Hyperspectral Classification With Noisy Label Detection via Superpixel-to-Pixel Weighting Distance

Bing Tu, *Member, IEEE*, Chengle Zhou, *Student Member, IEEE*, Danbing He, *Student Member, IEEE*, Siyuan Huang, *Student Member, IEEE*, and Antonio Plaza, *Fellow, IEEE*

***Abstract*— Classification is an important technique for remotely sensed hyperspectral image (HSI) exploitation. Often, the presence of wrong (noisy) labels presents a drawback for accurate supervised classification. In this article, we introduce a new framework for noisy label detection that combines a superpixel-to-pixel weighting distance (SPWD) and density peak clustering. The proposed method is able to accurately detect and remove noisy labels in the training set before HSI classification. It considers two weak assumptions when exploiting the spectral–spatial information contained in the HSI: 1) all the pixels in a superpixel belong to the same class and 2) close pixels in spectral space have the same label. The proposed method consists of the following steps. First, a superpixel segmentation step is used to obtain self-adaptive spatial information for each training sample. Then, a metric is utilized to measure the spectral distance information between each superpixel and pixel. Meanwhile, in order to overcome the first weak assumption, we use K nearest neighbors to obtain the closest neighborhoods of pixels around each superpixel, and a Gaussian weight is employed to mitigate the second weak assumption by adapting the original distance information. Next, the noisy labels in the original training set are removed by a density threshold-based decision function. Finally, the support vector machine (SVM) classifier is employed to evaluate the effectiveness of the proposed SPWD detection method in terms of classification accuracy. Experiments performed on several real HSI data sets demonstrate that the method can effectively improve the performance of classifiers trained with noisy training sets in terms of classification accuracy.**

***Index Terms*— Density peak (DP) clustering, Gaussian weighting, hyperspectral images (HSI), noisy labels, superpixel segmentation, support vector machines (SVMs).**

Manuscript received November 5, 2019; revised December 15, 2019; accepted December 16, 2019. 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 Hunan Provincial Innovation Foundation for Postgraduate under Grant CX20190914, 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. *(Chengle Zhou and Bing Tu contributed equally to this work.) (Corresponding author: Bing Tu.)*

Bing Tu, Chengle Zhou, Danbing He, and Siyuan Huang are with the School of Information Science and Engineering, Hunan Institute of Science and Technology, Yueyang 414000, China (e-mail: tubing@hnist.edu.cn; chengle_zhou@foxmail.com; danbing_he@163.com; ss24cs@126.com).

Antonio Plaza is with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politecnica, University of Extremadura, E-10003 Caceres, Spain (e-mail: aplaza@unex.es).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2019.2961141

I. Introduction

THE recent development of space remote sensing technology provides a wealth of usable remote sensing data for researchers. Hyperspectral images (HSIs) comprise rich spectral and spatial information, which allows for their exploitation in many domains, e.g., environmental monitoring [1], [2], precision agriculture [3]–[5], and military applications [6], [7], among others.

In many applications of HSIs, supervised classification methods such as support vector machines (SVMs) [8], sparse representation (SR) [9]–[11], random forests [12], [13], kernel-based methods [14]–[16], and deep learning [17], [18] play an important role. However, when limited labeled data are available, many of these classification methods are unable to achieve satisfactory classification accuracies. The reason is that, although the rich spectral and spatial information contained in HSIs provides the possibility to detect and classify various objects with better accuracy, the lack of sufficient labeled data in real applications leads to difficulties in the classification task. Besides, the acquisition of labeled data to be used for training by supervised methods tends to suffer from "noisy label" disturbances, due to problems in the acquisition device and manual labeling errors. In general terms, the noisy label phenomenon is inevitable, since it is caused by the following three main reasons: 1) the captured scenes are so complex that ground tasks are extremely difficult, leading to the misclassification of some land covers; 2) large regular areas often contain many complex spatial structures, resulting in some wrongly labeled anomalous regions; and 3) the spatial resolution of the scene depends on technology of the hyperspectral sensor, which may cause the pixel of an HSI to be inevitably mixed (i.e., composed by different land covers). Therefore, designing a method for the detection of noisy labels prior to supervised classification is a highly desirable objective.

In recent years, the problem of "noisy" labels in supervised classification has become quite important in the computer vision domain (in general) and also in remote sensing image processing (specifically). For instance, Xiao *et al.* [19] introduced a probabilistic graphical framework to train convolutional neural networks (CNNs) with only a limited number of clean labels and millions of noisy labels. Lu *et al.* [20] formulated a novel L_1 optimization-based sparse learning model to directly (and explicitly) detect noisy labels for

semantic segmentation. Yao *et al.* [21] presented a generative model (called *latent stability analysis*) to extract stable patterns from images with noisy labels. Mnih and Hinton [22] proposed two robust loss functions (obtained by training a deep neural network) to improve the robustness of deep learning classifiers. Foody [23] showed that noisy labels can affect the accuracy of classification obtained by discriminant analysis and SVMs on airborne thematic mapper data sets. In order to adjust the labels for noisy (or missing) labeled data sets, Song and Wang [24] provided a new effective label-refinement algorithm, which is useful for generating better labels. Hou *et al.* [25] proposed a semisupervised probability graphic-based classification framework to address the quantity and quality of the labeled training pixels. Although some researchers have tried to address the problem of noisy labels in related fields, the aforementioned methods cannot be directly extended to HSI classification because of the high dimensionality and nonlinear structure of HSI data.

In order to address the aforementioned issues, Kang *et al.* [26] first introduced the reasons for the formation of noisy labels in the supervised classification of HSI data, and further proposed an edge-preserving filtering (EPF) and spectral detection-based method to correct mislabeled training samples. Jiang *et al.* [27] proposed a random label propagation algorithm (RLPA) to cleanse the noisy labels of the training set. The RLPA method introduces graph theory to classification problems under noisy labels for the first time and is very robust to the label noise especially when the noise level is large. In [28], a new method that fuses the spectral angle and a local outlier factor (SALOF) was proposed to detect noisy labels in HSI classification. Tu *et al.* [29] presented a density peak (DP) clustering-based noisy label detection method to remove noisy labels. The experimental results show that the DP-based detection method can effectively promote classification accuracy. Moreover, Leng *et al.* [30] designed a label noise cleansing method with a sparse graph (SALP) which relies on the assumption that noise follows a Gaussian distribution. It was shown that SALP provides an effective mechanism to identify noisy labels in the training set prior to HSI classification.

In addition, the spatial density peak (SDP) clustering algorithm has been adopted to remove noisy labels [31]. The key idea of the SDP is to add neighboring samples to available (center) labeled samples to measure the anomalous nature of the center samples. However, although the introduction of spatial information by means of the SDP can effectively promote the accuracy of noisy label detection and improve the performance of HSI classification, there are still several aspects that need to be optimized. First, the incorporation of spatial information using fixed neighborhoods may introduce improper segmentation results in contextual edge information. Second, the correlation among samples is often determined by the average nearest neighbor distance, which may lead to ignoring the (weak) assumption that close pixels in spectral space often have the same label. In order to address the aforementioned problems, superpixel segmentation techniques and weighted joint nearest neighbor-based methods are investigated in this article. In previous developments, Fang *et al.* [32] used superpixel algorithms to adaptively extract spatial information in HSI classification. A superpixel correlation coefficient-based representation was also presented to improve the performance of HSI classification in [33]. Tu *et al.* [34] proposed a weighted joint nearest neighbor and sparse representation method to better characterize the distribution of samples when computing their distance.

