Collaborative Contrastive Learning for Hyperspectral and LiDAR Classification

Sen Jia, *Senior Member, IEEE*, Xi Zhou, Shuguo Jiang, and Ruyan He

***Abstract*— Using single-source remote sensing (RS) data for classification of ground objects has certain limitations; however, multimodal RS data contain different types of features, such as spectral features and spatial features of hyperspectral image (HSI) and elevation information of light detection and ranging (LiDAR) data, which can be used to extract and fuse high-quality features to improve the classification accuracy. Nevertheless, the existing fusion techniques are mostly limited by the number of labeled samples due to the difficulty of label collection in the multimodal RS data. In this article, a fusion method of collaborative contrastive learning (CCL) is proposed to tackle the abovementioned issues for HSI and LiDAR data classification. The proposed CCL approach includes two stages of pretraining (CCL-PT) and fine-tuning (CCL-FT). In the CCL-PT stage, a collaborative strategy is introduced into contrastive learning (CL), which can extract features from HSI and LiDAR data separately and achieve the coordinated feature representation and matching between the two-modal RS data without labeled samples. In the CCL-FT stage, a multilevel fusion network is designed to optimize and fuse the unsupervised collaborative features, which are extracted in the CCL-PT stage for the classification tasks. Experimental results on three real-world datasets show that the developed CCL approach can perform excellently on the small sample classification tasks, and CL is feasible for the fusion of multimodal RS data.**

***Index Terms*— Contrastive learning (CL), hyperspectral image (HSI), light detection and ranging (LiDAR).**

NOMENCLATURE

Symbol	*Meaning*
$\boldsymbol{X}_h$, $\boldsymbol{X}_l$	HSI and LiDAR data.
H, W, B	Height, width, and channels of the HSI or LiDAR data.
$\boldsymbol{x}_h$, $\boldsymbol{x}_l$	Patch cubes for HSI and patches for LiDAR data.
P	Neighborhood size.
f_h, f_l, f_{mlf}	Encoders for HSI, LiDAR data, and multilevel fusion.
g	Projector.
h	Predictor.
N	Number of stacked residual blocks.
$\boldsymbol{p}_h$, $\boldsymbol{p}_l$	Output vector of hyperspectral and LiDAR data through h.
$\boldsymbol{z}_h$, $\boldsymbol{z}_l$	Output vector of hyperspectral and LiDAR data through g.
stopgrad	Branches of the network without gradient back-propagation.
C_h, C_l	Output constants for stopgrad($\boldsymbol{z}_h$) and stopgrad($\boldsymbol{z}_l$).
q	Convolution blocks.
$\boldsymbol{a}_h$, $\boldsymbol{a}_l$	Features of different depths from hyperspectral and LiDAR data.
$\boldsymbol{a}_{\text{hl}}$	Concatenated features of $\boldsymbol{a}_h$ and $\boldsymbol{a}_l$.
$\boldsymbol{m}_{\text{hl}}$	Features of different depths extracted by the fusion network.
$\boldsymbol{r}_{\text{hl}}$	Output vector through q.
θ	Parameters of convolution layer.
$\hat{y}$, y	Predicted values and ground truth for pixels.

Manuscript received 29 October 2022; revised 8 February 2023; accepted 25 March 2023. Date of publication 31 March 2023; date of current version 11 April 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62271327 and Grant 41971300; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011290; and in part by the Shenzhen Science and Technology Program under Grant RCJC20221008092731042, Grant JCYJ20220818100206015, and Grant KQTD20200909113951005. *(Corresponding author: Ruyan He.)*

The authors are with the College of Computer Science and Software Engineering, the Guangdong–Hong Kong–Macau Joint Laboratory for Smart Cities, and the Key Laboratory for Geo-Environmental Monitoring of Coastal Zone, Ministry of Natural Resources, Shenzhen University, Shenzhen 518060, China (e-mail: senjia@szu.edu.cn; 992566968@qq.com; shuguoj@foxmail.com; luckhry106@163.com).

Digital Object Identifier 10.1109/TGRS.2023.3263511

I. INTRODUCTION

THE hyperspectral image (HSI) acquired by combining imaging technology and spectral technology contains rich features both in the spatial and spectral aspects [1], which have unique advantages in ground object recognition [2]. HSI has a higher spectral resolution than other remote sensing (RS) data and has been widely used for classification of ground objects. In the previous studies, researchers have proposed a large number of classification approaches for HSI [3], which are mainly based on the spectral features of pixels [4], [5], the neighborhood features of space [6], and the joint features of spectral and spatial features [7], [8], [9], [10]. However, there are some disadvantages if HSIs are only adopted for classification. On the one hand, the process of HSI acquisition is easily affected by factors, such as the atmosphere, which can result in the phenomenon of the same objects with different spectral response curves and the different objects with similar spectral response curves. On the other hand, the mixed pixels in HSI with the relatively low spatial resolution may make it difficult to achieve the goal of fine-grained classification. In order to overcome these shortcomings, several researchers try to take advantage of multimodal RS data to improve the HSI classification accuracy [11], [12]. In particular, light detection and ranging (LiDAR) data, as a common RS data, can provide high-precision digital elevation models [13] and

elevation information, which can provide useful features as complementary information for HSI to satisfy the different application requirements, such as complex area classification [14] and vegetation coverage analysis [15]. Also, the LiDAR data are not affected by the cloud, so that it can offer supplementary information to compensate for the details of cloudy areas in HSI.

The fusion of HSI and LiDAR data has received much attention. In general, the fusion techniques of the two-modal RS data can be divided into three levels according to the different image information processing methods, which are pixel-level fusion, feature-level fusion, and decision-level fusion [16]. Also, there are some studies combining the abovementioned fusion techniques for HSI and LiDAR data [17], [18], [19]. Pixel-level fusion is performed directly on raw HSI and LiDAR data. Feature-level fusion is to fuse the feature information, which is extracted from raw RS data. Decision-level fusion is a fusion process based on decision rules after a series of image processing, such as feature extraction, feature recognition, and decision classification. In particular, due to the obvious heterogeneity of HSI and LiDAR data, feature-level fusion is more common in the existing research works [20].

Pedergnana et al. [21] proposed a fusion method by superimposing features of raw spectral information and morphologically extended attribute profiles (EAPs) from HSI and LiDAR data. The way of directly stacking features has the problem of high feature dimension, and the "Hughes phenomenon" can be prone to occur in the training process [22]. Therefore, researchers have proposed different approaches to reduce the feature dimension for HSI and LiDAR data fusion [23], [24], [25]. Ghamisi et al. [26] performed principal component analysis (PCA) on HSI before fusion to extract the top n principal components. Total variational component analysis (TVCA) is used to map features from a high-dimensional space to a low-dimensional space for the fusion and classification [27]. Similarly, Rasti et al. [28] proposed a sparse low-rank component analysis (SLRCA) method to fuse the extracted features. Although the above studies reduce the redundancy of spectral information of HSI, there is the possibility of losing some effective information for classification [29]. Moreover, these methods depend on the handcrafted features, and the deep features are excavated insufficiently, which may not well fit the complex nonlinear relationship of ground object features in HSI and LiDAR data [30].

Deep learning (DL) methods, to some extent, can make up for the shortcomings of traditional fusion methods, which can automatically extract features and learn rich semantic information from multimodal RS data. Currently, scholars have carried out many DL-based studies on the fusion of HSI and LiDAR data for the purpose of classification. The DL-based methods can be roughly divided into two branches, including the supervised classification and the unsupervised classification. The supervised classification refers to the use of the feature prior knowledge obtained from a large number of labeled samples to fuse and classify for specific tasks, which is susceptible to human subjective factors and time-consuming for sample labeling. In contrast, the unsupervised classification provides some advantages for HSI and LiDAR data, since it does not require labeled samples for model training and learning the potential features from the given data.

Several supervised classification methods have been developed. Chen et al. [31] proposed a two-branch convolutional neural network (CNN) model to separately extract the deep features of HSI and LiDAR data for the fusion and classification. They considered the information in the spatial neighbors of a given pixel, while the spectral information of HSI is not fully utilized in the CNN. Xu et al. [32] changed the HSI branch CNN to a two-tunnel CNN framework extracting both the spectral and spatial features on the basis of two-branch CNNs model. Hang et al. [33] adopted a parameter-sharing strategy and determined the classification accuracy of each output through adaptive weights to assess the contribution of different feature information for classification. Li et al. [34] applied a three-branch CNNs model to extract spectral features, spatial features, and elevation features and transformed these features into a kernel space for HSI and LiDAR data fusion.

The unsupervised methods have also been proposed. Zhang et al. [35] proposed an unsupervised patch-to-patch CNN (PToP CNN) to extract the features of HSI and LiDAR data for classification. Hong et al. [36] designed a deep encoder–decoder network architecture (EndNet) by enforcing the fusion features of HSI and LiDAR data, which can better activate the multimodel features in the case of insufficient labeled samples. Mohla et al. [37] presented a feature fusion and extraction framework (FusAtNet), which extracted modality-specific features and ensemble features through the self-attention mechanism and cross-attention mechanism to generate attention maps for HSI and LiDAR data classification. Jia et al. [38] proposed a multiple feature-based superpixel-level decision fusion (MFSuDF) from the perspective of the impact of compression noise and achieved superior accuracy for classification task.

Contrastive learning (CL) emerging in the field of natural images has relatively low requirements for data annotation, which can directly learn general feature representations from unlabeled data with positive and negative samples constructed by data augmentation methods. In this article, cross-modal input consisting of HSI and LiDAR data is used as a pair of positive samples without data augmentation, because different RS data contain similar parts and heterogeneous parts for the same ground object. Then, inspired by SimSiam framework [39], we propose a collaborative CL (CCL) approach for HSI and LiDAR data fusion and classification, which includes the pretraining (CCL-PT) for collaborative feature representation and matching and the fine-tuning (CCL-FT) for multilevel fusion of the two-modal RS features.

