Received June 9, 2020, accepted June 25, 2020, date of publication July 1, 2020, date of current version July 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3006288

High-Resolution Remote Sensing Image Information Extraction and Target Recognition Based on Multiple Information Fusion

YI LIU[1,4,7], MIN CHANG[2,3,5,6], AND JIE XU[1,7]

[1] State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, China
[2] Shaanxi Provincial Land Engineering Construction Group Company Ltd., Xi'an 710075, China
[3] Key Laboratory of Degraded and Unused Land Consolidation Engineering, Ministry of Natural Resources, Xi'an 710021, China
[4] Cele National Station of Observation and Research for Desert Grassland Ecosystem in Xinjiang, Cele 848300, China
[5] Institute of Land Engineering and Technology, Shaanxi Provincial Land Engineering Construction Group Company Ltd., Xi'an 710075, China
[6] Shaanxi Provincial Land Consolidation Engineering Technology Research Center, Xi'an 710075, China
[7] University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding author: Jie Xu (xujie16@mails.ucas.edu.cn)

ABSTRACT The current research on multiple information fusion of remote sensing images is mainly aimed at remote sensing images of specific satellite sensors, and cannot be extended to other types of data source images. For high-resolution remote sensing images, when its surface coverage changes significantly, most of the mainstream algorithms are difficult to restore satisfactorily. The algorithm proposed in this paper combines the sparse representation and the spectral, spatial, and temporal features of remote sensing images for the first time to solve the above problems. The algorithm proposed in this paper first simulates the human visual mechanism, and obtains the spatial, spectral, and temporal features of the remote sensing image through the spatial spectral dictionary learning and the time-varying weight learning model. Secondly, local constraints are added to the extraction of temporal features to obtain temporal and geographical change information of heterogeneous remote sensing images. Then, a sparse representation model combining space-spectrum-time features is proposed to extract features of high-resolution remote sensing images. Finally, based on the VGG-16 network, this paper proposes a target recognition network with deep fully convolutional network, and uses the extracted feature map as the input of the target recognition network to realize the target recognition of the remote sensing image. Experimental results show that the method proposed in this paper can improve the accuracy of target recognition and improve the accuracy of recognition.

INDEX TERMS Multiple information, fusion, image, feature extraction, recognition.

I. INTRODUCTION

High-resolution remote sensing image target recognition is an important part of information extraction and processing of high-resolution ground observation system and automatic recognition system [1]–[3]. With the massive increase in the volume of high-resolution remote sensing data, the gradual diversification of data representation forms, and the complexity of remote sensing image scenes, artificially designed features have been unable to meet the precise classification and recognition tasks of high-resolution remote sensing images [4], [5]. How to make full use of the superiority of high spatial resolution and improve the recognition accuracy and target extraction reliability are of great significance.

Target detection in remote sensing images is of great significance in both military and civilian fields [6], [7]. However, due to the difference in the appearance of the target and the interference of the complex background and noise in the remote sensing image, in the remote sensing image with high spatial resolution, target detection is usually difficult. In order to detect targets in remote sensing images, many studies have been conducted [8]–[10]. All these work are focused on two issues, which are the characteristics to choose the target and how to efficiently select the region. Munoz-Mari *et al.* [11] used target contours, Zernike moments, and wavelet features, combined with support vector machines to detect aircraft targets from remotely sensed

The associate editor coordinating the review of this manuscript and approving it for publication was Zhihan Lv.

images. Xie *et al.* [12] used Gabor filtering and support vector machines to perform remote sensing image target detection. Li *et al.* [13] designed a fully automatic target detection system based on wavelet transform. Zhou *et al.* [14] used key points and spatially sparsely coded word bag models to detect targets. In addition, LBP [15] and HOG [16] are also commonly used features for target detection. However, the above methods have only achieved relatively good results in specific application scenarios. In some more complex scenarios, the effects achieved by these methods are limited. Because these characteristics are highly dependent on professional knowledge, they are often not enough to fully describe the goal, and the more essential characteristics of the goal cannot be obtained. In order to achieve a better recognition effect, Zhang *et al.* [17] identified planar roads from multispectral images fused with panchromatic bands, and used edges for auxiliary post-processing to remove non-road features connected to roads, and finally obtained a more accurate road network. Reza *et al.* [18] combined the extracted linear and planar target types, and extracted information such as road networks, agricultural plots, and residential areas from the fused high-resolution multi-source remote sensing image data. Shi *et al.* [19] proposed a method based on multi-feature learning, combining the linear and nonlinear features of hyperspectral remote sensing images to explore the linear and nonlinear boundaries between different features of remote sensing images. Shaaban *et al.* [20] proposed a multi-core learning method based on Bayesian theory, which can efficiently fuse hundreds or thousands of nuclear features. However, ignoring the spectral and spatial characteristics of the pixels at the recognition boundary also limits the further improvement of target recognition accuracy.

With the success of deep learning models in the field of computer vision, they have been gradually applied to the field of remote sensing. There are many related researches in remote sensing image scene recognition, target recognition and super-resolution reconstruction. Deep belief network is a good unsupervised feature learning model, and has achieved good results in speech recognition, image data set processing, etc. [21]–[23]. Saba *et al.* [24] proposed to use DBN to detect roads in high-resolution aerial remote sensing images, proving that the DBN model can extract image features. In order to solve the problem of complex data structure and limited number of training samples, Ghasemzadeh *et al.* [25] proposed a new feature extraction and recognition method for hyperspectral image interpretation. In the latest research, Zhang *et al.* [26] adopted a diversified DBN model, combined with two training processes of unsupervised pre-training and supervised fine-tuning to solve the problem of a small number of labeled samples, and then used diversification to deal with DBN hidden parameters. Compared with the original DBN model and other methods, the reflection and non-response situations have achieved higher recognition accuracy. Automatic encoders are also widely used in the field of remote sensing, and are mainly based on semi-supervised or unsupervised feature learning. SAE is widely used in hyperspectral images [27], [28]. It can reduce the dimension of hyperspectral remote sensing images, and can retain more original image information than dimensionality reduction methods such as principal component analysis, independent principal component analysis, and minimum noise separation transformation. Tao *et al.* [29] used the feature mapping function of the sparse stacked auto encoder to adaptively learn the feature representation from the labeled data. After that, the established sparse spectral features and multi-scale spatial features are identified using linear support vector machines. Experiments show that the learned spectral spatial feature representation is more discriminative and versatile. Yildirim *et al.* [30] built a deep network model based on an auto encoder, and used unsupervised greedy layer-by-layer training to train each layer to obtain a more robust feature expression, which effectively improved the accuracy of surface coverage recognition. Hamouda *et al.* [31] constructed a large-scale image processing recognition framework based on stacked auto encoders. The model parameters were adjusted and optimized according to the test sample, and the recognition accuracy was higher than that of random forest, support vector machine, and artificial neural network, which verified the advantages of SAE in land cover recognition. Liu *et al.* [32] established a stacked self-encoder identification method combining spectral and spatial information, demonstrating the great potential of deep learning in the accurate identification of hyperspectral data. Compared with DBN and SAE, convolutional neural network is a more efficient deep learning method, and has become a research hotspot in many scientific fields, especially in the field of pattern recognition and image processing. Bera *et al.* [33] applied deep convolutional neural networks to feature recognition of hyperspectral data. By reconstructing the spectral feature image and selecting a convolution filter of reasonable size, the spectral features of different land cover were extracted. When this method is applied to hyperspectral data in different situations, excellent recognition performance is obtained by adjusting parameters. Chang *et al.* [34] used convolutional neural networks to encode the spectral and spatial information of hyperspectral images. A multi-layer perceptron is used to perform the recognition task. The results on multiple experimental data sets show the potential of this method in hyperspectral image recognition. Huang *et al.* [35] studied how to transfer learning from the CNN features that have been successfully trained to scene recognition of high-resolution remote sensing images. CNN features are extracted through different layers of the network to generate image feature scenes. Experimental results on public scene data sets show that the features extracted by this method can obtain better performance. Peng *et al.* [36] used convolutional networks to identify high-resolution remote sensing images, which reduced the complexity of feature extraction and recognition, and improved recognition accuracy. To solve the problem of optical remote sensing image recognition, Zou *et al.* [37] use the optimized convolutional neural network to recognize the target on the 0.6m resolution remote sensing image.

Experiments show that the CNN model can achieve a higher accurate recognition rate of target features. In response to the problem of overfitting caused by the limited number of synthetic aperture radar training sets, Dong *et al.* [38] proposed a fully convolutional network that reduces the number of free parameters. The network only contains a sparse connection layer and does not use a fully connected layer. The test recognition accuracy on the benchmark data set can reach an average accuracy of 99%, which is significantly better than traditional target recognition methods.

The convolutional neural network directly takes the image as its input, without the need for complex pre-processing of the image. Compared with a standard backpropagation neural network of the same size, the number of connection parameters is smaller and training is easier. And the convolutional neural network has certain invariance to translation, distortion, and scaling. Therefore, based on the idea of multiple information fusion, this paper combines the sparse representation and the spectral, spatial, and temporal features of remote sensing images, and proposes an image extraction and recognition network based on multiple information fusion. First, it simulates the human visual mechanism, and obtains the spatial, spectral, and temporal features of remote sensing images through the spatial spectral dictionary learning and time-varying weight learning models. Secondly, local constraints are added to the extraction of temporal features to obtain temporal and geographical change information of heterogeneous remote sensing images. Then, a sparse representation model combining space-spectrum-time features is proposed to extract features of high-resolution remote sensing images. Finally, based on the VGG-16 network, this paper proposes a deep fully convolutional network to realize the target recognition of remote sensing images. Experimental results show that the network proposed in this paper has a good effect on efficiency and accuracy.

