Research on remote sensing image extraction based on deep learning

Zhao Shun [Equal first author, 1] , **Danyang Li** [Equal first author, 1] , **Hongbo Jiang** [1] , **Jiao Li** [1] , **Ran Peng** [1] , **Bin Lin** [1] , **QinLi Liu** [1] , **Xinyao Gong** [1] , **Xingze Zheng** [1] , **Tao Liu** [Corresp. 1]

[1] Sichuan Agricultural University, College of Information Engineering, Yaan, Sichuan, China

Corresponding Author: Tao Liu
Email address: liutao@sicau.edu.cn

Remote sensing technology has the advantages of fast information acquisition, short cycle, and a wide detection range. It is frequently used in surface resource monitoring tasks. However, traditional remote sensing image segmentation technology cannot make full use of the rich spatial information of the image, the workload is too large, and the accuracy is not high enough. To address these problems, this study carried out atmospheric calibration, band combination, image fusion, and other data enhancement methods for Landsat 8 satellite remote sensing data to improve the data quality. In addition, deep learning is applied to remote-sensing image block segmentation. An asymmetric convolution-CBAM (AC-CBAM) module based on the convolutional block attention module is proposed. This optimization module of the integrated attention and sliding window prediction method is adopted to effectively improve the segmentation accuracy. In the experiment of test data, the mIoU, mAcc, and aAcc in this study reached 97.34%, 98.66%, and 98.67%, respectively, which is 1.44% higher than that of DNLNet (95.9%). The AC-CBAM module of this research provides a reference for deep learning to realize the automation of remote sensing land information extraction. The experimental code of our AC-CBAM module is https://github.com/LinB203/remotesense.

1 Research on Remote Sensing Image Extraction

2 Based on Deep Learning

3 Shun Zhao[1, #], Danyang Li[1, #], Hongbo Jiang[3], Jiao Li[4], Ran Peng[5], Bin Lin[6], Qinli Liu[7], Xinyao

4 Gong[8], Xingze Zheng[9], Tao Liu[*]

7 [1] College of Information Engineering, Sichuan Agricultural University, Yaan, Sichuan, China

8 [1] College of Information Engineering, Sichuan Agricultural University, Yaan, Sichuan, China

9 [2] College of Information Engineering, Sichuan Agricultural University, Yaan, Sichuan, China

10 [3] College of Information Engineering, Sichuan Agricultural University, Yaan, Sichuan, China

11 [4] College of Information Engineering, Sichuan Agricultural University, Yaan, Sichuan, China

12 [5] College of Information Engineering, Sichuan Agricultural University, Yaan, Sichuan, China

13 [6] College of Information Engineering, Sichuan Agricultural University, Yaan, Sichuan, China

14 [7] College of Information Engineering, Sichuan Agricultural University, Yaan, Sichuan, China

15 [8] College of Information Engineering, Sichuan Agricultural University, Yaan, Sichuan, China

17 [#] These authors contributed equally to this work.

18 [*] Corresponding Author

20 No. 46, Xinkang Road, Yucheng District, Yaan, Sichuan, 625014, China

21 Email address: liutao@sicau.edu.cn

22 Abstract

23   Remote sensing technology has the advantages of fast information acquisition, short cycle,

24 and a wide detection range. It is frequently used in surface resource monitoring tasks. However,

25 traditional remote sensing image segmentation technology cannot make full use of the rich spatial

26 information of the image, the workload is too large, and the accuracy is not high enough. To

27 address these problems, this study carried out atmospheric calibration, band combination, image

28 fusion, and other data enhancement methods for Landsat 8 satellite remote sensing data to improve

29 the data quality. In addition, deep learning is applied to remote-sensing image block segmentation.

30 An asymmetric convolution-CBAM (AC-CBAM) module based on the convolutional block

31 attention module is proposed. This optimization module of the integrated attention and sliding

32 window prediction method is adopted to effectively improve the segmentation accuracy. In the

33 experiment of test data, the mIoU, mAcc, and aAcc in this study reached 97.34%, 98.66%, and
34 98.67%, respectively, which is 1.44% higher than that of DNLNet (95.9%). The AC-CBAM
35 module of this research provides a reference for deep learning to realize the automation of remote
36 sensing land information extraction. The experimental code of our AC-CBAM module is
37 https://github.com/LinB203/remotesense.
38 **Keywords:** attention mechanism, automatic extraction, band fusion, sliding window prediction,
39 semantic segmentation

Introduction

41 Semantic segmentation is critical tasks when processing remote sensing images. It is used to
42 segment objects on the earth, such as vegetation, roads, buildings, water bodies, and other objects.
43 Its main purpose is to extract semantic information effectively and divide the studied image pixels
44 into a number of object areas. Pixels belonging to the same category are classified as a specific
45 homogeneous area while ensuring heterogeneity between different object areas. Traditional
46 semantic segmentation methods are based on geometry and statistics, and before deep learning
47 methods became popular, semantic segmentation methods based on traditional machine learning
48 classifiers such as Texton Forest [1] and Random Forest [2] were more commonly used. More
49 recently, many models based on level sets [3,4,5], deep learning [6,7,8,9,10], and Markov random
50 field models have been proposed [11,12,13,14,15].
51 In recent years, many techniques and prior information have been added to improve the
52 accuracy of semantic segmentation technology. In particular deep learning technology has greatly
53 improved the performance of remote sensing image classification, remote sensing image retrieval,
54 remote sensing image target detection, and other related technologies. Therefore, based on deep
55 learning, the semantic segmentation of remote sensing images can be reasonably performed, the
56 label of each pixel in the image can be predicted, and then be accurately classified. Long et al. [16]
57 proposed a fully convolutional network (FCN) in 2014, applying deep learning to semantic
58 segmentation for the first time, achieving end-to-end, pixel-to-pixel training, and combining
59 traditional convolutional neural networks. The fully connected layer in the convolutional neural
60 network (CNN) [17] is transformed into a convolutional layer, and the fully connected layer is
61 removed to realize the input of any size image. The encoding part only obtains the feature layer
62 through convolution and pooling processing, and the decoding part restores the feature map
63 obtained by the last convolution layer through deconvolution and upsampling, thereby achieving
64 pixel-level semantic segmentation. However, the use of FCN-based semantic segmentation has
65 many problems such as chaotic sampling structure, loss of object details, or smoothness in the
66 receptive field by fixed segmentation.
67 Therefore, in 2015, Badrinarayanan et al. proposed the SegNet algorithm [18]. It uses a
68 symmetrical structure to improve accuracy and efficiency, and uses pooling indices to save the

69   contour information of the image, reducing the number of parameters. Chen et al. proposed the
70   Deeplab algorithm and continuously optimized it on this basis. The v1 series [19] not only
71   considers the DCNN output but also considers the pixel value around the pixel when classifying
72   each pixel so that the boundary of the semantic segmentation result is clear. To solve the problems
73   caused by insufficient accuracy, pooling, and down-sampling repetition, the V2 series [20] chose
74   to design a porous spatial pyramid pooling (ASPP) module based on the cavity convolution
75   algorithm to expand the receptive field. V3 [21] further improved the ASPP module. There is also
76   a v3+ series [22]. One approach is to design a decode module based on v3 to refine the result, the
77   other is to use modified Xception as the backbone. These are semantic segmentations based on the
78   encoder-decoder (up-sampling/deconvolution) structure. Gong et al. [23] proposed a new training
79   method, that was realized by adding a metric learning regularization term. Gong et al. [24] used
80   context coding loss for canonical model training and proposed a novel feature enhancement
81   network for detection. In addition, semantic segmentation models based on attention have been
82   developed to obtain context information, such as non-local neural network[25] and DANet [26].
83   These deep learning-based semantic segmentation algorithms can also be continuously optimized
84   by ResNet [27] to continuously improve segmentation performance.
85   In this study, the method of introducing the attention mechanism into the semantic
86   segmentation model is adopted to improve the accuracy. The attention mechanism was first
87   proposed in computer vision. In 2014, the Google Mind team used the attention mechanism [28]
88   on the RNN model for image classification. Since then, the attention mechanism has been widely
89   used in computer vision fields such as target detection and semantic segmentation.
90   Subsequently, Bahdanau et al. used a mechanism similar to that used to simultaneously
91   perform translation and alignment on machine translation tasks. Their work is regarded as the first
92   to apply the attention mechanism to the field of NLP [29].
93   Attention mechanism is widely used in various NLP tasks based on neural network models
94   such as RNN/CNN. In 2017, Google used the self-attention mechanism to learn text
95   representations [30]. Since then, the attention mechanism has become a research hotspot.
96   Aiming at the object classification and recognition of remote sensing images, this article
97   addresses the problems of traditional remote sensing image extraction such as low accuracy, poor
98   generalization ability, and low degree of automation. The original AC-CBAM module is created,
99   and a full convolution based on the DNLNet encoding and decoding structure is constructed. A
100 neural network is used for remote sensing image extraction, aiming to realize the precision and
101 intelligence of remote sensing image extraction. Our main contributions are as follows:
102   (1) Based on Landsat 8 OLI sensor data, this study preprocessed the image band. The
103 preprocessed image is combined with the original image to form a 4-channel image as the input.
104   (2) Based on the requirements of practical and experimental tasks, a new convolution
105 attention module, AC-CBAM, is proposed in this paper. The innovative module combines
106 asymmetric convolution and CBAM, which has better effect than CBAM and mainstream
107 segmentation network in object classification and segmentation of remote sensing images.

108   (3) Besides proposing the AC-CBAM module, this study analyzes how to effectively extract
109   its features from the things, so that accuracy can be improved without a large number of search
110   and experimental network structures. By analyzing the features of categories, in this study water
111   and roads were found to have unique shape features, and, therefore, the idea of extracting features
112   by using the uniqueness of asymmetric convolution is proposed.

113   (4) We use sliding window prediction that ignores the edges to predict the entire remote
114   sensing image, avoiding the problem of obvious splicing traces after the conventional grid cut
115   prediction. Experiments show that the entire remote sensing image has a good prediction result.

Data preprocessing

117   This chapter introduces the necessity and methods of remote sensing image preprocessing,
118   gives the preprocessing results, and describes the difficulties and concerns of remote-sensing
119   image block segmentation. Second, through Gaussian filtering, image fusion, and band
120   combination, the image quality and parcel information are enhanced, and the remote sensing image
121   parcel dataset is prepared, which lays a good data foundation.

2.1 Land parcel identification process

123   Remote sensing image classification technology provides a means to observe and analyze
124   specific phenomena in the fields of commerce, ecology, forestry, and urban and regional planning.
125   It has a wide application prospects in these fields. Land parcel segmentation using in-depth
126   learning technology for remote sensing images is an important research topic in all fields. Precise
127   land parcel extraction results will provide further research in all fields. This has great significance.
128   In a real application scenario, the result of land parcel information extraction is highly required,
129   but there are many interference factors in the land parcel information extraction of remote sensing
130   images. For example, there is more cloud cover in the remote sensing image, and some areas have
131   distinct brightness differences on the visual level, different bands perform differently on water,
132   and in some bands, land and water have the same visual performance, as well as the shadows of
133   land objects, which all increase the difficulty of land parcel extraction. If the remote sensing image
134   is simply preprocessed, instead of enhancing and highlighting the land parcel information, a large
135   number of pixel misclassifications will often occur. This causes unsatisfactory land extraction
136   results and fails in practical application scenarios. In addition to the above interference factors, the
137   remote sensing image itself has more complex and diverse feature information. Classifying water
138   bodies, vegetation, buildings, roads and other objects from these complex feature information not
139   only requires many human and material resources, but also makes it difficult to ensure extraction
140   accuracy. For efficiency and accuracy of land parcel extraction, this study excludes the ground
141   object information outside water, vegetation, buildings and roads. It only focuses on the land parcel
142   information of remote sensing images, simplifies the task of remote sensing image recognition,
143   focuses on the land parcel extraction of remote sensing images, and simplifies the multi-
144   classification problem of remote sensing images into five classification problems of water,
145   vegetation, buildings, roads and others while improving the efficiency and reduces the workload
146   of human visual interpretation, and improves the efficiency and accuracy of remote sensing image
147   block extraction.
148   The neural network obtains the global optimal weight matrix by forward error conduction

149 and reverse weight updating. The trained matrix is used to predict the remote sensing image and
150 obtain the final land parcel prediction result. Remote sensing land parcel identification is a pixel-
151 level classification task, which classifies each pixel point into five categories: water, vegetation,
152 buildings, roads, and others. The specific process of land parcel identification is as follows: (1) the
153 remote sensing image is input into the trained network model; (2) the output of the model is the
154 confidence level at which the pixel points belong to five categories, and the pixel point category
155 is determined according to the confidence level; (3) the color is filled according to the decision
156 category of the pixel points to achieve five classifications of the land masses of the remote sensing
157 image; (4) The classification results are smoothed and denoised by the median filter to obtain the
158 final remote sensing land mass prediction results. The output of the two-dimensional median filter
159 is:

160
161 $$g(x, y) = \mathrm{med}\{f(x - k, y - l), (k, l \in W)\} \#(1)$$
162
163   Where $f(x, y)$ and $g(x, y)$ are the original image and the processed image respectively.

