MRU-Net: A remote sensing image segmentation network for enhanced edge contour Detection

Jing Han[1], Weiyu Wang[1], Yuqi Lin[1], and Xueqiang LYU[1*]
[1] Beijing Key Laboratory of Internet Culture Digital Dissemination, Beijing Information Science and Technology University, Beijing 100101, China
[e-mail: icddtxyx@163.com]
[*]Corresponding author: Xueqiang LYU

Received July 25, 2022; revised June 19, 2023; revised August 28, 2023; accepted December 5, 2023; published December 31, 2023

Abstract

Remote sensing image segmentation plays an important role in realizing intelligent city construction. The current mainstream segmentation networks effectively improve the segmentation effect of remote sensing images by deeply mining the rich texture and semantic features of images. But there are still some problems such as rough results of small target region segmentation and poor edge contour segmentation. To overcome these three challenges, we propose an improved semantic segmentation model, referred to as MRU-Net, which adopts the U-Net architecture as its backbone. Firstly, the convolutional layer is replaced by BasicBlock structure in U-Net network to extract features, then the activation function is replaced to reduce the computational load of model in the network. Secondly, a hybrid multi-scale recognition module is added in the encoder to improve the accuracy of image segmentation of small targets and edge parts. Finally, test on Massachusetts Buildings Dataset and WHU Dataset the experimental results show that compared with the original network the ACC、mIoU and F1 value are improved, and the imposed network shows good robustness and portability in different datasets.

Keywords: Convolutional neural network, Image processing, Hybrid multiscale identification module, Micro residual structure, Remote sensing image segmentation

1. Introduce

Since important role in intelligent city construction and distribution of illegal buildings in exploration, the segmentation of remote sensing images has received more attention in recent years. How to obtain the building information contained in remote sensing images has become the focus of researchers. The ideal method of image segmentation is to simulate human vision System (HVS) to process various visual information in remote sensing images, such as edge, spectral intensity, texture and spatial relationship attributes [1], and segment the image into non-overlapping sub-regions corresponding to real objects. The feature of remote sensing image is affluent, however the large difference of terrain information, which lead to remote a great change of remote sensing image target dimension. The shape of the image segmentation of complex, edge blur, prone to errors of segmentation results, and tend to have weak light of different changes in remote sensing image, the light change makes the segmentation result is not fine and have poor edge segmentation effect[2]. How to better process remote sensing image information at model level is an urgent problem to be studied.

At present, deep convolutional neural network is widely used in computer vision tasks, natural language processing and other fields, such as FCN[3], SegNet[4], U-Net[5]and other neural networks can be applied to image segmentation. In 2015, Long propose Fully Connect Network (FCN) centering on image segmentation, which extracted deep semantic features through multiple convolution and pooling, and generates segmentation results through deconvolution and upsampling of the final feature graph directly. However, due to multiple pooling operations, the scale of feature maps is too small and the generated segmentation results are not fine enough. It influenced by FCN the deep neural network model is widely used in remote sensing image segmentation gradually. The idea of SegNet network is very similar to that of FCN network. On the left side of the network, convolutional extraction features were used to increase the receptive field through pooling; on the right side, deconvolution and sub-sampling are used to output segmentation graph, and the full connection layer is removed, which reduces the amount of computation to a certain extent. U-net network is a classical deep neural network model propose by Ronneberger[5], which was initially applied to the segmentation of two-dimensional medical images. The network consists of two symmetrical parts, namely, the encoding part and the decoding part. In the coding part, the features in the image are extracted by convolution and pooling and the feature map is gradually reduced. In this process, the number of channels is increasing, the receptive field is gradually enlarged, and the extracted features are gradually changed from low-level features to high-level features. Coding part is pooling and convolution operation of encoder to extract the characteristics of decoding, through many times of the previous layer decoding results and corresponding decoding encoder extracted features. Every time after decoding the figure size expanded, and the channel number is reduced, the last time after decoding is obtained by the convolution of a 1 x 1 layer segmentation result.

The above networks can obtain the maximum probability of each pixel in its category, and finally complete the pixel-level classification. However, they are not sensitive to the edge information of objects, and the effect of extracting buildings with obvious boundaries is bad. Moreover, complex building categories and varying scales are not conducive to image segmentation in remote sensing images. So it is difficult for convolutional neural networks to fully extract target features from deep features and shallow features to further improve the segmentation accuracy. Contour detection is a critical task in the field of

computer vision, aiming to accurately extract the edge profiles of objects from images. An efficient network for region-based accurate object detectionpropose by Guan[33],For achieving accurate object detection.

To solve the above problems, this paper designed a remote sensing image segmentation network integrating multi-scale recognition and residual connection on the basis of U-Net. In the U-Net encoder network part, add the hybrid multiscale identification module into the part. Each layer is the original model can increase the receptive field size, and will not reduce the spatial resolution of the middle figure, the characteristics of deep and shallow features and intermediate can be a very good extraction. In order to solve the problem that U-Net network is prone to gradient disappearance in the stage of remote sensing image feature extraction, which overfits the degradation model of deep neural network and leads to fuzzy segmentation of edge parts, the residual connection structure is added to the model inspired by ResNet[6]. At the same time, in order to reduce the calculation of the model, only BasicBlock structure is adopted to improve the accuracy of edge part extraction of the model.

The main contributions of the present paper include the following. At first, we solve the existing scale diversity issue in remote sensing images, put forward mix the inflation convolution module to extract context features and converge the global context information, which it better segmentation of small objects in remote sensing image. Secondly, we replace the convolutional layer structure of coding part with the BasicBlock structure in the residual connection structure, which makes it easier to obtain the deep features of the network and improves the segmentation accuracy. Finally, we apply an activation function to further improve the robustness of the proposed segmentation model, and experimental results show that our method is much better than other state-of-the-art networks.

The rest of our work is as follows. The second section introduces the main research methods of remote sensing image segmentation network. In the third section, we introduce the model of fusion multi-scale recognition module and residual connection structure proposed in this paper. In section four, we verify the validity of the proposed modular network on two basic remote sensing image datasets, and compare the proposed MRU-Net network with several popular lightweight neural networks. In section five, we summarize the network proposed in this paper.

