IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 February 2023, accepted 28 February 2023, date of publication 3 March 2023, date of current version 8 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3252400

RESEARCH ARTICLE

Deep Learning-Based Multiscale Pyramid Sieve and Analysis Module in Image Segmentation

DI CAO[1], JIAN-NONG CAO[2,3], QIAN ZHU[3], LI-JIAO LOU[4], AND NIAN-ZHONG XIAO[5]

[1]School of Earth Science and Resources, Chang'an University, Xi'an 710054, China
[2]Key Laboratory of Degraded and Unused Land Consolidation Engineering, Ministry of Land and Resources, Xi'an 710054, China
[3]School of Geology Engineering and Geomatics, Chang'an University, Xi'an 710054, China
[4]Yellow River Hydrological Survey and Mapping Bureau, Zhengzhou 450000, China
[5]Fourth Geological Exploration Institute of Henan Geology and Mineral Bureau, Zhengzhou 450000, China

Corresponding author: Di Cao (RenesmeeSwan@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 41571346; in part by the Key Laboratory of Degraded and Unused Land Consolidation Engineering of the Ministry of Land and Resources under Grant SXDJ2017-10 and Grant 2016KCT-23; in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2018JQ4027; and in part by the Fundamental Research Funds for the Central Universities under Grant 310826171014, Grant 310826160320, and Grant 300102268647.

ABSTRACT Image segmentation is an essential part of remote sensing preprocessing. This paper proposes a deep learning-based multiscale pyramid sieve and analysis module to solve the problems associated with multiscale object coexistence in the complex semantic environments of multispectral remote sensing images. The image information distribution features, which change with scales in different convolutional feature map pyramids, were investigated, and the relationship between informative scales and the variation function curve was examined. Then, the proposed sieve and analysis module was designed according to the results. The proposed module is described as follows. (1) First, a feature map pyramid is built via a convolution calculation. (2) Sieve informative feature layers from the feature map pyramid with variation analysis. (3) Fuse informative feature layers. (4) The final feature representation is input into subsequent calculations for image segmentation. Experimental results demonstrate that, compared to the control group, the precision of the experimental group improved by 3.57%-5.89%, and the intraclass conformity improved by 2.74%-5.58%. In addition, the intraclass chaos decreased by 1.5%-7.9%. The experimental results demonstrate that the proposed multiscale pyramid sieve and analysis module can separate informative feature layers and improve the hierarchical segmentation accuracy of remote sensing multispectral images.

INDEX TERMS Convolutional networks, deep learning, image segmentation, multiscale pyramid sieve, remote sensing.

I. INTRODUCTION

Segmentation is the primary task in remote sensing image processing. Prior to 2000, the segmentation methods used in remote sensing image processing can be categorized in to five types [1], i.e., threshold-based segmentation methods [2], region-based segmentation methods [3], [4], edge-based segmentation methods [5], texture features-based segmentation methods, and clustering-based segmentation methods. In 2006, Hinton [6] proposed deep neural network learning based on the human brain. Subsequently, the primary main segmentation methods moved toward graph theory-based methods, clustering-based methods, classification-based methods, and hybrid clustering and classification methods. With the fast development of deep learning theory, due to a lack of generalizability, traditional image segmentation methods, which rely on pure mathematical formula derivation, have been gradually replaced by deep neural network methods, especially for images with complex scenes. As shown in Table 1, the state-of-the-art deep neural network-based image segmentation method has differentiated hundreds of model structures [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29].

The associate editor coordinating the review of this manuscript and approving it for publication was Gerardo Di Martino.

TABLE 1. Typical deep learning-based image segmentation methods.

No	Segmentation method	Architecture	Contributions
1	AlexNet	CNN	Utilizes ReLU, dropout, LRN, and GPU
2	ZFNet	AlexNet	Deconvolution
3	VGGNet-16	VGGNet	Simple topology
4	GoogLeNet-Inception v1	Inception module	Inception
5	ResNet	Residual module	Gradient
6	Fully convolutional network [7]	VGG-16 (FCN)	Forerunner
7	U-net [8]	FCN	Complete set of encoders and decoder
8	Dilated convolutions [9]	FCN	Improved content aggregation without reducing resolution
9	RefineNet [10]	U-Net	Reduces memory usage, improves feature fusion in modules
10	PSPNet [11]	ResNet	Multi-receptive pyramid structure
11	Large kernel matters [12]	ResNet	Resolves “classification” and “partial”
12	SegNet [13]	VGG-16 + Decoder	Encoder-decoder
13	Bayesian SegNet [14]	SegNet	Uncertainty modeling
14	DeepLab [15]	VGG-16/ResNet-101	Standalone CRF, atrous convolutions
15	MINC-CNN [16]	GoogleNet (FCN)	Patchwise CNN, standalone CRF
16	CRFasRNN [17]	FCN-8s	CRF reformulated as RNN
17	Dilation [18]	VGG-16	Dilated convolutions
18	ENet [19]	ENet bottleneck	Bottleneck module for efficiency
19	ParseNet [20]	VGG-16	Global context feature fusion
20	ReSeg	VGG-16 + ReNet	Extension of ReNet to Semantic Segmentation
21	LSTM-CF [21]	Fast R-CNN + DeepMask	Fusion of contextual information from multiple sources
22	2D-LSTM [22]	MDRNN	Image context modeling
23	R CNN	MDRNN	Different input sizes, image context
24	DAG-RNN [23]	Elman network	Graph image structure for context modeling
25	SDS	R-CNN + BoxCNN	Simultaneous detection and segmentation
26	Deep mask [24]	VGG-A	Proposals generation for segmentation
27	Sharp mask [25]	DeepMask	Top-down refinement module
28	MultiPathNet [26]	Fast R-CNN + DeepMask	Multipath information flow through a network
29	Huang-3DCNN	3DCNN[27]	3DCNN for voxel point clouds
30	PointNet [28]	Own MLP-based	Segmentation of understanding point
31	Clockwork Convnet [29]	FCN	Clockwork scheduling for sequences
32	3DCNN-Zhang	Own 3DCNN	3D convolutions and graph cut for sequences

In the image segmentation task, context information has a significant impact on segmentation accuracy. For convolutional neural network-based segmentation algorithms, the acquisition of context information depends on the receptive field, i.e., the settings of the convolution kernel.

Among deep learning-based image segmentation methods, the AlexNet method [30] comprises five convolutional layers, three pooling layers, and three fully connected layers, and the sizes of the AlexNet convolution kernels in the convolutional layers are 11 × 11, 5 × 5, 3 × 3, 3 × 3, and 3 × 3. The ZFNet method, which is based on AlexNet, optimized the convolution kernel size of the first convolution layer to 7 × 7 and the stride, and, as a result, segmentation accuracy was improved [31]. For the Visual Geometry Group-16 (VGG-16) method [32], the 2014 ILSVRC (DET) champion, the hidden layers comprise thirteen convolutional layers, three fully connected layers, and five pooling layers. Here, the size of the convolution kernel in the convolutional layers is 3 × 3. Compared to AlexNet and ZFNet, this small kernel size and the multilayer nonlinear layers in the VGG method provide sufficient network depth to learn more complex patterns with fewer parameters. In addition, for the GoogLeNet-Inception v1 method, the kernel size in the first convolutional layer is 7 × 7. Compared to VGG, GoogLeNet has been shown to obtain good classification performance while controlling the number of calculations and parameters [33]. For the ResNet method [34], which was proposed in 2015, the kernel size in the first convolutional layer is also 7 × 7, and the kernel size in the remaining convolutional layers is 3 × 3. Segmentation algorithm models have constantly attempted to optimize the receptive field. For example, the dilated convolution technique [35] expands the receptive field by setting the interval in the convolution kernel element. In theory, the receptive field of algorithms, such as ResNet, can be larger than the size of the input image. However, Zhou [36] proved that the empirical receptive field of a CNN is smaller than the theoretical value. Thus, more effective and priori global descriptions are required in future research.

The pyramid scene parsing network (PSPNet) was proposed to solve this problem. PSPNet considers the context information under the multiscale receptive field (Fig. 1). PSPNet’s pyramid pooling module has four pooling sizes, i.e., 1 × 1, 2 × 2, 3 × 3, and 6 × 6. Then, 1 × 1 convolution layers are used to calculate the weight of each pyramid layer, and bilinear interpolation is utilized to restore the layers to the same size. Here, the size of the feature layers is 1/8 of the original image. Finally, all context information obtained via pooling is integrated through convolution to generate the

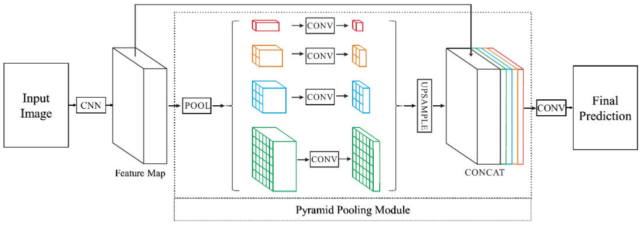

FIGURE 1. The architecture of PSPNet [11].

final segmentation result [37]. The PSPNet method considers the multiple scales of the receptive field but cannot filter the scales. However, the scale of ground objects varies greatly in informative remote sensing images, and the sieve and analysis of feature scales are particularly critical.

