

Morphological Convolution and Attention Calibration Network for Hyperspectral and LiDAR Data Classification

Zhongwei Li [✉], Hao Sui [✉], Cai Luo [✉], *Senior Member, IEEE*,
and Fangming Guo [✉], *Graduate Student Member, IEEE*

Abstract—Reasonable fusion of multimodal data can increase the accuracy of remote sensing classification. In this article, an effective morphological convolution and attention calibration network is proposed for the joint classification of the hyperspectral image (HSI) and light detection and ranging (LiDAR). First, we devise a morphological convolution block, which combines the dilation and erosion operations in morphology with convolution to better capture the feature from the HSI and LiDAR. Next, we designed a dual attention module that uses self-attention to calibrate features and cross attention to combine multisource complementary information, respectively. Finally, considering the features of semantic inconsistency and different scales, the adaptive feature fusion module is introduced to dynamically fuse multimodal features. To verify the progressiveness of the proposed network, we experiment on three common datasets and one self-made dataset. The result shows that our network performs better than the state-of-the-art models.

Index Terms—Attention mechanism, hyperspectral image (HSI), joint classification, light detection and ranging (LiDAR), morphological operations.

I. INTRODUCTION

BENEFITING from the vigorous development of sensor technology, the acquisition methods of remote sensing images are becoming increasingly diverse [1], [2]. Various sensors provide distinct ground feature information, which is widely used in surface observation, forest monitoring, environmental investigation, and other aspects [3], [4], [5]. A hyperspectral image (HSI) contains both rich spectral and spatial information,

and spectral information is encoded in dozens of continuous wavebands [6], [7], profoundly improving the ability to distinguish objects. However, an HSI has difficulty distinguishing objects at different altitudes with similar spectral responses [8], [9]. For example, roofs and roads are both made of concrete, which have similar spectral responses, but they belong to different categories. Instead, light detection and ranging (LiDAR) includes elevation information of the scene [10], making up for the shortage of HSI data. Consequently, the joint classification of the HSI and LiDAR has broad prospects and has received widespread attention [11].

Recently, some strategies of fusing the HSI and LiDAR images have been developed, among which feature-level fusion methods are widely used. Pedergrana et al. [12] applied morphological extended attribute profiles (EAPs) to obtain distinguishable features, which were then overlaid for classification. However, the splicing of features can lead to the Hughes phenomenon, especially when the trainable samples are limited. Therefore, principal component analysis (PCA) was adopted to reduce feature dimensions [13]. However, PCA has limitations in extracting local HSI structural information. To overcome this problem, Uddin et al. [14] applied FPCA on the highly correlated or spectrally separated bands' segments of the HSI, and then, proposed segmented FPCA (SFPCA) and spectrally SFPCA (SSFPCA), which further ameliorated feature extraction. Shemul et al. [15] proposed the segmented-sparse-PCA (SSPCA), which divided the entire dataset into multiple highly correlated spectral band subsets, and then, applied sparse-PCA to each subset, achieving satisfactory results. Rasti et al. [16] proposed a feature fusion method based on orthogonal total variation analysis, which maps the fused features to lower dimensions. It can improve the piece-wise smoothness while retaining the spatial structure, and finally, generate accurate feature maps. Another prevalent method is fusing features in the decision-making stage. A support vector machine (SVM) [17] with nonlinear kernel functions has been widely used. Xia et al. [18] assembled rotation forests and Markov random fields to obtain a higher classification accuracy, in which four different feature extraction approaches are applied. Peng et al. [19] proposed an SVM based on a region kernel that devises three mixed kernels to estimate the similarity between spatial and spectral features. Subsequently, combining the advantages

Manuscript received 29 March 2023; revised 3 May 2023 and 24 May 2023; accepted 7 June 2023. Date of publication 9 June 2023; date of current version 3 July 2023. This work was supported by the National Natural Science Foundation of China under Grant 62071491; in part by the Joint Funds of the National Natural Science Foundation of China under Grant U1906217 and the Innovation Fund Project for Graduate Students of China University of Petroleum (East China), and in part by the Fundamental Research Funds for the Central Universities under Grant 22CX04009A. (*Corresponding author: Cai Luo.*)

Zhongwei Li, Cai Luo, and Fangming Guo are with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China (e-mail: lizhongwei@upc.edu.cn; luo_cai@upc.edu.cn; guofangming@s.upc.edu.cn).

Hao Sui is with the College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China (e-mail: stu_suihao@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3284655

of the SVM and multiple classifier systems, Xia et al. [20] designed a novel rotation-based SVM to upgrade the performance on classified tasks. Nevertheless, these classical methods rely on the selection of parameters. Whenever faced with different classification tasks, the parameters need to be readjusted, which is undoubtedly time consuming.

Methods based on deep learning have achieved satisfactory achievements in the field of remote sensing in recent years [21], [22]. The convolutional neural network (CNN), standing out with its robust ability to process high-dimensional data, is especially suitable for solving the classification task of the HSI and LiDAR. Most existing methods based on the CNN use a dual branch structure to obtain heterogeneous information from multisource data. This dual-branch network structure has a preferable feature extraction ability that greatly improved the classification accuracy. Xu et al. [23] pioneered a dual-tunnel CNN structure to extract features in different dimensional spaces, and designed a CNN with cascaded blocks to extract features from LiDAR images. The experiment proved that the proposed dual-tunnel CNN has excellent classification performance. Afterwards, some unsupervised and semisupervised methods also achieved satisfactory results. Guan et al. [24] proposed an unsupervised cross-domain contrastive learning framework, which constructs signals in the spatial and spectral domains, respectively, and applies comparative learning to extract shared information between the two signals for the HSI representation. Duan et al. [25] developed a new semisupervised algorithm called a geodesic-based sparse manifold hypergraph, which built the manifold neighborhood of each sample and designed a pair of semisupervised hypergraphs to handle sparse correlations between samples.

Considering that there are pixels in the input patch that have different classification labels from the center pixel, which can affect classification accuracy, some researchers have applied attention mechanisms to remote sensing classification. The core of the attention mechanism is to assist the network to take more notice of the significant regions, which is generally reflected in the form of weight. It further helps the model improve the efficiency and accuracy of task processing. FusAtNet [26] adopts the dual attention mechanism, which uses self-attention to emphasize its own features, and takes cross attention to obtain the weight map from the LiDAR to optimize the spatial features of the HSI. MSNetSC [27] first proposed self-calibrated convolutional blocks, and combined them with multiscale structures to obtain optimized multiscale features, then applied self-attention to enhance features. Li et al. [28] developed a dual channel A3CLNN, which combines the multiscale structure, ConvLSTM, and self-attention to better describe features, and trained the network in stages to obtain classification results. EMFNet [29] adopted a feature tuning module to achieve cross optimization between multisource data, and designed a novel feature fusion module to assign appropriate weights to the features to be fused.

In fact, morphological methods can effectively complete the task of spatial feature extraction. Among them, the most widely used are morphological profiles (MPs) [30], attribute profiles (APs) [12], [31], and extension profiles (EPs) [3], [32]. Ghamisi

et al. [33] combined EPs with deep learning to obtain more accurate classification results. Hong et al. [34] using invariant attribute profiles (IAPs) to locally extract the spatial invariant features. Roy et al. [35] developed a novel local morphology pattern (LMP), which used opening and closing operations scales to accurately obtain location and contour information of the objects. Mellouli et al. [36] combined the CNN with morphological feature extraction to improve the image quality and achieved good results in image classification tasks. Franchi et al. [37] proposed a method of using morphological nonlinear operators in a deep learning framework. It has been proven in multiple applications that nonlinear morphological operations and convolutional layers are complementary. Roy et al. [38] constructed spatial morphological blocks using morphological expansion and erosion layers, and combined them with the CNN, which achieved good classification results.

However, there are still some problems with HSI-LiDAR classification. First, the current classification methods mostly use traditional convolution to extract spatial features, which has limitations in describing the boundaries and shapes of objects. Moreover, self-attention mechanisms are often utilized to optimize features, but there is no interaction between features, which further limits their semantic relevance. More importantly, the simple feature concatenation will ignore the correlation and complementarity between multimodal data and increase the feature dimension, which may result in dimension disaster.

In order to settle these issues, we devise multiple approaches to effectively utilize multimodal data for classification. Foremost, the dilation and erosion operations in morphology are combined with traditional convolution to better extract HSI spatial information and LiDAR elevation information, and the dual attention mechanism (self-attention and cross-attention) is applied to calibrate features and increase the cross guidance between multimodal features. Then, the nonlinear attention feature fusion module (AFF) [39] is introduced to dynamically fuse multimodal spatial features in the way of context-scale awareness. Experiments were carried out on four real HSI and LiDAR datasets, and results indicate that our method is more advanced in analyzing multimodal data.

In short, the main contributions to this article can be summarized into four aspects.