2. Related work

At present, remote sensing image segmentation methods can be divided into traditional remote sensing image segmentation method and convolutional neural network based segmentation method.

In recent years, with the continuous application of remote sensing images in various fields, how to obtain the building information contained in remote sensing images has become the focus of researchers. In the traditional remote sensing image segmentation method, it mainly depends on artificial features, such as the local image characteristics, texture characteristics and morphological characteristics[7] to get the information from remote sensing images, such as watershed segmentation algorithm[8], graph based segmentation algorithm[9], support vector machine[10] and genetic algorithm[11] which is widely used in remote sensing image building extraction. However, remote sensing image is affected by light、atmospheric conditions、sensor quality and other aspects, so there will have a lot of noise inevitably. Traditional segmentation methods need many

manual design, weak robustness and noise has a great impact on the generated results, by feature design just solve the problem of specific data, does not have a strong generalization performance. With the progress of deep learning in computer vision, more and more remote sensing images are segmented by deep learning. Persello[12] make fully convolutional networks perform semantic segmentation of remote sensing images and obtain pixel-level segmentation results of buildings. However, due to the insensitivity of FCN to image details, there is still room for improvement in segmentation accuracy. Song [13] segmente remote sensing images using the improved SegNet network model and extracted the buildings, improved the multi-scale recognition ability of high-level features by adding an enhanced void pyramid module at the end of the encoder, and added an attention fusion module to the decoder to reduce noise interference. Wang[14] use DeepLab network to extract buildings in remote sensing images and used conditional random fields to correct network output results, effectively improving the recognition accuracy. However, this method is not end-to-end because of the need for post-processing by conditional random fields. Su[15]method U-Net to achieve end-to-end pixel-level remote sensing image segmentation, and applied integrated learning strategy to post-process the segmentation results, and finally achieved good results on the dataset of "AI Classification and Recognition Contest of CCF Satellite Images". However, this method did not combine multi-scale method. Therefore, in practical use, the scale of the object still has an impact on the recognition effect.

In order to solve the problem of classification imbalance in remote sensing image segmentation. Wong[16] propose an improved segmentation method of full convolutional neural network, in which self-made data sets were trained in the optimized FCN network, so as to integrate more local information. It can segment some crack images with uneven illumination and complex background well. However, this method has poor effect on small target recognition. In 2022, Wang[17] improved residual blocks and dense global spatial pyramid pool modules in the network, and combined them with U-NET network for training. The improved residual block extracts multilevel features while the dense global spatial pyramid pool extracts context features. In this study, the local and global information of the scene can be extracted in parallel using the processing structure to obtain a more efficient holistic approach.

The traditional methods have poor generality and the classification performance depends on the low-level features of manual screening. The application of deep learning technology in remote sensing image detection has achieved preliminary results, but further improving the accuracy of target detection and edge segmentation is still one of the main difficulties in this field. Therefore, how to better process remote sensing image information at the model level is an urgent problem to be studied.

3. Proposed model:MRU-Net

In order to better describe the remote sensing image segmentation network integrating multi-scale modules and residual connections, the network structure with residual connection structure and multi-scale recognition module used in the network are respectively described in this section.

3.1 Network Architecture

Due to the problems of small targets and unclear segmentation edge parts in image segmentation for remote sensing images[18], the main frame of this paper adopts U-Net

network, which has a good effect on image segmentation at present, and its core idea is that continuous convolutional neural network processing will lead to image size shrinkage and resolution reduction. The segmentation process is to segment remote sensing images by using U-shape encoding and decoding network. The encoding part is divided into four layers, which are composed of four encoders. Each layer extracts features through multiple downsampling and convolution. Part of extracted features are transmitted to the next layer, and the other part is directly concat to the decoding part. The decoding part is also divided into four layers, composed of four decoders The segmentation result is finally obtained by decoding the downsampling feature map of remote sensing image and the features obtained by concating receiving encoding part.

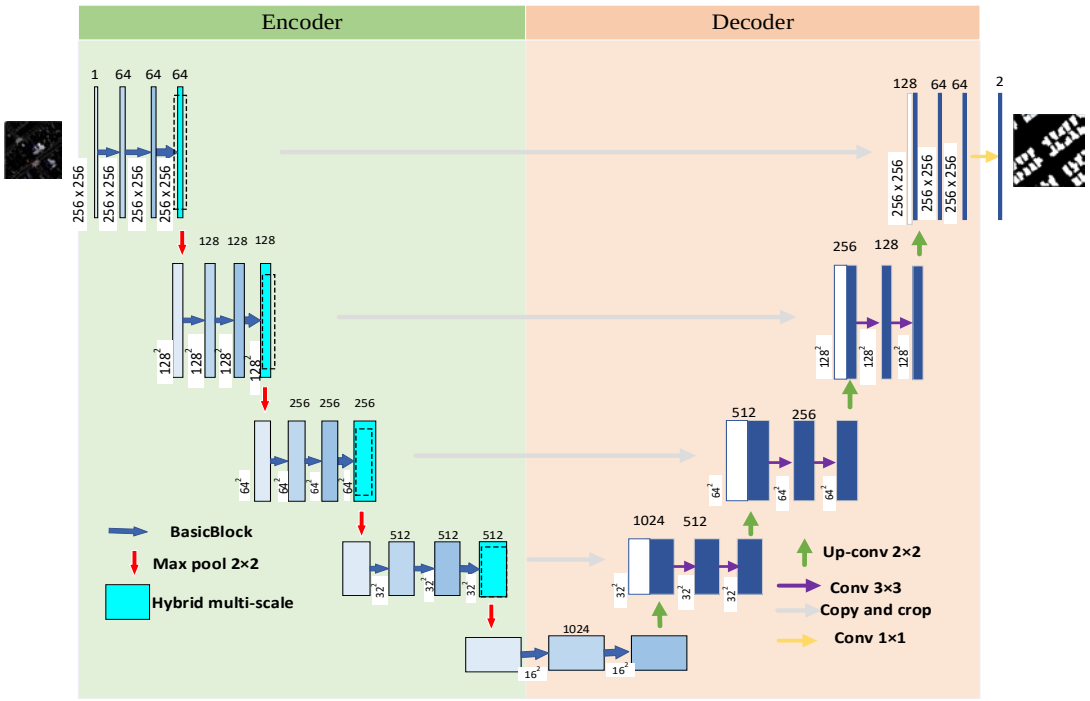


Fig. 1. Structure Diagram of MRU-Net Mode

However the commonly used remote sensing image segmentation network structure is relatively fixed. Faced with the characteristics of rich semantic information of remote sensing images, the higher emantic features are more likely to be lost, which affects the training of the network and the accuracy of feature extraction to a certain extent[19]. With the gradual deepening of the lower sampling layer and the gradual change of the mesoscale of remote sensing images. The original U-Net network only uses two convolution layers to extract features, and it cannot accurately segment remote sensing images with different scales. Therefore, there is still room for improvement in the accurate segmentation of remote sensing images using only U-Net network. As shown in the legend of **Fig. 1**.