Specifically, the technical contributions of our paper can be concluded as follows:

This paper proposes an image extraction and recognition network based on multiple information fusion. The network can improve the image restoration effect when the surface coverage changes greatly. At the same time, the problem that the fully connected layer in the traditional convolutional neural network compresses the feature image into one dimension and loses the spatial information is solved.

The rest of our paper was organized as follows. Related work was introduced in Section II. Section III described the structure of the convolutional neural network algorithm proposed in this paper. Experimental results and analysis were discussed in detail in Section IV. Finally, Section V concluded the whole paper.

II. RELATED WORKS

A. OVERVIEW OF HIGH-RESOLUTION REMOTE SENSING IMAGE RECOGNITION

Remote sensing image recognition refers to a comprehensive analysis of the spectral and spatial characteristics of various features in the image, based on some means to select the features that can express the features, and finally divides the features into different feature categories through a certain recognition algorithm or according to. Figure 1 shows the basic framework of the high-resolution remote sensing image recognition method. First, preprocess the high-resolution remote sensing image. Then extract and select various features such as space and texture according to the characteristics of the feature to be recognized, and use it as the input of the recognizer to train the recognizer and complete the prediction of the image. In object-oriented high-resolution remote sensing image recognition, image segmentation is required after preprocessing the original image, and then feature extraction and recognition are performed. In practical applications, considering abnormal points and spatial smoothness in the recognition results, maximum/minimum analysis, clustering processing, and clustering processing are often used for post-recognition processing to further improve the recognition accuracy [39].

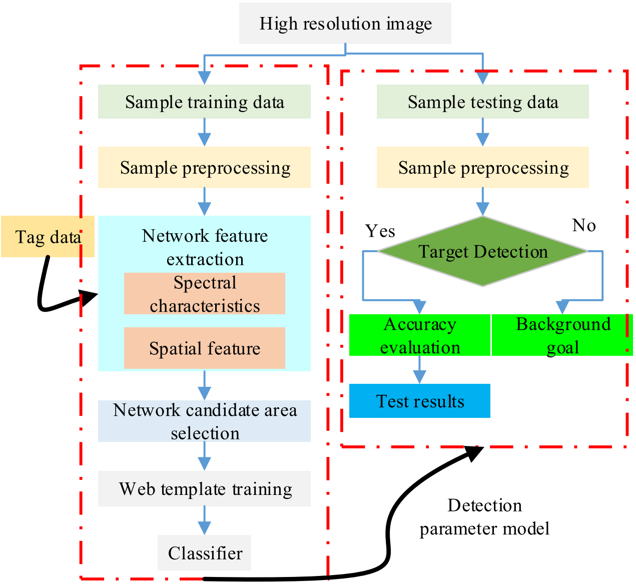


FIGURE 1. Flow chart of target recognition for high-resolution remote sensing images.

Remote sensing imaging is essentially a radiation transformation process from a three-dimensional space scene to a two-dimensional image plane. In the process of data collection, affected by the external weather conditions and the internal noise of the sensor, it causes a certain degree of geometric distortion and spectral distortion to the acquired high-resolution image. Therefore, pre-processing such as radiation correction, atmospheric correction, and geometric correction is required before image recognition.

Traditional remote sensing image recognition includes two key steps: recognition feature selection and recognition algorithm. Selecting appropriate feature variables is a key link to improve the accuracy of remote sensing image recognition. Commonly used identification features include spectral features, spatial features, temporal features, and polarization features. In the recognition process, a variety of features are usually selected to improve target recognition. The preprocessing, feature extraction and selection of images are summarized as feature expression, and the quality of feature expression is crucial to the performance of the recognizer.

Unlike traditional remote sensing image recognition, remote learning image recognition based on deep learning can integrate feature expression and recognition into one, and directly realize end-to-end learning and prediction. Deep learning takes the original image as input, and learns a highly nonlinear representation and a complex function representation from the original input through its deep neural network structure. Then, the recognizer connected through the network completes the recognition.

B. TWO REMOTE SENSING IMAGE RECOGNITION METHODS BASED ON DEEP LEARNING

As a kind of deep network, the convolutional neural network is a multi-layer network structure, and its feature extraction parameters are associated with the output. Unlike other deep networks, convolutional neural networks can directly and automatically extract spatial information in images [40].

1) IMAGE BLOCK RECOGNITION

In the early days of deep learning, the end of the convolutional neural network uses a fully connected layer, so that the test samples and training must maintain the same size.

Therefore, the image block recognition method is usually used to recognize the image. In order to identify a pixel, an image block around the pixel is used as the input of the network for training and prediction. Suppose the size of the image to be recognized is $M \times$N, and the size of the image block is $B\times$B. In order to recognize the pixels around the image, the image to be recognized needs to be filled. Normally, B is an odd number, and the number of width and height filled pixels of the image to be recognized is $(B-1)/2$. The process of image block recognition method is shown in Figure 2.

2) IMAGE SEMANTIC SEGMENTATION

Image semantic segmentation refers to grouping each pixel in the image according to the different semantic meanings expressed in the image. It can be seen from Figure 3 that the biggest difference with image block recognition is that the size of the image to be recognized and the image of the recognition result remain the same before and after the semantic segmentation of the image. In addition, the sliding step of the semantic segmentation of the image is the size of the image block, which can avoid the problem of repeated calculation of the pixels in the image block recognition.

C. EVALUATION INDEX OF REMOTE SENSING IMAGE TARGET DETECTION PERFORMANCE

In order to compare the performance of various target detection algorithms and avoid the subjective judgment algorithm, it is necessary to quantitatively evaluate the detection performance of each algorithm and give a comprehensive evaluation result.

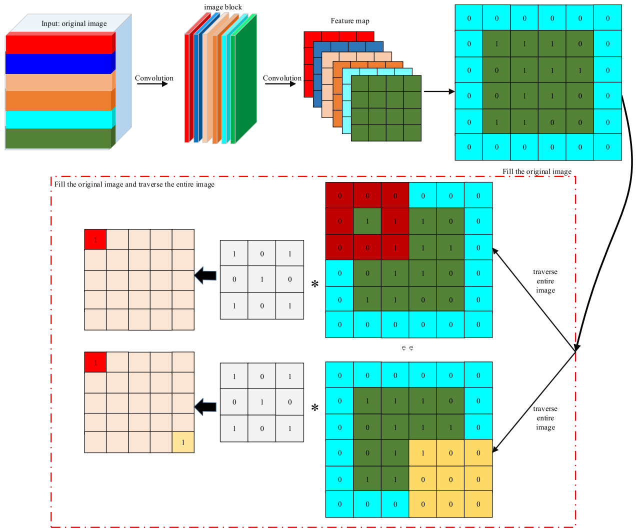


FIGURE 2. Classification and recognition of image blocks.

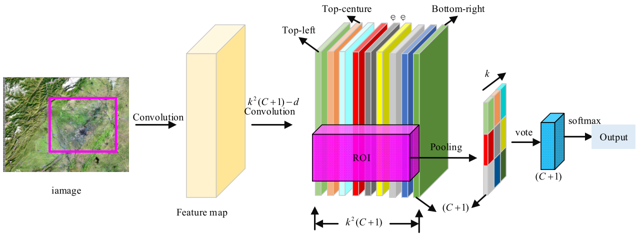


FIGURE 3. Image semantic segmentation.

1) RECALL RATE AND PRECISION RATE

Recall and precision are the most commonly used and basic evaluation indicators in the target detection process. Assuming that the prediction category is a positive sample and the prediction is correct, it is recorded as TP, and the prediction error is recorded as FP. The prediction category is negative sample and the prediction is correct as TN, and the prediction error as FN. Then the recall rate and precision rate are recorded as:

$$\mathrm{Pr}\,ecision = \frac{TP}{TP+FP} \tag{1}$$

$$\mathrm{Recall} = \frac{TP}{TP+FN} \tag{2}$$

2) KAPPA COEFFICIENT

When the uncertainty factor of the high-resolution remote sensing image recognition result is relatively large, the Kappa coefficient can avoid the excessive dependence of the overall accuracy on the number of target feature categories and the number of samples, and more realistically reflect the performance of the recognition algorithm.

The Kappa coefficient considers the influence of uncertainty on the recognition result, and the calculation formula is as follows:

$$Kappa = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}(x_{i+}x_{+i})}{N^2 - \sum_{i=1}^{r}(x_{i+}x_{+i})} \tag{3}$$

Among them, the variable r is the number of all categories. The variable x_{ii} is the number of pixels in row i and column i of the confusion matrix. The variable x_{i+} and the variable x_{+i} are the number of all cells in the row i and the column i, respectively. The variable N is the number of all cells used for evaluation.

III. IMAGE FEATURE EXTRACTION AND RECOGNITION NETWORK BASED ON MULTIPLE INFORMATION FUSION

Spectral, spatial, and temporal features are often used for remote sensing image analysis. With the development of remote sensing sensor technology, there is an urgent need to develop corresponding data fusion methods, combining different optical remote sensing images with spectral, spatial, and temporal features. Based on this, we first proposed a remote sensing image feature extraction network based on sparse representation. The network can simultaneously realize different tasks such as space-time data fusion, space-spectrum data fusion, spectrum-time data fusion, space-spectrum-time data fusion and so on. Secondly, in view of the problem that the fully connected layer in the traditional convolutional neural network compresses the feature image into one dimension and loses the spatial information, this paper proposes a modified deep fully convolutional network based on the VGG-16 network to realize the remote sensing image target recognition.

A. SPATIAL SPECTRUM FEATURE LEARNING AND TIME-VARYING FEATURE LEARNING

Suppose there are two kinds of feature data sets $X = \{x_1, x_2, \ldots, x_{N_1}\}$ and $Y = \{y_1, y_1, \ldots, y_{N_2}\}$ from different sensors x and y. Among them, the variable B_1 and the variable B_2 respectively represent the corresponding spectral channel dimensions. The variable N_1 and the variable N_2 represent the number of pixels.

