Hierarchical Structure-Based Noisy Labels Detection for Hyperspectral Image Classification

Bing Tu, *Member, IEEE*, Chengle Zhou, *Graduate Student Member, IEEE*, Xiaolong Liao, Zhi Xu, Yishu Peng, and Xianfeng Ou, *Member, IEEE*

***Abstract*—In hyperspectral image (HSI) classification, the performance of supervised learning tends to be affected by priori knowledge, i.e., quantity and quality of samples. However, it is inevitable to limit the performance of supervised classification due to the presence of noisy labels in the training samples. In this article, we first propose a hierarchical constrained energy minimum (HCEM) method to detect mislabeled samples (noisy labels) of original training set trained with supervised task and boost the performance of classifiers and spectral-spatial classification methods in HSI applications. The basic idea behind this work is that the filter output energy of noisy label spectrum is hierarchically suppressed based on the cascade of CEM detectors at different layers. The proposed HCEM method consists of four key steps: First, the distance information among samples is obtained to calculate the spectral similarity between samples per class in original training set. Then, the confidence spectra per class is constructed according to the maximum similarity domain. Next, the spectra of mislabeled samples is suppressed and the spectra of true samples is preserved by multilayers CEM detector. Finally, the noisy labels are detected and removed based on the output evaluated by the proposed HCEM method. Experimental results on four real hyperspectral datasets are verified by a series of spectral classifiers and spectral-spatial classification methods, it demonstrated that the proposed HCEM method, can accurately remove noisy labels of original training set and effectively improve the performance of supervised classification task.**

***Index Terms*—Constrained energy minimum (CEM), hierarchical structure, hyperspectral image (HSI), noisy label, supervised task.**

Manuscript received February 23, 2020; revised April 18, 2020; accepted May 6, 2020. Date of publication May 12, 2020; date of current version May 29, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61977022 and Grant 51704115, in part by the Natural Science Foundation of Hunan Province under Grant 2019JJ50212, in part by the Key Research and Development Program of Hunan Province under Grant 2019SK2102, in part by the Research Foundation of Education Bureau of Hunan Province under Grant 19B237, in part by the Hunan Institute of Science and Technology Innovation Foundation for Postgraduate under Grant YCX2019A11, in part by the Engineering Research Center on 3-D Reconstruction and Intelligent Application Technology of Hunan Province under Grant 2019-430602-73-03-006049, and in part by the Hunan Emergency Communication Engineering Technology Research Center under Grant 2018TP2022. *(Bing Tu and Chengle Zhou contributed equally to this work.) (Corresponding authors: Zhi Xu; Yishu Peng.)*

Bing Tu, Chengle Zhou, Xiaolong Liao, Yishu Peng, and Xianfeng Ou are with the College of Information and Communication Engineering, Hunan Institute of Science and Technology, Yueyang 414000, China (e-mail: tubing@hnist.edu.cn; chengle_zhou@foxmail.com; xiaolong_liao@vip.hnist.edu.cn; lovepys@hnist.edu.cn; ouxf@hnist.edu.cn).

Zhi Xu is with the Guangxi Key Laboratory of Images and Graphics Intelligent Processing, Guilin University of Electronics Technology, Guilin 541000, China (e-mail: xuzhi@guet.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2020.2994162

I. Introduction

With the increasing development of remote sensing technology, the demands of treatment to the numerous data obtained by sensor are higher and higher. Hyperspectral images (HSI) are composed of hundreds of contiguous spectral bands and be regarded as a 3-D image cube. The elements of spectral vectors in the image cube represent to the fine and unique spectral features of corresponding pixels in hyperspectral image, and the particularly spectral characteristic provide a possibility for discriminating the materials of ground coverings. With the superiority of HSI to feature property identification, HSI processing technologies [1]–[3] are developing rapidly and have been applied in various scenes, such as urban planning [4]–[6], precision agriculture [7], [8], and grassland species monitoring [9], [10]. Among those remote sensing applications, HSI classification technologies [11]–[13] play an important role. The purpose of classification task is to label each unlabel pixel according to given spectral and spatial information. In recent years, the supervised classification is widely used and become a new branch of the mainstream remote sensing classification methods since the classification accuracy and reliability of supervised classification far exceed unsupervised classification such as support vector machines (SVMs) [14]–[16], neural network [17], [18], decision tree [19], [20], and genetic algorithm [21], [22]. However, there is a problem in the application of supervised classification is that they usually assume that the training set is highly convincing. This assumption is not always held since labeling mistakes are unavoidable in real applications. Therefore, designing an efficient noise label detection method will be an extremely active research topic for HSI supervised classification.

Noise label problem in process of supervised learning has attracted wide attention of researchers both computer vision and remote sensing fields. To name a few, in the general computer vision, Wu *et al.* [23] introduced a light convolutional neural network learning framework to model a compact embedding on the large-scale face data with massive noisy labels. Experimental results show that the proposed framework can exploit large-scale noisy data to learn a light model that is useful in term of computational costs and storage spaces. Liu and Tao [24] designed a scheme to solve classification problem with noisy labels by using importance reweighting, with consistency assurance that the label noise does not ultimately hinder the search for the optimal classifier of the noise-free sample. Yao *et al.* [25] proposed

a generative model called latent stability analysis to discover stable patterns from images with noisy labels. In addition, To learn decision rule from weak and noisy labels for semantic segmentation, a L_1-optimisation-based sparse model is formulated by Lu *et al.*[26]. For remote sensing domain, Kang *et al.* [27] combined the mechanism and reasons that may generate the noisy labels in supervised task of HSI classification for the first time, and developed detection and correction of mislabeled training samples method based on the edge-preserving filtering (EPF) and spectral constraints. Experiments performed on real hyperspectral datasets demonstrate the effectiveness of the proposed method in improving classification performance with respect to the classifier trained with the original training set that contains a number of mislabeled samples. Jiang *et al.* [28] introduced a random label propagation algorithm (RLPA) method to detect the noisy labels of the original training set. The basic idea behind the RLPA algorithm is to utilize the superpixel-based spectral-spatial constrains from the observed HSIs based on graph theory and employ it to conduct label propagation. Experimental studies show that the RLPA can decrease the level of noisy label and demonstrate the advantages of the RLPA method over some classic classifiers with a significant margin-the gains in terms of classification accuracy. In [29], an noisy label detection method is designed to obtain good performance for the classification with mislabeled samples by fusing spectral angle and the local outlier factor (SALOF). Similarly, in order to promote supervised classification performance, Tu *et al.* [30]–[32] utilized density distribution to explore noisy labels in original training set, and proposed a series of noise label detection algorithms based on density peak (DP) clustering algorithm. The sufficient experimental results show that the DP-based detection methods can effectively remove noisy labels of original training set and promote classification accuracy of supervised task. Furthermore, Jie *et al.* [33] presented a joint spectral-spatial distributed sparse representation based noisy label detection method, which takes advantage of the intraband structure and the interband correlation for joint sparse representation and joint dictionary learning. In [34], a kernel entropy component analysis (KECA) based method is proposed to remove noisy labels of a training set with mislabeled samples and then improve performance of supervised classification.

Recently, a multitude of CEM-based approaches have been widely used in remote sensing data processing. For example, Chang and Wang [35] makes use of constrained energy minimization (CEM) to constrain a single band to calculate its priority for band selection. Gao *et al.* [36] proposed a novel adjusted spectral matched filter for hyperspectral target detection, and using the RX anomaly detector to adjust the output of supervised CEM detector. Zou *et al.* [37] designed a quadratic constrained energy minimization detector for HSI target detection. Experimental results on one real hyperspectral images and one synthetic image suggest their method significantly improves the performance of the original CEM detection algorithm. Moreover, a hybrid sparsity and CEM (HSCEM) detector for HSI is proposed to improve performance of target detection[38]. Experimental results illustrate the outperformance of the HSCEM detector over several classic statistics-based detectors and sparsity-based detectors.

In this article, we initially introduce a hierarchical structure-based CEM (HCEM) method to detect mislabeled samples (noisy labels) of original training set trained with supervised task and improve the effectiveness of classifiers and spectral-spatial classification methods in HSI processing. The motivation behind this work is that the filter output energy of noisy label spectrum is hierarchically suppressed according to the cascade of CEM detectors at different layers. The proposed HCEM method consists of four key steps: First, the distance information among samples is obtained to calculate the spectral similarity between samples per class in original training set. Then, the confidence spectra per class is constructed according to the maximum similarity domain. Next, the spectra of mislabeled samples is suppressed and the spectra of true samples is preserved by multilayers CEM detector. Finally, the noisy labels are detected and removed based on the output evaluated by the proposed HCEM method. The contributions of our work are summarized as the following three points.

1) We initially introduce multilayers structure-based CEM algorithm instead of a simple CEM detector into the noisy label detection problem for hyperspectral supervised learning. Meanwhile, we prove the convergence of the proposed method, and we also give a theoretical explanation on why we can achieve the gradually increasing detection performance through the hierarchical suppression process.
2) The proposed HCEM method is first applied for noisy label detection. It is found that the constructed target spectrum can enhance the robustness of the proposed method better than the maximum confidence target spectrum for each class.
3) Four representative distance metrics are analyzed in the proposed detection framework, in which the spectral angle mapping (SAM) is found to be a robust metric for detecting noisy labels of original training set.

The rest of this article is organized as follows. Section II reviews the traditional constrained energy minimum algorithm and different spectral similarity measures. Section III introduces the proposed HCEM-based noisy label detection method in detail. The theory-based performance analysis for the proposed HCEM method is given in Section IV. Experimental results on the real HSI dataset are presented in Section V. Conclusion and future work are discussed in Section VI.

II. Review of Related Works

In this section, we will review the traditional CEM algorithm and spectral similarity metrics.

A. Spectral Similarity Metrics

Let $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_N}\}$ is the spectral vector in the remote sensing image, N represents the number of pixel, $\boldsymbol{x_i} = [x_{i1}, x_{i2}, \ldots, x_{iL}]^T (1 \leq i \leq N)$ is L-dimensional column vector, L refers to the number of band. The definitions of mainstream distance metrics, i.e, spectral information divergence

(SID)[39], correlation coefficient (CC)[40], spectral gradient angle (SGA)[41], and SAM[42], are presented in the following.

1) *Spectral Information Divergence*:

$$\boldsymbol{D}_{\boldsymbol{SID}}^{\boldsymbol{ij}} = -\sum \boldsymbol{P_i} \log\left(\frac{\boldsymbol{P_i}}{\boldsymbol{P_j}}\right) - \sum \boldsymbol{P_j} \log\left(\frac{\boldsymbol{P_j}}{\boldsymbol{P_i}}\right) \tag{1}$$

where $\boldsymbol{P_i} = (\boldsymbol{x_i}/\sum_{a=1}^{L} x_{ia})$ and $\boldsymbol{P_j} = (\boldsymbol{x_j}/\sum_{a=1}^{L} x_{ja})$ refer to the desired probability vectors resulting from the pixel vector $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$.

2) *Correlation Coefficient*:

$$\boldsymbol{D}_{\boldsymbol{CC}}^{\boldsymbol{ij}} = \frac{\mathrm{cov}(\boldsymbol{x_i}, \boldsymbol{x_j})}{\sqrt{\mathrm{var}(\boldsymbol{x_i})} \cdot \sqrt{\mathrm{var}(\boldsymbol{x_j})}} \tag{2}$$

where $\mathrm{cov}(\boldsymbol{x_i}, \boldsymbol{x_j})$ represents the covariance of $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$, $\mathrm{var}(\cdot)$ is the standard deviation.

3) *Spectral Angle Mapper*:

$$\boldsymbol{D}_{\boldsymbol{SAM}}^{\boldsymbol{ij}} = \cos^{-1}\left(\frac{\boldsymbol{x_i}^T \boldsymbol{x_j}}{\|\boldsymbol{x_i}\| \, \|\boldsymbol{x_j}\|}\right) \tag{3}$$

where $\|\boldsymbol{x_i}\| = (\sum_{a=1}^{L} x_{ia}^2)^{\frac{1}{2}}$ and $\|\boldsymbol{x_j}\| = (\sum_{a=1}^{L} x_{ja}^2)^{\frac{1}{2}}$.

4) *Spectral Gradient Angle*:

$$\boldsymbol{D}_{\boldsymbol{SGA}}^{\boldsymbol{ij}} = \boldsymbol{D}_{\boldsymbol{SAM}}^{\boldsymbol{ij}}\{G(\boldsymbol{x_i}), G(\boldsymbol{x_j})\} \tag{4}$$

where $G(\boldsymbol{x_i})$ and $G(\boldsymbol{x_j})$ represent to the gradient vector of $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$. The formula is defined by

$$\begin{aligned} G(\boldsymbol{x_i}) &= [x_{i2} - x_{i1}, x_{i3} - x_{i2}, \ldots, x_{iL} - x_{i(L-1)}] \\ G(\boldsymbol{x_j}) &= [x_{j2} - x_{j1}, x_{j3} - x_{j2}, \ldots, x_{jL} - x_{j(L-1)}]. \end{aligned} \tag{5}$$

B. Constrained Energy Minimum

Assuming that $\boldsymbol{d}$ is the known prior information as the target spectral signal to be detected. The key idea of the CEM algorithm is to design a finite impulse response (FIR) linear filter so that the output energy of the filter is minimized when the following constraint condition is satisfied

$$\boldsymbol{d^T w} = \sum_{l=1}^{L} d_l w_l = 1 \tag{6}$$

where $\boldsymbol{w} = [w_1, w_2, \ldots, w_L]^T$ is an L-dimensional column vector consisting of filter coefficient $\{w_1, w_2, \ldots, w_L\}$.