1) Coding part: For better semantic features of different scale in remote sensing image segmentation, it need to carry on the omni-directional semantic information for different size of target feature extraction and segmentation. At the same time, to minimize the loss of the features in the process of extraction and segmentation, taking four layers will increase hybrid multiscale identification module encoder in the operation. Strengthen the extraction of different scale features in remote sensing images successively. At the

same time, the convolutional layer is replaced by BasicBlock structure in residual connection, which can reduce the computation and increase the extraction of edge features. As presented in **Fig. 2** the hybrid multiscale identification module can detect the semantic information of different scales in remote sensing image, and the micro residual structure can further extract the features of remote sensing image to avoid feature loss, so as to improve the accuracy and robustness of network in remote sensing image segmentation.

This part involves feature extraction operations of three different structures. The convolutional layer before the network is replaced with residual connection structure, and the ReLU function[20] is replaced with PReLU function[21] to normalize and activate the feature graph. At the same time, in order to better complete feature extraction, a multi-scale recognition module is added to the network except the last encoder, and the number of channels of the encoder doubles with the reduction of the feature graph, until it reaches 512 and no longer widens.

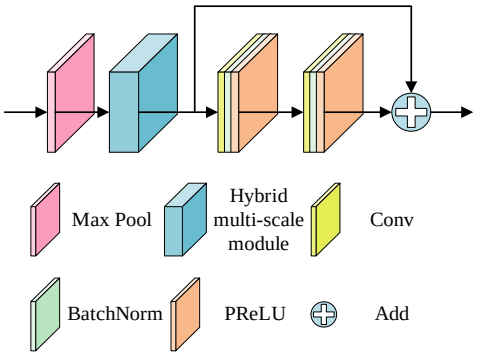


Fig. 2. Structure of encoder

2) Decoding part: **Fig. 3** shows this part also has 4 decoders, the input of each decoding operation is the previous output and the encoder result of the corresponding size. Since is the result of the splicing of two feature graphs, the number of input channels is twice that of the previous layer's output channels. As the decoding process goes on, the size of the feature graph gradually expands and the number of channels gradually shrinks. The number of output channels of the last decoder is 64. At the end of the network, a 1×1 convolution is used to map the output to the segmentation result graph of the desired category. In addition, the network adopts padding operation in the process of convolution to ensure that the output size of convolution is the same as the input size, so there is no need to perform the same clipping operation as U-Net, and the output size is the same as the input size.

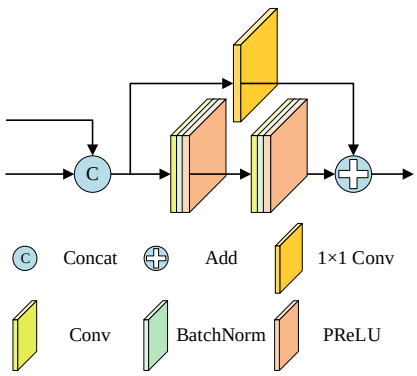


Fig. 3. Structure of decoder

3) Activation function: After the convolutional layer, in order to better extract features in remote sensing images, the network adopts batch normalization operation and activation function. The use of Batch Normalization (BN) operations[22] and activation functions effectively prevents over-fitting and thus improves the final identification accuracy. At present, the activation functions mainly include ReLU, ELU[23], PReLU and Swish[24], the PReLU activation function is adopted in this paper.

The formula of batch normalization is shown in Equation (1). The normalized operation can make the eigenvalues conform to a better distribution and avoid the change of data distribution during the training of the middle layer. At the same time, two parameters are added in this operation: scaling parameter and offset parameter, which gives the network the ability to adjust the feature distribution.

$$y = \frac{x - E(x)}{\sqrt{Var(x) - \in}} * \gamma + \beta \tag{1}$$

Where x represents a batch of input features, E(x) and Var(x) are the unbiased estimators of the mean and variance of this feature, γ is the scaling parameter, β is the offset parameter, ϵ is a minimum to prevent the occurrence of division by zero anomalies, and y is the output feature of BN layer.

PReLU activation function can be expressed by Equation (2), which is modified from ReLU, and most commonly used in the past. ReLU adjusted the negative value of the input part to 0, while the rest remained unchanged; PRelu activation function introduces learnable parameter A on the basis of ReLU function, which enables the network to respond to negative regions to a certain extent, so as to solve the "neuronal death" problem of ReLU, that is, the gradient is zero caused by negative input value, and improve the learning ability and adaptability of the network.

$$y_i = \begin{cases} x_i & if\ x_i \geq 0 \\ a_i x_i & if\ x_i < 0 \end{cases} \tag{2}$$

The x_i represents the input characteristic, y_i represents the output of PReLU, and a_i is a learnable parameter between 0 and 1.

3.2 Hybrid multiscale recongnition

Hybrid multiscale recognition(HMR) module is the basic principle of using the convolution of different inflation rate detection in the image objects of different scales, By adding a hybrid multi-scale identification module, the model can simultaneously use the feature information at multiple scales for target detection and segmentation, so as to improve the accuracy of small targets and edge components. The model can better capture the features such as the detail, texture and shape of the target, making the segmentation results more accurate and detailed. as the original U-Net network every layer of the encoder to semantic segmentation of remote sensing target feature extraction, in order to increase the accuracy of the extracted features step by step, and facilitate the final segmentation of different scale remote sensing image. Therefore, the multi-scale identification module is added to each layer of encoder.