- 1) To extract distinguishable spatial features, a morphological feature extraction block (MorFEB) is constructed, which combines the expansion and erosion operations in morphology with convolution to better capture the boundary shape information of arbitrary objects in complex regions and extract robust and differentiated spatial information.
- 2) A spectral calibration block (SpeCB) is built to correct spectral features. Specifically, the central pixel is extracted from the HSI patch, and then, the channel weight is calculated to calibrate the spectral features of the entire patch.
- 3) A position attention module (PAM) similar to the U-Net structure is projected to provide supplementary information guidance between multimodal data, which can make

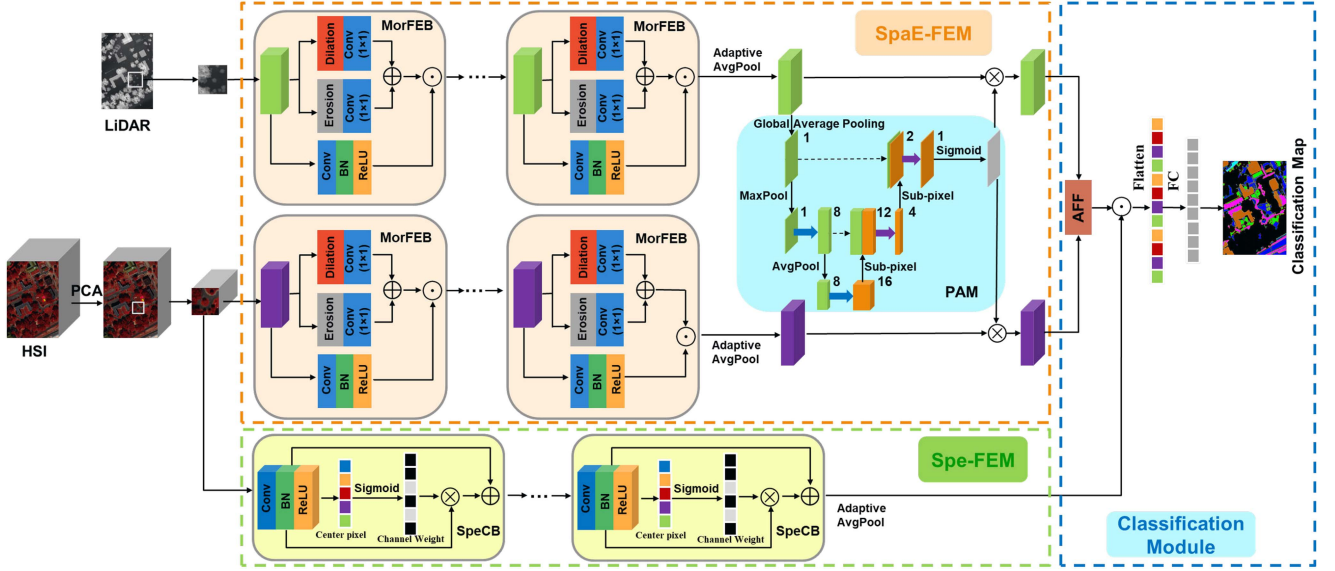


Fig. 1. Overall framework of the proposed network.

use of the advantages of LiDAR to make up for the disadvantages of the HSI and realize supplementary guidance between multimodal data.

- 4) Considering the inconsistency of semantics and scales, the AFF module is introduced into the proposed network, which fuses multimodal features dynamically and adaptively, and does not generate additional dimensions, and reduces information redundancy.

The rest of this article is organized as follows. Section II describes the details of the proposed network. The experimental results and analysis are shown in Section III. Finally, Section IV concludes this article.

II. METHODOLOGY

A. Architecture Overview

An HSI can capture subtle spectral differences based on its rich spectral characteristics to further distinguish various materials, while LiDAR can accurately model the surrounding environment in 3-D and provide elevation information that is missing from HSI. To properly solve the classification problems of HSI and LiDAR, it is necessary to correctly fuse the complementary features of multimodal data.

The proposed HSI and LiDAR joint classification framework is shown in Fig. 1. An SpaE-FEM is designed to extract significant complementary information from HSI and LiDAR. The MorFEB is the core component, which combines dilation and erosion operations in morphology with convolution to better capture robust and differentiated spatial/elevation information of arbitrary objects in complex regions, and then we devise a PAM to optimize the features and utilize the advantages of LiDAR to compensate for the shortcomings of the HSI. In addition, in order to obtain more accurate spectral features from the HSI, the Spe-FEM is proposed, in which the SpeCB is used to calibrate the weight between channels of the HSI.

B. Spatial-Elevation Feature Extraction Module

Being universally known, morphological operations are powerful nonlinear transformations that can capture the size, shape, and structure information of objects in more detail. Here, an MorFEB based on dilation and erosion operations is proposed, which takes structural elements of the size $(p \times p)$ as the core for operations. Fig. 2 describes the detailed mathematical process of dilation and erosion operations.

For example, using a LiDAR patch and an extended structural element SE_d for operation magnifies the object and increases the number of available feature pixels. In contrast, erosion is the process of eliminating all boundary points of significant objects, eliminating individual abnormal pixel points, and expanding the gaps between objects that allow them to be separated between adjacent areas. And convolution can preserve the original shape, boundaries, and other information of the object, which precisely serves as a supplement to dilation and erosion operations, making the extracted features more balanced. Given the input, LiDAR patch $X_L \in R^{M \times N}$ with spatial size $(M \times N)$. The dilation (\oplus) and erosion (\ominus) operations can be defined as an operation on the feature map centered on the spatial position (i, j) as

$$(X_L \oplus SE_d)(x, y) = \max_{(i, j) \in U} (X_L(x + i, y + j) + SE_d(i, j)) \quad (1)$$

$$(X_L \ominus SE_e)(x, y) = \min_{(i, j) \in U} (X_L(x + i, y + j) - SE_e(i, j)) \quad (2)$$

where $U = \{(i, j) \mid i \in \{1, 2, 3, \dots, s\}; j \in \{1, 2, 3, \dots, s\}\}$, and SE_d and SE_e represent the structural elements of dilation and erosion operations, respectively. Filling the image before dilation and erosion operations to ensure the same size of input and output. The feature F_{Morph} obtained by MorphFEB is derived

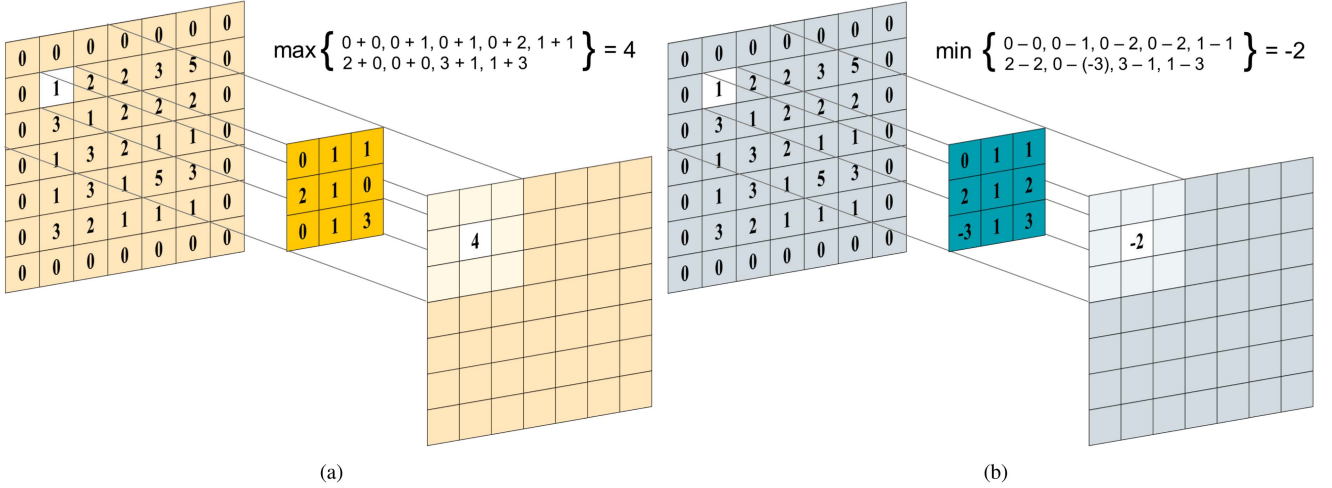


Fig. 2. Graphical visualization of the dilation and erosion operations where an input image patch of size $(7 \times 7 \times 1)$ is dilated and eroded with an SE of size $(3 \times 3 \times 1)$ and the obtained output size keep the same with a padding mechanism. (a) Dilation operation. (b) Erosion operation.

from the following formulas:

$$\mathcal{F}_d = \text{Conv}((X_L \oplus \text{SE}_d), k = 1) \quad (3)$$

$$\mathcal{F}_e = \text{Conv}((X_L \ominus \text{SE}_e), k = 1) \quad (4)$$

$$\mathcal{F}_{\text{Morph}} = (\mathcal{F}_d + \mathcal{F}_e) \odot \text{Conv}((X_L), k = 3) \quad (5)$$

where Conv represents convolution and k is the size of a convolution kernel, which projects morphological features into the same subspace as convolution to facilitate subsequent combination and \odot represents feature concatenation. Similarly, the proposed morphological operation can be applied to processing HSI channel by the channel.