The heterogeneous image feature data X and Y can be transformed into vectors $X \in R^{B_1 \times N_1}$ and $Y \in R^{B_2 \times N_2}$. We use $X(i,j) \in R^{p^2 \times B_1}$ to express a $p \times p \times B_1$ sized cube image at centered (i,j). Each element of the variable $X(i,j)$ can be expressed as $X(i,j)[n,l]$. The variables respectively represent the spatial position and spectral channel of the pixel. Usually variable $X(i,j) \in R^{p^2 \times B_1}$ can be transformed into the corresponding vector $x(i,j) \in R^{p^2 \times B_1}$.

$$x(i,j) = R(i,j)X \tag{4}$$

Among them, the variable $R(i,j) \in R^{p^2 \times B_1 \times N_1 \times B_1}$ is the image block extraction matrix.

Research on the human visual system shows that the receptive field of human eye cells sparsely selects a subset of structural primitives from over-complete coding set to encode natural images. Through the above findings, we can decompose high-dimensional remote sensing images $X(i,j)$ into a few components.

$$X(i,j)[n,l] = \sum_{m} a_m d_m[n,l] \tag{5}$$

Among them, the set $d_m[n,l] \in R^{p^2 \times B_1}$ is the basic set. The variable a_m is its correlation coefficient.

In the process of remote sensing image processing, the dictionary $d_m \in R^{p^2 \times B_1}$ of the original remote sensing image can be decomposed into corresponding products of spectral and spatial elements. In the experiment, the corresponding spectral and spatial training features are selected based on the image resolution. In this paper, the corresponding spectral primitives are trained with remote sensing images with high spectral resolution, and the corresponding spatial primitives are trained with remote sensing images with high spatial resolution.

$$d_m[n,l] = \phi_s[n]\theta_d[l] \tag{6}$$

Among them, variable $\{\phi_s\}_1^{p^2}$ and variable $\{\theta_d\}_1^{B_1}$ represent standard orthogonal basis, and describe the corresponding spatial and spectral characteristic distribution. Spectral and spatial primitives can be trained separately and combined into a joint function $d_m \in R^{p^2 \times B_1}$.

We express the corresponding spectral and spatial basis functions by using set $\Phi_s = \{\phi_1, \phi_2, \ldots, \phi_{p^2}\}$ and set $\Theta_d = \{\theta_1, \theta_2, \ldots, \theta_{B_1}\}$. In the sparse representation model, we use variables Φ_s and variables Θ_d to represent spatial and spectral dictionaries, respectively. The corresponding space spectrum joint basis vector can be expressed as $\phi_s \otimes \theta_d$. The process of solving spatial primitives is slightly different from that of spectral primitives. In the spectral domain, we can transform the corresponding hyperspectral remote sensing image X into $B_1 \times N_1$. Then learn the corresponding spectrum dictionary. In the spatial domain, we first use principal component analysis to map the spectral features of the corresponding image block to the spatial domain, and select the first principal component feature to train the spatial dictionary. In practical applications, we usually transform variables $d_m \in R^{p^2 \times B_1}$ into $D_m \in R^{p^2 \times B_1}$. Finally, we use the Kronecker product to obtain a joint space spectrum dictionary.

$$D = \Phi_s \otimes \Theta_d \tag{7}$$

Set $X_{t_1} \in R^{B_1 \times N_1}$ and set $X_{t_2} \in R^{B_1 \times N_1}$ represent two sets of data from the same sensor x at different times t1 and t2. We use set $X_{t_1}(\Omega_{ij})$ to represent the set of image blocks $X_{t_2}(i,j)$ that adjoin each other at time t1, where variable Ω_{ij} represents the set of image blocks $X_{t_2}(i,j)$ that adjoin each other at the center point (i, j). The variable Ω_{ijk} represents the k-th image block in the variable Ω_{ij}.

$$X_{t_2}(i,j) = \sum_{\Omega_{ijk}\in\Omega_{ij}} w_{ijk} X_{t_1}(\Omega_{ijk}) \tag{8}$$

In practical applications, usually each remote sensing image block has a great similarity with its neighboring image blocks. Inspired by this, we added local constraints on the basis of the sparse constraints of formula (9). The introduction of local constraints emphasizes that local constraints are more important than sparse constraints, which is consistent with the conclusion of locality constraint linear coding (LLC) [41].

Combining local constraints, we can express the time-varying characteristics of remote sensing images as:

$$\begin{aligned} &\min ||X_{t_2}(i,j) - X_{t_1}(\Omega_{ijk})W_{ij}||^2 + \lambda||Dis\tan ce_{ij}W_{ij}||^2 \\ &s.t.W_{ij} = 1 \end{aligned} \tag{9}$$

Among them, the variable W_{ij} is the weight of the temporal change corresponding to the image block $X_{t_2}(i,j)$. The variable $Dis\tan ce_{ij}$ is the Euclidean distance between the two sets of remote sensing image blocks.

B. REMOTE SENSING IMAGE FEATURE EXTRACTION NETWORK COMBINING SPACE-TIME SPECTRUM FEATURES

Usually, the image Z_{t_1} at time t1 is used to represent the original hyperspectral-high spatial resolution remote sensing image X_{t_1} corresponding to the hyperspectral resolution remote sensing image Y_{t_1} and the high spatial resolution remote sensing image. In this article, we use high spectral resolution image X_{t_1} and high spatial resolution image Y_{t_1} as examples. The remote sensing image of each sensor source can be expressed as a joint dictionary of space spectrum as:

$$X_{t_1} = Z_{t_1}H + n_x = DA_{t_1}H + n_x \tag{10}$$

$$Y_{t1} = GZ_{t_1} + n_y = \tilde{D}A_{t_1} + n_y \tag{11}$$

Among them, the equation $\tilde{D} = GD$ represents a dictionary of transformed low-spectral resolution. The variable G is the corresponding change matrix. The variable H is the corresponding spatial deburring and down sampling operator of the remote sensing image. Matrix n_x and matrix n_y represent corresponding optimization errors. We add sparse constraints on the basis of formula (10) and formula (11) to solve the sparse coefficient matrix A_{t_2} at time t2.

$$A_{t_2} = \arg\min ||Y_{t_2} - \tilde{D}A_{t_2}||_F^2 + ||X_{t_2} - DA_{t_2}H||_F^2 + \lambda||A_{t_2}|| \tag{12}$$

Among them, the symbol $||\cdot||_F^2$ represents Fresenius norm. We propose a new sparse representation model combining space-time spectral features to maintain the spatial consistency between adjacent image blocks and the time-varying features between image blocks at different times.

$$\begin{aligned} A_{t_2} = &\arg\min ||Y_{t_2} - \tilde{D}A_{t_2}||_F^2 + ||X_{t_2} - DA_{t_2}H||_F^2 \\ &+\lambda_1 \sum_{ij} ||DA_{t_2}(i,j) - \tilde{Z}_{t_2}W_{ij}(\Omega_{ij})||_2^2 + \lambda_2||A_{t_2}|| \\ &s.t.\ a_{t_2}(i,j) \geq 0 \end{aligned} \tag{13}$$

Among them, equation $\tilde{Z}_{t_2}(\Omega_{ij}) = D\tilde{A}_{t_2}(i,j)$ represents the remote sensing image block with high spectral resolution and high spatial resolution at (i,j) constructed in each iteration. Using remote sensing image blocks adjacent remote sensing image block sets $\tilde{Z}_{t_2}(\Omega_{ij})$, formula (13) can be expressed as

$$\begin{aligned} A_{t_2} = &\arg\min ||Y_{t_2} - \tilde{D}A_{t_2}||_F^2 + ||X_{t_2} - DA_{t_2}H||_F^2 \\ &+\lambda_1 \sum_{ij} ||DA_{t_2}(i,j) - U_{t_2}||_2^2 + \lambda_2||A_{t_2}|| \\ &s.t.\ a_{t_2}(i,j) \geq 0 \end{aligned} \tag{14}$$

$$U_{t_2} = \{\tilde{Z}_{t_2}(\Omega_{i_1,j_1})W_{i_1,j_1}, \tilde{Z}_{t_2}(\Omega_{i_2,j_2})W_{i_2,j_2}, \ldots, \tilde{Z}_{t_2}(\Omega_{i_n,j_n})W_{i_k,j_k}\} \tag{15}$$

We use Figure 4 to represent the sparse representation of remote sensing image feature extraction network structure combined with space-time spectral features. First, we use spectral and spatial feature extraction models to obtain feature primitives with high spectral resolution and high spatial resolution. Then, the time series feature change weight learning model is used to combine the spatial spectrum basis function and the time change feature. Finally, a sparse representation remote sensing image feature extraction network combined with space-time spectral features is proposed to explore the relationship between spatial spectral features and time-varying features and find the most useful features to serve the target recognition of remote sensing images.

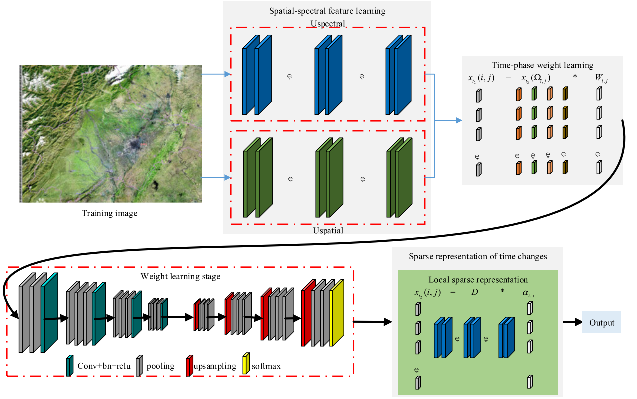

Spatial-spectral feature learning
Uspectral
Training image
Uspatial
Time-phase weight learning
$x_{t_2}(i,j)$ − $x_{t_1}(\Omega_{i,j})$ * $W_{i,j}$
Weight learning stage
Conv+bn+relu
pooling
upsampling
softmax
Sparse representation of time changes
Local sparse representation
$x_{t_2}(i,j)$ = D * $\alpha_{i,j}$
Output

FIGURE 4. Sparse representation fusion model combining multiple feature information.