2.2 Study Area

165   We chose Kunming in southwestern China as our research city. The longitude and latitude
166 of Kunming are 102°10' ~ 103°40' E and 24°23' ~ 26°22' N. The research area includes three large
167 lakes: Dianchi Lake, Fuxian Lake and Yangzonghai Lake. In this study, the Minjiang River and
168 part of the Jinsha River water bodies in Leshan City were selected. The longitude and latitude were
169 102°15' ~ 104°15 'E and 28°28 ~ 29°56' N. The mixed complex region mainly refers to the region
170 with complex block combinations, with latitude and longitude ranging from 115°47' to 116°45' E
171 and 28°22 'to 29°45' N. Blocks in this region were used to test the comprehensive extraction ability
172 of the model. Poyang Lake in southeast China was selected as the study area. The relevant
173 information is presented in Table 1.
174   The data used in this study are from the Landsat 8 satellite data. Landsat 8 remote sensing
175 satellite was launched by NASA in 2013. Compared with Landsat 7, Landsat 8 only adjusts the
176 bands of the data, eliminating the effects of atmospheric scattering and absorption, water vapor
177 and other factors on the data. In addition, two new bands, blue band(band 1 0.433-0.453 μm) and
178 short-band infrared band(band 9 1.360 − 1.390 μm), have been added. The original image of
179 remote sensing data is shown in Figure 1.
180   Landsat 8 satellite data are widely used in geological exploration, environmental
181 management, resource assessment and other domains, providing a large amount of remote sensing
182 observational data for various disciplines and related research. It has a total of eleven bands,
183 including eight multispectral bands, one panchromatic band and two thermal infrared bands, which
184 can address different application scenarios. The main parameters are listed in Tables 2 and 3.
185   In this study, the red, green, and near-infrared bands were selected from Landsat 8 OLI data
186 and fused with 15-meter panchromatic band to enhance resolution, covering three research areas.
187 In this study, automatic extraction and manual verification were used to complete the plot labeling.
188 The image label after manual interpretation was accurate enough, so we took the result of plot
189 marking as the mapping label of the sample image and the baseline of the test image.

2.3 Image preprocessing

191   Remote sensing image preprocessing involves eliminating and suppressing some influences
192   and errors in imaging, which mainly come from external conditions during imaging, such as the
193   satellite sensor itself, absorption and scattering of ultraviolet radiation by atmosphere, solar
194   altitude angle, earth surface topography and other factors. To obtain real reflectance data, methods
195   such as radiometric calibration and atmospheric correction are used to preprocess remote sensing
196   images to obtain real image data. In addition, in order to obtain richer remote sensing information
197   and higher-quality images, image enhancement is first carried out on remote sensing images to
198   highlight image edges and improve the overall quality of data products, thus laying a data
199   foundation for subsequent data set division, thus improving the accuracy of plot discrimination.
200   A Gaussian filter is a linear smoothing filter that is widely used in image processing. It is
201   mainly used to suppress noise and smooth images. Because noise is transmitted, it is a severe
202   problem in the later application of digital images. Then most of the noise in the image is Gaussian
203   noise, so the Gaussian filter is widely used in image denoising.
204   For remote sensing classification tasks, the selected size of Gaussian kernel is 7 × 7, the
205   average pixel weight of the Gaussian kernel is larger, and the surrounding pixels are smaller, which
206   reduces the importance of surrounding pixels and reduces image blur during anti-aliasing. The
207   main function of Gaussian filtering is to smooth the image. Based on this, the original image with
208   twice the size can be subtracted from the image after Gaussian filtering, and the resulting image
209   has more obvious block edges.
210   Means of improving the image quality based on existing Landsat 8 data must be considered
211   because the resolution of the satellite data is limited. In the data, the resolution of the panchromatic
212   band is higher than that of multispectral data, so image fusion is used to fuse multispectral data to
213   improve the image resolution. Brovey transformation fusion, which is a simple and effective
214   method, is mainly used for the process. The processed image not only highlights the spatial
215   resolution of the panchromatic band data, but also highlights the differences in the features of the
216   objects in different bands.
217   While being generated, remote sensing images will be distorted to a certain extent owing to
218   the influence of the satellite-sensor tilt angle, atmospheric scattering, water vapor, and other
219   factors. To produce remote sensing images as realistically as possible, certain methods must be
220   used to preprocess them to inhibit and eliminate interference. The image contents and different
221   research objectives require different pretreatment methods.

2.3.1 Radiometric Calibration and Atmospheric Correction

223   Remote sensing technology receives electromagnetic wave information reflected by various
224   objects on the Earth surface from outer space and carry out attribute analysis and imaging. The
225   satellite sensor receives the solar radiation reflected from the Earth surface, and the final original
226   image is a synthesis of all information due to atmospheric scattering, water vapor absorption,
227   surrounding objects and other factors. The process of atmospheric correction is the process of
228   separating specific spectral information from the study an object's attributes. Because there are
229   errors in sensor imaging, these errors need to be eliminated when using remote sensing images for
230   analysis and research, which is the process of radiation correction. The remote sensing data
231   processing software used in this study is ENVI 5.3, through which remote sensing images can be
232   atmospherically corrected. The ENVI FLAASH module was used for the atmospheric correction
233   of remote sensing images after radiometric calibration.

2.3.2 Remote Sensing Image Enhancement

The main purpose of remote sensing image enhancement is to enhance the image quality and information, so that the plot information can be observed more clearly and intuitively under the observation of human eyes, which plays a crucial role in the discrimination of plots. Common remote sensing image enhancement methods include image synthesis, numerical filtering and contrast stretching. Different enhancement methods have been adopted to match diverse research content and methods. In this study, numerical filtering was used to enhance image quality, band combination was used to enhance image block information, and the effect of block discrimination is strengthened.

(1) Image enhancement in spatial domain. In this study, a Gaussian low-pass filter was used to denoise and smooth the image. A low-pass filter is mainly used to carry out two-dimensional convolution calculations through a Gaussian checking image.

7×7 Gaussian kernel is adopted in this paper, and the specific structure is shown in Figure 2.

Gaussian low-pass filtering processing flow:

(a) Reading of various bands required.

(b) The selected band is processed by the Gaussian filter.

(c) The Gaussian kernel is moved across each pixel in the image, and the pixel value of the corresponding pixel is convolved by the kernel.

(d) The calculated value is taken as the final pixel value.

The actual processing results are shown in Figure 3.

(2) Band combination image enhancement. Generally, color images are RGB false color images. In remote sensing image classification, different RGB components are selected according to different classification targets to combine new standard false color images. Landsat8 has many different bands, and the combination of different bands also has different uses. The specific band combination scheme and uses are shown in Table 4.

For Landsat 8 OLI remote sensing products with atmospheric correction, which provide a variety of band data, including panchromatic, near-infrared, and green bands, we can easily produce standard false-color image data. The green, red, and infrared bands of the remote sensing data are assigned to the blue, green and red band in RGB respectively. It is well known that the colors of vegetation and water are similar in unprocessed remote sensing images. By synthesizing standard false color images, vegetation will be predominantly red, water will be green, blue, dark blue, and so on, depending on the number of microorganisms inside. One of the best benefits of converting remote sensing images to standard false color images is to distinguish vegetation from the color of water. According to the above description, we use five (NIR), four (Red) and three (Green) bands to enhance the band combination of remote sensing images and enhance the land parcel information. The results are shown in Figure 4.

2.3.3 Image fusion

Because the resolution of multispectral data is low, and the panchromatic band has a better resolution than multispectral data the panchromatic band was fused with the selected multispectral band. The new images have improved resolution, retain their multispectral characteristics, and the image quality and performance are enhanced.

There are information differences among the satellite sensors. To eliminate such differences, composite models are generally used to integrate data from different sensors to obtain clearer images and more pronounced ground object information. This is more conducive to remote sensing

279 classification. In the case of limited existing data, it is of great practical significance to make full
280 use of existing data to extract ground object information from remote sensing images. In a real
281 scene, data from different sources fill the entire experimental stage, and data from different sources
282 have different advantages and disadvantages. Image fusion can effectively complement the
283 advantages of data and compensate for the shortcomings of a single data source.
284   Landsat8 satellite multi-spectral data resolution is 30 m, band 8 is panchromatic and the band
285 resolution is 15 meters, the selection of multi-spectral data band 5 (near infrared), band 4 (red),
286 band 3 (green), and panchromatic band image fusion can effectively improve image clarity and
287 data quality. Existing image fusion methods include HIS transform, multiplication transform
288 fusion, wavelet transform fusion and Brovey transform fusion. The Brovey ratio transform fusion
289 method was used in this study.
290   The Brovey ratio transformation and fusion method is a multiplicative band operation with
291 a panchromatic band after the multispectral data has undergone normalized processing, and it
292 clearly represents surface objects such as mountains, vegetation, and water bodies. This fusion
293 method can improve the segmentation effect, and is, therefore, used in this study.
294   In this study, the Brovey transform fusion method uses the image and panchromatic bands
295 of the red band (R), green band (G), and near-infrared (NIR) to conduct fusion, finally obtaining
296 good spatial resolution from the information and retaining multispectral images. This method also
297 promotes the extraction of detail.
298   The fusion formula of Brovey transformation is as follows:
299

300
$$\begin{cases} BT_{red} = Pan \times \dfrac{R}{R + G + NIR} \\ BT_{gre} = Pan \times \dfrac{R}{R + G + NIR} \#(2) \\ BT_{nir} = Pan \times \dfrac{NIR}{R + G + NIR} \end{cases}$$

301
302   BTred, BTgre, and BTnir respectively are pixel values of red, green, and near-infrared images
303 after fusion. Where, R, G, and NIR are pixel values of red, green, and near-infrared bands in multi-
304 spectral images respectively, and Pan is a pixel value of panchromatic band of high-resolution
305 images. Images are processed according to the Brovey formula above, which improves visibility
306 to the human eye, sharpens plot features and achieve higher image quality.
307   In this study, only Landsat 8 satellite data were used as the data source, and Brovey
308 transformation was used for image fusion of partial multispectral band data and panchromatic
309 bands in the image. This effect is shown in Figure 5.