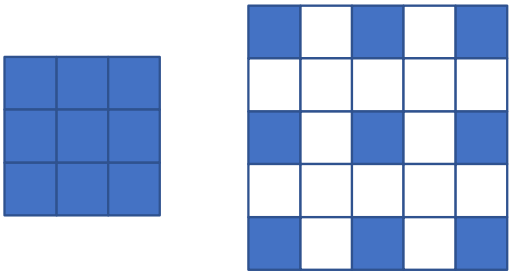
In order to segment image modules with small pixels, not only multi-scale context information is required, but also large enough output resolution is required. In convolutional neural network, feature maps of different levels and sizes can be obtained by convolution and pooling of original images. The shallow web focuses on details, while

the deep web focuses on semantic information, which helps us accurately detect targets. In order to obtain the context information of networks of different scales, the traditional neural network integrates successive sub-sampling layers, which will lose part of the resolution. Multi-scale context information can be aggregated by expansive convolution without reducing the size of the feature graph. Expansive convolution supports exponential growth of the size of the receptive field. The basic principle of the multi-scale recognition module is to detect objects of different scales in the image by applying the expansion convolution of different expansion rates. The expansion rate used by the module is related to the resolution of the input feature graph. The calculation formula of the expansion convolution is described in Equation (3). Where, y_p represents the output feature, K represents the convolution kernel size, $x_{p+r\times i,p+r\times j}$ represents the input feature graph, and $w_{i,j}$ represents the weight of the convolution kernel.

$$y_p = \sum_i^K \sum_j^K x_{p+r\times i,p+r\times j} w_{i,j} \tag{3}$$

This module is composed of several expansion convolution combinations with different expansion rates, and its output is the sum of the outputs of each expansion convolution combination [25]. In the process of extracting image features, convolution kernels of different receptive fields map features of different contents to highlight the feature diversity. Convolution kernels with large receptive fields can effectively extract features from remote sensing images with large scale, while convolution kernels with small receptive fields can analyze detailed features. During the experiment, it is found that if multiple expansion convolution with the same expansion rate is superimposed, many pixels in the sensing field are not utilized, which results in a large number of cavities. In this case, the continuity and integrity of data will be lost, which is not conducive to learning. Based on this the size of the dilatative convolution kernel used in the neural network is 3×3, and the expansion rate is 1、2、4 and 6. Therefore, the receptive field size of each branch will be 3、7、15 and 31 to further extract small targets and edge features.

Since the internal elements of the convolution kernel are adjacent in the expansion convolution, the elements of the convolution kernel are separated in the expansion convolution, and the size of the spacing depends on the expansion factor, not all pixels can be used for calculation in the feature extraction process, which will lose the continuity of information. Expansion convolution is shown in **Fig. 4**, in order to make small target feature extraction better by using different expansion coefficients, the receptive field can use more information. When the expansion rate is increased to 6 in the feature map with small size, it is found that the receptive field contained in each position has been saturated, and it is not suitable to use too large expansion rate. This design ensures that the network can have appropriate multi-scale recognition capability at different stages.



（a） (b)

((a)standard convolution kernel; (b)dilated convolutional kernel which dilate rate is 2)

Fig. 4. Introduce of dilated convolutional

3.3 Micro-residual structure

The formula of residual structure is shown in Equation (4). Where $\mathscr{F}$ consists of convolution layer, which has two forms, named BasicBlock and Bottleneck Block. The main difference is that BasicBlock uses two convolution layers of 3×3 size, whose result and input add element position. On the other hand, the 1×1 convolution layer is used to compress the channel size, and then the 3×3 convolution layer is used to extract features. Finally, the 1×1 convolution layer is used to restore the channel size, and the result is added with the input element position.

$$y = F(x,\{W_I\}) + X \tag{4}$$

represents the input, yrepresents the output, F represents the mapping of the input, and $W_i$represents the parameters required by the mapping. There will usually be multiple convolution operations, so there will be multiple parameters W.

The residual connection structure makes the optimization of remote sensing segmentation network easier and improves the convergence speed of the segmentation network. At the same time, this connection structure can alleviate the problem of gradient disappearance in deep networks, and U-Net network requires feature extraction through four-layer encoder [26]. Therefore, in this paper, the convolutional layer in the coding structure is replaced by the BasicBlock in the residual, which contains the residual branch and a short-cut branch, the structure is shown in **Fig. 5**. Micro residues address the difficulty of gradient disappearance in deep networks and the model training by adopting replacement convolution. The miniature residue structure proposes a mechanism to "skip" part of the layers and adds the skipped information with the transformed information to form the residual blocks. Compared with the traditional convolutional structure, one more short-cut branch is used to transmit low-level information, so that the network can be trained deeply. In the training process, the structure mainly uses two convolution of 3*3, and then carries out batch normalization, and then transmits the feature to the activation function. The residual between deep and shallow features of remote sensing segmentation network is extracted by replacing the original convolutional layer with BasicBlock structure, which makes it easier for the network to obtain rich deep features, so as to improve the accuracy of remote sensing image segmentation by the improved network.

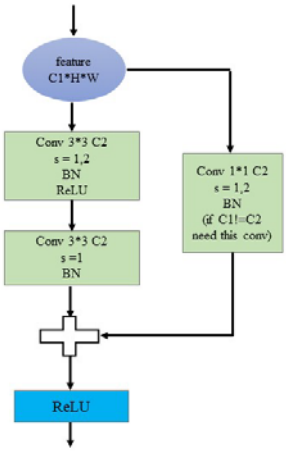


Fig. 5. Residual Connection Structure Diagram

4. Experimental analysis and comparison of results

4.1 Introduction of datasets

To verify the effectiveness of the improved network, the Massachusetts Building dataset and the WHU dataset were tested separately. The Massachusetts Buliding dataset and WHU dataset are widely used in the study of remote sensing image segmentation and edge detection. These datasets are strictly selected and processed, with certain typicality and representativeness, reflecting the common remote sensing image scenes and edge features in the real world, and can evaluate the robustness and adaptability of edge detection algorithms in different scenarios. Ensure that the experimental results are somewhat reliable and comparable. The two datasets have different image numbers and large differences in segmentation targets, so the performance of the improved network can be comprehensively analyzed from different perspectives. The details of the two datasets are described below.