C. HIGH-RESOLUTION REMOTE SENSING IMAGE TARGET RECOGNITION NETWORK

The traditional high-resolution remote sensing image recognition method has the problems that features cannot be automatically extracted and it is difficult to process big data.

The remote sensing image recognition method based on deep learning can automatically learn image features and achieve good recognition results. To this end, in this paper, first of all, through the remote sensing image feature extraction network structure, the sample feature set is obtained. Then according to the characteristics of the sample data, the network is improved on the basis of VGG-16 to better adapt to the recognition task of high-resolution remote sensing images. Finally, this paper further improves the recognition prediction method of the convolutional neural network, using image block recognition with a sliding step size greater than 1 and bilinear up sampling to make the recognition speed faster and the recognition accuracy higher.

In order to learn as much information as possible from a large number of remote sensing images, this paper builds a model with powerful learning ability based on VGG-16. In order to improve the efficiency of traditional image block recognition, this paper sets the sliding step size in image block recognition to greater than 1 to obtain the down-sampled probability recognition image. Finally, bilinear up sampling is used to obtain a recognition result map consistent with the original image resolution. The network structure is shown in Figure 5.

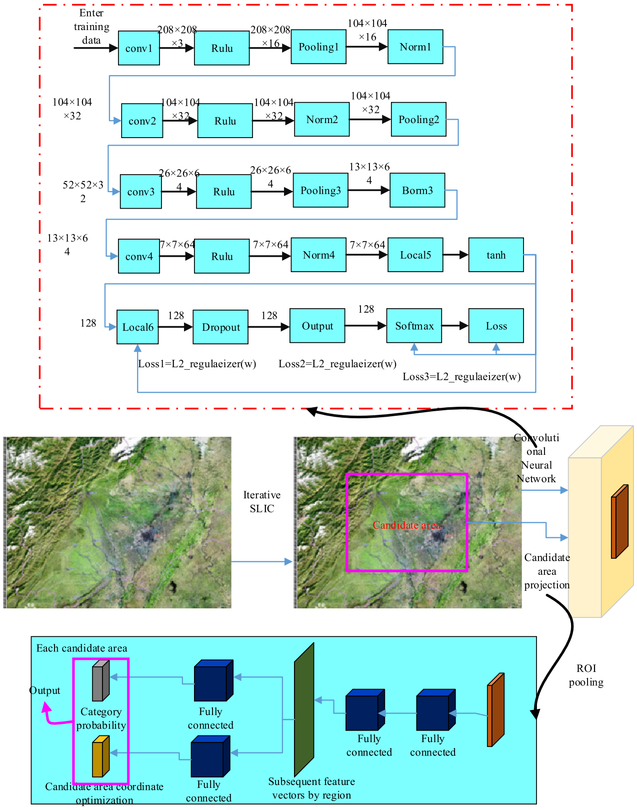


FIGURE 5. The proposed feature extraction and target recognition network structure.

First, first extract the network structure through the feature extraction of the remote sensing image to obtain the feature map of the sample. At the same time, iterative SLIC technology is used to generate a set of candidate regions. Then, each candidate region uses a ROI pooling layer to get a fixed-size feature vector from the feature map, and each feature vector will go through several consecutive fully connected layers. Finally, two outputs are obtained, which are the category probability of the candidate region and the optimization of the coordinates of the candidate region. From these two outputs, the final recognition result for the input image can be obtained.

1) CANDIDATE AREA GENERATION

Sliding windows are widely used for target detection in traditional methods. The sliding window needs to traverse the entire image, which is a time-consuming mechanism. In addition, in order to cover multi-scale targets, multiple sliding windows of different sizes are required, which increases the number of candidate regions by several times.

SLIC [42], that is, simple linear iterative clustering, is an image segmentation algorithm with simple ideas and convenient implementation. It converts the image from RGB color space to CIE-lab color space and 5-dimensional feature vector in XY coordinates. Then the distance metric is constructed on the 5-dimensional feature vector, and the process of local clustering of image pixels is performed. The SLIC algorithm can generate relatively compact and uniform super pixels, and has advantages in terms of calculation speed, object contour retention, and super pixel shape.

2) ROI POOLING

Feature extraction using convolutional networks is a computationally intensive and time-consuming operation. In this paper, after using SLIC to generate regions, we obtained about 2000 candidate regions. For each generated region, CNN was used to extract features for subsequent identification and recognition, which made the calculation amount very huge, and there were many repeated calculations. Therefore, this paper uses ROI pooling, which only extracts the convolution feature once for the image to be recognized, avoiding a lot of repeated calculations and greatly improving the efficiency of target recognition.

3) BILINEAR UP SAMPLING

In order to improve the efficiency of image recognition, the step size of the sliding window is set to be greater than 1 in the experiment, resulting in the result of the down sampling of the original image. In order to restore the recognition result to the original image resolution, up-sampling is needed, that is, by using the existing image data points to bring into the resampling function and sum. Figure 6 is realized by the recognition method combined with up sampling. When the sampling sliding step is 2, the probability map of down sampling is obtained by image block recognition. The dimension of the probability map is the number of categories. Then, the probability map of each category is up-sampled and the category corresponding to the maximum probability value of

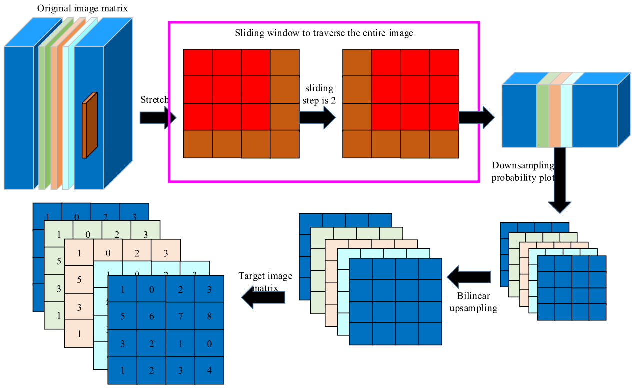


FIGURE 6. Classification and recognition method combined with up sampling.

each pixel position is taken to obtain the final recognition result.

Commonly used interpolation methods include three types: nearest neighbor interpolation, bilinear interpolation, and cubic convolution. The nearest neighbor pixel method has a small amount of calculation, but it is easy to cause discontinuous gray values on the image after interpolation, which may appear jagged. The bilinear interpolation method is computationally complex and can achieve satisfactory results. The cubic convolution method has the best interpolation effect, but the calculation amount is also the largest. Among them, the bilinear interpolation method does not need to be learned, the interpolation effect is good, and the calculation speed is fast, which has become a more commonly used method. Therefore, this paper selects bilinear resampling kernel.

4) CNN MODEL STRUCTURE AND TRAINING

Considering the size and complexity of the feature map constructed in the feature extraction network, this paper removes the last connected pooling layer and three layer convolutional layers in the VGG-16 network structure to prevent the sample data from being down sampled to negative values and excessive deep networks learn overexpression of simpler features.

The last few layers of VGG-16 are fully connected layers, which will squash the original two-dimensional matrix into one dimension, resulting in the loss of spatial information. The spatial feature information plays an important role in the task of remote sensing image recognition. In addition, VGG-16 uses more parameters and uses more memory. To this end, this paper connects two convolutional layers at the end of the network, while reducing the number of neurons. Finally, to reduce overfitting of convolutional neural networks, a dropout layer is added after each pooling layer and convolutional layer. The final network structure is shown in Figure 7.

5) TARGET RECOGNITION AND RESULT OUTPUT

After the target candidate region is generated, the image to be recognized is sent to the trained network model for feature extraction and recognition. After that, standard non-maximum suppression is used to fuse multiple overlapping detections of the same target, and the final result is obtained.

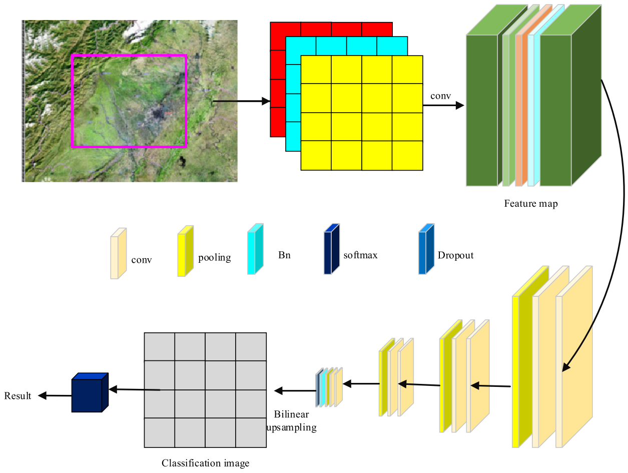


FIGURE 7. Target recognition network structure based on improved VGG-16.

D. LOSS FUNCTION AND OPTIMIZATION METHOD

The loss function is used to evaluate the error between the predicted value and the label value of the network, and then evaluate whether the network is suitable. This paper builds an overall cost error function with the sample's true labeling results. The formula is as follows:

$$J(w,b)=[\frac{1}{m}\sum_{i=1}^{m}||h_{w,b}(x^i)-y^i||^2]+\frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(w_{ji}^l)^2 \tag{16}$$

Among them, variable $J(w, b)$ is the mean square error. The second item is regularization. Regularization can reduce the proportion of weights and avoid overfitting. The function $h_{w,b}(x^i)$ is a nonlinear hypothesis model. In the recognition problem, $y = 0$ or 1 is used to represent two types of tags. In order to solve the minimum value of the non-convex function $J(w, b)$, this paper uses the truncated positive distribution to generate a relatively small random number, and assigns a random initialization value of parameter w_{ji}^l and parameter b_i^l close to 0, which is used to avoid too much weight in the network largely led to saturation. Then use the Adam gradient descent optimization algorithm for the objective function.

