Local Correntropy Matrix Representation for Hyperspectral Image Classification

Xinyu Zhang, Yantao Wei, Weijia Cao, Huang Yao, Jiangtao Peng, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*— The hyperspectral images (HSIs) classification technique has received widespread attention in the field of remote sensing. However, how to achieve satisfactory classification performance in the presence of a large amount of noise is still a problem worthy of consideration. In this article, a local correntropy matrix (LCEM)-based spatial–spectral feature representation method is proposed for HSI classification. Motivated by the successful application of information-theoretic learning (ITL), we propose to adopt correntropy matrix to represent the spatial–spectral features of HSI. Specifically, the dimension reduction is first performed on the original hyperspectral data. Then, for each pixel, we select its local neighbors within a sliding window using cosine distance for the construction of the LCEM. In this way, each pixel can be characterized as an LCEM. Finally, all the correntropy matrices are fed into a support vector machine (SVM) for final classification. In addition, we also propose a novel way to determine the size of the local window based on standard deviation. Because the LCEM as the feature descriptor can characterize discriminative spatial–spectral features, the proposed method has shown great interclass separability and intraclass compactness. Compared with other advanced approaches, the proposed LCEM method has achieved competitive performance in both evaluation indexes and visual effects, especially when the training size is very small.**

***Index Terms*— Correntropy matrix, feature extraction, hyperspectral image (HSI) classification.**

I. Introduction

With hundreds of continuous spectral bands, hyperspectral images (HSIs) contain large amounts of useful information, which can greatly enhance the reliability of remote sensing analysis [1]. Consequently, HSIs have been widely applied in many fields, such as military security [2], target detection [3], agriculture [4], and urban mapping [5]. As the basis for further applications, the HSI classification technique, which aims at assigning a certain label to each pixel, has attracted the broad interest of many researchers over the past few decades [6]. However, it is still challenging due to insufficient labeled samples, high dimensionality, and noise [7], [8].

Early studies mainly focused on the use of spectral information for classification, and these included matching spectral information-based methods [9], [10] and algorithms that used classifiers such as support vector machine (SVM) [11], random forest (RF) [12], and multinomial logistic regression (MLR) [13] to classify items of spectral information. However, they may suffer from the curse of dimensionality problem due to the high dimensionality of HSI [14]. Therefore, some dimension reduction strategies have also been used for HSI classification, such as principal component analysis (PCA) [15], locally linear embedding (LLE) [16], and maximum noise fraction (MNF) [17]. However, PCA only considers the global statistics, which usually fails to extract the local useful features of the HSI for classification [18]. Then, the spectrally segmented-PCA and the folded-PCA have also been proposed for better HSI feature extraction [19]. By maximizing the signal-to-noise ratio of the dimension reduced data, MNF generally outperforms PCA in HSI classification [20]. Nevertheless, although these spectral-based methods can classify labels by exploring spectral similarities, the classification performances are usually unsatisfactory because of the spectral variability caused by climate condition change and other environmental interferers [6].

It is well known that HSI is 3-D, whose spectral reflectance and spatial information are independent. Naturally, many attempts incorporating spatial information into HSI classification have been made, in which some successfully indicate spatial information is beneficial to improve the classification performance [21]. Therefore, a variety of spatial–spectral methods have been proposed [22]–[25]. For example, the extended morphological profiles (EMPs) method is put forward to adaptively extract spatial features in HSI [26]. In [27],

Manuscript received February 21, 2022; accepted March 10, 2022. Date of publication March 24, 2022; date of current version April 12, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42171351 and Grant 61502195, in part by the Natural Science Foundation of Hubei Province under Grant 2021CFA087 and Grant 2018CFB691, in part by the Self-Determined Research Funds of Central China Normal University (CCNU) From the Colleges' Basic Research and Operation of Ministry of Education (MOE) under Grant CCNU20TD005, and in part by the National Key Research and Development Program of China under Grant 2020YFA0714200. *(Corresponding authors: Yantao Wei; Weijia Cao; Huang Yao.)*

Xinyu Zhang is with the Hubei Research Center for Educational Informationization, Faulty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China, and also with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: zxy.one@mails.ccnu.edu.cn).

Yantao Wei and Huang Yao are with the Hubei Research Center for Educational Informationization, Faulty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China (e-mail: yantaowei@ccnu.edu.cn; yaohuang@mail.ccnu.edu.cn).

Weijia Cao is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, also with the Faculty of Science and Technology, University of Macau, Macau, China, also with the Yangtze Three Gorges Technology and Economy Development Company Ltd., Beijing 101100, China, and also with the Zhongke Langfang Institute of Spatial Information Applications, Langfang, Hebei 065001, China (e-mail: caowj@aircas.ac.cn).

Jiangtao Peng is with the Hubei Key Laboratory of Applied Mathematics, Faculty of Mathematics and Statistics, Hubei University, Wuhan, Hubei 430062, China (e-mail: pengjt1982@hubu.edu.cn).

Yicong Zhou is with the Faculty of Science and Technology, University of Macau, Macau, China (e-mail: yicongzhou@um.edu.mo).

Digital Object Identifier 10.1109/TGRS.2022.3162100

1558-0644 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Universidade de Macau. Downloaded on April 14,2022 at 00:44:45 UTC from IEEE Xplore. Restrictions apply.

a spectral–spatial classification framework is proposed to improve the performance achieved by pixel-wise classifier using edge-preserving filtering. Local binary patterns (LBPs) are also adopted to utilize the local contextual feature of HSI in [28]. To learn discriminative features of HSI, Zhou and Wei *et al.* [29] proposed a hierarchical spectral–spatial feature learning model to exploit robust features of HSI using multiscale adaptive weighted filtering. In addition, segmentation-based strategies, i.e., superpixel-based classification via multiple kernels (SCMK) [30] and multiscale superpixel segmentation [31], have also been introduced to obtain homogeneous area for spatial–spectral features extraction. In [32], the manifold learning-based spatial–spectral dimension reduction approach is developed for HSI classification. To make full use of local geometric structure and spatial correlation in HSI, a local linear spatial–spectral probabilistic distribution-based method has been proposed by Huang *et al.* [33].

In recent years, deep learning technology has achieved great success in diverse computer vision tasks, such as scene segmentation [34], image super-resolution [35], and object detection [36]. Motivated by these successful applications, some advanced deep neural network models are also employed to extract both linear and nonlinear features of HSI [37]–[39], i.e., deep brief network [40], convolutional neural network (CNN) [41], and graph convolutional network (GCN) [42]. It is a straightforward way to adopt 3-D CNN for HSI classification due to its capability of spatial–spectral feature learning. However, it suffers from high computational complexity and overfitting problem. To tackle this issue, works [43] and [44] have achieved competitive performance by the cooperation between 2-D CNN and 3-D CNN. In [39], a new regularization approach called neighboring region dropout based on 2-D CNN has been developed to further relieve the problem of overfitting. Nevertheless, though deep learning-based methods have led to a great breakthrough, a large amount of training data are usually required for robust classification performance. However, labeled samples are very limited in practice [8], [45]. The HSI classification is a pixel-level classification task, and it is tough to label an HSI that with hundreds of thousands of pixels. Besides, the labeled data generally need to be collected on the ground and in many cases requires experts to process. Thus, acquiring labeled hyperspectral data is always time-consuming and expensive. Consequently, the labeled samples are not sufficient in HSI classification [1].

In addition, the methods mentioned above do not consider the correlation between spectral bands, which can offer valuable information [46]–[49]. To exploit the discriminative spectral correlation features, Fang *et al.* [46] proposed a feature extraction method using local covariance matrix representation (LCMR), which presents each pixel as a covariance matrix. Taking the advantage of covariance matrix representation in representing data relationships, it achieved good classification performance. However, due to the high dimensionality and redundancy of hyperspectral signals and the interference of environmental factors during imaging, there is much noise in HSI and hyperspectral data are usually nonlinear; these hamper the classification performance of covariance matrix representation-based methods.

Inspired by the excellent performance of correntropy in dealing with nonlinear and non-Gaussian data [50], this article presents a new feature extraction method based on the local correntropy matrix (LCEM) representation to handle the problem. Recently, information-theoretic learning (ITL) has drawn great attention of researchers [51]–[53]. Correntropy, as a robust and simple similarity measurement [54], has been successfully applied on a variety of tasks. For instance, Peng and Du [55] incorporated correntropy into joint sparse representation to replace the traditional least-squares error. In [56], correntropy is also introduced to compute the affinity matrix for band selection. These works only adopt correntropy as a novel similarity measurement metric and have nothing to do with the category. In this article, we construct the LCEM for more discriminative spatial–spectral features representation. In another word, the LCEM can be regarded as the feature descriptor. In the correntropy matrix, each nondiagonal element represents the relationships of two different channels. Taking advantage of the kernel method, the correntropy in ITL is beneficial to characterize the nonlinear relationships in hyperspectral data and reduce the negative influence of noise. In summary, the main contributions of this article are given as follows.

1) This article proposes a new feature extraction approach based on the LCEM. We characterize each pixel in the image as an LCEM containing spatial–spectral information and then classify the matrices. The characterized features have shown great interclass separability and intraclass compactness. To the best of the author's knowledge, this is the first time to adopt the LCEM for the HSI feature representation and classification.
2) This article presents an adaptive way based on standard deviation to determine the size of the local window and the number of neighboring pixels. Furthermore, compared with other state-of-the-art methods, the proposed LCEM method has achieved competitive performance on three public datasets, especially when the training samples are very limited.

