Mathematical Biosciences and Engineering

MBE, 20 (7): 11763–11784.
DOI: 10.3934/mbe.2023523
Received: 22 February 2023
Revised: 24 April 2023
Accepted: 25 April 2023
Published: 09 May 2023

http://www.aimspress.com/journal/MBE

Research article

Urban scene segmentation model based on multi-scale shuffle features

Wenjuan Gu, Hongcheng Wang, Xiaobao Liu, Yanchao Yin* and Biao Xu

Faculty of Mechanical & Electrical Engineering, Kunming University of Science & Technology, Kunming, 650500, China.

* **Correspondence:** Email: yinyc@163.com.

Abstract: The monitoring of urban land categories is crucial for effective land resource management and urban planning. To address challenges such as uneven parcel distribution, difficulty in feature extraction and loss of image information in urban remote sensing images, this study proposes a multi-scale feature shuffle urban scene segmentation model. The model utilizes a deep convolutional encoder-decoder network with BlurPool instead of MaxPool to compensate for missing translation invariance. GSSConv and SE module are introduced to enhance information interaction and filter redundant information, minimizing category misclassification caused by similar feature distributions. To address unclear boundary information during feature extraction, the model applies multi-scale attention to aggregate context information for better integration of boundary and global information. Experiments conducted on the BDCI2017 public dataset show that the proposed model outperforms several established segmentation networks in OA, mIoU, mRecall, P and Dice with scores of 83.1%, 71.0%, 82.7%, 82.7% and 82.5%, respectively. By effectively improving the completeness and accuracy of urban scene segmentation, this study provides a better understanding of urban development and offers suggestions for future planning.

Keywords: urban scene; remote sensing image; segmentation; feature shuffle; multi-scale attention

1. Introduction

Remotely-sensed image maps have been widely utilized in various fields such as urban planning due to their real-time characteristics and rich spectral and location information. Semantic segmentation, which involves the recognition and classification of features, has now become an important method for remote-sensing image analysis, providing invaluable data support for regulating urban development, preserving state-owned land and coordinating urban spatial layout. Unfortunately, remotely-sensed images commonly possess distinct target scales, an irregular number of targets and obscured targets, posing tremendous challenges to their interpretation. Therefore, the quality of remote sensing image segmentation is crucial to the further development of this discipline [1–4].

Image segmentation refers to the process of categorizing each pixel and segmenting it into regions that indicate differences between regions and similarities within regions, rendering the image more comprehensible and analyzable [5]. Threshold segmentation is done by thresholding the image grayscale values and dividing the image into regions according to the threshold value [6]. Region segmentation is to select seed points based on certain similar properties and cluster them according to that property [7]. Also, edge segmentation is to find out the image edge pixel points and join the found pixel points together to form the desired region boundary [8]. These methods are the traditional methods of image segmentation that are commonly applied.

Presently, thresholding techniques using meta-heuristic equalization algorithms are easily used for image segmentation [9]. Region segmentation can be achieved by using a combination of Multiresolution Segmentation (MRS) and Simple Linear Iterative Clustering (SLIC) superpixel algorithms [10], and from mathematical and geometric perspectives [11], among other approaches to achieve target regions. Edge segmentation is more widely used in medical and remote sensing images. By performing recursive iterations on pre-defined categories based on regions, it enables regions with under-grown high-ranking memberships to be associated with predicted categories in time to improve the accuracy of remote sensing images [12]. Furthermore, the features of motion cracks are extracted from medical CT images by an edge extraction-based algorithm as a way to improve the segmentation accuracy after image enhancement [13]. Most of the above segmentation is based on features such as grey scale, color, texture and shape of the image, which is mainly based on expert cognition and understanding and manual design. However, this approach has limited generalizability for remotely-sensed images due to the high variability in image features and changing shots over time that make feature-based segmentation less effective [14].

In recent years, the use of convolutional neural networks (CNNs) has become prevalent in image segmentation due to the remarkable advancements in deep learning techniques [15–17]. Various CNN-based semantic segmentation networks have been proposed, including FCN [18], SegNet [19], PSPNet [20], UNet [21] and DeepLabV3 [22]. A large number of studies have been conducted to improve the network to achieve higher recognition accuracy for the object of study. Adding other modules to the above high-precision network or modifying the structure of the network itself can achieve better understanding and recognition of the model. For example, Xie et al. [23] introduced DUSegNet, a distinctive semantic segmentation model for open-pit mines at a pixel level. DUSegNet combines SegNet's pyramidal model and upsampling method of pooling indices with UNet's convolutional skip connection architecture and dilated convolution's intensifier. Wang et al. [24] introduced a PSPNet based semantic segmentation network for coal gangue images (SSNet_CG), which extracts feature information by embedding feature fusion channels, an attention mechanism and a three-layer pyramid pooling module, providing a new idea for fast coal gangue recognition. Su et al. [25] designed an improved UNet network model that combines the benefits of DenseNet, UNet, Dilated Convolution and DeconvNet to perform remote sensing image segmentation. Liu et al. [26] presented a Context-Transfer-UNet (CT-UNet) network to solve the problem of blurred building map boundaries after segmentation and inconsistencies within classes caused by the similarity between buildings and backgrounds. Yang et al. [27] addressed the untimely acquisition of wheat inversion information and facilitated the identification of inversion losses in wheat seed selection through a DeepLabV3+

wheat inversion detection model based on multi-headed self-attention. Lastly, Belhadi et al. [28] developed a medical segmentation model based on the UNet network to train complex medical data for Internet of Medical Things (IoMT) scenarios.

Although the above deep learning methods show promising segmentation performance, their classification targets are relatively single, mostly scene-specific and of low generalization ability. To address the problems of uneven number of target distribution, difficulty in feature extraction and loss of image information in remote sensing images a network model of GSEPNet is proposed in this paper. The model is based on UNet for remote sensing image segmentation, and introduces shuffle feature extraction module (SF Block), which consists of BlurPool, GSSConv convolution and SE module. First, MaxPool in the routine UNet model is replaced with BlurPool to make up for the missing translation invariance due to data enhancement across blocks and improve segmentation accuracy. Second, in order to solve the problem of difficult feature extraction among multiple features with high similarity in urban remote sensing images, the GSSConv and SE module are introduced. They enhance the interaction of information between parcels, filter redundant information and improve network segmentation performance. Due to the lack of boundary information caused by the sliced image map, a multi-scale attention (MsCA) is constructed to improve the adaptive capture of global contextual information and further enrich the deep semantic information features.

The organization of this paper is as follows: In Section 2, the process of building the GSEPNet network is described in detail and significance modules in remote sensing image segmentation is introduced. In Section 3, the process of making the image dataset is described and the feasibility of the model is verified through a large number of comparison experiments. Section 4 discusses the advantages and limitations of the algorithm. Section 5 concludes the study and provides an outlook for the future.

2. The proposed Methods

In this paper, the input remote sensing images are scaled to 512×512×3 and passed into the model based on the encoder-decoder UNet network model. The encoder part consists of 4 blocks, where each block consists of 1 BlurPool layer [29], GSSConv convolution and channel attention SE [30], to reconstruct the SF Block. The number of feature maps is doubled after each downsampling, and MsCA is inserted after block 4, and its output features are fused with the block 3 output feature channel splicing as the input features in the decoding stage, then decoding is performed with bilinear interpolation upsampling after completing two ordinary convolutions. The segmentation of remote sensing images is achieved by repeating the decoding four times and then going through ordinary convolution. The final network model GSEPNet is shown in Figure 1.

In this paper, the segmentation model was proposed to solve the problems of uneven number of remote sensing image targets, difficulties in feature extraction and loss of image information.

(1) In order to solve the problem of missing translation invariance caused by the input of data-enhanced remote sensing images into the network, the missing translation invariance of the MaxPool layer with a step size of 2 can be better mitigated by using the BlurPool layer in the Encoder stage

(2) Due to the segmentation difficulties caused by the existence of highly similar features to parcels in urban remote sensing images, the interaction information between channels is preserved by GSSConv. Then, the SE module is introduced to differentiate the importance of different channel features and to locate and identify target areas more accurately. GSSConv and SE module introduce

spatial location relations for the decoder stage not only help the model to detect and extract the same image features, but also improve the model to obtain finer segmentation results.

(3) The processing of remote sensing images in a slice leads to the destruction of the overall structure of the image and increases the difficulty of image category segmentation. The final stage of encoder uses MsCA to aggregate boundary information and various categories of contextual information, in this way, the feature information between different channels can be dynamically interconnected to achieve the ability of accurate feature positioning and improve the performance of network segmentation.