310 **2.3.4 Dataset Preprocessing**
311   The convolutional neural network has certain standards for the size of the input image, and
312 the size should be standard and uniform. Too large a size is not conducive to the extraction of
313 detailed features by the convolutional layer, and too little feature information extracted from too
314 small a size is not conducive to pixel classification. The particularity of the convolutional operation
315 requires that the length and width of the input feature map be consistent. Therefore, it is critical to
316 determine the input end of the model. If the size of the input end is larger, some image details will
317 be lost, and if the size of the input end is smaller, training errors will be generated because many

318   blocks of a certain type, thus affecting the accuracy of the model. As the spatial resolution of the
319   remote sensing image data obtained from Landsat 8 OLI satellite is 30 m, it has medium resolution.
320   Considering the image input size and label distribution comprehensively, image cutting should be
321   performed first. The cutting method randomly selects coordinates, which are taken as the origin of
322   interception, and sampling interception is carried out. Large remote sensing images and
323   corresponding labeled images are cut into sub-images in the above way, with sizes of 512 × 512,
324   256 × 256 and 128 × 128. Figure 6 shows only a sample image of the water segmentation.
325   Remote sensing images are too large to be placed directly into the neural network, and the
326   number of remote sensing images is small. Therefore, OpenCV is used to expand and add noise to
327   the images, to effectively improve the training accuracy and generalization ability of the network.
328   The main processing steps are as follows:
329   (1) The original and labeled images were rotated by 90, 180, and 270 degrees along the Y-
330   axis.
331   (2) Noise, including salt and pepper noise, Gaussian noise, and other noise transformations
332   were added to the original image transformation.
333   (3) Light transformation was performed on the original image, including for brightness.
334   (4) The original image was processed using fuzzy methods.
335   After image cutting and data expansion, the amount of data reached 10,000 pieces. In
336   addition to image expansion, corresponding label images were expanded. After expansion, the
337   classification distribution of the dataset was balanced, and the image resolution was sufficient to
338   meet the demands of training the model. After the cutting image was trained, the prediction image
339   was output. The predicted images were backtracked to the origin of the cutting and were
340   individually spliced into the original large image. After a simple adjustment, the final prediction
341   remote sensing image was obtained.
342   In this study, 512 pixels were selected as the reference scale, which, because of its rich
343   background information, is the most appropriate scale for the human eye to recognize differences.
344   We divide the complete and test images into three levels, namely 512 × 512, 256 × 256, 128 ×
345   128, thus obtaining the initial sample set and test set. See Figure 7.
346   We selected three scene images for data augmentation processing. Because the neural
347   network is sensitive to data with different directions, different colors, and with or without noise
348   points, we expand each image to a photo set with 128 × 128, 256 × 256, 512 × 512 sizes by
349   randomly cutting, angle rotating and adding noise points. Then, for the 128 × 128 data graph, set
350   the category proportion threshold to 50%, delete the images whose category proportion exceeds
351   the threshold, and form the following training set, as shown in Table 5.
352   For Landsat 8 OLI sensor data, the first four bands of image preprocessing obtain real remote
353   sensing images, by removing the interference of the satellite sensor itself as much as possible. The
354   tagged data produced by the third national census data of water conservancy projects are
355   automatically checked after expanding and dealing with the noise of the dataset, for it to have a
356   certain size and diversity.

Method

3.1 Attention Mechanism of Computer Vision

359   In human vision, the visual attention mechanism processes brain signals. Human vision
360   functions by rapidly scanning the global image to focus on the target area. Therefore, this area

361   should receive more attention to obtain detailed information about the target and suppress
362   irrelevant information. For example, when people view pictures, they tend to focus on noteworthy
363   objects but pay less attention to the background. There are three common types of attention
364   mechanism models: the spatial attention model, channel attention model, and spatial and channel
365   mixed attention model.

3.1.1 Channel Attention

367   A good example of the channel attention mechanism is SENet (squeeze and congestion net).
368   The SENet principle involves squeezing the tensor size. Then, it is changed into a tensor of 1×1
369    $\times$ C, for the feature graph of each channel to become a real number, thus completing feature
370   compression. The compressed tensor was used for the excitation operation. The excitation
371   operation uses the parameter W to generate weights for each characteristic channel. The parameter
372   W is learned to explicitly model the correlation between feature channels, thereby realizing the
373   weight distribution for each channel. SENet uses two layers of bottlenecks to multiply the
374   corresponding elements. Thus, the output tensor with the characteristic of the attention distribution
375   in the channel can be obtained. This structure is shown in Figure 8.
376   SENet can achieve notable network performance improvements with minimal additional
377   computation. Because this idea can be applied to various existing network structures, SENet has
378   many derivatives. A typical combination is ResNet and SENet.

3.1.2 Spatial Attention

380   The spatial attention mechanism differs slightly from the channel attention mechanism. If
381   the channel attention mechanism achieves the distribution of attention along the channel, the
382   spatial attention mechanism obtains the spatial distribution of attention, known as the distribution
383   on a two-dimensional picture. This structure is shown in Figure 9. Then a convolution operation
384   is used to transform the feature graph with the number of channels two into the feature graph with
385   the number of channels. To realize the principle of the spatial attention mechanism a feature graph
386   with a channel number of two must first be obtained from the channel dimension by compression.
387   Then, a convolution operation is used to transform the feature graph with two channels into the
388   feature graph with one channel. This obtains the spatial distribution of attention. The eigenmatrix
389   is then multiplied by the corresponding elements of the input tensor to obtain the output tensor
390   characterized by the attention distribution along the channel.

3.1.3 Integration of Channel and Spatial Attention Mechanism

392   The convolution block attention module (CBAM) is a simple and effective attention module.
393   Given an input tensor, the CBAM takes it in two directions: spatial attention and channel attention.
394   Finally, the attentional feature distribution is multiplied by the input tensor to achieve adaptive
395   feature optimization. The structure of CBAM is shown in Figure 10.
396   The advantage of the CBAM is that it is lightweight, and, considering its general purpose, it
397   can be integrated into any convolutional neural network. Second, compared with SENet, CBAM

398   adopts both global average pooling and global maximum pooling, and experiments confirm that,
399   when simultaneously using two types of pooling, the CBAM has a better effect than only using
400   one type of pooling.

3.2 Asymmetric Convolution

402   Asymmetric convolution operates with different lengths and widths of the convolution
403   kernel. Christian Szegedy's Rethinking Inception Architecture for computer vision is an example
404   of asymmetric convolution. Its structure is shown in Figure 11. Compared to ordinary convolution,
405   the significance of designing asymmetric convolution lies in several factors. First, the asymmetric
406   convolution with the convolution kernel size of n × 1 and then the asymmetric convolution with a
407   convolution kernel size of 1 × n is equivalent to the result of convolution with the convolution
408   kernel size of n × n. Meanwhile, asymmetric convolution requires only 2 × n multiplications, while
409   ordinary convolution requires n × n multiplications. Therefore, the larger convolution kernel
410   requires less computation.

3.3 AC-Bottleneck

412   The basic bottleneck structure is shown in Figure 12. First, it reduces the dimensions using a
413   1 × 1 convolution kernel. Then, it is convolved with a 3 × 3 convolution kernel. Finally, the
414   dimension is raised by a convolution kernel of size 1 × 1. The main function is to deepen the
415   network and reduce the amount of calculation by reducing the number of parameters. Feature
416   extraction is more effective after reducing the dimensionality.
417   After extensive experiments, the Swish activation function has proved superior to ReLU.
418   However, compared with ReLU, it has the disadvantage of a higher computational overhead. To
419   reduce the computational overhead of Swish, H-swish was proposed. Therefore, in this study, H-
420   swish directly replaces the ReLU activation function.
421

$$h - swish(x) = x\frac{ReLU6(x + 3)}{6}\#(3)$$

423
424   The input tensors were separately convolved asymmetrically twice, and the two results were
425   added. Then, the result activated by the h-swish function is passed to the CBAM. Finally, the result
426   was added to the residuals to obtain the final result. Asymmetric convolution, the CBAM, and
427   residual structure can greatly improve the training effect with a small amount of computational
428   overhead, as shown in Figure 13. Compared with the CBAM, the AC-CBAM introduced an h-
429   swish function. The introduction of the h-swish function reduced the computational cost of
430   asymmetric convolution and increased the computational speed. In addition, the AC-CBAM added
431   two asymmetric convolutions in front of the CBAM module compared with the unmodified
432   CBAM. In remote sensing images, water and roads were found to differ from others features, being
433   mainly in long strip shapes. The other features are mainly square, requiring the long-strip image

434   features to be extracted using asymmetric convolution.

Experimental process

4.1 Experimental Settings

437   The program in this study adopts Pytorch, the mainstream deep learning framework, and the
438   hardware and software configurations of the machine are shown in Table 6.

439   The parameters of the deep learning network model used in this study are shown in Table 7.

440   The ratio of the training set to the validation set in the dataset was 4:1. The initial learning
441   rate was set to 0.01. At the same time, due to the limitations of the experimental environment in
442   this paper, the batch size and iteration times (epoch) were set to 4 and 300, respectively, and the
443   performance was optimal. The optimizer selects SGD because it converges to the global optimal
444   solution.

4.2 Experimental Results

446   Experiment 1: Three CBAM modules were added to the convolution layer. Simultaneously,
447   the ReLU function is replaced by the h-swish function.

448   Experiment 2: After adding three CBAM modules to the convolution layer, the third CBAM
449   module was replaced with the AC-CBAM module. Simultaneously the ReLU function is replaced
450   by the h-swish function.

451   Experiment 3: The AC-CBAM module was added only after the third convolution layer.
452   Simultaneously, the ReLU function is replaced by the h-swish function.

453   Experiment 4: Based on Experiment 3, the AC-CBAM of the last convolution was replaced
454   by the CBAM.

455   The corresponding structure is shown in Figure 14:

456   The relevant experimental results are shown in Table 8 and Table 9.

457   By comparing Tables 8 and 9, it can be concluded from Experiments 1 and 2 is that under
458   the same conditions, adding the CBAM module to the first two convolutions and the AC-CBAM
459   module to the last convolution in Bottleneck is better than adding the CBAM module to the third
460   convolution.

461   Conclusions can be drawn from Experiments 1 and 3, confirm that, under the same other
462   conditions, the effect of not adding CBAM to the first two convolutions and adding AC-CBAM to
463   the last convolution is better than that of adding CBAM to the first two convolutions and adding
464   AC-CBAM to the last convolution. AC-CBAM alone was found to be more effective.

465   The conclusions drawn from Experiments 3 and 4, confirm that under the same conditions,
466   the effect of adding the AC-CBAM to the bottleneck structure is better than that of the CBAM.

467   The conclusions drawn from Experiments 2 and 4, confirm that under the same conditions,
468   the effect of adding the CBAM and AC-CBAM in the bottleneck is inferior to adding only the AC-

469   CBAM.To test the robustness and effectiveness of the model in this study, we compared the SOTA
470   model proposed in recent years with AC-CBAM, and the results are shown in Table 10. The mIoU
471   of the AC-CBAM model reached 97.34%, which is 1.44% higher than that of DNLNet (95.9%),
472   and the IoU of each category was the highest.

473   ## 4.3 Slide Window Prediction

474   In the process of prediction, the remote sensing image is too large to be directly predicted, so
475   it is necessary to segment the remote sensing image and predict the segmented sub-blocks. Finally,
476   the prediction image is spliced according to the segmentation sequence to obtain the prediction
477   image of the original remote sensing image.
478   However, through local amplification, we found that this effect was not ideal. Because of the
479   problem of the receptive field of the convolutional network, attention was focused on the middle
480   region, and the segmentation effect was poor for the edges of small blocks of images, and the sense
481   of abruptness after fusion was obvious. Using only a simple segmentation prediction an image will
482   exhibit a strong edge with the high error rate of edge prediction causing obvious abrupt changes,
483   and other shortcomings. Therefore, the sliding window prediction method was used for
484   enhancement.
485   In sliding window prediction, the overlapping sliding window strategy (sliding window step
486   size < sliding window size) was adopted. In prediction, only the central region of the prediction
487   results is retained, and the edges of images with inaccurate predictions are discarded. In this study,
488   we make a slide-window prediction based on the segmentation of the original image and obtain a
489   good prediction effect.
490   We set the length of the original image as x and the height as y, and the relevant filling and
491   sliding window settings are as follows:
492   1. Fill 1(yellow part): Fill the lower right boundary to an integer multiple of the size of the
493   sliding window prediction window to facilitate divisible cutting;
494

495
$$\text{Padding\_x\_yellow} = \left\lceil \frac{x}{stride} \right\rceil \times \text{stride} - x \#(4)$$

496

497
$$\text{Padding\_y\_yellow} = \left\lceil \frac{y}{stride} \right\rceil \times \text{stride} - y \#(5)$$

498
499   2. Fill 2(blue part): Fill the outer border with 1/2 sliding step size (considering the expansion
500   prediction of edge data). The filling effect is illustrated in Figure 15.
501

502
$$\text{Padding}_{\text{blue}} = \frac{stride}{2} \#(6)$$

503
504   3. Only the prediction result of the center stride × stride of the Sliding window is reserved for
505   each prediction, as shown in Figure 16.