In this article, a new method that combines a superpixel-to-pixel weighting distance (SPWD) and the DP clustering method is developed to address the problem of "*noisy labels*" in HSI classification. Our newly developed method is specifically designed to cleanse noisy labels in the training set and further improve the classification performance. Our method includes the following main steps. First, an entropy rate superpixel algorithm is introduced to obtain self-adaptive spatial contextual information around each training sample. Then, the spectral angle mapper (SAM) is employed as a distance measure between superpixels and pixels in the scene. Meanwhile, a KNN-based Gaussian weight is utilized to fully integrate the spectral–spatial information contained in the HSI. Next, noisy labels are removed from the original training set by a density threshold-based decision function. Finally, the SVM classifier is adopted to evaluate the effectiveness of the proposed SPWD method. Specifically, the main innovative contributions of the proposed approach can be summarized as follows.

1) The traditional (spatial) DP clustering algorithm is improved in this work by introducing (self-adaptive) superpixels which further consider the spatial information around each training sample. Moreover, we use a KNN-based Gaussian weighted distance (instead of the traditional average distance) as a more effective way to characterize the distance among different samples.
2) We introduce, for the first time in the literature, a method combining the superpixel-to-pixel weighting distance and DP clustering for noisy label detection. It is found that the self-adaptive spatial contextual information brought by superpixels is more effective than spatial neighborhoods with fixed shape when defining the local density of the training samples.

Our experimental evaluation, conducted using several HSI data sets, demonstrates that our newly developed method for noisy label detection can effectively improve the performance of classifiers trained with noisy training sets. In fact, the method can improve the classification performance of different supervised (spectral and spectral–spatial) HSI classifiers. Moreover, it is computationally efficient and prone to be exploited in real applications.

The rest of this article is organized as follows. In Section II, we briefly review the concepts of superpixel segmentation, Gaussian weighting, and DP clustering. Our newly developed method for detection noisy labels is described in detail in Section III. In Section IV, we present the experimental results (including a detailed comparison between several state-of-the-art detection methods). Finally, Section V concludes this article with some remarks and hints at plausible future research lines.

II. Related Work

In this section, we briefly review the superpixel segmentation method, the Gaussian weighting method, and the DP clustering method.

A. Superpixel Segmentation

Superpixel segmentation algorithms such as simple linear iterative clustering (SLIC) [35] and entropy rate superpixel (ERS) [36] have been widely used in HSI processing to exploit spatial contextual information around pixels. Let us assume that S_n refers to a predefined superpixel block. The graph-based ERS algorithm first maps the image to a graph $G = (V, E)$, where V is a set of vectors (representing pixels) and E is a set of edges (denoting pairwise similarities between adjacent pixels). Then, a subset of edges A is selected to segment the graph into S_n-related local regions. $H(\cdot)$ (the entropy rate term of the random walk) and $B(\cdot)$ (a balancing term that reduces small superpixels) are incorporated into the objective function to form balanced superpixels as follows:

$$\max_{A} \{H(A) + \lambda B(A)\} \quad \text{s.t.} \quad A \subseteq E \tag{1}$$

where $\lambda \geq 0$ is a weight that controls the contribution of the entropy rate term and the balancing term. The problem in (1) can be solved efficiently by a greedy algorithm, as mentioned in [37] and [38]. Finally, given a common label for each superpixel (obtained by measuring the distance between each superpixel and each training class), the base image U can be described as follows:

$$U = \bigcup_{s_i=0}^{S_n} M_{s_i} \quad \text{and} \quad M_{s_i} \cap M_{s_j} = \varnothing, \quad (s_i \neq s_j) \tag{2}$$

where M_{s_i} and M_{s_j} denote any two different superpixels in the base image U.

B. Gaussian Weighting

Gaussian weighting has been proven to be effective to characterize the spatial correlation among feature vectors [39]. The Gaussian weighting function has been used in HSI classification to increase the reliability of distance metrics [34]. Let us assume that we have a vector $\boldsymbol{Dist}^c$ containing the distance information from a sample of the cth class in the data with regard to other adjacent samples. In this case, the weighted coefficient vector $\boldsymbol{\omega}$ is defined as follows:

$$\boldsymbol{\omega}^c = e^{-\frac{\boldsymbol{Dist}^{c2}}{2\hat{\sigma}^2}} \tag{3}$$

where $\hat{\sigma}$ is the half-peak width that controls the attenuation parameters of the Gaussian weighted function. Gaussian weights are nonnegative and decrease with distance. In addition, if we consider the redundancy of distance information, the weighted K nearest neighbor (KNN) distance can be used as an effective way to balance the proximity of spatial samples as follows [34]:

$$\mathbf{Dist}^c_{\text{weight}} = \frac{\sum \mathbf{Dist}^{\mathbf{c}}(K)\boldsymbol{\omega}^c}{\sum \boldsymbol{\omega}^c} \tag{4}$$

where $\mathbf{Dist}^{\mathbf{c}}(K)$ represents the distance information for the K nearest neighbor samples of the cth class. In this way, it is obvious that the introduction of a weighted distance can effectively allocate a suitable weight to each sample, thus overcoming the imbalance problem in the original distance information.

C. DP Clustering

The DP clustering algorithm is based on the assumption that the cluster center is surrounded by low-density data points, and is far away from another high-density data point. As a result, each sample is analyzed in terms of its local density ρ and its distance δ to other data points with higher local density. The DP clustering algorithm [40] can be summarized as follows:

$$d_{ij} = \|r_i - r_j\|_2^2 \tag{5}$$

where r_i and r_j are the points that belong to a set $\mathbb{C} = \{r_\tau\}_{\tau=1}^n$, n represents the number of points, and d_{ij} is the Euclidean distance between points r_i and r_j. With the aforementioned definitions in mind, the kernel-based local density ρ_i of data point r_i can be calculated through one of the following two methods:

$$\rho_i = \begin{cases} \sum_j \chi(d_{ij} - d_c), & \text{Cut-off kernel} \\ \sum_j e^{-\left(\frac{d_{ij}}{d_c}\right)^2}, & \text{Gaussian kernel} \end{cases} \tag{6}$$

where d_c is the cut-off distance (which refers to the radius of the search region). The Gaussian kernel has the advantage of decreasing the negative impact of the statistical errors caused by the limited availability points. As a result, the Gaussian kernel-based local density has demonstrated to be successful in HSI data interpretation [41]–[43]. The local density ρ_i reflects the number of data points that are closer than d_c to the data point r_i.

Once ρ_i is obtained, another value δ_i is obtained as follows:

$$\delta_i = \begin{cases} \max_j(d_{ij}), & \text{if } \rho_i = \max(\rho) \\ \min_{j:\rho_j > \rho_i}(d_{ij}), & \text{Otherwise.} \end{cases} \tag{7}$$

In particular, if the point r_i has the highest local density, then δ_i is set to the maximum distance from the point r_i to any other point. Otherwise, δ_i is defined as the minimum distance from the point r_i to any other point with higher local density. Finally, a score γ_i used to find out cluster centers as follows:

$$\gamma_i = \rho_i \times \delta_i \tag{8}$$

where γ_i considers jointly ρ_i and δ_i (the larger the score of γ_i, the more likely that the associated point r_i is a cluster center). Therefore, a hint for choosing the number of centers is provided by the solution of γ_i, ordered in decreasing order [40].