The remainder of this article is organized as follows. In Section II, some related works are briefly reviewed. Section III details the proposed LCEM method for HSI classification. Section IV shows the experimental results on three real HSI datasets and presents the results of the comparison with other advanced classification methods. The conclusion is given in Section V.

II. Related Work

A. Maximum Noise Fraction

In hyperspectral data processing, in order to reduce computational effort and data redundancy, data dimensionality reduction is usually performed before feature extraction. In order to remove noise during dimensionality reduction, this article uses the MNF method for dimensionality reduction of HSIs. The MNF method maximizes the signal-to-noise ratio of the dimensionality-reduced data by finding a linear transformation

Authorized licensed use limited to: Universidade de Macau. Downloaded on April 14,2022 at 00:44:45 UTC from IEEE Xplore. Restrictions apply.

matrix A, thus achieving dimensionality reduction. Let S and N denote the uncorrelated signal and noise in data D, respectively. Also, we assume $D = S + N$, and then,

$$\mathrm{Cov}(D) = \mathrm{Cov}(S) + \mathrm{Cov}(N) \tag{1}$$

where $\mathrm{Cov}(\cdot)$ presents the covariance matrix. The matrix A can be solved by this formula

$$\arg\max_A \frac{A^T \sum_{S^A}}{A^T \sum_{N^A}} = \arg\max_A \frac{A^T \sum_{D^A}}{A^T \sum_{N^A}} - 1. \tag{2}$$

The matrix A is the eigenvector corresponding to the L largest eigenvalues of $\mathrm{Cov}(N)^{-1}\mathrm{Cov}(D)$, where L is the number of components to be retained. Next, the reduced dimensional data Y can be obtained from the formula $Y = A^T D$.

Compared to the PCA which maximizes the variance of the projected data by orthogonal vectors, the MNF method not only extracts the principal components effectively but also maximizes the signal-to-noise ratio and further reduces the effect of noise.

B. Correntropy

The concept of corentropy was first proposed by Prof. J. Principe of the University of Florida and his team in 2006, inspired by the relevant content of ITL, which is a generalized correlation function [54]. With the joint efforts of experts and scholars in related fields, this theoretical system has been developed and widely applied in signal and information processing in recent years.

The definition of correlation entropy is as follows. If there are two random variables x and y, the correntropy between them can be defined as

$$V_\sigma(x, y) = E[k_\sigma(x, y)] = \int k_\sigma(x, y) dF_{X,Y}(x, y) \tag{3}$$

where $E[\cdot]$ is the expected function and $dF_{X,Y}(x, y)$ denotes the joint probability density function of (x, y). $k_\sigma(\cdot)$ is the kernel function, which commonly uses the Gaussian kernel function

$$k_\sigma(x, y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-y)^2}{\sqrt{2\sigma^2}}\right). \tag{4}$$

However, in practical applications, the joint distribution function $dF_{X,Y}(x, y)$ is difficult to obtain or even unknown. Thus, the correntropy has to be estimated use limited available data $\{x_i, y_i\}_{i=1}^N$

$$V_\sigma(x, y) = E[k_\sigma(x, y)] \approx \frac{1}{N} \sum_{i=1}^{N} k_\sigma(x, y). \tag{5}$$

Thus, correntropy is always positive and symmetric. Then, the Taylor series expansion of the correntropy with Gaussian kernel can be expressed as follows:

$$V_\sigma(x, y) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n \sigma^{2n} n!} E\left[(x-y)^{2n}\right]. \tag{6}$$

It can be concluded from the above that the correntropy with Gaussian kernel is actually the mean value of the difference between two random variables after Gaussian transformation, which is a more generalized signal similarity measurement. In addition, compared with low-order statistics, it can extract high-level information and contains all the even-order moment information of $x - y$.

III. Proposed Method

To fully exploit the nonlinear characteristics of HSI, we propose a spatial–spectral feature representation method based on the LCEM. The flowchart of the proposed method is shown in Fig. 1. Specifically, MNF-based dimension reduction is first used to reduce noise and computational complexity. Then, the similar neighboring pixels of each pixel can be obtained based on cosine distance in a sliding window. The size of the sliding window is determined according to the standard deviation of the image. Then, the correntropy is applied on each pixel and its similar neighboring pixels to represent the correlation between two different spectral channels. Thus, each pixel can be characterized as a correntropy matrix. Finally, a set of matrices are fed into an SVM as extracted features for final classification.

A. Selection of Local Similar Pixels

Extracting the spatial information of HSI through the local window is a common strategy. However, there may be many kinds of pixels or noises in the same window. Therefore, in the dimension reduced image (using MNF), this article first uses the cosine distance-based K-nearest neighbor (KNN) strategy to select the most similar K neighboring pixels in the sliding window. Nevertheless, how to determine the size of the local window is still a challenging problem. Hence, we propose an adaptive approach based on standard deviation. Standard deviation reflects the distribution of pixels in an image. The larger the value of the standard deviation, the more obvious the image boundary, and the richer the image details [57]. Therefore, the standard deviation can be used to determine the size of the local window. When the standard deviation is large, the image details are more and the difference between pixels is large, and then, a large window is needed to select enough similar pixels to characterize the features belonging to the same class. On the contrary, when the standard deviation of the image is small, the edge details of the image are not rich. On the one hand, a small window is sufficient to select enough pixels of the same classes, and on the other hand, it can avoid the selection of more heterogeneous pixels. Thus, if the size of local window W is $T \times T$, then it can be determined by

$$T = T_\lambda \times \mathrm{std}(I) \tag{7}$$

where T_λ is the hyperparemeter and $\mathrm{std}(I)$ is the standard deviation of the image I, whose channel has been reduced to 1 using PCA.

Usually, the value of K is determined by manual experiments. In this article, we put forward a reasonable way to obtain the number of neighboring pixels K according to the size of local window. The value of K should not be fixed but should vary with the window size. Thus, it can be determined by

$$K = K_p \times T^2 \tag{8}$$

Authorized licensed use limited to: Universidade de Macau. Downloaded on April 14,2022 at 00:44:45 UTC from IEEE Xplore. Restrictions apply.

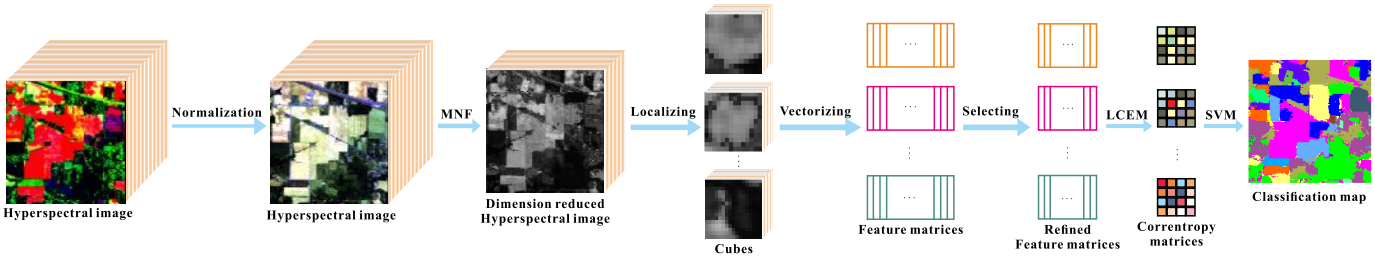


Fig. 1. Flowchart of the proposed LCEM method. Specifically, the proposed LCEM method mainly includes four steps: the MNF-based dimension reduction, the selection of local similar pixels, the construction of the LCEM, and classification. Among them, the selection of local similar pixels, including the localizing, vectorizing, and selecting steps, means that the most similar pixels of each pixel are selected using the cosine distance in the local window. Then, each pixel and its neighboring pixels can be presented as an LCEM. Finally, we obtain the classification maps by classifying the matrices.

where K_p is the proportion factor of total pixels in the local window. Then, we can employ cosine distance, which is a common similarity measurement for HSI processing, to measure the spectral similarity between different pixels [9]. If the center pixel in the window W is p_1, then its neighboring pixels are $p_i, i = 2, 3, \ldots, T^2$. The cosine distance between the pixel and its surrounding pixels can be obtained by the formula

$$\cos(p_1, p_i) = \frac{\langle p_1, p_i \rangle}{\| p_1 \|_2 \cdot \| p_i \|_2}, \quad i = 1, 2, \ldots, T^2 \tag{9}$$

where $\langle \cdot \rangle$ and $\| \cdot \|$ represent the inner product and Frobenius norm, respectively. By selecting the most similar first $K - 1$ pixels, we can get the neighboring pixels that are similar in space and spectrum. Thus, these pixels can be considered as the same kind of material as the central pixel.