The main contributions of this article are summarized as follows.

1) First, in the CCL-PT stage, we introduce a collaborative strategy into CL framework for the coordinated feature learning and matching of HSI and LiDAR data without labeled samples. More specifically, we design a novel

CCL network based on ResNet to capture the discriminative collaborative features from HSI and LiDAR data separately with no shared weights. After that, the collaborative features are projected into a shared abstract space where the potential complementary information of the two-modal RS data can be excavated deeply. Finally, the feature matching is achieved by prediction pretext task under the constraint of the similarity measure of CL method.

2) Second, in the CCL-FT stage, we design a multilevel fusion network to accurately construct a mapping relationship and hierarchically fuse the collaborative features of HSI and LiDAR data for classification. In particular, we first fine-tune the collaborative features with a small number of labeled samples to fully explore the semantic correlations between the two-modal RS data. Then, the two-modal features of different depths are interacted and fused based on cross channel via the multilevel fusion network instead of a simple concatenation strategy.
3) Third, our CCL approach reduces the reliance on data augmentation and eliminates the need to construct additional negative samples in the CL-based methods. The high robustness for training results is achieved with a large number of unlabeled samples by unsupervised CCL-PT. Extensive experiments on three real-world HSI and LiDAR datasets demonstrate that the proposed CCL approach is effective in solving the classification problems of single-source RS data and limited labeled samples and outperforms other state-of-the-art approaches, such as FusAtNet and CoupledCNNs.

The rest of this article is organized as follows. Section II provides a brief overview of related work on self-supervised CL and CL-based methods for HSI classification. The proposed CCL approach is presented with two stages of CCL-PT and CCL-FT in Section III. The experimental description and the result analysis on three real-world datasets are provided in Section IV. Finally, Section V summarizes the conclusions of this article.

II. Related Works

A. Self-Supervised CL

CL, a type of self-supervised learning without labeled samples, usually refers to using the pretext task to discriminate positive samples and negative samples and automatically extracts the features of samples for training models. The core idea is to make the distance between positive samples closer and the distance between negative samples farther in the sample feature space. In general, the CL methods include two stages: the pretraining stage and the fine-tuning stage. The input data are used to construct positive samples and negative samples as supervised information by data augmentation to learn features in the pretraining stage, and the high-quality feature extraction modules learned in the pretraining stage are transferred to downstream tasks for fine-tuning. With the development of CL, researchers have proposed many meaningful CL-based methods mainly from two aspects of increasing the number of negative samples and constructing the diversity of samples by different strategies of data augmentation.

The traditional CL-based methods generally store negative samples in memory banks to increase the number of negative samples, which has a disadvantage of inconsistency between old and new encoders. To solve that problem, He et al. [40] proposed a MoCo model that retained the memory banks, but they used queues and momentum encoders to ensure the consistency of the old and new encoders. Chen et al. [41] replaced the memory bank by adding nonlinear layers and enlarging batch size to increase the number of negative samples in the proposed SimCLR model. Furthermore, they experimented with a series of data augmentation strategies, such as flipping, scaling, and so on, to explore which strategy could extract high-quality features for representations in the learning process. After that, the MoCo model was optimized by adding a nonlinear layer to the network and improving the data augmentation method based on the SimCLR model [42], whereas the SimCLR model was modified via using a deeper network and a memory bank referenced to [43].

Another disadvantage of CL methods is that it requires many negative samples for contrast, which is time-consuming and memory-consuming. To overcome this problem, Caron et al. [44] proposed an SwAV model based on the prior information of cluster centers instead of constructing negative samples to distinguish the clusters of each classification. Grill et al. [45] also proposed a BYOL model without using negative samples for contrast, and they applied the representation of a view to predict other views of the same image. Moreover, in order to avoid the possibility of model collapse due to the missing negative samples, the BYOL model used a sliding average to update the parameters of another branch after the gradient return of one branch. Continuing the prediction idea of BYOL, Chen and He [39] designed a simple Siamese network, named SimSiam, to directly maximize the similarity of two views generated by a picture without the need for negative samples, cluster centers, and momentum encoders, and they found that stop gradient was the key operation to avoid model collapse.

B. CL for HSI Classification

In recent years, self-supervised CL methods have been applied to HSI classification. Hou et al. [46] proposed a two-stage training strategy based on self-supervised CL to use the information of a large number of unlabeled samples to tackle the problem of insufficient labeled information in HSI. Zhao et al. [47] also introduced a self-supervised method to solve the HSI classification problem with only a few labeled samples. These studies show that self-supervised CL methods can learn effective feature representations of RS data and make great achievements in the RS image domain. Therefore, other researchers have conducted in-depth study on this method for hyperspectral data.

One way is to use the valuable information in HSI to construct various types of inputs for learning cross-domain representations in CL. Lee and Kwon [48] proposed a cross-domain CNN by using different HSIs with multiple spectral

characteristics for learning representations, which can merge different HSIs to learn rich multidimensional features for classification. In order to make use of the semantic information in the spectral and spatial domains of HSI, Guan and Lam [49] developed an XDCL method to construct effective signals for cross-domain contrast by utilizing the spectral and spatial domains, respectively. Another way is to introduce new feature extraction modules in the existing CL framework to improve the effectiveness of the model. Zhu et al. [50] designed a lightweight feature extraction network, including multiple plug-and-play efficient asymmetric dilated convolution blocks, which can reduce the high-computational cost and solve the underutilization of pixel-level multiscale context information for feature learning. To take full advantage of the multiscale semantic representation of images, Huang et al. [51] added a 3-D-SwinT-based multiscale local CL module for hierarchical learning, and their method can extract pixel-level representations. In addition, some research works have investigated the possibility of combining deep subspace clustering methods with CL. For example, Cai et al. [52] presented a neighborhood contrastive subspace clustering network that can greatly improve the consistency between positive samples in the subspace, and it proved to be a scalable and robust method for HSI unsupervised classification.

C. Cross-Modal CL for HSI Classification

With the rapid development of CL methods, some researchers have extended the idea of CL methods to process the multimodal RS data, and the cross-modal CL methods may have great potential in the application of HSI classification. Hang et al. [53] designed a cross-modal contrastive loss to explore the semantic and structural information between HSI and LiDAR data, and they utilized a dual fine-tuning strategy to transfer the features of small samples into the reused branch of HSI for classification. Except for the abovementioned study, we have not found any other studies of cross-modal CL-based methods for the fusion of HSI and LiDAR data.

In this article, we also develop a cross-modal CL-based approach for HSI and LiDAR data classification. However, there are some differences between our work and the study by Hang et al. as follows. First, the pretext task is different in the cross-modal CL-based method. The study by Hang et al. is based on the contrast task of positive samples and negative samples to directly maximize the attribution relationship between representations of HSI and LiDAR data, which ignores the local details. Our approach adopts the prediction task to transform and match the two-modal features and focuses on the local feature matching. Second, our method removes the need for a large number of negative samples, since the negative samples constructed from RS data with the high similarity of spatial textures of adjacent regions may interfere with the feature learning of positive samples. Meanwhile, the collaborative strategy combined with stop-gradient operation is adopted to avoid model collapse by complementary feature learning and adjusting network parameters between HSI and LiDAR data. Third, our feature extraction module is independent and works by collaboratively processing the information of HSI and LiDAR data, which can preserve the intermodel and intramodal similarity structures. In [53], the PCA technique was applied for the HSI dimensionality reduction process to ensure the same dimension as the LiDAR data, but the operation of PCA may lead to the loss of useful information. At last, our fine-tuning strategy is different from that of [53]. We design a cross-channel-based multilevel network that can further fuse multimodal features and strengthen the correlation of composite information between HSI and LiDAR data, while the study by Hang et al. only focuses on the potential information in hyperspectral data.

III. Methodology

In this section, our proposed CCL approach will be introduced, which takes full advantage of CCL in the pretraining (CCL-PT) stage and the fine-tuning (CCL-FT) stage for HSI and LiDAR data fusion and classification. For the clarity and readability, all the abbreviations of the CCL approach are explained in the Nomenclature. The framework of CCL method is shown in Fig. 1.

A. Pretraining for Collaborative Features

CL can learn some features of data without labeled samples, which is beneficial to solve the classification problem of insufficient labeled samples in the RS data. In this article, HSI and LiDAR data covering the same area of the Earth's surface (the paired data) are used to directly construct positive sample pairs as the training samples for feature learning in CL, instead of the way of data augmentation. Then, a collaborative strategy is introduced into CL for the two-modal RS data, which can more effectively learn the features than the shared weights in the feature extraction module. Specifically, CCL-PT maps HSI and LiDAR data to different vector representation space by independent feature extraction network to learn coordinated representations and adopts the constraint of similarity measure to collaborate the correlation of the two-modal RS data in the CCL-PT stage. Therefore, the information of independence and complementarity between the features of HSI and LiDAR data can be coordinated in the feature space, and the learned feature extraction module can be transferred for subsequent multilevel fusion in the CCL-FT stage.