III. Proposed Method

Different from a fixed-shape spatial neighborhood, superpixels adaptively obtain spatially homogeneous regions that preserve the edge information of spatial structures. However,

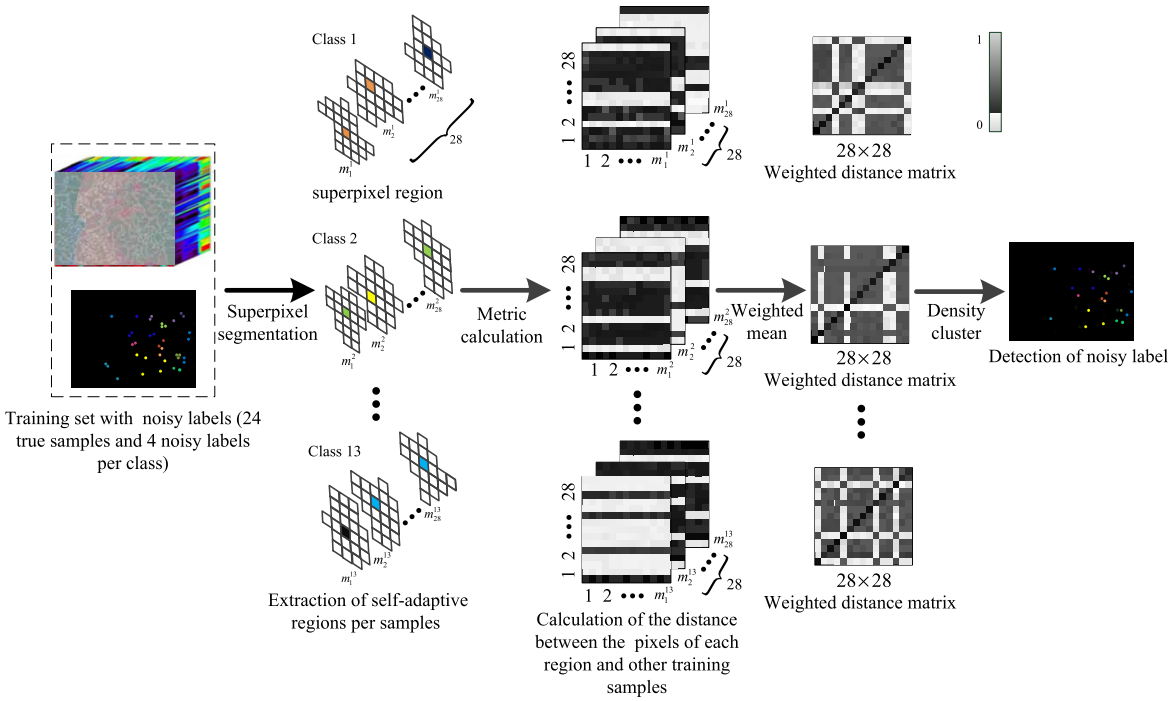


Fig. 1. Overview of the proposed SPWD method for noisy label detection prior to HSI classification. m_t^c represents the neighboring pixels of the tth training sample of the cth class.

it is difficult to guarantee that all pixels belonging to the same superpixel block share identical class information and that close pixels in spectral space have the same label. To address these issues, we develop a new noisy label detection method that combines an innovative superpixel-to-pixel weighting distance (SPWD) and the DP clustering algorithm, as illustrated in Fig. 1 and Algorithm 1. Our newly developed method consists of the following three parts: 1) construction of self-adaptive regions; 2) adjustment of spectral–spatial information; and 3) detection of noisy labels. In the following, we describe each of these parts in detail.

A. Construction of Self-Adaptive Regions

In our method, the spectral features (i.e., pixel values) are exploited to measure the difference among training samples. In fact, we use PCA to reduce the dimensionality of the input data only before applying superpixel segmentation, but then work with the full spectral information. We emphasize that our method is fully compatible with the processing of transformed spectral features (not only PCA but also other transformations). Our main reason to focus on the original spectral information is to illustrate the advantages of our newly proposed method in the most simple case. In order to obtain the superpixel segmentation on $\boldsymbol{I}$, we proceed as follows:

$$\boldsymbol{P}_e = \mathrm{PCA}(\boldsymbol{I}) \tag{9}$$

where $\boldsymbol{P}_e$ are the three first principal components retained to capture a significant amount of the spectral and spatial information of the original image $\boldsymbol{I}$.

Algorithm 1 SPWD

Inputs: 1) HSI; 2). Noisy training set; 3). Number of principal components.
Output: Improved training set T and classification results obtained by the SVM trained with T.

1: Obtain P_e from the original HSI using the PCA algorithm.
2: Construct a self-adaptive region for each sample by using the ERS algorithm to the P_e.
3: **for** training sample $\mathbf{x}_t^c$ **do**
4: Select a set of neighboring pixels from a superpixel block of $\mathbf{x}_t^c$, where the size can be adjusted by the parameter S_n.
5: Obtain the distance information $\mathbf{D_S^c}(\mathbf{u}, \mathbf{v})$ for each class between the samples in the class and all the pixels in each superpixel.
6: Calculate the weighted coefficients $\mathbf{W_S^c}(\mathbf{u}, \mathbf{v})$ based on the aforementioned distance information.
7: Redefine the distance information among training samples by applying KNN Gaussian weighting distance $\mathbf{D^c_{S_K,W}}(\mathbf{u}, \mathbf{v})$.
8: Use the weighted distance information to calculate the local density information by means of the DP algorithm.
9: Cleanse the original training set by using the decision function, and build an improved training set $\mathbf{T}$.
10: **end for**
11: Perform classification using the SVM trained with the improved training set $\mathbf{T}$.

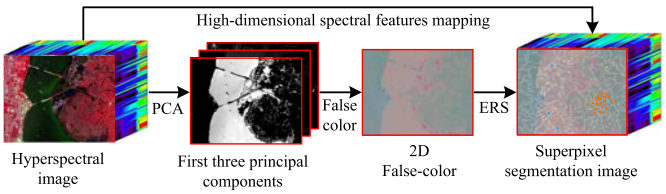


Fig. 2. Outline of the superpixel generation process for HSIs.

Then, the ERS is utilized on the $\boldsymbol{P}_e$ components to obtain S_n nonoverlapping 2-D superpixel regions. This is done as follows:

$$\boldsymbol{I}_s = \mathrm{ERS}_{S_n}(\boldsymbol{P}_e) \tag{10}$$

where $\boldsymbol{I}_s$ refers to the segmentation details corresponding to S_n blocks. Meanwhile, the original image $\boldsymbol{I}$ (with the full spectral content) is divided into blocks based on the aforementioned superpixel segmentation details (obtained in the reduced space given by $\boldsymbol{P}_e$). As a result, hereafter, we work with the full spectral dimensionality of the data. The outline of the superpixel generation process is given in Fig. 2.