B. Construction of LCEM Representation

After that, the total K pixels are used to construct a correntropy matrix. In this article, correntropy is used to represent both linear and nonlinear relationships between spectral bands of the same class. If b_i and b_j denote two different spectral bands, then the correntropy between them can be expressed as

$$C(b_i, b_j) = \frac{1}{K} \sum_{n=1}^{K} k_\sigma(b_{in}, b_{jn}) \tag{10}$$

where b_{in} and b_{jn}, respectively, represent the nth spectral value on the b_i th and b_jth spectral band, and there are k pixels. If we regard b_i and b_j as two spectral vectors, then the correntropy between the two psectral bands can also been expressed as

$$C(b_i, b_j) = \frac{1}{K} k_\sigma(b_i, b_j) \tag{11}$$

where

$$k_\sigma(b_i, b_j) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\| b_i - b_j \|_2^2}{\sqrt{2\sigma^2}}\right). \tag{12}$$

Then, the correntropy matrix can be obtained as follows:

$$M_C = \{C(b_i, b_j)\}_{i,j=1}^{B} \tag{13}$$

where B presents the number of spectral channels of the dimension reduced image, and in this article, B is set to 20.

Each nondiagonal element in the correntropy matrix represents the correntropy of different spectral bands, that is, the relationship between different spectral bands. With the moving of the window, each pixel in the hyperspectral data can obtain a correntropy matrix. The correntropy matrix not only reflects the spectral characteristics but also contains the information of spatial neighborhood. In this way, the spatial–spectral information of HSI is naturally integrated into a series of correntropy matrices, which provides more distinguishing features for classification. To ensure the positive definite of the correntropy matrix, regularization is used in this article.

C. Classification

Finally, a series of correntropy matrices are fed into SVM for final classification. As the correntropy matrix is a symmetric positive definite matrix, it is not located in the Euclidean metric space, but in the Riemannian manifold space. Therefore, the correlation entropy matrix cannot be directly input into SVM for classification. Fortunately, a kind of logm function can transform the points in manifold space into Euclidean space and complete the vector input of SVM through logarithm-Euclidean kernel function [58], [59]. The logarithm-Euclidean kernel can be defined by

$$k_{\text{logm}}(M_1, M_2) = \text{trace}\left[\text{logm}(M_1) \cdot \text{logm}(M_2)\right] \tag{14}$$

where M_1 and M_2 are two different correntropy matrices and trace[·] is the trace of a matrix. Given a symmetric positive definite matrix $C = U \sum U^T$, its logm can be defined as

$$\text{logm}(C) = U \log\left(\sum\right) U^T. \tag{15}$$

IV. Experimental Results and Discussion

In this section, we conduct extensive experiments on three real HSI datasets to verify the effectiveness of the proposed method. The hyperparameters of the proposed method are first analyzed. Then, on each dataset, the proposed LCEM is compared with some state-of-the-art methods, including LBP-based method [28], superpixel-based approach via multiple kernels (SCMK) [30], LCMR-based method [46], random patch network (RPNet) [60], neighboring region dropout for HSI classification (deepNRD) [39], hybrid spectral convolutional neural network (HybridSN) [44], double-branch dual-attention network (DBDA) [61], CNN-enhanced GCN (CEGCN) [42], and a semisupervised HSI feature extraction algorithm called joint and progressive subspace analysis (JPSA) [62]. Also, all the codes of comparison approaches come from the original articles. For deep learning-based methods, the number of epochs is set to 200.

Authorized licensed use limited to: Universidade de Macau. Downloaded on April 14,2022 at 00:44:45 UTC from IEEE Xplore. Restrictions apply.

TABLE I
DETAILS OF INDIAN PINES, PAVIA UNIVERSITY, AND SALINAS DATASET

Indian Pines			Pavia University			Salinas		
Class	Name	Samples	Class	Name	Samples	Class	Name	Samples
1	Alfalfa	46	1	Asphalt	6631	1	Brocoli_green_weeds_1	2009
2	Corn-notill	1428	2	Meadows	18649	2	Brocoli_green_weeds_2	3726
3	Corn-mintill	830	3	Gravel	2099	3	Fallow	1976
4	Corn	237	4	Trees	3064	4	Fallow_rough_plow	1394
5	Grass-pasture	483	5	Painted metal sheets	1345	5	Fallow_smooth	2678
6	Grass-trees	730	6	Bare Soil	5029	6	Stubble	3959
7	Grass-pasture-mowed	28	7	Bitumen	1330	7	Celery	3579
8	Hay-windrowed	478	8	Self-Blocking Bricks	3682	8	Grapes_untrained	11271
9	Oats	20	9	Shadows	947	9	Soil_vinyard_develop	6203
10	Soybean-notill	972				10	Corn_senesced_green_weeds	3278
11	Soybean-mintill	2455				11	Lettuce_romaine_4wk	1068
12	Soybean-clean	593				12	Lettuce_romaine_5wk	1927
13	Wheat	205				13	Lettuce_romaine_6wk	916
14	Woods	1265				14	Lettuce_romaine_7wk	1070
15	Buildings-Grass-Trees-Drives	386				15	Vinyard_untrained	7268
16	Stone-Steel-Towers	93				16	Vinyard_vertical_trellis	1807
Total		10249	Total		42776	Total		54129

Then, three commonly used quality indicators, overall accuracy (OA), average accuracy (AA), and Kappa coefficient k, are exploited to objectively evaluate the classification performance. To robustly evaluate the results, all the experimental results achieved by different methods are averaged ten times in our experiments. Furthermore, the visual classification maps of diverse methods are also shown in this article.

A. Datasets

In our experiments, three real HSI datasets are used to verify the proposed LCEM-based approach. The details are given as follows.

1) Indian Pines: The first dataset was recorded by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) sensor over Indian Pines test site in Northwestern Indiana in 1992. The size of it is $145 \times 145 \times 220$, and the spatial resolution is 20 meters per pixel (mpp). It should be mentioned that the original data, which contain 220 spectral bands, are used in our experiments. There are 10 249 labeled pixels in the image, which belong to 16 land-cover classes. The detailed information is tabulated in Table I.

2) Pavia University: The second image was acquired by the Reflective Optics Spectrographic Imaging System (ROSIS) sensor over the campus at the University of Pavia, Italy. This image contains 610×340 pixels with a spatial resolution of 1.3 mpp. There are 103 spectral channels used in our experiments. The ground truth is composed of nine land-cover classes.

3) Salinas: This image was gathered by the AVIRIS sensor over the Salinas Valley, California. This scene has a size of $512 \times 217 \times 224$. Also, the spatial resolution is 3.7 mpp. The ground-truth map covers 16 classes. It should be noted that we use the original data, which contains 224 spectral bands, for our experiments.

B. Parameters Setting

The proposed LCEM is a handcrafted feature extraction method, and thus, the parameters play a significant role. In this section, how the number of dimensions B after dimensionality reduction, the window factor T_λ, and the proportion factor K_p influence the classification performance is seriously investigated. Here, we randomly select 1% of the labeled samples per class for training, and the rest are used for testing on the Indian Pines dataset. Fig. 2 shows the effect of different values of B on the performance and computation time of the proposed LCEM on the Indian Pines dataset. As can be observed, the accuracy is very low when B is set to 10, while there is an obvious improvement in accuracy when B is increased to 20. Then, with the growth of B (from 20 to 80), the OA, AA, and Kappa obtained by LCEM improve slightly, but the corresponding running time increases heavily. In the proposed LCEM method, the construction of correntropy matrix representation is the most time-consuming part. According to (13), the time complexity of this part is $O(B^2)$. It means that the larger B, the larger size of the obtained feature matrices, and the more computational cost. Thus, B is set to 20 for all the following experiments.

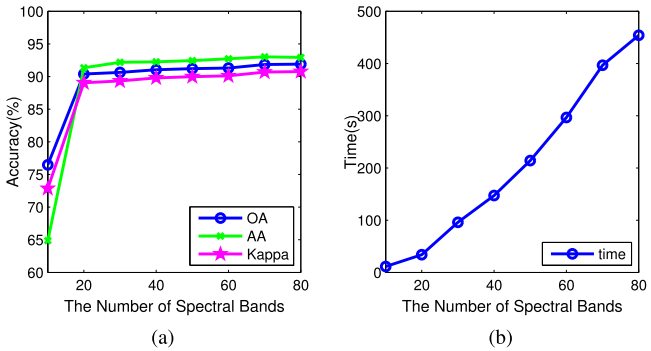


Fig. 2. Sensitivity analysis of parameter B in our proposed LCEM method. (a) Effect of the number of spectral bands on classification accuracy. (b) Effect of the number of spectral bands on running time. The experiments are performed on the Indian Pines dataset using 1% samples for training per class. As can be observed, the accuracy is very low when B is set to 10. With the growing of B (from 20 to 80), the OA, AA, and Kappa obtained by the LCEM improve slightly, but the corresponding running time increases heavily. Thus, B is set to 20 for our all experiments.

Fig. 3 shows the effect of different T_λ's (ranging from 10 to 100, with a step of 10) and K_p (ranging from 0.05 to 0.95, with a step of 0.1) on OA, AA, and Kappa on the Indian Pines dataset. For T_λ, the larger the value of T_λ, the larger the size of the local window. As can be seen from Fig. 3,

Authorized licensed use limited to: Universidade de Macau. Downloaded on April 14,2022 at 00:44:45 UTC from IEEE Xplore. Restrictions apply.