Thus, this paper proposes a multiscale pyramid sieve and analysis model based on PSPNet. The image's full-scale pyramid structure feature map is first constructed in this model. Then, the optimal scales set are screened through the sieve and analysis algorithm in the pyramid pooling module. Further convolution operations and feature extraction processes are then performed. Finally, image segmentation is performed.

PSPNet-based multiscale image segmentation methods are commonly used in conventional visual recognition tasks. However, for remote sensing images, the number of targeted studies has been increasing [38], [39], [40]. The proposed multiscale pyramid sieve and analysis strategy significantly improves the precision of multiscale image segmentation by 3.57%–5.89%. In addition, the proposed method is effective for remote sensing images, which have more complex semantic environments and larger scale differences.

The remainder of this paper is organized as follows. In Section II we provide the definitions of optimal scale and propose the sieve and analysis algorithm for optimal scales. To confirm the performance of the proposed algorithm, experiments in multispectral remote sensing images are presented in Section III. Section IV concludes the paper and discusses future research directions.

II. METHODOLOGY

In this section, we provide definitions for *multiscale* and *optimal scale*, propose the sieve and analysis algorithm for optimal scale, and give the verification process through single-scale images and recombination-scale images.

A. DEFINITION OF MULTISCALE AND OPTIMAL SCALE

Image multiscale refers to the spatial series $\{V_j\}_{j\in Z}$ in scale space $\mathrm{L}^2(\mathrm{R})$. The series satisfies the conditions of monotonicity, approximability, scalability, and translation invariance, and the Riesz base provides a generalized multiscale analysis theory that departs from the traditional theory. This new theory builds scales based on other scale functions in the scale space [41], [42]. (Fig. 2)

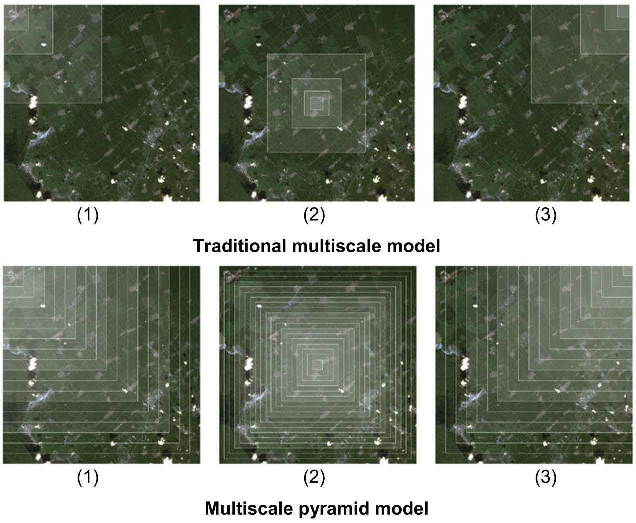

FIGURE 2. Different multiscale models. Traditional multiscale models have advantages in computation rate, decay rate, and the globality of information perception; however, the interval distance between two adjacent layers increases exponentially with the establishment of the multiscale model, which leads to missing information layers. Multiscale pyramid models are closer to the multiscale perception mechanism in human vision.

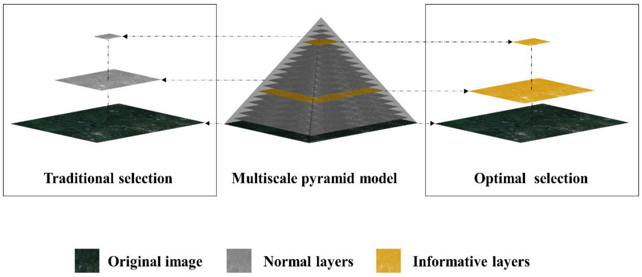

FIGURE 3. Traditional multiscale selection and optimal multiscale selection. The traditional selection method selects multiscale layers from a multiscale pyramid model with a fixed step. The optimal selection method sieves informative layers from a multiscale pyramid model with a variation function.

In generalized multiscale analysis theory and in scale space $\mathrm{L}^2(\mathrm{R})$, $\{V_j\}_{j\in Z}$ is a spatial series termed as $V_j \subset L^2\,(R)\,,\,,V_j{\subset}V_{j+1},j\in Z$. For each $\{V_j\}_{j\in Z}$, there exists $\varphi_j(x)$, $\left\{\varphi_j(x-k/2^j)\right\}_{k\epsilon Z}$, i.e., the orthonormal basis of V_j.

The optimal scale is formulated in single-scale image segmentation to detect and localize boundaries accurately. Here, the targets are segmented into one or more segments with low fragmentation and high feature homogeneity [43] (Fig. 3).

There are two modes to evaluate the optimal scales, i.e., posterior and prior models. The posterior model refers to images initially segmented under any scale and then brought into the evaluation system for segmentation results to filter out the optimal scale. The evaluation system includes qualitative evaluations, e.g., visual interpretation, and quantitative evaluations, e.g., homogeneity and inhomogeneity detection [44]. The prior model consists of various functions, local variance, objective functions, and RMAS detection.

Note that the posterior model is more applicable to traditional multiscale models and considers opportunity costs. In contrast, the prior model is more practical for continuous-scale convolutional deconstruction modes.

A variation function is employed to detect the spatial variation characteristics and spatial variation intensity of the regional variables. The equation is given as follows:

$$\boldsymbol{\gamma}\,(\boldsymbol{h},\boldsymbol{\alpha}) = \frac{\mathbf{1}}{\boldsymbol{V}}\int_{\boldsymbol{V}} [\boldsymbol{z}\,(\boldsymbol{x}) - \boldsymbol{z}(\boldsymbol{x}+\boldsymbol{h})]^{\mathbf{2}}\boldsymbol{dx} \tag{1}$$

here, γ is the continuous variable variation function, x is the regional variable, h is the delay distance, and α is the variable range.

In 1997, Atkinson and Kelly [45] introduced the variation function into image analysis. This function is used to reflect the relationship between the image scale and information distribution. This variation function is expressed as follows:

$$\gamma\,(h,\alpha) = \frac{1}{N(h)}\sum\nolimits_{i=1}^{N(h)} [z\,(x_i) - z(x_i+h)]^2 \tag{2}$$

here, γ is the discrete variable variation function, x is the regional variable, h is the delay distance, α is the variable range, and N is the resolution.

B. DEFINITION OF MOMENT

The moment is used to describe the characteristics of a random variable in mathematical statistics and applied to show the geometrical characteristics in image analysis. Stable geometrical characteristics are not affected by light, noise, and geometric deformation. Therefore, based on generalized multiscale analysis, we introduce the geometric moments for the subsequent multiscale analysis.

We set the gray distribution of the image target area D as:

$$f(x,y)\ \ (x,y)\in D \tag{3}$$

The origin moment of order p + q for D:

$$m_{pq} = \iint_D x^p y^q f(x,y)dxdy\ \ (p,q=0,1,2\cdots) \tag{4}$$

The central moment of order p + q for D:

$$\mu_{pq} = \iint_D (x-\bar{x})^p(y-\bar{y})^q f(x,y)dxdy\ \ (p,q=0,1,2\cdots) \tag{5}$$

The relationship between Origin moment and Central moment:

$$\mu_{pq} = \sum_{k=0}^{p}\sum_{l=0}^{q}\binom{p}{k}\binom{q}{l}(-1)^{k-l}m_{p-k,q-l}m_{10}^{k}m_{01}^{l}m_{00}^{-(k+l)} \tag{6}$$

The normalized central moments:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{r}}\ \ (r = \tfrac{p+q+2}{2}, p+q=2,3\cdots) \tag{7}$$

TABLE 2. Moment operators.

M	Moment Operator
M_a	$\phi_{01} = \eta_{01}+\eta_{10}$
M_b	$\phi_{02} = \eta_{01}+\eta_{10}$ （Noise Reduction）
M_c	$\phi_{03} = \eta_{11}$
M_d	$\phi_{H1} = \eta_{20}+\eta_{02}$
M_e	$\phi_{H2} = (\eta_{20}-\eta_{02})^2 + 4\eta_{11}^2$
M_f	$\phi_{H3} = (\eta_{30}-3\eta_{12})^2 + (\eta_{03}+3\eta_{21})^2$
M_g	$\phi_{H4} = (\eta_{30}+\eta_{12})^2 + (\eta_{21}+\eta_{03})^2$
M_h	$\phi_{H5} = (\eta_{30}-3\eta_{12})(\eta_{30}+\eta_{12})[(\eta_{30}+\eta_{12})^2-3(\eta_{21}+\eta_{03})^2]+$ $(3\eta_{21}-\eta_{03})(\eta_{21}+\eta_{03})[3(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2]$
M_i	$\phi_{H6} = (\eta_{20}-\eta_{02})[(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2]+$ $4\eta_{11}(\eta_{30}+\eta_{12})(\eta_{21}+\eta_{03})$
M_j	$\phi_{H7} = (3\eta_{21}-\eta_{03})(\eta_{30}+\eta_{12})[(\eta_{30}+\eta_{12})^2-3(\eta_{21}+\eta_{03})^2]+$ $(3\eta_{12}-\eta_{30})(\eta_{21}+\eta_{03})[3(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2]$

For the moment of the image, the lower its order, the more sensitive it is to the extreme value and the lower its computational complexity: $\phi_{01}, \phi_{02,}\phi_{03}$.