4.1.1 Massachusetts Road Dataset

The Massachusetts Buildings Dataset[27] covers a variety of building segmentation datasets for remote sensing images in urban, suburban and rural areas, with only Buildings and backgrounds in the annotation. Each remote sensing image is RGB three-channel with a size of 1500*1500 pixels. The dataset includes 137 aerial remote sensing images as a training set, 4 validation sets and 10 test sets. In order to increase the robustness and generalization ability of the model and prevent over-fitting during training, this paper performs data enhancement operations such as scaling, flipping, clipping, dropout, and randomly adding noise to remote sensing image data. Since the complete image has a large resolution, the whole remote sensing image is cut with step size of 100 to a resolution of 550×550 in the training stage, and 512×512 is cut randomly as the input. The image is vertically flipped with a probability of 50%, and brightness is adjusted according to a certain random probability. Then random rotation and Gaussian filtering are performed with a certain probability. Finally, normalization is used to make the data conform to the normal distribution with a mean of 0 and variance of 1. After the above image enhancement operation, the number of dataset has been increased to 13,700 training sets,

400 verification sets and 1000 image test sets. The sample dataset is shown in **Fig. 6**, with the top three original images and the bottom three labeled images.

Fig. 6. Sample Massachusetts Dataset

4.1.2 WHU building dataset

Remote sensing image dataset WHU Building Dataset[28] extracted 220,000 buildings for Christ Church area in New Zealand, including aerial and satellite images. In this paper, only have aerial image dataset, which has high annotation accuracy. The dataset consists of 8189 remote sensing images and their corresponding labels with pixels of 512 × 512, which are randomly divided into three parts: 4736 as a training set, 1036 as a verification set and 2416 as a test set. The dataset is shown in **Fig. 7**, with the top three images as original images and the bottom three as their corresponding labels.

Fig. 7. Sample WHU Dataset

4.2 Training Details

4.2.1 Experimental equipment and environment

The experimental training environment was GUN/Linux 16.04, video card version information was Tesla V100, video memory capacity was 16G, and CUDA version was 10.0. The program environment is configured as Python 3.8, and the Deep learning framework uses the PyTorch the version is 1.4.0.

In the experiment, Softmax+ cross entropy is used as the loss function, and its formula is shown in equation (5) and equation (6). AdamW optimizer is used to optimize the

network, and the initial learning rate is set to 0.0005. Dynamic attenuation learning rate scheduling algorithm is adopted. Attenuation rate is 0.1. After testing, the effects of model initialization and model-free initialization are similar, so parameter initialization is not adopted during the experiment. During the training, the model with the best training effect will be saved and tested. Where x is the predicted result of the network and k is the number of categories. y is the marked one-hot form, and ŷ is the predicted result.

$$Soft\max(x_i) = \frac{\exp(x_i)}{\sum_{j=0}^{k-1} \exp(x_j)} \tag{5}$$

$$loss = -\sum_{i=0}^{k-1} y_i \log_2(y_i) \tag{6}$$

4.3 Evaluation Indicators

The values ACC, mIoU and F1 were used in this experiment. The expressions of ACC, mIoU and F1 are listed by Equation (7), (8) and (9). According to the following equation, ACC can only represent the proportion of correctly classified pixels to the total number of pixels, but the number of background in the actual scene is much larger than the number of buildings, so mIoU can evaluate the model effect more accurately. mIoU represents the average intersection ratio, which is the ratio of the number of correctly identified pixels of a certain category to the sum of the number of recognized pixels of the category and the actual number of pixels of the category. mIoU is the average of the intersection ratio of all categories. Therefore, mIoU can more accurately describe the accuracy of results in the case of category imbalance. Where P is the confusion matrix calculated from the predicted result and the real value, TP is the number of true cases, FN is the number of false negative cases, FP is the number of false positive cases.

$$ACC = \frac{\sum_{i=1}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \tag{7}$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ij} - p_{ii}} \tag{8}$$

$$F1 = \frac{2 \times TP}{2 \times TP + FN + TP} \tag{9}$$

4.4 Evaluation Indicators

Based on the two datasets commonly used in remote sensing image field, the improved network is tested and compared. At the same time, the effects of hybrid multiscale identification module and micro residual structure were verified, and the ablation experiments were carried out to verify the effectiveness of different modules.

4.4.1 Comparison of Massachusetts dataset results

In order to verify the effectiveness of MRU-Net, this experiment compared the current commonly used semantic segmentation model and compared it with the existing remote