Then, let the output of the abovementioned FIR filter corresponding to the input $\boldsymbol{r_i}$ is $\boldsymbol{y_i}$, the formula is expressed as follows:

$$y_i = \sum_{l=1}^{L} w_l x_{iL} = \boldsymbol{w^T x_i} = \boldsymbol{x_i^T w}. \tag{7}$$

Next, for all pixels $\{x_1, x_2, \ldots, x_N\}$, the average energy output by the filter is defined as follows:

$$\begin{aligned} E &= \frac{1}{N}\sum_{i=1}^{N} y_i^2 = \frac{1}{N}\sum_{i=1}^{N} \left(\boldsymbol{x_i^T w}\right)^T \boldsymbol{x_i^T w} \\ &= \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w^T x_i x_i^T w} = \boldsymbol{w^T}\left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x_i x_i^T}\right)\boldsymbol{w} \\ &= \boldsymbol{w^T R w} \end{aligned} \tag{8}$$

where $\boldsymbol{R} = (1/N)\sum_{i=1}^{N} \boldsymbol{x_i x_i^T}$ represents the autocorrelation matrix. Therefore, the CEM algorithm can be regarded as an optimization problem with linear constraints, the formula is as follows:

$$\min_{w}\{E\} = \min_{w}\{\boldsymbol{w^T R w}\} \quad s.\,t. \quad \boldsymbol{d^T w} = 1. \tag{9}$$

The optimization problem with linear constraints can be solved by

$$F(\boldsymbol{w}) = \boldsymbol{w^T R w} + \lambda\left(\boldsymbol{d^T w} - 1\right) \tag{10}$$

where λ refers to lagrangian multiplier. To minimize the abovementioned function, let it be equal to zero after partial derivative of w. The formula is expressed as follows:

$$\begin{aligned} \frac{\partial F(\boldsymbol{w})}{\partial \boldsymbol{w}} &= \frac{\partial \boldsymbol{w^T R w}}{\partial \boldsymbol{w}} + \lambda \boldsymbol{d} \\ &= \left(\boldsymbol{R} + \boldsymbol{R^T}\right)\boldsymbol{w} + \lambda \boldsymbol{d} \\ &= 2\boldsymbol{R w} + \lambda \boldsymbol{d} \\ &= 0 \end{aligned} \tag{11}$$

where $\boldsymbol{R}$ is a symmetric matrix. Equation (6) can be, then, solved by

$$\boldsymbol{w} = -\frac{1}{2}\lambda \boldsymbol{R}^{-1}\boldsymbol{d}. \tag{12}$$

Meanwhile, the optimal solution to (4) can be defined by

$$\boldsymbol{w}_{\mathrm{CEM}} = \frac{\boldsymbol{R^{-1}d}}{\boldsymbol{d^T R^{-1} d}}. \tag{13}$$

Finally, the CEM detector, $\delta_{\mathrm{CEM}}(\boldsymbol{x})$ derived in [43], is specified by the weight vector $\boldsymbol{w}_{\mathrm{CEM}}$ in (13) and given by

$$\delta_{\mathrm{CEM}}(\boldsymbol{x_i}) = (\boldsymbol{w}_{\mathrm{CEM}})^T \boldsymbol{x_i}. \tag{14}$$

III. Description of the Proposed Approach

Different from the traditional detector CEM using single-layer, the proposed HCEM method consists of different layers of the CEM detectors, and the detectors of different layers are linked in series. The hierarchical structure based constrained energy minimum method is introduced for the first time in a noisy label detection framework for HSI supervised classification (see Fig. 1), which consists of three major components: 1) construction of confidence spectrum; 2) description of hierarchical structure based CEM; and 3) cleanse the training set with noisy labels. The pseudocode of the proposed HCEM method is

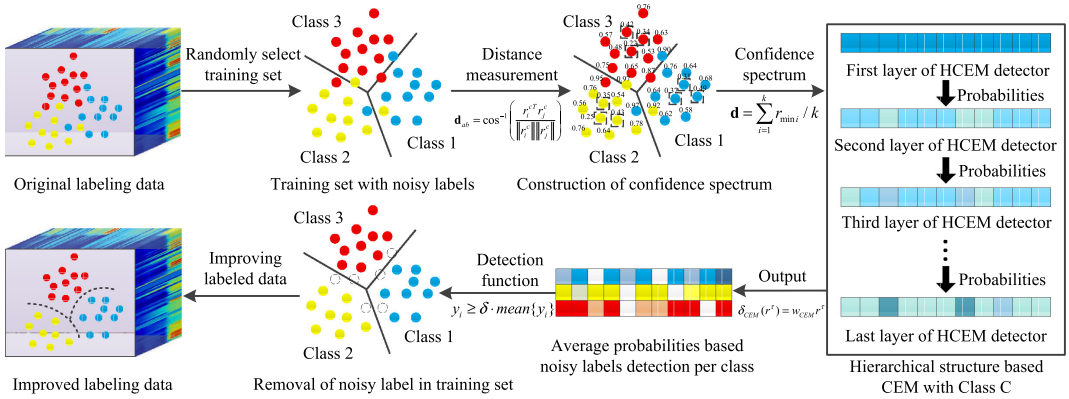


Fig. 1. Schematic diagram of the proposed method for noisy label detection in HSI. Different colors represent different classes of the training samples.

Algorithm 1: The Proposed HCEM Method.

Inputs: Original training set with noisy labels: $\boldsymbol{r} = \{\boldsymbol{r_1^c}, \boldsymbol{r_2^c}, ..., \boldsymbol{r_n^c}\} \in \mathbb{R}^{m \times n}$, $c \in (1, 2, ..., L)$

Outputs: Improved training set $\boldsymbol{u}$ and classification results obtained by the SVM trained with $\boldsymbol{u}$.

For each $c = 1 : L$ **do**.
  Calculate the SAM-based distance information $\boldsymbol{S_{ij}^c}$ among each sample for cth class.
  Select a few hypothetical true sample spectra according to the spectral similarity domain.
  Construct confidence spectrum $\boldsymbol{d}$ based on the mean operation of spectra of hypothetical true samples.
  For each $z = 1 : \tau$ **do**.
    Achieve the CEM-based filter output $\boldsymbol{y(z)^c}$ for each sample per class.
    Implement spectral nonlinear suppression $\boldsymbol{r(z+1)_i^c} = q\{\boldsymbol{y(z)_i^c}\}\boldsymbol{r(z)_i^c}$.
  End For
  Cleanse the original training set according to the hierarchical operation based filter output $\boldsymbol{y(z)^c}$.
  Obtain improved training set $\boldsymbol{u} = [\boldsymbol{u^1}, \boldsymbol{u^2}, ..., \boldsymbol{u^L}]$.
End For
Classification results can be obtained by feeding the improved training set to the SVM classifier.

illustrated in Algorithm 1. Each component contains some key steps, the details of which are shown as follows.

A. Construction of Confidence Spectrum

Let us assume that the original training set is given by $\boldsymbol{r} = \{\boldsymbol{r_1^c}, \boldsymbol{r_2^c}, \ldots, \boldsymbol{r_n^c}\} \in \mathbb{R}^{m \times n}$, where c refer to the cth class in original training set $c \in (1, 2, \ldots, L)$, n represents the nth samples in the cth class, and m is denoted as the dimension for each sample. The SAM-based distance information among samples per class can be first defined by

$$\boldsymbol{S_{ij}^c} = \cos^{-1}\left(\frac{\boldsymbol{r_i^{cT} r_j^c}}{\left\|\boldsymbol{r_i^c}\right\| \cdot \left\|\boldsymbol{r_j^c}\right\|}\right). \tag{15}$$

After calculating the distances among samples, an indicator that judge whether a sample is a true training sample or a noise label can be obtained by

$$\boldsymbol{D_i^c} = \sum_{j=1}^{n} \boldsymbol{S_{ij}^c} \quad (i = 1, 2, \ldots, n). \tag{16}$$

By this way, the confidence coefficient (inversely proportional to the distance information) of each sample can be constructed as $\boldsymbol{D^c} = [D_1^c, D_2^c, \ldots, D_n^c]$. Then, the top k samples $\hat{\boldsymbol{r}} = \{\boldsymbol{\hat{r}_1^c}, \boldsymbol{\hat{r}_2^c}, \ldots, \boldsymbol{\hat{r}_k^c}\}$ with greater confidence coefficient are selected based on the ascending order of the confidence coefficient $\boldsymbol{D^c}$. Finally, the confidence spectrum $\mathbf{d^c}$ can be achieved by conducting mean operation on the top k minimum spectral set. The formula is defined as follows:

$$\boldsymbol{d^c} = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{\hat{r}_i^c}. \tag{17}$$

B. Description of Hierarchical Structure-Based CEM

Taking into account the zth layer, the CEM-based filter output of samples $\boldsymbol{r^c}$ (called as $\boldsymbol{r(z)^c}$ in following equations) in original training set can be expressed by:

$$\boldsymbol{y(z)^c} = \frac{\boldsymbol{R_z^{-1} d}}{\mathbf{d}^T \boldsymbol{R_z^{-1} d}} \boldsymbol{r(z)^c} \tag{18}$$

where $\boldsymbol{r(z)^c}$ and $\boldsymbol{R_z}$ refer to the spectral matrix and the correlation matrix of the zth layer, respectively. Then, each training sample $\boldsymbol{r(z)_i^c}$ of original training set is transformed by multiplying a nonnegative number $\boldsymbol{y(z)_i^c}$ based on its output score as follows:

$$\boldsymbol{r(z+1)_i^c} = q(\boldsymbol{y(z)_i^c})\boldsymbol{r(z)_i^c} \tag{19}$$

where the $q(\cdot)$ as a nonlinear function is used to impose on the spectral vector of a training sample $\boldsymbol{r(z)_i^c}$. The function is regarded as a "soft-threshold" operation: retain the spectrum $\boldsymbol{r(z)_i^c}$ whose output score is large, while suppress the spectrum $\boldsymbol{r(z)_j^c}$ whose output score is small. By this way, the spectra of noisy labels is gradually suppressed after each detection of layer, while the spectra of true samples will maintain unchanged. The nonlinear suppression function is defined as follows:

$$q(\theta) = \begin{cases} 1 - e^{-\lambda\theta} & \theta \geq 0 \\ 0 & \theta < 0 \end{cases} \tag{20}$$

where λ refers to a free parameter controlling the shape of the function. Finally, the true samples spectra and the transformed spectra of noisy label will be used to construct the new CEM detector in the $(z+1)$th layer. The abovementioned steps will be repeated until the output $\boldsymbol{y^c}$ converges to a constant. In this article, we calculate δ_z, the error of the average output energy of the current layer and the previous layer, as follows:

$$\delta_z = \frac{1}{N}\left\|\mathbf{y}(\mathbf{z}+\mathbf{1})^c\right\|_2^2 - \frac{1}{N}\left\|\mathbf{y}(\mathbf{z})^c\right\|_2^2 \tag{21}$$

where the iteration will be stopped when $\delta_z < \epsilon$ (ϵ represents a small positive number).

C. Cleanse the Training Set With Noisy Labels

Once the output score of the training samples of each class has been obtained, the noisy label of the original training set can be easily detected and removed as follows:

$$\mathbf{D} = \begin{cases} \boldsymbol{r}_i, \text{ if } \mathbf{y}_i \geq \alpha \cdot \bar{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{c}} \\ \varnothing, \text{ Otherwise} \end{cases} \tag{22}$$

where α is a free parameter controlling the size of decision threshold for each class, which is set by the optimal results of experiments under the SVM trained with the improved training set. Finally, the improved training set is represented by $\boldsymbol{u} = [\boldsymbol{u^1}, \ldots, \boldsymbol{u^c}, \ldots, \boldsymbol{u^L}]$.

IV. Theory-Based Performance Analysis for the Proposed HCEM Method

In this section, the output energy of the proposed HCEM method is given theoretically to explain the reason why in each layer we can learn a better CEM detector than previous layers, and how the detection performance is gradually enhanced.

Suppose that $\boldsymbol{E}(z)$ refers to the residual error in the zth layer for the proposed HCEM method. The residual error can be represented as follows:

$$\begin{aligned} \boldsymbol{E}(z) &= \frac{1}{n}\left\|\boldsymbol{y(z)^c} - \boldsymbol{o}\right\|_2^2 \\ &= \frac{1}{n}(\left\|\boldsymbol{y(z)^c}\right\|_2^2 - 2\boldsymbol{y(z)^c o}^T + \left\|\boldsymbol{o}\right\|_2^2) \\ &= \frac{1}{n}\left\|\boldsymbol{y(z)^c}\right\|_2^2 - \frac{2}{n}\mathbf{w}_z^T\boldsymbol{r(z)^c o}^T + \frac{1}{n}\left\|\boldsymbol{o}\right\|_2^2 \end{aligned} \tag{23}$$

where $\boldsymbol{o} = [o_1, o_2, \ldots, o_N] \in \mathbb{R}^{1\times N}$ refers to the binary attribute label of the input sample $\boldsymbol{x_i^c}$, if the class of $\boldsymbol{x_i}$ is the same as the corrected class, $o_i = 1$, otherwise $o_i = 0$. To explain (23) clearly, $\boldsymbol{R}(z)$ will be broken down into three terms. The first term can be transformed as follows:

$$\begin{aligned} \frac{1}{n}\left\|\boldsymbol{y(z)^c}\right\|_2^2 &= \frac{1}{n}\left\|\mathbf{w}_z^T\boldsymbol{r(z)^c}\right\|_2^2 \\ &= \frac{1}{\boldsymbol{d}^T\boldsymbol{R}_z^{-1}\boldsymbol{d}}. \end{aligned} \tag{24}$$

For the second term, we suppose that all the spectrums of true sample are not be suppressed during the hierarchically filtering process. Then, the equation $\boldsymbol{r(z)^c o}^T = n_k\boldsymbol{d}$ can be gained easily, in which n_k refers to the number of spectra constructing confidence spectra. Furthermore, this term subject to $\mathbf{w}^T\mathbf{d} = 1$, the second term can be transformed as follows:

$$\begin{aligned} \frac{2}{n}\mathbf{w}_z^T\boldsymbol{r(z)^c o}^T &= \frac{2n_k}{n}\mathbf{w}_z^T\mathbf{d} \\ &= \frac{2n_k}{n}. \end{aligned} \tag{25}$$

