Multiscale and Multidirection Feature Extraction Network for Hyperspectral and LiDAR Classification

Yi Liu, Zhen Ye, Yongqiang Xi, Huan Liu, Wei Li, *Senior Member, IEEE*, and Lin Bai, *Member, IEEE*

***Abstract*—Deep learning (DL) plays an increasingly important role in Earth observation by multisource remote sensing. However, the current DL-based methods do not make a fully use of the complementary information among multisource remote sensing data, such as hyperspectral image and light detection and ranging data, and lack the consideration of multiscale, directional, and fine-grained features. To address these issues, a multiscale and multidirection feature extraction network is proposed in this article. Specifically, the multiscale spatial feature (MSSpaF) module is designed to extract the MSSpaFs, and then, these features are fused by feature concatenation operation. In addition, the multidirection spatial feature module is designed to further extract multidirection and frequency information, employing cross-layer connection and multiscale feature fusion strategy to improve the fineness of the proposed network. Moreover, the spectral feature module is employed to provide detailed spectral information for enhancing the expression ability of multiscale features. Experimental results on three different datasets demonstrate the superior classification performance of the proposed framework.**

***Index Terms*—Convolutional neural network (CNN), feature extraction, hyperspectral image (HSI), light detection and ranging (LiDAR), multisource remote sensing.**

I. Introduction

HYPERSPECTRAL image (HSI), which contains rich spectral and spatial information [1], has been widely used in resource management, urban planning, forest monitoring, military, and security fields [2], [3], [4], [5]. However, for many urban and rural scenes, it is difficult for HSI to distinguish objects with similar spectral features (SpeF). For example, roads and roofs of buildings are both made of concrete. Light detection and ranging (LiDAR) data record the height and structure of different surface objects, which can provide valuable supplementary information. Since LiDAR provides digital surface model (DSM) data that incorporate elevation information [6], joint classification of HSI and LiDAR data for object recognition is a promising branch [6], [7].

Compared with single-source data, multisource data can provide more comprehensive information and achieve better monitoring effects [8]. Some studies show that complementary information obtained by feature fusion of multisource data has great potential in the field of remote sensing classification [9], [10]. Since HSI and LiDAR data contain information of different attributes, simple concatenation or stacking of features may be limited in individual feature extraction. Therefore, effectively extracting and fusing multisource data are crucial for multisource classification.

Many methods have been proposed to combine HSI and LiDAR data for remote sensing classification. The traditional methods are based upon feature fusion strategies, which manually design feature extractors. The typical filtering algorithms include morphological profile (MP), attribute profile (AP), and extinction profile (EP). Liao et al. [11] used MP to extract features from HSI and LiDAR data, a support vector machine (SVM) for feature-level classification, and finally, a weighted majority vote for joint decision-level classification. Ghamisi et al. [12] used APs to extract spatial features from HSI and LiDAR data and connected the extracted features to obtain better classification results. In order to further improve classification performance, Ghamisi et al. [13] employed EP to extract spatial features from HSI and elevation features from LiDAR data. In [14], a technique called sparse and low-rank component analysis was proposed to fuse HSI and LiDAR data. Kang et al. [15] proposed an effective probabilistic optimization method based on an extended random walk for HSI and LiDAR data classification. Xia et al. [16] proposed semisupervised graph fusion, in which morphological filters were applied to the first several components of LiDAR and HSI data, and then, spectral, spatial, and elevation features were projected into a lower subspace to obtain joint features. In [17], multifeature-based super-pixel-level decision fusion was proposed to obtain discriminative Gabor features of HSI and LiDAR data. However, traditional feature extraction methods mentioned earlier are limited in extracting high-level semantic information and digging latent representations from raw data.

Feature extraction methods based on deep learning (DL) networks are receiving increasing attention [18], [19], [20], because they can learn deep semantic features in an end-to-end manner to promote classification performance [21], [22]. In [13], convolutional neural network (CNN) was early used for HSI and LiDAR joint classification. Spatial and elevation features of HSI

Manuscript received 29 January 2024; revised 23 April 2024 and 30 April 2024; accepted 8 May 2024. Date of publication 14 May 2024; date of current version 30 May 2024. This work was supported in part by the National Key Research & Development Program of China under Grant 2020YFC1512002 and in part by the Hangzhou Major Science and Technology Innovation under Project 2022AIZD0012. *(Corresponding author: Zhen Ye.)*

Yi Liu, Zhen Ye, Yongqiang Xi, and Lin Bai are with the School of Electronics and Control Engineering, Chang'an University, Xi'an 710064, China (e-mail: 2022232009@chd.edu.cn; yezhen525@126.com; 2020132069@chd.edu.cn; linbai@chd.edu.cn).

Huan Liu and Wei Li are with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: huanliu233@gmail.com; liwei089@ieee.org).

The source code of this method is available online at https://github.com/lyywowo/MSMD-Net.

Digital Object Identifier 10.1109/JSTARS.2024.3400872

and LiDAR data were extracted by EP and then fed into CNN to generate the final classification map. In [23], a dual-branch CNN was proposed to reduce model complexity through weight sharing, and feature and decision-level fusion strategies were used for the precise classification of multisource remote sensing data. In [24], a similar double-concentrated network was proposed for effectively extracting features and accurately classifying ground objects. The dual-centralized network captures the spectral and spatial features of HSI and expands the connection of LiDAR information based on the trained HSI branch. In [25], an interleaving perception CNN was proposed to extract and integrate HSI features and LiDAR features. A bidirectional autoencoder was designed to reconstruct HSI and LiDAR data, and finally, the concatenated feature map was sent to a two-branch CNN for classification.

Directly concatenating HSI and LiDAR features not only increases the dimensionalities of feature maps but also deteriorates classification accuracy due to ignoring the possible interaction information of multisource data in DL networks. How to effectively realize the information fusion and interaction for HSI and LiDAR data becomes a key for accurate classification. In [26], a cascade block is added to the two-branch network to realize spatial and spectral feature extraction. In [27], a three-branch CNN network was proposed, in which each CNN branch adopted the multilayer fusion (MLF) module to fuse shallow and deep features and then used the mutual guided attention (MGA) module to enhance the information of HSI and LiDAR data. In [28], a patch-to-patch CNN with an encoder–decoder structure was proposed by employing unsupervised learning methods to learn and integrate spatial, spectral, and elevation information from HSI and LiDAR data, generating superimposed vectors for final classification. In [29], a deep encoder–decoder network (EndNet) used an encoder network to extract features from multisource data and fused them to reconstruct multisource inputs. In [30], the CNNMRF method further utilized spatial information by introducing Markov random fields into convolutional networks. In [31], a new dual-channel spatial, spectral, and multiscale attention convolutional long short-term memory neural network (dual-channel A^3CLNN) was proposed to learn HSI and LiDAR features with multiscale attention mechanism and a transfer learning strategy. In [32], a nest-neighbor-based contrast learning network was proposed, which made full use of large amounts of unlabeled data to learn discriminant feature representations and designed bilinear attention modules to extract higher-order features of HSI and LiDAR data. In [33], an adaptive multiscale spatial–spectral enhancement network with multiple branches for HSI and LiDAR data classification was proposed. In this method, SpeFs of HSI were deeply mined by involutive operators, and spectral–spatial features were extracted by hierarchical fusion strategy.

As an alternative to using deep networks for HSI and LiDAR classification, there has been increasing interest in multiscale network architectures (e.g., [34], [35], [36]), which enlarge the range of the receptive fields of CNN-based networks. Typically, such a multiscale network consists of, in essence, multiple parallel branches of feature extraction in conjunction with some form of feature fusion prior to classification, with the branches being designed to independently extract features at differing scales. Commonly, each branch implements a cascade of two or more convolutional layers with feature summation or concatenation. In [34], a multiscale network and a single-branch backbone network were designed, and the proposed position-channel cooperative attention module adaptively enhanced the features extracted from a multiscale network, so as to obtain the comprehensive features of HSI and LiDAR data and reduce the semantic differences of heterogeneous features. In [35], a disentangled nonlocal network was proposed, which used multiscale modules to capture spectral and spatial information. In [37], a multiscale cross-level attention learning network was proposed to fully mine the global and local multiscale features for classification. In [36], a new multiscale network with self-calibrated convolution was proposed, using hierarchical residual structure and self-calibrated convolution to extract features with different receptive fields, which can enhance the ability to represent multisource data.