For extracted feature map, the information contained in different positions has different contributions to the recognition of the central pixel. Benefiting from the progress of deep learning technology, the attention mechanism has become an effective tool for remote sensing classification tasks [40], [41], [42]. Focusing on significant regions and suppressing irrelevant parts is more conducive to classification. Therefore, we propose a PAM to assign weights to different spatial positions. The PAM is similar to U-Net in structure. The position weight map generated by the PAM contains both shallow and deep information, which can more accurately find areas of interest.

In the down-sampling stage, given the input feature vector $F_L \in R^{C \times H \times W}$, where C , H , and W are the number of channels, height, and width of the feature, respectively. First, the global average pooling is used to aggregate features. Afterwards, the Maxpool is used to sample down to retain the most important information in the neighborhood, and then the feature F_L^1 is further extracted by convolution. After that, the Avgpool is applied to sample down, and then the generated results are sent to the convolution layer to obtain the output F_L^E of the encoder. The aforementioned process can be expressed as

$$F_L^1 = \text{Conv}(\text{MP}(\text{GAP}(F_L)), k = 3) \quad (6)$$

$$F_L^E = \text{Conv}(\text{AP}(F_L^1), k = 3) \quad (7)$$

where GAP, MP, and AP represent global average pooling, Maxpool, and Avepool, respectively.

In the upsampling phase, sub-pixel convolution is applied to upsample F_L^E , and the obtained result is connected with F_L^1 , and then, feed them to the convolution layer to obtain the feature F_L^2 . Repeating this process and applying a sigmoid function to generate position weight map F_L^{att} , the aforementioned process can be expressed as

$$F_L^2 = \text{Conv}([F_L^1, \text{SPC}(F_L^E)], k = 1) \quad (8)$$

$$F_L^{\text{att}} = \text{Sigmoid}(\text{Conv}([\text{GAP}(F_L), \text{SPC}(F_L^2)], k = 1)) \quad (9)$$

where SPC is subpixel convolution. Then, the weight $F_{L,i}^{\text{att}}$, is applied to optimize the LiDAR feature map F_L and HSI feature vector F_H , which provides supplementary information guidance from LiDAR to the HSI to improve classification results.

C. Spectrum Feature Extraction Module

The HSI has abundant spectral information that is encoded in dozens of continuous wavebands and numerous imaging channels, which has strong target recognition capability. However, the importance of spectral information contained in each channel varies, attention mechanisms need to be used to calibrate interchannel weights to obtain more effective spectral information. Hence, we design a SpeCB, which takes the spectral information of the central pixel extracted from the HSI data as the benchmark to establish the spectral weight of the central pixel. The aforementioned process can be expressed as

$$F_H^{\text{spe}} = \text{ReLU}(\text{BN}(\text{Conv3D}(X_H))) \quad (10)$$

$$F_H^{cb} = \text{Sigmoid}(\text{Center}(F_H^{\text{spe}})) \times F_H^{\text{spe}} + F_H^{\text{spe}} \quad (11)$$

where Center represents the center pixel extracted from the extracted spectral feature F_H^{spe} . The sigmoid function is applied to normalize channels to the range $[0, 1]$ to get the weight of each

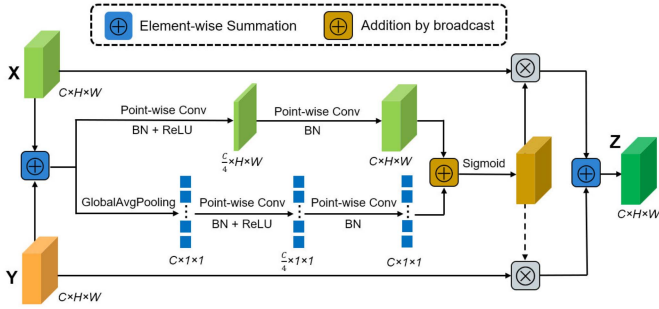


Fig. 3. Illustration of the AFF module.

channel. Finally, the weight is used to calibrate HSI patches, and the residual is introduced to complete the optimization.

Here, to extract pure spectral features, the convolution kernel size of 3D-CNN is $(7 \times 1 \times 1)$, which signifies that the convolution is only performed in the spectral dimension without leading to redundant spatial information.

D. Classification Module

Since there is overlapping information between the spatial features and the elevation features, and the spectral features are not related to them, we should take into account the correlation and difference between features when fusing features, and select the appropriate fusion strategy. Specifically, the AFF method is chosen to fuse spatial feature and elevation feature, which focuses on both global and local information, assigning an appropriate proportion to them and constantly changing the proportion during training to obtain more effective fusion results without increasing the dimension. The overall structure of the AFF is shown in Fig. 3.

Given two features to be fused $X, Y \in R^{C \times H \times W}$, the initial integration of the two features is realized by addition, which is recorded as $F = X + Y$. The AFF emphasizes both global information and local information, and selects point-by-point convolution (PWConv) to obtain local information and aggregate the inter channel relationships of each feature pixel. In the AFF, the local channel context $L_c \in R^{C \times H \times W}$ can be obtained by the following operation.

$$L_c(F) = \beta(\text{PWConv}_2(\delta(\beta(\text{PWConv}_1(F))))) \quad (12)$$

Similarly, the global channel context $G_c \in R^{C \times 1 \times 1}$ is calculated by the following formula:

$$G_c(F) = \beta(\text{PWConv}_4(\delta(\beta(\text{PWConv}_3(\text{GAP}(F))))) \quad (13)$$

where β and δ are the BN and ReLU activation functions, respectively. Here, $L_c(X + Y)$ does not change in size and retains lower level features rich in spatial information with higher resolution. After that, the attention weight is calculated by the following formula:

$$W = \text{Sigmoid}(L_c(X + Y) \oplus G_c(X + Y)) \quad (14)$$

Finally, the output of the AFF can be expressed as

$$Z = W \times X + (1 - W) \times Y \quad (15)$$

TABLE I
NUMBERS OF TRAINING AND TESTING SAMPLES FOR THE HOUSTON DATASET

No.	color	Class	Training	Test
1		Healthy grass	125	1 126
2		Stressed grass	125	1 129
3		Synthetic grass	70	627
4		Trees	124	1 120
5		Soil	124	1 118
6		Water	33	292
7		Residential	127	1 141
8		Commercial	124	1 120
9		Road	125	1 127
10		Highway	123	1 104
11		Railway	123	1 112
12		Parking lot 1	123	1 110
13		Parking lot 2	47	422
14		Tennis court	43	385
15		Running track	66	594
Total			1502	13527

where $Z \in R^{C \times H \times W}$ is the final fused feature map. In this fusion module, the fused feature Z integrates low resolution global features and high resolution local features. And during the training process, W will be continuously adjusted to obtain the most effective fusion features, further improving the classification effect.

Similarly, spatial features and spectral features are not related to each other, so we treat them as equally important, and then, use concatenation to connect them. So far, we have fused all features together, taking into account correlation and differences, and giving attention to both global and local features. Finally, the fused feature is fed into the fully connected layer to obtain the ultimate classification result.

III. RESULTS AND DISCUSSION

A. Dataset Description

To evaluate the effectiveness and progressiveness of the proposed method, three common datasets and a self-made multi-modal dataset are used for comparative classification experiments. The basic situation of the four datasets is described as follows.

1) *Houston Dataset*: This dataset was collected in June 2012 on the University of Houston campus and adjacent regions, with a spatial resolution of 2.5 m and a size of 349×1905 pixels. The band numbers for hyperspectral and lidar are 144 and 1, respectively. There are 15 available classes. We randomly selected 10% of the pixels from the dataset as a training set. Details are

TABLE II
NUMBERS OF TRAINING AND TESTING SAMPLES FOR THE TRENTO DATASET


















No.	color	Class	Training	Test
1		Apple trees	81	3 953
2		Buildings	58	2 845
3		Ground	10	469
4		Woods	182	8 941
5		Vineyard	210	10 291
6		Roads	63	3 111
Total			604	29 610

TABLE III
NUMBERS OF TRAINING AND TESTING SAMPLES FOR THE MUUFL GULFPORT







No.	color	Class	Training	Test
1		Trees	1 162	22 084
2		Mostly grass	214	4 056
3		Mixed Ground	344	6 538
4		Dirt and sand	91	1 735
5		Roads	334	6 353
6		Water	23	443
7		Shadows	112	2 121
8		Buildings	312	5 928
9		Sidewalks	69	1 316
10		Yellow curbs	9	174
11		Cloth Panels	14	255
Total			2 684	51 003

shown in Table I. The visualization of the HSI, LiDAR, and ground truth are shown in Fig. 9(a)–(c), respectively.