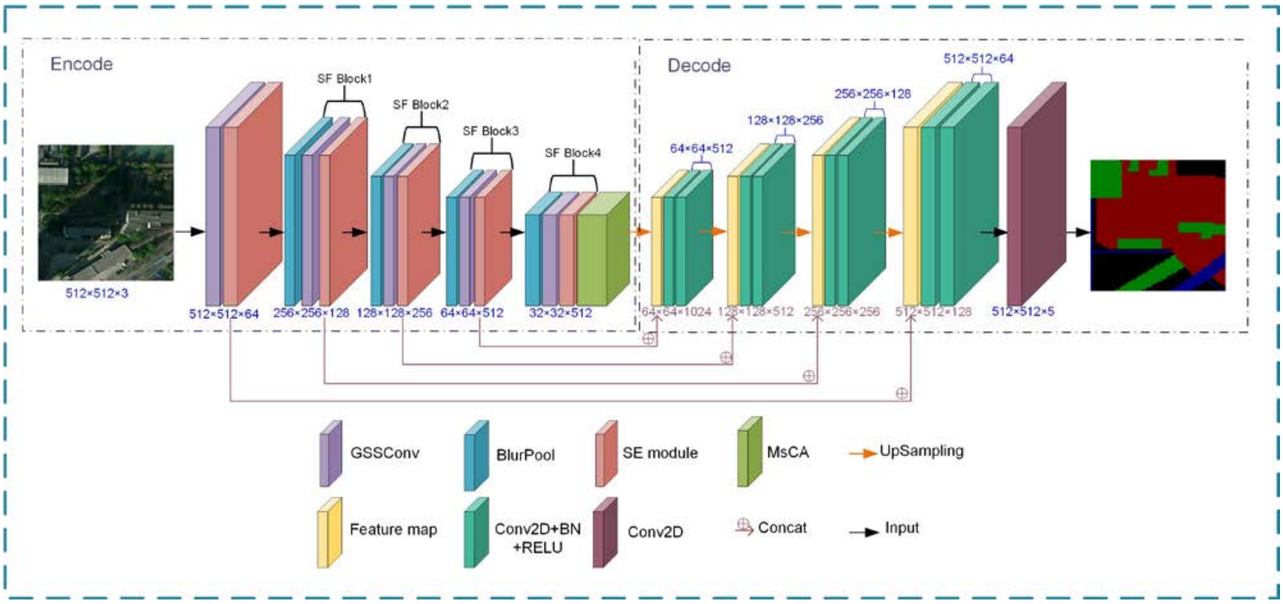


Figure 1. GSEPNet network model.

2.1. SF Block

Because of the large variety and uneven distribution of land parcels in urban remote sensing images, and the high similarity of interlaced distribution between individual parcels, especially for roads and background, it is likely to confuse target feature areas with non-target feature areas during feature extraction, leading to category misclassification. In the encoder stage of the UNet model, the feature extraction network is stacked with two 3×3 Standard Convolution (SC). Although the feature extraction capability and fusion capability of SC are efficient, the lack of information flow between groups and the retention of redundant information will limit the network segmentation performance.

In this paper, we reconstruct the feature extraction network of the model to alleviate the translation invariance of lost remote sensing images, enhance the information interaction between various types of parcels, and filter the redundant information to improve the classification accuracy of the network. The SF Block proposed in this paper consists of the following structures: First, the feature map is guaranteed to have translational invariance through BlurPool. Second, GSSConv is used to maintain the information interconnection between channels and space, which enhances the information transfer between channels and improves the training performance of the network. Finally, the SE module is embedded to weight the importance of the feature extracted images and enhance the classification of parcels with differential information extraction.

2.1.1. BlurPool

The original dataset was enhanced to solve the problem of uneven distribution of parcels across the remote sensing imagery, resulting in large differences among the accuracy of parcel segmentation. In most semantic segmentation networks, convolutional neural network and pooling are used to provide translation invariance. However, when the step size of the downsampling operation is greater than 2, the output will change drastically upon a small translation or transformation of the input. In this paper, BlurPool is used to alleviate the missing translation invariance of this downsampling and facilitate the system to produce identical responses.

MaxPool is most widely used because it can effectively reduce the amount of data in neural networks, simplify data and speed up data processing. Selecting the maximum value of the original data covered by the convolutional kernel to compress the feature map helps the network to remove redundant information, reduce the number of network parameters, enhance local information acquisition and retain texture information. As shown in formula (1) and formula (2).

$$H_{\text{out}} = \frac{H_{\text{in}} + 2 \times \text{padding}[0] - D[0] \times (K\_S[0] - 1) - 1}{S[0]} + 1 \tag{1}$$

$$W_{\text{out}} = \frac{W_{\text{in}} + 2 \times \text{padding}[1] - D[1] \times (K\_S[1] - 1) - 1}{S[1]} + 1 \tag{2}$$

Where H_{in} and H_{out} represent the input and output lengths respectively. W_{out} and W_{in} are input and output widths respectively. *padding* represents the feature image completion operation, *K_S* is the window size for MaxPool, *D* is the element step size in the control window and *S* is the step length.

The paper introduces BlurPool to proposed module, which combines low-pass filtering with anti-aliasing to retrieve to a certain extent the translation invariance lost by the convolution operation of MaxPool(stride > 1), as shown in Figure 2, where the difference in MaxPool's results is huge for just one pixel translation.

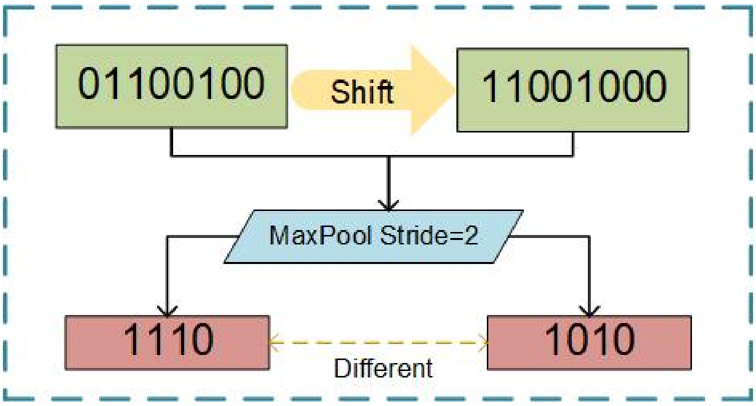


Figure 2. MaxPool operation.

The MaxPool (stride = 2) can be decomposed into two parts, MaxPool with stride = 1 and BlurPool with stride = 2. First, the shift isotropy is maintained by the Max operation being evaluated intensively in a sliding window. Second, an anti-aliasing filter with a kernel of m×m is added in combination with atomic sampling as a better representation of the intermediate signal, as shown in Figure 3.

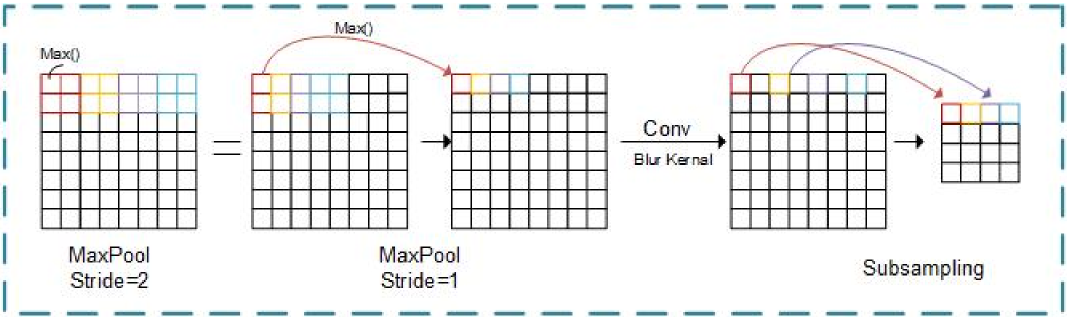


Figure 3. BlurPool operation.

2.1.2. GSSConv

The high degree of similarity between parcels inevitably leads to the accuracy of image segmentation, and most segmentation networks focus on contextual information linking of various categories of channel or spatial information, neglecting the fusion of information between categories in space or between channels, resulting in information lag between categories. This paper mainly uses GSSConv in the feature extraction network to make the information of each type of channel or space intermingled. First, the convolution compresses the input feature information to generate dense channel information. Then, the information is infiltrated into the feature information proposed by Depth-wise Separable Convolution (DSC) to achieve complete mixing of SC information into the output of DSC. Next, the channel shuffle is used to achieve dense mixed information of each category. Finally, the dense mixed information is compressed by 3×3 SC to expand the perceptual field and improve the generalization ability of the model, as shown in Figure 4.