Here is a paired HSI $\boldsymbol{X}_h \in \mathbb{R}^{H\times W\times B}$ and LiDAR image $\boldsymbol{X}_l \in \mathbb{R}^{H\times W}$, where H and W represent the height and width of the two images in the spatial dimension and B represents the number of spectral bands of HSI. We perform CCL-PT based on the prediction pretext task for the paired HSI and LiDAR data (see Fig. 1). First, the positive sample pairs are constructed based on the target pixels. Taking each target pixel as the center, several adjacent pixels are selected to form the HSI patch cube $\boldsymbol{x}_h \in \mathbb{R}^{P\times P\times B}$ and LiDAR patch $\boldsymbol{x}_l \in \mathbb{R}^{P\times P}$, respectively, where P represents the neighborhood size. In our experiments, P is set to 9, which can retain enough spatial context information of the data and reduce the amount of calculation. Then, the constructed positive sample pairs $\{(\boldsymbol{x}_h^{(i)}, \boldsymbol{x}_l^{(i)})|i = 1, 2, 3, \ldots, T\}$, where i refers to the selected patch according to the ith target pixel and T refers to the number of training samples, are processed by two encoders

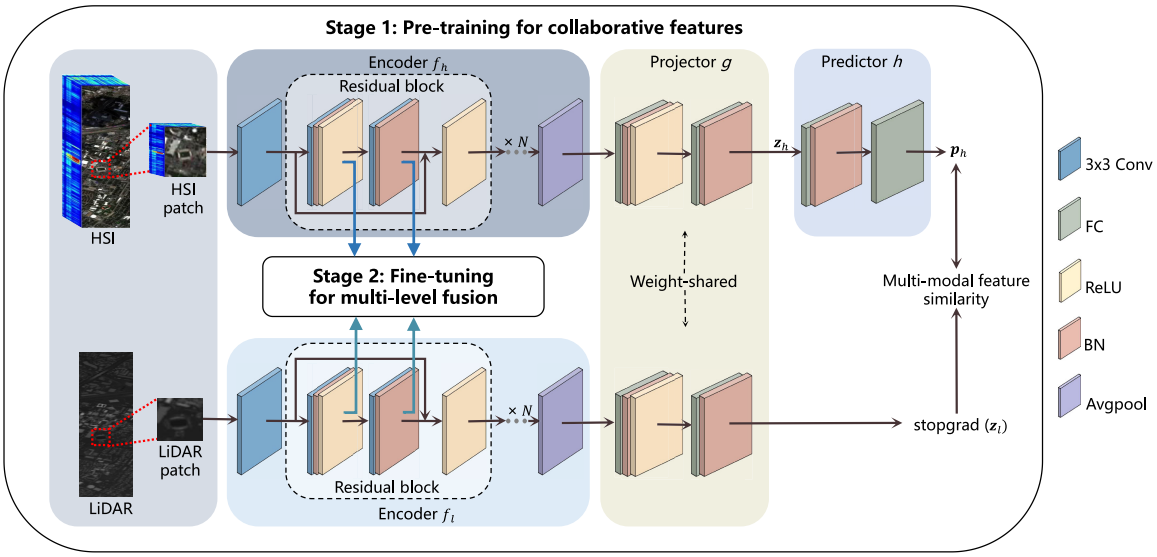

Fig. 1. Flowchart of the proposed CCL method.

of the feature extraction module, a projector of the projection module and a predictor of the prediction module. It is worth noting that we use the collaborative strategy to make the feature extraction module no longer share weights to better learn the features of HSI and LiDAR data.

According to our previous experience, using complex networks to train HSI and LiDAR data easily leads to overfitting due to the strong correlation between pixels of HSI. So, our encoder adopts a relatively simple network structure to extract discriminative features. The two encoders, denoted as f_h and f_l for HSI and LiDAR data, have the same network structure that consists of one convolutional layer, a stack of $N = 2$ residual blocks, and one average pooling layer. In detail, the inputs of HSI patch cube $\boldsymbol{x}_h$ and LiDAR patch $\boldsymbol{x}_l$ are first preprocessed by the convolutional layer to obtain initial feature representations with the same dimensions. High-quality features are extracted from initial feature representations in the residual blocks, and the redundant information is removed through the average pooling layer to reduce the amount of parameters. Each residual block consists of two cascaded convolutions, and the number of channels is 64. The cascaded convolutions use skip connections for identity mapping and obtain nonlinear outputs by activation function ReLU, which can effectively train network parameters via deep gradient propagation to upper layers. The module structures of projector and predictor are both composed of multilayer perceptron (MLP), denoted as g and h, respectively. The projector includes FC-BN-ReLU and FC-BN, and it recompresses collaborative features into another potential feature space to enhance the invariance of data transformation and learn higher-order features. The predictor consists of FC-BN and FC, which is adopted for feature matching of HSI and LiDAR data. The final two output vectors are denoted as $\boldsymbol{p}_h$ and $\boldsymbol{z}_l$ and represented as follows:

$$\boldsymbol{p}_h \triangleq h(g(f_h(\boldsymbol{x}_h))) \tag{1}$$

$$\boldsymbol{z}_l \triangleq g(f_l(\boldsymbol{x}_l)). \tag{2}$$

The loss function of CCL-PT is calculated by minimizing the negative value of the cosine similarity and can be expressed by the following equation:

$$D(\boldsymbol{p}_h, \boldsymbol{z}_l) = -\frac{\boldsymbol{p}_h}{\|\boldsymbol{p}_h\|_2} \cdot \frac{\boldsymbol{z}_l}{\|\boldsymbol{z}_l\|_2} \tag{3}$$

where $\|\cdot\|_2$ is the $\boldsymbol{l}_2$-norm and a distance metric function, which is equivalent to the mean-squared error of the vector elements.

To balance the respective influences of two-modal RS data, we define a symmetric loss for the patches generated by target pixels of HSI and LiDAR data

$$L_{\text{cl}} = \frac{1}{2}D(\boldsymbol{p}_h, \boldsymbol{z}_l) + \frac{1}{2}D(\boldsymbol{p}_l, \boldsymbol{z}_h). \tag{4}$$

In addition, we also adopt stop-gradient operation to avoid model collapse without negative samples and to efficiently learn eigenfeatures of two-modal RS data in CCL-PT. When features of HSI are used to predict LiDAR data features, the gradient backpropagation is not executed on the network branch of LiDAR data, and the output vector stopgrad ($\boldsymbol{z}_l$) can be denoted as a constant C_l. Similarly, when features of LiDAR data predict HSI features, the network branch of HSI does not perform gradient backpropagation either, and the output vector stopgrad($\boldsymbol{z}_h$) is denoted as a constant C_h. The formula for calculating the contrastive loss of each patch is as (5). The total loss is the sum of the losses of all patches generated by target pixels of HSI and LiDAR data in the pretraining stage

$$L_{\text{stopgrad}} = \frac{1}{2}D(\boldsymbol{p}_h, C_l) + \frac{1}{2}D(\boldsymbol{p}_l, C_h). \tag{5}$$

B. Fine-Tuning for Multilevel Fusion

In the CCL-FT stage, we design a multilevel fusion network structure and select a small proportion of labeled samples (ten for each class) as the training set for supervised learning to perform multilevel concatenation and cross-channel fusion on the existing unsupervised features extracted in the

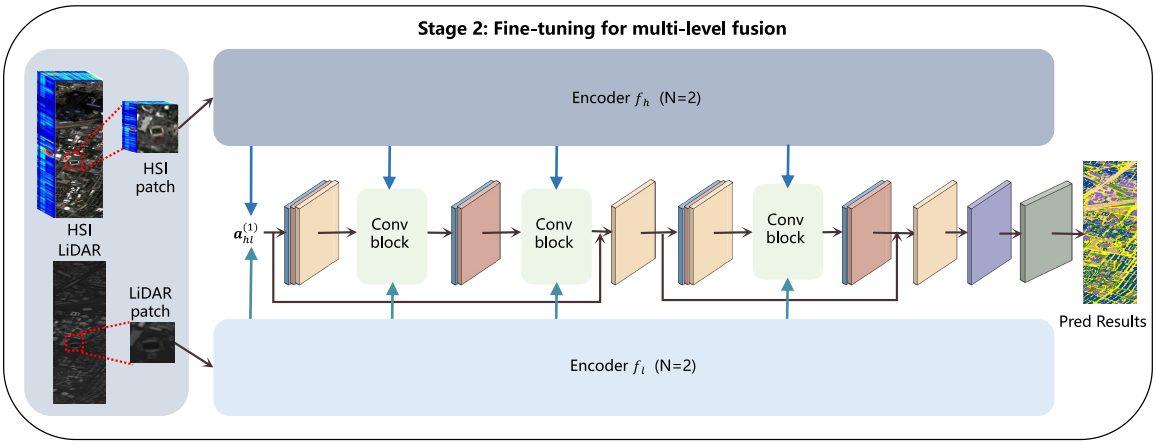

Fig. 2. Structure of the multilevel fusion network in the CCL-FT stage.

CCL-PT stage. CCL-FT can make full use of complementarity information of ground objects contained in multiattribute features of HSI and LiDAR data and combine those details with information of labeled samples for final classification. The network framework of multilevel fusion is denoted as f_{mlf} and shown in Fig. 2.

First, the feature extraction module learned in the CCL-PT stage is transferred to the multilevel fusion network for feature extraction of labeled samples in the CCL-FT stage. The network branches of HSI and LiDAR data can learn four features of different depths from the stack of $N = 2$ residual blocks f_h and f_l in the transferred feature extraction module, respectively, including two shallow features and two deep features, which preserve independence and complementarity of the two-modal RS data. Then, multilevel concatenation is carried out to generate four new fused features based on the abovementioned features. The formula of the new concatenated feature is as follows:

$$\boldsymbol{a}_{\text{hl}}^{(j)} = \boldsymbol{a}_{h}^{(j)} \oplus \boldsymbol{a}_{l}^{(j)}, \quad j = 1, 2, 3, 4 \tag{6}$$

where $\oplus$ represents the concatenation along the channel axis and j represents the feature learned from the jth depth.