2) *Trento Dataset*: This dataset was obtained in the southern part of Trento, Italy, with a spatial resolution of 1.0 m and a size of 600×166 pixels. There are 63 bands in HSI data and two bands in LiDAR data. The ground cover types are divided into six types. 2% of the labeled pixels are randomly selected as a training set on this dataset. The detailed information is shown in Table II. The visualization of the HSI, LiDAR, and ground truth are shown in Fig. 10(a)–(c), respectively.

3) *MUUFLL Gulfport*: This dataset was obtained from the University of Southern Mississippi Gulf Park, with a size of 325×220 pixels. Initially, the HSI included 72 channels, but the information contained in eight channels was confirmed to be noise, so 64 spectral channels were ultimately left for experimentation. The LiDAR included two bands. There are 11 different types of ground cover. Here, 5% of the labeled pixels are selected as a training set. The detailed information is shown in Table III. The visualization of the HSI, LiDAR, and ground truth are shown in Fig. 11(a)–(c), respectively.

TABLE IV
NUMBERS OF TRAINING AND TESTING SAMPLES FOR THE TAMARIX DATASET

No.	color	Class	Training	Test
1		Dry spartina alterniflora	276	27 315
2		Phragmites australis	45	4 472
3		Water	186	18 384
4		Suaeda salsa	70	6 899
5		Tamarix chinensis	109	10 826
6		Roads	10	1 030
Total			696	68 926

4) *Tamarix Dataset*: Tamarix dataset was collected on the Yellow River Delta National Nature Reserve in August 2022. There are 126 bands in HSI data and 1 band in LiDAR data. The spatial resolution of the Tamarix dataset is resampled to 1.0 m. The space size is 387×673 pixels, and there are 69 622 trainable pixels. The dataset contains six divergent classes. Here, we select 1% of the labeled pixels as a training set. Table IV shows the details. The visualization of the HSI, LiDAR, and ground truth are shown in Fig. 12(a)–(c), respectively.

B. Parameter Tuning

Our network is implemented in a PyTorch framework. All deep learning models in this article are implemented on computers equipped with Intel Core CPU i9-10900 k, 96-GB RAM, and NVIDIA RTX 3090. During the training process, the Adam algorithm is applied to optimize the network, and cross entropy is applied as the loss function of the network. At the same time, we selected overall accuracy (OA), average accuracy (AA), and Kappa coefficient as evaluation indicators to evaluate the classification performance.

The adjustment of parameters is crucial to the performance of deep learning models. Among all the parameters, patch size ($p \times p$), learning rate (lr), and the numbers of training samples have a greater impact on the classification effect. So next, we will focus on analyzing the patch size, learning rate, and the ratio of training samples. We set the default values of p and lr to 11 and 0.00001, respectively. The default proportions for training samples in the four datasets are set to 10%, 2%, 5%, and 1%, respectively. For other secondary parameters, we set the batch size and epoch to 64 and 100 based on experience.

1) *Selection of Patch Size*: The size of the input patch determines how much information it contains. We compared the classification results for different input patch sizes on four datasets. Theoretically, the larger the patch, the more information it contains. However, excessive patch size can affect the training speed of the model, so selecting a reasonable range is important. We decided to limit the patch size to 7, 9, 11, and 13, and fix other parameters to default values. As shown in Fig. 4, the OA of the algorithm initially increases with the enlargement in patch size, but after reaching the peak value, there will be a slight downward trend. Coincidentally, when the OA of the four datasets reaches the peak, the patch size is 11×11 .

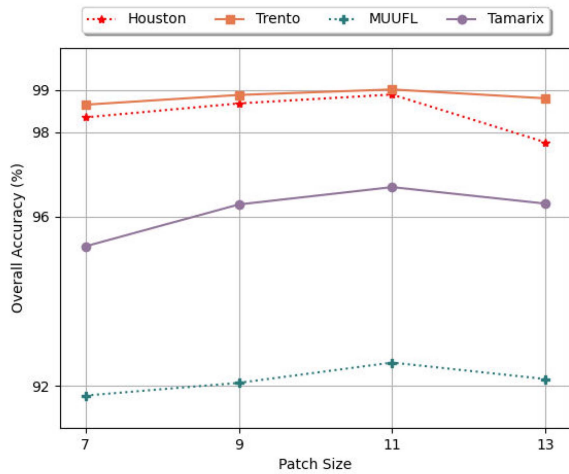


Fig. 4. OA with different patch size.

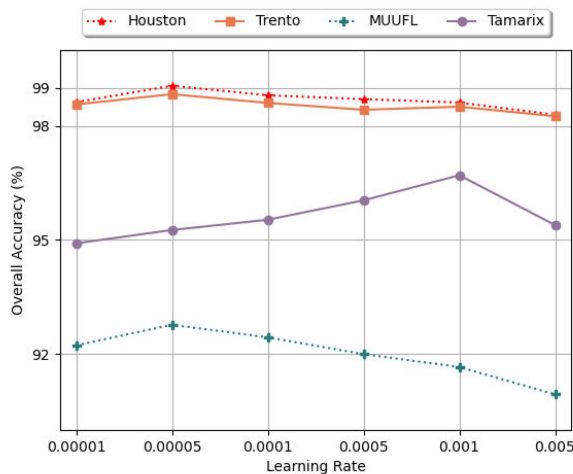


Fig. 5. OA with different learning rate.

2) *Selection of Learning Rate*: Learning rate is an important super parameter in deep learning, which determines the speed of convergence of our model. Generally speaking, the higher the learning rate, the faster the learning speed of the neural network. If the learning rate is too small, the network may mistake the local optimal solution for the global optimal solution; however, excessive learning rate may not find the local optimal solution. Based on previous experience, we decide to limit the learning rate to 0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, and fix other parameters to default values. Fig. 5 shows the performance of different learning rates on each dataset. When the learning rate is 0.00005, the model performs better in Houston, Trento, and MUUFL. In the Tamarix dataset, the model performs best when the learning rate is 0.001.

3) *Evaluation With Different Ratio of Training Samples*: Similarly, the proportion of training samples also requires experimental evaluation. On the one hand, if the number of training samples is insufficient, it may lead to underfitting of the model. On the other hand, excessive training samples will result in overfitting and poor adaptability. Therefore, it is crucial to

select an appropriate proportion of training samples for the performance of the model. Based on previous experience, we limited the proportion of training samples to a reasonable range, and conducted experiments with the proposed method and other comparative methods on four datasets to select the most suitable proportion based on the results.

Figs. 6–8 show the changes in OA, AA, and Kappa indicators using different proportions of training samples on four datasets. From the figure, we can see that as the proportion of training samples increases, OA, AA, and Kappa also increase accordingly. However, the rising speed gradually slows down, and some indicators are even at the same level as before. At this time, the increase in the proportion of training samples has little impact on the performance of the model. Our goal is to achieve satisfactory results with a smaller training sample ratio, so we set the training sample ratios for the Houston, Trento, MUUFL, and Tamarix datasets to 10%, 2%, 5%, and 1%, respectively.

C. Classification Performance

In order to verify the superiority of our proposed network over other methods in the field of multimodal remote sensing classification, some classic models and up-to-the-minute methods have been used to compare our network with experiments, such as SVM [17], the contextual deep CNN model (CDCNN) [43], 3-D deep learning approach (3D-CNN) [44], the two-branch CNN model TBCNN [23], CoupleNet [45], FusAtNet [26], and joint CNNs and morphological feature learning (MorphNet) [38].

Based on previous experience, we have supplemented some details of the network training process by conducting ten repeated experiments for each method, with 100 epochs per experiment, and taking the intermediate value as the result. Before training the model, we set the learning rate, patch size, and batch size of the proposed network to the optimal solution for the corresponding dataset. Similarly, the parameters of other methods are also their best choices. In addition, we make the number of MorFEB and SpeCB in the network to 3.

Tables V–VIII show the classification accuracy comparison results of each method on the Houston, Trento, MUUFL, and Tamarix datasets. The bold values in the table represent the best classification results for this category. Our conclusions are as follows.

The classification accuracy of algorithms based on deep learning is commonly better than that of classical machine learning algorithms. For instance, on the Houston dataset, even the worst performing CDCNN in deep learning methods has a 6.58% higher OA than SVM methods. This is because deep learning algorithms have more powerful feature extraction capabilities, which attempt to directly obtain high-level features from data.