Adam is a method of adapting different parameters to different learning rates. It can dynamically adjust the learning rate of each parameter in the network through the first and second moment estimation of the gradient. After offset correction, the learning rate has a certain range and relatively stable parameters. The attenuation method of this method is similar to momentum, and the calculation formula is as follows:

$$m_\tau = u^* m_{\tau-1} + (1-u)g_\tau(\theta) \tag{17}$$

$$n_\tau = v^* n_{\tau-1} + (1-v) g_\tau^2(\theta) \tag{18}$$

In the above formula, the variable $g_\tau(\theta)$ is the gradient. The variable m_τ and the variable n_τ are the first moment estimation and the second moment estimation of the gradient, respectively. The variable m_τ and the variable n_τ tend to the vector of 0, especially when the decay rates u and v are close to 1. In order to improve this problem, the deviation of the variable m_τ and the variable n_τ is corrected:

$$\tilde{m}_\tau = \frac{m_\tau}{1-u^\tau} \tag{19}$$

$$\tilde{n}_\tau = \frac{n_\tau}{1-v^\tau} \tag{20}$$

$$\Delta J(\theta_\tau) = -\frac{\tilde{m}_\tau}{\sqrt{\tilde{n}_\tau} + \varepsilon}\xi \tag{21}$$

Among them, variable μ, variable v, and variable ε are constant terms. Their size settings are as follows: $\mu = 0.9$, $v = 0.999$, $\varepsilon \ = 10e^{-8}$. The variable ε is used to ensure that the denominator is not zero. As can be seen from the above formula, Adam dynamically adjusts the learning rate and forms a certain range of constraints while not increasing the additional memory usage.

The key to update the two parameters of weight and offset is to calculate the gradient of the two, and use the backward propagation algorithm to calculate the gradient. First, use the forward propagation calculation formula to get the activation value of the input layer $L2$, $L3$ until the output layer L_{n_l}, for each output unit of the n_τ, the calculation formula of the residual is as follows:

$$\delta_i^l = \frac{\partial J(w,b;x,y)}{\partial z_i^{n_l}} = \frac{\partial (y_i - a_i^{n_i}) f'(z_i^{n_l})}{\partial z_i^{n_l}} \tag{22}$$

Among them, the variable n_l represents the number of layers of the network. The variable w_{ij}^l is the weight of the connection between the unit of layer l and the unit of layer l +1 of layer i. The variable a_i^l represents the activation value (output value) of the unit of the l layer. The variable z_i^l represents the input of the unit of layer l and includes the weighted sum of the offset units. The variable z_i^l calculation formula is as follows:

$$z_i^l = \sum_{j=1}^{n} w_{ij}^{l-1} x_j + b_i^{l-1} \tag{23}$$

$$a_i^l = f(z_i^l) \tag{24}$$

Among them, symbol $f(\cdot)$ means activation function. The activation function used in this paper is the unsaturated linear unit relu function to accelerate the network convergence speed. The deep convolutional neural network uses relu to calculate the residual items of the l layer and the i node for each layer of $l = n_l - 1, n_l - 2, \ldots, 2$ as:

$$\delta_i^l = (\sum_{j=1}^{s_{l+1}} w_{ij}^l \delta_j^{l+1}) f'(z_i^l) \tag{25}$$

Finally, calculate the required partial derivative according to the following method:

$$\frac{\partial J(w,b;x,y)}{\partial w_{ij}^l} = a_i^l \delta_i^{l+1} \tag{26}$$

$$\frac{\partial J(w,b;x,y)}{\partial b_i^l} = \delta_i^{l+1} \tag{27}$$

IV. EXPERIMENTS AND RESULTS

A. IMAGE DATA SET

In order to evaluate the recognition performance of the proposed network, we used two public datasets as experimental datasets. The UCMerced-Landuse dataset contains 21 high-resolution remote sensing image target categories. The HR dataset contains 19 remote sensing categories. The dataset was collected from Google Earth, and each category contains 50 high-resolution color images. For the accuracy of the experimental data, we split the training set and the test set according to the experimental rules of the data set. The ratio of training set to test set in the UCMerced-Landuse dataset is 80/20. The ratio of training set to test set in the RS dataset is 50/50.

B. PARAMETER SETTINGS

Parameter adjustment is a major difficulty in training convolutional neural networks. In order to reduce the training difficulty, the experiment only adjusts the most difficult and most important hyper parameter learning rate, and other parameters use the default values in most networks. The specific settings are: random inactivation coefficient is 0.8, regularization coefficient is 0.0001. The convolution kernels are all 3 × 3 in size, the batch size is 64, and the number of trainings is 10000. In order to set an appropriate learning rate, the strategy adopted in this paper is to set the initial learning rate to a very small value, and gradually increase the learning rate by observing the decreasing speed of the loss value curve, and the final learning rate is set to 0.0001.

Figure 8 shows the test accuracy, loss, and the weight and bias parameter changes in the network when training using the data set. It can be seen that in the first 1000 batches, the training accuracy, and loss of the training model rise and fall respectively faster, the accuracy reaches about 0.991, and the loss decreases to about 0.1%. Afterwards, the training accuracy and the loss both changed relatively slowly, and there was a small oscillation. The trained model is used for testing in the test set, and the test accuracy rate reaches 0.985.

C. THE EFFECT OF FUSION OF SPECTRAL FEATURES, SPATIAL FEATURES AND TEMPORAL FEATURES OF REMOTE SENSING IMAGES

In order to verify the effect of the proposed algorithm on the fusion of spectral features, spatial features and temporal features of remote sensing images, the experiments in this section are mainly divided into the following three parts.

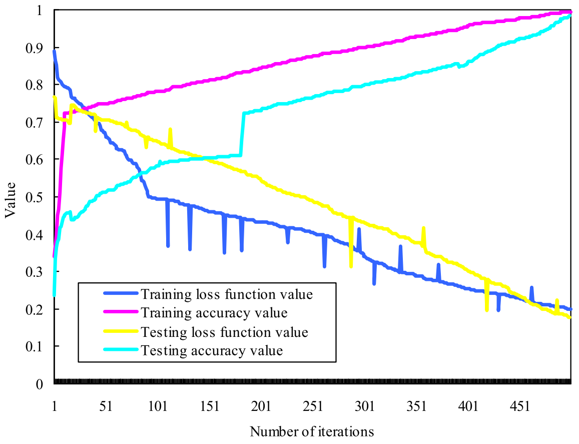


FIGURE 8. Accuracy and loss function values of training and test sets.

1) SPACE-SPECTRUM FEATURE FUSION

During the experiment, we first used a Gaussian blur function with an 8 × 8 standard deviation of 3 to blur the original image. Then, down sample every 4 pixels in the horizontal and vertical directions.

In the experiment, we adopted PSNR, CC, SSIM, ERGAS, SAM and Q-avg index, which measure the algorithm proposed in this paper can achieve mainstream remote sensing image fusion effect. Figures 9 and 10 compare qualitatively and quantitatively the indicators and compare the proposed method with five mainstream spatial spectral fusion methods: GSA [43], SFIM [44], GLP [45], ECCV-14 [46], and MAPSMM [47]. Experimental results show that all comparison algorithms can maintain the spectral and spatial characteristics of the remote sensing image.

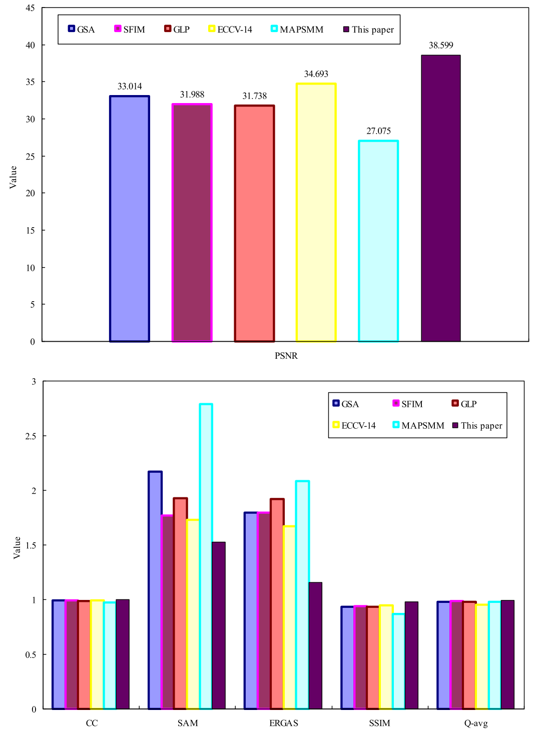


FIGURE 9. Space-spectrum quantitative evaluation experiment.

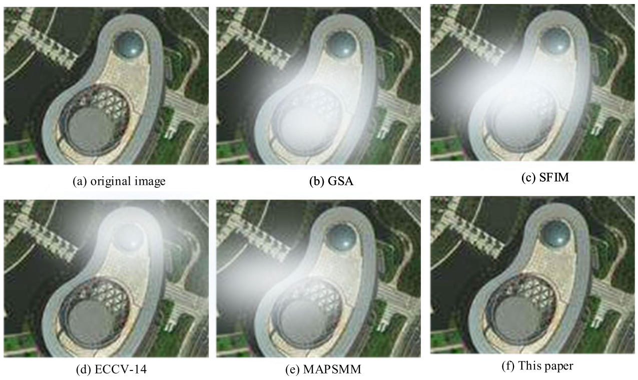

(a) original image
(b) GSA
(c) SFIM
(d) ECCV-14
(e) MAPSMM
(f) This paper

FIGURE 10. Space-spectral fusion image restoration results.