Second, we fine-tune these new concatenated features based on cross channel via the multilevel fusion network to deepen the semantic correlation between new features. Therefore, we insert three new convolution blocks Conv-BN-ReLu in the fusion network on the basis of the feature extraction module in the CCL-PT stage, which can change the dimension of the concatenated features of HSI and LiDAR data to adapt the number of channels of the following convolution layers. The new convolution blocks are denoted as q and shown in Fig. 3. The output vector $\boldsymbol{r}_{\text{hl}}$ can be expressed as follows:

$$\boldsymbol{r}_{\text{hl}}^{(j)} = q\left(\boldsymbol{m}_{\text{hl}}^{(j-1)} \oplus \boldsymbol{a}_{\text{hl}}^{(j)}\right), \quad j = 2, 3, 4 \tag{7}$$

where $\oplus$ represents the concatenation along the channel axis, j represents the feature learned from the jth depth, and $\boldsymbol{m}_{\text{hl}}$ here represents the features of different depths extracted by the fusion network.

Specifically, the multilevel fusion process is realized via a three-step cross-channel fusion, which can fuse and map

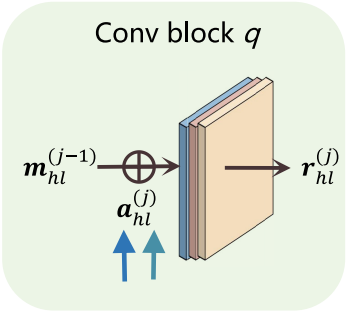

Fig. 3. Structure of convolution block.

the collaborative features of HSI and LiDAR data to the hidden feature space relevant to the classification task. First, the new concatenated feature $\boldsymbol{a}_{\text{hl}}^{(1)}$ is used as the initial input of the multilevel fusion network, and the feature representation $\boldsymbol{m}_{\text{hl}}^{(1)}$ is obtained after the first depth convolution layer. Then, in the three new convolution blocks, the feature representation obtained by the previous depth convolution layer and the new concatenated feature of the current depth convolution layer, such as $\boldsymbol{m}_{\text{hl}}^{(1)}$ and $\boldsymbol{a}_{\text{hl}}^{(2)}$ in the second depth layer, are reduced dimensionally and fused in the hidden feature space to obtain the current depth feature representation, which is used for the input of the next depth convolution layer. The output vectors of different depth convolution layers in the multilevel fusion network based on cross channel are expressed by the following formula:

$$\boldsymbol{m}_{\text{hl}}^{(j)} = \begin{cases} f_{\text{mlf}}\left(\theta_j, \boldsymbol{a}_{\text{hl}}^{(j)}\right), & j = 1 \\ f_{\text{mlf}}\left(\theta_j, \boldsymbol{r}_{\text{hl}}^{(j)}\right), & j = 2, 3, 4 \end{cases} \tag{8}$$

where θ refers to the parameters of convolution layer at different depths in the multilevel fusion network.

In order to speed up the training process of the multilevel fusion network, skip connections are applied to assist gradient backpropagation. After the three-step cross-channel fusion, the obtained feature representations are compressed through an average pooling layer to integrate global information. Finally, the integrated features of HSI and LiDAR data are output to a fully connected layer for classification. The loss function of

TABLE I
NUMBER OF SAMPLES FOR TRAINING AND TEST OF EACH CLASS ON THE TRENTO DATASET

Class No.	Categories	No. of Samples	Training	Test
C1	Apple trees	4034	10	4024
C2	Buildings	2903	10	2893
C3	Ground	479	10	469
C4	Woods	9123	10	9113
C5	Vineyard	10501	10	10491
C6	Roads	3374	10	3364
Total		30414	60	30354

the classification process is as follows:

$$L_{\text{mlf}} = -\log\left(\frac{\exp(\hat{y}[y])}{\sum_i \exp(\hat{y}[i])}\right) \tag{9}$$

where $\hat{y}$ is the predicted value and y is the ground truth.

IV. EXPERIMENTS

A. Dataset Description

In this article, three widely used HSI and LiDAR datasets are adopted to conduct a large number of experiments to demonstrate the effectiveness and robustness of our proposed CCL approach. The three datasets are described below.

1) Trento Data: The first dataset is collected in the area of southern Trento, Italy. There are six land cover classes and a total of 30 414 labeled samples. The HSI data consist of 63 spectral bands ranging from 400 to 980 nm. Each band is 600 × 166 pixels with a spatial resolution of 1.0 m [54]. Likewise, the LiDAR data have the same spatial size but only one band. The detailed information is given in Table I and Fig. 4.

2) Houston2013 Data: The second dataset is collected over the University of Houston campus and its neighboring areas. There are 15 land cover classes with a total of 15 029 labeled samples. The HSI data contain 144 spectral bands that range from 380 to 1050 nm. Each band is 349 × 1905 pixels with a spatial resolution of 2.5 m [55]. Similarly, the size of corresponding LiDAR data is also 349 × 1905, including the height information of the surface material. The details are described in Table II and Fig. 5.

3) MUUFL Gulfport Data: The third dataset is collected at the Gulf Park Campus of the University of Southern Mississippi. There are 11 land cover classes and 53 687 labeled samples. The original HSI data suffer from severe noise and contain a region of invalid data. Therefore, eight spectral bands are removed, and the original HSI is cropped in the spatial extent as a new dataset. The new HSI data consist of 64 spectral bands ranging from 380 to 1080 nm. Each band is 325 × 220 pixels with a spatial resolution of 1.0 m [56], [57]. Meanwhile, the LiDAR data have the same spatial size and resolution. The details are presented in Table III and Fig. 6.

B. Experimental Setup

To demonstrate the superiority of our proposed CCL approach for classification of HSI and LiDAR data, we carry out a series of compared experiments, including two single-modal models of CNN-HSI [58] and 3DVSCNN [59], three state-of-the-art fusion networks of CoupledCNNs [33], EndNet [36], and FusAtNet [37], and three models fused at different times of Early-Fusion, Middle-Fusion, and Late-Fusion. To ensure the fairness for all experiments, we use the LiDAR images with the same band. So, the attribute profiles are not applied to expand the number of bands of LiDAR images in EndNet, Early-Fusion, Middle-Fusion, and Late-Fusion models as described in the original papers. The comparison methods are described below.

1) *CNN-HSI:* The 2-D convolution network was adopted to jointly extract spectral and spatial features to avoid information loss after the use of 3-D convolution and PCA [58].

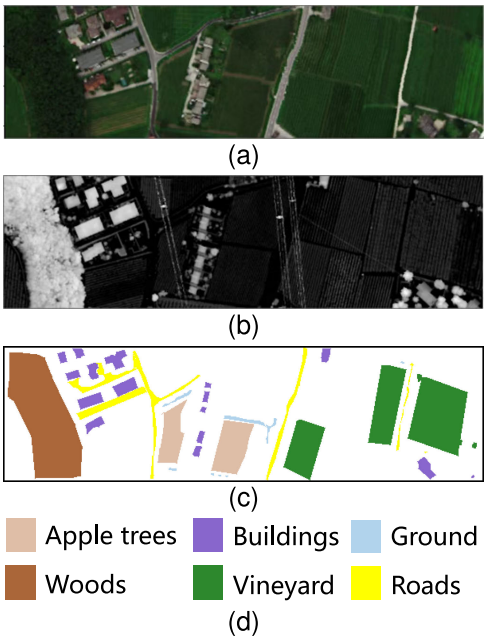

Fig. 4. Visualization of the Trento dataset. (a) Pseudo-color image of HSI. (b) Gray-scale image of LiDAR data. (c) Ground-truth map. (d) Legend.

TABLE II
NUMBER OF SAMPLES FOR TRAINING AND TEST OF EACH CLASS ON THE HOUSTON2013 DATASET

Class No.	Categories	No. of Samples	Training	Test
C1	Health grass	1251	10	1241
C2	Stressed grass	1254	10	1244
C3	Synthetic grass	697	10	687
C4	Trees	1244	10	1234
C5	Soil	1242	10	1232
C6	Water	325	10	315
C7	Residential	1268	10	1258
C8	Commercial	1244	10	1234
C9	Road	1252	10	1242
C10	Highway	1227	10	1217
C11	Railway	1235	10	1225
C12	Parking lot 1	1233	10	1223
C13	Parking lot 2	469	10	459
C14	Tennis court	428	10	418
C15	Running track	660	10	650
Total		15029	150	14879

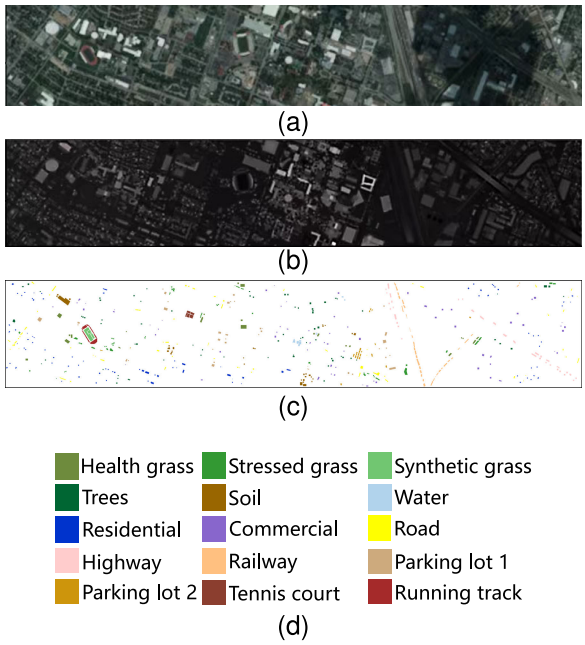

Fig. 5. Visualization of the Houston2013 dataset. (a) Pseudo-color image of HSI. (b) Gray-scale image of LiDAR data. (c) Ground-truth map. (d) Legend.