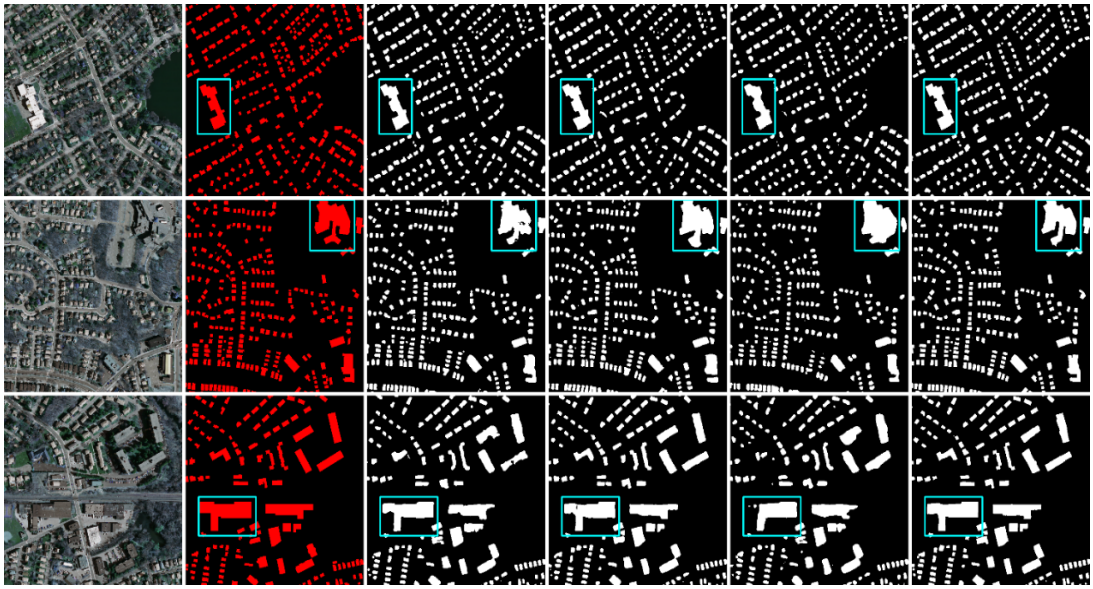
sensing image segmentation model. HRNet+OCRNet[29] focuses on the extraction and expression of context information in semantic segmentation, associating and expressing the information of a pixel through the expression of different target areas, so as to establish more discriminative context information and divide the label of a pixel and its corresponding target category. Deeplab V3+[30] network uses encoder-decoder results to extract semantic information for high-level features, and encoder extracts boundary information step by step, which improves segmentation effect while paying attention to boundary information. BBRNet[31] network is a convolutional neural network based on complex shape buildings for high resolution segmentation, which is roughly divided into two parts: prediction module and residual refinement module to improve the accuracy of building extraction. Meanwhile, Dice Loss is used to alleviate data imbalance. ENRU-Net network[32] is a segmentation network proposed based on the fine details of building structures with high-resolution images and buildings of different scales. A dense spatial pyramid pool was designed to extract dense and multi-scale features at the same time, and focal Loss was used to suppress the influence of error labels on ground truth, making the training stage more stable. **Table 1** compares the results of MRU-Net with other models.

The performance in the three accuracy indicators is better than other networks. The network for the comparison experiment is the current mainstream semantic segmentation network, and the three commonly used semantic segmentation networks, U- Net, Deeplab V3+ and HRNet+OCRNet, whichare retrained respectively. The experimental results are shown in **Table 1**.

Table 1. Comparison of Massachusetts dataset results

Model	ACC	mIoU	F1
U-Net	93.6	70.0	82.1
HRNet+OCRNet	94.9	79.7	81.0
DeepLab v3+	94.2	-	81.5
BBRNet	-	73.3	84.5
ENRU-Net	94.1	73.0	84.4
SegNet	94.3	74.1	83.7
PSPNe	94.8	75.2	84.1
UNet++	94.9	75.8	-
MRU-Net(Ours)	**95.9**	**81.2**	**84.7**

According to the data in the analysis **Table 1**, the ACC value, mIoU value and F1 value of the original network can be increased by 2.1 percentage points, 11.2 percentage points and 2.6 percentage points by the improved network. The optimization effect is significant, which confirms the accuracy of the comparison network in remote sensing image segmentation. The results show that the hybrid multi-scale recognition module plays an important role in improving the accuracy of image segmentation of small targets and edge parts. At the same time, the ACC mIoU and F1 evaluation indexes of HRNet+OCRNet, Deeplab V3+, BBRNet, ENRU-Net, PSPNet and UNet++ were compared, and the results showed that the values in different remote sensing image segmentation networks all improved. It can be seen that this network can recognize buildings well in remote sensing image segmentation.



a)Original image b)True label c)U-Net d)DeepLab V3+ e）HRNet+OCR f）Ours

Fig. 8. Experimental Results on the Massachusetts Dataset

In order to further verify the effectiveness of the improved network, part of the experimental results are shown in **Fig. 8** The image shows the effects of four cropped images. The left is the original image of the data set used in the experiment, and the red image next to it is the label image of the remote sensing image data set. Then from left to right are the comparison experiments conducted on U-Net, DeepLab V3+ and HRNet+OCRNet, and the results predicted by the MRU-Net model in this paper are at the far right. By comparing these extraction effects, it can be found that the model presented can improve the accuracy of remote sensing image segmentation to a certain extent and has a high edge contour recognition effect.

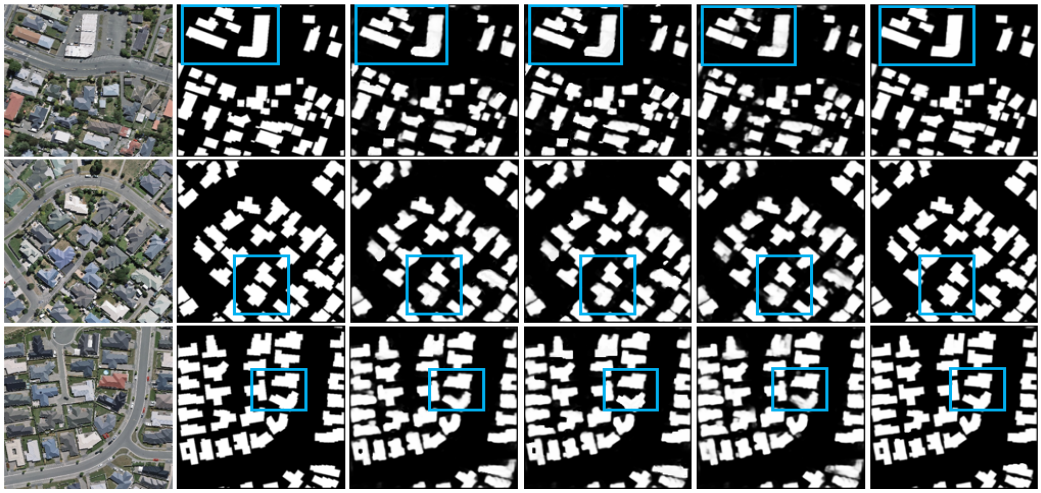
4.4.2 Comparison of WHU dataset results

In order to further verify the detection effect of MRU-Net on remote sensing image datasets, this paper uses the network to train the WHU dataset and test whether it has a good detection effect when applied to different types of datasets. The comparison results between the proposed model and the mainstream model are shown in **Table 2**. The ACC, mIoU and F1 of the proposed model are 95.8%, 68.0% and 78.9%, respectively, which are 0.5%, 1.2% and 0.9% higher than the original U-Net. In addition, the proposed model is compared with the latest models SegNet, Deeplab v3+ and UNet++, and the results show that the MRU-Net network has better effect on remote sensing image edge segmentation. By increasing the segmentation effect comparison of MRU-Net network and other advanced segmentation networks on WHU dataset, the proposed algorithm is robust and generalized in remote sensing segmentation images.