The third term can be transformed by

$$\frac{1}{n}\left\|\boldsymbol{o}\right\|_2^2 = \frac{n_k}{n}. \tag{26}$$

Finally, the simplified form $\boldsymbol{R}(z)$ can be achieved by substituting (24)–(26) into (23), the formula is defined by

$$\boldsymbol{E}(z) = \frac{1}{\mathbf{d}^T\boldsymbol{R}_z^{-1}\mathbf{d}} - \frac{n_t}{n}. \tag{27}$$

According to (20), the residual error $\boldsymbol{E}(z+1)$ has a negative relationship with the correlation matrix inverse $\boldsymbol{R}_{z+1}^{-1}$. If $\boldsymbol{R}_{z+1}^{-1}$ is higher than $\boldsymbol{R}_z^{-1}$, the residual error $\boldsymbol{E}(z+1)$ will be smaller compare with $\boldsymbol{E}(z)$. Here, to further prove the performance of $z+1$th detector will get better effect than last layer, The relationship of correlation matrices between the $z+1$th layer and the zth layer is clearly deduced as follows:

$$\begin{aligned} \boldsymbol{R}_{z+1} - \boldsymbol{R}_z &= \sum_{i=1}^{n}\frac{q_i^2\boldsymbol{x_i^z}\boldsymbol{x_i^z}^T - \boldsymbol{x_i^z}\boldsymbol{x_i^z}^T}{n} \\ &= \boldsymbol{V}\boldsymbol{V}^T \end{aligned} \tag{28}$$

where $\boldsymbol{V} = [\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots\ldots, \boldsymbol{v_n}] \in \mathbb{R}^{m\times n}$, and $\boldsymbol{v_i} = \sqrt{(1-q_i^2)/n}\cdot\boldsymbol{r(z)_i^c}$. Suppose that $\boldsymbol{R}_z$ and $(\boldsymbol{R}_z + \boldsymbol{v_i}\boldsymbol{v_i}^T)$ are all invertible. Then, based on the Shermann–Morrison–Woodbury formula[44], we can get the equation easily

$$(\boldsymbol{R_z} + \boldsymbol{v_i}\boldsymbol{v_i}^T)^{-1} = \boldsymbol{R_z}^{-1} - \frac{\boldsymbol{R_z}^{-1}\boldsymbol{v_i}\boldsymbol{v_i}^T\boldsymbol{R_z}^{-1}}{1+\boldsymbol{v_i}^T\boldsymbol{R_z}^{-1}\boldsymbol{v_i}}. \tag{29}$$

After that, we can get

$$\begin{aligned} \boldsymbol{d^T}\left(\boldsymbol{R_{z+1}^{-1}} - \boldsymbol{R_z^{-1}}\right)d &= \left\{\boldsymbol{R_{z+1}^{-1}} - \right. \\ &\quad \left.\left[\boldsymbol{R_{z+1}^{-1}} + (-\boldsymbol{V}\boldsymbol{V^T})\right]^{-1}\right\}\boldsymbol{d} \\ &= \sum_{i=1}^{n}\frac{\boldsymbol{d}^T\boldsymbol{R_{z+1}}^{-1}\boldsymbol{v_i}\boldsymbol{v_i}^T\boldsymbol{R_{z+1}}^{-1}\boldsymbol{d}}{1+\boldsymbol{v_i}^T\boldsymbol{R_{z+1}}^{-1}\boldsymbol{v_i}} \\ &= \sum_{i=1}^{n}\frac{\left\|\boldsymbol{d}^T\boldsymbol{R_{z+1}}^{-1}\boldsymbol{v_i}\right\|_2^2}{1+\boldsymbol{v_i}^T\boldsymbol{R_{z+1}}^{-1}\boldsymbol{v_i}}. \end{aligned} \tag{30}$$

Since $\|\boldsymbol{d}^T\boldsymbol{R_{z+1}}^{-1}\boldsymbol{v_i}\|_2^2 \geq 0$ and $\boldsymbol{v_i}^T\boldsymbol{R_{z+1}}^{-1}\boldsymbol{v_i} \geq 0$, the relationship $\boldsymbol{d}^T(\boldsymbol{R_{z+1}}^{-1} - \boldsymbol{R_z}^{-1})\boldsymbol{d} \geq 0$ can be determined. Thus,

$$\frac{1}{\boldsymbol{d}^T\boldsymbol{R_z}^{-1}\boldsymbol{d}} \geq \frac{1}{\boldsymbol{d}^T\boldsymbol{R_{z+1}}^{-1}\boldsymbol{d}}. \tag{31}$$

Due to the residual error $\boldsymbol{E}(z+1)$ has a negative relationship with the correlation matrix inverse $\boldsymbol{R_{z+1}}^{-1}$, the relationship can

be achieved by

$$\boldsymbol{E}(z) \geq \boldsymbol{E}(z+1). \tag{32}$$

According to the abovementioned theory derivation, it can be obtained that, with the number of layer increase, the residual error will reduce gradually until the number of layers reach the threshold. Then, the residual error will converge to zero. The difference residual error caused different output of each layer, the residual error is close to zero, it means that the output energy of the proposed HCEM method will converge to the output energy of last layer detector. In general, the output energy in each layer will increase compare with last layer. When the number of layer reach the threshold, the output energy will close to a constant.

V. Experimental Result on Real HSI Datasets

A. Experimental Setup

In this article, the performance of the proposed HCEM methods is evaluated using four hyperspectral real datasets. The hyperspectral image of University of Houston was acquired with NSF-funded Center for Airborne Laser Mapping (NCALM). The hyperspectral scene of Salinas Valley was acquired by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over Salinas Valley in California, U.S. The third hyperspectral image is the Kennedy Space Center (KSC) dataset acquired by the AVIRIS over the Kennedy Space Center, Floride. The hyperspectral scene of the Washington DC dataset was acquired by the NCALM) over the University of Houston campus and the neighboring urban area. The four datasets are detailed as follows.

1) *University of Houston Dataset*: The Houston University image was acquired over the Houston University campus and its neighboring area, which was used in the 2013 GRSS Data Fusion Contest. The hyperspectral data contains 144 spectral bands in the 380 nm–1050 nm region, and 349×1905 pixels with a spatial resolution of 2.5 m. This image is an urban datasets whose most of the land covers are man-made objects, which contains fifteen classes. Fig. 2(a)–(c) shows the false-color composite of the University of Houston image, the corresponding reference data, and color coding, respectively.

2) *Salinas Valley dataset*: The Salinas Valley image was also acquired by the AVIRIS sensor over Salinas Valley, California, U.S. The image is of size $512 \times 217 \times 224$ with a spatial resolution of 3.7 m per pixel. 20 water absorption spectral bands (no. 108–112, 154–167, and 224) were removed and the reference image has 16 different classes. Fig. 3(a)–(c) shows the color composite of the Salinas Valley image, the corresponding reference data, and color coding.

3) *KSC dataset*: The image was acquired by the AVIRIS over the Kennedy Space Center, Floride. The image is of size 512×614 pixels, in which 48 bands are removed as water absorption and low SNR bands. The false color composite of the KSC image, the reference classification map, and color coding are shown in Fig. 4(a)–(c).

4) *Washington DC dataset*: The Washington DC image was acquired by the NCALM) over the University of Houston campus and the neighboring urban area. The image consists of 280×307 pixels, each pixel including 210 spectral bands. Bands ranging ranges from 0.4 to 2.5 μm and the spatial resolution of the image is 3 m per pixel. In the experiments, bands ranging from 0.9 to 1.4 μm, where the atmosphere of these bands is opaque, are discarded from the dataset, leaving 191 bands. Fig. 5(a)–(c), respectively, demonstrates the false-color composite of the Washington DC image, the corresponding reference data, and color coding, which considers six classes of interest.

In this article, a pixelwise classifiers SVM, as a baseline, is employed to test the effectiveness of the proposed HCEM method for noisy label detection. Specifically, the SVM is implemented with the LIBSVM library[45] by adopting the radial basis function kernel. The parameters of the SVM are determined using five-fold cross-validation step. In addition, in order

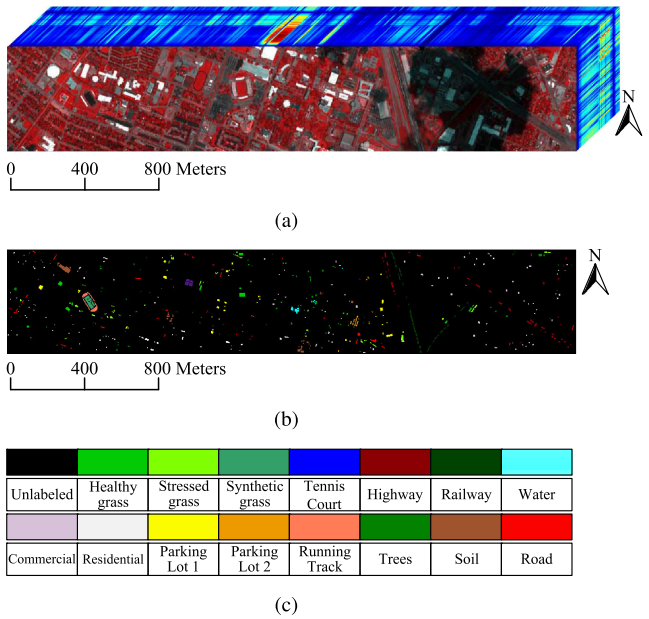


Fig. 2. University of Houston dataset. (a) Three-band color composite. (b) Reference data. (c) Color coding.

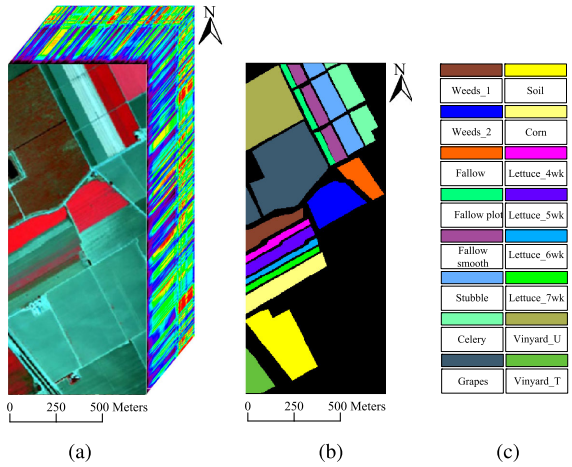


Fig. 3. Salinas Valley dataset. (a) Three-band color composite. (b) Reference data. (c) Class coding.

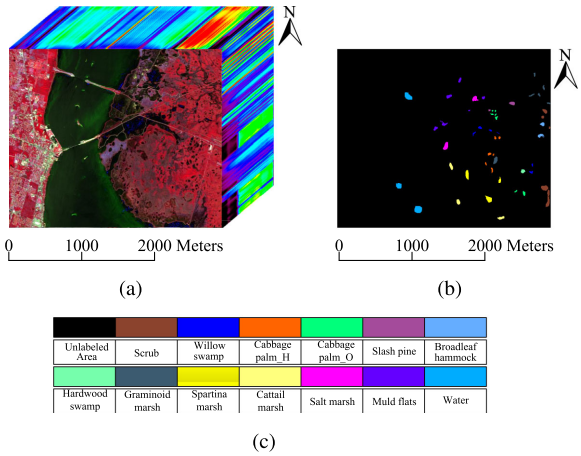


Fig. 4. KSC dataset. (a) Three-band color composite. (b) Reference data. (c) Class coding.

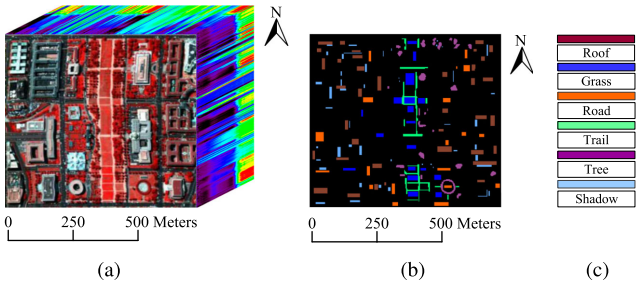


Fig. 5. Washington DC dataset. (a) Three-band color composite. (b) Reference data. (c) Color coding.

to make the comparison objective and fair, in all experiments, the widely used quality indexes, such as overall accuracy (OA), average accuracy (AA), Kappa coefficient (Kappa), are computed by averaging the results achieved after ten repeated Monte Carlo experiments with various randomly selected training samples and noisy labels, the mean and the standard deviation after such experiments are given in our experiments. However, no ready specific sources of noisy labels are available for our analysis. In order to construct the mislabeled samples of these three hyperspectral datasets and to further simulate the problem of "noisy labels", randomly selected samples from other classes are added to each class.

B. Analysis of the Influence of Spectral Similarity Metrics

In this section, an experiment is performed to analyze the influence of the different spectral similarity metrics to the performance of the proposed HCEM method. The basic idea behind this experiment is to construct the confidence spectral vector in each class of the original training set through the best spectral similarity metric. The confidence level of the preset confidence spectrum will determine the performance of the HCEM for noisy label detection. This experiment is conducted on the different HSI datasets with 25 true samples and five noisy labels for each class to evaluate the measure performance of the SID, CC, SGA, and SAM metrics in noisy label detection. The experiment results are represented in Table I. It can be observed that the HCEM based on different metrics has improved the classification accuracy of the SVM to different degrees on the University of Houston, Salinas Valley, KSC, and Washington DC datasets, respectively. Specifically, the SVM trained with training set improved by the SAM-based HCEM method can achieve higher classification accuracy with respect to SID, CC, and SGA metrics. In addition to comparing classification accuracy, we counted the running time that obtaining the classification results of the SVM classifier trained using improved training sets. It can be seen from Table I that the SAM-based HCEM method takes less uptime on the different dataset. Therefore, according to the optimal experimental results, the SAM metric is selected as a component of the proposed HCEM method in follow-up experiments.

C. Analysis of the Influence of the Parameters

In this section, the influence of the parameters relevant to the performance of the proposed HCEM method is analyzed, such as decision threshold for removing noisy labels α, controlling the shape of the nonlinear function λ, the number of layer z, proportion (in %) of top k minimum spectrum per class, and residual between two layers based stopping criterion ϵ. The experiments are, respectively, conducted on four real hyperspectral datasets. For the University of Houston, Salinas Valley, KSC, and Washington DC datasets, the original training set consists of 25 true training samples and five noisy labels per class, respectively.