Hu. M. K proposed the concept of invariant moments., and constructed seven invariant moments using the second-order normalized central moments and the third-order normalized central moments: $\phi_{H1}, \phi_{H2,}\phi_{H3}, \phi_{H4}, \phi_{H5}, \phi_{H6}, \phi_{H7}$. Invariant moments have rotation invariance, scaling invariance and translation invariance. But in keeping image characteristics, $\phi_{H1}, \phi_{H2}, and\phi_{H3}$ show higher stability than $\phi_{H4}, \phi_{H5}, \phi_{H6}$, *and* ϕ_{H7}. Especially, ϕ_{H2}, ϕ_{H3} show well in keeping boundaries in multiscale image segmentation.

Considering the high requirements for the stability of multiscale image segmentation, this paper uses three sets of moments $\mathrm{M_c}$, $\mathrm{M_e}$, and $\mathrm{M_f}$ as the basic operators of the C. multiscale optimal scales sieve and analysis model, as shown in Table 2.

C. MULTISCALE OPTIMAL SCALES SIEVE AND ANALYSIS MODEL

There are different types of objects in remote sensing images. Objects with attributes, including scale, shape, pattern, and color, comprise an object system (As shown in Fig. 4: multiscale pyramid 1 - multiscale pyramid N). When guided by a specific application direction, an optimal scale always exists for mono-scale object systems.

However, the regularity between object and image scales is somewhat suppressed for images with a complex object system. The SingleScale image was developed previously to analyze the frequency between object and image scales. The complexity of image object systems is gradually increased to match the complexity of remote sensing images. Convolutional operators for feature maps are defined in Table 2.

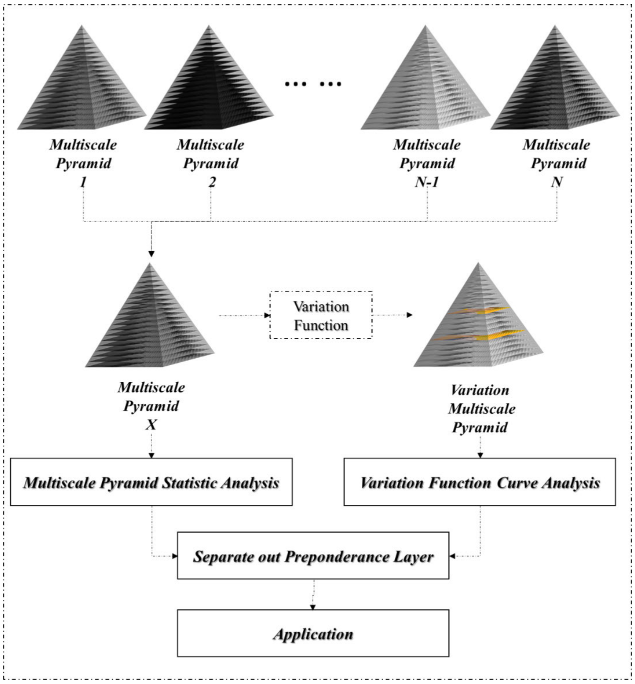

FIGURE 4. Multiscale Optimal scales sieve and analysis model.

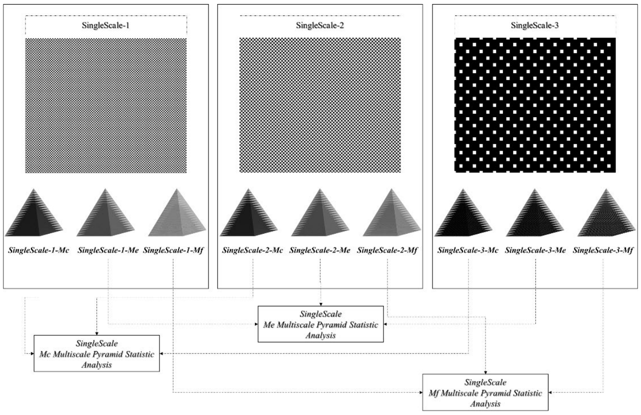

FIGURE 5. SingleScale set analysis model.

D. CONVOLUTIONAL MULTISCALE PYRAMID MODEL ANALYSIS FOR SINGLESCALE IMAGE

As shown in Fig. 5, the following set of SingleScale images was created:

{SingleScale-1; SingleScale-2; SingleScale-3}

Here, each image in the set contains one object type, and each object has the same scale, shape, and pattern. However, the overall scale is a unique variable among the images. Three continuous-scale convolutional deconstruction models were constructed for each image using three convolutional operators, i.e., M_c, M_e, and M_f, as shown in the equation at the bottom of the next page.

Statistical analysis was then performed for the three convolutional feature map frameworks M_c, M_e, and M_f, as shown in the equation at the bottom of the next page.

As shown in Fig. 6, the brightness distribution for 50 scale levels in the convolutional feature pyramid Mc reveals that there was only one steep section of change for all three SingleScale images. The wave was observed to be stable with scale-level growth. In addition, the steep section was followed to move toward a high scale level between images with increased object size. The amplitude also increased, and slight fluctuations were intensified around the steep section.

As shown in Fig. 7, the brightness distribution for 50 scale levels in the convolutional feature pyramid Me demonstrates that the three curves had similar trends (for scale-level growth) for all three SingleScale images. Here, there was a single change section, and the wave was observed to be stable. The first peak was observed to move toward a high scale level with the increase in object size between images. In addition, the amplitude revealed no significant difference, and the wavelength was increased. The second wave was observed to move toward a higher scale level, and the amplitude and wavelength increased.

As shown in Fig. 8, the brightness distribution for 50 scale levels in the M_f convolutional feature pyramid revealed that the three curves for all SingleScale images exhibited similar trends as the Me convolutional feature pyramid. Here, several effects occurred with the increased object size among images. The first peak reached a high scale level, and the amplitude showed no significant difference. The wavelength was also observed to increase. We found that the second wave advanced toward a higher scale level, and the amplitude and wavelength increased.

Thus, we conclude the following from the analysis of the statistical information for the SingleScale images in the three convolutional feature pyramids.

1) The information distribution changes with image convolutional feature pyramid scales in a nonuniform variation mode. A significant difference may exist between adjacent scale levels for statistical information in the images.

2) The distribution of information has different features under different convolutional feature pyramids.

3) There is a similar distribution of information under the same convolutional feature pyramid when the object scale is a unique variable between images.

4) The object scales can influence the information distribution in the image's convolutional feature pyramid.

E. CONVOLUTIONAL MULTISCALE PYRAMID MODEL ANALYSIS FOR RECOMBINATION-SCALE IMAGE

A set of SingleScale images was created, including

{SingleScale-1, SingleScale-2, SingleScale-3}.

Here, each image in the set has a single object type, and each object has the same scale, shape, and pattern. In addition, the scale of the objects is a unique variable between images. A set of RecombinationScale images was created based on the SingleScale image set, including:

RecombinationScale-1:
{RecombinationScale-1) | SingleScale-2 ∪ SingleScale-3}
RecombinationScale-2:
{RecombinationScale-2) | SingleScale-1 ∪ SingleScale-3}
(Fig. 9)

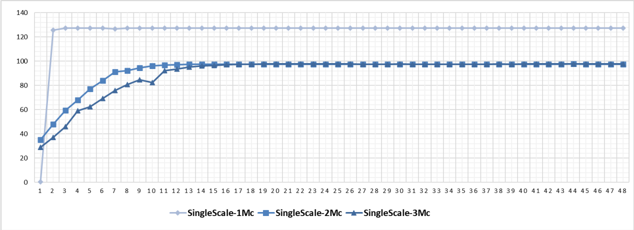

FIGURE 6. Distribution of SingleScale set brightness in the Mc model.

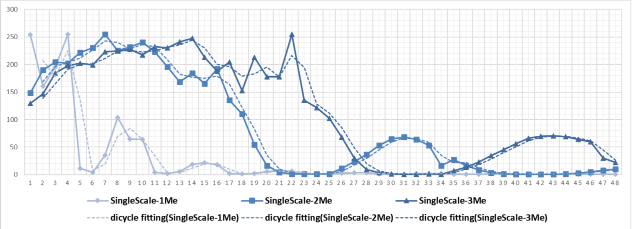

FIGURE 7. Distribution of SingleScale set brightness in the Me model.

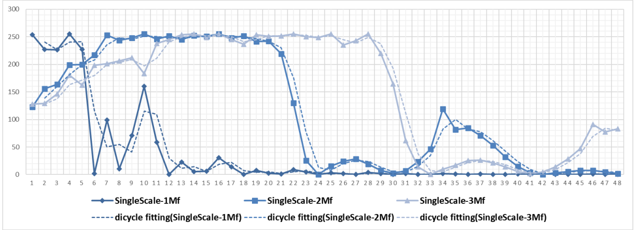

FIGURE 8. Distribution of SingleScale set brightness in the Mf model.