Although the methods mentioned above can perform well in multiscale feature extraction, there is a lack of direction information representation. Combining the Gabor filter into the CNN can enhance the learning of multiscale and multidirection features and favor texture information representation [38]. In [39], the proposed fractional Gabor transformation enables data analysis in both the real space and the frequency domain simultaneously. The spatial pattern can be rotated as the fractional order changes. In [40], the fractional Gabor convolutional network (FGCN) was proposed for HSI and LiDAR classification, using Octave convolution to reduce redundant low-frequency information and fractional Gabor convolutional layers to extract multidirection information, improving feature diversity and integrity. The previous methods, such as FGCN, simply used 1-D convolution to extract SpeFs of HSI data. In addition, the independent extraction of spatial features by HSI and LiDAR channels will degrade the classification performance due to the large differences in heterogeneous features. Third, widely used fusion methods are relatively simple, limiting the representation of semantic information. Considering the previous problems, a multiscale and multidirection feature extraction network is proposed in this article. Specifically, the multiscale spatial feature (MSSpaF) module is designed to extract spatial and elevation features for HSI and LiDAR data with a cross-channel connection improving the interaction of heterogeneous features. Additionally, the multidirection spatial feature (MDSpaF) module is designed for multidirection representation of the spatial and elevation features, effectively mining fine-grained features through cross-layer connection and multiscale fusion. Finally, the SpeF module is designed to allocate channel weights for obtaining detailed spectral information.

The primary contributions of this article are as follows.

1) In this article, multiscale features are extracted from HSI and LiDAR data via a double-branch capitalizing on the MSSpaF module with hierarchical residual connections. The multiscale nature of the feature extraction is finely grained by using information interaction of multisource data.

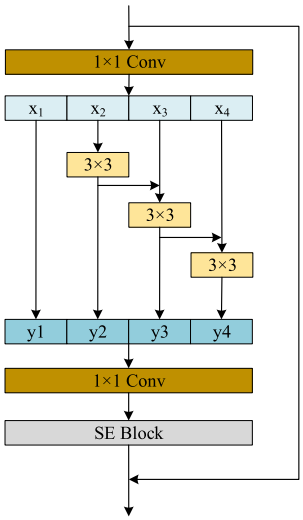

Fig. 1. Structure diagram of Res2Net.

2) The proposed MDSpaF module uses fractional Gabor filters to modulate convolution, enabling the extraction of multiscale and multidirection features. In different feature extraction stages, multiscale fractional Gabor filters can improve the complementarity of heterogeneous features. Our approach is, to the best of our knowledge, the first to apply multiscale and multidirection feature extraction for joint classification of HSI and LiDAR data.
3) The SpeF module is designed to fully extract useful SpeFs by adding a weight allocation mechanism. Instead of simply stacking or adding spatial and SpeFs, the semantic correlation of multisource data is strengthened by spatial feature weight consideration in the feature fusion stage.

The rest of this article is organized as follows. Section II introduces the relevant theories of Res2Net and Gabor transform. Section III describes the proposed network structure in detail. Section IV gives the experimental results and analysis. Finally, Section V concludes this article.

II. Related Work

A. Res2Net

Gao et al. [41] proposed a novel CNN building block, Res2Net, by constructing layered residual-like connections in a single residual block. It represents fine-grained multiscale features and enlarges each layer's receptive fields. The structure of the Res2Net module is shown in Fig. 1. Instead of using a set of 3×3 filters in a bottleneck block, the module seeks an alternative architecture with stronger multiscale feature extraction capability while maintaining similar computational complexity. Specifically, after conducting a 1×1 convolution, the input feature map is evenly divided into s subbands, represented by x_i, where $i \in \{1, 2, \ldots, s\}$. After division, the size of the subband feature map remains unchanged while the number of bands is $1/s$. Except for the first subband feature graph x_1, the other subbands x_i have a 3×3 convolution layer, which is represented by C_i. The output of subband features x_i is represented by y_i, which can be expressed as

$$y_i = \begin{cases} x_i, & i = 1 \\ C_i x_i, & i = 2 \\ C_i(x_i + y_{i-1}), & 2 < i \leq s \end{cases} \tag{1}$$

where each 3×3 convolution can obtain information from subbands, so the output features have larger receptive fields. Moreover, the output of the Res2Net module contains feature combinations with different numbers and scales, enabling the Res2Net module to fully extract spatial features of remote sensing images.

B. Gabor Transform

Although CNN has a strong ability to learn features from multisource remote sensing images, it lacks the description of direction and scale information. Because the Gabor filter has good characteristics of time-domain and frequency-domain transformation, it can guide CNN to obtain MSSpaF and MDSpaFs. In the backpropagation process, only a few parameters need to be updated due to manually modulated convolution kernels, reducing computational complexity. The complex form of the Gabor function can be regarded as the product of the Gaussian and sine function

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma}\right) \exp\left[j\left(2\pi \frac{x'}{\lambda} + \psi\right)\right] \tag{2}$$

where

$$\begin{cases} x' = x\cos\theta + y\sin\theta \\ y' = -x\sin\theta + y\cos\theta. \end{cases} \tag{3}$$

The different parameters of the Gabor function have different influences on the filter. θ represents the direction of the filter, λ represents the wavelength of the filter, ψ represents the phase shift of the sine function in the Gabor function, σ denotes the standard deviation of the Gaussian factor, and γ denotes the aspect ratio.

III. Methodology

A. Overall Architecture

As shown in Fig. 2, the network comprises three parts, i.e., MSSpaF extraction module, MDSpaF extraction module, and SpeF module. The MSSpaF extraction module contains three residual structures, each of which has three Res2Net layers to extract and fuse the spatial features of the HSI and LiDAR data, respectively. The MDSpaF extraction module contains three fractional Gabor convolution blocks, which can extract texture, direction, and transformation information. To obtain fine-grained features, feature reuse operations are conducted in each fractional Gabor convolution block. The third part introduces the SpeF module and classification framework. The SpeF module is employed to extract detailed spectral information as a supplement. The classification results are obtained by assigning

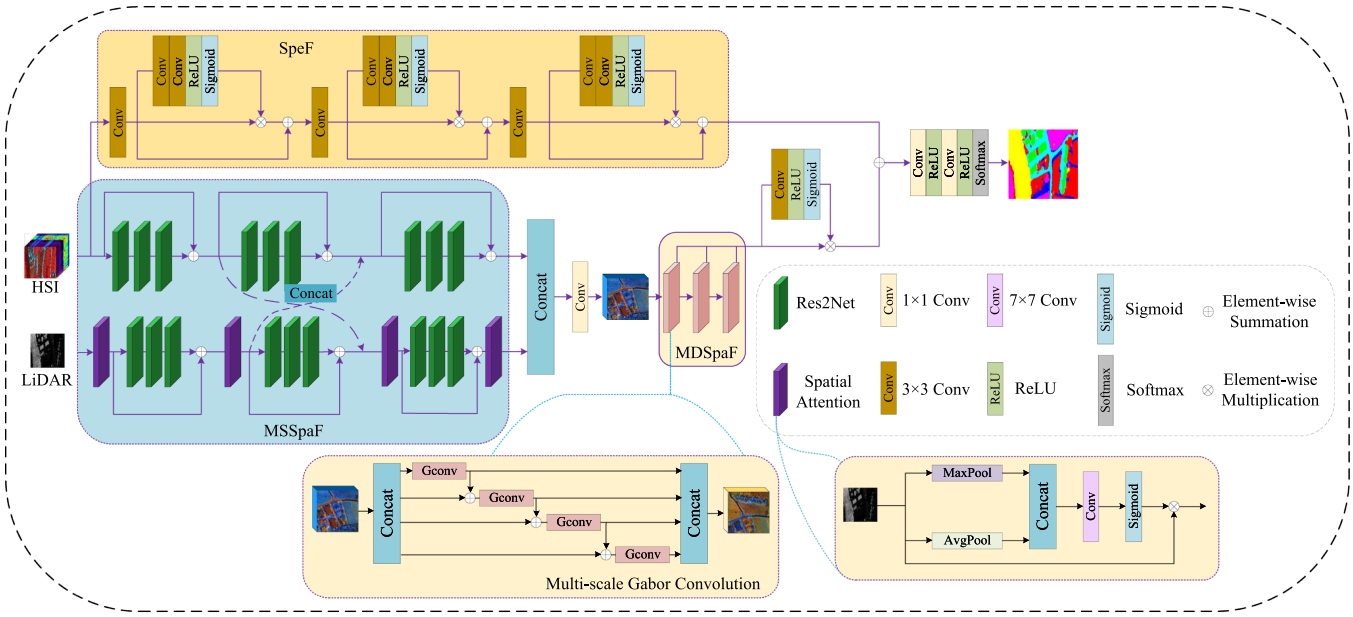

Fig. 2. Framework of the proposed method.

weights to spatial features and concatenating them with SpeFs, which are then fed into the Softmax layer.