Results and discussions

5.1 Contribution to semantic segmentation of remote sensing

Data preprocessing is essential before model training. A superior processing method effectively improves model accuracy. Because the sensor records from orbit, the imaging process will inevitably be affected by many factors, such as the action of external radiation and the atmosphere, resulting in a degree of error between the remote sensing image and the original view. Here, we adopt radiometric calibration and atmospheric correction to eliminate maximally the influence of error. There are usually many different bands in remote sensing images, and different bands are combined depending on the land parcel category. The practice of band combination in this study will provide a reference for the field of remote sensing parcel segmentation. Panchromatic band images have the advantage of higher resolution. Further to enhance the information contained in the image, we fuse the selected multispectral band data with panchromatic band data. In this study, blocks of vegetation, water, and roads are highlighted by the Brovey transform fusion method to improve segmentation accuracy. This paper provides a reference for the selection of image fusion data sources and fusion methods.

The dataset used in this experiment includes five categories: forest, building, water body, rationale, and other categories. After training with GCNet, PointRend, OCRNet, DANet, APCNet, and other semantic segmentation networks, it was found that the pixel points of the roads category were mainly distributed in thin strips and had fewer numbers than other categories when inferring the prediction images. The OCRNet model prediction map for a remote sensing image is shown in Figure 17.

The main reason for the above results is that in the process of feature extraction for images, roads are mainly strip-shaped and occupy a relatively small area. If a larger general symmetric convolution kernel is used to extract features from images, the effect may change because the convolution kernel size is too large to achieve accurate feature extraction and other unnecessary information will be introduced. For recognizing and segmenting traditional remote sensing images the method of introducing an attention mechanism effectively improves the model training accuracy and can also solve a series of problems in object classification. The introduction of the attention mechanism could also alleviate problems in remote sensing image extraction. However, because the road shape is a strip, the image can be extracted using the asymmetric convolution core of the bar shape, and the attention mechanism module of CBAM can be used to obtain the target area requiring attention. Simultaneously, the two introduced structures can enhance the outcome of the entire training process. Therefore, the original AC-CBAM module was created. A comparison of the predictions is shown in Figure 18.

The experiments show that the entire remote sensing image predicts well. This method effectively extracts a complete image. Overall, this study undertook the following research.

(1) Introduction of related theory and technology: Related theories such as neural networks and their training processes were introduced to commonly used methods of remote sensing image

544   segmentation.

545   (2) Study area selection and remote sensing image preprocessing: To improve the image
546   quality, remote sensing image is preprocessed.

547   (3) Dataset preparation: Label data were generated by automatic labeling and manual
548   validation for data preprocessed from remote sensing images, and the dataset is expanded to 10,000
549   pieces.

550   (4) In the decoder part, we created the AC-CBAM module and added the h-swish activation
551   function. The experimental results show that the performance of the model was somewhat
552   improved.

553   (5) Rethinking the network structure from object features: According to the experimental
554   results, the AC-CBAM module and sliding window prediction can improve the accuracy of
555   extraction to some extent, which proves that the combination of computer vision and remote
556   sensing knowledge will bring new ideas to solve many problems in the field of remote sensing.

557   5.2 Different from existing methods

558   The CBAM has proven to be an attention mechanism module that can be widely used in
559   target detection and semantic segmentation tasks. However, through experiments, we found that
560   the unmodified CBAM is not suitable for the semantic segmentation of remote sensing images.
561   Because the CBAM is an attention mechanism module based on ordinary convolution, a
562   symmetrical square convolution kernel was used in feature extraction, whereas the improved AC-
563   CBAM used a strip asymmetric convolution kernel, which is more in line with the category
564   characteristics in the experimental task. When extracting features from the convolution kernel,
565   most of the information obtained in the convolution area is acquired from categories; thus, when
566   integrating information, the variance in different information is greater and more conducive to
567   identification.

568   We obtained the results of semantic segmentation of OcrNet, Deeplabv3+, and other
569   networks, and compared the effects of each method. First, we used OcrNet. The purpose of the
570   object context information proposed by OcrNet is explicitly to enhance the object information.
571   First, the regional feature expression of a group of objects is calculated, and then the representation
572   of these regional-feature objects is propagated to each pixel according to the similarity between
573   the regional-feature object's representation and the pixel feature's representation, to achieve the
574   intended effect. However, for the category of very narrow roads, the advantages of OcrNet were
575   not fully reflected in this experiment. Then, we used Deeplabv3+, which introduces hole
576   convolution, for each convolution output to contain a large range of information without losing
577   any. However, a large range of information does not necessarily contain sufficient information,
578   and will not meet the experimental requirements. Thereafter, we continued to use other methods,
579   and finally, through experimentation, obtained the best effect with the DNLNET method and
580   further improved it.

5.3 Limitations and future work

The image quality extracted from satellites is limited by the complexity of cloud and fog masking geographical features which reduces the accuracy of the obtained images. Owing to the irregular distribution of plots and the mutual embedding of various plots, the traditional convolutional neural network cannot better extract the edge feature information of junctions, thus having a considerable impact on the model's recognition capability.

In the future, this research will further improve the channel separation attention-detection network. There are two approaches to improving the network structure: one is the enhancement algorithm of defogging by fusing the improved dark channel and Retinex [39]. By using the single-scale Retinex algorithm to enhance the restored image the problem of low accuracy caused by the incorrect estimation of transmittance could be somewhat alleviated. The enhanced image is more consistent with human visual characteristics, and the image details are more obvious. Another direction is to introduce a deformable convolution [40]. In this method, the fixed-shape convolution matrix is transformed into a variable convolution matrix that can adapt to the shape of the object for the network structure itself to adapt. In addition, the edge information of the image is better extracted further to improve the fitting effect of the model.

References

[1] Shotton J, Johnson M, Cipolla R. Semantic texton forests for image categorization and segmentation[C]//2008 IEEE conference on computer vision and pattern recognition. IEEE, 2008: 1-8.

[2] Schroff F, Criminisi A, Zisserman A. Object Class Segmentation using Random Forests[C]//BMVC. 2008: 1-10.

[3] Osher S, Sethian J A. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations[J]. Journal of computational physics, 1988, 79(1): 12-49.

[4] Ball J E, Bruce L M. Level set hyperspectral image classification using best band analysis[J]. IEEE Transactions on Geoscience and Remote Sensing, 2007, 45(10): 3022-3027.

[5] Ma H, Yang Y. Two specific multiple-level-set models for high-resolution remote-sensing image classification[J]. IEEE Geoscience and Remote Sensing Letters, 2009, 6(3): 558-561.

[6] Längkvist M, Kiselev A, Alirezaie M, et al. Classification and segmentation of satellite orthoimagery using convolutional neural networks[J]. Remote Sensing, 2016, 8(4): 329.

[7] Zhao W, Du S, Emery W J. Object-based convolutional neural network for high-resolution imagery classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(7): 3386-3396.

[8] De S, Bruzzone L, Bhattacharya A, et al. A novel technique based on deep learning and a synthetic target database for classification of urban areas in PolSAR data[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 11(1): 154-170.

[9] Tang X, Zhang X, Liu F, et al. Unsupervised deep feature learning for remote sensing image retrieval[J]. Remote Sensing, 2018, 10(8): 1243.

[10] Zhao W, Du S, Wang Q, et al. Contextually guided very-high-resolution imagery classification with semantic segments[J]. ISPRS journal of photogrammetry and remote sensing, 2017, 132: 48-60.

622 [11] Li S Z. Markov random field models in computer vision[C]//European conference on
623 computer vision. Springer, Berlin, Heidelberg, 1994: 361-370.
624 [12] Besag J. On the statistical analysis of dirty pictures[J]. Journal of the Royal Statistical
625 Society: Series B (Methodological), 1986, 48(3): 259-279.
626 [13] Nishii R. A Markov random field-based approach to decision-level fusion for remote
627 sensing image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2003,
628 41(10): 2316-2319.
629 [14] Feng W, Jia J, Liu Z Q. Self-validated labeling of Markov random fields for image
630 segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(10):
631 1871-1887.
632 [15] Zheng C, Zhang Y, Wang L. Semantic segmentation of remote sensing imagery using an
633 object-based Markov random field model with auxiliary label fields[J]. IEEE Transactions on
634 geoscience and remote sensing, 2017, 55(5): 3015-3028.
635 [16] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic
636 segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern
637 recognition. 2015: 3431-3440.
638 [17] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling
639 sentences[J]. arXiv preprint arXiv:1404.2188, 2014.
640 [18] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder
641 architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine
642 intelligence, 2017, 39(12): 2481-2495.
643 [19] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep
644 convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.
645 [20] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with
646 deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on
647 pattern analysis and machine intelligence, 2017, 40(4): 834-848.
648 [21] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic
649 image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.
650 [22] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution
651 for semantic image segmentation[C]//Proceedings of the European conference on computer
652 vision (ECCV). 2018: 801-818.
653 [23] Cheng G, Lang C, Wu M, et al. Feature Enhancement Network for Object Detection in
654 Optical Remote Sensing Images[J]. Journal of Remote Sensing, 2021, 2021.
655 [24] Cheng G, Yang C, Yao X, et al. When deep learning meets metric learning: Remote sensing
656 image scene classification via learning discriminative CNNs[J]. IEEE transactions on geoscience
657 and remote sensing, 2018, 56(5): 2811-2821.
658 [25] Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proceedings of the
659 IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.
660 [26] Xue H, Liu C, Wan F, et al. Danet: Divergent activation for weakly supervised object
661 localization[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision.
662 2019: 6589-6598.
663 [27] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings
664 of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
665 [28] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]//Advances in neural
666 information processing systems. 2014: 2204-2212.

667   [29] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and
668   translate[J]. arXiv preprint arXiv:1409.0473, 2014.
669   [30] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural
670   information processing systems. 2017: 5998-6008.
671   [31] Kirillov A, Wu Y, He K, et al. Pointrend: Image segmentation as rendering[C]//Proceedings
672   of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9799-9808.
673   [32] Yuan Y, Chen X, Wang J. Object-contextual representations for semantic
674   segmentation[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK,
675   August 23–28, 2020, Proceedings, Part VI 16. Springer International Publishing, 2020: 173-190.
676   [33] Peng C, Zhang X, Yu G, et al. Large kernel matters--improve semantic segmentation by
677   global convolutional network[C]//Proceedings of the IEEE conference on computer vision and
678   pattern recognition. 2017: 4353-4361.
679   [34] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution
680   for semantic image segmentation[C]//Proceedings of the European conference on computer
681   vision (ECCV). 2018: 801-818.
682   [35] Wu X, Wu Z, Guo H, et al. DANNet: A One-Stage Domain Adaptation Network for
683   Unsupervised Nighttime Semantic Segmentation[C]//Proceedings of the IEEE/CVF Conference
684   on Computer Vision and Pattern Recognition. 2021: 15769-15778.
685   [36] He J, Deng Z, Zhou L, et al. Adaptive pyramid context network for semantic
686   segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
687   Recognition. 2019: 7519-7528.
688   [37] Zhu Z, Xu M, Bai S, et al. Asymmetric non-local neural networks for semantic
689   segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision.
690   2019: 593-602.
691   [38] Yin M, Yao Z, Cao Y, et al. Disentangled non-local neural networks[C]//European
692   Conference on Computer Vision. Springer, Cham, 2020: 191-207.
693   [39] Zhou Y W, Chen Q, Sun Q S. Remote sensing image enhancement based on dark channel
694   prior and bilateral filtering[J]. Journal of Image and Graphics, 2014, 19(2): 313-321.
695   [40] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE
696   international conference on computer vision. 2017: 764-773.
697

Figure 1

The original image

The original image

Figure 2

Gauss kernel structure

0.0000	0.0000	0.0003	0.0006	0.0003	0.0000	0.0000
0.0000	0.0011	0.0079	0.0153	0.0079	0.0011	0.0000
0.0000	0.0079	0.0563	0.1082	0.0563	0.0079	0.0003
0.0000	0.0153	0.1082	0.2079	0.1082	0.0153	0.0006
0.0000	0.0079	0.0563	0.1082	0.0563	0.0079	0.0003
0.0000	0.0011	0.0079	0.0153	0.0079	0.0011	0.0000
0.0000	0.0000	0.0003	0.0006	0.0003	0.0000	0.0000