Two convolutional feature pyramid models were constructed for images using operators $\mathrm{M_e}$ and $\mathrm{M_f}$, including: SingleScale-1, SingleScale-2, RecombinationScale-1, and RecombinationScale-2, as shown in the equation at the bottom of the next page.

A statistical analysis was performed in two groups for the Me and Mf convolutional feature pyramid.

Group <1>

$$\begin{cases}\text{Me : SingleScale\_2\_Me; RecombinationScale\_1\_Me} \\ \text{Mf : SingleScale\_2\_Mf; RecombinationScale\_1\_Mf}\end{cases}$$

Group <2>

$$\begin{cases}\text{Me : SingleScale\_1\_Me; RecombinationScale\_2\_Me} \\ \text{Mf : SingleScale\_1\_Mf; RecombinationScale\_2\_Mf}\end{cases}$$

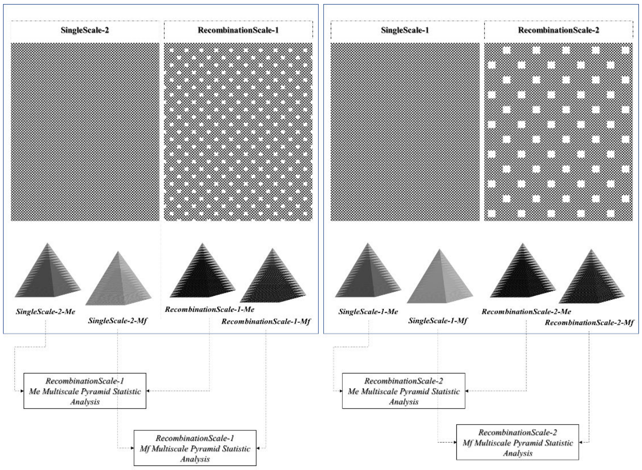
SingleScale-2
RecombinationScale-1
SingleScale-1
RecombinationScale-2
SingleScale-2-Me
SingleScale-2-Mf
RecombinationScale-1-Me
RecombinationScale-1-Mf
SingleScale-1-Me
SingleScale-1-Mf
RecombinationScale-2-Me
RecombinationScale-2-Mf
RecombinationScale-1 Me Multiscale Pyramid Statistic Analysis
RecombinationScale-1 Mf Multiscale Pyramid Statistic Analysis
RecombinationScale-2 Me Multiscale Pyramid Statistic Analysis
RecombinationScale-2 Mf Multiscale Pyramid Statistic Analysis

FIGURE 9. Recombination-scale set analysis model.

Group <1>

The brightness distribution for 50 scale levels in the Me and Mf convolutional feature pyramids was assessed. The two curves exhibited a similar trend when in the same convolutional feature pyramid for the SingleScale-2 and RecombinationScale-1 images (Fig. 10 and Fig. 11, respectively). The brightness values for the SingleScale image were higher than those of the RecombinationScale image before level 20, and the opposite was observed after level 20. In addition, there was a more significant amount of fluctuation in the RecombinationScale image curve.

Group <2>

The brightness distribution for 50 scale levels in the Me and Mf convolutional feature pyramids was assessed, and we found that the two curves exhibited a similar trend for images SingleScale-1 and RecombinationScale-2 when in the same convolutional feature pyramid. The brightness values for the SingleScale image were higher than the RecombinationScale image before level 10, and the RecombinationScale image featured higher brightness values after level 10. We also observed a more significant fluctuation in the RecombinationScale image curve. (Fig. 12, Fig. 13)

The analysis of the statistical information for the RecombinationScale images in the two convolutional feature pyramids produced the following conclusions.

1) The information distribution changes with image convolutional feature pyramid scales in a nonuniform variation

$$\begin{cases} \boldsymbol{SingleScale-1: SingleScale-1-Mc; SingleScale-1-Me; SingleScale-1-Mf} \\ \boldsymbol{SingleScale-2: SingleScale-2-Mc; SingleScale-2-Me; SingleScale-2-Mf} \\ \boldsymbol{SingleScale-3: SingleScale-3-Mc; SingleScale-3-Me; SingleScale-3-Mf} \end{cases}$$

$$\begin{cases} \boldsymbol{Mc: SingleScale-1-Mc; SingleScale-2-Mc; SingleScale-3-Mc} \\ \boldsymbol{Me: SingleScale-1-Me; SingleScale-2-Me; SingleScale-3-Me} \\ \boldsymbol{Mf: SingleScale-1-Mf; SingleScale-2-Mf; SingleScale-3-Mf} \end{cases}$$

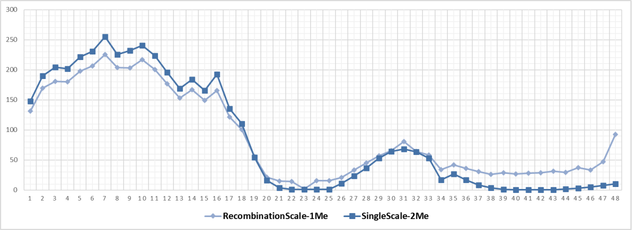

FIGURE 10. Brightness distribution for recombination-scale set 1 in the M_e model.

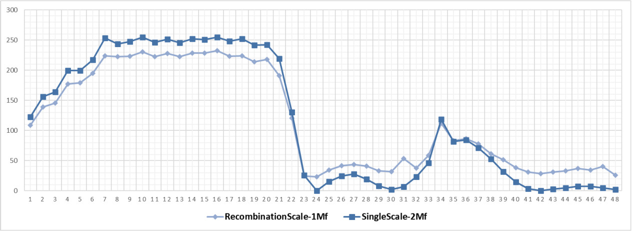

FIGURE 11. Brightness distribution for recombination-scale set 1 in the Mf model.

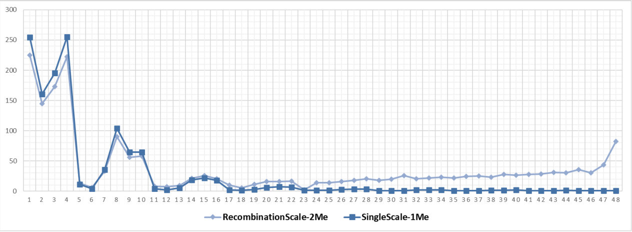

FIGURE 12. Brightness distribution for RecombinationScale set 2 in the Me model.

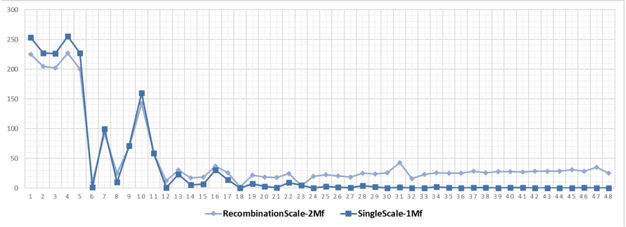

FIGURE 13. Brightness distribution for RecombinationScale set 2 in the Mf model.

mode. In addition, a significant difference may exist between adjacent scale levels in the statistical information in images.

2) The distribution of information has different features under different convolutional feature pyramids.

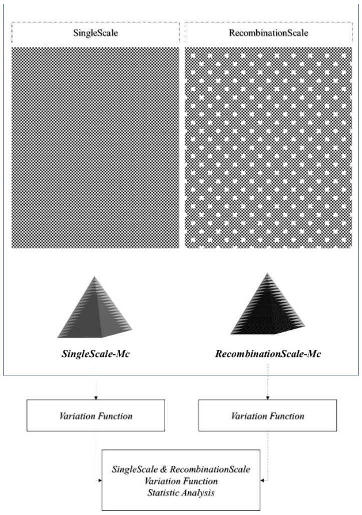

FIGURE 14. Variation function analysis model.

3) There is a similar distribution of information for the SingleScale and RecombinationScale images while under the same convolutional feature pyramid.

4) There is a feature level for overlaying multiple mono-scale object systems. Here, the statistical information is suppressed and promoted before and after the feature level, respectively. The scale of the superimposed mono-scale object system influences the position of the feature level.

F. VARIATION FUNCTION ANALYSIS

A set of SingleScale images were created, including (Fig. 14): {SingleScale-1;SingleScale-2}.

Here, each image in the set has a single object type, and each object has the same scale, shape, and pattern. The scale of the objects is a unique variable between the images.

The following set of RecombinationScale images was created based on a single-scale image set:

RecombinationScale-1:

$$\{RecombinationScale\_1 \mid SingleScale\_1 \cup SingleScale\_2\}$$

Using the Mc operator, a convolutional feature pyramid model was created for SingleScale-2 and RecombinationScale-1 images.

$$\{\text{SingleScale\_2} : \text{SingleScale\_2\_Mc}$$
$$\{\text{RecombinationScale}_1 : \text{RecombinationScale\_1\_Mc}$$

Variation function analysis was then performed for the Mc convolutional feature pyramid.

$$\{\text{SingleScal\_2} : \text{SingleScale\_2\_Me; SingleScale\_2\_Mf}$$
$$\{\text{SingleScal\_1} : \text{SingleScale\_1\_Me; SingleScale\_1\_Mf}$$
$$\begin{cases}\text{RecombinationScale\_1} : \text{RecombinationScale\_1\_Me; RecombinationScale\_1\_Mf}\\ \text{RecombinationScale\_2} : \text{RecombinationScale\_2\_Me; RecombinationScale\_2\_Mf}\end{cases}$$

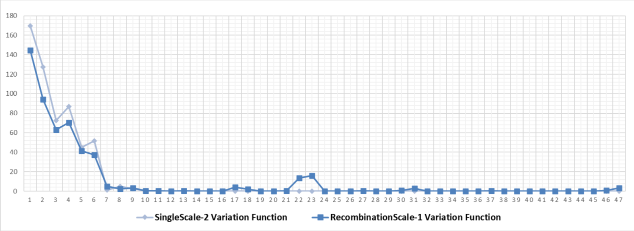

FIGURE 15. Variation function analysis.