B. Multiscale Spatial Feature

Fundamentally, we use Res2Net blocks to extract the spatial features of the HSI and LiDAR, as shown in Fig. 1. Each layer of the Res2Net block is added with a residual structure, which enlarges the receptive fields of each layer. In this step, the HSI can be represented as $X_h \in \mathbb{R}^{r\times c\times b_h}$, which means that there are b_h bands with $r \times c$ pixels in each band. The LiDAR image of the same region is defined as $X_l \in \mathbb{R}^{r\times c\times b_l}$ with b_l bands. In each 3×3 convolution layer, the parameters are set as $stride = 1$ and $padding = 1$, and the number of output channels (y_1, y_2, y_3, and y_4) is the same as the number of input channels (x_1, x_2, x_3, and x_4). After the convolution operation, the feature sizes of the input channels and the output channels remain the same. Therefore, the splicing operation can be directly carried out, and then, the features of output channels are linearly integrated. Finally, the Squeeze-and-Excitation (SE) module that imposes different weights on different channels is adopted to enhance or suppress the information of different channels. For HSI data, the number of subsets in Res2Net blocks is set to 4. For the LiDAR image, the number of subsets in Res2Net blocks is set to 2. As shown in Fig. 2, every three Res2Net blocks act as a basic spatial feature extraction module that operates through residual connections, which can alleviate problems such as gradient disappearance and gradient explosion. Taking the HSI branch as an example, $X_h^1 \in \mathbb{R}^{r\times c\times b_{h1}}$, $X_h^2 \in \mathbb{R}^{r\times c\times b_{h2}}$, and $X_h^3 \in \mathbb{R}^{r\times c\times b_{h3}}$ are obtained by basic spatial feature extraction modules in different layers. Because the LiDAR image mainly contains elevation information and weakly contains spatial information, a spatial attention block is added at intervals of three Res2Net blocks to help enhance spatial feature representation. Thus, $X_l^1 \in \mathbb{R}^{r\times c\times b_{l1}}$, $X_l^2 \in \mathbb{R}^{r\times c\times b_{l2}}$, and $X_l^3 \in \mathbb{R}^{r\times c\times b_{l3}}$ are obtained on different layers of the LiDAR branch.

It is worth noting that in a two-branch network, the concatenation of features without considering the inherent differences of multisource data often leads to significant differences in features, which reduces classification accuracy. To solve this problem, two cross-connections are set up between HSI and LiDAR branches for information interaction, which is shown as dotted lines in the MSSpaF module in Fig. 2. Before the third basic feature extraction module, the features of the two branches are concatenated as the input

$$X_h^3 = \text{Concat}\left[X_l^1, X_h^2\right] \tag{4}$$

$$X_l^3 = \text{Concat}\left[X_h^1, X_l^2\right]. \tag{5}$$

This is because features X_h^1 and X_l^1 contain spatial information of HSI data and elevation information of LiDAR data, respectively. If the cross-connections are performed for first or third basic feature extraction modules, feature redundancy or feature stacking will occur. It is also noted that the features concatenation of two branches will lead to the number of input feature channels increasing. That is to say, although the number of input and output feature channels will not change during the process of the spatial feature extraction, there are $b_{h1} = b_{h2} \neq b_{h3}$ and $b_{l1} = b_{l2} \neq b_{l3}$ after the third spatial feature extraction module. Therefore, a 1×1 convolution operation is added to keep the number of feature channels consistent for feature fusion.

C. Multidirection Spatial Feature

Just as the Gabor transform is derived from the Fourier transform, the fractional Gabor transform is derived from the fractional Fourier transform. The fractional Fourier transform

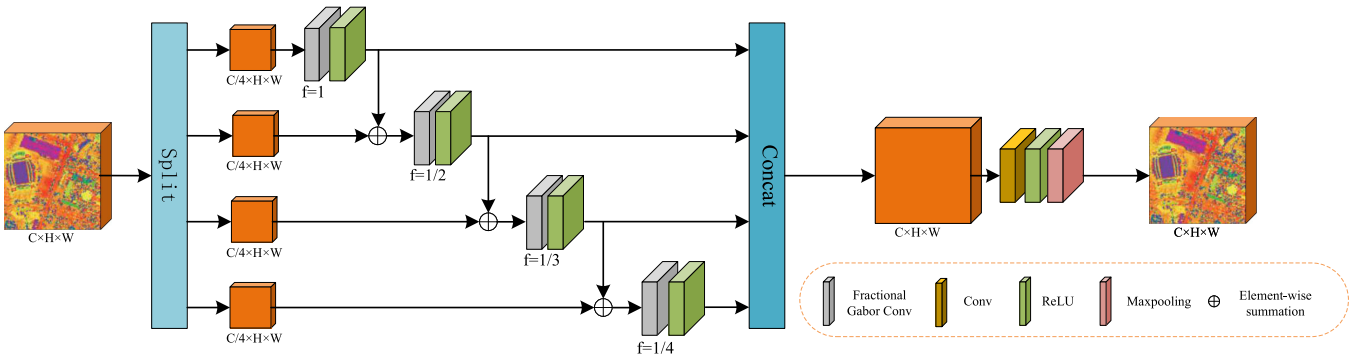

Fig. 3. Structure of the multiscale Gabor convolution module.

can be expressed as

$$X_p(u) = \int_{-\infty}^{+\infty} f(x) K_{px}(x,u) dx \tag{6}$$

where the kernel function $K_{px}(x,u)$ can be defined as

$$K_{px}(x,u) = \begin{cases} A_p \exp\left(j\left(\frac{x^2+u^2}{2\tan p} - \frac{xu}{\sin p}\right)\right), & p \neq n\pi \\ \delta(x-u), & p = 2n\pi \\ \delta(x+u), & p = (2n+1)\pi \end{cases} \tag{7}$$

where $A_p = \sqrt{(1-j\cot p)/2\pi}$, and p represents the fractional order.

The 2-D fractional Fourier transform can be conducted by

$$\begin{aligned} &F_{x,y}(x,y,u,v) \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y) K_{px}(x,u) K_{py}(y,v) dxdy \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y) K_{px,py}(x,y,u,v) dxdy. \end{aligned} \tag{8}$$

The obtained 2-D fractional Gabor transform can be described as

$$\begin{aligned} &G_{x,y}(x,y,u,v) \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y) K_{px}(x,u) K_{py}(y,v) h(x,y) dxdy \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y) K_{px,py}(x,y,u,v) h(x,y) dxdy \end{aligned} \tag{9}$$

where (x,y) represents the spatial position, u and v represent the spatial position corresponding to x and y in a fractional Gabor domain, $f(x,y)$ represents the input image, $K_{px,py}(x,y,u,v)$ represents the 2-D fractional Gabor transform kernel, px and py represent the fractional components in the horizontal and vertical directions, and $h(x,y)$ represents the 2-D Gaussian function. Gabor filter has two essential parameters: 1) scale and 2) direction. In order to extract multiscale and multidirection features of HSI and LiDAR, frequency (f) and direction angle (θ) are determined as parameters by the grid search method. In fact, the classification accuracy will increase with the number of f and θ. Considering the complexity of the network, these two parameters are set as

$$\theta \in \left\{0, \frac{\pi}{16}, \frac{\pi}{8}, \frac{3\pi}{16}, \ldots, \frac{15\pi}{16}\right\} \tag{10}$$

$$f \in \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}\right\}. \tag{11}$$

In this work, the MDSpaF module is designed to extract the spatial features of HSI, the elevation features of the LiDAR image. As shown in Fig. 2, MDSpaF module is the second part of the proposed network. And the details are presented in Fig. 3. An interlayer residual structure is added in each fractional Gabor convolution block; a jump connection is added between two adjacent channels. With the addition of a jump connection, the receptive field of the fractional Gabor convolution block is expanded. Followed by a Conv layer, ReLU layer, and MaxPooling layer, multiscale and multidirection features are obtained. Notably, previous feature extraction methods show that low-level features contain rich spatial information but lack semantic information while high-level features are the opposite. Multiscale feature fusion can introduce more semantic information into low-level features and embed more spatial information into high-level features, to obtain more comprehensive multiscale features. Therefore, the proposed MDSpaF module employs three fractional Gabor convolution blocks to extract spatial feature map f_{spat}, balancing spatial and semantic information.

D. SpeF and Classification

HSI contains rich spectral bands, and most existing SpeF extraction methods focus on 1-D stack convolution, so they will not be able to make full use of much spectral information. In this work, we propose an SpeF module, in which the spectral channel attention mechanism is considered to enhance or suppress different spectral channels by assigning weights for complementary spectral information. As shown in Fig. 2, by employing two 1×1 convolutional layers, a ReLU layer, and a Sigmoid layer, the weights of spectral channels will be obtained. Assigning different weights to different channels helps the network balancing information contributions between different layers. And, operating residual connections achieves information supplement since some useful information may be

lost during feature extraction. The SpeF map can be obtained by

$$f_i^{\text{spec}} = (W_i^{\text{spec}} \otimes f_{i-1}^{\text{spec}}) \oplus f_{i-1}^{\text{spec}} \tag{12}$$

where W_i^{spec} represents the spectral weight, which can be expressed as

$$W_i^{\text{spec}} = \text{Sigmoid}(\text{ReLU}(\text{Conv}(\text{Conv}(f_{i-1}^{\text{spec}})))) \tag{13}$$

where f_i^{spec} represents the SpeF of the ith layer, $\otimes$ represents elementwise multiplication, and $\oplus$ represents the element addition operation. The final SpeF f_{spec} is obtained by the SpeF module. f_{spec} contains SpeFs obtained by assigning weights to spectral channels, which provides supplementary information for multisource data joint classification.