Our proposed algorithm performs better than other methods based on deep learning. For the Houston, Trento, MUUFL, and Tamarix datasets, we have achieved OA of 99.09%, 99.01%, 92.86%, and 96.70% OA, respectively. Similarly, AA and Kappa metrics are also higher than other algorithms. Taking the Houston dataset as an example, OA is 5.29%, 3.17%, 2.61%, 2.72%, 2.40%, and 2.51% higher than CDCNN, 3D-CNN, two-branch CNN, CoupleNet, FusAtNet, and MorphNet, respectively. The

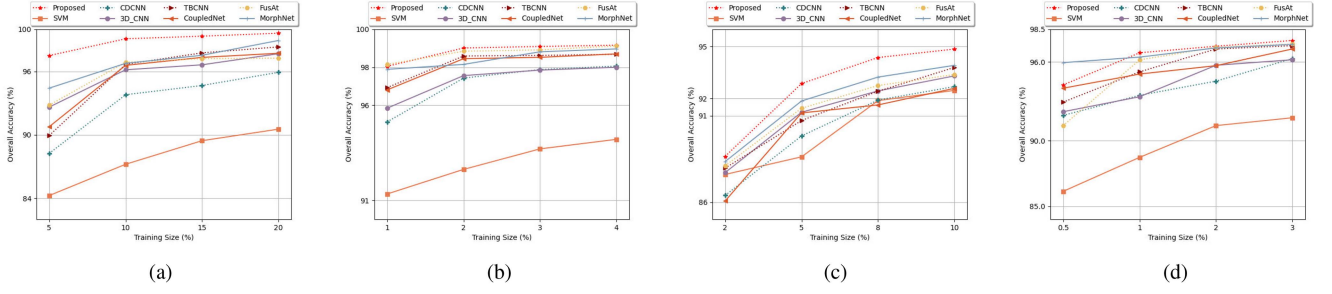


Fig. 6. OA achieved by different methods with varying training sample sizes which are randomly taken from (a) Houston (b) Trento (c) MUUFL, and (d) Tamarix datasets.

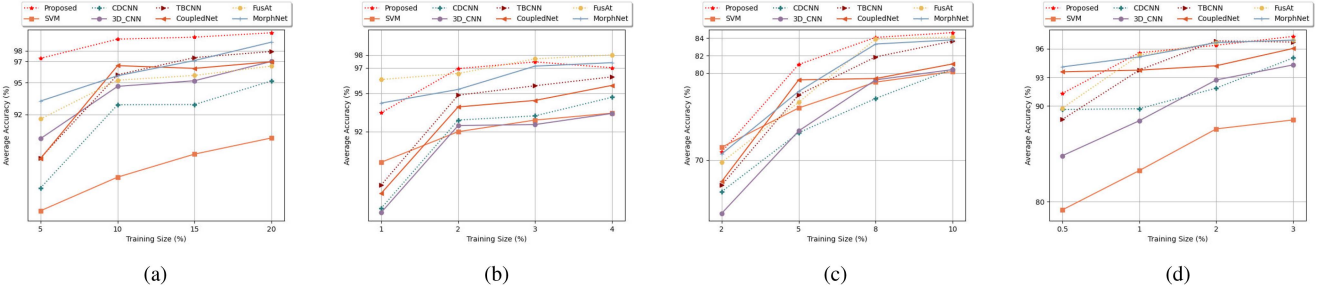


Fig. 7. AA achieved by different methods with varying training sample sizes which are randomly taken from (a) Houston, (b) Trento, (c) MUUFL, and (d) Tamarix datasets.

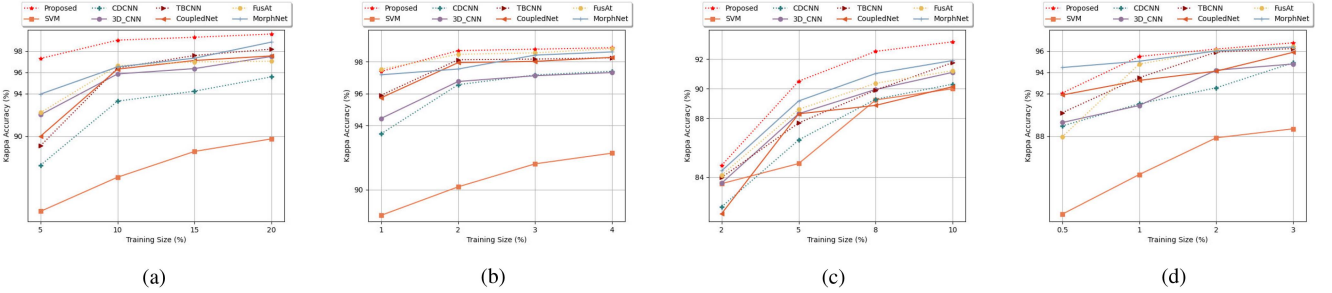


Fig. 8. Kappa (k) Accuracy achieved by different methods with varying training sample sizes which are randomly taken from (a) Houston (b) Trento (c) MUUFL, and (d) Tamarix datasets.

TABLE V
CLASSIFICATION RESULT OBTAINED BY DIFFERENT METHODS ON HOUSTON DATASET

Class	SVM	CDCNN	3D-CNN	Two-branch CNN	CoupleNet	FusAtNet	MorphNet	Proposed
1	92.81	98.67	97.96	96.98	96.98	96.00	97.42	97.78
2	98.67	98.85	98.94	98.49	98.67	99.11	99.29	99.91
3	99.20	98.72	99.84	100.00	100.00	100.00	100.00	100.00
4	98.84	95.09	99.02	99.91	96.88	99.11	99.91	99.64
5	98.48	99.19	99.64	99.82	99.55	99.46	99.82	99.82
6	84.25	88.01	81.51	85.96	94.52	81.51	84.93	98.29
7	86.50	94.30	96.23	99.04	96.84	98.69	99.39	99.12
8	78.48	92.95	95.98	96.61	98.39	98.39	98.04	99.55
9	80.04	83.67	93.52	89.71	91.57	95.65	91.84	97.69
10	82.97	93.84	97.10	96.65	92.75	98.64	95.74	99.64
11	82.19	88.94	95.86	97.30	95.86	97.66	95.86	99.28
12	76.31	96.58	96.85	97.39	95.86	96.04	96.76	97.93
13	35.31	69.67	69.19	80.33	94.08	76.54	78.44	97.63
14	97.92	95.32	99.22	98.70	97.14	91.69	98.18	100.00
15	99.66	100.00	98.82	99.16	99.83	100.00	99.16	100.00
OA	87.22	93.80	96.16	96.68	96.58	96.87	96.77	99.09
AA	86.11	92.92	94.65	95.74	96.60	95.23	95.65	99.09
Kappa	86.16	93.30	95.85	96.41	96.30	96.62	96.51	99.02

The bold values denote the best results.

TABLE VI
CLASSIFICATION RESULT OBTAINED BY DIFFERENT METHODS ON TRENTO DATASET

Class	SVM	CDCNN	3D-CNN	Two-branch CNN	CoupleNet	FusAtNet	MorphNet	Proposed
1	78.35	96.33	98.50	96.00	97.62	97.65	97.62	98.19
2	95.22	92.79	92.92	97.78	95.99	97.89	95.92	97.81
3	94.88	77.61	73.08	82.69	85.14	87.63	86.19	88.46
4	99.83	99.99	99.99	100.00	100.00	100.00	100.00	100.00
5	91.06	99.77	99.96	99.92	99.92	99.84	99.87	99.97
6	92.61	90.94	90.37	97.12	95.02	96.30	92.20	97.15
OA	92.63	97.43	97.56	98.59	98.45	98.84	98.15	99.01
AA	91.99	92.90	92.47	95.59	95.62	96.55	95.30	96.93
Kappa	90.17	96.56	96.76	98.12	97.94	98.46	97.54	98.69

The bold values denote the best results.

TABLE VII
CLASSIFICATION RESULT OBTAINED BY DIFFERENT METHODS ON MUUFL GULFPORT

Class	SVM	CDCNN	3D-CNN	Two-branch CNN	CoupleNet	FusAtNet	MorphNet	Proposed
1	95.84	97.22	97.73	97.38	96.76	97.74	97.72	98.22
2	74.95	74.01	76.23	73.62	81.83	75.81	78.38	82.99
3	81.17	80.50	85.41	83.88	79.72	87.78	84.80	86.34
4	85.36	86.51	84.84	89.97	93.89	86.40	89.45	92.05
5	93.40	91.36	94.30	93.70	94.03	93.88	95.09	94.25
6	85.33	85.33	85.55	97.52	90.52	89.39	94.58	97.07
7	75.01	90.33	86.19	84.02	93.64	86.80	87.93	87.74
8	93.29	96.32	96.69	95.12	96.61	96.54	96.79	97.37
9	38.53	49.24	57.60	55.85	49.24	46.20	56.99	58.81
10	25.86	13.22	13.22	17.82	17.82	8.62	14.94	13.22
11	87.06	40.39	29.02	63.53	77.25	73.73	60.00	82.75
OA	88.62	89.83	91.20	90.72	91.16	91.43	91.85	92.86
AA	75.98	73.13	73.34	77.49	79.21	76.63	77.88	80.98
Kappa	84.93	86.53	88.31	87.68	88.30	88.60	89.17	90.51

The bold values denote the best results.