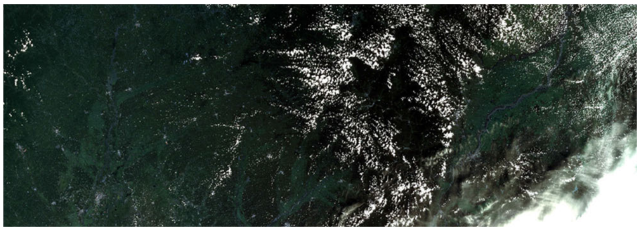

FIGURE 16. Remote sensing satellite multispectral image.

A variation function was superimposed on two Mc models. The variation function curve for the SingleScale and RecombinationScale images was compared within the convolutional feature pyramid (Fig. 15). We found that both curves exhibited the same trend except for an extra peakon the latter curve, which indicates a steep change in brightness value. The additional wave was revealed on level 20, which is the same as the feature level of group <1> in the RecombinationScale image convolutional feature pyramid analysis. According to the result of the analysis, the informative scales demonstrate steep changes in the variation curve. Thus, we can fix the informative levels using the variation curve analysis.

Hence, the optimal multiscale pyramid sieve analysis mechanism was considered to involve the following steps: (1) First, a multiscale feature map pyramid is built. (2) Sieve optimal scales (informative feature layers) from the feature map pyramid with variation analysis. (3) Fuse optimal scales (informative feature layers). (4) The final feature representation is then input into subsequent calculations.

III. APPLICATIONS AND ANALYSIS

A prototype of the proposed pyramid multiscale sieve analysis model was implemented to evaluate its feasibility and performance. Here the proposed model was applied to a remote sensing image, and an optimal-scale image was included in the multiscale image segmentation algorithm. We then evaluated the segmentation results using different evaluation indices.

A. DATA PREPARATION

The remote sensing image used in the verification was acquired from the National Defense Technology Industry Administration of China and the National Space Administration of China (Fig. 16).

The dataset includes remote sensing satellite multispectral images with eight bands (covering 850 kilometers × 300 kilometers) and ground truth data (main target including corn, soybeans, and rice), used explicitly for crop type recognition. (https://dianshi.bce.baidu.com/)

B. ANALYSIS

Overview of the proposed model: (1) remote sensing image preprocessing, (2) convolutional calculation of feature map, (3) multiscale pyramid sieve and analysis pool, (4) upsampling and multiscale convolution, (5) evaluation.

In this experiment, the remote sensing satellite multispectral images included band alpha 1, band alpha 2, band alpha 3, band alpha 4, band alpha5, band R, band B, and band G. The size of the experiment image slices was 1024 × 1024 pixels. After the multispectral remote sensing image was preprocessed, the CNNs extracted feature maps for each spectral. The convolution kernel included $\mathrm{M_a}$, $\mathrm{M_b}$, $\mathrm{M_c}$, $\mathrm{M_d}$, $\mathrm{M_e}$, $\mathrm{M_f}$, $\mathrm{M_g}$, $\mathrm{M_h}$, $\mathrm{M_i}$, and $\mathrm{M_j}$, which focus on different remote sensing features. Here, $\mathrm{M_a}$ and $\mathrm{M_b}$ focused on texture details, and $\mathrm{M_h}$, $\mathrm{M_i}$, and $\mathrm{M_j}$ contained fragmentation information. The operators $\mathrm{M_c}$, $\mathrm{M_e}$, $\mathrm{M_f}$, and $\mathrm{M_g}$ were sensitive to boundary information, which is essential for the segmentation task.

Given the feature maps, as shown in Fig. 18, we utilized the multiscale pyramid sieve and analysis pool (Fig. 17) to gather image features. Note that the pyramid pool fuses features under different pyramid scales.

Take the $\mathrm{M_c}$ feature pyramid of band alpha 3 as an example, this paper calculated the brightness value of the feature pyramid from levels 1 to 50, shown in Fig. 19, which revealed that the curve was steady with no steep changes and had several small fluctuations. Note that this result does not clearly suggest excessive information enrichment at a specific scale. A variation function was subsequently applied to the statistical results. As shown in Fig. 20, after the variation calculation, informative levels were labeled (Level<2>, Level<4>, Level<11>, Level<23>, Level<34>). In consideration of the real ground target's granularity and maintaining a reasonable representation gap, we set a threshold before screening out unsuitable levels (Level<23> and Level<34> were coarse, and Level<2> and Level<4> were too close to the original image under this rule). Note that Level<11> was labeled at last.

The sieve and analysis process was repeated on all feature maps of all bands. Then, the upsampling and concatenation layers were used to form the final feature representation, which carried local and global context information. Finally, the representation was input to a convolution layer and filters to obtain the final prediction.

Thus, we conducted a control group experiment to verify the effectiveness of the multiscale pyramid sieve and analysis pool. Here, the experimental procedures and parameters were unchanged, and we removed the sieve and analysis process for informative levels. To ensure the strictness of the experiment, the control groups covered levels from Level<3> to Level<22>, in which all feature maps of all bands used the

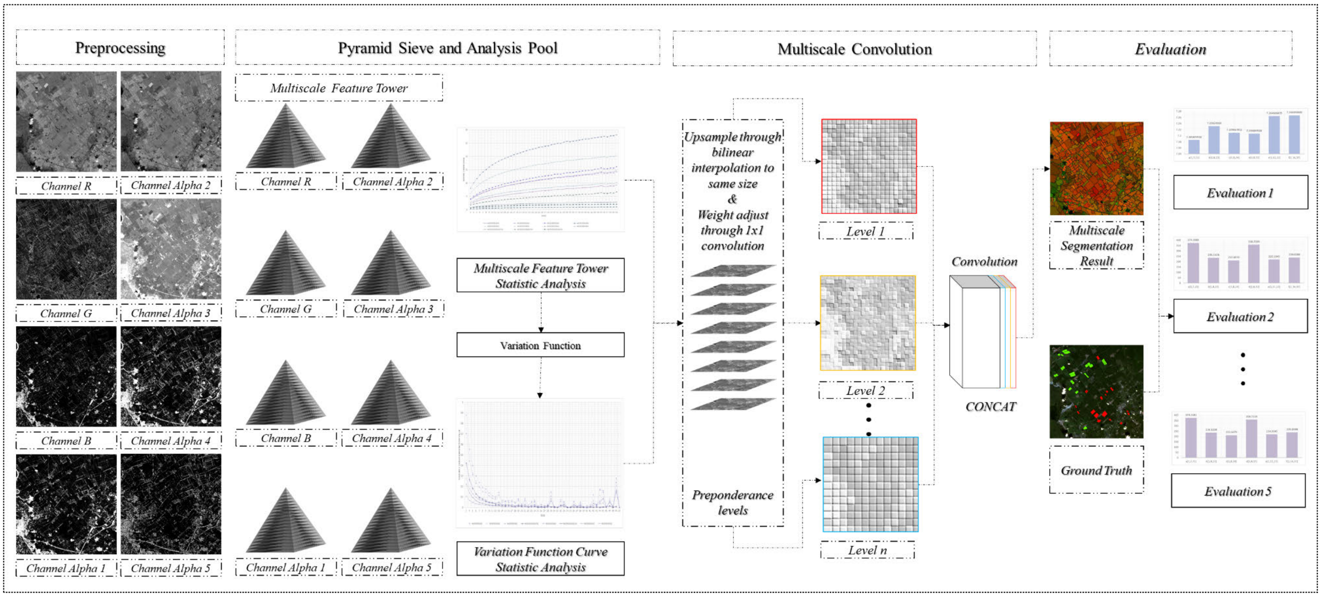

FIGURE 17. Remote sensing image segmentation model.