By now, we have obtained the SpeF map f_{spec} for HSI and the fused spatial feature map f_{spat} for HSI and LiDAR data. If the spatial features and SpeFs are simply stacked or added, their semantic relevance will be limited and the effect of feature fusion will be greatly reduced. Considered spatial attention mechanism, the fused feature map can be expressed as f_{fuse}

$$f_{\text{fuse}} = f_{\text{spec}} + W_{\text{spat}} \otimes f_{\text{spat}} \tag{14}$$

where

$$W_{\text{spat}} = \text{Sigmoid}(\text{ReLU}(\text{Conv}(f_{\text{spat}}))). \tag{15}$$

Multiplying W_{spat} by spatial feature f_{spat} can allocate weights to spatial features to increase useful spatial information and suppress redundant spatial information. Thus, the feature fusion is implemented by (13), where W_{spat} is obtained by performing a 3×3 convolution operation, a ReLU operation, and a Sigmoid operation. Finally, classification is processed by two 1×1 convolutions, two ReLU activation functions, and a softmax activation. The steps of the proposed network are summarized in Algorithm 1.

E. Motivations Analysis of the Proposed Method

HSIs contain rich spectral and spatial information, which can identify ground targets. LiDAR data provide elevation information, recording the height and structure of objects. The combined application of HSI and LiDAR data allows for more detailed and precise identification of ground objects. On this basis, the proposed network fully considers spatial–spectral characteristics of HSI and spatial-elevation characteristics of LiDAR data. The MSSpaF module extracts the spatial features of HSI and LiDAR data respectively, retains the LiDAR elevation information, and combines these features. MDSpaF module is used to extract the multidirection features, which can be obtained by combining the modulated Gabor kernel with CNN, yielding accurate classification results, especially for the objects with different shapes. In addition, the SpeF module is designed to extract SpeFs of HSI as a complement to the fused feature map.

IV. Experiments and Analysis

A. Dataset Description

MUUFL Gulfport [42]: The MUUFL Gulfport dataset was collected in November 2010 at the University of Southern Mississippi campus in Bay Park, Long Beach, Mississippi, USA.

Algorithm 1: Pseudocode for our method Training.

Input: HSI data X_h, LiDAR data X_l, Ground truth Y.
Output: Classification map M.

```
Parameter setting and weight initialization.
for epochs do
    Extract the MSSpaFs X_h^3 and X_l^3 of HSI data and LiDAR data according to Eq. (5) and Eq. (6), respectively, and feed these features into the MDSpaF module. Then, MSSpaF and MDSpaF map f_spat will be obtained.
    Extract SpeFs f_spec from HSI data according to Eq. (8)
    Take f_spat multiplied by spatial weights and f_spec as to obtain f_fuse and then get the classification map through the Softmax classifier.
    Train network as shown in Fig. 2.
end for
Obtain the probability distribution and classification map.
```

TABLE I
Number of Training and Testing Samples for the MUUFL Data

Class		Number of samples	
No.	Name	Train	Test
1	Trees	100	23246
2	Mostly grass	100	4270
3	Mixed ground surface	100	6882
4	Dirt and sand	100	1826
5	Road	100	6687
6	Water	100	466
7	Building shadow	100	2233
8	Building	100	6240
9	Sidewalk	100	1385
10	Yellow curb	100	183
11	Cloth panels	100	269
Total		1100	53 687

HSI data were collected by the Compact Airborne Spectrographic Imager (CASI)-1500 sensor, covering the wavelength range of 367.7–1043.4 nm. The original HSI contains 72 spectral bands. Eight noise bands were removed, and the remaining 64 bands were used for experiments. HSI is composed of 325×220 pixels, and the spatial resolution is 0.54 m. The DSM of LiDAR data was obtained by the airborne laser terrain mapper sensor using a laser with a wavelength of 1064 nm. MUUFL Gulfport dataset covers 11 classes, as shown in Table I.

Houston 2013 [6]: The Houston 2013 dataset is composed of HSI and LiDAR-based DSM. The size of the data is 349×1905 pixels. The HSI scene is acquired by the ITRES CASI-1500 sensor, which consists of 144 spectral bands with wavelengths ranging from 0.38 to 1.05μm, including 15 classes, as shown in Table II. The spatial resolution of both HSI and LiDAR-based DSM is 2.5 m.

Trento [43]: The Trento dataset covers a rural area in southern Trento, Italy. The HSI data were collected by the AISA Eagle sensor, covering 63 spectral bands with a spectral range of 402.89−989.09 nm and a spatial resolution of 1 m. The DSM of LiDAR data was collected by the Optech Airborne Laser

TABLE II
NUMBER OF TRAINING AND TESTING SAMPLES FOR THE HOUSTON DATA

Class		Number of Samples	
No.	Name	Train	Test
1	Health grass	100	1251
2	Stressed grass	100	1254
3	Synthetic grass	100	697
4	Tress	100	1244
5	Soil	100	1242
6	Water	100	325
7	Residential	100	1268
8	Commercial	100	1244
9	Road	100	1252
10	Highway	100	1227
11	Railway	100	1235
12	Parking lot 1	100	1233
13	Parking lot 2	100	469
14	Tennis court	100	428
15	Running track	100	660
Total		1500	15029

TABLE III
NUMBER OF TRAINING AND TESTING SAMPLES FOR THE TRENTO DATA

Class		Number of Samples	
No.	Name	Train	Test
1	Apple trees	50	4034
2	Buildings	50	2903
3	Ground	50	479
4	Woods	50	9123
5	Vineyard	50	10501
6	Roads	50	3174
Total		300	30214

Topographer (ALTM) 3100EA sensor, which is composed of 166 $\times$600 pixels and the spatial resolution is 1 m. This dataset has six classes, as shown in Table III.

B. Experimental Setup

All DL methods are implemented in TensorFlow and the high-level API Keras framework. To optimize them, we use the Adam algorithm. Keras is a simplified interface to TensorFlow, an open-source software library for numerical calculations using data flow diagrams. All experiments were conducted on a PC equipped with Ubuntu 18.04 and a GTX-2080 GPU.

To gauge performance, we calculate overall accuracy (OA), average accuracy (AA), class-specific accuracy (CA), and Kappa coefficient (κ). OA is the ratio of the number of correct predictions to the total number of pixels in all test sets, AA denotes the average of individual class accuracies. The calculation of κ and CA is based on a confusion matrix, which is an indication of consistency. To avoid any bias induced by random sampling, we conducted ten trials and reported average results along with standard deviation.

C. Algorithm Configuration and Parameter Tuning

According to the mainstream CNN-based HSI and LiDAR joint classification methods, the fundamental hyperparameters are convolution kernel size *r*, fractional order *p*, and the learning rate lr [23], [26]. The tenfold cross-validation is adopted to select the optimal hyperparameters.

During the training of the proposed network, we divide the training/validation set randomly into two equal-size parts-one part is used to train the network while the other part is a validation set used to tune the network hyperparameters. To make the network model fully converge, the number of training rounds was set as $num\_epoch = 2000$. The value ranges of the learning rate are $\{1e-5, 1e-4, 1e-3, 1e-2, 1e-1\}$. According to experimental results, the proposed method is insensitive to this hyperparameter. The learning rate is $\mathrm{lr} = 1e-3$ in the following experiments. To evaluate the influence of other hyperparameters, i.e., fractional order *p* and convolution kernel size *r*, the value range of *r* is 3–13 and that of *p* is 0.05−0.5, respectively. As shown in Fig. 4, the network is sensitive to *p* and *r*. For example of the Trento dataset, 25 samples of the training set and the validation set were selected individually for parameter adjustment. The results show that when $r = 13$ and $p = 0.1$ OA is the highest, i.e., 0.9821. For the MUUFL dataset and the Houston 2013 dataset, 50 samples from the training and validation sets were selected for parameter adjustment. When $r = 3$ and $p = 0.2$, the OAs for these datasets are the highest, with values of 0.8883 and 0.9725, respectively. The experimental results indicate that features extracted using different convolutional kernels produce different classification performances. For scenes containing complex spatial texture information, a convolutional kernel with a relatively big size expects better classification results. For example, the proposed method reaches a satisfactory OA for the Houston dataset using a convolutional kernel with a size of $r = 13$. The Trento and MUUFL datasets require smaller convolutional kernels. For parameter *p*, a larger fractional order implies Gabor features close to the feature map in the frequency domain, and small fractional orders retain more spatial information. The Houston dataset contains more classes and the distribution of features is more complex, so smaller fractional order is helpful for extracting more useful spatial features. For the MUUFL and Trento datasets with relatively simple terrain distribution, the larger fractional orders were chosen.

D. Experimental Results and Analysis

In comparative experiments, multisource data classification performance was evaluated quantitatively and qualitatively. OA, AA, CA, and κ were used to evaluate the classification performance of the proposed and several comparable methods while classification maps were also shown in a qualitative perspective. The comparable methods include SVM [44], as well as five advanced DL methods, namely deep EndNet [29], two branch CNN (TBCNN) [26], two-channel CNN (Coupled CNN) [23], Markov-random-field-based CNN (CNNMRF) [30], and FGCN [39]. This article focuses on MSSpaF and MDSpaF extraction by incorporating feature reuse operations into fractional Gabor convolution and designing the SpeF module to supplement multisource features. To make a fair comparison, we selected the optimal hyperparameters of the corresponding articles.