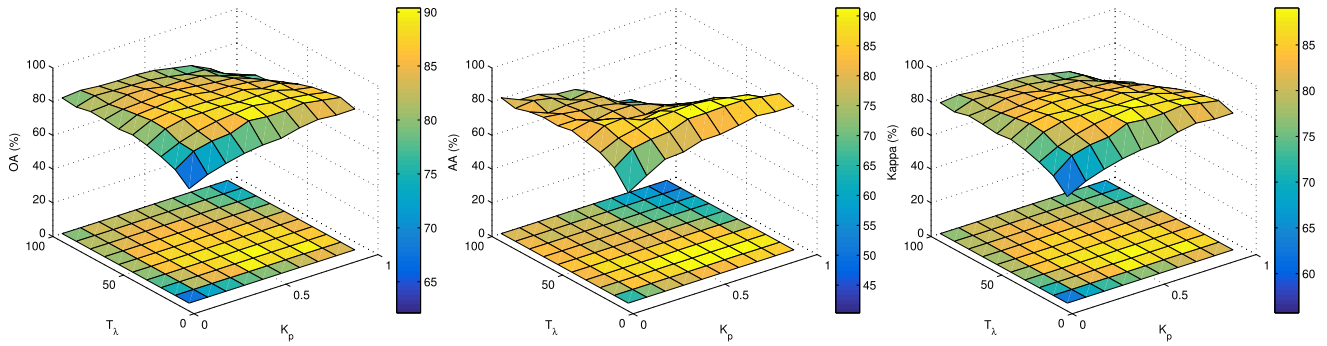


Fig. 3. Effects of different T_λ's (ranging from 10 to 100, with a step of 10) and K_p (ranging from 0.05 to 0.95, with a step of 0.1) on (a) OA, (b) AA, and (c) Kappa on the Indian Pines dataset. As can be observed, when the values of T_λ and K_p are 30 and 0.55, respectively, the best classification accuracy can be obtained. Hence, the values of T_λ and K_p are set to 30 and 0.55, respectively. In addition, the values of these two parameters remain unchanged in all our subsequent experiments.

TABLE II

AVERAGE RESULTS (%) OF TEN REPEATED EXPERIMENTS ON THE INDIAN PINES DATASET (220 SPECTRAL BANDS) PRODUCED BY DIVERSE METHODS WITH 1% TRAINING SAMPLES PER CLASS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Class	JPSA	LBP	SCMK	HybridSN	deepNRD	RPNet	CEGCN	DBDA	LCMR	LCEM
1	24.22	91.33	88.00	48.48	40.67	2.22	2.70	96.05	98.44	**100.00**
2	42.64	80.58	80.33	59.18	49.00	66.39	63.63	87.23	83.97	**87.04**
3	36.35	75.05	68.79	54.01	47.20	39.62	28.01	84.28	70.83	**86.38**
4	22.91	82.61	60.64	36.87	55.00	7.56	8.46	**84.43**	66.67	83.59
5	62.82	74.46	69.77	70.10	32.30	49.96	36.80	**95.59**	90.96	92.09
6	82.66	84.46	96.16	81.81	62.33	75.60	**98.29**	94.58	93.77	96.39
7	67.41	**95.56**	90.74	48.15	65.56	21.11	10.10	56.39	80.00	94.81
8	84.90	**98.88**	97.19	87.89	91.75	48.35	91.46	97.85	96.87	96.36
9	46.32	**85.26**	65.79	78.95	70.00	6.32	4.21	80.18	81.58	82.63
10	45.95	77.61	70.23	48.65	23.53	51.50	38.25	**83.50**	76.47	80.76
11	59.00	89.38	85.54	68.43	65.75	**95.95**	90.18	88.10	85.93	91.67
12	22.98	75.33	65.71	39.06	34.46	15.72	17.89	79.76	69.45	**87.55**
13	82.82	72.87	96.68	71.42	58.02	69.26	72.17	94.62	**99.11**	98.17
14	83.51	94.73	91.94	77.59	85.26	87.52	**97.47**	96.25	96.88	95.05
15	23.77	83.48	76.99	47.05	67.72	19.50	29.60	88.50	78.01	**93.30**
16	79.02	52.72	**97.83**	32.91	23.81	1.96	7.87	92.45	87.28	95.43
OA	56.00	84.09	81.63	69.90	57.09	64.44	65.32	88.38	84.80	**90.38**
AA	54.21	82.14	81.40	65.35	54.52	41.16	43.57	87.49	84.76	**91.33**
Kappa	49.78	81.86	79.00	65.53	50.63	57.22	59.03	86.75	82.68	**89.04**

with the increase of T_λ, the OA, AA, and Kappa increase first and then decreases slightly. This is because when the size of the window is very small, there are few similar pixels in the window, which is insufficient for feature representation. When the size is too large, more heterogeneous pixels will be selected, which will affect the classification performance. Besides, the value of K_p controls the number of neighboring pixels. Similarly, we can also observe that selecting too few or too many pixels in the window will significantly affect the classification results. As can be observed, competitive results can be obtained within the reasonable value range of the two parameters. When the values of T_λ and K_p are 30 and 0.55, respectively, however, the best classification accuracy can be obtained. Thus, the values of T_λ and K_p are set to 30 and 0.55, respectively. To verify the effectiveness of the proposed approach, the values of these two parameters remain unchanged in all our subsequent experiments.

C. Comparison With State-of-the-Art Methods

This section compares the proposed method with several advanced techniques in terms of objective accuracy and visual classification map.

1) Experiments on Indian Pines Image: The first experiment is performed on the Indian Pines dataset. In this case, 1% samples per class are randomly selected for training, and the remaining samples are used for testing. Table II lists the quantitative mean results of different methods, including OA, AA, kappa, and the average classification accuracy of each class. It is worth noting that, in our experiment, the noisy bands in the Indian Pines image are not removed in advance. As can be observed, the JPSA achieves the lowest accuracy. It fails to effectively extract the discriminative feature from original data when there are limited samples used for training. Besides, we can also observe that the classification accuracies of HybridSN, deepNRD, and CEGCN that based on deep learning are much low, which are lower than handcrafted methods (i.e., LBP and SCMK). The reason is that the training samples selected for training in this article are very few, which can easily lead to overfitting. Although the convolution kernels of RPNet do not need training, the recognition ability of random blocks cannot be guaranteed when the training size is small. Thus, it does not provide satisfactory results. Among the comparison methods, the DBDA approach has achieved the competitive classification results by capturing

Authorized licensed use limited to: Universidade de Macau. Downloaded on April 14,2022 at 00:44:45 UTC from IEEE Xplore. Restrictions apply.

plenty of spectral–spatial features using the double-branch dual-attention mechanism, but it is still inferior to the proposed LCEM method. Both the LCMR method and our proposed LCEM method extract the correlation features between spectral bands, but the accuracy of the proposed LCEM approach is nearly 6% higher than that of the LCMR. The reason is that our proposed LCEM algorithm exploits the spectral feature using correntropy, which is more appropriate to represent the nonlinear relationships between spectral bands in hyperspectral data than covariance. Then, a set of correntropy matrices obtained from the local window extract more discriminative spatial–spectral features, giving a boost to the final classification performance.

In addition to the objective quantitative analysis and comparison, this article also verifies the advantages of the proposed method from the visual effect. Fig. 4 shows the full classification maps of different methods, which is generated by one of the random experiments. It is easy to observe that the proposed LCEM has the best performance. Also, it is worth noting that, in our experiment, the noisy bands in the Indian Pines image are not removed in advance. Thus, the JPSA, SCMK, CEGCN, and LCMR have produced noisy classification maps when with a small training size, while for the LBP method, the LBP descriptor and Gabor filter could greatly suppress the negative effect of noise. However, it fails to represent the spectral features, resulting in an oversmooth classification map. Besides, there are many misclassifications in the classification maps of the HybridSN and deepNRD methods caused by the overfitting problem. For the RPNet approach, its classification map fails to capture the structure of the image. Taking the advantage of the attention mechanism, the DBDA method has produced a smooth map, but the classification result is not so accurate. Compared with more isolated noises in the classification map of LCMR, the proposed LCEM method leads to much smoother results. As can be seen, the borders between different classes can be well preserved and the regions consisting of the pixels belonging to the same class are smoother. This is mainly because the correntropy can extract the high-order moment information and correntropy matrix representation is robust to noise and intraclass variations.

The running time for one experiment of diverse approaches on the Indian Pines dataset is reported in Table III. In the experiment, we recorded the time from loading the dataset to generating the classification map. All the experiments are conducted on the device with a 2.6-GHz CPU, 64 GB of RAM, and an RTX-3090 GPU. Note that HybridSN, deepNRD, CEGCN, and DBDA are performed on PyCharm using GPU and Python 1.9. Also, the others are conducted on MATLAB 2021b. From Table III, we can observe that the LBP is the most time-consuming because it extracts a very high-dimensional feature for each pixel. Among the deep learning-based methods, the CEGCN takes the whole HSI as the network inputs, resulting in a shorter running time than other deep learning-based methods that consider the small patches as inputs. However, the proposed LCEM is more time-consuming than LCMR. This is because the calculation of correntropy matrix in the proposed LCEM is computationally

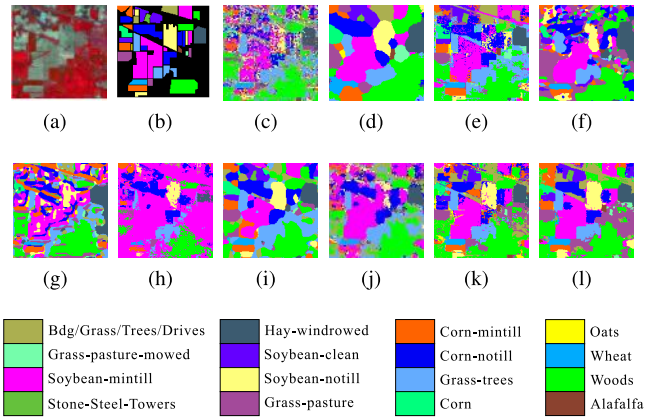


Fig. 4. Full classification maps on the Indian Pines dataset. (a) False-color map. (b) Ground truth. (c) JPSA. (d) LBP. (e) SCMK. (f) HybridSN. (g) deepNRD. (h) RPNet. (i) DBDA. (j) CEGCN. (k) LCMR. (l) Proposed LCEM.