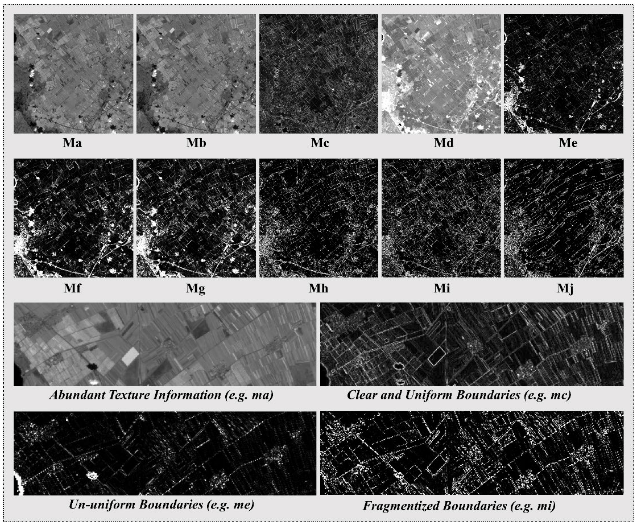
Ma
Mb
Mc
Md
Me
Mf
Mg
Mh
Mi
Mj
Abundant Texture Information (e.g. ma)
Clear and Uniform Boundaries (e.g. mc)
Un-uniform Boundaries (e.g. me)
Fragmentized Boundaries (e.g. mi)

FIGURE 18. Convolutional feature maps.

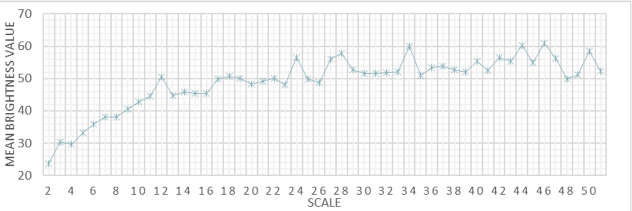

FIGURE 19. Distribution of brightness values for band Alpha 3 in Mc.

same level.

C-level<03> C-level<04> C-level<05> C-level<06>
C-level<07> C-level<08> C-level<09> C-level<10>
C-level<12> C-level<13> C-level<14> C-level<15>
C-level<16> C-level<17> C-level<18> C-level<19>
C-level<20> C-level<21> C-level<22>

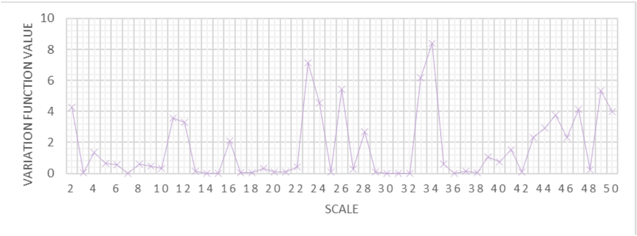

FIGURE 20. Variation function for band Alpha 3 in Mc.

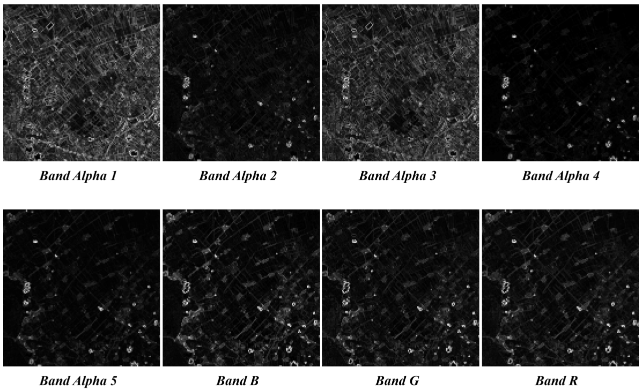

FIGURE 21. Level<11> feature maps of eight bands in Mc. As mentioned previously, the distribution of information has different features under different convolutional feature pyramids. Thus, the sieve and analysis process should be repeated.

C. RESULTS AND EVALUATION

Multispectral remote sensing image segmentation results were shown in Fig. 22, which includes 20 sub-images:

Experiment group:

O-level<11>

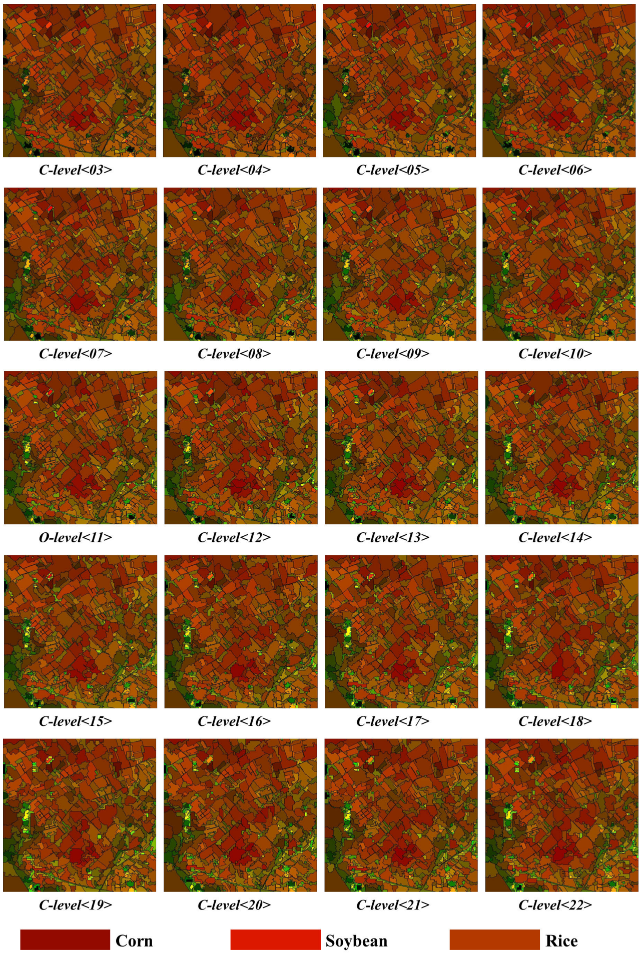

FIGURE 22. Multispectral remote sensing image segmentation results.

Control group:

C-level<03> C-level<04> C-level<05> C-level<06>
C-level<07> C-level<08> C-level<09> C-level<10>
C-level<12> C-level<13> C-level<14> C-level<15>
C-level<16> C-level<17> C-level<18> C-level<19>
C-level<20> C-level<21> C-level<22>

Because the Multispectral remote sensing image segmentation results contain many broken patches, we cannot visually observe the difference between the experiment group and the control group. Then we introduced the ground truth data and evaluation indexes.

In the ground truth data, red represents corn, blue represents soybeans, and green represents rice. In this study, we converted the ground truth data into binary images for a subsequent evaluation (Fig. 23).

In this evaluation, we considered seven different evaluation indices, including the Dice ratio, the Hausdorff distance, the Jaccard index, the average perpendicular distance, the conformity coefficient, precision, and recall, which are described as follows.

1) The Dice ratio, which is derived from a reliability measure known as the kappa statistic, computes the balance of the intersection area divided by the mean sum of each individual area [46], [47]. With this evaluation index, the optimal group O-level<11>, compared to the control groups, demonstrated the highest similarity (86%) with the ground truth data (Table 3).

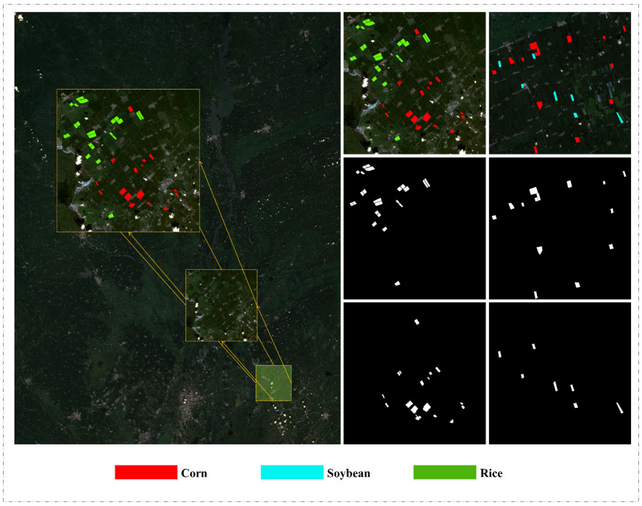

FIGURE 23. Ground truth data and conversion process.

2) The Hausdorff distance [48], which is a measure of the maximum discrepancy between two true subsets in space, demonstrates the inhomogeneity between segmented blocks in image segmentation. A higher Hausdorff distance value indicates higher inhomogeneity between segmented blocks. Here, we found that the optimal group O-level<11> obtained the shortest Hausdorff distance value (333.00); thus, it exhibited the lowest intraclass chaos. Note that C-level<10> and C-level<12> (361.26 and 361.51) demonstrated poor performance. Accordingly, compared to the control groups, intraclass chaos was reduced by 1.5%–7.9% in the experiment group (Table 3).

3) The Jaccard index measures the ratio of the intersection area of two sets divided by the area of their union [49]. Note that the Jaccard index is similar to the Dice ratio. Here, we found that, compared to the control groups, the optimal group O-level<11> obtained the highest Jaccard index value (0.45) and demonstrated the highest similarity with the ground truth data (Table 3).

4) The average perpendicular index measures the average vertical distance between the experimental result and the ground truth data. Here, compared to the control groups, the optimal group O-level<11> exhibited the shortest distance 76.58 (Table 3).

5) The conformity coefficient is a global similarity coefficient that is used to measure the ratio of correctly segmented pixels to the number of incorrectly segmented pixels. Here, the results were similar to those for the Dice ratio and Jaccard index. We found that the optimal group O-level<11> obtained a high conformity coefficient value

TABLE 3. Multispectral remote sensing image segmentation evaluation.