In the first experiment, the influence of the decision threshold for removing noisy labels α and the controlling the shape of the nonlinear function λ on the performance of the proposed HCEM method is verified in the classification of the aforementioned hyperspectral datasets. The variation range of the parameter α and λ are set to $\{0, 0.04, \ldots, 0.3\}$ and $\{1, 3, \ldots, 11\}$, respectively. As shown in Fig. 6, it can be observed that the OAs of the classification results achieved by the SVM trained with training set improved by the HCEM show close relationship with the variation of the parameter value. For example, when the parameter α is fixed, the OAs will first increase and then decrease with the change of λ. In fact, it indicates that the shape of the suppression function will exist a direct impact. Similarly, when fixing λ and focusing on α, it can be observed from Fig. 6 that the variation of the α has a more significant effect on classification accuracy. The reason is that the α is the control amount of the noise label removal degree per class in the original training set, which can directly determine the true sample purity of the improved training set. Therefore, an optimal parameter value can be determined based on the highest classification accuracy of SVM, using the improved training set. It can be seen from Fig. 6 that the highest OA of SVM using the improved training set on the University of Houston, Salinas Valley, KSC, and Washington DC datasets is OA = 82.15% ($\alpha = 0.14$, $\lambda = 1$), OA = 87.50% ($\alpha = 0.28$, $\lambda = 4$), OA = 87.24% ($\alpha = 0.14$, $\lambda = 3$), and OA = 86.18% ($\alpha = 0.20$, $\lambda = 2$), respectively. Therefore, these values of α and λ are set to default parameters corresponding to different datasets in this article.

TABLE I
CLASSIFICATION PERFORMANCE OBTAINED BY THE HCEM METHOD USING DIFFERENT SPECTRAL DISTANCE METRICS WITH 25 TRUE SAMPLES AND FIVE NOISY LABELS. NUMBER IN THE PARENTHESIS REPRESENTS THE STANDARD VARIANCE OF THE ACCURACIES OBTAINED IN TEN REPEAT EXPERIMENTS

Date Set	SVM 25 (T)	SVM 25 (T)+5(N)	The different spectral metrics methods based HCEM			
			SID [39]	CC [40]	SGA [41]	SAM [42]
University of Houston	84.47(1.37)	78.29(1.63)	80.09(1.32)	77.71(1.06)	79.07(1.31)	**81.08(1.19)**
Salinas Valley	85.81(1.52)	81.79(1.69)	84.64(2.07)	84.64(1.02)	84.96(0.99)	**85.64(1.31)**
KSC	88.75(0.96)	83.57(1.74)	84.36(1.81)	83.66(1.72)	80.31(4.59)	**86.25(1.72)**
Washington DC	87.93(1.76)	85.61(1.33)	82.37(3.79)	83.33(2.35)	80.61(4.12)	**87.13(2.81)**
Time(s)	11.4∼105.1	12.5∼93.1	7.9∼87.8	7.1∼81.4	6.5∼78.5	7.1∼75.8

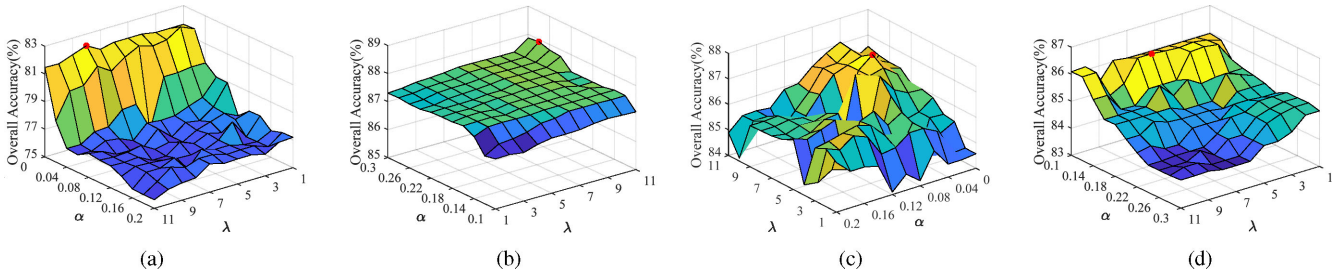


Fig. 6. Influence of the parameter α and λ to the HCEM method on different real HSI datasets with 25 true samples and 5 noisy labels per class. (a) University of Houston dataset. (b) Salinas Valley dataset. (c) KSC dataset. (d) Washington DC dataset.

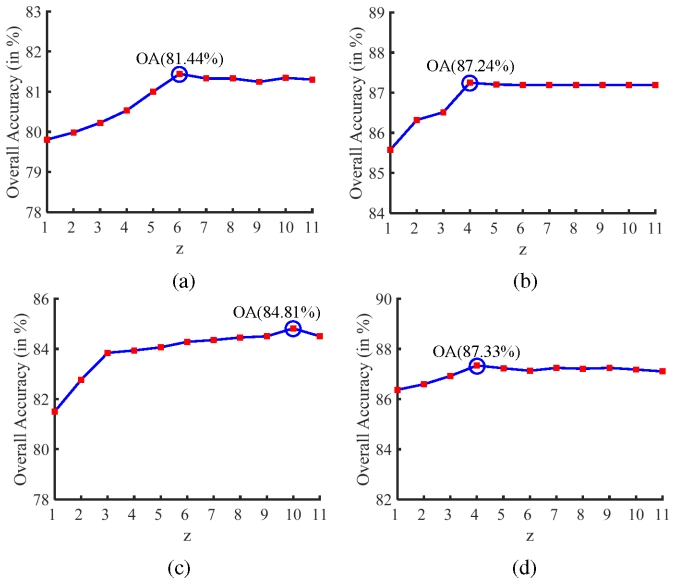


Fig. 7. Influence of parameters z on the performance of the proposed HCEM method on different datasets with 25 true samples and 5 noisy labels per class. (a) University of Houston dataset. (b) Salinas Valley dataset. (c) KSC dataset. (d) Washington DC dataset.

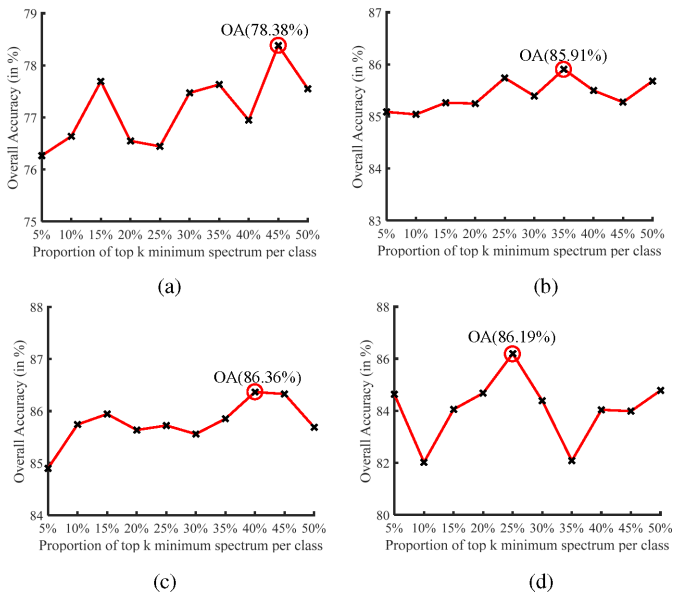


Fig. 8. Influence of parameters k on the performance of the proposed HCEM method on different datasets with 25 true samples and 5 noisy labels per class. (a) University of Houston dataset. (b) Salinas Valley dataset. (c) KSC dataset. (d) Washington DC dataset.

In the second experiment, the effectiveness of the number of layer z and proportion (in %) of top k minimum spectrum per class on the performance of the proposed HCEM method is tested in the classification before-mentioned hyperspectral datasets. The number of layer z is chosen from $z = 1$ to $z = 11$. Note that the detection performance of the CEM detector similarity to the HCEM of one layer. As shown in Fig. 7, it can be observed that the parameter z can make the classification accuracy of the SVM classifier trained with improved training set tends to increase gradually, but converge to a constant finally. Focusing on Fig. 7(a), the classification accuracy achieves the highest value OA = 81.4% when z is set to six on the University of Houston dataset. For the other datasets such as the Salinas Valley (OA = 87.30%), KSC (OA = 84.81%), and Washington DC (OA = 87.20%) datasets, the optimal OAs will be achieved when z is set 4, 10, and 4, respectively. Similarly, the optimal proportion of the top k minimum spectra is determined based on the similar spectral metric (see Fig. 8) on the various HSI real datasets. Therefore, default parameters (the number of layer z and proportion (in %) of top k minimum spectrum per class)

TABLE II
DETECTION PERFORMANCE (NUMBERS) OF NOISY LABELS FOR THE PROPOSED HCEM METHOD ON DIFFERENT DATASETS. WHERE $T_n \times L$ IS THE TOTAL NUMBER OF NOISY LABELS IN TRAINING SET, T_n IS THE NUMBER OF MISLABELED SAMPLES PER CLASS, AND THE NUMBER IN THE PARENTHESIS REFERS TO THE STANDARD VARIANCE OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Datasets	University of Houston		Salinas Valley		KSC		Washington DC	
Total	5×15	15×15	5×16	15×16	5×13	15×13	5×6	15×6
Correct	42.60 (1.94)	128.3 (2.08)	52.60 (2.54)	143.5 (4.85)	36.00 (3.82)	92.00 (10.4)	15.80 (3.82)	40.60 (5.44)
Incorrect	23.00 (2.44)	7.660 (3.05)	8.200 (1.98)	3.200 (2.69)	19.10 (5.36)	15.50 (3.62)	6.800 (5.22)	2.500 (1.95)

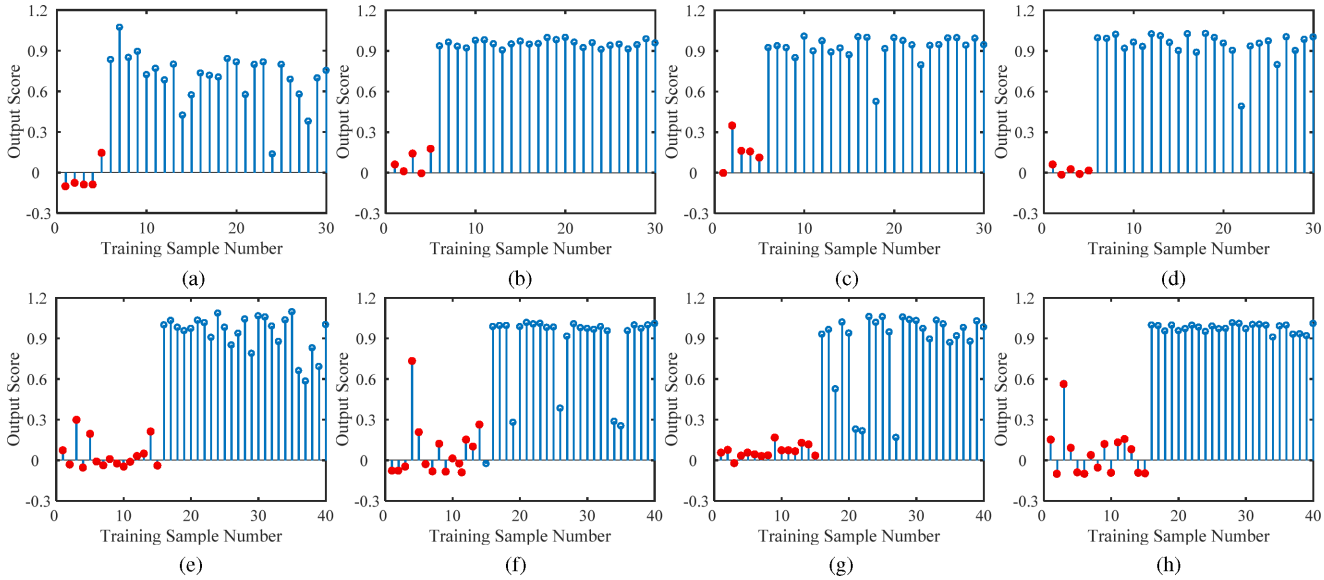


Fig. 9. Illustration of the HCEM-based output score in different classes for the four hyperspectral real datasets. (a) and (e) University of Houston (25 true samples and five/fifteen noisy labels). (b) and (f) Salinas Valley (25 true samples and five/fifteen noisy labels). (c) and (g) KSC (25 true samples and five/fifteen noisy labels). (d) and (h) Washington DC dataset (25 true samples and five/fifteen noisy labels).

of the proposed HCEM method for the different datasets are confirmed based on the highest OA.

Furthermore, in addition to analyzing the number of layer z and the proportion of top k minimum spectrum per class, here, we have added a discussion of the residual between two layers based stopping criterion ϵ. Since the parameters ϵ and z have an opposite relationship to the performance of HCEM, namely, the larger the parameter ϵ, the lesser the HCEM layering will be. This article focuses on the hierarchical performance of the proposed HCEM method, so according to the best classification result the parameters ϵ are set to 10^{-4} (University of Houston), 10^{-2} (Salinas Valley), 10^{-3} (KSC), and 10^{-2} (Washington DC), respectively.

D. Evaluation of Detection Performance

In this section, detection performance (number) of the proposed HCEM method is analyzed to demonstrate effectiveness of the proposed method in HSI supervised classification. The numbers of noisy labels that are detected correctly and falsely on the University of Houston, Salinas Valley, KSC, and Washington DC datasets are shown in Table II. It can be seen that a small number of samples are detected falsely (see the first row). This means that the proposed method can be widely used in the preprocessing of HSI supervised classification. In addition, the score (see 22) distribution (certain class of four datasets) on the four real datasets is visualized to show the difference in sample output score after the HCEM method processing. In fact, the output score serves as an objective reflection of the suppression effect of the HCEM method on the spectrum (true samples and noise labels). In other words, the output score will directly affect the detection performance of noisy labels of the HCEM method. As shown in Fig. 9, it can be observed that the output scores of the true samples (blue circle) and the noise labels (red dot) are significantly distinguishable, regardless of whether the original training set contains 5 noise labels or 15 noise labels in each class. However, there is also a phenomenon that the true training samples have relatively low output score in some classes [see Fig. 9(a), (f), and (g)]. The reason is that the true training sample in the original training set may be a sample with a high noise ratio, resulting in a lower output score.