TABLE III
COMPUTATIONAL TIME (SECONDS) FOR ONE EXPERIMENT OF DIFFERENT METHODS ON THE INDIAN PINES DATASET WITH 1% TRAINING SAMPLES PER CLASS

JPSA	LBP	SCMK	HybridSN	deepNRD
7.27	117.08	4.10	35.75	112.21
RPNet	CEGCN	DBDA	LCMR	LCEM
13.77	9.30	70.35	22.47	34.02

expensive. According to (11) and (12), it nonlinearly maps the difference of two spectral bands to another feature space through the Gaussian kernel function. Compared with covariance, the correntropy can extract high-level information of the difference and reduce the negative effect of noise. According to (13), its computational complexity is $O(B^2)$. Although more computational complexity is needed than LCMR, higher classification accuracy and more accurate classification map are achieved by the LCEM.

2) Experiments on Pavia University Image: The second experiment is conducted on Pavia University image, where only 0.1% samples per class are randomly selected for training and the rest of the samples are taken as testing samples. Table IV reports the average classification results of each method. Similar conclusions can be drawn from Table IV. The comparison methods still perform poorly due to the lack of sufficient samples. However, the proposed LCEM algorithm based on correntropy is still the best, which is about 5% higher than LCMR in terms of AA.

Fig. 5 shows the complete classification maps of the comparison methods when only 0.1% of the samples per class are used for training. It can be obviously observed that in the case of few samples used for training, some of the classification maps of the comparison methods have generated more noise and some have lost the details of the image. However, the proposed LCEM method not only preserves the image structure information but also reduces the existence of isolated noise as much as possible.

3) Experiments on Salinas Image: In this case, we only randomly select 0.1% of the labeled samples from each class

Authorized licensed use limited to: Universidade de Macau. Downloaded on April 14,2022 at 00:44:45 UTC from IEEE Xplore. Restrictions apply.

TABLE IV

AVERAGE RESULTS (%) OF TEN REPEATED EXPERIMENTS ON THE PAVIA UNIVERSITY DATASET PRODUCED BY DIVERSE METHODS WITH 0.1% TRAINING SAMPLES PER CLASS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Class	JPSA	LBP	SCMK	HybridSN	deepNRD	RPNet	CEGCN	DBDA	LCMR	LCEM
1	53.39	60.90	**91.33**	79.03	71.42	80.00	90.92	82.61	86.51	85.42
2	87.99	96.90	83.10	95.85	93.28	96.23	96.04	86.95	96.07	**98.54**
3	45.77	37.82	58.16	92.70	19.06	72.71	57.77	75.38	70.50	**77.36**
4	65.80	28.33	85.26	46.47	88.8	80.06	68.60	90.95	**93.49**	86.52
5	99.53	97.77	95.98	**100.00**	60.36	42.76	99.62	97.80	87.37	89.61
6	37.14	54.61	67.86	75.81	36.52	63.25	93.15	77.54	84.08	**95.94**
7	45.28	30.10	75.30	0.30	15.48	76.37	59.47	76.68	70.05	**94.95**
8	62.50	56.89	83.77	7.48	84.99	63.94	**85.78**	66.51	63.08	69.98
9	52.09	11.90	74.70	21.35	50.19	39.28	39.60	**95.77**	67.91	62.39
OA	69.03	71.16	81.55	75.10	74.14	81.19	87.91	82.19	87.19	**90.65**
AA	61.05	52.80	79.50	57.67	57.79	68.29	76.77	83.35	79.90	**84.52**
Kappa	58.34	58.87	75.87	65.95	65.14	74.20	83.93	75.99	83.04	**87.58**

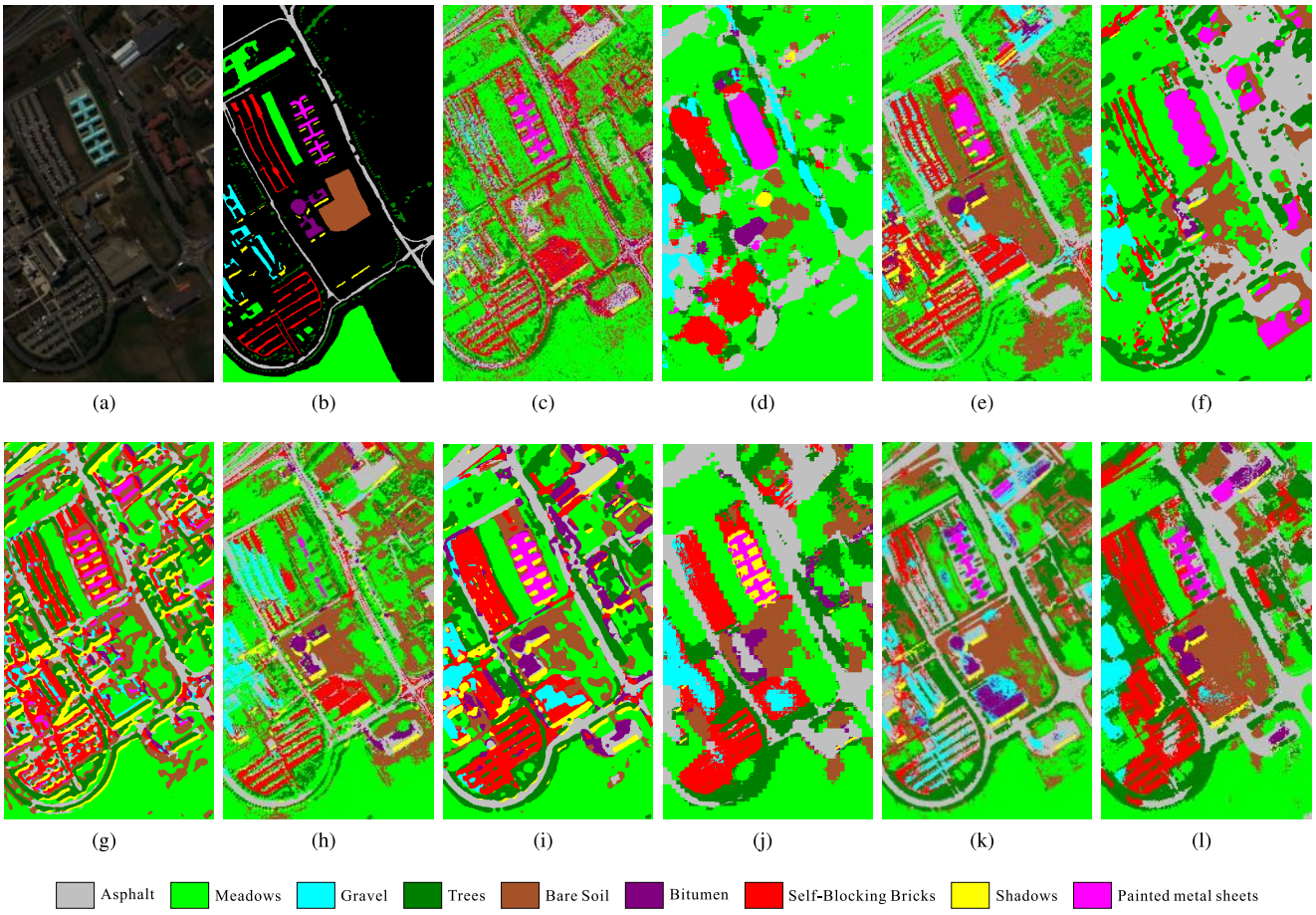


Fig. 5. Full classification maps on the Pavia University dataset. (a) False-color map. (b) Ground truth. (c) JPSA. (d) LBP. (e) SCMK. (f) HybridSN. (g) deepNRD. (h) RPNet. (i) DBDA. (j) CEGCN. (k) LCMR. (l) Proposed LCEM.

for training. The detailed average results of several methods are shown in Table V. It can be seen from the table that the proposed LCEM method has achieved the best results in terms of OA, AA, and Kappa coefficient. Fig. 6 presents the visual classification maps of all the comparison methods. It is possible to state that our proposed approach yields the best visual effect in such a small training size. Considering all the above experimental results, the effectiveness of our proposed LCEM approach is fully proved. Besides, the experimental results also show the unique potentiality and advantage of the correntropy matrix as a feature descriptor in extracting the spectral–spatial of hyperspectral data.

D. Effect of Different Training Samples on Classification Performance

To more comprehensively demonstrate the effectiveness of the proposed LCEM, we further analyze the effect of different training samples on classification performance.

On the Indian Pines dataset, we randomly select 1%, 2%, 3%, 4%, and 5% labeled samples per class for training. Fig. 7

Authorized licensed use limited to: Universidade de Macau. Downloaded on April 14,2022 at 00:44:45 UTC from IEEE Xplore. Restrictions apply.