TABLE III
NUMBER OF SAMPLES FOR TRAINING AND TEST OF EACH CLASS ON THE MUUFL GULFPORT DATASET

Class No.	Categories	No. of Samples	Training	Test
C1	Trees	23246	10	23236
C2	Grass pure	4270	10	4260
C3	Grass groundsurface	6882	10	6872
C4	Dirt and sand	1826	10	1816
C5	Road Materials	6687	10	6677
C6	Water	466	10	456
C7	Building's shadow	2233	10	2223
C8	Buildings	6240	10	6230
C9	Sidewalk	1385	10	1375
C10	Yellow curb	183	10	173
C11	Cloth panels	269	10	259
	Total	53687	110	53577

2) *3DVSCNN:* The 3-D convolutional network was designed based on the idea of active learning to extract spatial–spectral features for HSI classification [59].
3) *CoupledCNNs:* The coupled dual-branch CNN was developed to fully integrate the heterogeneous features of HSI and LiDAR data for classification [33].
4) *EndNet:* The deep encoder–decoder architecture was applied to reconstruct the multimodal inputs and fuse the multimodal features for HSI and LiDAR classification [36].
5) *FusAtNet:* The LiDAR-derived attention map was used to emphasize the spatial features of HSI, and the spatial–spectral information was extracted for classification task by the attention mechanism [37].
6) *Early-Fusion, Middle-Fusion, and Late-Fusion:* The output multimodal features were concatenated and fused at the different stage of the early, middle, and late stages in the networks for classification [60].

We also investigate the contribution of four innovative modules in our CCL approach to the classification accuracy,

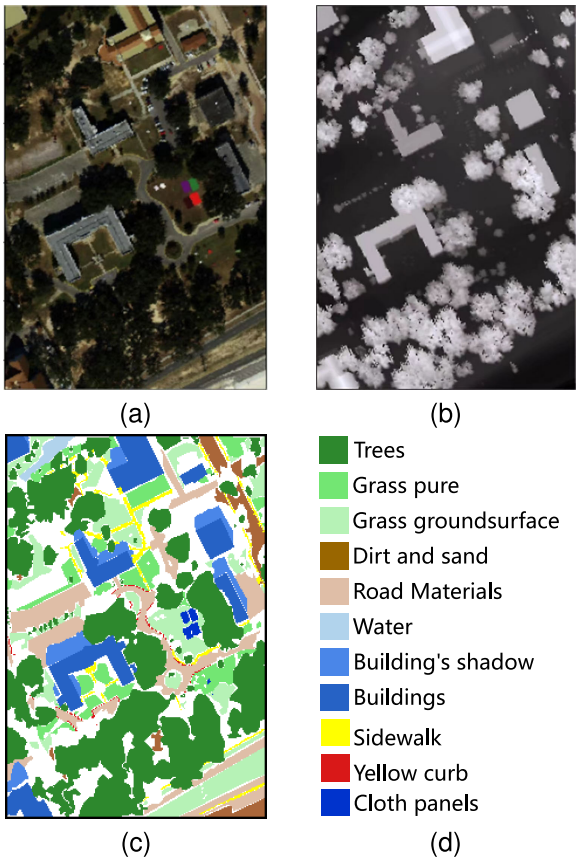

Fig. 6. Visualization of the MUUFL Gulfport dataset. (a) Pseudo-color image of HSI. (b) Gray-scale image of LiDAR. (c) Ground-truth map. (d) Legend.

which includes Single-Encoder-HSI, Single-Encoder-LiDAR, CCL-FT, and CCL-PT. Three ablation experiments are performed on the real-world datasets, and the detailed description is as follows.

1) *Single-Encoder-HSI and Single-Encoder-LiDAR:* The single-modal encoder is adopted based on ResNet with a shallow network depth to extract features of HSI and LiDAR data, respectively. This experiment does not include the modules of CCL-FT and CCL-PT.
2) *CCL-FT:* The modules of Single-Encoder-HSI and Single-Encoder-LiDAR are used for feature extraction, and our designed network of the cross-channel-based multilevel fusion is added for fusion and classification. This experiment does not include the module of CCL-FT.
3) *CCL:* Our proposed CCL method includes two stages of the CCL-PT and the CCL-FT, which is used to compare with CCL-FT to evaluate the contribution of the CCL-PT module.

The aim of this article is to address the problem of insufficient labeled samples for HSI and LiDAR data classification. In the proposed CCL approach, 80% of the unlabeled samples in each dataset are used for collaborative feature extraction in the pretraining (CCL-PT) stage, and ten labeled samples are randomly selected from each class as the training set for multilevel fusion in the fine-tuning (CCL-FT) stage. Specifically, there are 60, 150, and 110 labeled samples

TABLE IV
CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS ON THE TRENTO DATASET. ALL THE BEST ACCURACIES ARE IN BOLD

Class No.	CHH-HSI	3DVSCNN	CoupledCNNs	EndNet	FusAtNet	Early-Fusion	Middle-Fusion	Late-Fusion	CCL
C1	95.79	80.14	98.49	60.35	97.09	98.75	99.28	99.28	**99.49**
C2	86.55	60.50	95.04	94.68	95.19	95.44	**98.35**	97.21	97.60
C3	97.18	69.59	95.48	86.31	90.49	94.07	95.88	91.66	**97.53**
C4	98.43	98.27	99.74	98.68	99.59	99.82	99.97	99.91	**100.00**
C5	96.78	87.63	**99.91**	85.05	99.06	99.28	99.45	99.54	99.77
C6	83.14	72.63	93.06	86.42	91.88	91.35	92.82	90.85	**96.08**
OA(%)	94.74	85.39	98.41	86.94	97.70	98.09	98.73	98.36	**99.17**
AA(%)	92.98	78.13	96.95	85.25	95.55	96.45	97.63	96.41	**98.41**
Kappa	0.9299	0.8064	0.9788	0.8255	96.9300	0.9746	0.9831	0.9781	**0.9890**

TABLE V
CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS ON THE HOUSTON2013 DATASET. ALL THE BEST ACCURACIES ARE IN BOLD

Class No.	CNN-HSI	3DVSCNN	CoupledCNNs	EndNet	FusAtNet	Early-Fuison	Middle-Fusion	Late-Fusion	CCL
C1	**91.56**	89.49	87.70	86.16	81.50	91.31	87.89	89.39	89.61
C2	92.53	86.98	92.11	89.08	81.25	**93.14**	90.68	90.01	90.13
C3	88.69	89.94	96.89	99.80	95.88	97.58	**99.80**	99.04	99.72
C4	94.10	91.41	97.74	92.03	92.02	94.64	96.56	94.95	**98.68**
C5	99.13	99.03	99.20	96.85	89.04	99.40	99.19	97.30	**99.58**
C6	89.51	86.32	91.14	87.24	75.59	85.40	86.67	84.41	**92.03**
C7	69.12	71.43	86.88	80.13	72.67	84.72	85.27	80.98	**94.90**
C8	55.54	57.36	73.87	70.62	65.43	70.40	73.63	74.16	**88.06**
C9	72.75	75.07	75.47	72.62	61.52	73.78	70.30	70.10	**83.47**
C10	80.52	70.66	82.70	64.24	59.33	71.03	70.85	71.69	**90.54**
C11	71.13	67.61	90.54	77.01	72.11	74.66	82.06	81.65	**90.98**
C12	64.06	70.26	79.34	66.65	68.14	79.99	77.60	72.04	**83.65**
C13	80.45	83.27	95.40	46.80	79.46	92.48	93.88	93.33	**96.23**
C14	97.41	90.74	**99.67**	97.37	96.94	98.30	99.19	97.56	98.78
C15	**99.89**	96.12	99.75	98.32	92.18	98.42	99.88	98.86	99.55
OA(%)	81.23	80.02	88.35	80.98	76.93	85.42	85.76	84.57	**92.15**
AA(%)	83.09	81.71	89.89	81.66	78.87	87.02	87.58	86.36	**93.06**
Kappa	0.7973	0.7841	0.8741	0.7944	0.7508	0.8423	0.8461	0.8332	**0.9151**

selected for Trento, Houston2013, and MUUFL Gulfport datasets, respectively. With regard to parameter settings, we refer to the SimSiam framework to set the CCL-PT parameters. The stochastic gradient descent (SGD) is adopted for pretraining, and the batch size is 512. An Adam optimizer is employed for optimization in the CCL-FT stage. Three common metrics of overall accuracy (OA), average accuracy (AA), and Kappa coefficient are calculated to evaluate the performance of different approaches for classification. Moreover, to avoid the influence of accidental factors, training samples are randomly selected ten times from each dataset, and the classification accuracy is obtained by the average of ten experimental results. All experiments are implemented on NVIDIA Tesla P100, 16-GB memory, and four GPUs.