B. Adjustment of Spectral–Spatial Information

Let us assume that the original training set is given by $\mathbf{X} = \{\mathbf{X}^c, \mathbf{Q}^c\}_{c=1}^{l} \in (\mathbb{R}^{m\times n} \times \mathbf{Q})^c$, where $\mathbf{X}^c = \{\mathbf{x}_t^c\}(t = 1, 2, \ldots, n)$, $\mathbf{Q}^c$ refers to the cth class label, l is the number of classes, n is the number of training samples in the cth class, and m is the number of samples in the training set. First, we compute the distance between the training samples in each class using the spectral angle mapper as follows:

$$\mathbf{D}^{\mathbf{c}}(\mathbf{u}, \mathbf{v}) = \arccos\left(\frac{\langle \mathbf{x}_{\mathbf{u}}^{\mathbf{c}}, \mathbf{x}_{\mathbf{v}}^{\mathbf{c}} \rangle}{\|\mathbf{x}_{\mathbf{u}}^{\mathbf{c}}\| \|\mathbf{x}_{\mathbf{v}}^{\mathbf{c}}\|}\right) \tag{11}$$

where $\mathbf{x}_{\mathbf{u}}^{\mathbf{c}}$ and $\mathbf{x}_{\mathbf{v}}^{\mathbf{c}}$ refer to the uth and vth classes; note that $u, v \in (1, 2, \ldots, n)$, training sample of cth class, respectively. Here, taking into account the advantages of spatial contextual information, the neighboring pixels from the same superpixel block of the training samples are introduced in the calculation to adjust the original distance information, which is represented as follows:

$$\mathbf{D}_{\mathbf{S}}^{\mathbf{c}}(\mathbf{u}, \mathbf{v}) = [\mathbf{D}^{\mathbf{c}}(\mathbf{u}, \mathbf{v}_1), \mathbf{D}^{\mathbf{c}}(\mathbf{u}, \mathbf{v}_2), \ldots, \mathbf{D}^{\mathbf{c}}(\mathbf{u}, \mathbf{v}_N)] \tag{12}$$

where $\mathbf{D}^{\mathbf{c}}(\mathbf{u}, \mathbf{v}_\xi)$ denotes the distance information from $\mathbf{x}_{\mathbf{u}}^{\mathbf{c}}$ to the ξth neighborhood of $\mathbf{x}_{\mathbf{v}}^{\mathbf{c}}$ for the training samples of the cth class, and N refers to the number of training samples of the superpixel block where the $\mathbf{x}_{\mathbf{v}}^{\mathbf{c}}$ is located.

Then, to reduce the redundancy of spatial information from a superpixel and improve the balance of the spectral feature-based distance information among pixels that belong to the same block, a Gaussian weighting is applied to the distance calculation as follows:

$$\mathbf{W}_{\mathbf{S}}^{\mathbf{c}}(\mathbf{u}, \mathbf{v}) = e^{-\frac{\mathbf{D}_{\mathbf{S}}^{\mathbf{c}}(\mathbf{u},\mathbf{v})^2}{2C_t{}^2}} \tag{13}$$

where C_t refers to the half-peak width that controls the attenuation of weighting. Therefore, the distance between $\mathbf{x}_{\mathbf{u}}^{\mathbf{c}}$ and $\mathbf{x}_{\mathbf{v}}^{\mathbf{c}}$ is defined by

$$\mathbf{D}_{\mathbf{S_K},\mathbf{W}}^{\mathbf{c}}(\mathbf{u}, \mathbf{v}) = \frac{\sum_{b=1}^{K} \mathbf{W}_{\mathbf{S_K}}^{\mathbf{c}}(\mathbf{u}, \mathbf{v_b}) \cdot \mathbf{D}_{\mathbf{S_K}}^{\mathbf{c}}(\mathbf{u}, \mathbf{v_b})}{\sum_{b=1}^{K} \mathbf{W}_{\mathbf{S_K}}^{\mathbf{c}}(\mathbf{u}, \mathbf{v_b})} \tag{14}$$

where $\mathbf{D}_{\mathbf{S_K}}^{\mathbf{c}}(\mathbf{u}, \mathbf{v_b})$ represents the bth distance information when selecting K minimum distances from the $\mathbf{D}_{\mathbf{S}}^{\mathbf{c}}(\mathbf{u}, \mathbf{v})$. $\mathbf{W}_{\mathbf{S_K}}(\mathbf{u}, \mathbf{v_b})$ refers to the weight corresponding to $\mathbf{D}_{\mathbf{S_K}}^{\mathbf{c}}(\mathbf{u}, \mathbf{v_b})$. A general flowchart of the KNN-based Gaussian weighting distance is illustrated in Fig. 3.

C. Detection of Noisy Labels

Once the superpixel-to-pixel distance is obtained, the cutoff distance can be obtained as follows:

$$\mathbf{D}_{\mathbf{S_K},\mathbf{W-CD}}^{\mathbf{c}}(\mathbf{u}, \mathbf{v}) = E^c(\theta) \text{ s.t. } \theta = \left\langle \frac{n \cdot (n-1)}{100} \cdot p \right\rangle \tag{15}$$

where E^c is a matrix that sorts the non-zero elements in the upper triangular matrix of $\mathbf{D}_{\mathbf{S_K},\mathbf{W}}^{\mathbf{c}}(\mathbf{u}, \mathbf{v})$ from the smallest to the largest, parameter p adjusts the size of the cutoff distance, and $< \cdot >$ refers to the round operation.

After obtaining the cutoff distance, the Gaussian kernel-based local density $\boldsymbol{\rho}^c = \{\rho_1^c, \ldots, \rho_u^c, \ldots, \rho_n^c\}$ can be calculated for each training sample in the cth class of a noisy training set as follows:

$$\boldsymbol{\rho}^c = \sum \exp\left\{-\left(\frac{\mathbf{D}_{\mathbf{S_K},\mathbf{W}}^{\mathbf{c}}(\mathbf{u}, \mathbf{v})}{\mathbf{D}_{\mathbf{S_K},\mathbf{W-CD}}^{\mathbf{c}}(\mathbf{u}, \mathbf{v})}\right)^2\right\}. \tag{16}$$

With the aforementioned local density, the noisy labels of each class can be effectively detected and removed by

$$\mathbf{T}^{\mathbf{c}} = \begin{cases} \mathbf{x}_{\mathbf{t}}^{\mathbf{c}}, & \text{if } \rho_t^c \geq \lambda \cdot \bar{\rho}^c \\ \varnothing, & \text{Otherwise} \end{cases} \tag{17}$$

where $\mathbf{T} = \{\mathbf{T}^c, \mathbb{Q}^c\}_{c=1}^{l}$ refers to the improved training set in which noisy labels have been detected and removed, and λ is a free parameter controlling the decision threshold.

IV. Experimental Results

A. Data Sets Description

To verify the effectiveness of the proposed SPWD method, experiments have been conducted on four HSI data sets, described as follows.

1) *University of Pavia*: This image was obtained over the campus of the University of Pavia, Italy, by the Reflective Optics System Imaging Spectrometer (ROSIS-3). The image is of size 610 × 340 pixels with a spatial resolution of 1.3 m per pixel (mpp) and 115 spectral bands. After removing 12 noisy bands, experiments were conducted on the remaining 103 bands. Fig. 4(a)–(c), respectively, shows the false-color composite of the University of Pavia image, the corresponding labeled data, and the class labels.
2) *Kennedy Space Center*: This image was collected by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) over the Kennedy Space Center, Florida. The image has 512 × 614 pixels and 224 spectral bands.

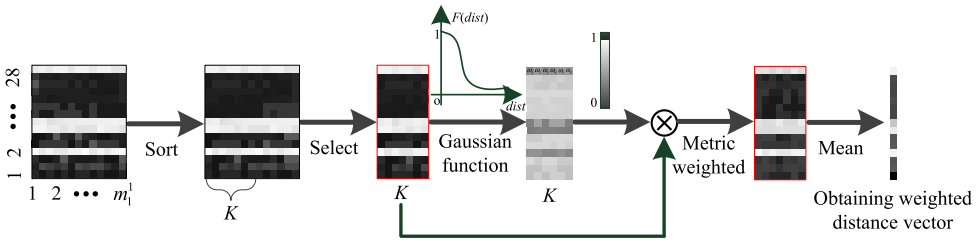


Fig. 3. Illustration of the KNN-based Gaussian weighting process in the SPWD algorithm.