TABLE VIII
CLASSIFICATION RESULT OBTAINED BY DIFFERENT METHODS ON TAMARIX DATASET

Class	SVM	CDCNN	3D-CNN	Two-branch CNN	CoupleNet	FusAtNet	MorphNet	Proposed
1	97.09	97.08	97.13	96.24	98.31	97.24	98.24	97.96
2	82.31	91.30	86.87	90.44	95.15	91.94	92.72	95.73
3	98.97	98.64	97.63	99.81	99.39	98.79	99.06	99.42
4	37.56	86.14	82.28	86.91	84.34	93.01	92.19	88.76
5	84.52	82.10	87.64	92.15	86.01	92.53	91.17	94.35
6	99.03	83.08	79.23	96.92	99.33	98.37	97.40	97.12
OA	88.73	93.46	93.35	95.25	95.08	96.16	96.37	96.70
AA	83.25	89.72	88.46	93.75	93.75	95.31	95.13	95.56
Kappa	84.41	91.05	90.89	93.50	93.24	94.77	95.04	95.49

The bold values denote the best results.

data fusion method adopted by the CDCNN method and the 3D-CNN method is quite simple, just overlaying HSI and LiDAR data and feeding them into the network, which ignores the differences between multimodal data, and the efficiency of information fusion is not satisfactory. In the two-branch CNN method, only simple convolutions are used to extract features from the HSI and LiDAR, respectively, which lacks a more powerful feature extraction method; CoupleNet fuses data at both the feature level and the decision level, but lacks targeted fusion methods; FusAtNet applies cross attention to multimodal remote sensing classification for the first time, but the model has multitudinous parameters and a long running time; MorphNet applies the dilation and erosion operations in morphology to multimodal remote sensing classification, but the extraction of spectral features is too simple and requires more effective

spectral feature extraction methods. Compared with other methods, our network has been improved accordingly. On the one hand, our model combines dilation and erosion operations in morphology with convolution to extract spatial and elevation information more effectively; on the other hand, our model enhances features through a self-attention mechanism, and optimizes the HSI using LiDAR's elevation information through cross attention. In addition, we also consider the correlation and difference between features and adopt targeted fusion strategies.

Finally, in order to display the experimental results more intuitively, we use different colors to label different ground types. Figs. 9–12 show the classification results of various algorithms on four datasets. It is evident that our proposed method is closer to the corresponding ground truth and shows fewer error marks than other algorithms.

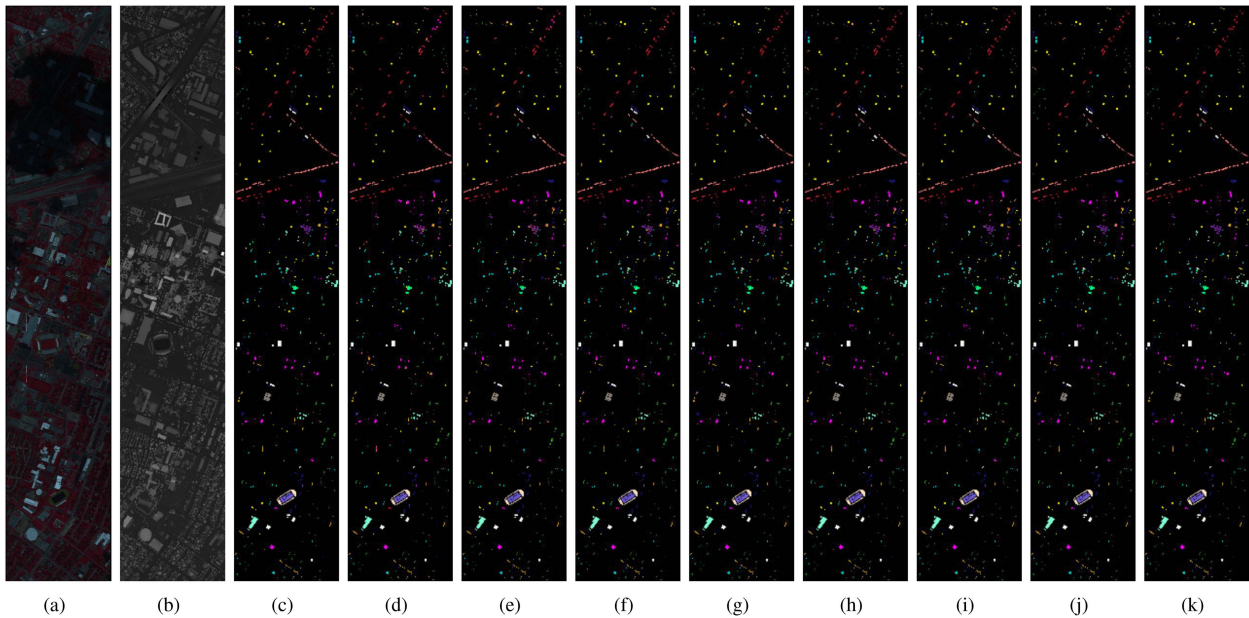


Fig. 9. Houston data visualization and classification maps obtained by different models. (a) False-color image for the HSI over bands 60, 27, and 11, respectively. (b) Grayscale image for LiDAR data. (c) Ground truth. (d) SVM. (e) CDCNN. (f) 3D-CNN. (g) Two-branch CNN. (h) CoupleNet. (i) FusAtNet. (j) MorphNet. (k) Proposed method.

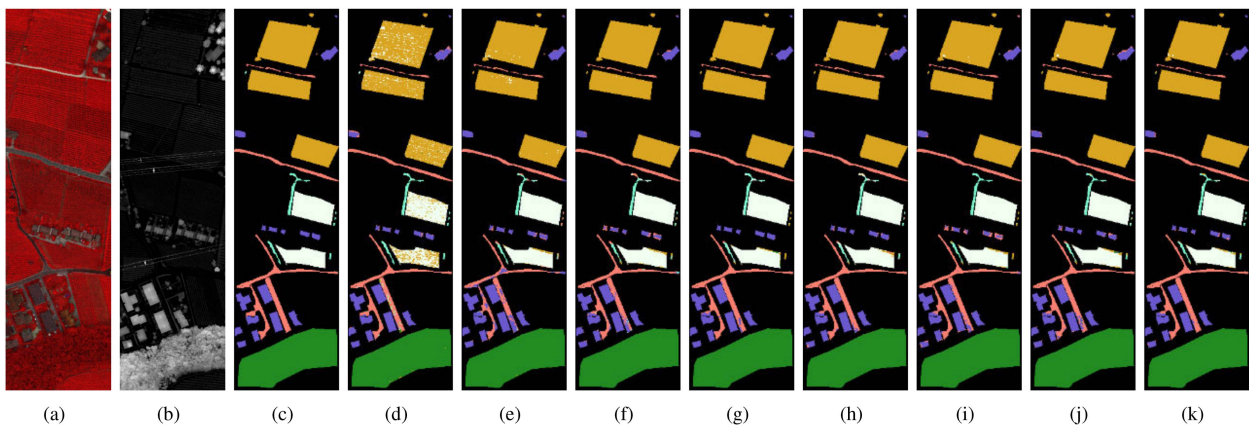


Fig. 10. Trento data visualization and classification maps obtained by different models. (a) False-color image for the HSI over bands 60, 27, and 11, respectively. (b) Grayscale image for LiDAR data. (c) Ground truth. (d) SVM. (e) CDCNN. (f) 3D-CNN. (g) Two-branch CNN. (h) CoupleNet. (i) FusAtNet. (j) MorphNet. (k) Proposed method.

D. Ablation Study

Finally, we conducted ablation experiments to verify the role of each module in the network. The ablation experiment is divided into two parts. The first part removes MorFEB, SpeCB, and PAM modules from the network one by one, and observes the changes in classification results after removing each module. The second part verifies the impact of different fusion strategies on the results. We conducted ablation experiments on all four datasets, and recorded the OA of the experimental results in Tables IX and X.

To demonstrate the powerful feature extraction capabilities of dilation and erosion, we replace the MorFEB with traditional convolution. The comparison result in Table IX shows that the

TABLE IX
ABLATION EXPERIMENTS ABOUT DIFFERENT MODULE ON
DIFFERENT DATASETS

MorFEB	SpeCB	PAM	OA (%)			
			Houston	Trento	MUUFLL	Tamarix
			98.19	98.10	90.25	94.25
✓			98.44	98.33	90.85	95.13
	✓		98.40	98.23	90.92	94.98
		✓	98.36	98.27	90.70	94.70
✓	✓		98.64	98.57	92.04	96.08
✓		✓	98.80	98.70	91.37	95.97
	✓	✓	98.59	98.58	91.14	95.68
✓	✓	✓	99.09	99.01	92.86	96.70

The bold values denote the best results.