Figure 3

Comparison before and after enhancement

Comparison before and after enhancement

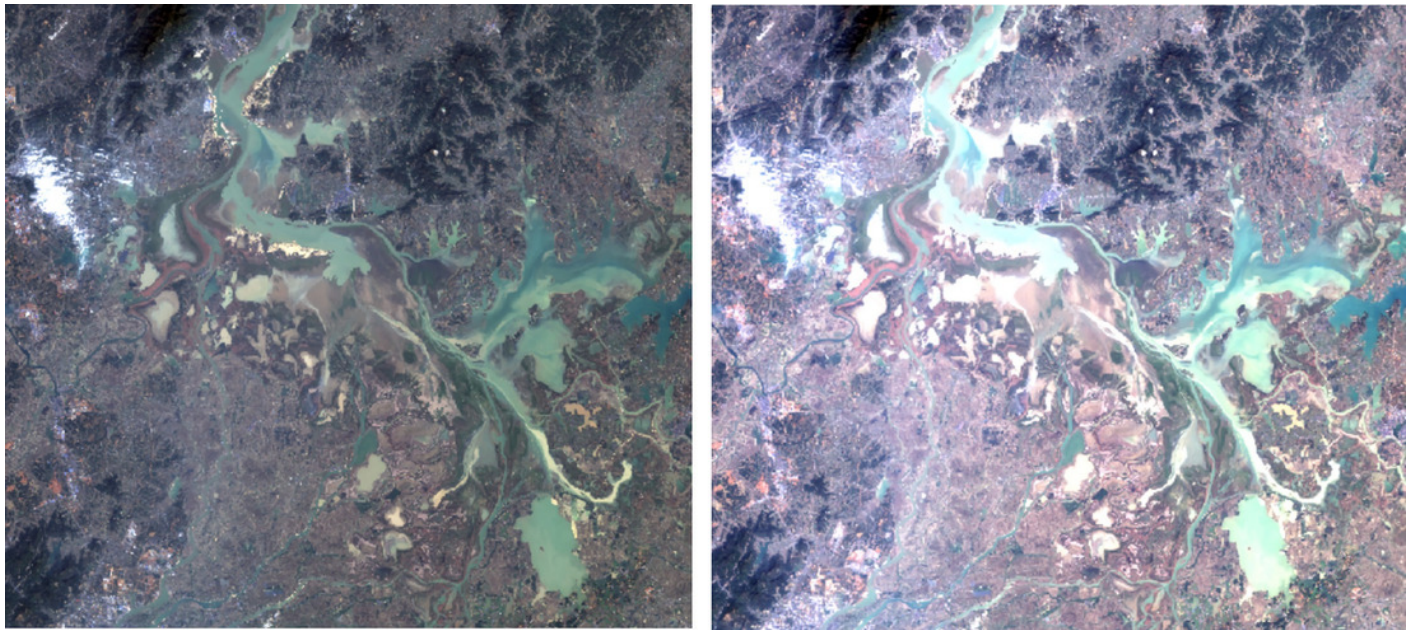



Figure 4

Before and after band combination with Gaussian filtering

a is the original image, B is the image after band combination and C is the image after Gaussian filtering

(a)

(b)

(c)

Figure 5

Before and after fusion

the left is the image before fusion, and the right is the image after fusion

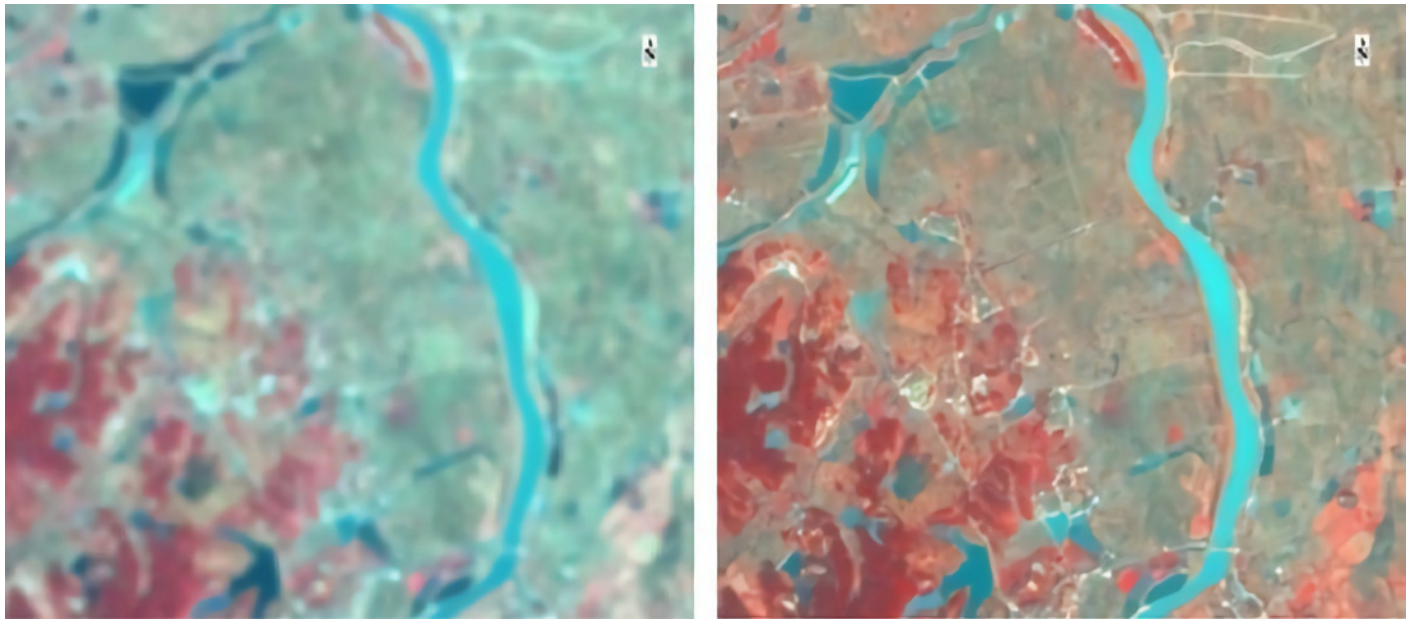



Figure 6

Visualization of training set image and corresponding label taking water body as an example

