolution and attention to preserve more local detailed features and enhance the representation of global features while reducing network parameters. We designed the multiscale grouped convolutional feature extraction module (GAM) with attention to enhance the representation of detailed features. A convolutional PEG is added to redetermine the number of transformers, which solves the problems of loss of local feature information and network convergence difficulties. CAM of the decoder enhances the salient information of the features and solves the problem of information redundancy after feature fusion. The experimental results show that the network has significant accuracy improvement and good application prospects compared with other ordinary networks. Further research will be carried out in the future on the lightweight and efficient processing of the model and the application of engineering deployment, to solve the problems of the transformer structure relying on a large amount of training data and the redundancy of model parameters.

Although, MATUnet achieves better results in building extraction, there are still some limitations: (1) The samples of MATUnet come from semantically segmented labels, and labels need to be input manually, which makes MATUnet have a larger sample collection cost, (2) Transformer in MATUnet still needs to compute attention on the whole graph, which is different from the convolution-based models, and (3) the recognition effect for dense buildings in mountainous areas needs to be improved. Based on the above problems, further research will be carried out on the lightweight and efficient processing of the model as well as engineering deployment applications in the future, to solve the problem that the transformer structure relies on a large amount of training data and the redundancy of model parameters. We expect Transformer-based lightweight networks to be integrated on UAV hardware or satellite sensor devices to improve the real-time of remote sensing semantic segmentation tasks.

Data Availability

The WHU building dataset used to support the findings of this study have been deposited in website (http://gpcv.whu.edu.cn/data/building_dataset.html).The Massachusetts dataset used to support the findings of this study have been deposited in website (https://www.cs.toronto.edu/~vmnih/data/).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant 42271090, the National High-Resolution Earth Observation Major Project under grant 31-Y30F09-9001-20/22, and the Fundamental Research Funds of the Institute of Earthquake Forecasting, CEA under grant numbers CEAIEF2022050504 and CEAIEF20230202.

References

[1] M. Uzar, "Automatic building extraction with multi-sensor data using rule-based classification," *European Journal of Remote Sensing*, vol. 47, no. 1, pp. 1–18, 2014.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, IEEE, Boston, MA, USA, Jun 2015.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, pp. 234–241, 2015.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: a deep convolutional encoder–decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[5] S.-J. Xu, P.-Y. Ouyang, X.-Y. Guo, M. Khan Taha, and Z.-X. Duan, "Building segmentation in remote sensing image based on multiscale-feature fusion dilated convolution resnet," *Optics and Precision Engineering*, vol. 28, no. 7, pp. 1588–1599, 2020.

[6] J. Wang, Q. Qin, X. Ye, J. Wang, X. Qin, and X. Yang, "A survey of building extraction methods from optical high resolution remote," *Sensing Imagery*, vol. 31, no. 4, pp. 653–662+701, 2016.

[7] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[8] C. Li, F. Huang, W. Hu, and J. Zeng, "Building extraction from high-resolution remote sensing image based on Res_AttentionUnet," *Journal of Geo-Information Science*, vol. 23, no. 12, pp. 2232–2243, 2022.

[9] Y. Wang, Q. Zhao, Y. Wu, W. Tian, and G. Zhang, "SCA-Net: multiscale contextual information network for building extraction based on high-resolution remote sensing images," *Remote Sensing*, vol. 15, no. 18, Article ID 4466, 2023.

[10] Y. Sun, J. Chen, X. Huang, and H. Zhang, "Multi-level perceptual network for urban building extraction from high-resolution remote sensing images," *Photogrammetric Engineering & Remote Sensing*, vol. 89, no. 7, pp. 427–434 8, 2023.

[11] S. Chan, Y. Wang, Y. Lei, X. Cheng, Z. Chen, and W. Wu, "Asymmetric cascade fusion network for building extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023.

[12] X. Dai, M. Xia, L. Weng, K. Hu, H. Lin, and M. Qian, "Multiscale location attention network for building and water segmentation of remote sensing image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–19, 2023.

[13] Y. Sun, F. Bi, Y. Gao, L. Chen, and S. Feng, "A multi-attention UNet for semantic segmentation in remote sensing images," *Symmetry*, vol. 14, no. 5, Article ID 906, 2022.

[14] Z. Che, L. Shen, L. Huo et al., "MAFF-HRNet: multi-attention feature fusion HRNet for building segmentation in remote sensing images," *Remote Sensing*, vol. 15, no. 5, Article ID 1382, 2023.

[15] J. Wang, K. Sun, T. Cheng et al., "Deep high-resolution representation learning for visual recognition," *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, vol. 43, no. 10, pp. 3349–3364, 2021.

[16] Y. Zhao, G. Sun, L. Zhang, A. Zhang, X. Jia, and Z. Han, "MSRF-Net: multiscale receptive field network for building detection from remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[17] T. Yu, P. Tang, B. Zhao et al., "ConvBNet: a convolutional network for building footprint extraction," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[18] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966–11976, IEEE, New Orleans, LA, USA, Jun 2022.