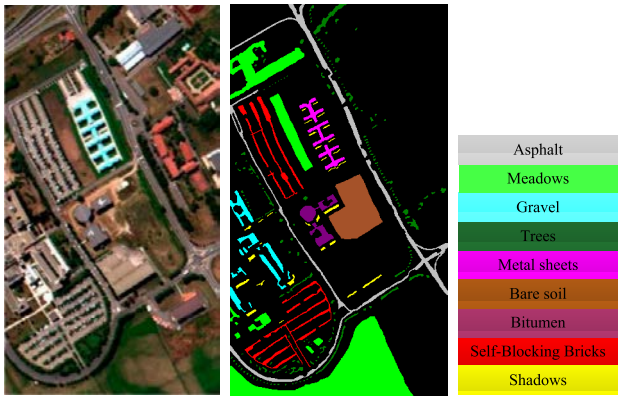


Fig. 4. University of Pavia data set. (a) Three-band color composite. (b) Labeled data. (c) Class names.

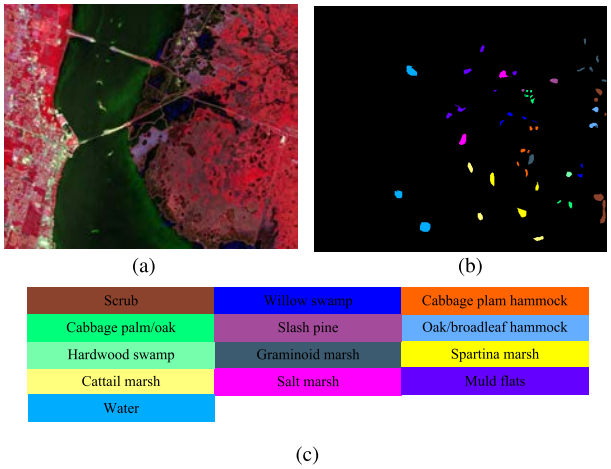


Fig. 5. Kennedy Space Center data set. (a) Three-band color composite. (b) Labeled data. (c) Class names.

Forty-eight bands have been removed due to water absorption and low signal to noise ratio. The false-color composite of the Kennedy Space Center image, the corresponding labeled data, and the class labels are, respectively, shown in Fig. 5(a)–(c).

3) *Salinas Valley*: This image was acquired by the AVIRIS over Salinas Valley in California. The image contains 224 bands and 512 × 217 pixels, with a spatial resolution of 3.7 mpp. Twenty water absorption bands are removed. Fig. 6(a)–(c), respectively, shows a false-color composite of the Salinas Valley image, the corresponding labeled data, and the class labels.
4) *Washington DC*: This image was collected by the hyperspectral digital image collection experiment (HYDICE) sensor over the Washington DC Mall. The sensor measures 210 bands (in our experiments, we removed 19 bands in the spectral range 0.9–1.4 × 10 μm). The data set contains 280 scan lines and 307 pixels in each scan line. A false-color composite and the corresponding labeled data and labels (containing six ground reference classes) are, respectively, shown in Fig. 7(a)–(c).

The SVM is one of the most widely used pixel-wise classifiers and has been adopted in this work as a baseline to evaluate the performance of the proposed SPWD method. The SVM is implemented with the LIBSVM library [44] by using the radial basis function kernel. Moreover, the parameters of the SVM are determined using five-fold cross-validation. To make the comparison fair, the represented quality indexes

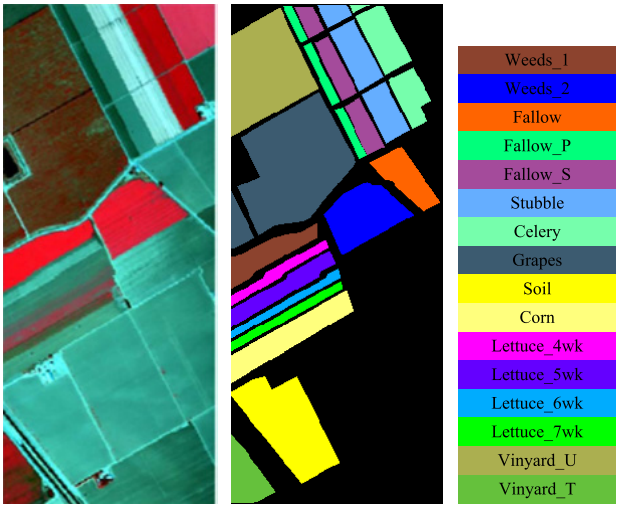


Fig. 6. Salinas Valley data set. (a) Three-band color composite. (b) Labeled data. (c) Class names.

of overall accuracy (OA), average accuracy (AA), Kappa coefficient (κ), and class individual accuracies are calculated by averaging the results obtained after ten repeated Monte Carlo experiments with different randomly selected training samples and noisy labels, and the mean and the standard deviation after such experiments are provided in our experiments. Note that OA and AA indicators are used in percentage formation to present experimental results in this article. The training

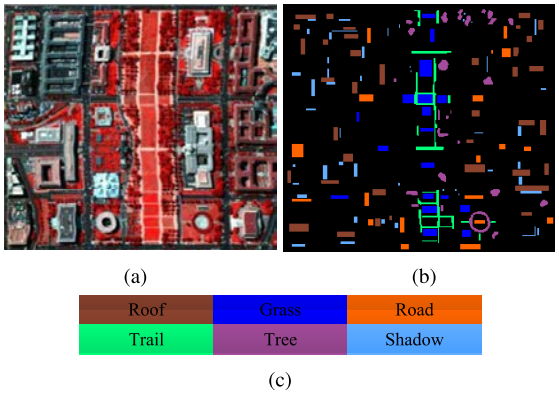


Fig. 7. Washington DC data set. (a) Three-band color composite. (b) Labeled data. (c) Class names.

sets are constructed using the labeled samples available in the ground truth. For each class, some pixels (randomly selected from other classes) are added to simulate the problem of "noisy label."

B. Parameter Settings

In this section, the influence of the parameters relevant to the performance of the proposed method is analyzed, such as the number of superpixel blocks S_n, the half-peak C_t of the Gaussian weight function, the threshold parameter λ of decision function, and the nearest neighbor K. The experiments are respectively performed on four real data sets, i.e., University of Pavia, Kennedy Space Center, Salinas Valley, and Washington DC data sets. For the University of Pavia data set, the training set consists of 52 true training samples and 8 noisy labels for each class. For the KSC, Salinas, and Washington DC data sets, the training set consists of 24 true training samples and 4 noisy labels for each class, respectively.