Number		Different Scales			
Classification Result	Evaluation index	C-level<3>	C-level<4>	C-level<5>	C-level<6>
No[3]C[3]F[c]E[d]	Dice Ratio	0.838964921	0.843697855	0.843672065	0.814768453
No[3]C[3]F[c]E[h]	Hausdorff Distance	336.5249	339.29929	333.60007	333.0015
No[3]C[3]F[c]E[j]	Jaccard Index	0.437849478	0.44042835	0.440414294	0.424685233
No[3]C[3]F[c]E[a]	Avg_Perpen Distance	81.287849	84.544876	80.83878	79.522629
No[3]C[3]F[c]E[c]	Conformity Coefficient	19.16110102	19.29483127	19.29410664	18.45314878
No[3]C[3]F[c]E[p]	Precision Index	0.438380497	0.44092571	0.441036622	0.425580153
No[3]C[3]F[c]E[r]	Recall Index	0.973079538	0.975028325	0.968955359	0.95282121
Number		Different Scales			
Classification Result	Evaluation index	C-level<7>	C-level<8>	C-level<9>	C-level<10>
No[3]C[3]F[c]E[d]	Dice Ratio	0.815042389	0.825425563	0.815060039	0.809807833
No[3]C[3]F[c]E[h]	Hausdorff Distance	335.5965	338.8923	338.8923	361.26028
No[3]C[3]F[c]E[j]	Jaccard Index	0.424834084	0.430479208	0.424843675	0.421990476
No[3]C[3]F[c]E[a]	Avg_Perpen Distance	81.09906	82.262383	81.838707	82.658691
No[3]C[3]F[c]E[c]	Conformity Coefficient	18.46139899	18.77007437	18.46193037	18.30278272
No[3]C[3]F[c]E[p]	Precision Index	0.425789543	0.431194746	0.425583174	0.423069885
No[3]C[3]F[c]E[r]	Recall Index	0.949830048	0.962882393	0.960707002	0.94298663
Number		Different Scales			
Classification Result	Evaluation index	O-level<11>	C-level<12>	C-level<13>	C-level<14>
No[3]C[3]F[c]E[d]	Dice Ratio	0.855492875	0.817934341	0.807905156	0.832245444
No[3]C[3]F[c]E[h]	Hausdorff Distance	333.0015	361.50797	333.0015	336.2038
No[3]C[3]F[c]E[j]	Jaccard Index	0.445958446	0.426405767	0.420957255	0.43419037
No[3]C[3]F[c]E[a]	Avg_Perpen Distance	76.582077	79.001259	79.203712	80.162552
No[3]C[3]F[c]E[c]	Conformity Coefficient	19.26203404	18.54815957	18.24461895	18.96862782
No[3]C[3]F[c]E[p]	Precision Index	0.446617123	0.426964499	0.421734941	0.43493404
No[3]C[3]F[c]E[r]	Recall Index	0.966553365	0.970224337	0.958033084	0.962111942
Number		Different Scales			
Classification Result	Evaluation index	C-level<15>	C-level<16>	C-level<17>	C-level<18>
No[3]C[3]F[c]E[d]	Dice Ratio	0.834096713	0.815139467	0.831812664	0.820288476
No[3]C[3]F[c]E[h]	Hausdorff Distance	333.0015	333.0015	333.0015	338.8923
No[3]C[3]F[c]E[j]	Jaccard Index	0.435198227	0.424886835	0.433954786	0.427685513
No[3]C[3]F[c]E[a]	Avg_Perpen Distance	78.063805	79.293564	81.569107	77.269958
No[3]C[3]F[c]E[c]	Conformity Coefficient	19.02196521	18.46432137	18.95612466	18.61833357
No[3]C[3]F[c]E[p]	Precision Index	0.435748309	0.425772655	0.43468563	0.428426839
No[3]C[3]F[c]E[r]	Recall Index	0.97181056	0.953319737	0.96270111	0.961114888
Number		Different Scales			
Classification Result	Evaluation index	C-level<19>	C-level<20>	C-level<21>	C-level<22>
No[3]C[3]F[c]E[d]	Dice Ratio	0.822385691	0.814696424	0.823145916	0.828458626
No[3]C[3]F[c]E[h]	Hausdorff Distance	336.04761	335.05969	333.28967	335.05969
No[3]C[3]F[c]E[j]	Jaccard Index	0.428825858	0.424646095	0.429239286	0.432129379
No[3]C[3]F[c]E[a]	Avg_Perpen Distance	80.574104	76.529305	81.677155	82.656616
No[3]C[3]F[c]E[c]	Conformity Coefficient	18.68051073	18.45097853	18.70297123	18.85878237
No[3]C[3]F[c]E[p]	Precision Index	0.429697857	0.425515573	0.429867945	0.432678915
No[3]C[3]F[c]E[r]	Recall Index	0.954815318	0.954090188	0.967051892	0.971447995

(19.26), and the intraclass conformity was improved by 2.74%–5.58%.

6) Precision is the ratio of the number of correct-segmented pixels to the number of correct-segmented and incorrect-segmented pixels. As shown in Fig. 24, the precision index curve shows evaluation results similar to the Dice ratio index and Jaccard index. The optimal group O-level<11>, compared with control groups, has the highest precision (0.45), which improved by 3.57%–5.89% (Table 3).

7) Recall is the ratio of the number of correctly segmented pixels to the number of correctly segmented and correctly unsegmented pixels. Here, the optimal and control groups obtained similar results (0.97), except C-level<10> (0.94) (Table 3).

The results shown in Table 3 and Fig. 24 demonstrate that the proposed multiscale pyramid sieve and analysis module outperformed other methods, based on the PSPNet framework, with notable advantages in multispectral remote sensing image segmentation.

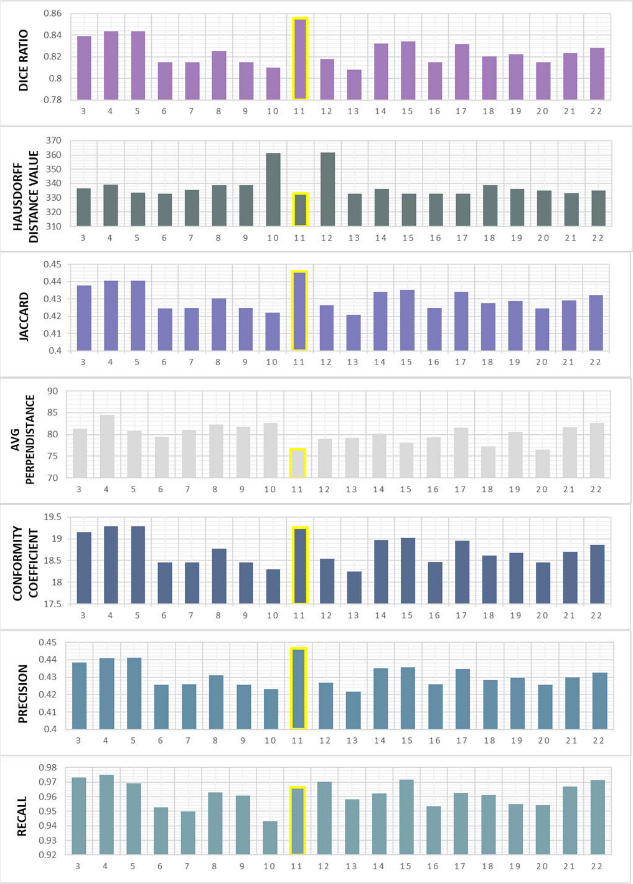

FIGURE 24. Multispectral remote sensing image segmentation evaluation histogram in 3–22 scales.

IV. CONCLUSION

In this paper, we have proposed a multiscale pyramid sieve and analysis module. The proposed module is based on the observation that the feature map information distribution changes with convolutional scales in a nonuniform variation mode and different feature map pyramids have different distributions. Based on CNN feature maps, the proposed multiscale pyramid sieve and analysis pool fuses features under different scales of different feature pyramids through variation analysis to obtain the final feature representation, which carries local and global context information. Experimental results obtained on authentic multispectral remote sensing images verified the effectiveness of the proposed module. An experimental segmentation evaluation demonstrated that, compared to the control group, the precision of the experimental group was improved by 3.57%–5.89%, the intraclass conformity was improved by 2.74%–5.58%, and the intraclass chaos was reduced by 1.5%–7.9%. Overall, the experimental results reveal that the proposed module can improve both segmentation flexibility and precision. In the future, we plan to apply the proposed module to segment different types of remote sensing images and other tasks, e.g., feature extraction and target identification.

A potential limitation of this study was that remote sensing images have the multispectral resolution, multi-time resolution, and multi-spatial resolution, and constructing a multiscale pyramid will increase the amount of data geometrically. In practical applications, massive data operations incur high computational costs, which will affect the efficiency of remote sensing data processing. Thus, in the future, we plan to investigate ways to construct a lightweight pyramid model and simplify the entire module.

REFERENCES

[1] L. Zhou and F. Jiang, "Survey on image segmentation method," *Appl. Res. Comput.*, vol. 34, no. 7, pp. 1921–1928, 2017.

[2] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 1, pp. 285–296, 1975.