[19] X. Shi, H. Huang, C. Pu, Y. Yang, and J. Xue, "CSA-UNet: channel-spatial attention-based encoder–-decoder network for rural blue-roofed building extraction from UAV imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[20] Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Al Dayil, and N. Al Ajlan, "Vision transformers for remote sensing image classification," *Remote Sensing*, vol. 13, no. 3, Article ID 516, 2021.

[21] K. Han, Y. Wang, H. Chen et al., "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2023.

[22] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems, in NIPS'17*, pp. 6000–6010, Curran Associates Inc., Red Hook, NY, USA, December 2017.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16 x 16 words: transformers for image recognition at scale," 2021.

[24] Z. Liu, Y. Lin, Y. Cao et al., "Swin transformer: hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, IEEE, Montreal, QC, Canada, October 2021.

[25] S. Zheng, J. Lu, H. Zhao et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6877–6886, IEEE, Nashville, TN, USA, June 2021.

[26] W. Yuan and W. Xu, "MSST-Net: a multi-scale adaptive network for building extraction from remote sensing images based on swin transformer," *Remote Sensing*, vol. 13, no. 23, Article ID 4743, 2021.

[27] W. Yuan, J. Wang, and W. Xu, "Shift pooling PSPNet: rethinking PSPNet for building extraction in remote sensing images from entire local feature pooling," *Remote Sensing*, vol. 14, no. 19, Article ID 4889, 2022.

[28] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" in *Advances in Neural Information Processing Systems*, vol. 34, pp. 12116–12128, NeurIPS 2021, 2021.

[29] O. Semih Kayhan and J. C. Van Gemert, "On translation invariance in CNNs: convolutional layers can exploit absolute spatial location," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14262–14273, IEEE, Seattle, WA, USA, June 2020.

[30] L. Xu, Y. Li, J. Xu, Y. Zhang, and L. Guo, "BCTNet: Bi-branch cross-fusion transformer for building footprint extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[31] J. Chen, Y. Lu, Q. Yu et al., "TransUNet: transformers make strong encoders for medical image segmentation," 2021.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE, Las Vegas, NV, USA, June 2016.

[33] Q. Zhu, Y. Zhang, L. Wang et al., "A global context-aware and batch-independent network for road extraction from VHR satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, pp. 353–365, 2021.

[34] Z. Xu, H. Guan, D. Peng, Y. Yu, X. Lei, and H. Zhao, "A dual-attention capsule network for building extraction from high-resolution remote sensing imagery," *National Remote Sensing Bulletin*, vol. 26, no. 8, pp. 1636–1649, 2022.

[35] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, IEEE, Honolulu, HI, USA, July 2017.

[36] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31×31: revisiting large kernel design in CNNs," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11953–11965, IEEE, New Orleans, LA, USA, June 2022.

[37] X. Gao, S. Ran, G. Zhang, and Y. Yang, *Building Extraction Based on Multi-Feature Fusion and Object-Boundary Joint Constraint Network*, Geomatics and Information Science of Wuhan University, 2022.

[38] X. Pan, F. Yang, L. Gao et al., "Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms," *Remote Sensing*, vol. 11, no. 8, Article ID 917, 2019.

[39] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: marrying convolution and attention for all data sizes," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 3965–3977, NeurIPS 2021, 2021.

[40] W. Wang and Z. Tu, "Rethinking the value of transformer components," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6019–6029, International Committee on Computational Linguistics, Barcelona, Spain (Online), December 2020.

[41] M. A. Islam, S. Jia, and N. D. B. Bruce, "How much position information do convolutional neural networks encode?" in *International Conference on Learning Representations*, ICLR 2020, September 2019.

[42] M. A. Islam, M. Kowal, S. Jia, K. G. Derpanis, and N. D. B. Bruce, "Padding and predictions: a deeper look at position information in CNNs," 2024.

[43] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Computer Vision – ECCV 2018*, pp. 3–19, Springer, Cham, 2018.

[44] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2019.

[45] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–5995, IEEE, Honolulu, HI, USA, July 2017.

[46] H. Wu, B. Xiao, N. Codella et al., "CvT: introducing convolutions to vision transformers," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22–31, IEEE, Montreal, QC, Canada, October 2021.

[47] W. Qiu, L. Gu, F. Gao, and T. Jiang, "Building extraction from very high-resolution remote sensing images using refine-UNet," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.
[48] X. Chu, Z. Tian, B. Zhang, X. Wang, and C. Shen, "Conditional positional encodings for vision transformers," 2023.
[49] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2023.
[50] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016.
[51] M. Lin, Q. Chen, and S. Yan, "Network in network," 2014.
[52] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE, Stanford, CA, USA, October 2016.
[53] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 6169–6181, 2021.
[54] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: fusing transformers and CNNs for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*, pp. 14–24, Springer, Cham, 2021.
[55] N. Ullah, T. Mahmood, S. G. Kim, S. H. Nam, H. Sultan, and K. R. Park, "DCDA-Net: dual-convolutional dual-attention network for obstructive sleep apnea diagnosis from single-lead electrocardiograms," *Engineering Applications of Artificial Intelligence*, vol. 123, Part C, Article ID 106451, 2023.