In the first experiment, the impact of the superpixel blocks S_n and the half-peak width C_t on the performance of the proposed method is tested in the classification of the above-mentioned hyperspectral data sets. The ranges of S_n and C_t are set to 4.5×10^3 to 9.5×10^3 and 5.0×10^{-2} to 1.5×10^{-1}, respectively. As shown in Fig. 8, it can be observed that the OAs of the classification results achieved by the proposed method exhibit a close relationship with the variation of the parameter value. For instance, if we fix S_n and observe C_t, it can be found from Fig. 8(a)–(c) that the OAs will decrease as the value of C_t increases, due to the fact that C_t controls the extreme value of the Gaussian weighting function, leading to a different degree of distance weight between different samples. Similarly, when C_t is fixed, it can be observed that the change of superpixel parameter S_n exhibits different degrees of influence on the OAs, which indicates that S_n controls the amount of spatial information that is effective for noisy label detection. Furthermore, an optimal parameter value can be determined based on the topmost classification accuracy of SVM, using the improved training set. It can be seen from Fig. 8 that the highest OA of SVM using the improved training set on the University of Pavia, Kennedy Space Center, Salinas Valley, and Washington DC data sets is 83.72% ($C_t = 0.08$, $S_n = 5000$), 87.36% ($C_t = 0.11$, $S_n = 9000$), 82.56% ($C_t = 0.13$, $S_n = 5500$), and 85.16% ($C_t = 0.09$, $S_n = 5500$), respectively. Therefore, these values of C_t and S_n are set to default parameters corresponding to different data sets in this article.

In the second experiment, the effectiveness of the threshold parameter λ and the nearest neighbor K on the performance of the proposed method is evaluated in the classification of the aforementioned hyperspectral data sets. The threshold parameter is chosen from $\lambda = 1.0 \times 10^{-2}$ to $\lambda = 11.0 \times 10^{-2}$ to solve (17) and the nearest neighbor K is selected from $K = 2$ to $K = 22$ to solve (14). As shown in Fig. 9, it can be observed that the value of OA rises first and then falls as the λ changes in the interval of $\lambda = 1.0 \times 10^{-2}$ to $\lambda = 11.0 \times 10^{-2}$ when the K is fixed. The reason is that the size of the λ controls the number of samples to be removed for each class. Moreover, when the parameter λ is fixed, it can be seen from Fig. 8 that the value of OA shows an increasing (and then decreasing) trend with an increase of K. The reason is that a small neighborhood may lack sufficient spatial information, and a large neighborhood may lead to the introduction of spatial information with dissimilar pixels. Therefore, according to the highest OA obtained by the SVM on various data sets, the default parameters of the proposed method on the University of Pavia, Kennedy Space Center, Salinas Valley, and Washington DC data sets are set to ($\lambda = 0.11$, $K = 12$), ($\lambda = 0.1$, $K = 6$), ($\lambda = 0.1$, $K = 4$), and ($\lambda = 0.1$, $K = 10$), respectively.

C. Component Analysis

In this section, we conduct experiments with the Kennedy Space Center data set (with 24 true samples and 4 noisy labels per class) and with the University of Pavia data set (with 52 true samples and 8 noisy labels per class). The first experiment is implemented to analyze the effectiveness of the proposed method with different kinds of superpixel algorithms, i.e., SLIC [35] and ERS [36]. As shown in Tables I and II, under the same experimental conditions, it can be observed that the ERS algorithm leads to better performance in terms of classification accuracy and time consumption with respect to the SLIC algorithm. In addition, schematic segmentation results obtained by the ERS and SLIC algorithm on the above KSC and University of Pavia scenes are shown in Figs. 10 and 11, respectively. Focusing on Fig. 10, the superpixel blocks achieved by the ERS algorithm are more shape-adaptive than those obtained by the SLIC algorithm. Therefore, the ERS-based superpixel segmentation approach is employed for building the proposed method and used in the following experiments.

The second experiment is conducted on the Kennedy Space Center data set with a training set that is made up of 24 true samples and 4 noisy labels for each class. In this experiment, the performance of the proposed method (using different distance metrics, such as Euclidean [45], spectral information divergence (SID) [38], correlation coefficient (CC) [46], and SAM [47], is represented in Table III. It can be seen that the proposed method achieves the highest classification accuracies

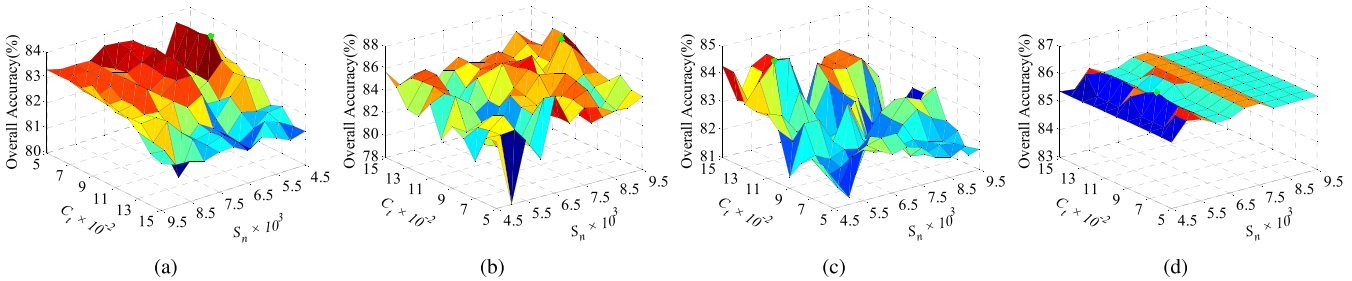


Fig. 8. Influence of parameters S_n and C_t on the performance of the proposed SPWD method. (a) University of Pavia data set (with 52 true and 8 noisy labels per class). (b) Kennedy Space Center data set (with 24 true and 4 noisy labels per class). (c) Salinas Valley data set (with 24 true and 4 noisy labels per class). (d) Washington DC data set (with 24 true and 4 noisy labels per class).

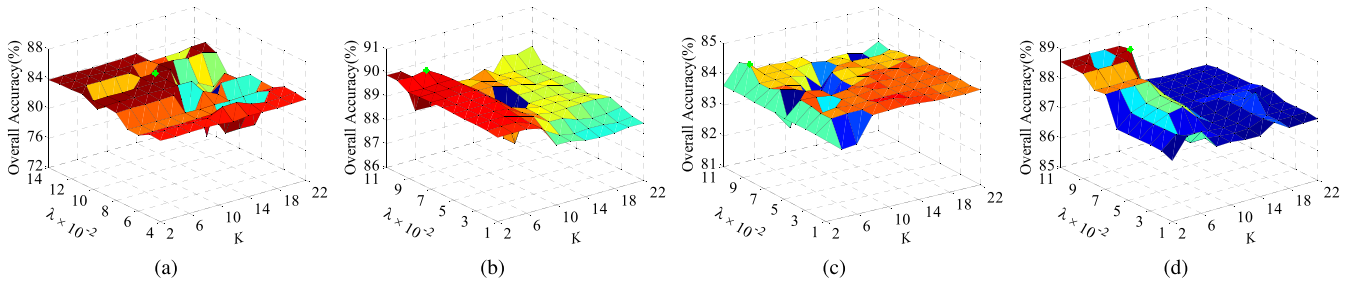


Fig. 9. Influence of parameters K and λ on the performance of the proposed SPWD method. (a) University of Pavia data set (with 52 true and 8 noisy labels per class). (b) Kennedy Space Center data set (with 24 true and 4 noisy labels per class). (c) Salinas Valley data set (with 24 true and 4 noisy labels per class). (d) Washington DC data set (with 24 true and 4 noisy labels per class).