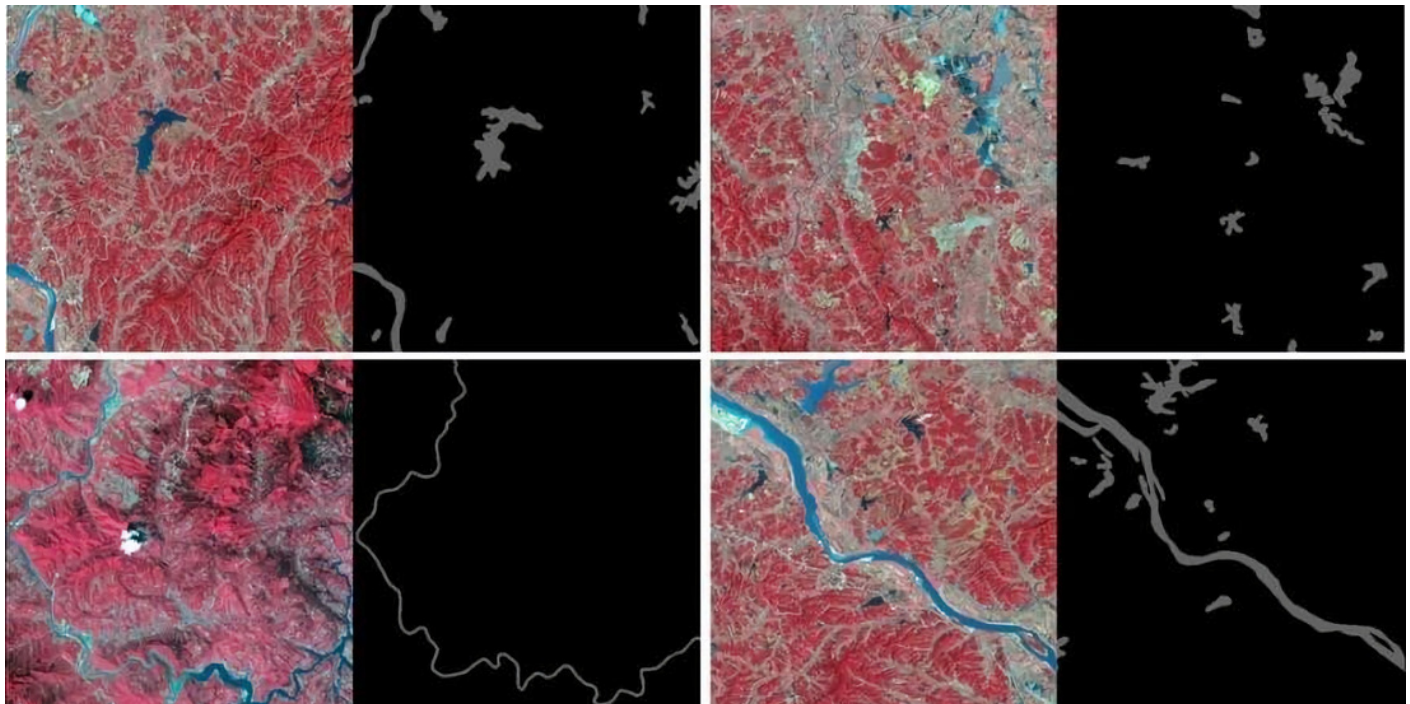



Figure 7

Part of the training set image processed by data enhancement

Part of the training set image processed by data enhancement

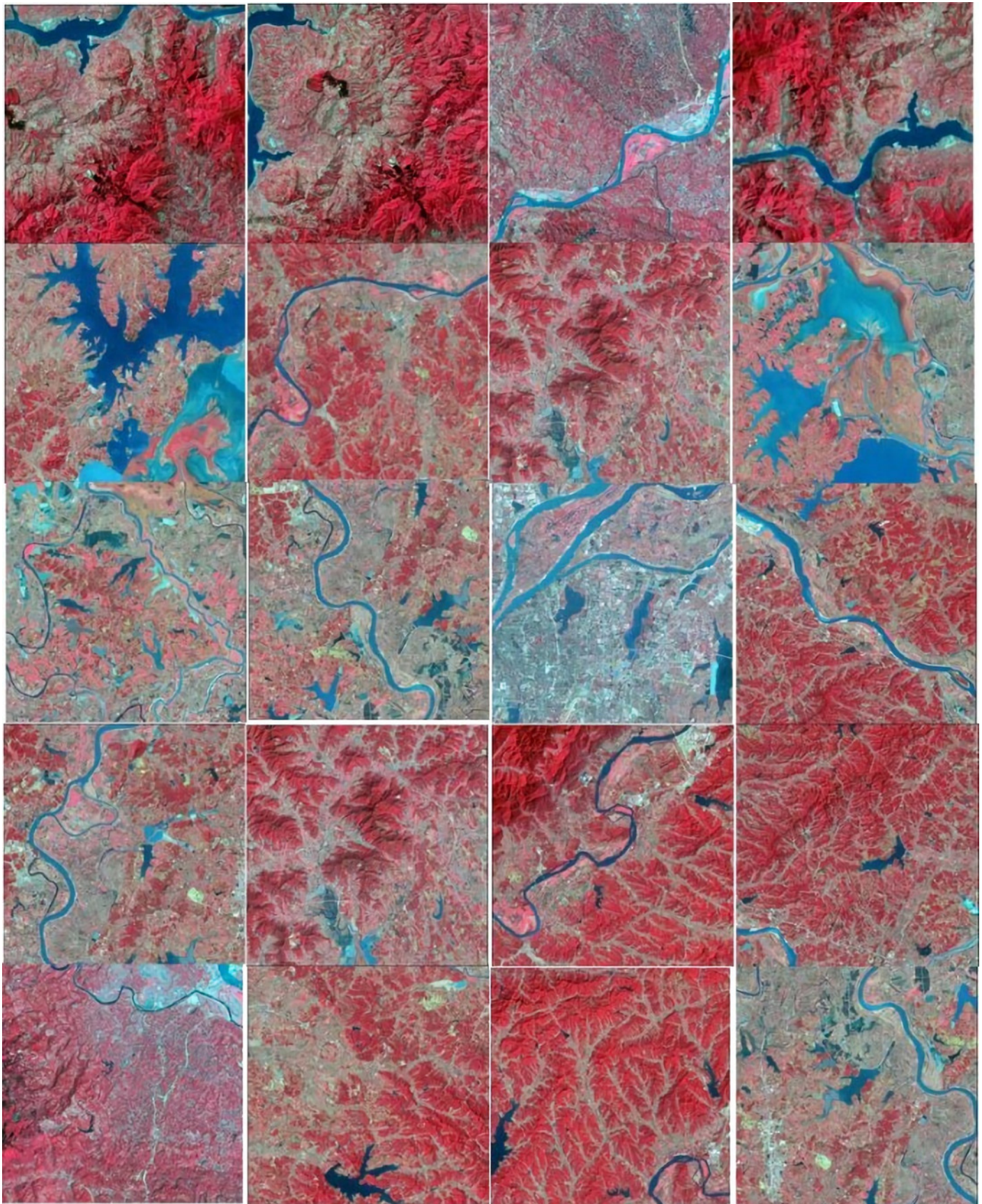



Figure 8

Channel Attention Mechanism structure

Channel Attention Mechanism structure

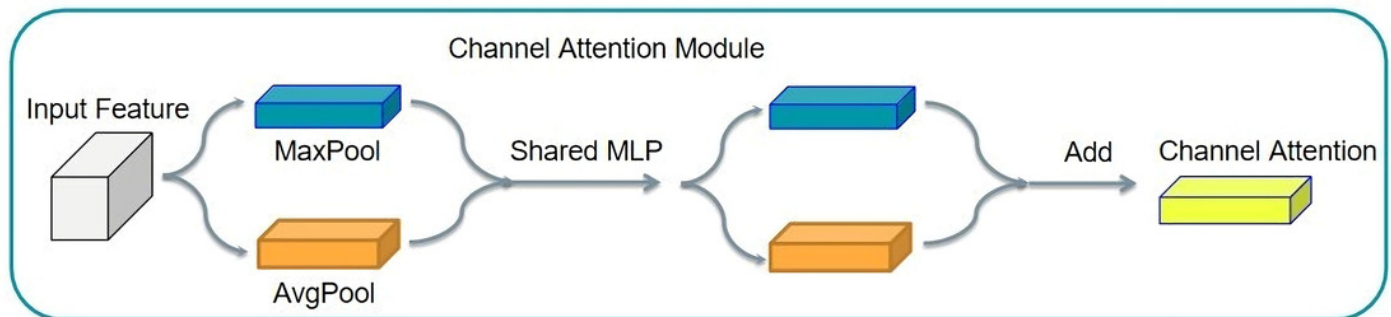



Figure 9

Spatial Attention Mechanism structure

Spatial Attention Mechanism structure

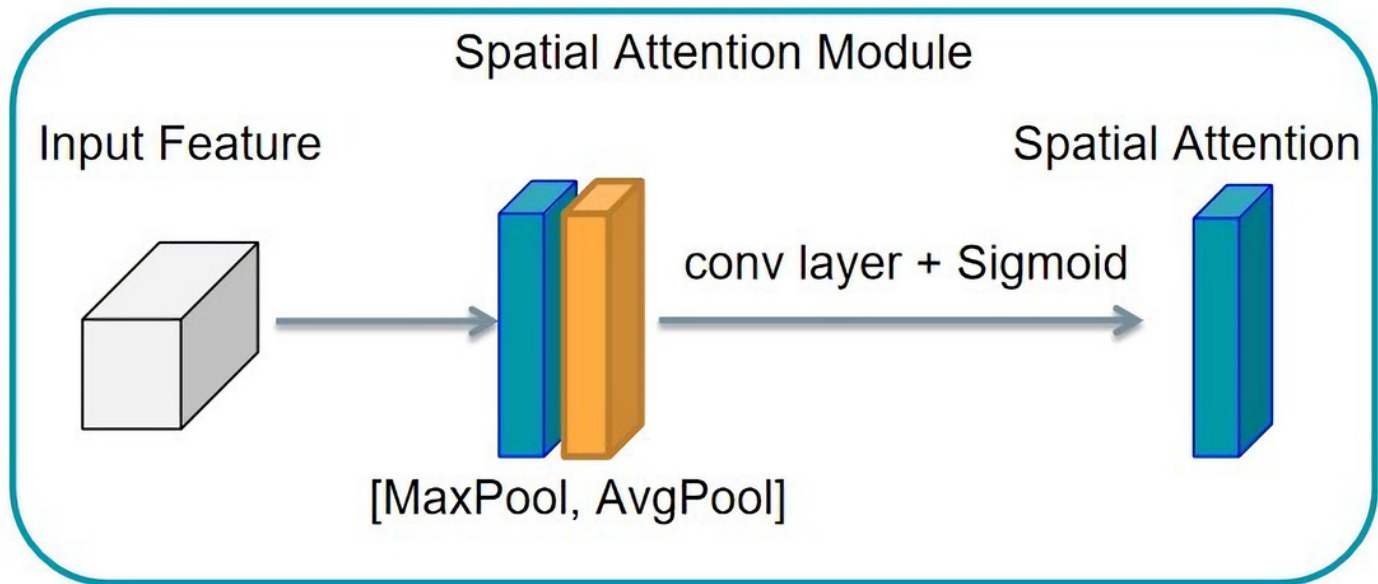



Figure 10

Convolution Block Attention Module structure

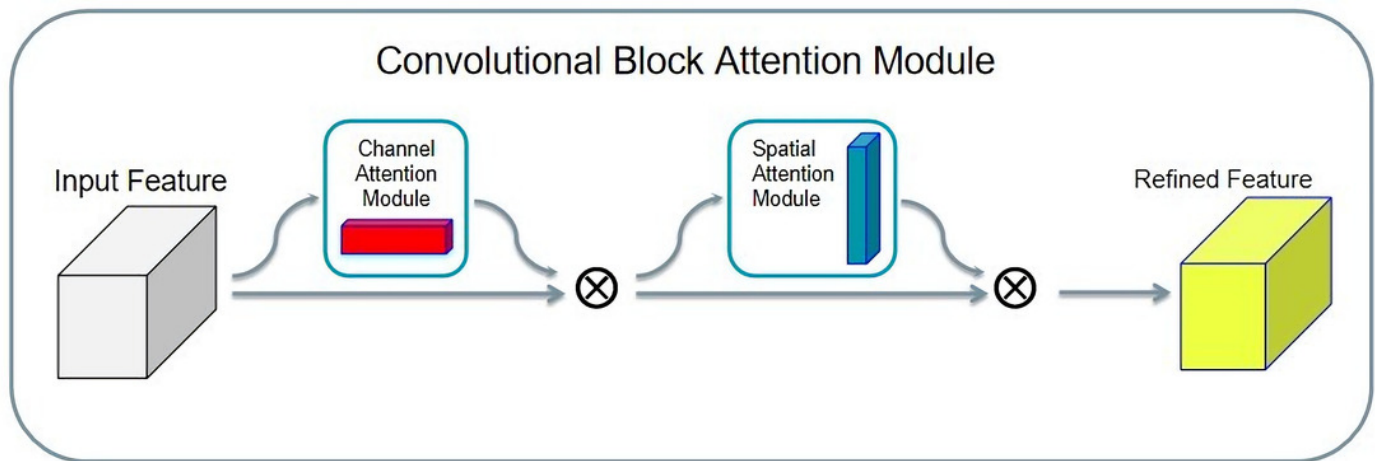



Figure 11

Asymmetric Convolution Module structure

Asymmetric Convolution Module structure