All six comparison algorithms can obtain part of the spatial and spectral information from the remote sensing image, and then use the obtained spatial spectrum features to reconstruct the hyperspectral-high spatial resolution remote sensing image. Compared with the other five algorithms, the algorithm proposed in this paper has the smallest error and the highest similarity between the reconstructed image and the original image. The reconstruction effect of SFIM and ECCV-14 algorithm is second only to the algorithm proposed in this paper, but it surpasses the other three algorithms. It can be seen from Table 1 that the algorithm proposed in this paper also has the least use time.

TABLE 1. Calculation time of space-spectrum experiments.

Method	Running time /s
GSA	25.267
SFIM	24.669
GLP	31.356
ECCV-14	109.006
MAPSMM	71.271
This paper	19.031

2) SPACE-TIME FEATURE FUSION

The main purpose of this set of experiments is to demonstrate the fusion effect of the algorithm proposed in this paper on the data of space-time feature changes. Figure 11 shows the scatterplot between the real value and the predicted value. It can be clearly seen from Figure 11 that the algorithm proposed in this paper can obtain approximately indistinguishable experimental results compared to the remaining three algorithms.

3) SPACE-SPECTRUM-TIME FEATURE FUSION EXPERIMENT

Figures 12 and 13 show the accuracy of the remote sensing image with high spatial-high spectral-high temporal resolution predicted by the algorithm proposed in this paper.

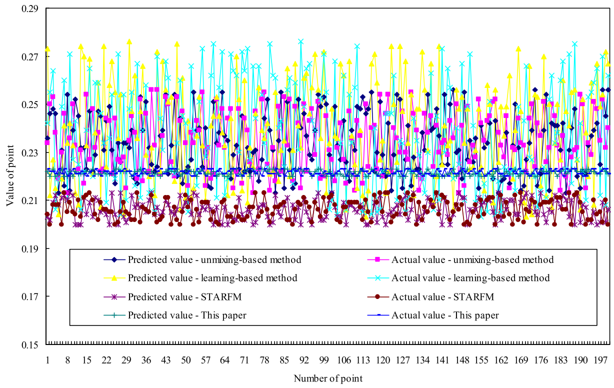


FIGURE 11. Scattering between the predicted and true values of space-time fusion.

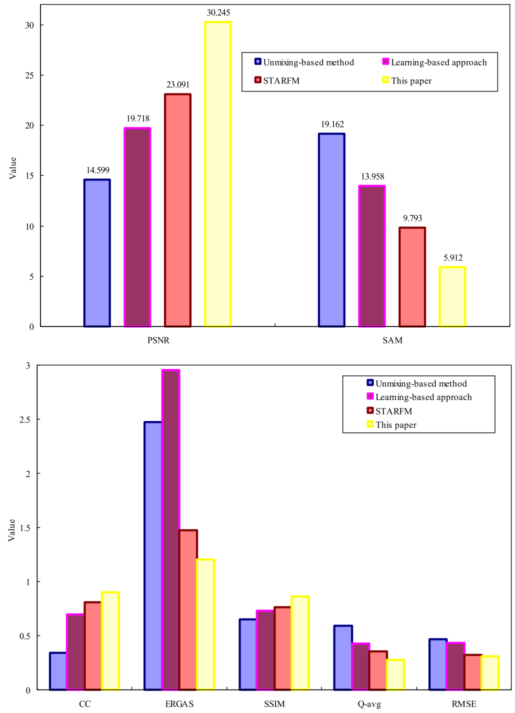
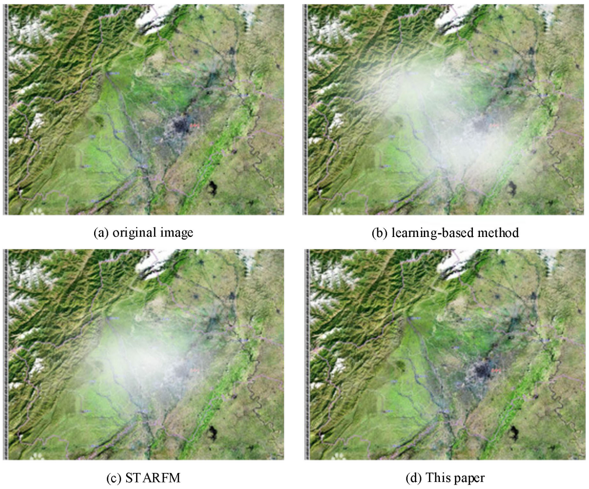


FIGURE 12. Space-time quantitative evaluation experiment.

It can be seen from the experimental results that the restored images obtained by other algorithms are fuzzy, and the remote sensing images restored by the algorithm proposed in this paper are the closest to the real images. This also indirectly shows that the algorithm proposed in this paper can deal well with the time variation of surface coverage.

Through the qualitative description indicators, the algorithm proposed in this paper can maximize the spatial structure information and retain the spectral information. The scatter plot between the true and predicted values in the near infrared band shows that the satellite image restored by the algorithm proposed in this paper is closest to the original image as shown in Figure 14.


(a) original image

(b) learning-based method

(c) STARFM

(d) This paper

FIGURE 13. Reconstructed image of a remotely sensed image with high spatial-high spectral-high temporal resolution.

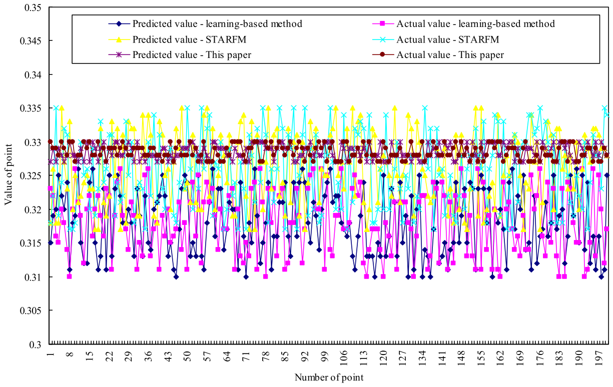


FIGURE 14. Scattering between the predicted and true values of the space-spectrum-time fusion experiment.

Compared with the previous space-time feature fusion results, the experimental results using the space-time spectral features can retain more detailed information on the types of features covered. The quantitative description of the experimental results in Table 2 shows that the algorithm proposed in this paper can fuse the spectral, spatial, and temporal features of the remote sensing image well, and obtain the fusion results of the remote sensing data without error. Although in practical

TABLE 2. Space-spectrum-time quantitative evaluation experiment.

Method	Learning-based approach	STARFM	This paper
PSNR	26.071	22.077	38.985
CC	0.443	0.681	0.987
SAM	0.901	0.811	0.441
ERGAS	4.678	2.793	0.547
SSIM	0.301	0.801	0.959
Q-avg	0.165	0.439	0.867

applications, the pixel prediction value is not only affected by the space-time spectral characteristics of the remote sensing image, but also by the redundant features of the image and the quality of the image imaging, but the algorithm proposed in this paper shows its application potential in multi-source data fusion of remote sensing images.

D. ALGORITHM RECOGNITION PERFORMANCE

Use the test data set to evaluate and analyze the results of deep learning building algorithm detection. For the same algorithm, setting different detection thresholds will result in different detection recall rates and accuracy rates. Therefore, set a detection confidence score threshold. When the score of a certain detection frame greater or equal to score threshold, this detection frame will be retained as the final target recognition result in the image. When score is less than score threshold, the corresponding detection frame will be filtered out.

Therefore, for the algorithm proposed in this paper, we set different confidence score thresholds and count the detection results of false detections, so as to obtain different detection recall rates and accuracy rates, and corresponding Fp values. And use this to get the optimal detection threshold of the algorithm, the result is shown in Figure 15. Here, under the same detection conditions and environment, different categories of confidence thresholds have obtained corresponding Fp values. As can be seen from Figure 15, as the score threshold continues to decrease from 0.9 to 0.1, the algorithm's detection recall rate is getting higher and higher, that is, the total number of detected targets is increasing. The recall rate continued to rise from 0.897 to 0.939. However, due to the stability of the algorithm, the accuracy rate has been maintained at a fairly high level, and there has been no significant decline with the increase of the recall rate, and it has continued to maintain a height of about 0.985. Therefore, due to the increase in the recall rate, the value of Fp also increased from 0.942 to 0.961. Therefore, in combination with the specific situation and the statistics of the experimental results, the experiment sets score threshold is 0.1, so as to obtain reliable and excellent detection results as much as possible.

In order to illustrate the accuracy and robustness of the target recognition algorithm used in this paper, we compare the algorithm of this paper with DBN, SAE, reference [36], reference [37], and reference [38]. Respectively compare the recall rate, accuracy rate, Fp, overall recognition accuracy and Kappa coefficient between the detection results of each algorithm, as shown in Figure 16.

As shown in Figure 16, the recognition result of the network structure proposed in this paper is the highest among all comparison algorithms, which fully illustrates the superiority of this method for target recognition of high-resolution remote sensing images. At the same time, the method of this paper is also ahead of other comparison methods in terms of recall rate and accuracy, reaching 0.939 and 0.985, respectively. Although the algorithms are used based on reference [37] and reference [38], the accuracy is almost close to the accuracy of the method in this paper. But the recall rate and Kappa coefficient are completely behind the method of this article. In addition, from the data in Figure 16, we can also find that the network structure proposed in this paper is better than the DBN network and SAE network in terms of recall rate. This is because the input of the DBN is one-dimensional information, ignoring the two-dimensional structure information of the image, and the image features of high-dimensional

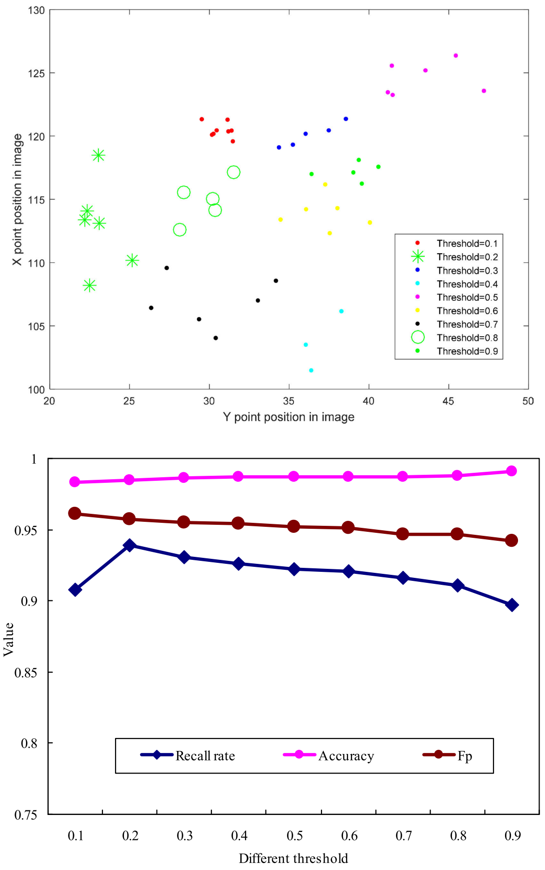


FIGURE 15. Detection results with different thresholds.