TABLE V

AVERAGE RESULTS (%) OF TEN REPEATED EXPERIMENTS ON THE SALINAS DATASET (224 SPECTRAL BANDS) PRODUCED BY DIVERSE METHODS WITH 0.1% TRAINING SAMPLES PER CLASS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Class	JPSA	LBP	SCMK	HybridSN	deepNRD	RPNet	CEGCN	DBDA	LCMR	LCEM
1	98.56	86.29	96.95	90.81	70.00	91.93	91.62	**99.66**	98.55	97.84
2	94.20	76.05	98.25	90.52	87.72	97.30	**100.00**	99.87	98.00	96.49
3	51.26	41.17	71.66	64.50	55.46	68.80	97.35	94.69	75.77	**99.96**
4	99.00	84.95	98.70	63.15	83.66	99.39	**99.93**	92.66	99.51	99.31
5	98.34	77.23	87.85	73.04	84.30	94.65	90.03	89.08	92.26	**98.87**
6	99.68	85.59	99.75	90.83	86.83	97.88	**99.89**	99.82	98.49	99.80
7	97.45	85.15	97.92	82.55	91.37	98.51	98.75	95.49	99.10	**99.97**
8	67.29	89.89	80.78	81.82	65.08	76.13	**98.74**	85.90	89.59	90.75
9	97.33	95.21	99.85	90.86	89.96	97.76	**100.00**	97.73	99.22	**100.00**
10	72.82	79.85	80.89	75.52	57.65	86.89	**93.04**	90.40	82.28	84.64
11	77.59	64.74	81.29	66.88	79.21	90.62	81.90	91.06	79.02	**98.64**
12	90.87	78.76	77.87	65.38	33.93	98.61	**100.00**	97.84	92.91	95.91
13	97.81	34.26	61.39	82.03	32.32	**98.97**	93.07	98.64	61.44	73.70
14	87.45	46.68	91.69	81.81	62.34	84.81	**100.00**	90.71	94.24	95.79
15	52.11	81.72	59.92	59.01	69.39	73.00	48.52	74.23	83.35	**86.95**
16	74.29	48.13	91.41	59.29	25.35	65.52	75.55	**99.66**	75.92	91.26
OA	80.36	80.47	85.12	85.66	71.44	86.49	90.13	90.60	90.56	**93.98**
AA	84.75	72.23	86.01	83.73	67.16	88.80	91.77	93.58	88.73	**94.37**
Kappa	78.13	78.06	83.41	83.97	68.18	84.95	88.95	89.54	89.47	**93.30**

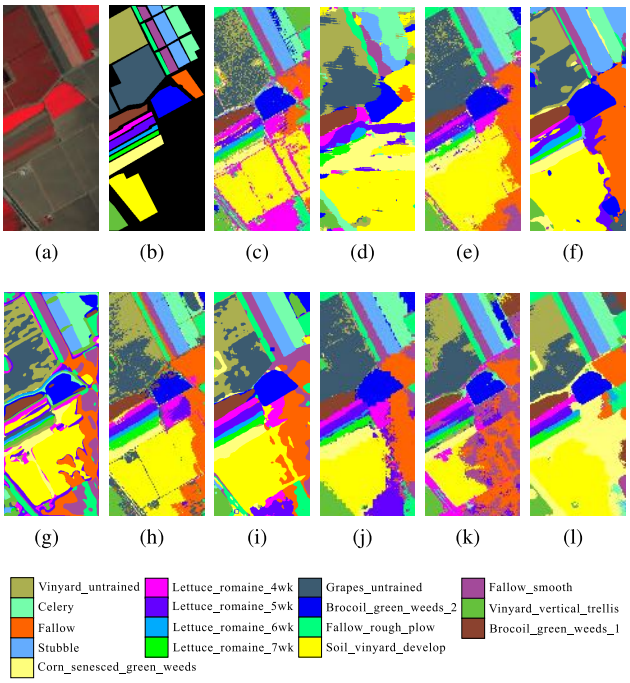


Fig. 6. Full classification maps on the Salinas dataset. (a) False-color map. (b) Ground truth. (c) JPSA. (d) LBP. (e) SCMK. (f) HybridSN. (g) deepNRD. (h) RPNet. (i) DBDA. (j) CEGCN. (k) LCMR. (l) Proposed LCEM.

presents the influence of different training samples on OA, AA, and Kappa. It can be seen that classification accuracies of all the methods steadily improve as the training size increases, and only when the training size is large enough, the comparison methods perform similar classification results. However, the proposed method always performs the best, especially when the training size is very small. In addition, Figs. 8 and 9 show the effect of the training sample size on classification performance on the Pavia University and Salinas datasets, respectively. On these two datasets, only 0.1%, 0.15%, 0.2%, 0.25%, and 0.3% samples per class are randomly chosen for training, and the remaining samples are used for testing. Similarly, we can conclude that the proposed method still achieves competitive results in terms of OA, AA, and kappa with limited training samples. Considering all the results, we can conclude that the correntropy matrix has the obvious advantage in capturing the hyperspectral features and can be beneficial to classification, especially when with limited training samples.

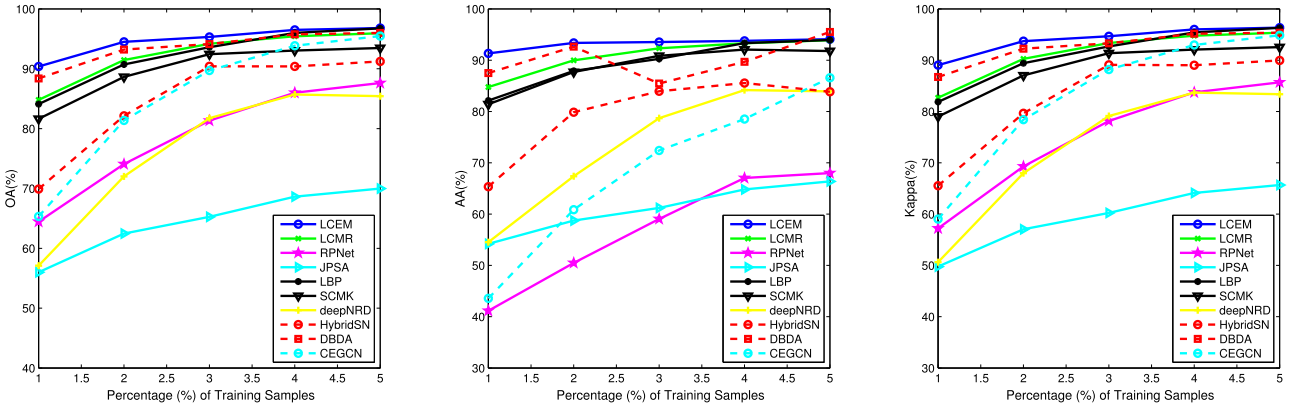
E. Feature Space Analysis

To further analyze the class separation capability of the proposed LCEM method, we also provide the visualization results of features learned by the LCMR and the proposed LCEM methods on three datasets. The 2-D projection of features is shown in Fig. 10. The maps in the first row are the results performed on the Indian Pines dataset. Fig. 10(a) shows the projection of original features. The blue color and the pink color represent the corn-notill and Soybean-mintill classes, respectively. As can be observed, it is challenging to classify the pixels accurately on this dataset due to the spectral aliasing. Fig. 10(b) and (c) shows the projection of features learned by the LCMR method and the proposed LCEM, respectively. Compared with the LCMR method, our proposed LCEM method can better separate pixels of different classes (i.e., the pixels marked blue and other colors). From the projection of the Pavia University dataset (the second row in Fig. 10), we can also easily observe that our proposed LCEM approach [Fig. 10(f)] can better cluster the pixels belonging to the same class, for example, the pixels marked green and brown. Similarly, it can also be observed from Fig. 10(h) and (i) that the proposed LCEM can not only better separate pixels belonging to different classes but also better cluster pixels belonging to the same class. Taking all the results into account, it is possible to state that our proposed LCEM method can characterize the discriminative spatial–spectral features for classification.

F. Effect of Diverse Classifiers and Dimensionality Reduction Algorithms

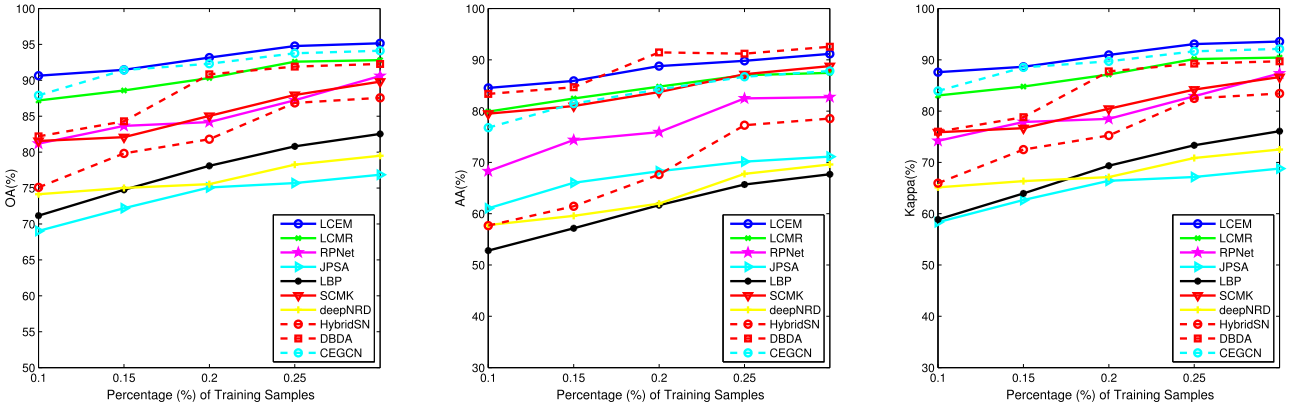
Furthermore, we also analyze the effect of diverse classifiers and dimensionality reduction algorithms on our proposed

Authorized licensed use limited to: Universidade de Macau. Downloaded on April 14,2022 at 00:44:45 UTC from IEEE Xplore. Restrictions apply.