C. Experimental Results and Analysis

1) Classification Results of Compared Methods: The classification accuracy of different methods is summarized in Tables IV–VI, and the classification maps are shown in Figs. 7–9. The detailed analysis of the compared results is as follows.

a) Trento Dataset: Table IV shows the classification results of different methods on the Trento dataset, and it can be noticed that all methods achieve good classification results, because the classes of ground objects in the Trento dataset are highly different and discriminative, and whether it is a single-modal method or a multimodal method that can extract features containing rich information of ground objects. The best classification result obtained by the single-modal methods is CNN-HSI, and the accuracy can reach 94.74%. Except for EndNet, the classification accuracy of the multimodal methods is above 97.7%, which is 2.96% higher than that of CNN-HSI. This demonstrates that the complementary information contained in multiattribute features contributes to the improvement of the classification performance in most cases. The reason why EndNet does not perform as well as CNN-HSI on this dataset may be that EndNet cannot adequately fuse the features reconstructed from HSI to LiDAR data with few labeled samples. Overall, the classification accuracy of our proposed CCL approach is higher than that of all other methods, especially on Wood (class 4), which can achieve a stable accuracy of 100%, and the OA, AA, and Kappa coefficient of our model are 99.17%, 98.41%, and 0.989, respectively.

b) Houston2013 Dataset: Table V displays the classification results of different methods on the Houston2013 dataset. In general, as performed on the Trento dataset, the classification accuracies of the methods with the multimodal input are higher than those with the single-modal input on the Houston2013 dataset, i.e., the OA and AA of 85.76% and 87.58% by Middle-Fusion and the corresponding accuracies of 81.23% and 83.09% by CNN-HSI. Our proposed CCL approach can learn the most discriminative features for almost every class, with the accuracies varying from 83.47% for class 9 to 99.58% for class 5. The OA, AA, and Kappa coefficient of our method can achieve 92.15%, 93.06%, and 0.9151, which

TABLE VI

CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS ON THE MUUFL GULFPORT DATASET. ALL THE BEST ACCURACIES ARE IN BOLD

Class No.	CNN-HSI	3DVSCNN	CoupledCNNs	EndNet	FusAtNet	Early-Fuison	Middle-Fusion	Late-Fusion	CCL
C1	72.78	73.21	81.49	77.88	73.49	80.16	79.66	80.52	**87.95**
C2	80.98	62.20	68.44	70.67	54.00	64.23	64.24	64.77	**74.25**
C3	67.69	49.37	56.59	**73.54**	51.76	66.94	65.97	64.57	62.64
C4	71.26	71.62	89.10	75.39	65.12	73.40	80.62	73.06	**90.41**
C5	75.99	62.79	78.61	**88.04**	83.80	82.92	83.31	78.32	80.09
C6	96.74	85.42	**100.00**	99.23	98.25	94.45	99.82	99.65	**100.00**
C7	**94.53**	63.47	89.35	89.38	72.69	92.15	92.46	90.55	81.25
C8	73.75	61.85	87.89	88.15	85.39	86.33	**88.29**	83.16	85.33
C9	51.74	39.85	47.41	**54.78**	39.39	47.69	49.80	45.59	47.92
C10	**79.84**	38.03	65.32	76.37	46.18	57.51	56.53	60.12	61.56
C11	73.61	81.31	76.25	70.32	74.98	86.22	85.91	85.83	**86.55**
OA(%)	73.84	65.37	77.43	79.15	70.76	77.77	78.06	76.65	**81.11**
AA(%)	76.26	62.65	76.40	**78.52**	67.73	75.64	76.97	75.10	78.00
Kappa	0.6748	0.5685	0.7133	0.7366	0.6337	0.7168	0.7217	0.7027	**0.7562**

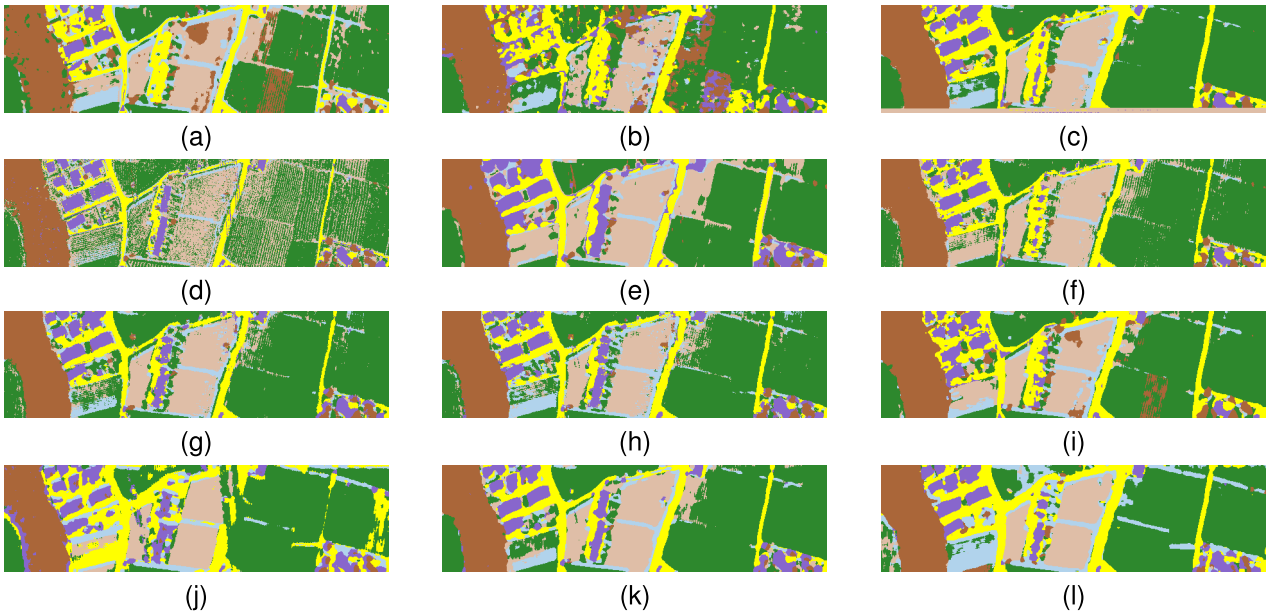

Fig. 7. Classification maps of different methods on the Trento dataset. (a) CNN-HSI (94.74%). (b) 3DVSCNN (85.39%). (c) CoupledCNNs (98.41%). (d) EndNet (86.94%). (e) FusAtNet (97.70%). (f) Early-Fusion (98.09%). (g) Middle-Fusion (98.73%). (h) Late-Fusion (98.36%). (i) Single-Encoder-HSI (96.93%). (j) Single-Encoder-LiDAR (93.00%). (k) CCL-FT (99.02%). (l) CCL (99.17%).

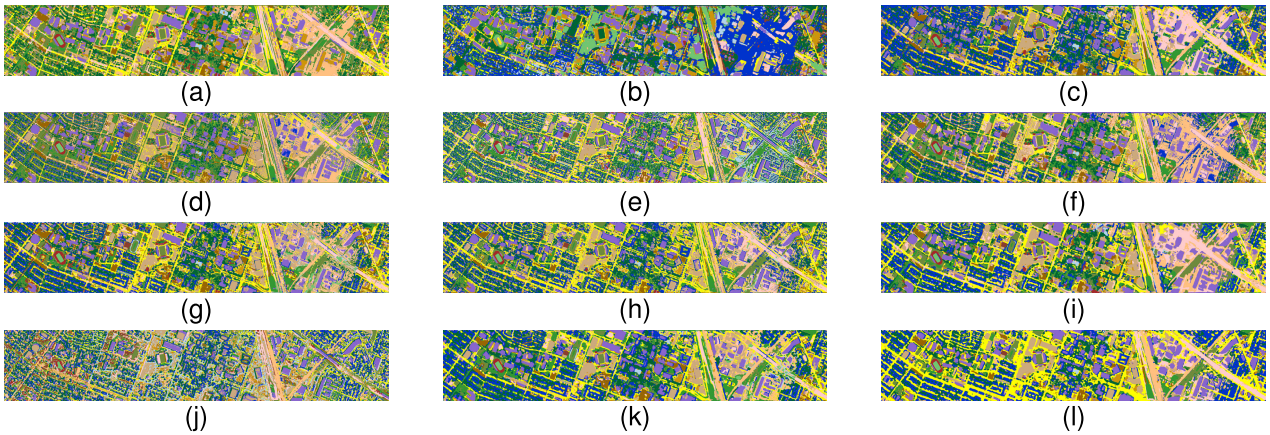

Fig. 8. Classification maps of different methods on the Houston2013 dataset. (a) CNN-HSI (81.23%). (b) 3DVSCNN (80.02%). (c) CoupledCNNs (88.35%). (d) EndNet (80.98%). (e) FusAtNet (76.93%). (f) Early-Fusion (85.42%). (g) Middle-Fusion (85.76%). (h) Late-Fusion (84.57%). (i) Single-Encoder-HSI (85.66%). (j) Single-Encoder-LiDAR (50.97%). (k) CCL-FT (88.61%). (l) CCL (92.15%).

are improved by 3.80%, 3.17%, and 0.041 compared with those of the typical and widely used method of CoupledCNNs (the OA, AA, and Kappa coefficient of 88.35%, 89.89%, and 0.8741, respectively). Therefore, the multiattribute features extracted by CL can reduce the loss of information of HSI and LiDAR data, and the fused features by CCL contain richer