Tables IV–VI list the average values with a standard deviation of OA, AA, κ, and CA for the three mentioned datasets. The

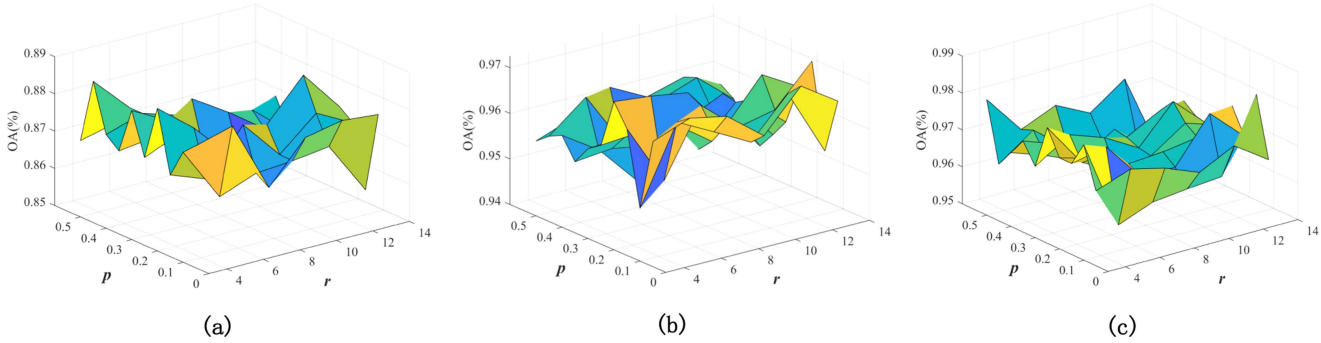

Fig. 4. Parameter tuning of p and r for the proposed network using three datasets. (a) MUUFL. (b) Houston 2013. (c) Trento.

TABLE IV
CLASSIFICATION ACCURACY (AVERAGE AND STANDARD DEVIATION) FOR THE MUUFL DATASET

No.	Class Name	Methods						
		SVM [45]	EndNet [30]	CNNMRF [31]	TBCNN [27]	Coupled-CNN [24]	FGCN [41]	Proposed
1	Trees	80.06±0.23	82.19±2.44	87.63±1.61	84.70±3.09	83.77±3.64	90.39±1.84	**90.58±1.94**
2	Mostly grass	70.26±0.33	77.10±3.00	75.91±2.94	79.13±2.35	74.45±3.29	82.39±5.04	**89.92±2.91**
3	Mixed ground surface	65.36±0.11	66.30±3.25	69.50±1.99	69.16±3.32	67.19±3.73	74.49±4.50	**81.85±3.82**
4	Dirt and sand	84.99±0.26	81.82±4.00	87.09±2.44	86.99±2.99	83.27±4.01	91.09±2.78	**93.51±2.69**
5	Road	82.27±0.31	86.06±2.12	76.58±2.13	83.12±1.88	83.41±2.76	85.25±1.47	**89.40±3.32**
6	Water	91.36±0.60	98.40±1.08	96.97±1.72	98.65±1.12	99.01±1.36	99.04±0.74	**99.34±0.61**
7	Building shadow	87.32±0.37	87.04±3.47	92.64±1.61	88.84±2.35	90.42±2.06	93.63±2.02	**94.50±2.61**
8	Building	76.34±0.41	89.25±1.55	77.81±3.48	92.95±1.70	90.70±2.36	**95.21±1.24**	94.64±1.79
9	Sidewalk	70.26±0.71	71.04±2.86	76.42±5.36	75.69±3.38	75.75±3.56	79.71±2.41	**85.55±3.50**
10	Yellow curb	86.43±0.38	93.15±3.24	92.20±4.08	88.25±3.83	88.65±4.62	92.33±3.79	**95.24±3.74**
11	Cloth panels	96.32±0.34	95.67±1.95	98.23±1.67	98.20±1.11	98.07±1.83	**98.94±1.07**	98.94±1.35
	OA	84.44±0.14	81.13±1.09	81.82±0.61	83.18±1.74	81.88±1.94	87.62±1.02	**89.72±0.96**
	AA	79.27±0.11	75.82±1.23	76.52±0.71	78.37±2.09	76.74±2.33	84.20±1.55	**86.62±1.21**
	κ	80.90±0.12	84.65±0.94	84.63±0.65	85.97±1.04	84.97±1.32	89.32±0.71	**91.96±0.89**

The value with the highest classification accuracy for each category is bold for the reader's convenience.

bold text in these tables represents the best values for the corresponding rows. The results of qualitative analysis for different datasets show that the traditional methods are easily affected by noise due to the lack of spatial information. For example, it can be seen from Trento datasets in Fig. 7(d) and (e) that it is difficult for SVM and EndNet to maintain spatial continuity. Specifically, certain ground objects, such as the *Vineyard* in the Trento dataset, were mistakenly classified as *Apple Trees* and *Woods*. Due to the deep encoder–decoder structure, EndNet can better extract features in the multisource data fusion stage, compared with SVM. The methods using spatial information, such as CNNMRF, TBCNN, and Coupled CNN, yield higher OAs than in previous two method shown in Tables IV–VI and smoother classification maps shown in Fig. 7(f)–(h). Compared with Fig. 7(d) and (e), the classification effect of *Vineyard*, *Apple Trees*, and *Woods* were significantly improved. However, their classification performance are limited when dealing with the classes with small samples, such as the *Ground* of the Trento dataset. Coupled CNN not only extracts spatial features but also adopts feature-level fusion strategy and decision-level feature fusion strategy for multisource data classification, further improving classification performance. Due to the residual structure adopted by TBCNN, its network depth can be increased to some extent while reducing the occurrence of the overfitting phenomenon. It has been proved that deep networks have greater potential in HSI and LiDAR joint classification. As our previous work, FGCN extracts both high-frequency and low-frequency features of HSI and LiDAR data by Octave convolution and extracts multiscale and multidirection features by fractional Gabor convolution improving classification performance to a certain degree. As shown in Tables IV–VI, OA of the MUUFL dataset is 87.62%, that of the Houston dataset is 97.06%, and that of the Trento dataset is 98.31%.

According to the classification results of Tables IV–VI, it is obvious that the proposed network can get the best classification performance. The reasons may be that, MSSpaFs are extracted by Res2Net with cross-connections between HSI branch and LiDAR branch enhancing information interaction; feature reuse is beneficial to obtaining fine-grained information of multisource data; residual structure can guarantee the depth of network at the same time, reducing the overfitting phenomenon; multidirection features can be extracted by fractional Gabor convolution, in which multiscale feature fusion is operated to enlarge the receptive fields of convolution.

TABLE V
CLASSIFICATION ACCURACY (AVERAGE AND STANDARD DEVIATION) FOR THE HOUSTON DATASET

No.	Class Name	Methods						
		SVM [45]	EndNet [30]	CNNMRF [31]	TBCNN [27]	Coupled-CNN [24]	FGCN [41]	Proposed
1	Health grass	97.30±0.32	97.72±1.27	95.44±1.49	97.18±1.09	97.46±1.80	97.33±1.34	**99.02±0.85**
2	Stressed grass	97.24±1.53	97.42±1.29	97.69±1.61	97.99±1.08	97.78±1.12	**98.62±0.57**	98.60±0.37
3	Synthetic grass	99.28±0.11	98.68±0.89	98.23±1.29	**99.71±0.68**	99.60±0.43	99.29±0.72	99.39±0.77
4	Tress	97.59±2.11	92.06±2.33	96.45±2.37	97.66±1.46	98.09±1.08	98.26±0.59	**99.02±1.01**
5	Soil	98.79±0.54	96.11±1.55	97.88±1.02	96.57±2.19	87.87±1.03	**99.12±0.63**	98.45±0.37
6	Water	96.92±1.64	95.55±1.67	97.53±1.96	97.73±1.85	96.83±2.30	**99.51±0.57**	99.27±0.56
7	Residential	94.49±2.65	88.99±2.14	91.93±1.70	94.17±2.45	92.82±2.26	96.81±0.82	**97.16±0.49**
8	Commercial	89.96±5.96	93.87±1.69	74.70±5.22	92.42±2.23	93.11±1.81	95.54±1.51	**97.25±1.93**
9	Road	87.20±8.23	76.21±3.63	84.90±2.62	84.95±7.56	89.25±3.14	94.52±3.35	**96.83±0.68**
10	Highway	91.43±4.32	90.30±2.60	90.99±2.66	92.230±2.30	94.92±1.53	96.54±2.44	**98.06±0.84**
11	Railway	89.88±6.32	87.40±2.74	88.62±2.48	95.67±2.25	94.59±2.08	95.56±2.08	**97.58±0.82**
12	Parking lot 1	84.62±8.39	78.53±4.69	87.75±2.67	88.95±3.35	94.45±1.56	95.13±2.22	**96.16±0.98**
13	Parking lot 2	80.22±4.65	71.82±4.94	97.54±1.51	92.01±3.46	96.42±2.44	97.02±1.90	**97.25±1.50**
14	Tennis court	98.84±0.68	99.15±0.69	99.45±0.39	**99.71±0.44**	99.39±0.53	99.22±0.99	99.22±0.11
15	Running track	99.32±0.21	98.73±0.62	98.98±0.61	99.165±0.65	98.61±0.89	99.38±0.69	**99.38±0.60**
	OA	92.75±4.92	90.69±1.00	91.79±0.79	94.43±1.17	95.55±0.50	97.06±0.50	**98.01±0.25**
	AA	92.15±4.38	89.92±1.09	91.11±0.86	93.97±1.26	95.19±0.55	96.82±0.54	**97.78±0.30**
	κ	93.54±3.96	90.84±1.92	93.21±0.63	95.07±0.99	96.08±0.51	97.44±0.48	**98.15±0.21**

The value with the highest classification accuracy for each category is bold for the reader's convenience.