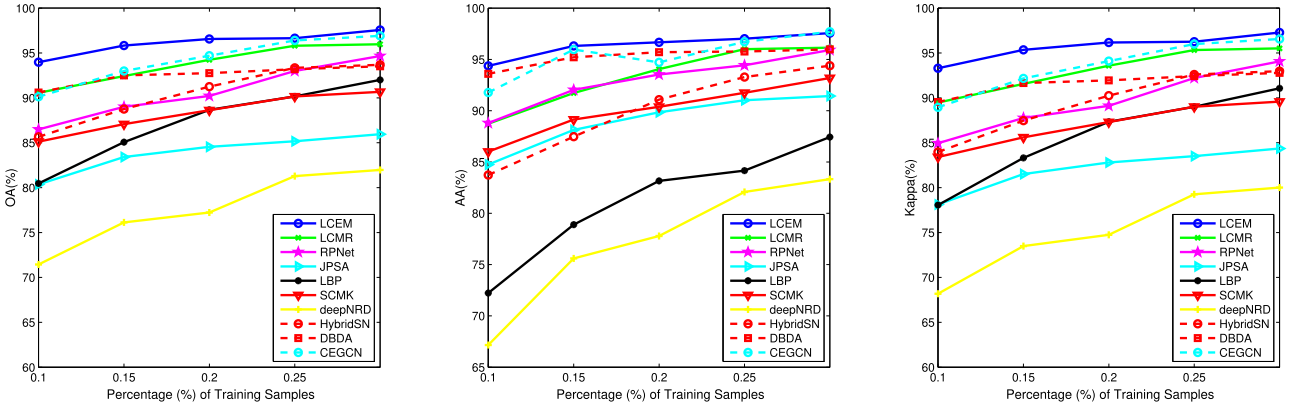
(a) (b) (c)

Fig. 7. Experimental results of different methods with different numbers of training samples on the Indian Pines dataset. (a) OA. (b) AA. (c) Kappa.



(a) (b) (c)

Fig. 8. Experimental results of different methods with different numbers of training samples on the Pavia University dataset. (a) OA. (b) AA. (c) Kappa.



(a) (b) (c)

Fig. 9. Experimental results of different methods with different numbers of training samples on the Salinas dataset. (a) OA. (b) AA. (c) Kappa.

LCEM approach. In this section, we use four classifiers to evaluate the effectiveness of the features learned by the proposed LCEM, including logarithm-Euclidean kernel-based SVM (SVM-logmE), radial basis function kernel-based SVM (SVM-RBF), KNNs, and RF classifier. The fivefold cross validation is employed in the SVM-RBF classifier and the

Authorized licensed use limited to: Universidade de Macau. Downloaded on April 14,2022 at 00:44:45 UTC from IEEE Xplore. Restrictions apply.

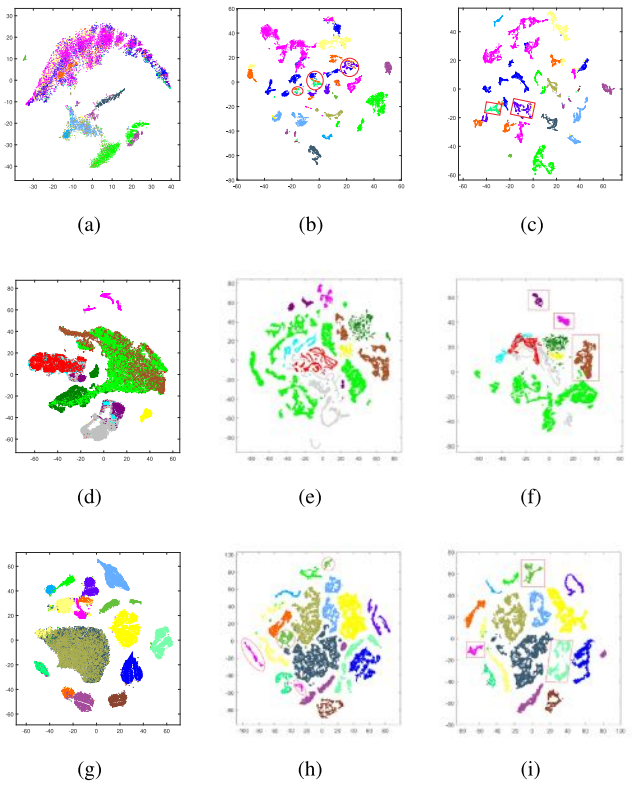


Fig. 10. 2-D projection of features learned by the LCMR and the proposed LCEM methods on three datasets using t-distributed stochastic neighbour embedding (t-SNE). The feature projections on (a)–(c) Indian Pines, (d)–(f) Pavia University, and (g)–(i) Salinas dataset. From left to right, the first column represents the projection of the original dataset, the second column represents the projection of the features learned by the LCMR method, and the last column represents the projection of the features learned by the proposed LCEM approach. From the figure, we can see that the proposed LCEM can better cluster pixels belonging to the same class and separate pixels of different classes.

TABLE VI
AVERAGE RESULTS (%) OF DIFFERENT CLASSIFIERS AND DIMENSIONALITY REDUCTION ALGORITHMS PERFORMED ON THE INDIAN PINES DATASET WITH 1% TRAINING SAMPLES PER CLASS

		SVM-logmE	SVM-RBF	KNN	RF
MNF	OA	**90.38**	89.74	89.98	86.60
	AA	**91.33**	90.47	91.25	80.82
	Kappa	**89.04**	88.29	88.58	84.57
PCA	OA	**84.95**	83.90	84.84	80.38
	AA	83.40	82.99	**84.38**	75.68
	Kappa	**82.84**	81.57	82.74	77.36

cosine distance is used in KNN. In addition, two different dimensionality reduction algorithms, MNF and PCA, are considered in our comparison experiments.

The experimental results conducted on the Indian Pines dataset with 1% training samples per class are reported in Table VI. As can be observed, all the classification methods have achieved competitive results, especially when the dimensionality reduction is the MNF. However, the RF classifier has achieved the lowest accuracy. The reason for that may be that it suffers from the class imbalance problem, as we can see that the AA produced by it is extremely low. Also, the SVM-logmE, SVM-RBF, and KNN classifiers have shown a similar performance. Nevertheless, the logarithm-Euclidean kernel function-based SVM is more appropriate in theory for the classification of correntropy feature matrix. Thus, we adopt the SVM-logmE classifier as the classifier in this article. Besides, we can also find that the MNF outperforms PCA in our proposed model. This is mainly because the MNF is committed to maximizing the signal-to-noise ratio of dimensionality-reduced data, which is beneficial to subsequent feature extraction. Furthermore, the experimental results of diverse classifiers also strongly prove the effectiveness of the features extracted by LCEM.

V. CONCLUSION

In this article, an effective spectral–spatial feature extraction method based on the LCEM representation is proposed for HSI classification, where correntropy is used to represent the correlation between spectral bands. In addition, we also propose a standard deviation-based parameter adaptive method to determine the window size and the number of neighboring pixels. The experimental results on three public datasets demonstrate that the proposed method is effective, especially in the case of very limited samples. Besides, the following two points are proved. First, taking the advantage of the kernel method, the correntropy can better describe the nonlinear relationships in hyperspectral data and suppress the negative impact of noise. Second, the LCEM as a local feature descriptor can provide the discriminative features of different land covers. Due to these two advantages, the proposed method has achieved better performance in classification accuracy and visual effects. At the same time, we have also noticed the application potential of information theory learning in HSI processing. In the future, we will further explore the application of the theory of ITL on HSI. However, how to make full use of spatial information of HSI is still a problem worthy of attention. We consider adopting the multiscale superpixel segmentation to capture the structure of HSI and then perform the feature extraction by a correntropy matrix.

REFERENCES

[1] 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.
[2] M. Shimoni, R. Haelterman, and C. Perneel, "Hypersectral imaging for military and security applications: Combining myriad processing and sensing techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 101–117, Jun. 2019.
[3] A. W. Bitar, L.-F. Cheong, and J.-P. Ovarlez, "Sparse and low-rank matrix decomposition for automatic target detection in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5239–5251, Aug. 2019.
[4] K. L.-M. Ang and J. K. P. Seng, "Big data and machine learning with hyperspectral information in agriculture," *IEEE Access*, vol. 9, pp. 36699–36718, 2021.
[5] J. Aravinth, A. Bharadwaj, K. Harikrishna, and N. Vignajeeth, "Classification of urban objects from HSR-HTIR data using CNN and random forest classifier," in *Proc. 3rd Int. Conf. Commun. Electron. Syst. (ICCES)*, Oct. 2018, pp. 388–391.
[6] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.
[7] J. Jiang, J. Ma, Z. Wang, C. Chen, and X. Liu, "Hyperspectral image classification in the presence of noisy labels," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 851–865, Feb. 2019.