E. Performance Verification With SVM Classifier

In this section, the classification performance of various methods (i.e., SALOF[29], DP[30], KECA[34], RLPA[28], and the proposed HCEM method) is compared by using the SVM classifier trained with different improved training sets on the

TABLE III
CLASSIFICATION PERFORMANCE OF THE SVM, SALOF, DP, RLPA, KECA, AND HCEM METHODS FOR THE UNIVERSITY OF HOUSTON DATASET WITH 25 TRUE SAMPLES AND DIFFERENT NUMBER OF NOISY LABELS PER CLASS AS TRAINING SET. NUMBER IN PARENTHESES REPRESENTS THE STANDARD VARIANCE OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Class	**SVM** 25 (true)	**The number of true samples and mislabeled samples**											
		25(true)+5(noisy)						25(true)+15(noisy)					
		SVM	SALOF	DP	RLPA	KECA	HCEM	SVM	SALOF	DP	RLPA	KECA	HCEM
1	93.60	81.97	86.94	88.16	81.00	90.20	**94.43**	87.84	87.64	86.82	76.64	88.16	**92.25**
2	94.29	83.68	85.96	**93.87**	86.14	92.98	93.46	88.64	89.07	87.37	87.04	**91.21**	90.63
3	98.08	80.24	96.77	99.00	**99.40**	95.01	97.24	83.63	84.21	89.19	**94.70**	87.82	89.32
4	95.17	92.77	92.37	95.65	87.97	94.64	**96.85**	90.71	87.88	94.18	98.00	92.34	**100.0**
5	91.13	89.52	91.70	**92.00**	86.27	89.03	90.57	88.63	89.11	**91.24**	86.91	90.11	85.81
6	93.13	69.31	77.31	92.87	**94.90**	83.06	84.45	78.41	71.97	80.50	69.62	87.80	**89.61**
7	82.69	76.96	72.27	70.23	79.44	79.12	**82.66**	71.56	73.20	69.01	72.87	78.48	**85.11**
8	79.28	77.84	71.76	72.19	73.40	76.95	**79.87**	77.06	72.19	77.32	87.05	77.50	**80.75**
9	73.99	66.53	70.36	**73.47**	76.02	69.69	64.18	66.71	68.99	67.66	67.80	**70.50**	60.52
10	80.44	**79.05**	73.43	73.97	67.67	77.12	74.52	67.63	**71.01**	65.89	65.51	75.89	70.32
11	77.18	74.53	72.19	74.04	77.18	73.17	**77.46**	59.90	64.55	66.33	68.36	**69.77**	69.15
12	73.86	71.82	69.44	68.65	**77.58**	70.78	67.22	62.72	65.89	64.90	63.84	**68.63**	58.19
13	47.89	48.98	39.70	39.11	28.62	47.88	**50.65**	33.35	39.57	42.09	30.50	46.05	**61.93**
14	88.54	77.66	86.22	79.77	83.19	**86.67**	84.90	83.77	84.48	80.77	87.73	83.17	**94.61**
15	99.39	85.07	96.18	**99.16**	97.74	97.16	97.45	93.81	93.25	94.04	92.87	92.99	**97.35**
OA	84.47	78.29	78.57	79.87	79.10	81.82	**82.01**	74.83	76.13	76.65	76.38	80.25	**80.66**
	(1.37)	(1.63)	(1.96)	(1.57)	(1.55)	(1.02)	(**1.19**)	(1.65)	(1.82)	(2.15)	(1.83)	(2.43)	(**2.09**)
AA	84.58	77.06	78.84	80.81	79.77	81.56	**82.39**	75.63	76.20	77.15	76.63	80.03	**81.70**
	(1.07)	(1.97)	(2.00)	(1.53)	(1.27)	(1.58)	(**1.57**)	(1.94)	(2.18)	(2.26)	(1.75)	(2.01)	(**1.94**)
Kappa	83.21	76.55	76.84	78.25	77.42	80.34	**80.55**	72.82	74.22	74.77	74.50	78.65	**79.08**
	(1.48)	(1.76)	(2.11)	(1.69)	(1.66)	(1.09)	(**1.29**)	(1.77)	(1.95)	(2.32)	(1.97)	(2.62)	(**2.25**)

University of Houston, Salinas Valley, KSC, and Washington DC datasets, respectively. All comparison method use the default parameters given in the reference, and the proposed method adopts the parameter setting presented earlier. For the University of Houston, Salinas Valley, KSC, and Washington DC datasets, the experiments are, respectively, performed with 25 true samples and different numbers of noisy labels in the range of 5 to 15 per class. As shown in Fig. 10, the classification results of the SVM trained with the different training sets are given, which include the average value and the standard deviation of the achieved OAs, AAs, and Kappas according to ten repeated experiments. It can be observed from Fig. 10 that the classification results obtained by the SVM on the improved training set are significantly better than the classification results obtained by the SVM on the original training set. It is worth mentioning that the SVM achieves best classification results on the training set improved by the proposed HCEM method in terms of OA, AA, and Kappa with respect to state-of-the-art detection methods (such as SALOF, DP, RLPA, and KECA). The experimental results on the University of Houston, Salinas Valley, KSC, and Washington DC dataset also demonstrate that the proposed HCEM method is more robust than SALOF, DP, RLPA, and KECA detection methods in the process of noisy label detection.

Moreover, the classification results obtained by the SVM trained with different training sets on the University of Houston dataset are presented in Table III. It can be observed from Table III that the proposed HCEM method can effectively improve the classification accuracies for most of the classes with respect to other detection methods. For example, focusing on healthy grass and stressed grass classes, the HCEM can promote the classification accuracy of the SVM classifier by about 4%–12% when each class of training samples contains five noisy labels. Specifically, when the noisy labels of each class of training samples are increased to 15, the classification accuracy of the SVM for the stressed grass class achieves 100%. Moreover, the HCEM achieves better classification results compared with other state-of-the-art detection methods in terms of the three important objective metrics (OA, AA, and Kappa). For example, when the original spectrum training set contains five noisy labels per class, OA, AA, and Kappa are increased from 78.29%, 77.06%, and 76.55% to 82.01%, 82.39%, and 80.55%, respectively. Specifically, as the number of noisy labels increases, the performance of the HCEM becomes more significant. When the original training set contains 15 noise labels, the three important objective metrics increase about 6%–7%. The visual performance comparison of different methods are represented in Fig. 11. As can be seen that the SVM trained with training set improved by the HCEM provides the best visual classification result compared with other test methods (see the zoom regions of Fig. 11). Therefore, It is fully demonstrated that the proposed HCEM method has obvious advantages of noisy label detection in HSI supervised classification.

Table IV and Fig. 12(a)–(k) show the experimental results of the state-of-the-art methods on the Salinas Valley dataset. As shown in Table IV, it can be seen that, when the number of the noisy labels increases, the improvement of classification accuracies of the SVM supervised classifier becomes more conspicuous. For instance, when the noisy original training set contains 25 true samples and five noisy label caused by mislabeling, the classification accuracy of the SVM trained with training set improved by the proposed HCEM method can be promoted about 4% with respect to the SVM trained with original training set, and improved about 2% compared with other detection methods (such as SALOF, DP, RLPA, and KECA methods). Specifically, the classification performance of

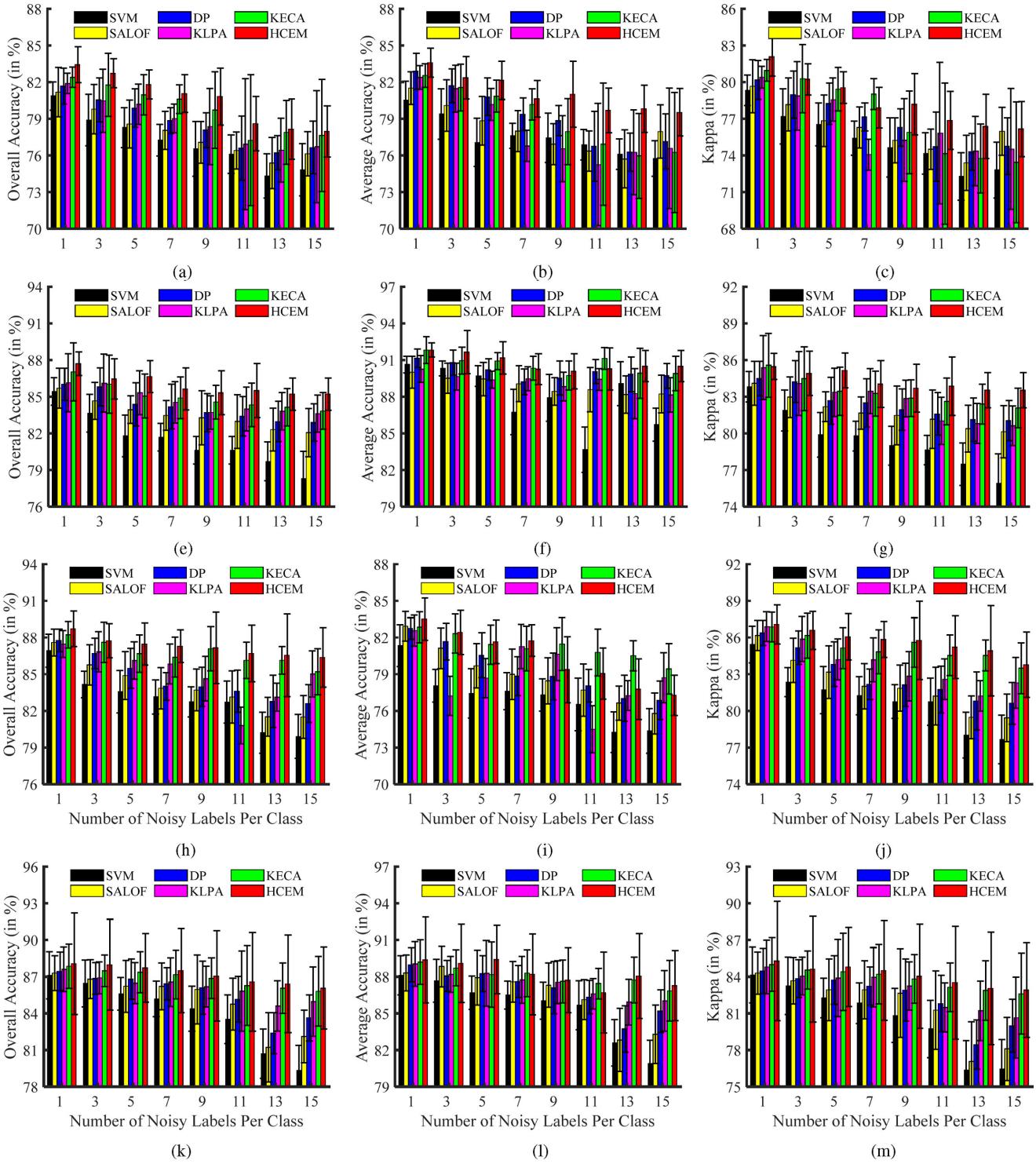


Fig. 10. Classification performance of the SVM (trained using the original training sets, and trained using the improved training sets obtained by the SALOF, DP, RLPA, KECA, and HCEM methods) in terms of OA (first column), AA (second column), and Kappa (third column). (a)–(c) Experiments on the University of Houston dataset with different number of noisy labels (varying from 1 to 15) and 25 true samples per class. (e)–(g) Experiments on the Salinas Valley dataset with different numbers of noisy labels (varying from 1 to 15) and 25 true samples per class. (h)–(j) Experiments on the KSC dataset with different number of noisy labels (varying from 1 to 15) and 25 true samples per class. (k)–(m) Experiments on the Washington DC dataset with different number of noisy labels (varying from 1 to 15) and 25 true samples per class.

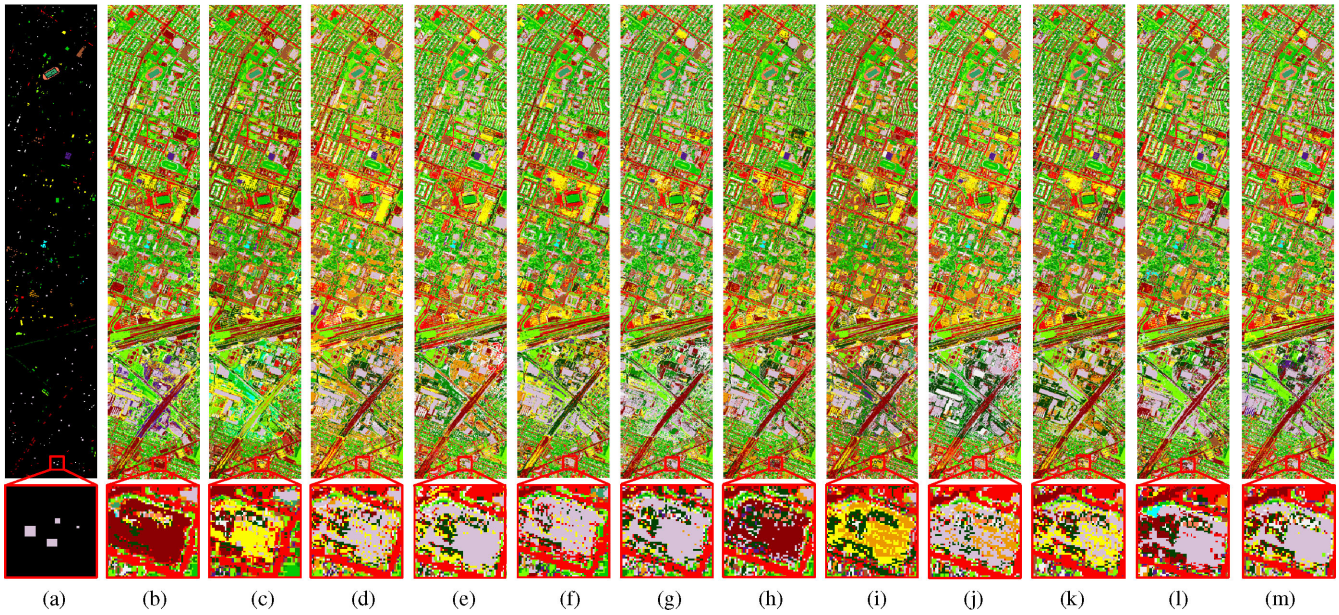


Fig. 11. Classification results (%) of comparing different method on the University of Houston dataset. The classification maps are obtained by the SVM [(b) and (h)], DP [(c) and (i)], SALOF [(d) and (j)], RLPA [(e) and (k)], KECA [(f) and (i)], and HCEM [(g) and (m)] trained with 25 true samples and different numbers of noisy labels. (b)– (f) contain 5 noisy labels and (g)–(k) contain 15 noisy labels. (a) Reference, (b) OA = 78.29%, (c) OA = 78.57%, (d) OA = 79.87%, (e) OA = 80.35%,(f) OA = 81.82%, (g) OA = 82.01%, (h) OA = 74.83%, (i) OA = 76.13%, (j) OA = 76.65%, (k) OA = 77.54%, (l) OA = 80.25%, and (m) OA = 80.66%.