TABLE VI
CLASSIFICATION ACCURACY (AVERAGE AND STANDARD DEVIATION) FOR THE TRENTO DATASET

No.	Class name	Methods						
		SVM [44]	EndNet [29]	CNNMRF [30]	TBCNN [26]	Coupled-CNN [23]	FGCN [40]	Proposed
1	Apple trees	64.07±0.30	88.08±4.15	98.78±0.45	97.79±1.83	97.72±1.13	97.39±1.87	**99.23±0.76**
2	Buildings	74.53±0.22	95.30±2.73	87.89±2.59	97.45±1.29	96.15±2.38	**97.97±1.79**	97.31±1.73
3	Ground	93.04±0.75	97.24±1.30	99.30±0.96	97.44±0.66	98.96±1.20	98.54±1.38	**99.55±0.80**
4	Woods	96.97±0.64	99.07±0.54	98.86±0.64	99.97±0.04	99.69±0.48	99.96±0.06	**99.99±0.01**
5	Vineyard	88.93±0.34	85.55±3.11	98.84±0.96	98.09±2.46	98.42±0.92	98.00±1.42	**99.81±0.28**
6	Roads	69.16±0.83	91.90±2.44	87.76±2.94	93.98±1.78	94.72±2.22	96.00±1.89	**96.85±1.90**
	OA	84.64±0.17	91.65±1.24	96.54±0.40	98.12±0.95	98.11±0.45	98.31±0.62	**99.24±0.24**
	AA	79.37±0.13	88.93±1.62	95.38±0.54	97.50±1.26	97.47±0.60	97.75±0.83	**98.98±0.33**
	κ	81.12±0.16	92.80±1.12	95.19±0.51	97.45±0.77	97.61±0.52	97.98±0.62	**98.79±0.35**

The value with the highest classification accuracy for each category is bold for the reader's convenience.

Classification maps of the proposed and comparable methods are shown in Figs. 5–7. It is obvious that the boundary continuity and regional smoothness of our classification map are better than those of the comparative methods. For example, in the bottom-right corner of Fig. 5(j), the boundary of the *Road* is more continuous than that of Fig. 5(i), and the proposed method has more obvious advantages in this respect than the other five methods. As can be seen from the top-left corner of Fig. 5(j), the classification smoothness of the proposed method for the *Mixed Ground Surface* also provides further improvement of smoother classification maps. The same conclusion can be reached for the *Highway* in Fig. 6 and the *Woods* in Fig. 7. In addition, the proposed method can obtain a more distinct classification map for the classes with multiple regions with different sizes. For example, the *Trees* are widely distributed and cluttered in the MUUFL dataset. Compared with other classification maps in Fig. 5, it can be found that the classification region obtained by the proposed method is clearer. Similarly, the *Parking lot 1* in the Houston dataset and the *Roads* in the Trento dataset also contain multiple regions with different sizes. It can be seen from Figs. 6 and 7 that our method can obtain a classification map closer to the ground truth for these classes. In addition, Figs. 5–7 show that the proposed method can yield classification maps with fewer erroneous outliers, such as the *Dirt and sand* in the top-right part and the bottom-right part of Fig. 5(j), the *Highway* in the upper half of Fig. 6(j), and the *Woods* in Fig. 7(j). In general, the classification maps obtained by the proposed method are smoother, with clearer boundaries and fewer erroneous outliers. No matter the Trento dataset with evenly terrain distribution or Houston and MUUFL datasets with complex terrain distribution, the proposed method obtains more accurate classification maps.

TABLE VII
ABLATION ANALYSIS OF THE PROPOSED METHOD IN TERMS OF OA (%) AND KAPPA (%)

Network	SpeF	MSSpaF	MDSpaF	Houston		MUUFL		Trento	
				OA	κ	OA	κ	OA	κ
1	✓			54.52	50.89	66.16	57.68	72.28	63.84
2		✓		95.69	95.34	87.07	83.75	98.04	97.37
3			✓	94.95	95.43	86.16	81.99	98.26	97.67
4	✓	✓		95.83	95.47	87.96	84.38	98.43	97.92
5	✓		✓	96.60	96.44	86.49	82.41	98.83	98.44
6		✓	✓	96.94	96.69	87.60	83.97	98.98	98.63
7	✓	✓	✓	98.01	97.78	89.72	86.62	99.24	98.98

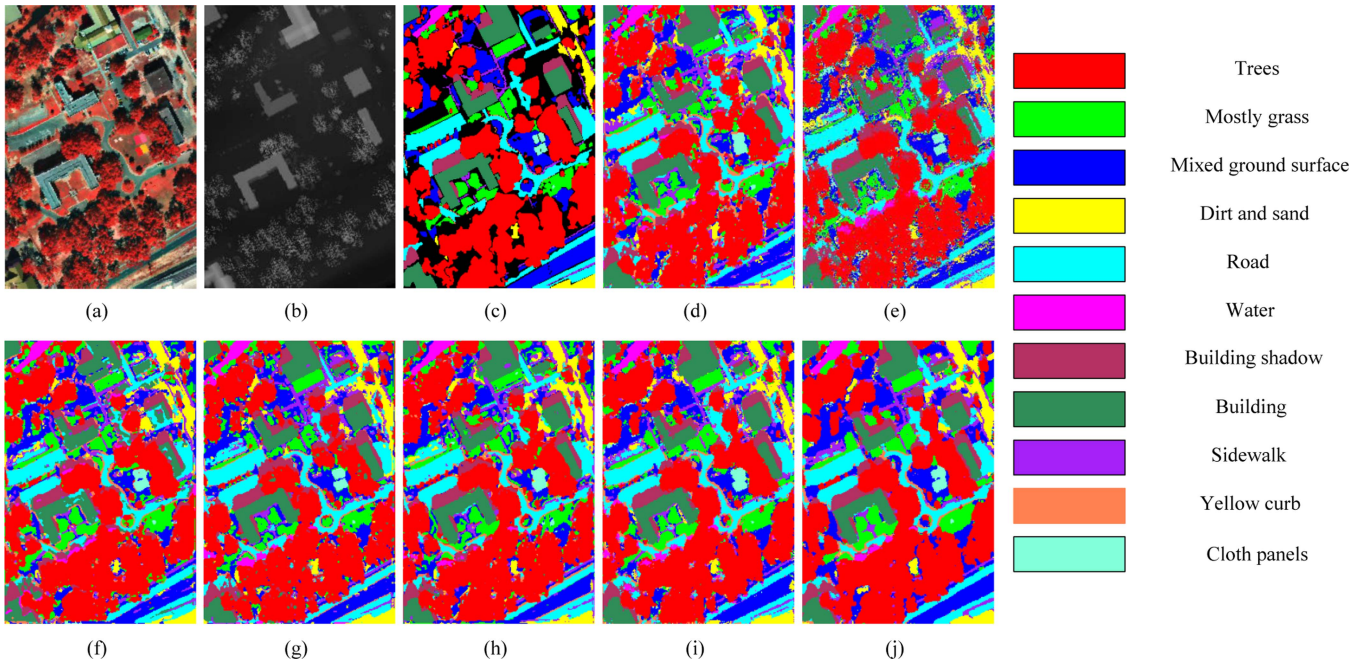

Fig. 5. Classification maps for the MUUFL dataset. (a) Pseudocolor image for HSI. (b) LiDAR-based DSM. (c) Ground truth map. (d) SVM (84.44%). (e) EndNet (81.13%). (f) CNNMRF (81.82%). (g) TBCNN (83.18%). (h) Coupled-CNN (81.88%). (i) FGCN (87.62%). (j) Proposed method (89.72%).