Authorized licensed use limited to: Universidade de Macau. Downloaded on April 14,2022 at 00:44:45 UTC from IEEE Xplore. Restrictions apply.

[8] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 279–317, Dec. 2019.

[9] C.-I. Chang, "An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis," *IEEE Trans. Inf. Theory*, vol. 46, no. 5, pp. 1927–1932, Aug. 2000.

[10] G. Camps-Valls, "Kernel spectral angle mapper," *Electron. Lett.*, vol. 52, no. 14, pp. 1218–1220, 2016.

[11] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[12] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.

[13] M. Pal, "Multinomial logistic regression-based feature selection for hyperspectral data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 14, no. 1, pp. 214–220, Feb. 2012.

[14] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[15] S. Prasad and L. M. Bruce, "Limitations of principal components analysis for hyperspectral target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 625–629, Oct. 2008.

[16] B.-Y. Sun, X.-M. Zhang, J. Li, and X.-M. Mao, "Feature fusion using locally linear embedding for classification," *IEEE Trans. Neural Netw.*, vol. 21, no. 1, pp. 163–168, Jan. 2010.

[17] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 65–74, Jan. 1988.

[18] M. P. Uddin, M. A. Mamun, M. I. Afjal, and M. A. Hossain, "Information-theoretic feature selection with segmentation-based folded principal component analysis (PCA) for hyperspectral image classification," *Int. J. Remote Sens.*, vol. 42, no. 1, pp. 286–321, Jan. 2021.

[19] M. P. Uddin, M. Al Mamun, and M. A. Hossain, "Effective feature extraction through segmentation-based folded-PCA for hyperspectral image classification," *Int. J. Remote Sens.*, vol. 40, no. 18, pp. 7190–7220, Sep. 2019.

[20] M. P. Uddin, M. A. Mamun, and M. A. Hossain, "PCA-based feature reduction for hyperspectral remote sensing image classification," *IETE Tech. Rev.*, vol. 38, no. 4, pp. 377–396, Jul. 2021.

[21] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral–spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[22] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 809–823, Mar. 2012.

[23] Y. Wei, Y. Zhou, and H. Li, "Spectral–spatial response for hyperspectral image classification," *Remote Sens.*, vol. 9, no. 3, p. 203, Feb. 2017.

[24] J. Li *et al.*, "Multiple feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1592–1606, Mar. 2015.

[25] L. He, C. Liu, J. Li, Y. Li, S. Li, and Z. Yu, "Hyperspectral image spectral–spatial-range Gabor filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4818–4836, Jul. 2020.

[26] 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.

[27] 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, May 2014.

[28] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015.

[29] Y. Zhou and Y. Wei, "Learning hierarchical spectral–spatial features for hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 46, no. 7, pp. 1667–1678, Jul. 2016.

[30] L. Fang, S. Li, W. Duan, J. Ren, and J. A. Benediktsson, "Classification of hyperspectral images by exploiting spectral–spatial information of superpixel via multiple Kernels," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6663–6674, Dec. 2015.

[31] T. Zhan, Z. Lu, M. Wan, and G. Yang, "Multiscale superpixel kernel-based low-rank representation for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 9, pp. 1642–1646, Sep. 2020.

[32] H. Huang, G. Shi, H. He, Y. Duan, and F. Luo, "Dimensionality reduction of hyperspectral imagery based on spatial–spectral manifold learning," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2604–2616, Jun. 2020.

[33] H. Huang, Y. Duan, H. He, and G. Shi, "Local linear spatial–spectral probabilistic distribution for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1259–1272, Feb. 2020.

[34] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, "Scene segmentation with dual relation-aware attention network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2547–2560, Jun. 2021.

[35] J. Xue, Y.-Q. Zhao, Y. Bu, W. Liao, J. C.-W. Chan, and W. Philips, "Spatial–spectral structured sparse low-rank representation for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 3084–3097, 2021.

[36] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[37] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, Jan. 2021.

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

[39] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Neighboring region dropout for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 1032–1036, Jun. 2020.

[40] P. Zhong, Z. Gong, S. Li, and C.-B. Schönlieb, "Learning to diversify deep belief networks for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 55, no. 6, pp. 3516–3530, Jun. 2017.

[41] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[42] Q. Liu, L. Xiao, J. Yang, and Z. Wei, "CNN-enhanced graph convolutional network with pixel- and superpixel-level feature fusion for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8657–8671, Oct. 2021.

[43] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, "A simplified 2D-3D CNN architecture for hyperspectral image classification based on spatial–spectral fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2485–2501, 2020.

[44] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.

[45] J. T. Peng, W. Sun, L. Ma, and Q. Du, "Discriminative transfer joint matching for domain adaptation in hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 972–976, Jun. 2019.

[46] L. Fang, N. He, S. Li, A. J. Plaza, and J. Plaza, "A new spatial–spectral feature extraction method for hyperspectral images using local covariance matrix representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3534–3546, Jun. 2018.

[47] N. He *et al.*, "Feature extraction with multiscale covariance maps for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 755–769, Feb. 2019.

[48] X. Zhang, Y. Wei, H. Yao, and Y. Zhou, "Improved local covariance matrix representation for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Sep. 2020, pp. 68–71.

[49] X. Zhang, Y. Wei, H. Yao, Z. Ye, Y. Zhou, and Y. Zhao, "Locally homogeneous covariance matrix representation for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9396–9407, 2021.

[50] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, Nov. 2007.

[51] S. Yu, L. S. Giraldo, and J. Principe, "Information-theoretic methods in deep neural networks: Recent advances and emerging opportunities," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 4669–4678.

Authorized licensed use limited to: Universidade de Macau. Downloaded on April 14,2022 at 00:44:45 UTC from IEEE Xplore. Restrictions apply.

[52] S. Yu, A. Shaker, F. Alesiani, and J. Principe, "Measuring the discrepancy between conditional distributions: Methods, properties and applications," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 2777–2784.

[53] Y. Wei, S. Yu, L. S. Giraldo, and J. C. Príncipe, "Multiscale principle of relevant information for hyperspectral image classification," *Mach. Learn.*, pp. 1–26, Jun. 2021.

[54] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: A localized similarity measure," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jul. 2006, pp. 4919–4924.

[55] J. Peng and Q. Du, "Robust joint sparse representation based on maximum correntropy criterion for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7152–7164, Dec. 2017.

[56] W. Sun, J. Peng, G. Yang, and Q. Du, "Correntropy-based sparse spectral clustering for hyperspectral band selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 3, pp. 484–488, Mar. 2020.

[57] Y. Cui, Y. An, W. Sun, H. Hu, and X. Song, "Multiscale adaptive edge detector for images based on a novel standard deviation map," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[58] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM J. Matrix Anal. Appl.*, vol. 29, no. 1, pp. 328–347, 2007.

[59] R. Wang, H. Guo, L. S. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2496–2503.

[60] Y. Xu, B. Du, F. Zhang, and L. Zhang, "Hyperspectral image classification via a random patches network," *ISPRS J. Photogramm. Remote Sens.*, vol. 142, pp. 344–357, Aug. 2018.

[61] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, "Classification of hyperspectral image based on double-branch dual-attention mechanism network," *Remote Sens.*, vol. 12, no. 3, p. 582, Feb. 2020.

[62] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Joint and progressive subspace analysis (JPSA) with spatial–spectral manifold alignment for semisupervised hyperspectral dimensionality reduction," *IEEE Trans. Cybern.*, vol. 51, no. 7, pp. 3602–3615, Jul. 2021.

Xinyu Zhang is currently pursuing the master's degree with the School of Artificial Intelligence, Xidian University, Xi'an, China.

He was with the Hubei Research Center for Educational Informationization, Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan, China. His research interests include computer vision and machine learning.

Yantao Wei is currently an Associate Professor with the Hubei Research Center for Educational Informationization, Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan, China. His research interests include educational artificial intelligence, computer vision, and machine learning.

Weijia Cao received the master's and Ph.D. degrees in computer science with the University of Macau, Macau, China, in 2013 and 2017, respectively.

She is currently an Assistant Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. Her main research interests revolve around multimedia encryption, machine learning, and remote sensing image processing.

Huang Yao is currently a Lecturer with the Hubei Research Center for Educational Informationization, Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan, China. His research interests include educational artificial intelligence, computer vision, and machine learning.

Jiangtao Peng received the B.S. and M.S. degrees from Hubei University, Wuhan, China, in 2005 and 2008, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011.

He is currently a Professor with the Faculty of Mathematics and Statistics, Hubei University. His research interests include machine learning and hyperspectral image processing.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree from Hunan University, Changsha, China, in 1992, and the M.S. and Ph.D. degrees from Tufts University, Medford, MA, USA, in 2008 and 2010, respectively, all in electrical engineering.

He joined the Department of Computer and Information Science, University of Macau, Macau, China, in 2011, as an Assistant Professor, where he is currently a Full Professor and the Director of the Vision and Image Processing Laboratory. His research interests include image processing, computer vision, machine learning, and multimedia security.

Authorized licensed use limited to: Universidade de Macau. Downloaded on April 14,2022 at 00:44:45 UTC from IEEE Xplore. Restrictions apply.