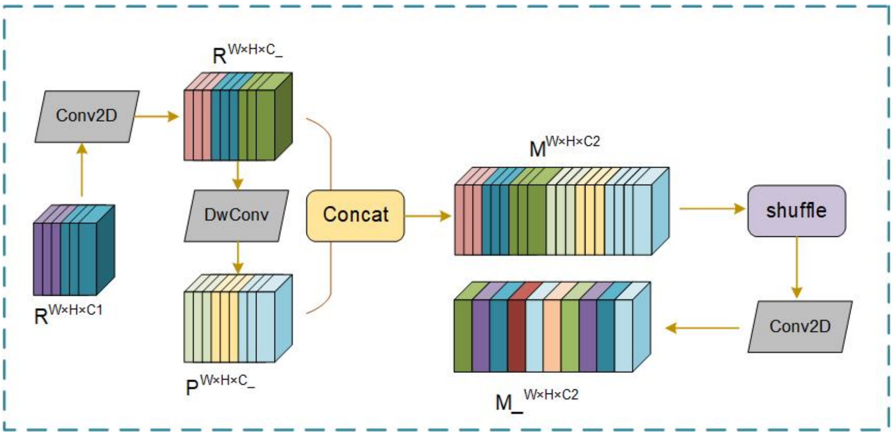


Figure 4. GSSConv structure.

The specific steps are as follows, input the feature map $R^{W\times H\times C1}$ from the original image or the output of the pooling layer into GSSConv, where $W\times H$ is the spatial dimensional size of the feature map and $C1$ represents the feature map tensor.

First, the input feature map tensor $C1$ is halved to $C\_$, and the dense feature information $O^{W\times H\times C\_}$ is extracted by taking the feature map $R^{W\times H\times C\_}$ and passing it into the SC with a convolution kernel size of 1. Second, the output feature information is input into the DSC to obtain the output feature information $P^{W\times H\times C\_}$. Then, the feature information $P^{W\times H\times C\_}$ and $R^{W\times H\times C\_}$ are stitched in dimension, so that the DSC and SC information are intermingled, achieving the output information $M^{W\times H\times C2}$. Next, the features of each group are dispersed to different groups by channel shuffle, which enables the output features to contain features from each group. Finally, the feature information is compressed by a 3×3 SC to extract more semantic information and enhance the network segmentation performance. The main formula for its GSSConv is shown below.

$$\begin{aligned} F = & (2\times C\_\times K_1^{\,2}-1)\times H\times W\times C\_+[(2\times K_2^{\,2}\times C\_/g+1)\times H\times W\times C\_/g]\times g \\ & +(2\times C_2\times K_3^{\,2}-1)\times H\times W\times C_2 \end{aligned} \tag{3}$$

where K_{n} represents the size of the convolution kernel for each convolution and g represents the number of input feature map subgroups.

2.1.3. SE Block

When using deep learning methods for semantic segmentation of images, the loss of space or channel information in the feature layer can limit the network segmentation performance. Channel attention SE is introduced in this model, which can assign the importance of information on the channel, focus on the information that is more critical to the current task among the numerous input information, suppress useless information, and improve the representation ability of the network model.

The SE module structure is shown in Figure 5, with the following steps.

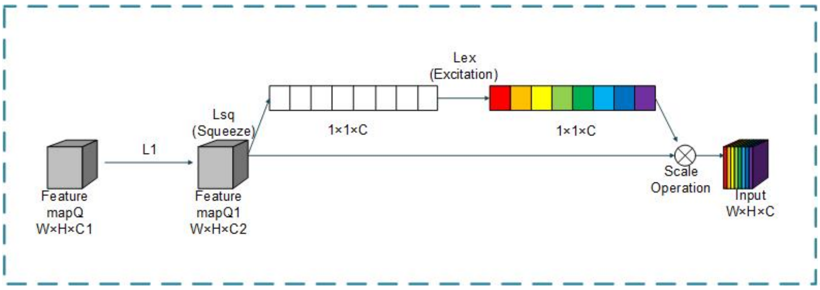


Figure 5. SE block structure.

The SE module includes three parts: *Squeeze*, *Excitation* and *Reweight*. The feature map of the input module is Q, whose dimension is $W\times H\times C1$. $W\times H$ is the spatial dimension of the feature map, and $C1$ represents the feature map tensor. The SE module is implemented through the following. First, through the convolution operation $L1$, the feature tensor of the input Q is converted from $C1$ to $C2$. Then, through the *Squeeze* operation L_{sq}, by global average pooling and the feature map $Q^{W\times H\times C2}$ is compressed into $X^{1\times 1\times C2}$. Next, the *Excitation* operation L_{ex} is performed, using the fitted correlation between the channels of the two fully connected layers. Finally, the *Reweight* operation multiplies the weight value of each channel calculated by the SE module with the two-dimensional matrix of the corresponding channel of the original feature map, and the result is outputted. The mapping relationship is shown in formula (4) to formula (6).

$$s_{\mathrm{q}} = L_{sq}(u_c) = \frac{1}{H \times W}\sum\nolimits_{i=1}^{H}\sum\nolimits_{j=1}^{W}\beta_c(i,j) \tag{4}$$

$$e_{\mathrm{x}} = L_{ex}(s_{\mathrm{q}},W) = \eta(g(s_{\mathrm{q}},W)) = \eta(W_2\delta(W_1 s_{\mathrm{q}})) \tag{5}$$

$$k_c = L_{scale} = (u_c, e_c) = e_c u_c \tag{6}$$

β_c represents each feature channel. δ is the Relu function. *W1* and *W2* are the parameters of full-connections operation. The sigmoid activation function represented by η.

2.2. MsCA Module

Slicing of remote sensing images can cause loss of contextual semantics and boundary information of the sliced images. Most of the semantic segmentation networks obtain contextual semantic information by fusing multi-scale feature information, but they can't extract features adaptively to link contextual information. This paper embeds the MsCA module to associate network context through the attention to assign weights to different scales of information. The MsCA module is shown in Figure. 6. First, the feature map *m* output from Block4 is subjected to convolution operation to generate feature map *q*. Then the output feature *q* is subjected to multi-scale feature extraction, i.e. the feature map *q* is pooled according to 3 different convolutional kernel sizes. By this method, three different scale feature maps are obtained. Finally, the multi-scale feature map is implemented through an attention mechanism to aggregate boundary information and contextual information between spaces and channels.

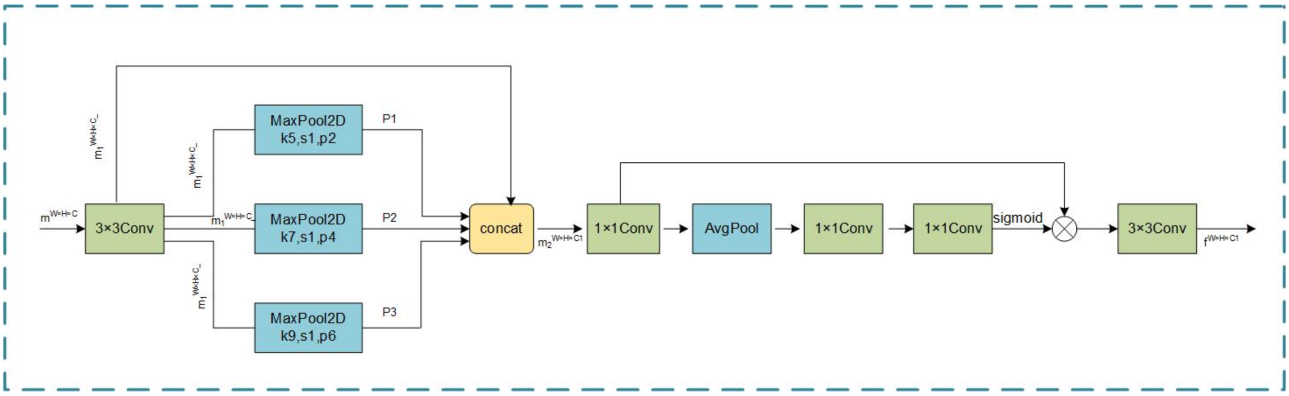


Figure 6. MsCA structure.