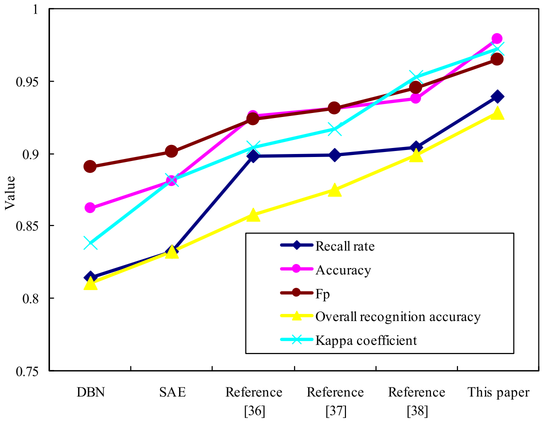


FIGURE 16. Performance comparison of different algorithms.

data cannot be directly extracted. However, SAE only uses the spectral characteristics of the image and ignores the spatial information. The algorithm in this paper extracts multiple features of the image and merges them to better utilize the correlation between the features. And the target recognition part of the network in this paper adopts a multi-level structure. This structure can reduce the parameters of multiple features extracted from the previous features, thereby achieving high-efficiency recognition of the target in the image.

E. OPERATING EFFICIENCY

The algorithms of this paper, DBN, SAE, reference [36], reference [37] and reference [38] use the same learning rate, and batch size and optimization function to perform time efficiency analysis on the data set.

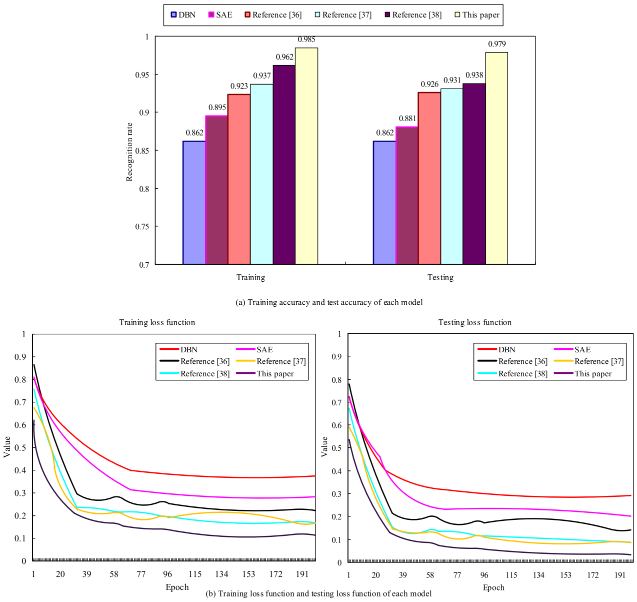
Table 3 shows the time required for one iteration of the six algorithms, that is, the time it takes to traverse a training set when training the network. The time-consuming relationship of the six models is: SAE > DBN > Reference [38] > Reference [36] > Reference [37] > the algorithm in this paper. By analyzing the structure of the four models, it is not difficult to know that the convolution part of VGG-16 is used in the model structure of the algorithm, reference [36], reference [37] and reference [38], so these four models the training speed is faster. There are many parameters in the network structure of DBN and SAE, which results in a long training time of the network.

TABLE 3. Comparison of training time performance of different algorithms.

Method	Training time / epoch
DBN	2178s
SAE	2256s
Reference [36]	753s
Reference [37]	891s
Reference [38]	941s
This paper	512s

To further analyze the influence of the model structure on the training accuracy, the experiment trained 200 epochs on the six network models, and recorded the training accuracy, training loss and test accuracy, and test loss change curves of each model. The results are shown in Figure 17. After training for 200 epochs, the training accuracy of the algorithms in this paper, DBN, SAE, reference [36], reference [37], and reference [38] are 0.985, 0.862, 0.895, 0.923, 0.937, and 0.962, respectively. And the test accuracy of the corresponding algorithm is 0.979, 0.862, 0.881, 0.926, 0.931, and 0.938 respectively.

In the first 50 epochs of model training, the algorithms in this paper and references [36], [37], and [38] have a faster loss of accuracy. The loss of DBN and SAE is slower. Throughout the training process, the decline trend of the loss functions of the six algorithms was roughly the same. However, the loss functions of the other five algorithms have experienced more fluctuations, especially the jitter of the DBN network far exceeds the other four models. But as the model training increases, this beating tends to be stable.



(a) Training accuracy and test accuracy of each model

(b) Training loss function and testing loss function of each model

FIGURE 17. Variation curves of training accuracy, training loss and test accuracy, test loss of each model.

V. CONCLUSION

Traditional image recognition algorithms are not ideal in remote sensing image recognition, mainly because of the limitation of algorithm structure, insufficient representation ability for complex data, and lack of generalization ability. Taking deep feature learning as the main line and high-resolution remote sensing image classification as the research object, this paper gives full play to the advantages of deep convolutional neural network in information extraction and feature expression, and proposes an image extraction and recognition network based on multiple information fusion. Firstly, the spatial, spectral and temporal characteristics of remote sensing images are obtained by spatial spectral dictionary learning and temporal weight learning model. Secondly, local constraints are added in the time feature extraction process to obtain the temporal and geographic variation information of heterogeneous remote sensing images. Then, a sparse representation model combining space-spectrum-time features is proposed to extract features of high-resolution remote sensing images. Finally, a deep convolutional network based on VGG-16 network is proposed to realize target recognition in remote sensing image. Experimental results show that the proposed network has good efficiency and accuracy. The remote sensing image recognition algorithm proposed in this paper has good fault tolerance and self-learning ability, which makes it suitable for complex remote sensing image processing and has great research value in practical application.

REFERENCES

[1] B. Liu, T. Chen, P. Fu, Y. Zhen, and J.-S. Pan, ''Method of target recognition in high-resolution remote sensing image based on visual saliency mechanism and ROI region extraction,'' *J. Internet Technol.*, vol. 20, no. 5, pp. 1333–1341, 2019.

[2] C. Li, H. Gao, Y. Yang, X. Qu, and W. Yuan, ''Segmentation method of high-resolution remote sensing image for fast target recognition,'' *Int. J. Robot. Autom.*, vol. 34, no. 3, pp. 4597–4618, 2019.

[3] Z. Shao, L. Wang, Z. Wang, and J. Deng, ''Remote sensing image super-resolution using sparse representation and coupled sparse autoencoder,'' *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2663–2674, Aug. 2019.

[4] Q. Wang, X. Zhang, G. Chen, F. Dai, Y. Gong, and K. Zhu, ''Change detection based on faster R-CNN for high-resolution remote sensing images,'' *Remote Sens. Lett.*, vol. 9, no. 10, pp. 923–932, Oct. 2018.

[5] W. Jiang, H. Xiao, Z. Zhao, and J. Zhou, ''A boundary parallel-like index for high-resolution remotely sensed imagery classification,'' *Int. J. Pattern Recognit. Artif. Intell.*, vol. 33, no. 4, pp. 1–15, 2019.

[6] W. Diao, X. Sun, X. Zheng, F. Dou, H. Wang, and K. Fu, ''Efficient saliency-based object detection in remote sensing images using deep belief networks,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 137–141, Feb. 2016.

[7] Y. Lin, H. He, Z. Yin, and F. Chen, ''Rotation-invariant object detection in remote sensing images based on radial-gradient angle,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 746–750, Apr. 2015.

[8] S. Zhang, G. He, H.-B. Chen, N. Jing, and Q. Wang, ''Scale adaptive proposal network for object detection in remote sensing images,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 864–868, Jun. 2019.

[9] P. Ding, Y. Zhang, P. Jia, and X.-L. Chang, ''A comparison: Different DCNN models for intelligent object detection in remote sensing images,'' *Neural Process. Lett.*, vol. 49, no. 3, pp. 1369–1379, Jun. 2019.

[10] Q. Shaohua, G. Wen, J. Liu, Z. Deng, and Y. Fan, ''Unified partial configuration model framework for fast partially occluded object detection in high-resolution remote sensing images,'' *Remote Sens.*, vol. 10, no. 3, pp. 464–487, 2018.

[11] J. Mũnoz-Marí, F. Bovolo, L. Gómez-Chova, L. Bruzzone, and G. Camp-Valls, ''Semisupervised one-class support vector machines for classification of remote sensing data,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 8, pp. 3188–3197, Aug. 2010.

[12] X. Xie, Y. Zhang, X. Ling, and X. Wang, ''A novel extended phase correlation algorithm based on log-Gabor filtering for multimodal remote sensing image registration,'' *Int. J. Remote Sens.*, vol. 40, no. 14, pp. 5429–5453, Jul. 2019.

[13] Q. Li, W. Zhang, M. Li, J. Niu, and Q. M. J. Wu, ''Automatic detection of ship targets based on wavelet transform for HF surface wavelet radar,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 714–718, May 2017.