TABLE IV
CLASSIFICATION PERFORMANCE OF THE SVM, SALOF, DP, RLPA, KECA, AND HCEM METHODS FOR THE SALINAS DATASET WITH 25 TRUE SAMPLES AND DIFFERENT NUMBER OF NOISY LABELS PER CLASS AS TRAINING SET. NUMBER IN PARENTHESES REPRESENTS THE STANDARD VARIANCE OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Class	**SVM** 25 (true)	**The number of true samples and mislabeled samples**											
		25(true)+5(noisy)						25(true)+15(noisy)					
		SVM	SALOF	DP	RLPA	KECA	HCEM	SVM	SALOF	DP	RLPA	KECA	HCEM
1	98.80	98.87	99.41	98.78	**100.0**	94.40	99.20	98.90	96.32	98.88	**99.75**	96.63	99.54
2	99.22	**99.33**	99.17	**99.33**	98.79	94.14	99.28	94.65	97.59	**98.29**	99.29	98.09	97.81
3	90.66	90.24	90.79	88.47	88.85	91.86	**94.56**	91.72	88.31	89.35	86.74	88.31	**94.91**
4	97.57	**97.38**	97.20	97.26	97.00	93.86	96.59	97.03	93.68	95.91	97.15	94.46	**99.19**
5	98.52	98.18	98.54	98.75	98.92	**99.37**	94.40	94.70	98.15	98.05	**99.72**	98.85	89.12
6	100.0	99.44	**100.0**	**100.0**	99.82	99.95	98.92	99.77	**100.0**	98.17	99.92	98.54	99.74
7	98.47	98.57	**98.58**	98.56	95.90	96.82	97.94	96.96	96.10	97.49	97.25	**98.25**	98.14
8	74.03	71.15	72.03	72.67	71.27	75.37	**77.46**	59.59	71.02	70.70	68.08	**73.39**	72.64
9	99.22	99.27	99.22	**99.28**	99.05	98.66	99.26	99.42	98.98	98.51	99.51	**99.59**	99.23
10	80.00	77.35	79.84	78.32	84.15	82.32	**89.84**	84.80	84.44	81.33	79.33	80.73	**87.83**
11	85.46	87.78	85.95	88.42	89.33	83.89	**91.36**	83.47	88.73	88.15	82.52	85.47	**90.40**
12	94.81	92.28	95.08	**95.52**	94.20	94.78	90.33	95.07	94.09	94.66	92.40	**95.51**	85.46
13	93.48	92.58	90.42	94.00	90.72	**94.70**	86.83	**95.19**	92.03	91.96	94.89	91.02	88.57
14	90.97	**91.68**	85.56	88.18	93.86	89.61	87.18	53.79	79.27	89.20	65.98	90.83	**92.83**
15	58.13	47.47	54.31	54.72	53.72	57.05	**60.34**	37.16	50.18	50.87	55.02	53.49	**57.48**
16	95.79	93.52	85.09	91.17	**98.90**	95.22	95.19	89.21	82.57	93.80	**99.14**	94.87	95.09
OA	85.81	81.79	83.94	84.38	84.71	85.06	**86.64**	78.29	82.07	82.89	83.86	83.83	**85.22**
	(1.52)	(1.69)	(1.12)	(1.76)	(0.80)	(1.79)	**(1.31)**	(2.23)	(1.98)	(1.54)	(1.19)	(1.50)	**(1.32)**
AA	90.95	89.69	89.45	90.21	90.91	90.24	**91.17**	85.71	88.22	89.71	88.54	89.89	**90.50**
	(0.97)	(0.83)	(1.37)	(0.87)	(1.17)	(0.70)	**(1.33)**	(1.34)	(1.54)	(1.01)	(1.78)	(1.42)	**(1.28)**
Kappa	84.24	79.88	82.18	82.67	83.01	83.44	**85.15**	75.91	80.14	81.04	82.05	82.08	**83.53**
	(1.66)	(1.82)	(1.21)	(1.91)	(0.85)	(1.97)	**(1.43)**	(2.42)	(2.14)	(1.66)	(1.27)	(1.62)	**(1.45)**

overall accuracy can be improved about 7% by employing the proposed noisy label detection method with respect to the SVM trained with original training set on the Salinas Valley dataset. As shown in the close-up comparisons presented in Fig. 12, the classification maps obtained by the SVM trained with training set improved by the proposed HCEM method are more similar as the reference data. This demonstrates that the noisy labels in the original training set can be effectively removed by the proposed HCEM method with respect to the SALOF, DP, RLPA, and KECA methods.

As shown in Table V, the classification accuracy of the SVM assisted by the HCEM can be increased by 4.85%–6.93% under

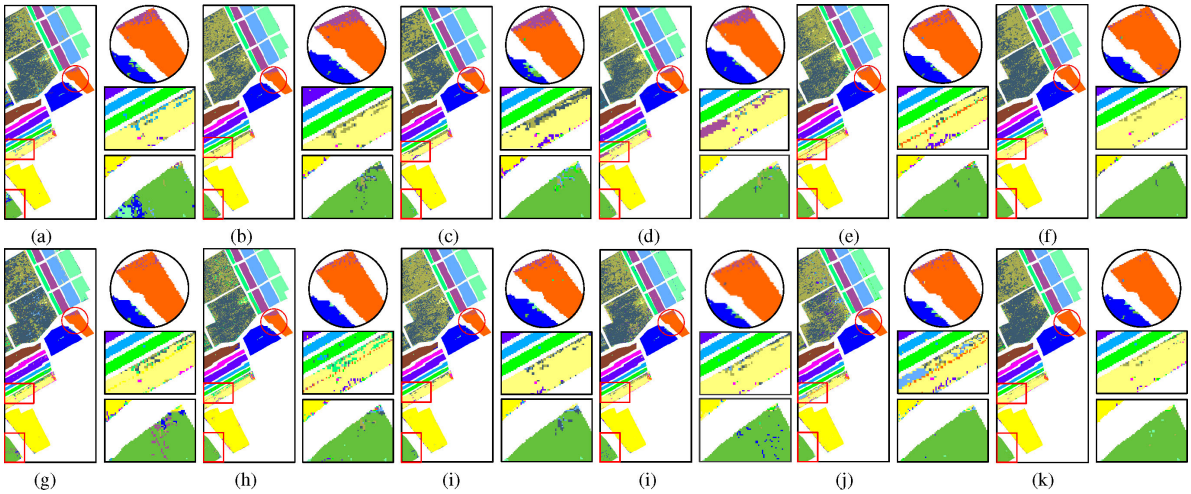


Fig. 12. Classification results (%) of comparing different method on the Salinas Valley dataset. Classification maps obtained by the SVM (the 1–2 columns), SALOF (the 3–4 columns), DP (the 5–6 columns), RLPA (the 7–8 columns), KECA (the 9–10 columns), and HCEM (the 11–12 columns) trained with 20 true samples and different numbers of mislabeled samples. The experiments of the first row all contain 5 noisy labels, and the experiments of the second row all contain 15 noisy labels.

TABLE V
CLASSIFICATION PERFORMANCE OF THE SVM, SALOF, DP, RLPA, KECA, AND HCEM METHODS FOR THE KSC DATASET WITH 25 TRUE SAMPLES AND DIFFERENT NUMBER OF NOISY LABELS PER CLASS AS TRAINING SET. NUMBER IN PARENTHESES REPRESENTS THE STANDARD VARIANCE OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Class	**SVM** 25 (true)	**The number of true samples and mislabeled samples**											
		25(true)+5(noisy)						25(true)+15(noisy)					
		SVM	SALOF	DP	RLPA	KECA	HCEM	SVM	SALOF	DP	RLPA	KECA	HCEM
1	94.92	93.53	**95.90**	95.30	94.30	93.29	92.87	**96.11**	95.43	95.73	94.41	93.38	87.79
2	89.05	79.86	83.34	83.77	76.85	85.60	**89.90**	77.02	74.58	74.53	78.37	79.93	**86.39**
3	87.57	**91.84**	85.19	84.55	82.60	82.76	83.67	**86.19**	84.14	84.34	76.35	81.06	84.16
4	67.59	58.72	64.28	64.64	**68.69**	62.60	61.99	56.02	55.63	57.00	59.83	**60.02**	54.48
5	65.38	51.88	62.20	61.92	**64.90**	62.45	63.28	57.14	52.38	57.82	59.31	56.72	**75.00**
6	61.91	51.72	54.59	**57.08**	56.93	55.13	54.59	36.10	50.07	48.74	50.34	**57.81**	51.60
7	71.02	56.83	68.49	**70.00**	62.21	66.96	65.14	53.19	63.78	**65.48**	59.78	60.36	64.65
8	84.20	**82.58**	79.89	78.09	82.43	80.34	78.24	52.58	66.88	69.17	71.35	80.82	**89.49**
9	90.42	89.27	86.95	84.96	**92.94**	89.07	92.67	**88.55**	85.28	82.63	87.46	86.75	87.52
10	93.02	80.85	83.85	92.55	92.54	94.48	**100.0**	83.21	83.16	87.27	92.03	**94.13**	91.72
11	90.08	86.72	86.37	87.33	80.29	92.60	**94.29**	**97.78**	88.79	90.71	89.44	90.83	90.59
12	94.77	83.62	86.77	90.24	**95.49**	93.48	84.92	82.76	86.28	87.10	88.59	**92.81**	84.88
13	100.0	99.18	98.08	96.62	98.26	99.60	**100.0**	**100.0**	98.73	98.65	99.15	98.07	96.33
OA	88.75	83.57	84.88	85.47	85.97	86.67	**87.52**	79.90	81.51	82.60	83.28	85.19	**85.23**
	(0.96)	(1.74)	(1.94)	(1.64)	(0.85)	(1.52)	(**1.72**)	(1.80)	(1.76)	(1.55)	(0.81)	(1.88)	(**2.44**)
AA	83.84	77.43	79.68	80.54	80.65	81.42	**81.66**	74.36	75.78	76.86	77.46	79.44	**80.35**
	(1.23)	(2.03)	(1.82)	(1.83)	(1.18)	(1.60)	(**1.78**)	(1.83)	(1.67)	(1.58)	(1.20)	(2.06)	(**1.64**)
Kappa	87.45	81.71	83.16	83.81	84.38	85.14	**86.07**	77.66	79.42	80.62	81.39	83.49	**83.45**
	(1.06)	(1.93)	(2.16)	(1.82)	(0.95)	(1.68)	(**1.90**)	(2.00)	(1.95)	(1.71)	0.90	(2.08)	(**2.68**)

different of training sets with noisy labels. Furthermore, to prove the effectiveness of the proposed HCEM method in noisy label detection, the experimental classification results obtained for the Washington DC dataset are shown in Table VI. Here, the classification performance obtained by the SVM trained with the improved training set provided by the proposed HCEM method can still obtain the highest accuracies for most of the classes. In particular, when the tree and shadow classes, respectively, contain 15 noisy labels, the classification accuracies achieve the highest OAs (97.69% and 99.32%, respectively). This demonstrates that the proposed HCEM method can effectively remove noisy labels and improve the classification performance of the SVM.