The specific steps are as follows. First, the feature map $m^{W\times H\times C}$ output from Block4 is halved in dimension to *C_* and the 3×3 convolution operation is used to output the feature information as $\mathrm{m}_1^{\mathrm{W\times H\times C\_}}$. Different scale pooling operations are used on the output feature map to obtain feature maps of different scales. The pooling module is 3 layers, each of size 5×5, 7×7 and 9×9, respectively, and MaxPool is applied to the feature maps to obtain P1, P2 and P3, 3 pooled feature maps of different sizes with the same dimension size as *C_*. Then, pooled feature maps are stacked with $m_1^{W\times H\times C\_}$ on the channel to output the feature $m_2^{W\times H\times C1}$. Next, $m_2^{W\times H\times C1}$is sequentially subjected to a Squeeze operation with global pooling and an Excitation operation consisting of two full joins, and the channel attention weights are extracted using the Sigmoid function for the output features, then the weights are mapped to the feature information $m_2^{W\times H\times C1}$ to obtain the output feature information. Finally, the spatial information of all channels is blended by 3×3 SC compression of the feature information. The MsCA module relationship is as following Eq(7).

$$Ms = Sigmoid[Avgpool(\varphi(m_2^{W\times H\times C1}))W_2\delta(W_1 s_q)]\times m_2^{W\times H\times C1}\times\varphi(\mathrm{f}) \tag{7}$$

where $\varphi()$ denotes a one-dimensional convolutional channel interaction, Avgpool represents the global average pooling operation, δ is the Relu function, W_1 and W_2 are the parameters of full-connections operation.

3. Experimental Analysis and Discussion

3.1. Experimental platform construction

The model was trained on an 16G RAM, AMD R5-3600 CPU and NVIDIA GeForce RTX 2080Ti 12GB GPU provided by Ubuntu 18.04 OS. The deep learning framework is Pytorch 1.6.0, using CUDA Toolkit 10.0 and CUDNN V 7.6.5 as the model training acceleration toolkit.

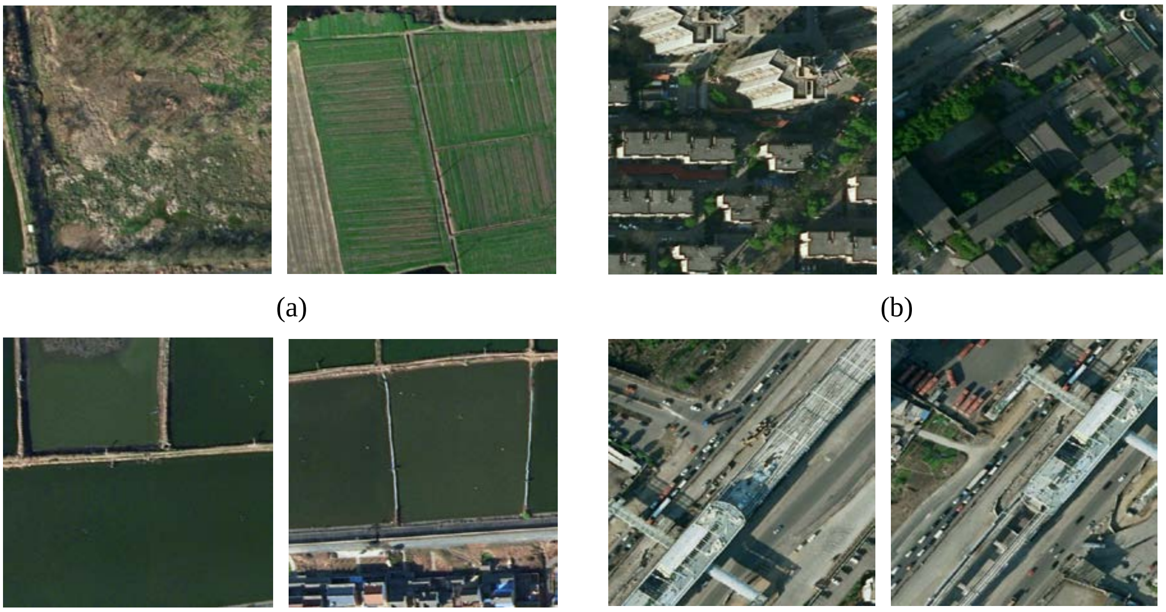
3.2. Dataset

The dataset utilized in this paper was obtained from the AI Classification and Recognition of Satellite Images competition held during the 2017 CCF Big Data and Computational Intelligence Conference (BDCI 2017). The dataset consists of land cover samples visually interpreted from high-resolution remote sensing imagery depicting a southern Chinese region, captured using sub-meter resolution and visible spectral bands (R, G, B).

3.2.1. Data acquisition

The BDCI2017 dataset comprises of three remote sensing images that have been annotated. The dimensions of the images are 4011×2470 pixels, 5664×5142 pixels and 3357×6116 pixels, respectively. The training samples are categorized into five classes: vegetation (marker 1), buildings (marker 2), water bodies (marker 3), roads (marker 4) and others (marker 0). Cropland, forest land and grassland are classified as vegetation. The labels were given pseudo-colors to help visualize subsequent experiments. Figure 7 illustrates a part of the schematic diagram.

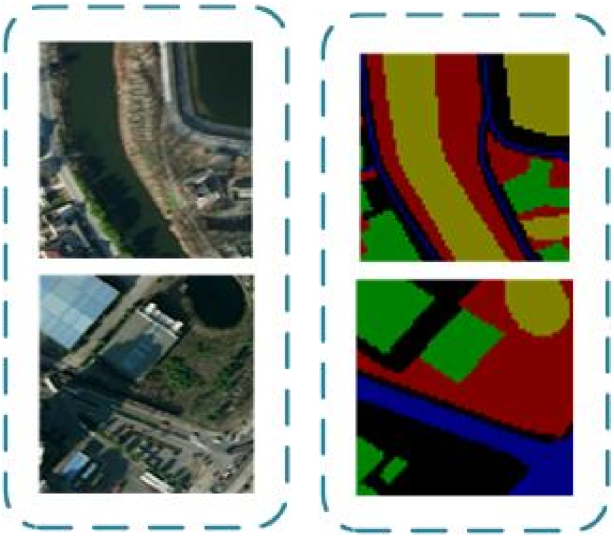
The dataset labeling follows a color scheme where red areas represent vegetation, green areas reflect buildings, yellow areas denote water bodies, blue areas indicate roads and black areas serve as the background. Refer to Figure 8 for illustration.



(a) (b)

(c) (d)

Figure 7. Some example images from BDCI2017 dataset. (a) Vegetation; (b) Buildings; (c) Water bodies; (d) Roads.



(a) (b)

Figure 8. Partially labelled datasets. (a) Indicates the original image of the urban remote sensing image; (b) Denotes the urban remote sensing label map corresponding to the original image.

3.2.2. Data pre-processing

The dataset has many challenges related to image extraction. For instance, direct processing of remote sensing images can cause memory shortages, direct cutting of images can lead to loss of contextual semantics. Additionally, the clipped image samples have the disadvantage of insufficient data volume. Consequently, the samples need to be processed for sliding segmentation and data enhancement. The image samples are pre-processed as follows.

(1) Sliding segmentation: the original image is sliced to save internal storage and enhance contextual semantic information. Select a sliding window of 512×512 and cut one piece per sliding 128 pixels, where the image size is 512×512. The sliding spacing of 128 pixels can effectively eliminate the loss of semantic information between features due to image segmentation. Total 874 remote sensing images were segmented and stored in the format of "jpg". The segmentation effect is shown in Figure 9.

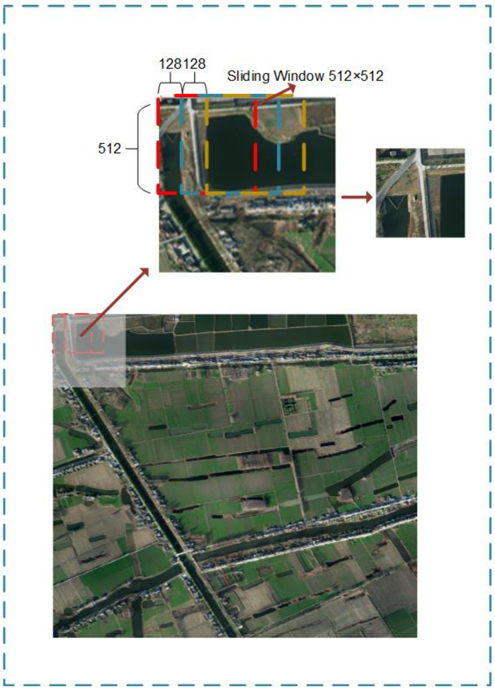


Figure 9. Sliding segmentation process.

(2) Data enhancement: at the beginning of the model training, image data are enhanced by rotating, panning and mirroring the data set to prevent underfitting of the model while expanding the sample size of remote sensing, so that the model has high detection capability of the target and generalization of the model. The enhanced sample size reaches 3104. These enhanced urban remote sensing images were divided into training and validation sets (at 9:1 ratio), and datasets were produced in PASCAL VOC format to facilitate subsequent experiments.

3.3. Evaluation indicators

The segmentation report of the proposed model is generated through the evaluation of several measures, including Overall Accuracy (OA), Mean IoU (mIoU), Mean Recall (mRecall), Mean Precision (P) and Dice coefficient (Dice).