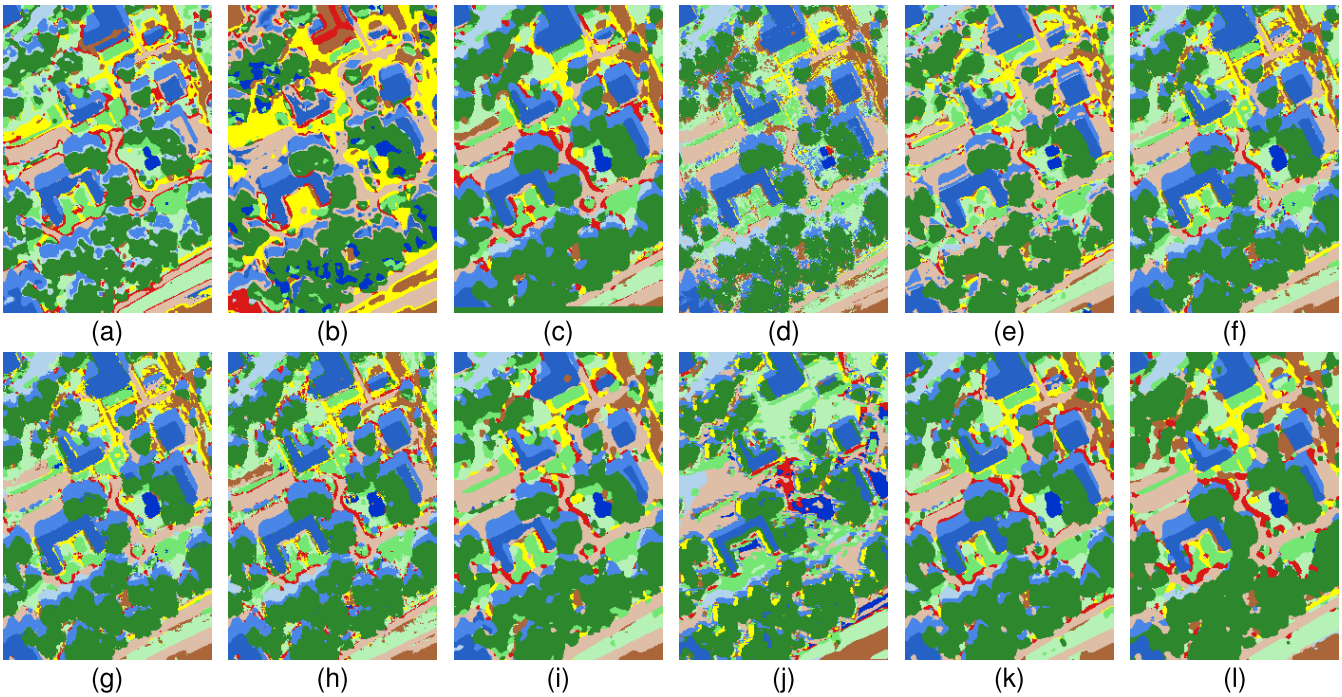

Fig. 9. Classification maps of different methods on the MUUFL Gulfport dataset. (a) CNN-HSI (73.84%). (b) 3DVSCNN (65.37%). (c) CoupledCNNs (77.43%). (d) EndNet (79.15%). (e) FusAtNet (70.76%). (f) Early-Fusion (77.77%). (g) Middle-Fusion (78.06%). (h) Late-Fusion (76.65%). (i) Single-Encoder-HSI (77.83%). (j) Single-Encoder-LiDAR (56.27%). (k) CCL-FT (79.00%). (l) CCL (81.11%).

and more accurate details of ground objects than other fusion results.

c) MUUFL Gulfport Dataset: Table VI lists the classification accuracies of different methods on the MUUFL Gulfport dataset. Due to the unbalanced distribution of sample classes in the MUUFL Gulfport dataset, the methods used for comparison in our experiments cannot well construct the nonlinear distribution of ground objects in the multiattribute feature space with limited labeled samples, resulting in poor classification accuracies. However, our proposed CCL approach can learn certain multiattribute features of nonlinear transformations through an MLP in the pretraining stage, and the multilevel fusion network can continuously optimize the learned feature representations and map them to the sample space effectively. The OA and Kappa coefficient of our method are 81.11% and 0.7562, which are achieved by an improvement of 1.96% and 0.0196 compared with EndNet. EndNet can also learn more discriminative features, such as on the mixed ground (class 3) and roads (class 5) in the MUUFL Gulfport dataset, and has a relatively high AA of 78.52%. In addition, from the results in Tables IV–VI, the classification accuracies of Middle-Fusion on the three real-world datasets are slightly higher than those of Early-Fusion and Late-Fusion. For example, the OA and AA of Middle-Fusion are 78.06% and 76.97% and the corresponding accuracies of 77.77% and 75.64% by Early-Fusion and 76.65% and 75.10% by Late-Fusion on the MUUFL Gulfport dataset. It may be because the features extracted by Early-Fusion tend to capture the spatial texture information, and the features in the Late-Fusion are too abstract in the case of insufficient labeled samples, while the Middle-Fusion can better balance the heterogeneity between features of HSI and LiDAR data.

Figs. 7–9 show the final classification results of different methods on three datasets. Compared with the false-color images of HSI, gray-scale images of LiDAR and ground-truth maps in Figs. 4–6, we can clearly find that our proposed CCL approach achieves better performance than the other methods, and the classification maps obtained by CCL are the most realistic. Taking the classification results of the Houston2013 dataset as an example, the methods, such as 3DVSCNN and EndNet, have produced many abnormal classification results, which do not correctly distinguish the boundaries of ground objects due to the impact factors, such as noise. As a result, the classification results are inconsistent with the real scenes. Our CCL approach can handle details more robustly to classify ground objects with the adaptive features learned in the pretraining stage, and the classification map can present a smoother visual effect.

2) Classification Results of Ablation Experiments: The classification accuracy of the ablation experiments is shown in Table VII, and the result analysis is detailed as follows.

The comparison results between Single-Encoder-HSI and Single-Encoder-LiDAR show that the Single-Encoder-HSI method can achieve better classification performance, i.e., the OA and AA of 96.93% and 94.77% on the Trento dataset and the corresponding accuracies of 93.00% and 87.25% by Single-Encoder-LiDAR, which indicates that rich spatial and spectral features from HSI are more beneficial to the classification of ground objects than the information from LiDAR data. However, compared with CCL-FT and CCL methods, the Single-Encoder-HSI method obtains a relatively low accuracy. Taking the Houston2013 dataset as an example, the OA of Single-Encoder-HSI is 85.66%, while the OAs of CCL-FT and CCL are 88.61% and 92.15%, respectively.

TABLE VII
ABLATION STUDIES ON DIFFERENT DATASETS. ALL THE BEST ACCURACIES ARE IN BOLD

Trento				
Class No.	Single-Encoder -HSI	Single-Encoder -LiDAR	CCL -FT	CCL
OA(%)	96.93	93.00	99.02	**99.17**
AA(%)	94.77	87.25	98.13	**98.41**
Kappa	0.9590	0.9066	0.9869	**0.9890**
Houston2013				
Class No.	Single-Encoder -HSI	Single-Encoder -LiDAR	CCL -FT	CCL
OA(%)	85.66	50.97	88.61	**92.15**
AA(%)	87.51	53.17	90.00	**93.06**
Kappa	0.8449	0.4731	0.8769	**0.9151**
MUUFL Gulfport				
Class No.	Single-Encoder -HSI	Single-Encoder -LiDAR	CCL -FT	CCL
OA(%)	77.83	56.27	79.00	**81.11**
AA(%)	73.91	40.60	75.95	**78.00**
Kappa	0.7170	0.4368	0.7309	**0.7562**

The results demonstrate that the features of LiDAR data, such as elevation, can be used as complementary information of HSI to improve HSI classification accuracy after fusion. CCL-FT has higher classification accuracies on the three datasets than Single-Encoder-HSI and Single-Encoder-LiDAR, i.e., the OAs of 79.00% and 56.27% of CCL-FT and Single-Encoder-LiDAR on the MUUFL Gulfport dataset. This shows that our multilevel fusion network based on cross channel can fully fuse the features provided by HSI and LiDAR data and reflect the effectiveness for classification tasks. After further adding CCL-PT on the basis of CCL-FT, our CCL method can learn more implicit and effective information of HSI and LiDAR data and achieves the best classification performance, especially on the Houseton2013 dataset with an improvement of OA by 3.54%.

In addition, the results can also verify the effectiveness of our designed feature extraction module by comparing the accuracies between Single-Encoder-HSI in Table VII and two single-modal methods of CHH-HSI and 3DVSCNN in Tables IV–VI. The OAs of Single-Encoder-HSI are 96.93%, 85.66%, and 77.83% on the Trento, Houseton2013, and MUUFL Gulfport datasets, and the corresponding accuracies are 94.74%, 81.23%, and 73.84% of the CHH-HSI method and 85.39%, 80.02%, and 65.37% of the 3DVSCNN method, respectively.

3) Visualization Maps of Extracted Features From HSI: We use t-distributed stochastic neighbor embedding (t-SNE) to visualize the feature distribution and investigate the divisibility between classes of ground objects by using the extracted features before and after the collaborative pretraining from HSI of the three datasets. Fig. 10 presents the feature distribution in the 2-D space. Specifically, the features in Fig. 10(a), (c), and (e) are generated by t-SNE from the original HSI high-dimensional features before the collaborative pretraining, while the features in Fig. 10(b), (d), and (f) are produced by t-SNE from features learned by the feature extraction module of the pretrained HSI branch. It can be clearly seen that the feature distribution in Fig. 10(a), (c), and (e) is more concentrated. However, the learned features after CCL in Fig. 10(b), (d), and (f) can improve the similarity of samples within a class and increase the difference of samples between classes. The results illustrate that the proposed CCL approach has a strong ability of feature extraction for HSI images, and the extracted features with large differences are more beneficial to downstream classification tasks.

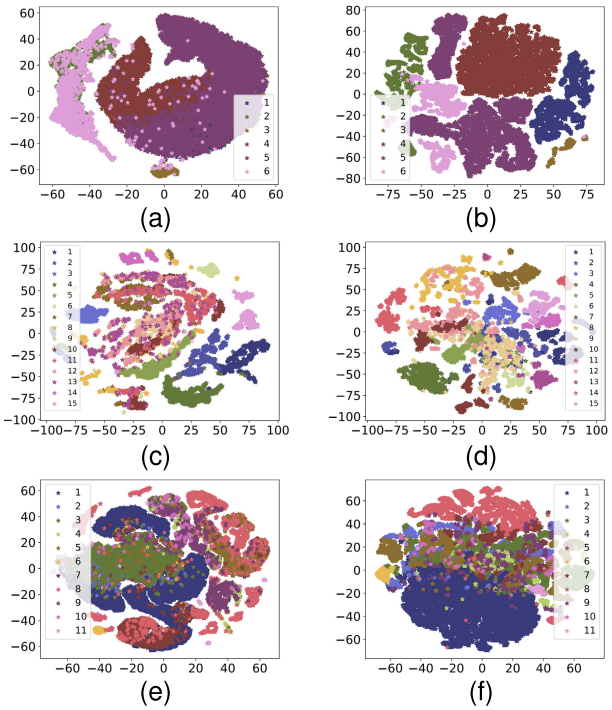

Fig. 10. Feature visualization results of encoders for HSI on three real-world datasets. On the Trento dataset (a) before pretraining and (b) after pretraining. On the Houston2013 dataset (c) before pretraining and (d) after pretraining. On the MUUFL Gulfport dataset (e) before pretraining and (f) after pretraining.