[3] R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 6, pp. 641–647, Jun. 1994, doi: 10.1109/34.295913.

[4] F. Meyer, "Skeletons and watershed lines in digital spaces," in *Proc. SPIE*, 1990, pp. 85–102.

[5] L. S. Davis, "A survey of edge detection techniques," *Comput. Graph. Image Process.*, vol. 4, no. 3, pp. 248–270, 1975, doi: 10.1016/0146-664X(75)90012-X.

[6] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006, doi: 10.1162/neco.2006.18.7.1527.

[7] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[8] J. Li, K. V. Sarma, K. C. Ho, A. Gertych, B. S. Knudsen, and C. W. Arnold, "A multi-scale U-Net for semantic segmentation of histological images from radical prostatectomies," in *Proc. AMIA Annu. Fall Symp.*, vol. 2017, 2017, pp. 1140–1148.

[9] Q. Zhang, Q. Yuan, J. Li, Z. Yang, and X. Ma, "Learning a dilated residual network for SAR image despeckling," *Remote Sens.*, vol. 10, no. 2, p. 196, Jan. 2018, doi: 10.3390/rs10020196.

[10] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "RefineNet: Refining object detectors for autonomous driving," *IEEE Trans. Intell. Vehicles*, vol. 1, no. 4, pp. 358–368, Dec. 2016, doi: 10.1109/TIV.2017.2695896.

[11] H. Fang and F. Lafarge, "Pyramid scene parsing network in 3D: Improving semantic segmentation of point clouds with multi-scale contextual information," *ISPRS J. Photogramm. Remote Sens.*, vol. 154, pp. 246–258, Aug. 2019, doi: 10.1016/j.isprsjprs.2019.06.010.

[12] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4353–4361.

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

[14] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder–decoder architectures for scene understanding," *Comput. Sci.*, vol. abs/1511.02680, Nov. 2015.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[16] H. Zhang, "Deep TEN: Texture encoding network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nov. 2017, pp. 2896–2905.

[17] Z. Yan, H. Zhang, Y. Jia, T. Breuel, and Y. Yu, "Combining the best of convolutional layers and recurrent layers: A hybrid network for semantic segmentation," 2016, *arXiv:1603.04871*.

[18] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7268–7277.

[19] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.0214*.

[20] W. Liu, A. Rabinovich, and A. C. Berg, "Looking wider to see better," in *Proc. ICLR*, Jun. 2015, doi: 10.48550/arXiv.1506.04579.

[21] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, and L. Lin, "LSTM-CF: Unifying context modeling and fusion with LSTMs for RGB-D scene labeling," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016.

[22] W. Byeon and T. M. Breuel, "Supervised texture segmentation using 2D LSTM networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2015, pp. 4373–4377.

[23] P. Baldi and G. Pollastri, "The principled design of large-scale recursive neural network architectures–DAG-RNNs and the protein structure prediction problem," *J. Mach. Learn. Res.*, vol. 4, pp. 575–602, Dec. 2003.

[24] P. O. Pinheiro, R. Collobert, and P. Dollár, "Learning to segment object candidates," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–9.

[25] P. O. Pinheiro, T. Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *Proc. 14th Eur. Conf.*, 2016, pp. 75–91.

[26] S. Zagoruyko, A. Lerer, T. Y. Lin, P. O. Pinheiro, S. Gross, S. Chintala, and P. Dollár, "A multipath network for object detection," 2016, *arXiv:1604.02135*.

[27] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

[28] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 652–660.

[29] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2017, pp. 405–420.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[31] M. D. Zeiler and R. Fergus, "Stochastic pooling for regularization of deep convolutional neural networks," 2013, *arXiv:1301.3557*.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[35] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2016, *arXiv:1511.07122*.

[36] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," 2014, *arXiv:1412.6856*.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2014, doi: 10.1109/TPAMI.2015.2389824.

[38] Y. Liu, L. Gross, Z. Li, X. Li, X. Fan, and W. Qi, "Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder–decoder with spatial pyramid pooling," *IEEE Access*, vol. 7, pp. 128774–128786, 2019, doi: 10.1109/ACCESS.2019.2940527.

[39] Y. Liu, J. Zhou, W. Qi, X. Li, L. Gross, Q. Shao, Z. Zhao, L. Ni, X. Fan, and Z. Li, "ARC-Net: An efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 8, pp. 154997–155010, 2020, doi: 10.1109/ACCESS.2020.3015701.

[40] J. Zhou, Y. Liu, G. Nie, H. Cheng, X. Yang, X. Chen, and L. Gross, "Building extraction and floor area estimation at the village level in rural China via a comprehensive method integrating UAV photogrammetry and the novel EDSANet," *Remote Sens.*, vol. 14, no. 20, p. 5175, Oct. 2022, doi: 10.3390/rs14205175.

[41] D. Y. S. Yun, "Research on evaluation model of human vision in stereo video based on multi-scale analysis and similarity," M.S. thesis, Jilin Univ., Changchun, China, 2011, p. 72.

[42] G. Y. Jing, "General multiscale analysis and Isocratic multiscale analysis and application," M.S. thesis, Jilin Univ., China, 2000.

[43] L. X. Cheng and L. I. Y. Shu, "A study of the segmentation scale of high-resolution remotely sensed data in Chengdu plain," *Remote Sens. Land Resour.*, vol. 2, pp. 7–11, Jun. 2010.

[44] L. Wang and Q. Y. Liu, "The methods summary of optimal segmentation scale selection in high-resolution remote sensing images multi-scale segmentation," *Geomatics Spatial Inf. Technol.*, vol. 38, no. 3, pp. 166–169, Mar. 2015.

[45] P. M. Atkinson and R. E. J. Kelly, "Scaling-up point snow depth data in the U.K. for comparison with SSM/I imagery," *Int. J. Remote Sens.*, vol. 18, no. 2, pp. 437–443, Jan. 1997, doi: 10.1080/014311697219178.

[46] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, "Morphometric analysis of white matter lesions in MR images: Method and validation," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 716–724, Dec. 1994, doi: 10.1109/42.363096.

[47] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. Tempany, M. R. Kaus, S. J. Haker, W. M. Wells III, F. A. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index," *Acad. Radiol.*, vol. 11, no. 2, pp. 178–189, Feb. 2004, doi: 10.1016/s1076-6332(03)00671-8.

[48] H. H. Chang, A. H. Zhuang, D. J. Valentino, and W. C. Chu, "Performance measure characterization for evaluating neuroimage segmentation algorithms," *NeuroImage*, vol. 47, no. 1, pp. 122–135, 2009, doi: 10.1016/j.neuroimage.2009.03.068.

[49] P. Jaccard, "The distribution of flora in the Alpine zone," *New Phytol.*, vol. 11, no. 2, pp. 37–50, 1912, doi: 10.1111/j.1469-8137.1912.tb05611.x.

DI CAO was born in Henan, China, in 1988. She received the B.S. degree in geographic information systems and the M.S. degree in geological engineering from Chang'an University, Xi'an, China, in 2012 and 2014, respectively, where she is currently pursuing the Ph.D. degree in geoscience information systems.

Her research interests include image understanding, remote sensing image analysis, and geographic information systems.

Ms. Cao's awards and honors include the Hibiscus Student Fellowship from the School of Earth Science and Resources, Chang'an University.

JIAN-NONG CAO was born in Xi'an, China, in October 1963. He received the B.S. degree from Wuhan University, Wuhan, China, in 1987, the M.S. degree from Northwest University, Xi'an, in 2000, and the Ph.D. degree from Wuhan University, in 2005.

Since 2005, he has been with Chang'an University. He has participated in national level project and provincial level topics, including "Research on feature structured multiscale analysis method for information extraction from high-resolution remote sensing images" and "Hyperspectral image segmentation based on high-dimensional Markov network structure statistics." His research interests include remote sensing technology and applications (including image understanding and image pattern recognition), high-resolution remote sensing information extraction technology, LiDAR point cloud information extraction, photogrammetry and remote sensing technology (including photogrammetry and oblique photography), geographic information systems and applications, 3S basic theory, and technology application.

Prof. Cao is a member of the Jiu San Community.

QIAN ZHU was born in Suining, Sichuan, China, in 1983. She received the B.S. degree in resource environment and urban and rural planning management, in 2004, the M.S. degree in cartography and geographic information engineering, in 2007, and the Ph.D. degree in geoinformation engineering from Chang'an University, Xi'an, in 2014.

From 2015 to 2022, she was a Research Assistant with the College of Geological Engineering and Geomatics, Chang'an University.

LI-JIAO LOU was born in Xinzheng, in November 1988. She received the B.S. degree from the North China University of Water Resources and Electric Power, Zhengzhou, China, in 2011, and the M.S. degree from the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, in 2014.

Since 2015, she has been with Yellow River Hydrological Survey and Mapping Bureau. Her research interests include remote sensing technology and applications and software development.

NIAN-ZHONG XIAO was born in Xinyang, in January 1987. He received the B.S. degree from North China University of Water Resources and Electric Power, Zhengzhou, China, in 2011.

Since 2012, he has been with the Fourth Geological Exploration Institute of Henan Geology and Mineral Bureau. His research interests include hydrogeological engineering and environmental geology.