TABLE I

CLASSIFICATION PERFORMANCE OBTAINED BY THE SPWD METHOD USING DIFFERENT SUPERPIXEL SEGMENTATION ALGORITHMS FOR THE UNIVERSITY OF PAVIA DATA SET (WITH 52 TRUE SAMPLES AND 8 NOISY LABELS PER CLASS). THE NUMBER IN THE PARENTHESIS REPRESENTS THE STANDARD VARIATION OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Indexes	SLIC [35]	ERS [36]
OA(%)	80.96(1.96)	**81.38(1.33)**
AA(%)	77.83(1.80)	**78.06(2.81)**
κ	0.757(0.02)	**0.761(0.02)**
Time(s)	26.08	25.83

TABLE II

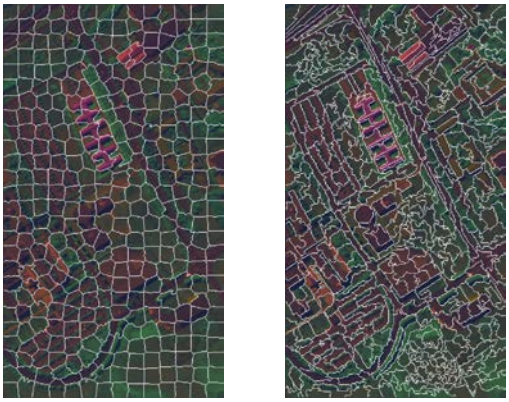
CLASSIFICATION PERFORMANCE OBTAINED BY THE SPWD METHOD USING DIFFERENT SUPERPIXEL SEGMENTATION ALGORITHMS FOR THE KENNEDY SPACE CENTER DATA SET (WITH 24 TRUE SAMPLES AND 4 NOISY LABELS PER CLASS). THE NUMBER IN THE PARENTHESIS REPRESENTS THE STANDARD VARIATION OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Indexes	SLIC [35]	ERS [36]
OA(%)	86.29(1.89)	**87.57(1.11)**
AA(%)	81.50(2.02)	**82.55(1.28)**
κ	0.847(0.02)	**0.862(0.01)**
Time(s)	30.85	30.62

using the SAM. Therefore, in this article, the SAM metric is utilized.

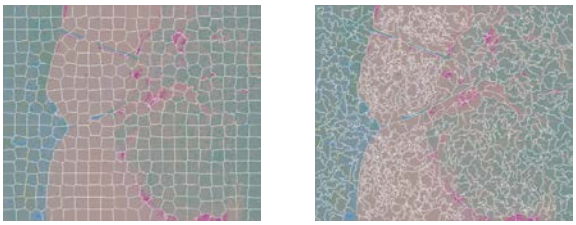
D. Detection Performance Analysis

To further illustrate that the proposed method can effectively detect and remove noisy labels in the training set, an experimental analysis of detection performance is conducted for different data sets. For the University of Pavia scene, the initial training set contains 52 true samples and a different number of noisy labels for each class. For the Kennedy Space Center,



(a) (b)

Fig. 10. Segmentation map (400 blocks) obtained by different superpixel segmentation methods for the University of Pavia data set. (a) SLIC method. (b) ERS method.



(a) (b)

Fig. 11. Segmentation map (400 blocks) obtained by different superpixel segmentation methods for the Kennedy Space Center data set. (a) SLIC method. (b) ERS method.

TABLE III

CLASSIFICATION PERFORMANCE OBTAINED BY THE SPWD METHOD USING DIFFERENT METRICS FOR THE KENNEDY SPACE CENTER DATA SET (WITH 24 TRUE SAMPLES AND 4 NOISY LABELS PER CLASS). THE NUMBER IN THE PARENTHESIS REPRESENTS THE STANDARD VARIATION OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Metric	ED [45]	SID [38]	CC [46]	SAM [47]
OA(%)	86.62(1.82)	86.83(1.87)	87.25(1.72)	**87.97(2.14)**
AA(%)	81.50(2.30)	82.12(1.97)	82.16(1.47)	**82.45(2.04)**
κ	0.851(0.02)	0.853(0.02)	0.858(0.02)	**0.866(0.02)**

TABLE IV

DETECTION PERFORMANCE (NUMBER) OF NOISY LABELS FOR THE PROPOSED METHOD ON DIFFERENT DATA SETS. $T_n \times l$ REFERS TO THE TOTAL NUMBER OF NOISY LABELS IN THE TRAINING SET. T_n REPRESENTS THE NUMBER OF NOISY LABELS PER CLASS. THE NUMBER IN THE PARENTHESIS REPRESENTS THE STANDARD VARIATION OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Dataset	Total noise	Detection performance of noisy label		
		Correct	Error	Not detected
University of Pavia	8×9	40.53(4.67)	8.00(3.60)	31.47(4.67)
	16×9	52.71(2.71)	7.14(3.87)	91.29(2.71)
	24×9	59.29(7.69)	5.29(1.83)	156.71(7.69)
KSC	4×13	37.33(0.47)	5.33(2.05)	14.67(0.47)
	8×13	53.25(4.49)	5.75(2.77)	50.75(4.49)
	12×13	67.40(2.33)	2.60(1.74)	88.60(2.33)
Salinas	4×16	52.14(1.73)	6.86(1.88)	11.86(1.73)
	8×16	89.57(4.87)	5.29(1.75)	38.43(4.87)
	12×16	99.71(5.75)	1.86(1.36)	92.29(5.75)
Washington DC	4×6	21.14(0.99)	7.57(2.44)	2.86(0.99)
	8×6	32.86(2.90)	2.71(2.12)	15.14(2.90)
	12×6	42.29(3.92)	1.29(0.88)	29.71(3.92)

Salinas Valley and Washington DC scenes, the initial training sets contain 24 true samples and various number of noisy labels for each class, respectively.

The first experiment is conducted to analyze the influence of the number of iterations on the performance of the proposed method. The proposed method repeats (12)–(17), and the key idea of the iterative process is that the previous output is used as the next input until a stopping criterion is met. As shown in Table IV, the proposed method obtains a low false detection rate (see third column), which refers to the fact that a small number of true samples are wrongly detected, and achieves pretty good detection performance when the number of noisy labels in each class of the training set is less than the number of true samples. However, there are still some noise labels in the improved training set (see the fourth column), especially when a multitude of noisy labels are still present in the original training set. The reason is that a decision threshold-based removal solution is not satisfactory with an original training set that contains a large number of noisy labels.

In addition, an iterative detection process (based on the proposed method) is introduced into the original training set to further remove the noisy labels. As shown in Fig. 12, it can be seen that the OA decreases as the number of iterations increases when the original training set contains a low number of noisy labels (see red column) on the different data sets. However, when the training set contains more noisy labels (see green and blue columns), the OAs rise first and then fall with the number of iterations. This means that iterative detection can achieve better detection performance in a training set with a large number of noisy labels. In order to balance the tradeoff between efficiency and classification accuracy, the number of iterations of the proposed method is set to one for the four considered data sets.