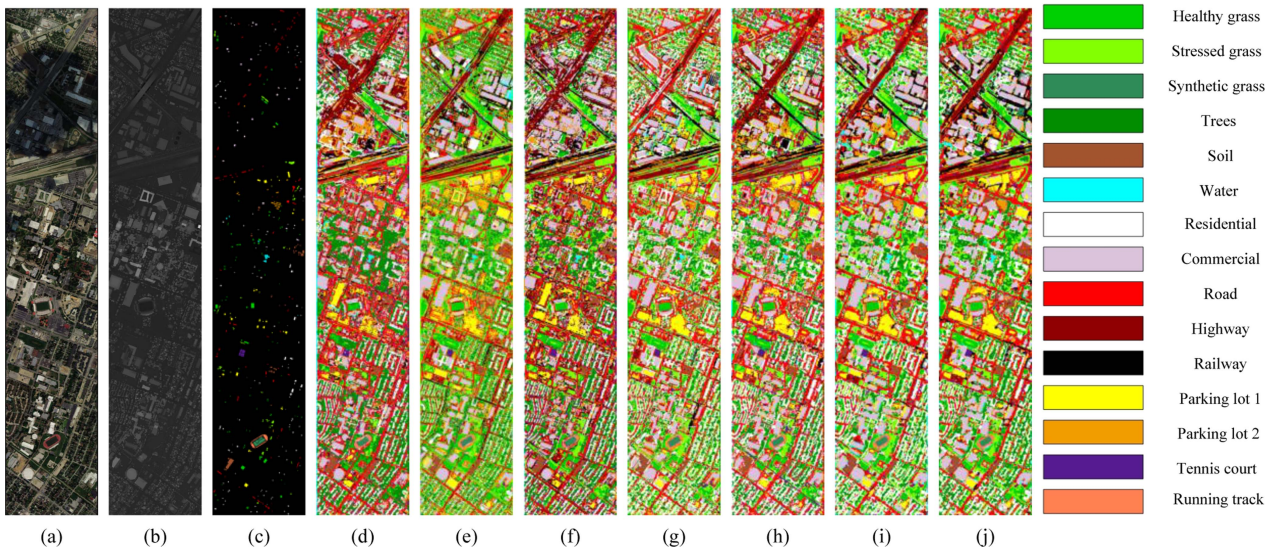

Fig. 6. Classification maps for the Houston dataset. (a) Pseudocolor image for HSI. (b) LiDAR-based DSM. (c) Ground truth map. (d) SVM (92.75%). (e) EndNet (90.69%). (f) CNNMRF (91.79%). (g) TBCNN (94.43%). (h) Coupled-CNN (95.55%). (i) FGCN(97.06%). (j) Proposed method (98.01%).

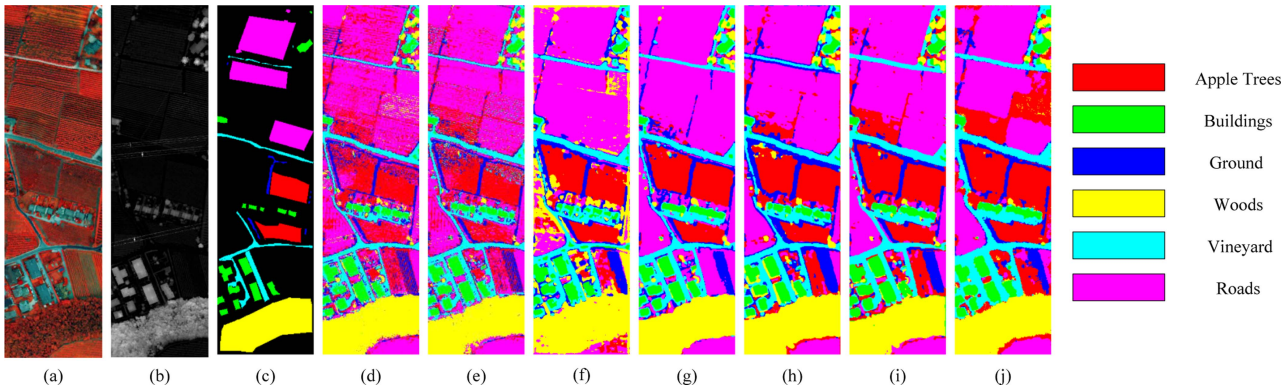

Fig. 7. Classification maps for the Trento dataset. (a) Pseudocolor image for HSI. (b) LiDAR-based DSM. (c) Ground truth map. (d) SVM (84.64%). (e) EndNet (91.65%). (f) CNNMRF (96.54%). (g) TBCNN (98.12%). (h) Coupled-CNN (98.11%). (i) FGCN (98.31%). (j) Proposed method (99.24%).

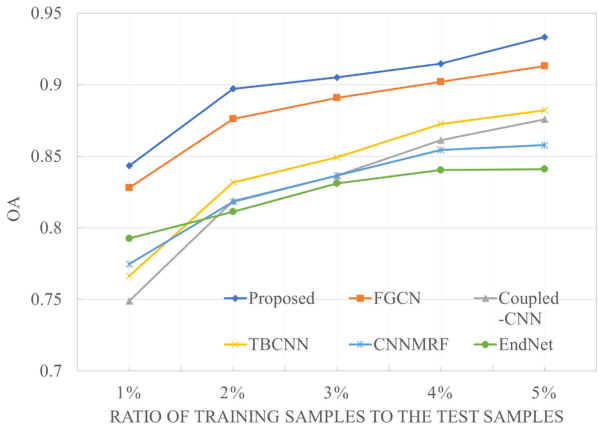

Fig. 8. Classification performance for varying number of training samples for the MUUFL dataset.

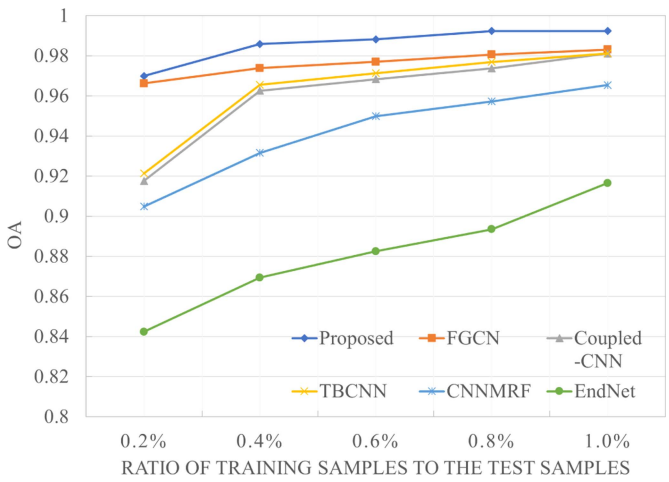

Fig. 9. Classification performance for varying number of training samples for the Trento dataset.

Figs. 8 and 9 depict the classification performance of the various algorithms as the training ratio varies. It can be seen that classification performance for all methods under comparison degrades with the decreasing number of training samples, which is as expected. However, compared to the other techniques, the proposed approach always achieves higher classification accuracy even with an exceedingly small training ratio (e.g., 1%).

E. Ablation Analysis

We conduct a battery of ablation experiments to verify various factors of the proposed method design. Specifically, we look at the impact of using the SpeF, the MSSpaF, and the MDSpaF modules. As discussed in Section III-D, the SpeF module employs a spectral channel attention mechanism to assign weights for extracting SpeFs as complementary information. MSSpaF module extracts MSSpaFs and fuses them through a residual connection while MDSpaF module extracts multidirection and multifrequency information and through cross-layer connection and multiscale feature fusion strategy. We design six additional networks, as given in Table VII, to gauge the effects of these three modules within the proposed approach. The OA and Kappa are shown for the three mentioned datasets. As can be seen from the results in this table, the classification results for the three datasets are worst only considering SpeFs. This is because the SpeF module is designed to complement the multiscale feature module and multidirectional feature module. Using this module alone cannot fully reflect the network design, but it is expected to enhance the effect of combining SpeFs with spatial features. When MSSpaF or MDSpaF module is used, the classification performance will improve, because these two modules are based on MSSpaF extraction, which can expand convolutional receptive fields and extract much more fine-grained features. When the SpeF module is combined with the MSSpaF or MDSpaF module, the classification performance will also be improved to some extent. Without considering SpeFs, the joint use of MSSpaF and MDSpaF modules for spatial feature extraction can yield satisfactory classification results. Furthermore, the combined application of the proposed three modules can improve classification performance.

V. Conclusion

In this article, a multiscale and multidirection feature extraction network is proposed for HSI and LiDAR data classification. First, the MSSpaF module is designed to extract multiscale spatial information of HSI and LiDAR data, expanding the receptive

fields of convolution in a more fine-grained manner than that of other architectures in the literature. In addition, multiscale and multidirection features are extracted using the proposed MDSpaF module, and then, the receptive fields of fractional Gabor convolution are enlarged through hierarchical jumping connections. The SpeF module extracts SpeFs of HSI data as complementary information following assigning spatial and spectral information weights, and finally, multisource feature maps are fused and fed into a classifier. Ablation experiments proved the ability and effectiveness of the proposed network to extract multiscale and multidirection features from HSI and LiDAR datasets while a battery of experimental results compared to other state-of-the-art networks demonstrated that the proposed network achieves outstanding classification performance even with limited training data.

REFERENCES

[1] D. Hong, N. Yokoya, J. Chanussot, and X. Zhu, "CoSpace: Common subspace learning from hyperspectral-multispectral correspondences," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4349–4359, Jul. 2019.