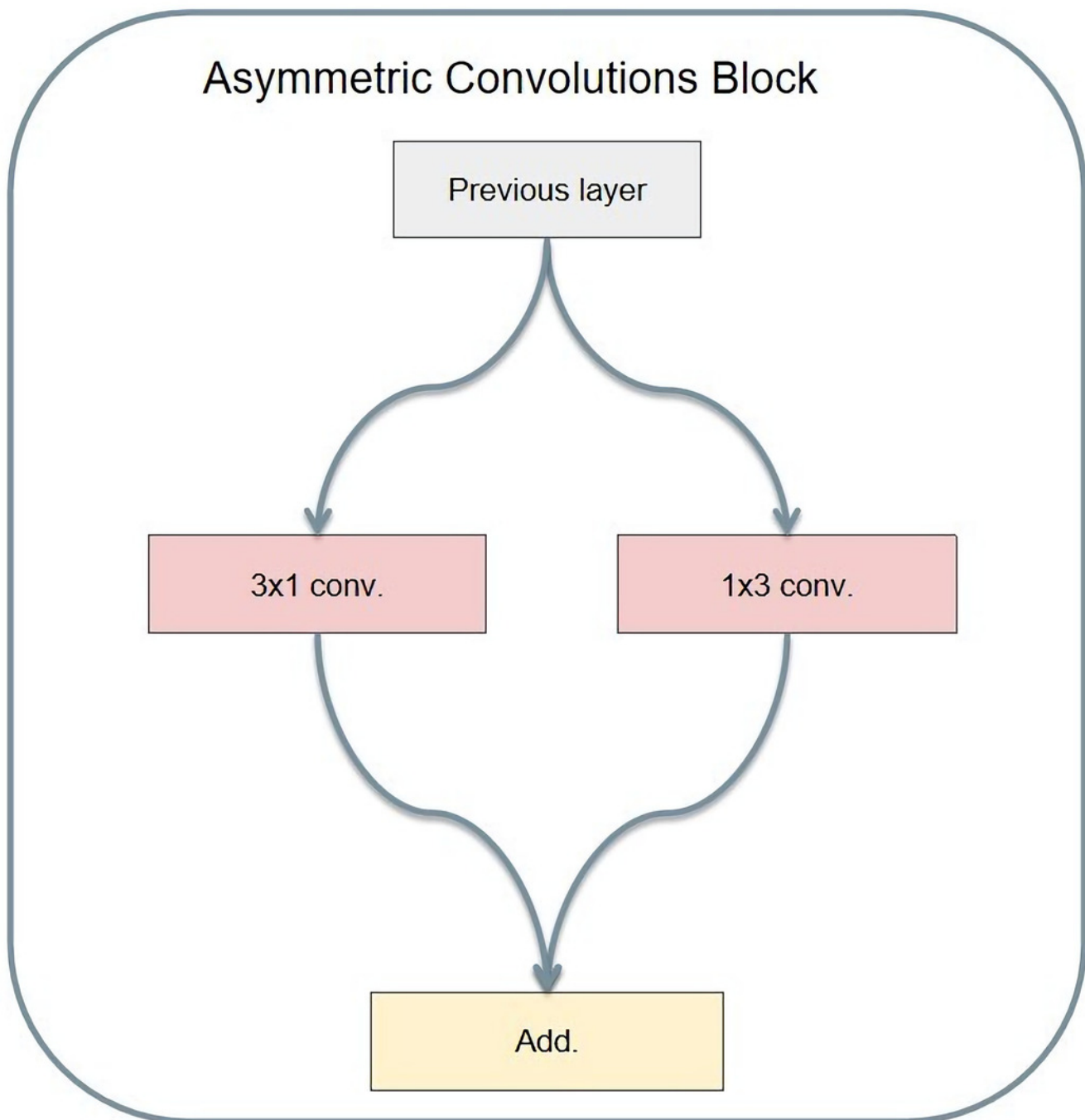



Figure 12

A new Bottleneck structure developed in this paper

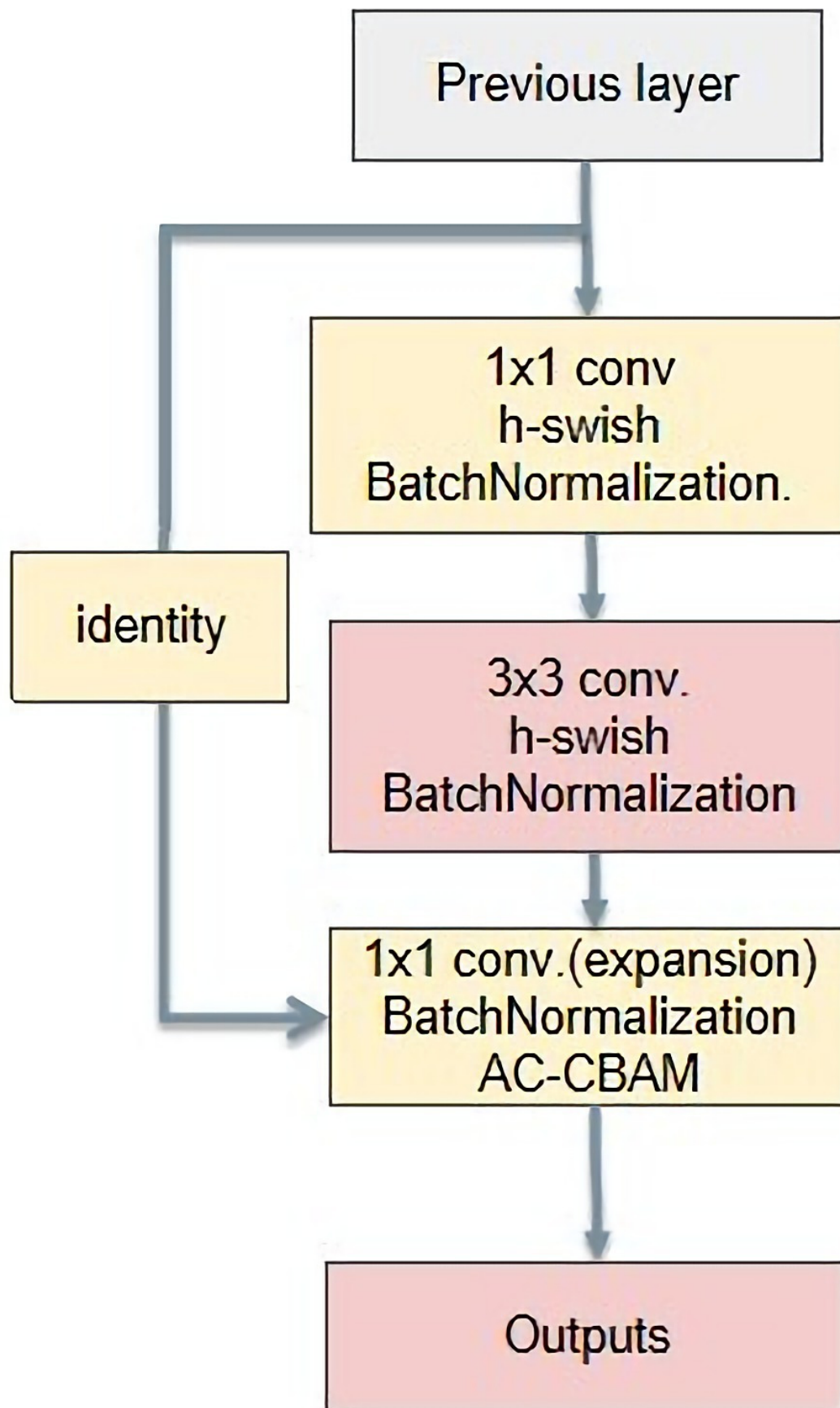



Figure 13

A New AC-CBAM Structure

A New AC-CBAM Structure

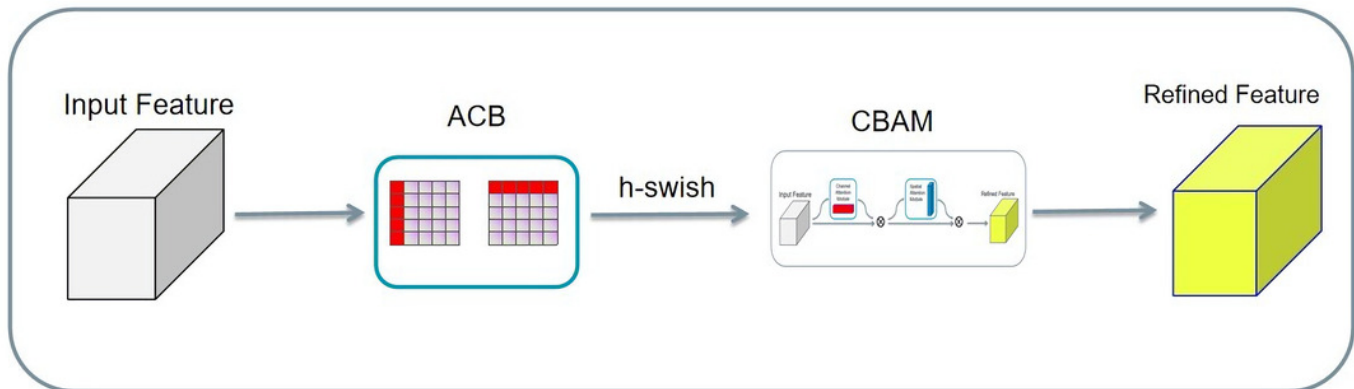



Figure 14

Compare the three Botttleneck diagrams of the experiments

with Experiment 1 on the left, Experiment 2 in the middle and Experiment 3 on the right

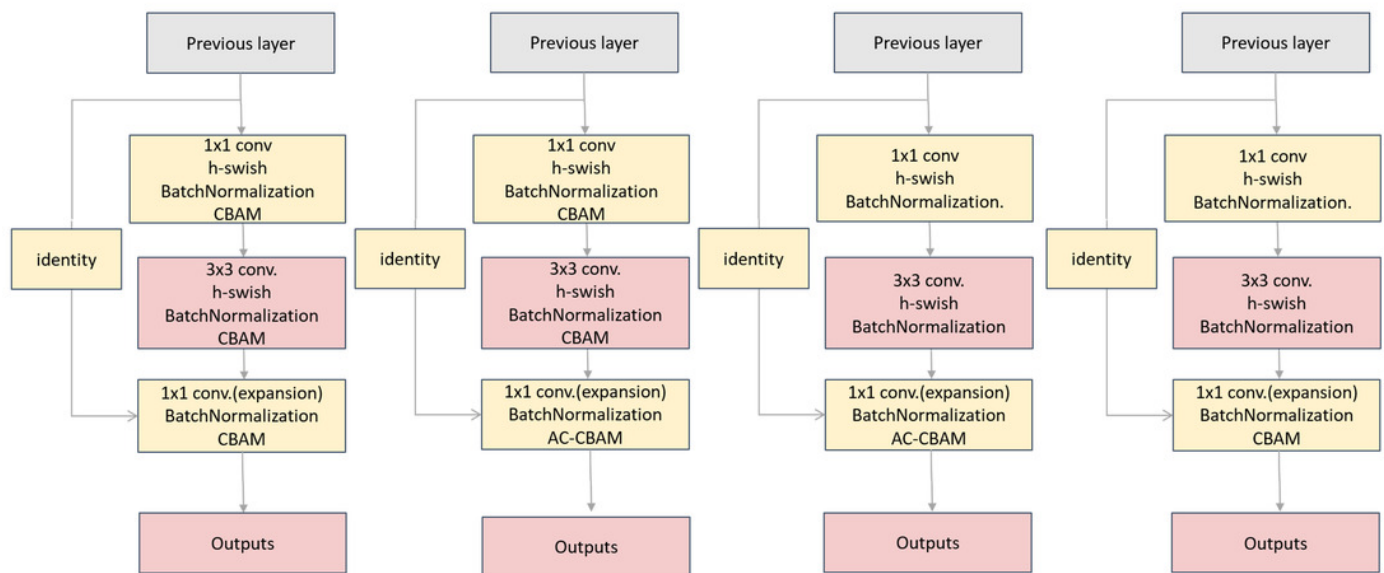



Figure 15

Filling diagram

Filling diagram

Figure 16

A sketch of a sliding window

with the horizontal sliding window predicting process on the left and the vertical sliding

window predicting process on the right.