OA measures the overall classification accuracy of all samples. On the other hand, mIoU computes the average ratio of the intersection of the predicted and true values of all classes to their concurrent sets. mRecall, on the other hand, represents the probability that the prediction is positive among all positive samples. Furthermore, mPrecision captures the probability that the true class is positive among all samples with positive predictions. Finally, Dice measures the similarity between the sets of predicted positive classes and true positive classes. Meanwhile, the memory (GPU) used by the training model is counted. In summary, these metrics are described as follows:

$$OA = \frac{S_{correct}}{S} \tag{8}$$

$$mIoU = \frac{[(C_{S1} \cap P_{S1}) / (C_{S1} \cup P_{S1}) + (C_{S2} \cap P_{S2}) / (C_{S2} \cup P_{S2}) ... + (C_{Sn} \cap P_{Sn}) / (C_{Sn} \cup P_{Sn})]}{N} \tag{9}$$

$$mRecall = \frac{TP}{TP + TN} \tag{10}$$

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$Dice = \frac{2 \times TP}{FN + TP + TP + FP} \tag{12}$$

where $S_{correct}$ represents all correctly classified samples and S represents the total number of samples. Moreover, C_{Sn}, P_{Sn} and N represent the true sample of a class, the sample predicted to be that class and all classes, respectively. TN, TP, FN and FP as shown in Table 1.

Table 1. Relationship between TN, TP, FN and FP.

The Real Deal	**Predicted results**	
	Positive examples	**Counter examples**
Positive examples	TP	FN
Counter examples	FP	TN

3.4. Experimental parameter setting

In this paper, UNet was chosen as the base model and the network was trained for 100 epochs. We train the model from scratch in the urban image segmentation phase. In order for the network to train the dataset correctly, the images were resized to 512×512 and normalized uniformly before being fed in the model. To ensure the validity of the experimental results, the training and validation sets were randomly shuffled while multiple cross-validations were performed before entering the network. An iterative reduction algorithm was applied to the gradient of this network to decrease the learning rate and improve segmentation performance.

The stochastic gradient descent (SGD) optimizer was used to update the model parameters, with an initial learning rate of 0.00001. This learning rate was adjusted for each step using the "Warmup" method, with a weight decay of 1e-4, momentum of 0.9 and dampening of 0. Due to computer hardware limitations and graphics card memory, the batch size was set to 4. Moreover, we used the dice loss function to mitigate the negative impact of foreground-background imbalance in the sample.

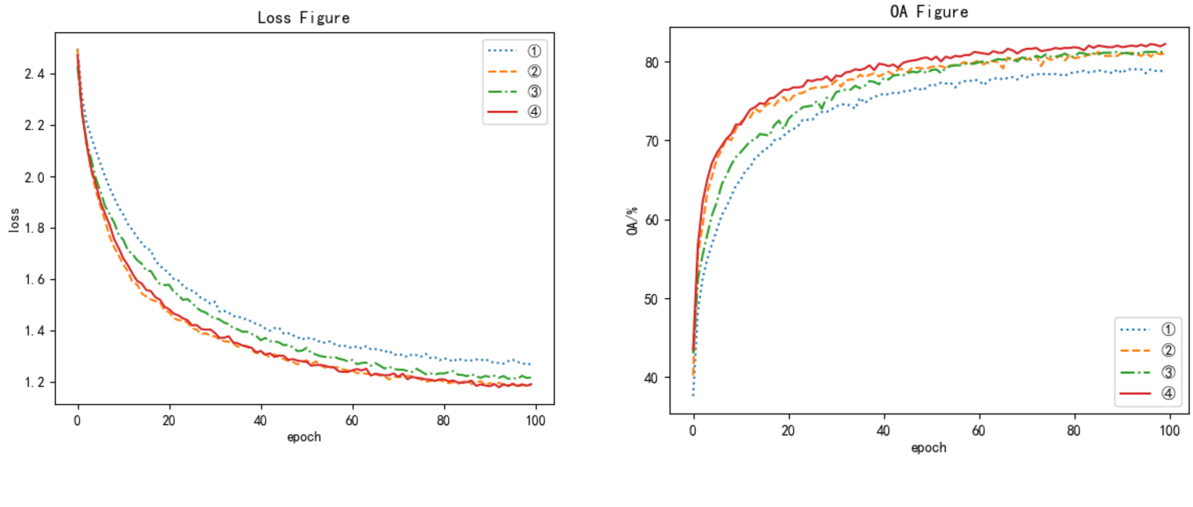
3.5. Analysis of experimental results

3.5.1. Ablation experiments

To demonstrate that the GSEPNet model designed in this paper improves translation invariance of feature images, interactivity of feature information and dynamic capture of spatial location information in remote sensing segmentation of urban images, the SF Block feature extraction module and MsCA module were used for ablation experiments under the framework of the network, respectively. To ensure the validity of the experimental results, we conducted all experiments on the same dataset and experimental environment built in Section 3.1 of this paper. The experimental results are shown in Table 2, and a comparison of the loss rate and OA for each model is shown in Figure 10.

Table 2. Results of ablation experiments with different modules.

Number	UNet	SF Block	MsCA	OA/%	mIoU/%	mRecall/%	P/%	Dice/%
①	✓			79.0	65.3	77.4	79.1	78.2
②	✓	✓		81.6	69.2	80.7	81.9	81.1
③	✓		✓	81.3	68.7	80.4	81.3	80.8
④	✓	✓	✓	83.1	71.0	82.7	82.7	82.5



(a) (b)

Figure 10. Comparison of loss rate and OA for each model. (a) Comparison of loss rates between models; (b) Comparison of the overall accuracy between models.

Table 2 shows the performance improvement when replacing the feature extraction module of the original UNet network with the SF Block module. The SF Block module effectively addresses segmentation challenges arising from the lack of translation invariance and highly similar features of remote sensing images. The improved OA, mIoU, mRecall, P and Dice are 81.6%, 69.2%, 80.7%, 81.9% and 81.1%, respectively. To further address the original network's bias towards local information extraction, MsCA is introduced to enhance contextual information fusion. The refined network improves OA, mIoU, mRecall, P and Dice by 2.3%, 3.4%, 3.3%, 2.2% and 2.6%, respectively, over the UNet model. Figure 10 shows that the GSEPNet network has a better segmentation performance than the UNet network. Both the SF Block and MsCA modules positively impact the segmentation effect. In summary, compared to the original network model, GSEPNet improves OA, mIoU, mRecall, P and Dice by 4.1%, 5.7%, 5.3%, 3.6% and 4.3%, respectively.

3.5.2. Shuffle feature extraction Block experiment

To validate the effect of different positions and numbers of SF Block features on network segmentation accuracy, this paper investigates how the position of the shuffle feature extraction module in the network can affect segmentation performance. The feature extraction Blocks of the original UNet network were replaced with various numbers and positions of SF Blocks to conduct the experiments, and the results are presented in Table 3. The positions of SF Blocks in Table 3 are indicated with 0 and 1, representing whether the respective four positions in the network were replaced with SF Blocks. The value of 0 indicates no replacement, while the value of 1 indicates replacement.

Table 3. Comparative experiments on the performance of using SF block.

SF Block Location	OA/%	mIoU/%	mRecall/%	P/%	Dice/%
(0,0,0,0)	79.0	65.3	77.4	79.1	78.2
(1,0,0,0)	79.7	66.3	79.0	79.4	78.9
(1,1,0,0)	80.6	67.7	80.0	80.3	80.1
(1,1,1,0)	81.0	68.0	80.4	80.6	80.3
(1,1,1,1)	81.6	69.2	80.7	81.9	81.1

The findings presented in Table 3 demonstrate that the stacked replacement feature extraction network has a positive impact on network segmentation, particularly when deploying a higher number of SF Blocks. Adopting a fully SF Block-constructed feature extraction network yields significant improvements, with OA increasing to 81.2% and mIoU to 68.3%, alongside mRecall, P and Dice enhancements of 3.3%, 2.8% and 2.9%, respectively. Our experiments confirm that the SF Block-reconstructed feature extraction network employed in the paper not only addresses the missing translational invariance of data expansion, but also boosts information intermingling and differential information extraction between different locations. Therefore, employing SF Block as the feature extraction module enables optimal feature information acquisition, leading to improved network accuracy.

3.5.3. Experimental analysis of multiscale channel attention mechanisms

In this study, we evaluate the performance of two models, GS-SPPNet and GSEPNet, which incorporate the SPP [31] and MsCA modules, respectively, at the end of the SF Block feature extraction network in a typical encode-decode structural framework for network segmentation. Our experimental comparison, detailed in Table 4, shows that incorporating the multiscale structure has a positive impact on segmentation performance. Specifically, the GSEPNet model outperforms the GS-SPPNet model in terms of OA, mIoU, mRecall, P and Dice by 0.5%, 0.6%, 0.3%, 0.8% and 0.4%, respectively. These results confirm that the MsCA module used in this paper effectively extracts feature-linked contextual information, fuses local features with global information and enhances the aggregation of local object-to-boundary information. Ultimately, our findings suggest that multiscale image fusion using the MsCA module can significantly improve the model's accuracy for network segmentation.