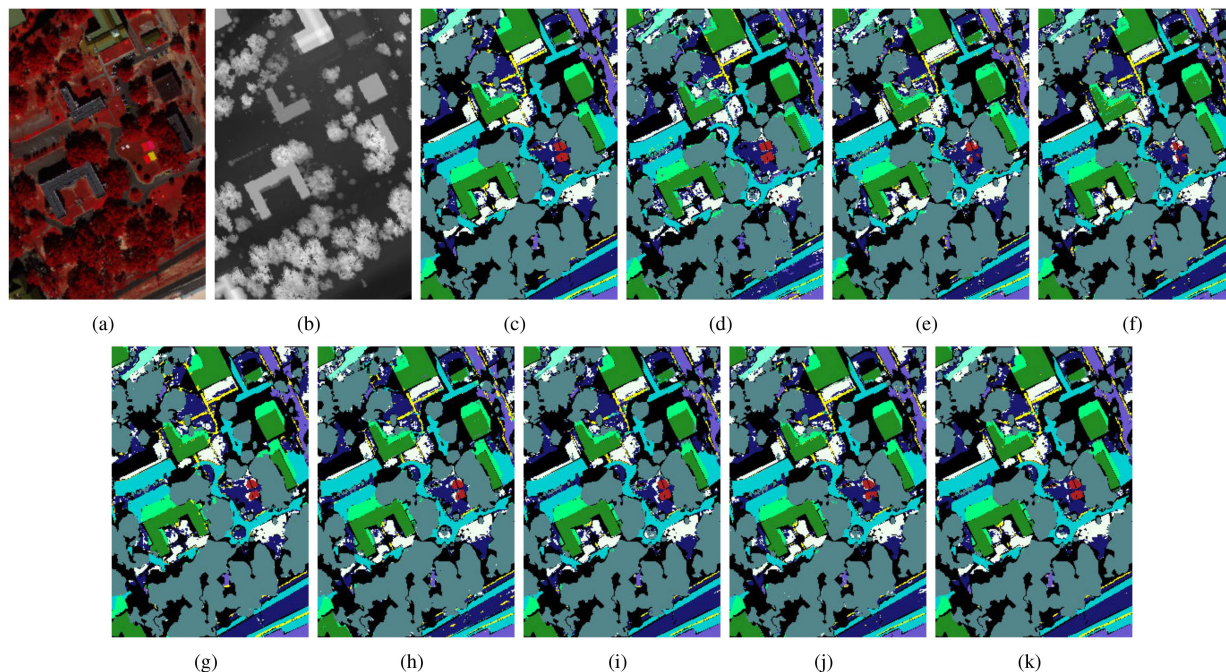


Fig. 11. MUUFL data visualization and classification maps obtained by different models. (a) False-color image for the HSI over bands 60, 27, and 11, respectively. (b) Grayscale image for LiDAR data. (c) Ground truth. (d) SVM. (e) CDCNN. (f) 3D-CNN. (g) Two-branch CNN. (h) CoupleNet. (i) FusAtNet. (j) MorphNet. (k) Proposed method.

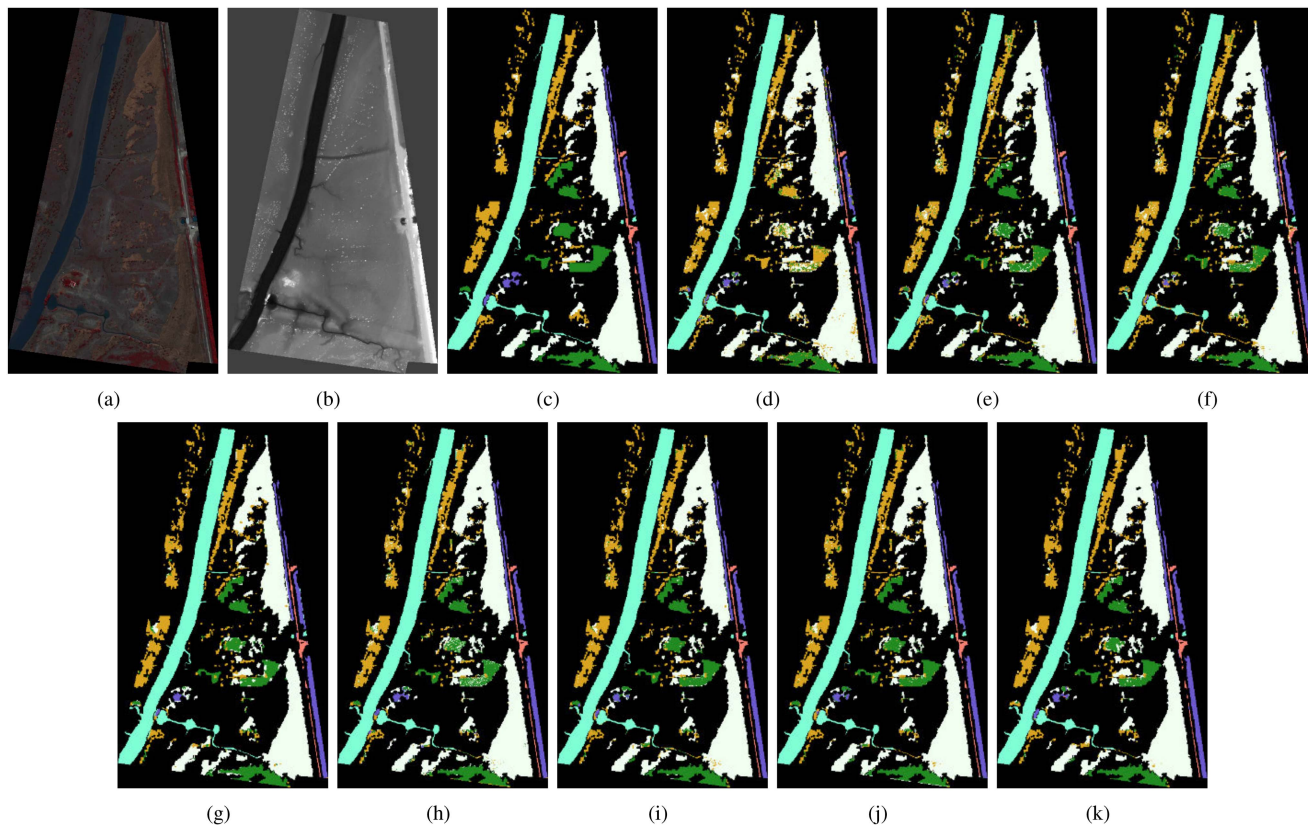


Fig. 12. Tamarix data visualization and classification maps obtained by different models. (a) False-color image for the HSI over bands 100, 60, and 40, respectively. (b) Grayscale image for LiDAR data. (c) Ground truth. (d) SVM. (e) CDCNN. (f) 3D-CNN. (g) Two-branch CNN. (h) CoupleNet. (i) FusAtNet. (j) MorphNet. (k) Proposed method.

TABLE X
ABLATION EXPERIMENTS ABOUT DIFFERENT FUSION STRATEGIES ON
DIFFERENT DATASETS

Spa and Ele	Spa-Ele and Spe	OA (%)			
		Houston	Trento	MUUFL	Tamarix
Concatenate	Concatenate	98.79	98.47	91.65	95.54
AFF	AFF	98.21	97.53	90.06	94.83
Concatenate	AFF	97.86	98.13	90.48	95.08
AFF	Concatenate	99.09	99.01	92.86	96.70

The bold values denote the best results.

classification accuracy of the network after removing MorFEB is lower than before, which proves the effectiveness of the MorFEB.

The SpeCB in the network can improve the resolution of the central pixel. We deleted SpeCB from the network to test the effectiveness of SpeCB. Through comparative experiments, we find a slight decrease in classification accuracy after removing SpeCB, which confirms the effectiveness of SpeCB.

To verify the effectiveness of the PAM module in the network, we directly remove the PAM from the network, and no longer optimize the spatial features and elevation features. Based on the results in Table IX, we can conclude that using the PAM can increase the information interaction between multimodal data and obtain better classification results.

In addition, we sequentially use concatenation and AFF at different feature fusion sites (which may require changing some parameters) to evaluate the effectiveness of AFF. The experimental results in Table X indicate that using appropriate fusion strategies can achieve the expected results, and also confirm that attention feature fusion can more effectively fuse multimodal features.