[14] W. Zhou, C. Wang, B. Xiao, and Z. Zhang, ''Action recognition via structured codebook construction,'' *Signal Process., Image Commun.*, vol. 29, no. 4, pp. 546–555, Apr. 2014.

[15] Q. U. Zhong, Z. Kang, and Q. Gao-Yuan, ''Research on algorithm of moving target detection and tracking based on MB-LBP feature extraction and particle filter,'' *Comput. Sci.*, vol. 659, no. 12, pp. 75–78, 2013.

[16] K. Takagi, K. Tanaka, S. Izumi, H. Kawaguchi, and M. Yoshimoto, ''A real-time scalable object detection system using low-power HOG accelerator VLSI,'' *J. Signal Process. Syst.*, vol. 76, no. 3, pp. 261–274, Sep. 2014.

[17] G. Zhang, F. Fang, A. Zhou, and F. Li, ''Pan-sharpening of multi-spectral images using a new variational model,'' *Int. J. Remote Sens.*, vol. 36, no. 5, pp. 1484–1508, Mar. 2015.

[18] R. Maalek, D. D. Lichti, and J. Y. Ruwanpura, ''Robust segmentation of planar and linear features of terrestrial laser scanner point clouds acquired from construction sites,'' *Sensors*, vol. 18, no. 3, pp. 819–849, 2018.

[19] L.-K. Shi, H. Zhou, and W.-H. Liu, ''Multi-feature fusion and visualization of pavement distress images based on manifold learning,'' *J. Highway Transp. Res. Develop.*, vol. 11, no. 1, pp. 14–22, Mar. 2017.

[20] A. Shaaban, M. Sayed, M. F. O. Hameed, H. I. Saleh, L. R. Gomaa, Y.-C. Du, and S. S. A. Obayya, ''Fast parallel beam propagation method based on multi-core and many-core architectures,'' *Optik*, vol. 180, pp. 484–491, Feb. 2019.

[21] Y. Chen, X. Zhao, and X. Jia, ''Spectral–spatial classification of hyperspectral data based on deep belief network,'' *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015.

[22] H. Shao, H. Jiang, H. Zhang, W. Duan, T. Liang, and S. Wu, ''Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing,'' *Mech. Syst. Signal Process.*, vol. 100, pp. 743–765, Feb. 2018.

[23] J. Zheng, X. Fu, and G. Zhang, ''Research on exchange rate forecasting based on deep belief network,'' *Neural Comput. Appl.*, vol. 31, no. S1, pp. 573–582, Jan. 2019.

[24] F. Saba, M. J. V. Zoej, and M. Mokhtarzade, ''Optimization of multiresolution segmentation for object-oriented road detection from high-resolution images,'' *Can. J. Remote Sens.*, vol. 42, no. 2, pp. 75–84, Mar. 2016.

[25] A. Ghasemzadeh and H. Demirel, ''3D discrete wavelet transform-based feature extraction for hyperspectral face recognition,'' *IET Biometrics*, vol. 7, no. 1, pp. 49–55, Jan. 2018.

[26] X. Zhang and H. Dai, ''Significant wave height prediction with the CRBM-DBN model,'' *J. Atmos. Ocean. Technol.*, vol. 36, no. 3, pp. 333–351, Mar. 2019.

[27] Y. Feng, Q. Chen, C. Li, and W. Hao, ''Research on algorithm for partial discharge of high voltage switchgear based on speech spectrum features,'' *Adv. Model. Anal. B*, vol. 60, no. 2, pp. 403–415, Jun. 2017.

[28] F. Huang, Y. Chen, W. Yin, W. Lin, X. Ye, W. Guo, and A. Reykowski, ''A rapid and robust numerical algorithm for sensitivity encoding with sparsity constraints: Self-feeding sparse SENSE,'' *Magn. Reson. Med.*, vol. 64, no. 4, pp. 1078–1088, Oct. 2010.

[29] C. Tao, H. Pan, Y. Li, and Z. Zou, ''Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2438–2442, Dec. 2015.

[30] O. Yildirim, R. S. Tan, and U. R. Acharya, ''An efficient compression of ECG signals using deep convolutional autoencoders,'' *Cognit. Syst. Res.*, vol. 52, pp. 198–211, Dec. 2018.

[31] M. Hamouda, K. S. Ettabaa, and M. S. Bouhlel, ''Hyperspectral imaging classification based on convolutional neural networks by adaptive sizes of windows and filters,'' *IET Image Process.*, vol. 13, no. 2, pp. 392–398, Feb. 2019.

[32] B. Liu, Q. Zhang, Y. Li, W. Chang, and M. Zhou, ''Spatial–spectral jointed stacked auto-encoder-based deep learning for oil slick extraction from hyperspectral images,'' *J. Indian Soc. Remote Sens.*, vol. 47, no. 12, pp. 1989–1997, Dec. 2019.

[33] S. Bera and V. K. Shrivastava, ''Analysis of various optimizers on deep convolutional neural network model in the application of hyperspectral remote sensing image classification,'' *Int. J. Remote Sens.*, vol. 41, no. 7, pp. 2664–2683, Apr. 2020.

[34] Y. Chang, L. Yan, H. Fang, S. Zhong, and W. Liao, ''HSI-DeNet: Hyperspectral image restoration via convolutional neural network,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 667–682, Feb. 2019.

[35] F. Huang, Y. Yu, and T. Feng, ''Automatic extraction of impervious surfaces from high resolution remote sensing images based on deep learning,'' *J. Vis. Commun. Image Represent.*, vol. 58, pp. 453–461, Jan. 2019.

[36] C. Peng, Y. Li, L. Jiao, Y. Chen, and R. Shang, ''Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation,'' *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2612–2626, Aug. 2019.

[37] Z. Zou and Z. Shi, ''Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images,'' *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1100–1111, Mar. 2018.

[38] M. Dong, S. Wen, Z. Zeng, Z. Yan, and T. Huang, ''Sparse fully convolutional network for face labeling,'' *Neurocomputing*, vol. 331, pp. 465–472, Feb. 2019.

[39] R. Pradeep and K. S. Rao, ''Incorporation of manner of articulation constraint in LSTM for speech recognition,'' *Circuits, Syst., Signal Process.*, vol. 38, no. 8, pp. 3482–3500, Aug. 2019.

[40] D.-T. Hoang and H.-J. Kang, ''Rolling element bearing fault diagnosis using convolutional neural network and vibration image,'' *Cognit. Syst. Res.*, vol. 53, pp. 42–50, Jan. 2019.

[41] Z. Li, D. Jiang, Y. Liu, and H. Li, ''Spatial context and locality-constraint based linear feature coding,'' *Jisuanji Fuzhu Sheji Yu Tuxingxue Xuebao/J. Comput. Aided Des. Comput. Graph.*, vol. 29, no. 2, pp. 254–261, 2017.

[42] R. M. Bommisetty, O. Prakash, and A. Khare, ''Video superpixels generation through integration of curvelet transform and simple linear iterative clustering,'' *Multimedia Tools Appl.*, vol. 78, no. 17, pp. 25185–25219, Sep. 2019.

[43] E. K. Ghasrodashti, "Hyperspectral image classification using a spectral–spatial random walker method," *Int. J. Remote Sens.*, vol. 40, no. 10, pp. 3948–3967, May 2019.

[44] H. Li, L. Jing, Y. Tang, and L. Wang, "An image fusion method based on image segmentation for high-resolution remotely-sensed imagery," *Remote Sens.*, vol. 10, no. 5, pp. 790–811, 2018.

[45] C. Li, M. Yang, X. Wang, H. Zhang, C. Yao, S. Sun, Q. Liu, H. Pan, S. Liu, Y. Huan, S. Li, J. Cao, X. Wang, Y. Guo, N. Guo, S. Jing, C. Zhang, and Z. Shen, "Glutazumab, a novel long-lasting GLP-1/anti-GLP-1R antibody fusion protein, exerts anti-diabetic effects through targeting dual receptor binding sites," *Biochem. Pharmacol.*, vol. 150, pp. 46–53, Apr. 2018.

[46] K. Zhang, M. Wang, S. Yang, and L. Jiao, "Spatial–spectral-graph-regularized low-rank tensor decomposition for multispectral and hyperspectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1030–1040, Apr. 2018.

[47] J. Feng, J. Chen, L. Liu, X. Cao, X. Zhang, L. Jiao, and T. Yu, "CNN-based multilayer spatial–spectral feature fusion and sample augmentation with local and nonlocal constraints for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1299–1313, Apr. 2019.

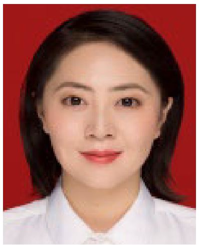
[48] H. Tuna, O. Arikan, and F. Arikan, "Model based computerized ionospheric tomography in space and time," *Adv. Space Res.*, vol. 61, no. 8, pp. 2057–2073, Apr. 2018.

[49] D. Hong and X. X. Zhu, "SULoRA: Subspace unmixing with low-rank attribute embedding for hyperspectral data analysis," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 6, pp. 1351–1363, Dec. 2018.

[50] H. Yeom, Y. Ko, and J. Seo, "Unsupervised-learning-based keyphrase extraction from a single document by the effective combination of the graph-based model and the modified C-value method," *Comput. Speech Lang.*, vol. 58, pp. 304–318, Nov. 2019.

YI LIU received the M.S. degree from Heriot-Watt University, U.K., in 2014, and the Ph.D. degree from Xinjiang University, China, in 2019. His research interests include hydrology and water resources management and simulation model.

MIN CHANG received the M.S. degree from the Shaanxi University of Science and Technology, in 2018. Her research interest includes land engineering.



JIE XU received the Ph.D. degree from Xinjiang University, China, in 2020. Her research interests include phytoremediation and ecological restoration.