Table 4. Comparison of experimental performance of different multi-scale modules.

Network Model	OA/%	mIoU/%	mRecall/%	P/%	Dice/%
GS-SPPNet	82.5	70.4	82.4	81.9	82.1
GSEPNet	83.1	71.0	82.7	82.7	82.5

3.5.4. Comparison tests of different models

To evaluate the effectiveness of our proposed GSEPNet model for segmentation on the BDCI2017 dataset, we compared it against several other models, including FCN [18], PSPNet [20] , DeepLabV3 [22], DeepLabV3+ [32], LR-ASPP [33], MobileNetV2-DeepLabV3+ [34] and UNet [21]. All experimental comparisons were conducted on the experimental environment built in Section 3.1 of this paper to ensure the validity of the results. The experimental outcomes are reported in Table 5.

Table 5. Comparison of different segmentation network models.

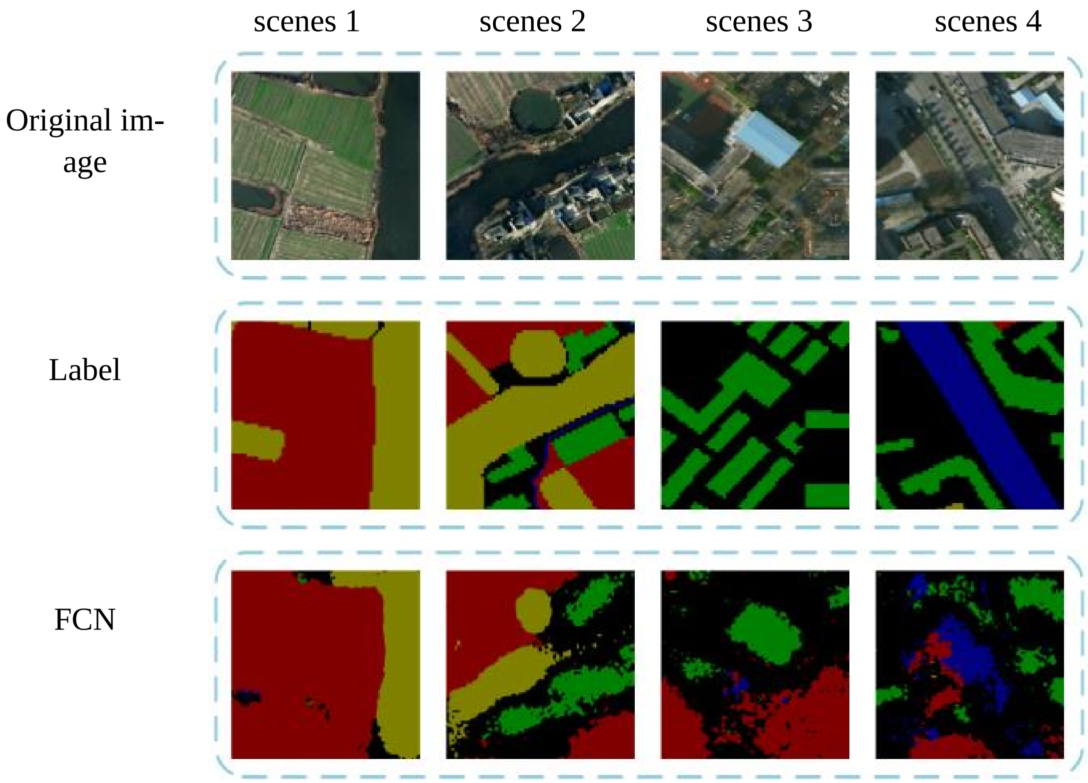
Network Model	OA/%	mIoU/%	mRecall/%	P/%	Dice/%	Other IoU/%	Vegetation IoU/%	Architecture IoU/%	Water bodies IoU/%	Roads IoU/%	GPU/G
FCN	65.1	43.3	55.3	68.3	57.9	43.9	69.7	31.3	56.9	14.8	6.7
PSPNet	81.2	68.5	81.2	80.4	80.7	59.0	84.0	53.0	81.0	66.0	2.6
DeepLabV3	65.1	43.1	57.6	60.6	58.0	42.8	70.2	32.2	52.6	17.5	6.5
DeepLabV3+	81.0	65.0	81.0	76.7	77.9	49.0	82.0	53.0	82.0	59.0	9.2
Iraspp	62.2	37.3	49.7	64.2	50.5	42.8	67.1	27.3	46.4	3.0	2.6
MobileNetV2-DeepLabV3+	82.8	65.7	82.7	77.2	78.3	48.0	83.0	56.0	83.0	58.0	4.8
UNet	79.0	65.1	77.4	79.1	78.2	55.8	81.6	49.0	79.1	60.1	9.6
GSEPNet	83.1	71.0	82.7	82.7	82.5	63.0	84.9	56.1	82.3	68.5	10.4

As can be seen from Table 5, with the same sample pool for model training, the GSEPNet model proposed in this paper achieves an overall classification accuracy of 83.1%, an mIoU of 71.0%, an mRecall of 82.7%, an P of 82.7% and an Dice of 82.5%, which are higher than other encode-decode segmentation networks, indicating that the GSEPNet model designed in this paper can be effectively used for urban parcel classification and recognition scenarios.

GSSNet occupies the best index scores in more categories, especially in the vegetation category, where it has the highest segmentation IoU of more than 80% compared with the other six models; for the water body category, the proposed model in this paper has only 0.7% lower IoU than the MobileNetV2-DeepLabV3+ model, and also achieves higher segmentation results compared with the remaining models. Analysis of the reasons for this shows that the dataset has a concentrated distribution of vegetation and water bodies, with more samples collected. Roads have a low IoU among all parcel segmentation due to their scattered distribution and high similarity to other targets, where GSEPNet reduces the classification error by introducing the SF Block feature extraction module, which increases the IoU of road segmentation to 68.5%. The building plots have the lowest segmentation IoU because they contain more scattered small targets and the lack of edge information makes it difficult to identify the targets.

As shown in Figure. 11, the urban remote sensing image segmentation m odels of FCN, PSPNet, DeepLabV3, DeepLabV3+, LR-ASPP, MobileNetV2-DeepLabV3+, UNet and GSEPNet were obtained based on the training models, and 4 different urban remote sensing image object block segmentations were selected for analysis, where scenes 1 to 4 are shown from left to right respectively. Comparing the overall segmentation results, the former six networks are less effective than the GSEPNet model and show misclassification in different cases, while the boundary information is handled more ambiguously. GSEPNet uses the SF Block feature extraction module to enhance differential feature extraction and reduce feature errors, and the MsCA module can effectively aggregate multi-scale feature information and capture boundary information.

In scenes 1 and 2, the vegetation and water targets are more widely distributed, GSEPNet handles the target outline and other target boundary information more clearly. In scenes 3 and 4, there are too many small targets in the building category and the targets are not evenly distributed. GSEPNet effectively retains the detail information of small-scale targets compared with other networks. In Scene 4, the road plots is scattered, which makes it difficult to identify each target and the boundary contours are difficult to be segmented, while GSEPNet performs well and is capable of handling complex segmentation scenes.