[2] F. Salem, M. Kafatos, T. El-Ghazawi, R. Gomez, and R. Yang, "Hyperspectral image assessment of oil-contaminated wetland," *Int. J. Remote Sens.*, vol. 26, no. 4, pp. 811–821, 2005.

[3] L. M. Dale et al., "Hyperspectral imaging applications in agriculture and agro-food product quality and safety control: A review," *Appl. Spectrosc. Rev.*, vol. 48, 2013, Art. no. 142.

[4] A. Song and Y. Kim, "Deep learning-based hyperspectral image classification with application to environmental geographic information systems," *Korean J. Remote Sens.*, vol. 33, no. 6_2, pp. 1061–1073, 2017.

[5] X. Briottet et al., "Military applications of hyperspectral imagery," *Proc. SPIE*, vol. 6239, 2006, Art. no. 62390B.

[6] C. Debes et al., "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.

[7] P. Ghamisi et al., "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 1, pp. 6–39, Mar. 2019.

[8] G. Yang, Y. He, X. Feng, X. LI, J. Zhang, and Z. Yu, "Methods and new research progress of remote sensing monitoring of crop disease and pest stress using unmanned aerial vehicle," *Smart Agriculture*, vol. 4, no. 1, 2022, Art. no. 1.

[9] Z. Wang, J. Wang, Y. Wu, J. Xu, and X. Zhang, "UNFusion: A unified multi-scale densely connected network for infrared and visible image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3360–3374, Jun. 2021.

[10] Y. Yang, J. Liu, S. Huang, W. Wan, W. Wen, and J. Guan, "Infrared and visible image fusion via texture conditional generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 12, pp. 4771–4783, Dec. 2021.

[11] W. Liao, R. Bellens, A. Pižurica, S. Gautama, and W. Philips, "Combining feature fusion and decision fusion for classification of hyperspectral and LiDAR data," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2014, pp. 1241–1244.

[12] P. Ghamisi, J. A. Benediktsson, and S. Phinn, "Land-cover classification using both hyperspectral and LiDAR data," *Int. J. Image Data Fusion*, vol. 6, no. 3, pp. 189–215, 2015.

[13] P. Ghamisi, B. Höfle, and X. X. Zhu, "Hyperspectral and LiDAR data fusion using extinction profiles and deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 3011–3024, Jun. 2017.

[14] B. Rasti, P. Ghamisi, J. Plaza, and A. Plaza, "Fusion of hyperspectral and LiDAR data using sparse and low-rank component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6354–6365, Nov. 2017.

[15] X. Kang, S. Li, L. Fang, M. Li, and J. A. Benediktsson, "Extended random walker-based classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 144–153, Jan. 2015.

[16] J. Xia, W. Liao, and P. Du, "Hyperspectral and LiDAR classification with semisupervised graph fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 666–670, Apr. 2019.

[17] S. Jia et al., "Multiple feature-based superpixel-level decision fusion for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1437–1452, Feb. 2020.

[18] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[19] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.

[20] X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[22] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, and N. Carvalhais, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

[23] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and LiDAR data using coupled CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4939–4950, Jul. 2020.

[24] Y. Zhu, W. Li, M. Zhang, Y. Pang, R. Tao, and Q. Du, "Joint feature extraction for multi-source data using similar double-concentrated network," *Neurocomputing*, vol. 450, pp. 70–79, 2021.

[25] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, "Information fusion for classification of hyperspectral and LiDAR data using IP-CNN," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5506812.

[26] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.

[27] T. Zhang, S. Xiao, W. Dong, J. Qu, and Y. Yang, "A mutual guidance attention-based multi-level fusion network for hyperspectral and LiDAR classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 5509105.

[28] M. Zhang, W. Li, Q. Du, L. Gao, and B. Zhang, "Feature extraction for classification of hyperspectral and LiDAR data using patch-to-patch CNN," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 100–111, Jan. 2020.

[29] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder-decoder networks for classification of hyperspectral and LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 5500205.

[30] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu, and J. Paisley, "Hyperspectral image classification with Markov random fields and a convolutional neural network," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2354–2367, May 2018.

[31] H. Li, W. Hu, W. Li, J. Li, Q. Du, and A. Plaza, "A^3CLNN: Spatial, spectral and multiscale attention convLSTM neural network for multisource remote sensing data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 747–761, Feb. 2022.

[32] M. Wang, F. Gao, J. Dong, H.-C. Li, and Q. Du, "Nearest neighbor-based contrastive learning for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501816.

[33] H. Gao, H. Feng, Y. Zhang, S. Xu, and B. Zhang, "AMSSE-Net: Adaptive multiscale spatial–spectral enhancement network for classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5531317.

[34] L. Zhou, J. Geng, and W. Jiang, "Joint classification of hyperspectral and LiDAR data based on position-channel cooperative attention network," *Remote Sens.*, vol. 14, no. 14, 2022, Art. no. 3247.

[35] W. Liu, F. Gao, and J. Dong, "Disentangled non-local network for hyperspectral and LiDAR data classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2397–2400.

[36] Z. Xue, X. Yu, X. Tan, B. Liu, A. Yu, and X. Wei, "Multiscale deep learning network with self-calibrated convolution for hyperspectral and LiDAR data collaborative classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5514116.

[37] F. Xu, G. Zhang, C. Song, H. Wang, and S. Mei, "Multiscale and cross-level attention learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501615.

[38] S. Luan, C. Chen, B. Zhang, J. Han, and J. Liu, "Gabor convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4357–4366, Sep. 2018.

[39] Y. Zhang et al., "Fractional Gabor transform," *Opt. Lett.*, vol. 22, no. 21, pp. 1583–1585, 1997.

[40] X. Zhao, R. Tao, W. Li, W. Philips, and W. Liao, "Fractional Gabor convolutional network for multisource remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5503818.

[41] S. Gao, M. Cheng, K. Zhao, X. Zhang, M. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.
[42] P. Gader, A. Zare, R. Close, J. Aitken, and G. Tuell, "MUUFL Gulfport hyperspectral and LiDAR airborne data set," Univ. Florida, Gainesville, FL, USA, Tech. Rep. REP-2013-570, 2013.
[43] B. Rasti, P. Ghamisi, and R. Gloaguen, "Hyperspectral and LiDAR fusion using extinction profiles and total variation component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3997–4007, Jul. 2017.
[44] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 1998.

Yi Liu received the B.S. degree in automation from Chang'an University, Xi'an, China, in 2020, where she is currently working toward the M.S. degree in control science and engineering with the School of Electronics and Control Engineering.

Her research interests include hyperspectral image analysis and multisource remote sensing.

Zhen Ye received the B.S., M.S., and Ph.D. degrees in information and communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2007, 2010, and 2015, respectively.

She spent one year as an Exchange Student with Mississippi State University, Mississippi State, MS, USA. She is currently an Associate Professor with the School of Electronics and Control Engineering, Chang'an University, Xi'an. Her research interests include remote-sensing image processing, pattern recognition, and machine learning.

Yongqiang Xi received the B.S. degree in automation from Chang'an University, Xi'an, China, in 2020, where he is currently working toward the M.S. degree in control science and engineering with the School of Electronics and Control Engineering.

His current research interest includes multisource remote-sensing image analysis.

Huan Liu received the B.S. degree in optical information science and technology from China University of Mining and Technology, Xuzhou, China, in 2013, and the M.S. degree in physics and Ph.D. degree in information and communication engineering from Beijing Institute of Technology, Beijing, China, in 2016 and 2022, respectively.

His research interests include image processing, image classification, and transfer learning.

Wei Li (Senior Member, IEEE) received the B.E. degree in telecommunications engineering from Xidian University, Xi'an, China, in 2007, the M.S. degree in information science and technology from Sun Yat-Sen University, Guangzhou, China, in 2009, and the Ph.D. degree in electrical and computer engineering from Mississippi State University, Starkville, MS, USA, in 2012.

He spent one year as a Postdoctoral Researcher with the University of California, Davis, CA, USA. He is currently a Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. He has authored or coauthored more than 160 peer-reviewed articles and 100 conference papers. His research interests include hyperspectral image analysis, pattern recognition, and target detection.

Dr. Li is currently an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS). He was an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS) and the IEEE SIGNAL PROCESSING LETTERS. He is a recipient of the JSTARS Best Reviewer in 2016 and TGRS Best Reviewer Award in 2020 from the IEEE Geoscience and Remote Sensing Society and the Outstanding Paper Award at IEEE International Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (Whispers) in 2019.

Lin Bai (Member, IEEE) received the B.S. degree in electronic information science and technology from Northwest University, Xi'an, China, in 2003, the M.S. degree in electronic science and technology from Xidian University, Xi'an, in 2006, and the Ph.D. degree in signal and information processing from Northwestern Polytechnical University in 2011.

He is currently an Associate Professor with the School of Electronic and Control Engineering, Chang'an University, Xi'an. His research interests include machine learning and remote-sensing image processing.