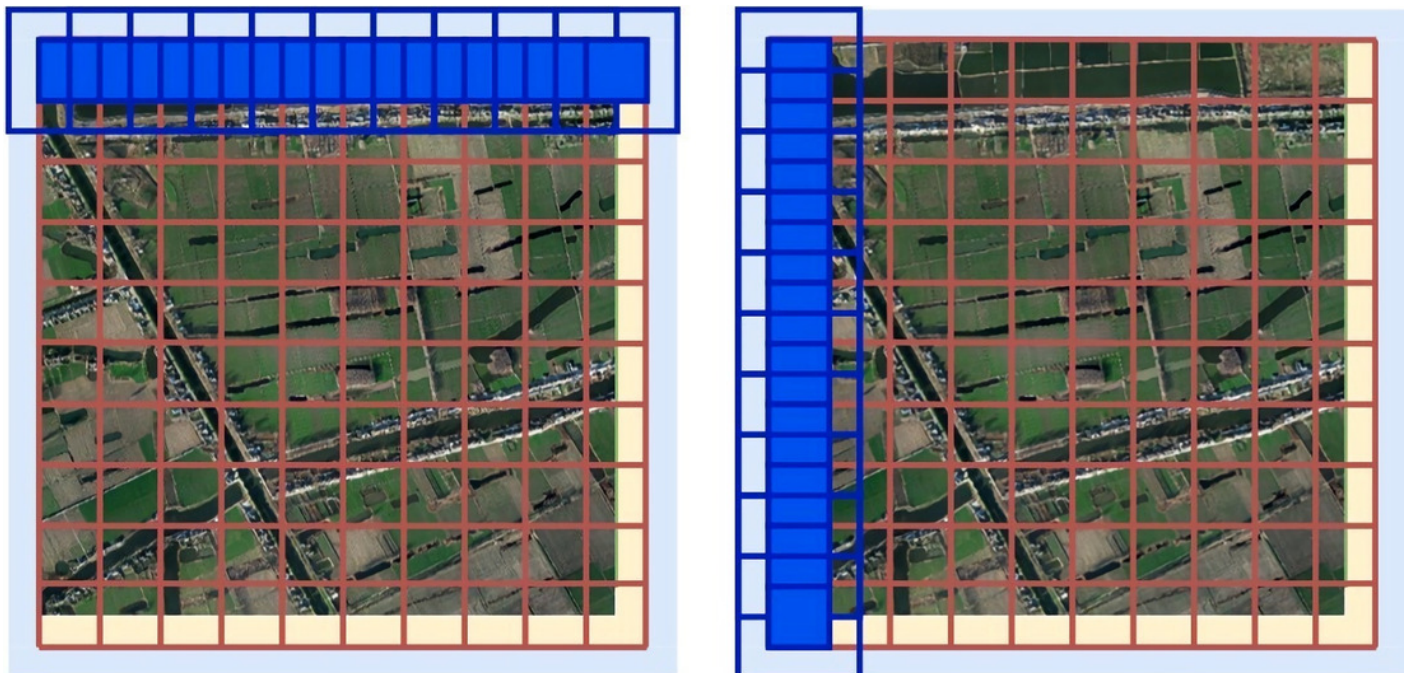



Figure 17

OCRNet Prediction Map

OCRNet Prediction Map

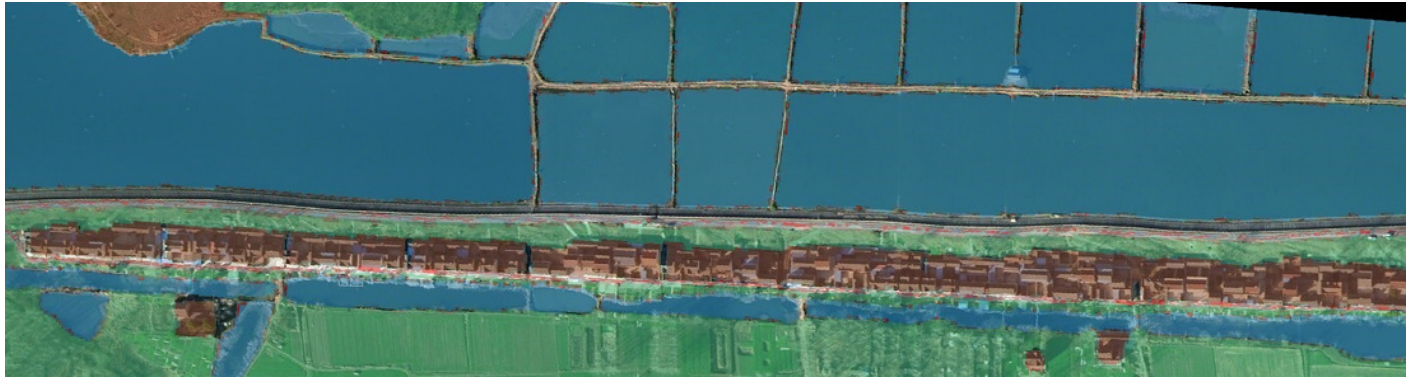



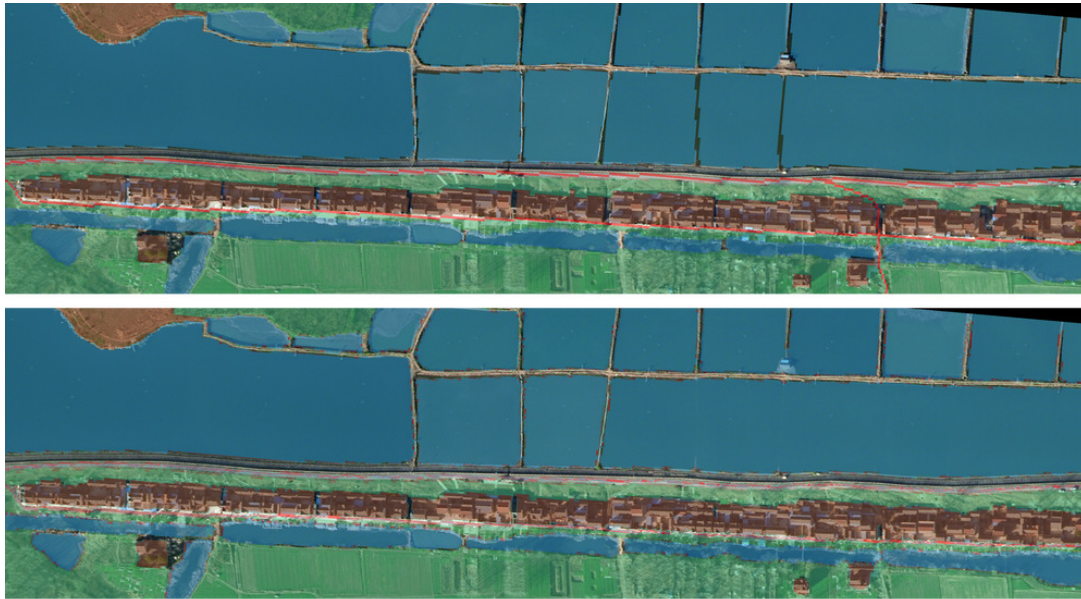
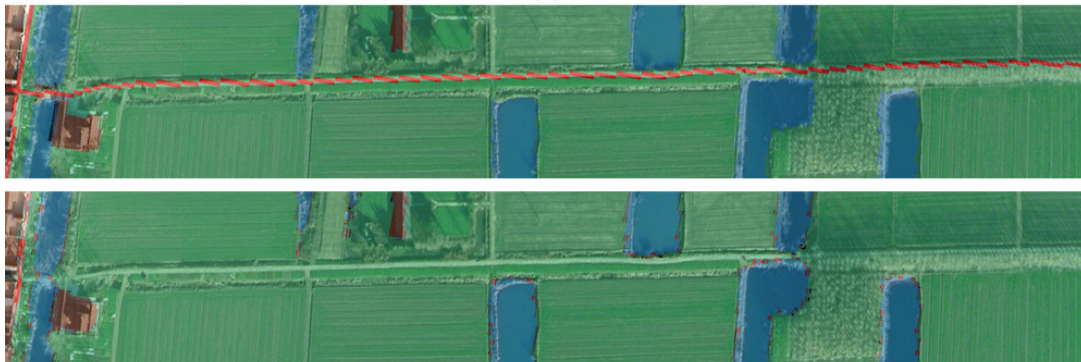
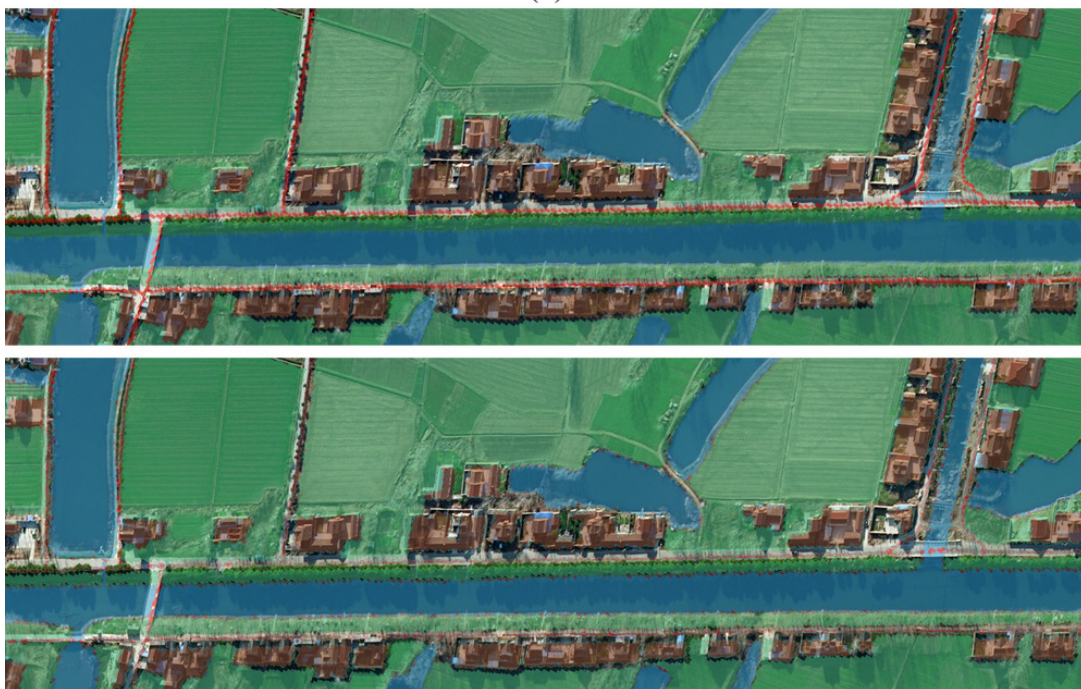
Figure 18

the comparison groups of four different regions.

a, b, c are the comparison groups of three different regions. For each group, the figure above is the ACB prediction chart and the figure below is OCRNet prediction chart

(a)

(b)

(c)

Study the basic information table of the image

Study the basic information table of the image

1   **Table 1:**

2   Study the basic information table of the image

The study area	Image acquisition time	Cloud content (%)
Kunming Dianchi Lake, Fuxian Lake, Yangzonghai	2020.1.18	0.03
Pan Yang Lake	2019.1.23	1.6
Leshan city, Yibin city area	2018.4.18	0.09

Landsat8 satellite sensor parameters

1 **Table 2:**

2 Landsat8 satellite sensor parameters

The sensor	Band	Wavelength range /μm	Signal to noise ratio	Spatial resolution /m
	1-COASTAL/AEROSOL	0.43-0.45	130	30
	2-Blue	0.45-0.51	130	30
	3-Green	0.53-0.59	100	30
	4-Red	0.64-0.67	90	30
OLI	5-NIR	0.85-0.88	90	30
	6-SWIR1	1.57-1.65	100	30
	7-SWIR2	2.11-2.29	100	30
	8-PAN	0.50-0.68	80	15
	9-Cirrus	1.36-1.38	50	30
	10-TIR	10.60-11.19	0.4k	100
TIRS	11-TIR	11.50-12.51	0.4k	100

Data product parameters

Data product parameters

1 **Table 3:**

2 Data product parameters

The parameter types	Values
The product type	Level 1T terrain correction image
The output format	GeoTIFF
Sampling methods	Cubic convolution algorithm
Map projection	UTM-WGS84 Projected Coordinate System
Fastest return cycle	> 72 hours
Angle	98.2 degrees
Operation cycle	98.9 minutes
Rail type	Near polar solar synchronous orbit
Track height	705km

Band combination scheme table

Band combination scheme table

1 **Table 4:**

2 Band combination scheme table.

RGB Combination	Main uses
4(Red), 3(Green), 2(Blue)	Natural True Color
7(SWIR2), 6(SWIR1), 4(Red)	City
5(NIR), 4(Red), 3(Green)	Standard false color images, vegetation, water
6(SWIR1), 4(NIR), 3(Blue)	Agricultural
7(SWIR2), 6(SWIR1), 5(NIR)	Penetrating the atmosphere, surveying special geological structures
5(NIR), 6(SWIR1), 2(Blue)	Healthy vegetation
5(NIR), 6(SWIR1), 4(Red)	Land
7(SWIR2), 5(NIR), 3(Green)	Remove the natural surface of the atmospheric image
7(SWIR2), 5(NIR), 4(Red)	Short wave infrared
6(SWIR1), 5(NIR), 4(Red)	Vegetation analysis

Number of samples at different scales

Number of samples at different scales

1 **Table 5:**

2 Number of samples at different scales

Image Size	Number
128 × 128	5000
256 × 256	3000
512 × 512	2000

Hardware and software configuration table

Hardware and software configuration table

1 **Table 6:**

2 Hardware and software configuration table

configuration project	detail
CPU	Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz
RAM	128G
Graphics card	2×NVIDIA Quadro RTX 5000
Operating system	64-bit Windows 10
CUDA	Cuda 10.2
Data processing	Python 3.8

Network model parameter list

Network model parameter list

1 **Table 7:**

2 Network model parameter list

parameters	value
epochs	150
batch_size	4
activation	h-swish
Input_tensor	(512,512,4)

Results of three experimental schemes.

IoU represents the intersection ratio between image prediction results and real labels, and Acc represents the accuracy of pixel prediction.

1 **Table 8:**

2 Results of three experimental schemes. IoU represents the intersection ratio between image
3 prediction results and real labels, and Acc represents the accuracy of pixel prediction.

Experiment 1			Experiment 2		
Class	IoU	Acc	Class	IoU	Acc
others	95.65	97.93	others	94.35	96.97
forest	98.33	99.14	forest	97.8	98.79
road	97.34	98.47	road	96.52	98.24
building	98.01	98.91	building	97.27	98.74
water	95.43	97.84	water	93.88	97.31
Experiment 3			Experiment 4		
Class	IoU	Acc	Class	IoU	Acc
others	96.21	97.97	others	76.21	84.41
forest	98.58	99.3	forest	87.63	93.67
road	97.68	98.87	road	83.33	92.29
building	98.32	99.19	building	88.79	94.69
water	95.9	97.99	water	79.08	89.49

Average accuracy of the three schemes.

mIoU represents the average IoU of all categories, mAcc represents the proportion of correctly classified categories calculated separately, and aAcc represents the average mAcc of each category.

1 **Table 9:**

2 Average accuracy of the three schemes. MIoU represents the average IoU of all categories, mAcc
3 represents the proportion of correctly classified categories calculated separately, and aAcc
4 represents the average mAcc of each category.

Experiment 1				Experiment 2			
Scope	mIoU	mAcc	aAcc	Scope	mIoU	mAcc	aAcc
global	96.95	98.46	98.46	global	95.96	98.01	97.97
Experiment 3				Experiment 4			
Scope	mIoU	mAcc	aAcc	Scope	mIoU	mAcc	aAcc
global	97.34	98.66	98.67	global	83.01	90.91	90.35

Relevant models are compared with AC-CBAM accuracy.

Relevant models are compared with AC-CBAM accuracy, where IoU.A represents the value of IoU in category A.

1   **Table 10:**

2   Relevant models are compared with AC-CBAM accuracy, where IoU.A represents the value of
3   IoU in category A.

Model	Backbone	mIOU	IoU.others	IoU.forest	IoU.road	IoU.building	IoU.water
PointRend[31]	ResNet101	0.948	0.926	0.972	0.953	0.956	0.924
OCRNet[32]	HRNet48	0.936	0.906	0.966	0.938	0.959	0.912
GCNet[33]	ResNet101	0.940	0.914	0.968	0.944	0.961	0.916
Deeplabv3+[34]	ResNet101	0.945	0.921	0.975	0.949	0.963	0.922
DANet[35]	ResNet101	0.948	0.926	0.972	0.953	0.965	0.924
APCNet[36]	ResNet101	0.952	0.933	0.974	0.959	0.969	0.929
ANN[37]	ResNet101	0.934	0.903	0.965	0.936	0.959	0.907
DNLNet[38]	ResNet101	0.959	0.942	0.978	0.964	0.973	0.937
Ours AC-CBAM	ResNet101	**0.973**	**0.962**	**0.986**	**0.977**	**0.983**	**0.959**