Table 2. Results of WHU dataset results

Model	ACC	mIOU	F1
U-Net	95.3	66.8	78.0
Deeplab V3+	95.6	67.3	78.0
HRNet+OCRNet	94.9	65.1	78.4
SegNet	95.4	67.1	78.2
PSPNet	95.3	67.3	78.5
UNet++	95.6	67.6	78.6
MRU-Net(Ours)	**95.8**	**68.0**	**78.9**

The segmentation results are shown in **Fig. 9**. It can be seen from the figure that compared with the original network, the improved network in this paper has slightly improved the overall definition of remote sensing image and the accuracy of image edge detection. Although the effect on WHU dataset is not very good, but the generalization ability is strong, and there is some improvement regardless of the type of dataset. The experimental results show that the improved network can clearly observe the optimization results when using the remote sensing image segmentation proposed in this paper, which proves the effectiveness and generalization of the improved network.



a)Original image b)true label c)U-Net d)DeepLab V3+ e)HRNet+OCR f)Ours

Fig. 9. Experimental Results on the WHU Dataset

4.4.3 Ablation results

To verify the effectiveness of the proposed MRU-Net and each module of the network. The proposed network model is a U-shaped network based on the encoding and decoding structure, so U-Net is chosen as the experimental baseline model. The ablation experiments were trained networks with Massachusetts datasets and WHU datasets, evaluated the segmentation results using the test set test and measured with the evaluation indexes ACC, mIoU and F1. As can be seen from **Table 3** and **Table 4**, the improved U-Net network improves over the baseline U-Net; adding the micro residual structure (MRS) improves the performance of the network and alleviates gradient disappearance during the training process; adding the hybrid multi-scale recognition (HMR) to improve the

accuracy of image segmentation of small targets. Finally, both MRS and HMR modules are added to the baseline network, that is, the proposed U-Net, and the detection results are better than using one of the modules alone, that is, both MRS and HMR modules improve the segmentation accuracy of buildings.

Table 3. Ablation experiments of Massachusetts dataset results

Model	ACC	mIoU	F1
U-Net	95.2	79.7	80.0
Improved U-Net	95.3	80.0	82.5
with HMR	95.3	80.2	83.7
with MRS	95.3	80.0	83.3
HMR+MRS	95.4	80.9	83.8
MRU-Net	**95.5**	**81.2**	**84.7**

Table 4. Ablation experiments of WHU dataset results

Model	ACC	mIoU	F1
U-Net	95.3	78.0	66.8
Improved U-Net	95.1	78.1	65.7
with HMR	94.9	78.2	65.0
with MRS	95.0	78.1	65.0
HMR+MRS	95.2	79.5	67.3
MRU-Net(Ours)	**95.5**	**80.2**	**68.2**

5. Conclusion

Due to the multifarious types of ground objects in remote sensing images, the scale of different types of targets varies greatly, and the detection effect of small target edge part is bad. In this paper, we are proposed to improve the multi-scale recognition capability by integrating micro-residual structure in the network and using hybrid multi-scale module in the coding stage based U-Net network, and finally solve the above problems effectively. Experimental results show that the improved U-Net network proposed in this paper has better effect and strong generalization than the current mainstream network, and the addition of ablation experiment also proves that this network can effectively improve the segmentation effect of remote sensing images. The experimental results show that the improved U-Net network proposed in this paper is better and stronger than the current mainstream network. At the same time, the added ablation experiment also proves that this network can effectively improve the segmentation effect of remote sensing images. However, in our work, there is still room for improvement in the training time and inference time of the network. In the future, we will try to use a more lightweight feature extraction network, focusing on solving the problem of inaccurate edge feature capture in the lightweight network. In addition, we will try to apply the research results to the urban road flow monitoring and management system to further promote the construction of smart city.

Acknowledgement

This work was supported in part by The National Natural Science Foundation of China (62171043), Beijing Natural Science Foundation of China (4212020) and Scientific Research Project of National Language Commission(ZDI145-10).

References

[1] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. of the 15th European Conference onComputerVision(ECCV2018)*, Munich, GERMANY, pp.833-851, 2018. Article (CrossRef Link)

[2] J. C. Wang, L. Shen, W. F. Qiao, Y. S. Dai and Z. L. Li, “Deep Feature Fusion with Integration of Residual Connection and Attention Model for Classification of VHR Remote Sensing Images,” *Remote Sensing*, vol. 11, no. 13, pp. 1617, 2019. Article (CrossRef Link)

[3] E. Shelhamer, J. Long, T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640-651, 2017. Article (CrossRef Link)

[4] V. Badrinarayanan, A. Kendall, 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. Article (CrossRef Link)

[5] O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of MICCAI 2015*, Munich, GERMANY, pp. 234-241, 2015. Article (CrossRef Link)

[6] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in *Proc. of CVPR 2016*, Las Vegas, NV, USA, pp. 770-778, June 2016. Article (CrossRef Link)

[7] Dunaeva AV, Kornilov FA, “Specific shape building detection from aerial imagery in infrared range,” *Vestnik Yuzhno-Ural'skogo Gosudarstvennogo Universiteta. Seriya" Vychislitelnaya Matematika i Informatika"*, vol. 6, no. 3, pp. 84-100, 2017. Article (CrossRef Link)

[8] Q. Guo, Y. Wang, S. Yang and Z. Xiang, “A method of blasted rock image segmentation based on improved watershed algorithm,” *Scientific Reports*, vol.12, no.1, pp. 1-21, 2022. Article (CrossRef Link)

[9] X. W. Zhao, J. Y. Liang, and J. Wang, “A community detection algorithm based on graph compression for large-scale social networks,” *Information Sciences*, vol. 551, pp. 358-372, 2021. Article (CrossRef Link)

[10] X. Wang, S. Wang, Y. Zhu and X. Meng, “Image segmentation based on support vector machine,” in *Proc. of 2012 2nd International Conference on Computer Science and Network Technology*, IEEE, pp. 202-206, 2012. Article (CrossRef Link)