IV. CONCLUSION

In this article, a network for joint classification of the HSI and LiDAR is proposed. This network applies morphological dilation and erosion operations, which capture more boundary shape information of arbitrary objects in complex regions and extract differentiated spatial information. Second, the PAM is used to correct spatial features. Specifically, we use a position weight map obtained from LiDAR to optimize the spatial features of the HSI; the SpeCB is used to extract the channel weight of the center pixel to calibrate the spectral features. In addition, we introduce the adaptive feature fusion module, AFF, which considers the problem of inconsistent feature semantics and different scales and does not increase the feature dimension. In addition to the three common multimodal datasets, we also compared the algorithm proposed in this article with other advanced methods on a self-made multimodal dataset, and the results show that our algorithm is relatively preferable. In fact, our algorithm relies on the accuracy of labeled samples, and accurate labeling of samples requires a lot of time. In future work, we will consider working towards semisupervised or unsupervised directions.

REFERENCES

- [1] C. Ge, Q. Du, W. Li, Y. Li, and W. Sun, "Hyperspectral and LiDAR data classification using kernel collaborative representation based residual fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1963–1973, Jun. 2019.
- [2] Y. Dong, T. Liang, Y. Zhang, and B. Du, "Spectral-spatial weighted kernel manifold embedded distribution alignment for remote sensing image classification," *IEEE Trans. Cybern.*, vol. 51, no. 6, pp. 3185–3197, Jun. 2020.
- [3] M. Zhang, P. Ghamisi, and W. Li, "Classification of hyperspectral and LiDAR data using extinction profiles with feature fusion," *Remote Sens. Lett.*, vol. 8, no. 10, pp. 957–966, 2017.
- [4] C. Chen, X. Zhao, W. Li, R. Tao, and Q. Du, "Collaborative classification of hyperspectral and LiDAR data with information fusion and deep nets," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 2475–2478.
- [5] W. Li, Q. Du, and B. Zhang, "Combined sparse and collaborative representation for hyperspectral target detection," *Pattern Recognit.*, vol. 48, no. 12, pp. 3904–3916, 2015.
- [6] M. Ahmad et al., "Hyperspectral image classification—Traditional to deep models: A survey for future prospects," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 968–999, Dec. 2021, doi: [10.1109/JSTARS.2021.3133021](https://doi.org/10.1109/JSTARS.2021.3133021).
- [7] D. Hong et al., "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 52–87, Jun. 2021.
- [8] 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.
- [9] H. Su, Y. Yu, Q. Du, and P. Du, "Ensemble learning for hyperspectral image classification using tangent collaborative representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3778–3790, Jun. 2020.
- [10] L.-Z. Huo et al., "Supervised spatial classification of multispectral LiDAR data in urban areas," *PLoS One*, vol. 13, no. 10, 2018, Art. no. e0206185.
- [11] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.
- [12] M. Pedernana, P. R. Marpu, M. Dalla Mura, J. A. Benediktsson, and L. Bruzzone, "Classification of remote sensing optical and LiDAR data using extended attribute profiles," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 7, pp. 856–865, Nov. 2012.
- [13] W. Sun, G. Yang, J. Peng, and Q. Du, "Lateral-slice sparse tensor robust principal component analysis for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 1, pp. 107–111, Jan. 2020.
- [14] M. P. Uddin, M. A. Mamun, M. A. Hossain, and M. Ibn Afjal, "Improved folded-PCA for efficient remote sensing hyperspectral image classification," *Geocarto Int.*, vol. 37, no. 25, pp. 9474–9496, 2022.
- [15] M. S. H. Shemul, M. M. Rahman, S. Ahmed, M. A. Marjan, M. P. Uddin, and M. I. Afjal, "Segmented-sparse-PCA for hyperspectral image classification," in *Proc. 4th Int. Conf. Elect., Comput. Telecommun. Eng.*, 2022, pp. 167–170.
- [16] 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.
- [17] M. A. Hearst, S. T. 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.
- [18] J. Xia, J. Chanussot, P. Du, and X. He, "Spectral-spatial classification for hyperspectral data using rotation forests with local feature extraction and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2532–2546, May 2015.
- [19] J. Peng, Y. Zhou, and C. P. Chen, "Region-kernel-based support vector machines for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 9, pp. 4810–4824, Sep. 2015.
- [20] J. Xia, J. Chanussot, P. Du, and X. He, "Rotation-based support vector machine ensemble in classification of hyperspectral data with limited training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1519–1531, Mar. 2016.
- [21] Z. Xiong, Y. Yuan, and Q. Wang, "AI-NET: Attention inception neural networks for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 2647–2650.
- [22] X. Gao, T. Chen, R. Niu, and A. Plaza, "Recognition and mapping of landslide using a fully convolutional DenseNet and influencing factors," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7881–7894, Aug. 2021, doi: [10.1109/JSTARS.2021.3101203](https://doi.org/10.1109/JSTARS.2021.3101203).

- [23] 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.
- [24] P. Guan and E. Y. Lam, "Cross-domain contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, May 2022, doi: [10.1109/TGRS.2022.3176637](https://doi.org/10.1109/TGRS.2022.3176637).
- [25] Y. Duan, H. Huang, and T. Wang, "Semisupervised feature extraction of hyperspectral image using nonlinear geodesic sparse hypergraphs," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.
- [26] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "FusAtNet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and LiDAR classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 92–93.
- [27] 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, pp. 1–16, Sep. 2021, doi: [10.1109/TGRS.2021.3106025](https://doi.org/10.1109/TGRS.2021.3106025).
- [28] H.-C. Li, W.-S. 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.
- [29] C. Li, R. Hang, and B. Rasti, "EMFNet: Enhanced multisource fusion network for land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4381–4389, Apr. 2021, doi: [10.1109/JS-TARS.2021.3073719](https://doi.org/10.1109/JS-TARS.2021.3073719).
- [30] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.
- [31] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.
- [32] P. Ghamisi, R. Souza, J. A. Benediktsson, X. X. Zhu, L. Rittner, and R. A. Lotufo, "Extinction profiles for the classification of remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5631–5645, Oct. 2016.
- [33] 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, May 2017.
- [34] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. X. Zhu, "Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3791–3808, Jun. 2020.
- [35] S. K. Roy, B. Chanda, B. B. Chaudhuri, D. K. Ghosh, and S. R. Dubey, "Local morphological pattern: A scale space shape descriptor for texture classification," *Digit. Signal Process.*, vol. 82, pp. 152–165, 2018.
- [36] D. Mellouli, T. M. Hamdani, M. B. Ayed, and A. M. Alimi, "Morph-CNN: A morphological convolutional neural network for image classification," in *Proc. Neural Inf. Processing: 24th Int. Conf.*, 2017, pp. 110–117.
- [37] G. Franchi, A. Fehri, and A. Yao, "Deep morphological networks," *Pattern Recognit.*, vol. 102, 2020, Art. no. 107246.
- [38] S. K. Roy, A. Deria, D. Hong, M. Ahmad, A. Plaza, and J. Chanussot, "Hyperspectral and LiDAR data classification using joint CNNs and morphological feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, May 2022, doi: [10.1109/TGRS.2022.3177633](https://doi.org/10.1109/TGRS.2022.3177633).
- [39] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3560–3569.
- [40] L. Mou and X. X. Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 110–122, Jan. 2020.
- [41] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019.
- [42] L. Sun, Y. Fang, Y. Chen, W. Huang, Z. Wu, and B. Jeon, "Multi-structure KELM with attention fusion strategy for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, Sep. 2022, doi: [10.1109/TGRS.2022.3208165](https://doi.org/10.1109/TGRS.2022.3208165).
- [43] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.
- [44] A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018.
- [45] 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.



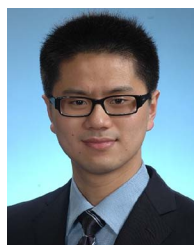
Zhongwei Li received the Ph.D. degree in oil and gas well engineering, in 2011, from the China University of Petroleum (East China), Qingdao, China, where he is currently working toward the master degree with the College of Computer science and Technology.

His current research interests include remote sensing image processing, ocean numerical forecasting, and cloud computing.



Hao Sui received the B.E. degree in computer science and technology, in 2021, from the China University of Petroleum (East China), Qingdao, China, where he is currently working toward the master degree with the College of Computer science and Technology.

His current research interest is hyperspectral and LiDAR joint classification.



Cai Luo (Senior Member, IEEE) received the Ph.D. degree in electronic and computer engineering, robotics and telecommunication from the University of Genoa, Genoa, Italy, in 2012.

He is currently a Lecturer with the China University of Petroleum (East China), Qingdao, China. His current research interests include unmanned aerial vehicle biomimetic design and dynamics and control of robotic systems.



Fangming Guo (Graduate Student Member, IEEE) received the B.E. degree in computer science and technology, in 2018, from the China University of Petroleum (East China), Qingdao, China, where he is currently working toward the doctor degree with the College of Oceanography and Space Informatics.

His current research interests in the hyperspectral images classification.