In the abovementioned experiments, all experimental programs are operated on a laptop computer with an Intel Core i5-6300HQ, CPU 2.30 GHz, and 8 GB of RAM, and the software platform is MATLAB R2017a (MathWorks, Natick, Massachusetts, America). The run times of the HCEM and other competitive methods are reported on the different real HSI datasets with 25 true training samples and 5 and 15 noisy labels. As can be observed from Table VII, the HCEM has time-consumption advantages with respect to the SALOF, DP, RLPA, and KECA methods in most instances. Specifically, as the number of noisy labels in the original training set increases, the proposed method still maintains superiority in execution time. It proves that the HCEM can effectively detect and remove noise labels in the noise training set, thereby reducing the training time and classification time of the SVM. Furthermore, the computing times of the proposed method in detection process are less than 1 s usually. The reason is that the major computing burden of

TABLE VI
CLASSIFICATION PERFORMANCE OF THE SVM, SALOF, DP, RLPA, KECA, AND HCEM METHODS FOR THE WASHINGTON DC DATASET WITH 25 TRUE SAMPLES AND DIFFERENT NUMBER OF NOISY LABELS PER CLASS AS TRAINING SET. NUMBER IN PARENTHESES REPRESENTS THE STANDARD VARIANCE OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Class	**SVM** 25 (true)	**The number of true samples and mislabeled samples**											
		25(true)+5(noisy)						25(true)+15(noisy)					
		SVM	SALOF	DP	RLPA	KECA	HCEM	SVM	SALOF	DP	RLPA	KECA	HCEM
1	90.28	86.08	84.12	88.43	88.91	89.28	**89.50**	74.98	84.70	87.45	85.22	**88.89**	86.15
2	95.97	95.78	95.63	93.15	91.16	95.99	**97.62**	80.04	89.74	89.30	**96.41**	94.25	92.43
3	71.22	71.67	70.40	68.80	66.52	**71.90**	71.40	68.55	67.30	66.07	70.19	69.69	**75.58**
4	81.67	77.41	85.52	83.48	80.61	81.95	**91.73**	79.43	76.85	76.95	**84.13**	79.17	71.13
5	95.38	93.52	93.48	96.65	**99.14**	96.77	91.83	85.62	87.80	94.11	85.99	92.98	**97.69**
6	98.94	95.67	98.38	**98.99**	98.26	98.19	95.58	96.73	93.35	97.26	87.23	96.10	**99.32**
OA	87.93	85.61	86.22	86.80	86.22	87.36	**88.87**	79.34	82.11	83.64	84.71	85.81	**86.08**
	(1.76)	(1.33)	(2.12)	(1.64)	(2.57)	(2.67)	(**2.81**)	(2.03)	(2.15)	(1.88)	(6.35)	(2.84)	(**3.35**)
AA	88.91	86.69	87.92	88.25	87.43	89.01	**89.61**	80.89	83.29	85.19	84.86	86.85	**87.05**
	(1.64)	(1.37)	(2.24)	(1.48)	(2.76)	(2.68)	(**2.80**)	(1.90)	(2.39)	(1.75)	(2.81)	(2.47)	(**2.85**)
Kappa	85.14	82.27	82.93	83.74	83.03	84.41	**86.24**	74.44	78.10	79.99	81.18	82.60	**82.78**
	(2.10)	(1.62)	(2.51)	(1.99)	(3.01)	(3.16)	(**3.24**)	(2.42)	(2.58)	(2.16)	(7.31)	(3.31)	(**3.88**)

TABLE VII
COMPARISON OF TIME-CONSUMPTION (SECONDS) OF DIFFERENT METHODS FOR UNIVERSITY OF HOUSTON, SALINAS, KSC, WASHINGTON DC DATASET CONTAIN 25 TRUE SAMPLES AND 5 OR 15 MISLABELED SAMPLES. WHERE THE DETECTION TIME IS MARKED BY "D-TIME" AND THE CLASSIFICATION TIME IS MARKED "C-TIME"

Training condition	*Methods*	**Detection and classification time of variable methods on the different datasets**							
		University of Houston		Salinas		KSC		Washington DC	
		D-Time	C-Time	D-Time	C-Time	D-Time	C-Time	D-Time	C-Time
25(T)+5(N)	SVM	–	93.14	–	91.49	–	82.65	–	12.65
	SALOF	0.20	63.44	0.16	61.87	0.13	59.76	0.10	7.14
	DP	0.34	65.45	0.54	36.12	0.63	58.92	0.16	4.06
	RLPA	0.36	121.32	0.41	48.36	0.29	69.54	0.18	11.06
	KECA	0.38	89.75	0.36	35.18	0.06	87.19	0.14	7.07
	HCEM	0.26	82.70	0.32	39.76	0.38	52.40	0.21	6.59
25(T)+15(N)	SVM	–	213.15	–	83.39	–	117.57	–	16.63
	SALOF	0.35	158.20	0.18	63.45	0.29	95.47	0.13	15.23
	DP	0.34	117.53	0.84	64.57	0.47	56.22	0.27	11.54
	RLPA	0.33	179.85	0.79	78.56	0.28	88.95	0.33	14.56
	KECA	0.31	117.45	0.45	75.33	0.06	87.19	0.18	12.25
	HCEM	0.29	108.11	0.36	61.33	0.39	77.37	0.21	9.22

the HCEM method is to calculate the SAM among samples. The complexity of obtaining SAM is $O(1)$, and the complexity of CEM detection is $O(\sum_{c=1}^{L} n^2)$, where n is the number of training samples in the cth class and L is the number of classes. Then, the complexity of the HCEM detection is $O(z \cdot \sum_{c=1}^{L} n^2)$, z is the number of layer.

F. Performance Verification With Other Classifiers and Spectral-Spatial Classification Methods

In this section, some other classic spectral classifiers, such as the basic thresholding classifier (BTC)[46], the kernel BTC (KBTC) [47], the sparse representation classifier (SRC) [48], and the extreme learning machine (ELM)[49], are adopted to further practicability of the proposed HCEM method in supervised learning of HSI classification. The experimental results are presented in Table VIII, which are achieved from the Salinas Valley dataset with 25 true training samples and different numbers of noisy labels per class. For objective and fair comparison, the experiment was repeated 10 times to achieve the average value and the standard deviation of the classification results. It can be observed from Table VIII that the spectral classifiers trained with the improved training set invariably acquire better classification results than those trained using the original noisy training set. Specifically, the spectral classifiers trained with the training set improved by the proposed HCEM method obtains the good performance in terms of the highest OAs, AAs, and Kappas with respect to spectral classifiers trained using other improved training sets. This experiment further demonstrates the robustness of the proposed HCEM method in the process of the noisy label detection.

In addition to analyzing the detection performance of the proposed HCEM in spectral classifiers, another experiment is conducted to verify the effectiveness of the proposed HCEM method for spectral-spatial classification methods, such as extended morphological profiles (EMP)[50], logistic regression and multilevel logistic (LMLL)[51], the joint sparse representation classifier (JSRC)[52], and the edge-preserving filtering (EPF) [53], on the Salinas Valley dataset with 25 true training samples and different numbers of noisy labels for each class. Similarly, in order to present a fair competition result, the experiment is repeated ten times to obtain the mean and standard deviation. Table IX gives the classification accuracies of various spectralspatial classification methods trained using the original noisy training set and the improved training sets (achieved by SALOF, DP, RLPA, KECA, and HCEM, respectively). As shown in Table IX, it can be seen that the spectral-spatial classification methods

TABLE VIII
CLASSIFICATION ACCURACY OF SPECTRAL CLASSIFIERS ON THE SALINAS VALLEY DATASET. THE SPECTRAL CLASSIFIERS OBTAINED THE CLASSIFICATION ACCURACY OF 25 TRUE SAMPLES AND THE NLA (NO PROCESSING), SALOF, DP, RLPA, KECA, AND HCEM METHODS TO COMPARE THE CLASSIFICATION ACCURACY UNDER 25 TRUE SAMPLES AND DIFFERENT NUMBERS OF NOISY LABELS PER CLASS. NUMBER IN PARENTHESES REPRESENTS THE STANDARD VARIANCE OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Index	*Classifier*	25(true)+5(noisy)				25(true)+15(noisy)			
		BTC	**KBTC**	**SRC**	**ELM**	**BTC**	**KBTC**	**SRC**	**ELM**
OA(%)	***NLA***	76.23(1.21)	75.82(1.15)	67.82(1.23)	69.54(0.98)	49.46(1.16)	57.33(1.75)	50.55(1.25)	54.38(1.52)
	SALOF	77.76(1.53)	80.71(2.24)	74.66(2.37)	76.17(2.17)	51.62(2.19)	62.01(1.54)	54.09(2.14)	58.13(1.54)
	DP	78.91(1.94)	82.78(1.88)	77.48(1.15)	79.72(1.36)	49.33(1.08)	61.39(1.63)	52.85(2.19)	57.84(1.86)
	RLPA	79.45(0.73)	83.75(1.46)	78.78(1.67)	80.06(1.25)	56.78(1.15)	64.52(0.98)	56.63(1.36)	59.03(1.15)
	KECA	80.15(1.65)	85.11(1.95)	79.48(1.58)	80.74(1.72)	58.99(2.06)	67.22(1.01)	58.89(1.42)	60.76(1.55)
	HCEM	**88.34(1.25)**	**86.25(1.45)**	**80.89(1.58)**	**83.54(1.65)**	**64.69(1.72)**	**68.62(1.27)**	**65.54(1.64)**	**70.09(1.32)**
AA(%)	***NLA***	76.35(0.88)	76.46(1.12)	79.11(1.21)	68.98(1.52)	48.13(1.75)	55.96(1.65)	64.06(1.71)	52.63(1.15)
	SALOF	80.78(1.45)	85.16(2.59)	84.49(1.49)	78.00(2.39)	50.53(2.21)	60.99(1.49)	68.45(1.84)	56.43(1.52)
	DP	82.66(2.63)	87.82(1.85)	87.42(0.63)	83.48(1.55)	47.88(1.27)	61.01(1.61)	66.39(2.48)	56.11(1.96)
	RLPA	82.78(1.69)	86.54(1.78)	81.36(1.25)	83.26(0.69)	59.77(1.86)	67.35(1.12)	59.03(1.26)	62.11(1.07)
	KECA	83.37(1.66)	89.67(2.05)	88.02(0.77)	83.53(1.55)	47.88(2.41)	66.93(1.71)	72.38(1.39)	59.32(1.38)
	HCEM	**93.74(1.21)**	**93.38(1.24)**	**88.77(0.95)**	**85.65(1.44)**	**77.41(1.82)**	**80.92(1.31)**	**75.84(1.45)**	**70.88(1.24)**
Kappa	***NLA***	72.77(1.31)	73.39(1.25)	64.84(1.32)	66.67(1.05)	45.66(1.56)	53.74(1.87)	46.62(1.36)	50.64(1.59)
	SALOF	75.44(1.71)	78.65(2.42)	72.00(2.54)	73.79(2.34)	47.61(2.31)	58.73(1.85)	51.21(2.28)	54.71(1.65)
	DP	76.69(1.54)	80.93(2.08)	75.16(1.23)	77.63(1.45)	45.07(1.65)	58.13(1.73)	49.02(2.38)	54.43(1.91)
	RLPA	77.66(1.37)	81.96(1.24)	76.54(1.52)	78.68(1.58)	54.62(1.76)	62.77(1.45)	54.69(1.77)	57.15(1.44)
	KECA	78.06(1.79)	83.53(2.14)	77.32(1.69)	78.45(1.84)	55.31(2.17)	64.18(1.03)	55.44(1.51)	57.41(1.64)
	HCEM	**87.04(1.33)**	**84.72(1.37)**	**78.87(1.47)**	**81.81(1.36)**	**61.77(1.63)**	**65.68(1.34)**	**62.22(1.26)**	**67.14(1.51)**

TABLE IX
CLASSIFICATION ACCURACY OF SPECTRAL-SPATIAL CLASSIFICATION METHODS ON THE SALINAS VALLEY DATASET. THE SPECTRAL-SPATIAL CLASSIFICATION METHODS OBTAINED THE CLASSIFICATION ACCURACY OF 25 TRUE SAMPLES AND THE NLA (NO PROCESSING), SALOF, DP, RLPA, KECA, AND HCEM METHODS TO COMPARE THE CLASSIFICATION ACCURACY UNDER 25 TRUE SAMPLES AND DIFFERENT NUMBERS OF NOISY LABELS PER CLASS. NUMBER IN PARENTHESES REPRESENTS THE STANDARD VARIANCE OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Index	*Methods*	25(true)+5(noisy)				25(true)+15(noisy)			
		LMLL	**EPF**	**EMP**	**JSRC**	**LMLL**	**EPF**	**EMP**	**JSRC**
OA(%)	***NLA***	88.45(1.86)	84.83(3.00)	85.92(2.70)	72.64(3.07)	85.89(2.48)	82.00(3.12)	83.34(2.45)	49.96(2.11)
	SALOF	92.11(1.82)	88.35(4.51)	89.46(1.53)	79.09(2.92)	90.22(1.66)	86.15(4.40)	86.53(1.90)	53.36(4.12)
	DP	92.03(2.07)	89.24(2.99)	91.30(2.79)	81.31(2.43)	91.82(1.95)	87.05(3.10)	87.41(2.00)	55.19(2.56)
	RLPA	94.37(1.64)	86.02(3.45)	92.25(2.31)	81.75(1.21)	91.68(2.65)	86.11(1.72)	87.47(1.59)	61.57(2.35)
	KECA	94.84(1.78)	91.48(3.71)	92.28(1.02)	82.61(2.92)	92.66(1.86)	89.20(3.82)	88.75(1.36)	62.43(1.71)
	HCEM	**95.16(1.52)**	**92.45(0.71)**	**92.37(0.49)**	**82.82(1.82)**	**96.33(1.56)**	**91.77(2.68)**	**90.76(1.62)**	**72.30(1.68)**
AA(%)	***NLA***	94.94(0.76)	94.38(0.66)	93.10(0.99)	73.85(2.55)	93.48(1.11)	93.12(0.83)	92.79(0.81)	55.87(3.29)
	SALOF	95.67(0.64)	91.31(1.73)	94.46(0.57)	78.04(2.11)	94.20(0.90)	94.33(1.58)	92.79(0.88)	59.51(2.99)
	DP	95.72(1.01)	95.30(0.67)	95.54(0.88)	79.53(1.73)	95.45(1.22)	94.04(1.11)	94.28(0.61)	60.77(2.99)
	RLPA	97.25(1.25)	94.38(2.56)	95.57(2.54)	83.24(1.25)	94.77(2.39)	94.36(1.65)	93.64(3.25)	72.48(2.68)
	KECA	96.95(0.87)	95.94(1.12)	95.74(0.36)	80.44(0.95)	96.07(0.72)	95.09(1.03)	94.43(0.51)	74.82(1.57)
	HCEM	**97.40(0.67)**	**96.12(0.83)**	**96.03(0.19)**	**85.05(1.24)**	**98.15(0.68)**	**95.32(1.21)**	**94.48(0.68)**	**77.93(1.36)**
Kappa	***NLA***	87.23(2.05)	84.29(3.29)	84.45(2.97)	69.98(3.28)	84.33(2.75)	80.13(3.41)	81.54(2.69)	46.01(2.28)
	SALOF	91.24(2.02)	87.10(4.93)	88.30(1.70)	76.97(3.16)	89.13(1.82)	84.68(4.90)	85.09(2.09)	49.70(4.25)
	DP	91.12(2.31)	88.07(3.28)	90.35(3.07)	79.33(2.67)	90.91(2.17)	85.50(3.41)	86.08(2.19)	51.14(2.69)
	RLPA	93.73(2.15)	84.55(1.56)	91.37(1.29)	78.49(1.98)	90.73(2.15)	84.65(0.97)	86.04(1.15)	57.83(3.24)
	KECA	94.27(1.97)	90.54(4.09)	91.40(1.11)	80.73(3.11)	91.84(2.06)	88.03(4.18)	87.51(1.49)	59.17(1.77)
	HCEM	**94.59(2.16)**	**91.59(0.69)**	**91.51(0.55)**	**80.93(2.05)**	**95.90(1.86)**	**90.83(3.24)**	**89.72(1.36)**	**69.60(1.68)**

trained with training set obtained by the HCEM can achieve the highest classification accuracies on the samples condition of both containing five noisy labels and existing fifteen noisy labels. This experiment results further suggest that the proposed HCEM method indeed works effectively in process of noisy label detection and that it is also high efficiency in promoting the classification accuracy of spectral-spatial classification methods.