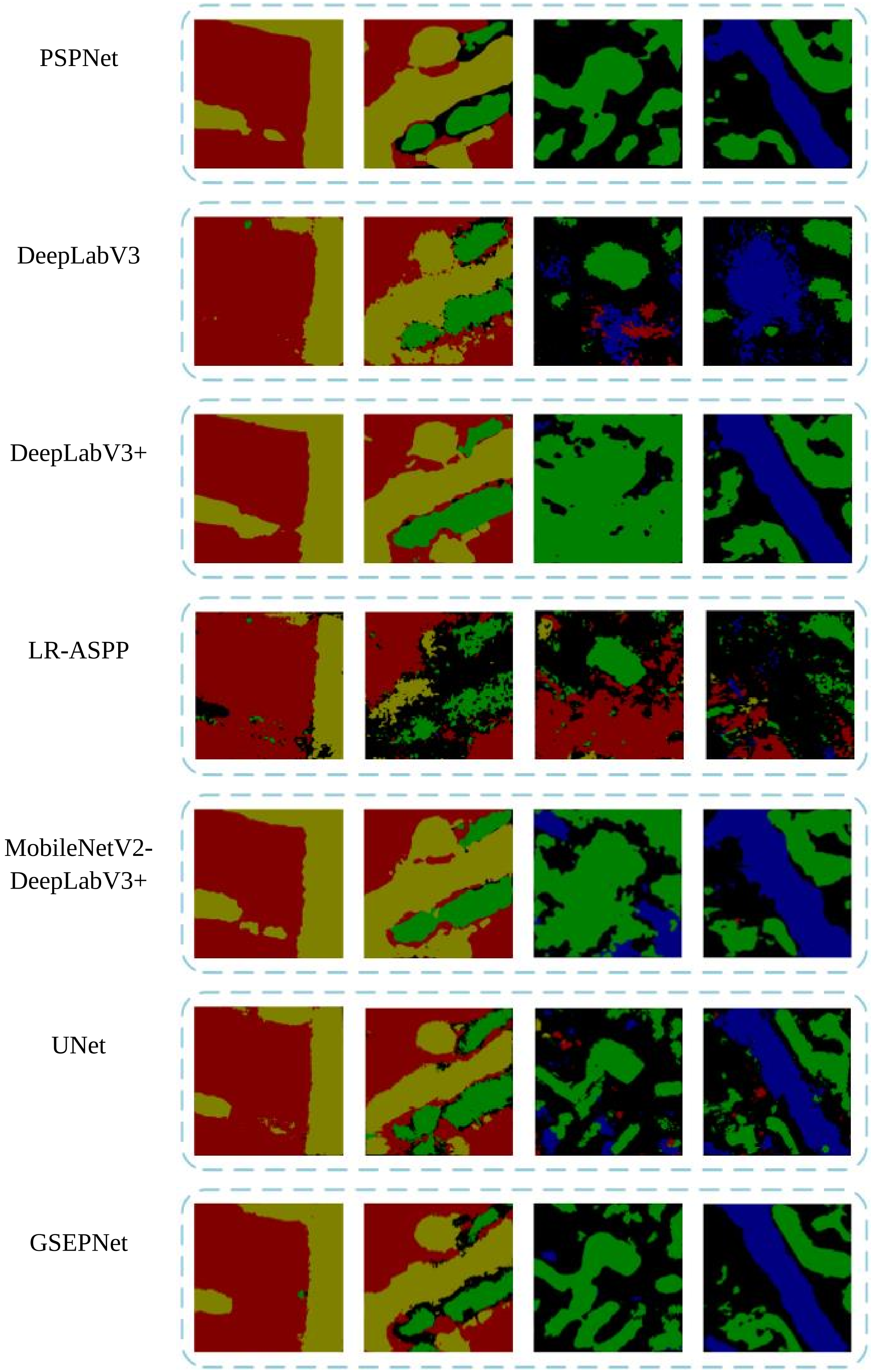


Figure 11. Comparison of different scenarios of segmentation without using network models.

3.5.5. Comparison tests of models in other urban remote sensing datasets

To verify the segmentation performance and model generalization ability of GSEPNet on other urban remote sensing images, the HD-Maps dataset is selected for model validation [35]. The dataset contains a total of 15 labeled images, including parking (marker 0), road (marker 1), building (marker 2), sidewalk (marker 3) and background (marker 4). In this experiment, the original and labeled maps of the dataset are cut into 512×512 size for training and validation using the sliding segmentation in Section 3.2.2, and the training and validation sets are divided in the ratio of 9:1. The GSEPNet model is compared with PSPNet, DeepLabV3+, ResUNet and UNet segmentation models. To ensure the validity of the experimental results, these experiments were conducted in the experimental environment constructed in Section 3.1 of this paper, and the experimental results are shown in Table 6. The segmentation effect of GSEPNet model is shown in Figure 12. Due to the uneven distribution of the number of categories in the HD-Maps dataset and the sporadic distribution of some categories, it is difficult to segment the tiny targets. The training results show that the GSEPNet model can segment better than the other four segmentation models without any processing of the dataset categories, so the GSEPNet model has higher generalization in the segmentation of urban remote sensing images.

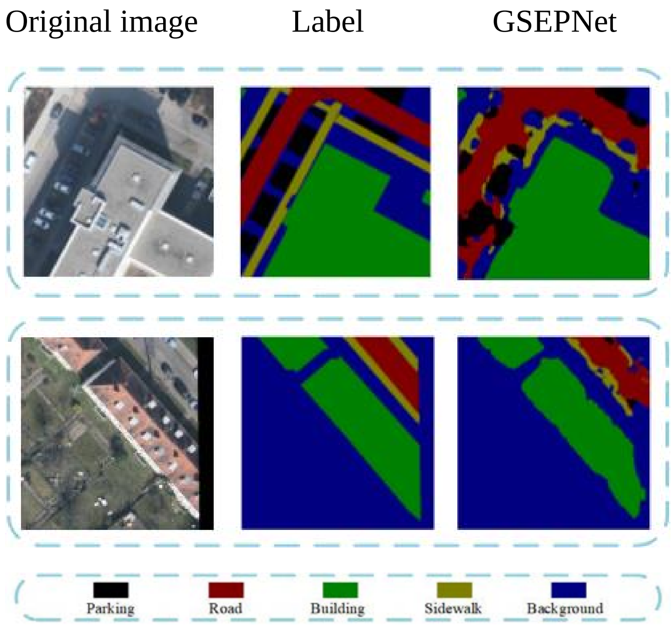


Figure 12. Segmentation results on HD-Maps dataset.

Table 6. Training results of HD-Maps dataset.

Network Model	OA/%	mIoU/%	mRecall/%	P/%	Dice/%	GPU/G
PSPNet	59.9	48.7	59.9	67.5	61.4	2.7
DeepLabV3+	66.2	53.8	66.2	67.4	66.4	9.4
MobileNetV2-DeepLabV3+	75.7	57.5	75.7	67.8	70.9	4.8
UNet	84.5	49.9	60.1	68.8	62.4	9.7
GSEPNet	86.2	53.8	64.0	72.2	66.6	10.5

4. Discussion

In semantic segmentation, the effectiveness of segmentation is related to the selection of networks, the optimization of parameters and the construction of models. This study is focused on the recognition and segmentation of urban features. The recognition of features in remote sensing images relies on spatial location relationships and textures, and requires a combined shallow and deeper network layer for feature extraction, so UNet is chosen as the basis of the segmentation model.

The advantages and disadvantages of the method are discussed. The advantages of this study's approach to urban feature segmentation are as follows.

(1) The image dataset is more informative as it uses publicly available datasets.

(2) The algorithm is also computationally inexpensive as the equipment used to train it is simple, and most computer devices are capable of supporting the model.

(3) GSEPNet has excellent performance in segmenting urban features.

The proposed algorithm also has the following limitations:

(1) The training set has only 2130 images, a small sample of the dataset. Expansion of the dataset can improve the network performance.

(2) Due to the limitation of experimental equipment, the segmentation effect of the model will be improved if the network parameters, such as batch_size and epoch, are further optimized.

5. Conclusion

This paper proposes a deep learning-based urban remote sensing image segmentation network named GSEPNet. It addresses the challenge of uneven distribution of quantities, undifferentiability among parcels and loss of boundary information by incorporating a shuffle feature extraction module and a multi-scale attention mechanism into the UNet network. The GSEPNet model proposed in this paper, which yielded OA of 83.1%, mIoU of 71.0%, mRecall of 82.7%, P of 82.7% and Dice of 82.5% achieves relatively high segmentation results on the BDCI2017 public dataset. Compared to other models such as FCN, PSPNet, DeepLabV3, DeepLabV3+, Iraspp, MobileNetV2-DeepLabV3+ and UNet, GSEPNet achieved the highest segmentation accuracy and average overlap among the eight networks. Further, it is proved experimentally that using shuffle feature extraction module and multi-scale attention in the GSEPNet model can improve the image segmentation effect of the network. The segmentation model proposed in this paper can better segment the feature classes in remote sensing images, and is highly feasible. The feature segmentation in urban remote sensing images mainly relies on spatial location information, and the image information is constantly increasing. Therefore, future research needs to consider how to further increase the depth of the network, reduce the amount of computation in the network and challenge more complex datasets.

Acknowledgments

This research was funded by Project supported by the National Natural Science Foundation, China, grant number 52065033

Conflict of interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this paper.