V. CONCLUSION

In this article, our proposed CCL approach can efficiently improve the classification accuracy of HSI and LiDAR data under limited labeled samples. In the pretraining (CCL-PT) stage, a collaborative strategy is introduced into CL to learn discriminative features from two-modal RS data without labeled samples. Then, we adopt the prediction pretext task to match the learned features to improve the complementarity of features and acquire the collaborative feature representations. Finally, these collaborative features are fused based on cross channel via our designed multilevel fusion network to obtain the integrated multiattribute fusion features in the fine-tuning (CCL-FT) stage. Also, a small number of labeled samples are used for the supervised training to enhance the semantic relevance of fusion features for classification. Extensive experimental results and visualized classification maps consistently demonstrate that our proposed CCL approach has excellent performance on the fusion of HSI and LiDAR data and can improve the classification accuracy on the three real-world datasets, especially on the dataset of Houston2013.

REFERENCES

[1] B. Rasti et al., "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, Dec. 2020.

[2] X. Kang, S. Li, and J. A. Benediktsson, "Spectral–spatial hyperspectral image classification with edge-preserving filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2666–2677, Jul. 2013.

[3] S. Jia, S. Jiang, Z. Lin, N. Li, M. Xu, and S. Yu, "A survey: Deep learning for hyperspectral image classification with few labeled samples," *Neurocomputing*, vol. 448, pp. 179–204, Aug. 2021.

[4] G. Licciardi, P. R. Marpu, J. Chanussot, and J. A. Benediktsson, "Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 447–451, May 2012.

[5] W. Li, S. Prasad, and J. E. Fowler, "Noise-adjusted subspace discriminant analysis for hyperspectral imagery classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 6, pp. 1374–1378, Nov. 2013.

[6] S. Jia, Z. Lin, B. Deng, J. Zhu, and Q. Li, "Cascade superpixel regularized Gabor feature fusion for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1638–1652, May 2020.

[7] Z. Ye, J. E. Fowler, and L. Bai, "Spatial–spectral hyperspectral classification using local binary patterns and Markov random fields," *J. Appl. Remote Sens.*, vol. 11, no. 3, Jul. 2017, Art. no. 035002.

[8] S. Jia et al., "A lightweight convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4150–4163, May 2020.

[9] L. He, J. Li, A. Plaza, and Y. Li, "Discriminative low-rank Gabor filtering for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1381–1395, Mar. 2017.

[10] S. Jia, S. Jiang, S. Zhang, M. Xu, and X. Jia, "Graph-in-graph convolutional network for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 20, 2022, doi: 10.1109/TNNLS.2022.3182715.

[11] J. Zhou, W. Sun, X. Meng, G. Yang, K. Ren, and J. Peng, "Generalized linear spectral mixing model for spatial–temporal–spectral fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5533216.

[12] K. Ren, W. Sun, X. Meng, G. Yang, J. Peng, and J. Huang, "A locally optimized model for hyperspectral and multispectral images fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.

[13] J. Zhang, X. Lin, and X. Ning, "SVM-based classification of segmented airborne LiDAR point clouds in urban areas," *Remote Sens.*, vol. 5, no. 8, pp. 3749–3775, Jul. 2013.

[14] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS J. Photogramm. Remote Sens.*, vol. 178, pp. 68–80, Aug. 2021.

[15] M. Dalponte, L. Bruzzone, and D. Gianelle, "Fusion of hyperspectral and LiDAR remote sensing data for classification of complex forest areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1416–1427, May 2008.

[16] H. Ghassemian, "A review of remote sensing image fusion methods," *Inf. Fusion*, vol. 32, pp. 75–89, Nov. 2016.

[17] 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.*, Jul. 2014, pp. 1241–1244.

[18] Q. Man, P. Dong, and H. Guo, "Pixel- and feature-level fusion of hyperspectral and LiDAR data for urban land-use classification," *Int. J. Remote Sens.*, vol. 36, no. 6, pp. 1618–1644, 2015.

[19] 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.

[20] L. Zhang and H. Shen, "Progress and future of remote sensing data fusion," *Int. J. Remote Sens.*, vol. 20, no. 5, pp. 1050–1061, 2016.

[21] M. Pedergnana, P. R. Marpu, M. D. 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.

[22] M. He, W. Chang, and S. Mei, "Advance in feature mining from hyperspectral remote sensing data," *Spacecraft Recovery Remote Sens.*, vol. 34, no. 1, pp. 1–12, 2013.

[23] W. Liao, R. Bellens, A. Pizurica, S. Gautama, and W. Philips, "Graph-based feature fusion of hyperspectral and LiDAR remote sensing data using morphological features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, vol. 7, Jul. 2013, pp. 4942–4945.

[24] M. Khodadadzadeh, J. Li, S. Prasad, and A. Plaza, "Fusion of hyperspectral and LiDAR remote sensing data using multiple feature learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2971–2983, Jun. 2015.

[25] 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.

[26] P. Ghamisi, J. A. Benediktsson, and S. Phinn, "Fusion of hyperspectral and LiDAR data in classification of urban areas," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Jul. 2014, pp. 181–184.

[27] 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.

[28] 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.

[29] F. Xiong, J. Zhou, S. Tao, J. Lu, and Y. Qian, "SNMF-Net: Learning a deep alternating neural network for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2021.

[30] F. Xiong, J. Zhou, S. Tao, J. Lu, J. Zhou, and Y. Qian, "SMDS-Net: Model guided spectral–spatial network for hyperspectral image denoising," *IEEE Trans. Image Process.*, vol. 31, pp. 5469–5483, 2022.

[31] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, "Deep fusion of remote sensing data for accurate classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1253–1257, Aug. 2017.

[32] 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.

[33] 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.

[34] H. Li, P. Ghamisi, U. Soergel, and X. X. Zhu, "Hyperspectral and LiDAR fusion using deep three-stream convolutional neural networks," *Remote Sens.*, vol. 10, no. 10, p. 1649, Oct. 2018.

[35] 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.

[36] 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, pp. 1–5, 2020.

[37] 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 (CVPRW)*, Jun. 2020, pp. 92–93.

[38] 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.

[39] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15750–15758.

[40] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[41] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[42] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[43] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 22243–22255.

[44] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9912–9924.

[45] J.-B. Grill et al., "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

[46] S. Hou, H. Shi, X. Cao, X. Zhang, and L. Jiao, "Hyperspectral imagery classification based on contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.

[47] L. Zhao, W. Luo, Q. Liao, S. Chen, and J. Wu, "Hyperspectral image classification with contrastive self-supervised learning under limited labeled samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[48] H. Lee and H. Kwon, "Self-supervised contrastive learning for cross-domain hyperspectral image representation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 3239–3243.
[49] P. Guan and E. Y. Lam, "Cross-domain contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528913.
[50] M. Zhu, J. Fan, Q. Yang, and T. Chen, "SC-EADNet: A self-supervised contrastive efficient asymmetric dilated network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5519517.
[51] X. Huang, M. Dong, J. Li, and X. Guo, "A 3-D-Swin transformer-based hierarchical contrastive learning method for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5411415.
[52] Y. Cai et al., "Superpixel contracted neighborhood contrastive subspace clustering network for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5530113.
[53] R. Hang, X. Qian, and Q. Liu, "Cross-modality contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528913.
[54] S. Jia, Z. Zhan, and M. Xu, "Shearlet-based structure-aware filtering for hyperspectral and LiDAR data classification," *J. Remote Sens.*, vol. 2021, Jan. 2021, Art. no. 9825415.
[55] 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.
[56] P. Gader, A. Zare, R. Close, J. Aitken, and G. Tuell, "Muufl Gulfport hyperspectral and LiDAR airborne data set," Dept. Elect. Comput. Eng., Univ. Florida, Gainesville, FL, USA, Tech. Rep., 2013-570, 2013.
[57] X. Du and A. Zare, "Technical report: Scene label ground truth map for MUUFL Gulfport data set," Dept. Elect. Comput. Eng., Univ. Florida, Gainesville, FL, USA, Tech. Rep., 20170417, 2017.
[58] S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing*, vol. 219, pp. 88–98, Jan. 2017.
[59] L. Hu, X. Luo, and Y. Wei, "Hyperspectral image classification of convolutional neural network combined with valuable samples," *J. Phys., Conf.*, vol. 1549, no. 5, Jun. 2020, Art. no. 052011.
[60] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, Aug. 2020.

Sen Jia (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the College of Computer Science, Zhejiang University, Hangzhou, China, in 2002 and 2007, respectively.

Since 2008, he has been with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, where he is currently a Full Professor. His research interests include hyperspectral image processing, signal and image processing, and machine learning.

Xi Zhou received the B.E. degree from Jiangxi Normal University, Nanchang, China, in 2021. She is currently pursuing the master's degree in computer science and technology with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

Her research interests include hyperspectral and light detection and ranging (LiDAR) classification and deep learning.

Shuguo Jiang received the B.E. degree in software engineering from the Xiamen University of Technology, Xiamen, China, in 2020. He is currently pursuing the master's degree in software engineering with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

His research interests include hyperspectral image classification, machine learning, and pattern recognition.

Ruyan He received the B.S. degree in geographic information system from Liaoning Technical University, Fuxin, China, in 2012, and the Ph.D. degree in photogrammetry and remote sensing from the China University of Mining and Technology at Beijing, Beijing, China, in 2019.

She was a Visiting Scholar with the University of California (UC) at Davis, Davis, CA, USA, from 2015 to 2017. She is currently an Associate Research Fellow with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. Her research interests include remote sensing image processing and deep learning.