VI. CONCLUSION

In this article, a hierarchical structure based constrained energy minimum is first proposed to cleanse noisy labels of training process in HSI supervised classification task. The motivation behind this work is exploiting sample energy distribution in cascaded CEM to detect noisy labels of the original training set. Experimental results on four real hyperspectral datasets are evaluated by a series of spectral classifiers and spectral-spatial classification methods, it demonstrated that the proposed HCEM method can effectively remove noisy labels of original training set and improved the performance of supervised classification task. However, one limitation of the proposed method is that the HCEM method combined with sample context information has no experimental and theoretical verification in this article due to workload and time complexity. Therefore, developing a noisy label detection method combining spectral-spatial information of HSI samples to further improve the detection performance of noisy labels will be the interesting focus in our future research directions. Furthermore, extending the proposed HCEM method to the processing of multilabel samples of hyperspectral remote sensing images[54] is a research point that we focus on.

ACKNOWLEDGMENT

The authors would like to thank the Editor-in-chief, the anonymous Associate Editor, and the reviewers for their insightful comments and suggestions, which significantly improved the quality and presentation of this article.

REFERENCES

[1] L. Zhang, W. Wei, Y. Zhang, C. Shen, A. Van Den Hengel, and Q. Shi, "Cluster sparsity field: An internal hyperspectral imagery prior for reconstruction," *Int. J. Comput. Vision*, vol. 126, no. 8, pp. 797–821, Mar. 2018.

[2] S. Li, Q. Hao, G. Gao, and X. Kang, "The effect of ground truth on performance evaluation of hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7195–7206, Dec. 2018.

[3] W. Wei, L. Zhang, Y. Jiao, C. Tian, C. Wang, and Y. Zhang, "Intracluster structured low-rank matrix analysis method for hyperspectral denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 866–880, Feb. 2019.

[4] Y. Hu, Q. Zhang, Y. Zhang, and H. Yan, "A deep convolution neural network method for land cover mapping: A case study of qinhuangdao, china," *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 2053.

[5] P. Wang, L. Wang, and J. Chanussot, "Soft-then-hard subpixel land cover mapping based on spatial-spectral interpolation," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1851–1854, Dec. 2016.

[6] Z. Wu, Y. Li, A. Plaza, J. Li, F. Xiao, and Z. Wei, "Parallel and distributed dimensionality reduction of hyperspectral data on cloud computing architectures," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2270–2278, Jun. 2016.

[7] K. Berger, C. Atzberger, M. Danner, M. Wocher, W. Mauser, and T. Hank, "Model-based optimization of spectral sampling for the retrieval of crop variables with the prosail model," *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 2063.

[8] I. Aneece and P. Thenkabail, "Accuracies achieved in classifying five leading world crop types and their growth stages using optimal earth observing-1 hyperion hyperspectral narrowbands on Google Earth engine," *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 2027.

[9] A. Marcinkowska-Ochtyra, A. Jarocińska, K. Bzdega, and B. Tokarska-Guzik, "Classification of expansive grassland species in different growth stages based on hyperspectral and lidar data," *Remote Sensing*, vol. 10, no. 12, 2018, Art. no. 2019.

[10] A. Marcinkowska-Ochtyra, K. Gryguc, A. Ochtyra, D. Kopeć, A. Jarocińska, and Ł. Sławik, "Multitemporal hyperspectral data fusion with topographic indicesimproving classification of Natura 2000 grassland habitats," *Remote Sens.*, vol. 11, no. 19, 2019, Art. no. 2264.

[11] C. Pelletier, S. Valero, J. Inglada, N. Champion, C. Marais Sicre, and G. Dedieu, "Effect of training class label noise on classification performances for land cover mapping with satellite image time series," *Remote Sens.*, vol. 9, no. 2, 2017, Art. no. 173.

[12] Y. Zhong and L. Zhang, "An adaptive artificial immune network for supervised classification of multi-/hyperspectral remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 894–909, Mar. 2011.

[13] B. Tu, X. Yang, N. Li, C. Zhou, and D. He, "Hyperspectral anomaly detection via density peak clustering," *Pattern Recognit. Lett.*, vol. 129, pp. 144–149, 2020.

[14] 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.

[15] 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.

[16] B. Guo, S. R. Gunn, R. I. Damper, and J. D. Nelson, "Customizing kernel functions for SVM-based hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 17, no. 4, pp. 622–629, Apr. 2008.

[17] F. Ratle, G. Camps-Valls, and J. Weston, "Semisupervised neural networks for efficient hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2271–2282, May 2010.

[18] 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.

[19] H. R. Bittencourt, D. A. de Oliveira Moraes, and V. Haertel, "A binary decision tree classifier implementing logistic regression as a feature selection and classification method and its comparison with maximum likelihood," *IEEE Int. Geosci. Remote Sens. Symp.*, pp. 1755–1758, Jul. 2007.

[20] J. Maschler, C. Atzberger, and M. Immitzer, "Individual tree crown segmentation and classification of 13 tree species using airborne hyperspectral data," *Remote Sens.*, vol. 10, no. 8, 2018, Art. no. 1218.

[21] D. G. Stavrakoudis, G. N. Galidaki, I. Z. Gitas, and J. B. Theocharis, "A genetic fuzzy-rule-based classifier for land cover classification from hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 1, pp. 130–148, Jan. 2011.

[22] A. Zehtabian and H. Ghassemian, "Automatic object-based hyperspectral image classification using complex diffusions and a new distance metric," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 4106–4114, Jul. 2016.

[23] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[24] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 38, no. 3, pp. 447–461, Mar. 2016.

[25] Y. Yao, L. Wang, L. Zhang, Y. Yang, P. Li, R. Zimmermann, and L. Shao, "Learning latent stable patterns for image understanding with weak and noisy labels," *IEEE Trans. Cybern.*, vol. 49, no. 12, pp. 4243–4252, Dec. 2019.

[26] Z. Lu, Z. Fu, T. Xiang, P. Han, L. Wang, and X. Gao, "Learning from weak and noisy labels for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 39, no. 3, pp. 486–500, Mar. 2017.

[27] X. Kang, P. Duan, X. Xiang, S. Li, and J. A. Benediktsson, "Detection and correction of mislabeled training samples for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5673–5686, Oct. 2018.

[28] 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. 2018.

[29] B. Tu, C. Zhou, W. Kuang, L. Guo, and X. Ou, "Hyperspectral imagery noisy label detection by spectral angle local outlier factor," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 9, pp. 1417–1421, Sep. 2018.

[30] B. Tu, X. Zhang, X. Kang, G. Zhang, and S. Li, "Density peak-based noisy label detection for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1573–1584, Mar. 2019.

[31] B. Tu, X. Zhang, X. Kang, J. Wang, and J. A. Benediktsson, "Spatial density peak clustering for hyperspectral image classification with noisy labels," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5085–5097, Jul. 2019.

[32] B. Tu, C. Zhou, D. He, S. Huang, and A. Plaza, "Hyperspectral classification with noisy label detection via superpixel-to-pixel weighting distance," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2019.2961141.

[33] L. Jie, Q. Yuan, H. Shen, and L. Zhang, "Noise removal from hyperspectral image with joint spectralspatial distributed sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5425–5439, Sep. 2016.

[34] B. Tu, C. Zhou, J. Peng, W. He, X. Ou, and Z. Xu, "Kernel entropy component analysis-based robust hyperspectral image supervised classification," *Remote Sens.*, vol. 11, no. 23, Nov. 2019, Art. no. 2823.

[35] Chein-I Chang and Su Wang, "Constrained band selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1575–1585, Jun. 2006.

[36] L. Gao, B. Yang, Q. Du, and B. Zhang, "Adjusted spectral matched filter for target detection in hyperspectral imagery," *Remote Sens.*, vol. 7, no. 6, pp. 6611–6634, 2015.

[37] Z. Zou, Z. Shi, J. Wu, and H. Wang, "Quadratic constrained energy minimization for hyperspectral target detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2015, pp. 4979–4982.

[38] Y. Zhang, B. Xie, J. Sun, and Y. Peng, "A hybrid sparsity and constrained energy minimization detector for hyperspectral images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2017, pp. 1137–1140.

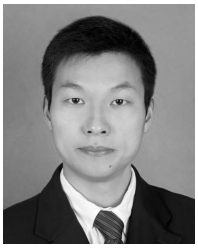
[39] C.-I. Chang, "Spectral information divergence for hyperspectral image analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp*, Jun. 1999, pp. 509–511.

[40] B. Tu, X. Zhang, X. Kang, G. Zhang, J. Wang, and J. Wu, "Hyperspectral image classification via fusing correlation coefficient and joint sparse representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 340–344, Mar. 2018.

[41] Y. Yuan, J. Lin, and Q. Wang, "Dual-clustering-based hyperspectral band selection by contextual analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1431–1445, Mar. 2015.

[42] 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.

[43] H. Ren and C.-I. Chang, "Target-constrained interference-minimized approach to subpixel target detection for hyperspectral imagery," *Opt. Eng.*, vol. 39, no. 12, pp. 3138–3145, Dec. 2000.
[44] X. Chen, C. Gu, Y. Zhang, and R. Mittra, "Analysis of partial geometry modification problems using the partitioned-inverse formula and Sherman–Morrison–Woodbury formula-based method," *IEEE Trans. Antennas Propag.*, vol. 66, no. 10, pp. 5425–5431, Oct. 2018.
[45] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27-1–27-27, May 2011.
[46] M. A. Toksz and I. Ulusoy, "Hyperspectral image classification via basic thresholding classifier," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 4039–4051, Jul. 2016.
[47] M. A. Toksz and I. Ulusoy, "Hyperspectral image classification via kernel basic thresholding classifier," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 715–728, Feb. 2017.
[48] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[49] X. D. R. Z. G.B. Huang, H. Zhou, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst. Man Cybernet.*, vol. 42, no. 2, Apr. 2012, Art. no. 513529.
[50] J. A. Benediktsson, M. Pesaresi, and K. Amason, "Classification and feature extraction for remote sensing images from urban areas based on morphological transformations," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 9, pp. 1940–1949, Sep. 2003.
[51] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new bayesian approach with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3947–3960, Oct. 2011.
[52] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.
[53] 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.
[54] Q. Hao, S. Li, and X. Kang, "Multilabel sample augmentation-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2019.2962014.

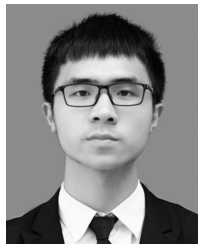


Bing Tu (Member, IEEE) received the M.S. degree in control science and engineering from the Guilin University of Technology, Guilin, China, in 2009 and the Ph.D. degree in mechatronic engineering from the Beijing University of Technology, Beijing, China, in 2013.

From 2015 to 2016, he was a Visiting Researcher with the Department of Computer Science and Engineering, University of Nevada, Reno, NV, USA, which is supported by the China Scholarship Council. Since 2018, he has been an Associate Professor with the School of Information Science and Engineering, Hunan Institute of Science and Technology, Yueyang, China. His research interests include sparse representation, pattern recognition, and analysis in remote sensing.

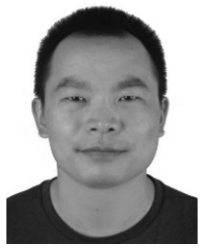
Chengle Zhou (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering and automation in 2019 from the Hunan Institute of Science and Technology, Yueyang, China, where he is currently working toward the M.S. degree in information and communication engineering with Hunan Institute of Science and Technology, Yueyang, China.

His research interests include image processing, pattern recognition, hyperspectral image classification, anomaly detection, and noisy label detection.



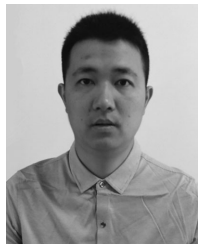
Xiaolong Liao is currently working toward the B.S. degree in automation with Hunan Institute of Science and Technology, Yueyang, China.

His research interest includes hyperspectral image analysis.



Zhi Xu received the B.S. degree in material physics from Lanzhou University, Lanzhou, China, in 2006, and the M.S. and Ph.D. degrees in communication and information system from Sichuan University, Chengdu, China, in 2009 and 2013, respectively.

He is currently an Associate Professor with the School of Computer Science and Information Safety, Guilin University of Electronic Technology. His research interests include computer vision, machine learning, and pattern recognition.



Yishu Peng received the B.S., M.S., and Ph.D. degrees from Northeastern University, Shenyang, China, in 2009, 2011, and 2017, respectively, all in mechanical design and theory.

From 2017 to 2019, he joined the School of Mechanical and Engineering, Hunan Institute of Science and Technology, Yueyang, China, and then he turned to the School of Information Science and Technology until now. His research interests include the image processing, object detection, and target tracing.

Xianfeng Ou (Member, IEEE) received the B.S. degree in electronic information science and technology and the M.S. degree in communication and information system from Xinjiang University, Urumchi, China, in 2006 and 2009, respectively, and the Ph.D. degree in communication and information system from Sichuan University, Chengdu, China, in 2015.

He was a visiting researcher with the Internet Media Group, Polytechnic di Torino, Turin, Italy, from January to April 2014, working on distributed video coding and transmission. His main research interests include machine vision and artificial intelligence, object detection, and image and video coding process technologies.