[11] L. Khrissi, N.El Akkad, H. Satori, and K. Satori, “Image Segmentation based on k-means and genetic algorithms,” in *Proc. of Embedded Systems and Artificial Intelligence: Proceedings of ESAI 2019*, Fez, Morocco, pp. 489-497, 2020. Article (CrossRef Link)

[12] C. Persello, A. Stein, “Deep fully convolutional networks for the detection of informal settlements in VHR images,” *IEEE geoscience and remote sensing letters*, vol.14, no.12, pp. 2325-2329, 2017. Article (CrossRef Link)

[13] Song Tingqiang, Li Jixu and Zhang Xinye, “Building Recognition in High-Resolution Remote Sensing Image Based on Deep Learning,” *Computer Engineering and Applications*, vol.56, no.8, pp. 26-34, 2020. Article (CrossRef Link)

[14] M. Wang, M. Wang, G. Yang, et al., “Remote Sensing Image Building Extraction Method Based on Deep Learning,” *Journal of Physics: Conference Series*, IOP Publishing, vol.1631, no.1, pp. 012010, 2020. Article (CrossRef Link)

[15] Su Jianmin, Yang Lanxin and Jing Weipeng., “U-Net Based Semantic Segmentation Method for High Resolution Remote Sensing Image,” *Computer Engineering and Applications*, vol.55, no.7, pp.207-213, 2019. Article (CrossRef Link)

[16] P. Weng, Y. H. LU, X. B. QI and S. Y. Yang, "Pavement crack segmentation technology based on improved fully convolutional network," *Computer Engineering and Applications*, vol.55, no.16, pp.235-239, 2019. Article (CrossRef Link)
[17] H. Wang, F. Miao, "Building extraction from remote sensing images using deep residual U-Net," *European Journal of Remote Sensing*, vol.55, no.1, pp.71-85, 2022. Article (CrossRef Link)
[18] M. S. Wang, M. C. Wang, G. D. Yang and Z. W. Liu, "Remote Sensing Image Building Extraction Method Based on Deep Learning," *Journal of Physics: Conference Series*, vol.1, no.1, pp. 012010, 2020. Article (CrossRef Link)
[19] W. P. Jing, M. W. Zhang, D. X. Tian, "Improved U-Net model for remote sensing image classification method based on distributed storage," *Journal of Real-Time Image Processing*, vol. 18, no.5, pp. 1607-1619, 2021. Article (CrossRef Link)
[20] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun; "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proc of ICCV*, pp.1026-1034, 2015. Article (CrossRef Link)
[21] X. Glorot, A. Bordes, Y. Bengio, "Deep Sparse Rectifier Neural Networks," *Journal of Machine Learning Research*, vol.15, no.2, pp. 315-323, 2011. Article (CrossRef Link)
[22] S. Loffe, C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of International conference on machine learning*, vol. 37, pp. 448–456, July. 2015. Article (CrossRef Link)
[23] D. A. Clevert, T. Unterthiner, S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," in *Proc. of ICLR 2016*, 2015. Article (CrossRef Link)
[24] P. Ramachandran, B. Zoph, Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017. Article (CrossRef Link)
[25] M. Mubashar, H. Ali, C. Grönlund, S. Azmat, "R2U++: a multiscale recurrent residual U-Net with dense skip connections for medical image segmentation," *Neural Computing and Applications*, vol. 34, pp. 17723-17739, 2022. Article (CrossRef Link)
[26] S. Wang, X. Hou, X. Zhao, "Automatic Building Extraction From High-Resolution Aerial Imagery via Fully Convolutional Encoder-Decoder Network With Non-Local Block," *IEEE Access*, vol. 8, pp. 7313-7322, 2020. Article (CrossRef Link)
[27] Mnih, Volodymyr, "Machine learning for aerial image labeling," *University of Toronto*, Canada, 2013.
[28] S. Ji, S. Wei, L. Meng, "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, 2018. Article (CrossRef Link)
[29] J. Wang, K. Sun, T. Cheng, B. Jiang and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349-3364, 2021. Article (CrossRef Link)
[30] L. C. Chen, Y. K. Zhu, G. Papandreou, F. Schroff and H. Adam, "Encoder-decoderwith atrous separable convolution for semantic image segmentation," in *Proc of ECCV 2018*, pp. 833-851, 2018. Article (CrossRef Link)
[31] Z. F. Shao, P. H. Tang, Z. Y. Wang, Saleem, Nayyer, Yam, Sarath, Sommai and Chatpong, "BRRNet: A Fully Convolutional Neural Network for Automatic Building Extraction From High-Resolution Remote Sensing Images," *Remote Sensing*, vol.12, no.6, pp.1050, 2020. Article (CrossRef Link)
[32] S. Wang, X. Hou and X. Zhao, "Automatic Building Extraction From High-Resolution Aerial Imagery via Fully Convolutional Encoder-Decoder Network With Non-Local Block," *IEEE Access*, vol. 8, pp. 7313-7322, 2020. Article (CrossRef Link)
[33] Y. Guan, M. Aamir, Z. Hu, et al., "A Region-Based Efficient Network for Accurate Object Detection," *Traitement du Signal*, vol.38, no.2, pp. 481-494, 2021. Article (CrossRef Link)

Jing Han, born in Handan City, Hebei Province in 1990, she received her Ph.D. degree from the University of Science and Technology Beijing in 2020 and has been working at Beijing Information Science and Technology University since 2020.Her research interests include object detection, semantic segmentation, etc.

Weiyu Wang, born in Shangqiu City, Henan Province in 1997, she received a bachelor's degree from Beijing Information Science and Technology University in 2022. Her main research interests are target detection, semantic segmentation, etc.

Yuqi Lin, born in 1996, graduated from Shangqiu, Henan province, he received a bachelor's degree from Beijing Information Science and Technology University in 2021. his main research interests are target detection, semantic segmentation, etc.

Xueqiang LYU, born in Fushun City, Liaoning Province in 1970, he received his Ph.D. from Northeastern University in 2003 and has been working at Beijing Information Science and Technology University since 2005.His research interests include multimedia information processing, computer vision, etc.