References

1. Z. Z. Fan, S. Wang, H. Zhang, R. L. Shi, W. J. Fu, M. Z. Li, W-Net-Based segmentation for remote sensing satellite image of high resolution, *J. South China Uni. Technol. (Natural Science Edition)*, **48** (2020**)**, 114–124. https://doi.org/10.12141/j.issn.1000-565X.200365
2. J. X. Zhang, L. X. Wang, Image segmentation models of remote sensing using full residual connection and multiscale feature fusion, *N. Remote Sens Bull.*, **24** (2020), 1120–1133. https://doi.org/10.11834/jrs.20208365
3. M. M. Li, A. Stein, M. K. de Beurs, A bayesian characterization of urban land use configurations from VHR remote sensing images, *Int. J. Appl. Earth Obs. Geoinf.*, **92** (2020), 102175. https://doi.org/10.1016/j.jag.2020.102175
4. D. L. Mao, Z. Zheng, F. X. Meng, C. Y. Zhou, J. P. Zhao, H. Z. Yang, et al., Large-scale automatic identification of urban vacant land using semantic segmentation of high-resolution remote sensing images, *Landscape Urban Plan*, **222** (2022), 104384. https://doi.org/10.1016/j.landurbplan.2022.104384
5. D. H. Cheng, H. X. Jiang, Y. Sun, L. J. Wang, Color image segmentation: Advances and prospects, *Pattern Recogn.*, **34** (2001), 2259–2281. https://doi.org/10.1016/S0031-3203(00)00149-7
6. S. S. Al-amri, V. N. Kalyankar, S. D. Khamitkar, Image segmentation by using threshold techniques, *Comput. Vis. Pat. Recog. (CVPR) (cs.CV)*. arXiv: 1005. 4020 [cs.CV]. https://doi.org/10.48550/arXiv.1005.4020
7. F. Meyer, Color image segmentation, *Intl. Conf. Im. Prcsg. Appls.*, Maastricht, Netherlands, 1992. https://ieeexplore.ieee.org/abstract/document/785528/
8. K. G. Hassana, J. B. Zou, Region-Based segmentation versus edge detection.5 *Intl. Conf. Intell. Info. Hdg & MM Sig. Prcsg.*, Kyoto, Japan, 2009. https://doi.org/10.1109/IIH-MSP.2009.13
9. M. Abdel-Basset, V. Chang, R. Mohamed, A novel equilibrium optimization algorithm for multi-thresholding image segmentation problems, *Neural Comput. Appl.*, **33** (2021), 10685–10718. https://doi.org/10.1007/s00521-020-04820-y
10. O. Csillik, Fast segmentation and classification of very high resolution remote sensing data using SLIC superpixels, *Remote Sens*, **9** (2017), 243. https://doi.org/10.3390/rs9030243
11. X. B. Liu, S. S. Wang, J. C. W. Lin, S. Liu, An algorithm for overlapping chromosome segmentation based on region selection, *Neural Comput. Appl.*, (2022). https://doi.org/10.1007/s00521-022-07317-y
12. P. M. Cipolletti, A. C. Delrieux, G. M. E. Perillo, M. P. Cintia, Superresolution border segmentation and measurement in remote sensing images, *Comput. Geosci.*, **40** (2012), 87–96. https://doi.org/10.1016/j.cageo.2011.07.015
13. Q. Nie, Yb. Zou, J. C. W. Lin, Feature Extraction for Medical CT Images of Sports Tear Injury, *Mobile Netw Appl*, **26** (2021), 404–414. https://doi.org/10.1007/s11036-020-01675-4
14. K. Z. Wu, S. Zhao, W. H. Li, R. Y. Jiang, Spatial global context information network for semantic segmentation of remote sensing image, *J. Zhejiang Uni. (Engineering Science)*, **56** (2022). 795–802. https://doi.org/10.3785/j.issn.1008-973X.2022.04.019
15. C. T, Tian, X. Y. Zhang, J. C. W. Lin, W. M. Zuo, Y. N. Zhang, C. W. Liu, Generative Adversarial Networks for Image Super-Resolution: A Survey. *Img. Vid. Prcsg. (eess.IV); Comput. Vis. Pat. Recog. (CVPR) (cs.CV)*. arXiv: 2204. 13620 [cs.CV]. https://doi.org/10.48550/arXiv.2204.13620

16. U. Ahmed, J. CW. Lin, G. Srivastava, Ensemble-based deep meta learning for medical image segmentation, *J. Intell. Fzy. Syst.*, **42 (**2022), 4307–4313. https://doi.org/10.3233/JIFS-219221
17. W. Z. Liu, P. Luo, G. X. Wang, O. X. Tang, Deep learning face attributes in the wild, *Intl. Conf. Comput. Vis. (ICCV),* Santiago, Chile, 2015. https://doi.org/10.1109/ICCV.2015.425
18. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, *Comput. Vis. Pat. Recog. (CVPR),* Boston, USA, 2015. https://doi.org/10.1109/CVPR.2015.7298965
19. V. Badrinarayanan, A. Kendall, R. CipollaI, SegNet: A deep convolutional encoder-decoder architecture for image segmentation, *Trans. Pat. Anal. Mach. Intell.*, **39** (2017), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615
20. S. H. Zhao, J. Shi, J. X. Qi, G. X. Wang, Y. J. Jia, Pyramid scene parsing network, *Comput. Vis. Pat. Recog. (CVPR),* Honolulu, USA, 2017. https://doi.org/10.1109/CVPR.2017.660
21. O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, *Comput. Vis. Pat. Recog. (CVPR) (cs.CV).* arXiv:1505.04597 [cs.CV]. https://doi.org/10.48550/arXiv.1505.04597
22. LC. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, *Comput. Vis. Pat. Recog. (CVPR) (cs.CV).* arXiv: 1706. 05587 [cs.CV]. https://doi.org/10.48550/arXiv.1706.05587
23. B. H. Xie, Z. Y. Pan, H. J. Luan, X. Yang, W. Y. Xi, Open-pit mining area segmentation of remote sensing images based on DUSegNet, *J. Indian Soc. Remote*, **49** (2021), 1257–1270. https://doi.org/10.1007/s12524-021-01312-x
24. X. Wang, C. Y. Guo, S. Wang, G. Cheng, Q. X. Wang, L. He, Rapid detection of incomplete coal and gangue based on improved PSPNet, *Meas.*, **201** (2022), 111646. https://doi.org/10.1016/j.measurement.2022.111646
25. B. Z. Su, W. Li, Z. Ma, R. Gao, An improved U-Net method for the semantic segmentation of remote sensing images, *Appl Intell*, **52** (2022), 3276–3288. https://doi.org/10.1007/s10489-021-02542-9
26. S. Liu, R. H. Ye, K. Jin, H. H. Cheng, CT-UNet: Context-Transfer-UNet for building segmentation in remote sensing images, *Neural Process Lett*, **53** (2021), 4257–4277. https://doi.org/10.1007/s11063-021-10592-w
27. Q. S. Yang, F. P. Wang, S. Wang, S. Y. Tang, F. J. Ning, J. Y. Xi, Detection of wheat lodging in UAV remote sensing image based on multi-head self-attention Deeplab v3+, *Trans. Chin. Soc. Agric. Mach.*, **53** (2022), 213–219. https://doi.org/710.6041/j.issn.1000-1298.2022.08.022
28. A. Belhadi, JO. Holland, A. Yazidi, G. Srivastava, J. CW. Lin, Y Djenouri, BIoMT-ISeg: Blockchain internet of medical th ings for intelligent segmentation, *Front. Physiol.*, **13** (2023). https://doi.org/10.3389/fphys.2022.1097204
29. Z. Richard, Making Convolutional networks shift-Invariant again, *Comput. Vis. Pat. Recog. (CVPR) (cs.CV); Machine Learning (cs.LG).* arXiv: 1904. 11486 [cs.CV]. https://doi.org/10.48550/arXiv.1904.11486
30. J. Hu, L. Shen, S. Albanie, G. Sun, H. E. Wu, Squeeze-and-Excitation Networks, *Trans. Pat. Anal. Mach. Intell.*, **42** (2019), 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372
31. M. K. He, Y. X. Zhang, Q. S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, *Trans. Pat. Anal. Mach. Intell.*, **37** (2015), 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824

32. C. L. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adadm, Encoder-Decoder with atrous separable convolution for semantic image segmentation, *Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018. https://doi.org/10.1007/978-3-030-01234-2_49
33. A. Howard, M. Sandler, B. Chen, J. W. Wang, C. L. Chen, X. M. Tan, et al., Searching for MobileNetV3, *Intl. Conf. Comput. Vis. (ICCV)*, Seoul, Korea, 2019. https://doi.org/10.1109/ICCV.2019.00140
34. X. B. Liu, B. Xu, W. J. Gu, Y. C. Yin, H. C. Wang, Plant leaf veins coupling feature representation and measurement method based on DeepLabV3+. *Front. Plant Sci.*, **13** (2022). https://doi.org/10.3389/fpls.2022.1043884
35. G. Máttyus, S. L. Wang, S. Fidler, U. Raquel, Hd maps: Fine-grained road segmentation by parsing ground and aerial images, *Comput. Vis. Pat. Recog. (CVPR)*, Las Vegas, USA, 2016. https://doi.org/10.1109/CVPR.2016.393