TABLE V

DETECTION PERFORMANCE OF THE SPWD WITH DIFFERENT PARTITION STRATEGIES FOR GENERATING THE TRAINING AND TEST SETS FOR THE SALINAS DATA SET (USING 24 TRUE SAMPLES AND 8 NOISY LABELS FOR EACH CLASS). THE NUMBER IN THE PARENTHESIS REPRESENTS THE STANDARD VARIANCE OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Indexes	SVM 24 (true)	SVM 24(true) + 8(noise)	SPWD (random way)
OA(%)	84.77(1.86)	81.16(1.01)	**83.94(1.46)**
AA(%)	90.81(0.93)	89.14(1.64)	**89.63(1.77)**
κ	0.831(0.02)	0.792(0.01)	**0.821(0.02)**
Indexes	SVM 24 (true)	SVM 24(true) + 8(noise)	SPWD (disjoint way)
OA(%)	76.07(1.82)	71.12(3.19)	**79.10(1.44)**
AA(%)	83.72(1.56)	78.45(3.74)	**87.09(1.72)**
κ	0.737(0.02)	0.683(0.03)	**0.769(0.02)**

The second experiment is performed on the Salinas Valley data set (with 24 true samples and 8 noisy labels for each class). In this experiment, the detection performance of the proposed SPWD method has been evaluated using different partition methods for generating the training and test sets. As shown in Fig. 13(a) and (b), the training samples are first selected randomly from the available ground truth, and then we use the remaining labeled samples for testing. In Fig. 13(c) and (d), another way of creating the training and testing set is considered, in which the training and the test samples are always disjoint, as indicated in [48]. These disjoint data sets are available from the GRSS Data and Algorithm Standard Evaluation website: http://dase.grss-ieee.org.

The experimental results reported in Table V indicate that the proposed SPWD method exhibits better performance in terms of classification accuracies than the SVM trained with the disjoint training set. The reason is that the difference between true samples and noisy labels is more obvious in the disjoint sample set. Taking into account the universality of the proposed SPWD method, the random strategy for generating the training and test sets is adopted to conduct the subsequent experiments on the proposed method.

E. Classification Performance Evaluation With the SVM

In this section, the classification performance of different methods (such as DP, SDP, KSDP, and SPWD) is evaluated by using an SVM classifier (trained with different improved training sets) on the University of Pavia, Kennedy Space Center, Salinas Valley, and Washington DC data sets. For the University of Pavia data set, the experiments are conducted with 52 true samples and different numbers of noisy labels in the range of 8–24 per class. For the Kennedy Space Center, Salinas Valley, and Washington DC data sets, the experiments are conducted with 24 true samples and

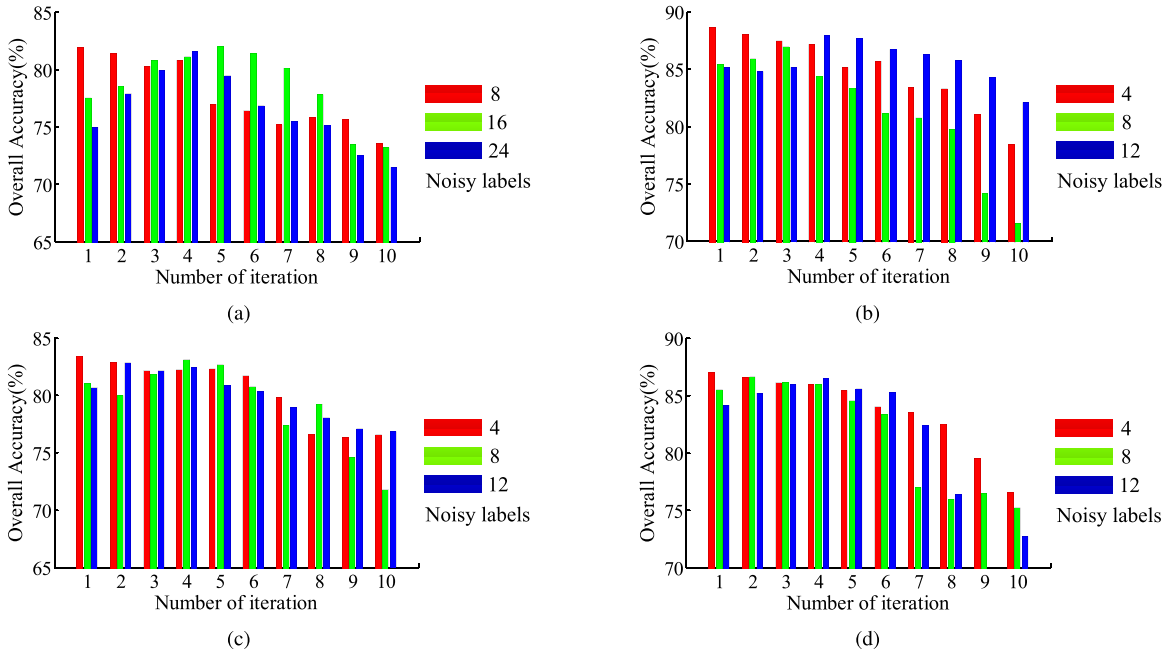


Fig. 12. Classification accuracy achieved by the SPWD with different numbers of iterations. (a) University of Pavia data set (with 52 true samples and a different number of noisy labels per class). (b) Kennedy Space Center data set (with 24 true samples and a different number of noisy labels per class). (c) Salinas Valley data set (with 24 true samples and different numbers of noisy labels per class). (d) Washington DC data set (with 24 true samples and different numbers of noisy labels per class).

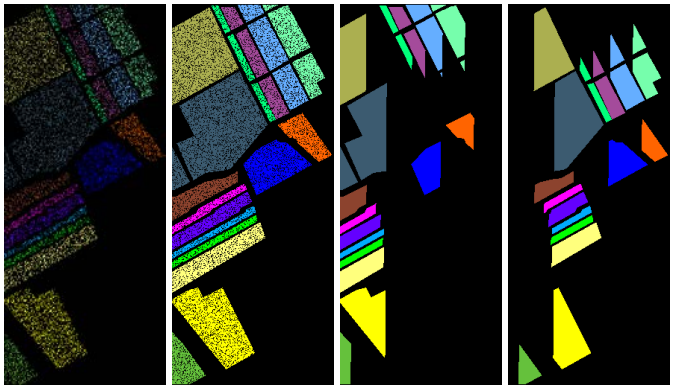


Fig. 13. Illustration of different train/test splits on the Salinas Valley data set. (a) Train(random). (b) Test(random). (c) Train(disjoint). (d) Test(disjoint).

different numbers of noisy labels in the range of 4–12 for each class. In Fig. 14, the classification performance of the SVM (trained using the different training sets) is provided. It can be observed that the classification results obtained by the SVM trained with improved training sets are higher than those obtained using the original training set. Specifically, the SVM trained using the training set improved by the proposed SPWD method can always achieve better classification results with respect to the improved training sets provided by other methods. This indicates that the introduction of shape-adaptive spatial information and a weighted distance metric can further improve the detection accuracy of noisy labels and, subsequently, the classification performance of the SVM classifier.

In addition, the classification results obtained by the SVM trained using different training sets on the University of Pavia data set are given in Table VI. It can be seen from Table VI that the proposed method can effectively improve the classification accuracies for most of the classes. For example, when the training set contains eight noisy labels for each class, the classification accuracy of the SVM trained using training set improved by the SPWD increases from 81.95% to 93.06% for the metal sheets class and from 76.10% to 81.28% for self-blocking bricks. Meanwhile, OAs can be increased by about 4%. This demonstrates that the noisy labels in the original training set can be effectively removed by the proposed SWPD method.

The experimental results on the Kennedy Space Center data set are shown in Fig. 15 and Table VII, respectively. Focusing on Fig. 15, we can see that the SVM trained using the with noisy labels shows several misclassifications in classes such as spartina marsh, cattail marsh, and salt marsh. Meanwhile, the classification maps of the SVM trained with the training set improved by DP, SDP, and KSDP show some improvements, but the most significant improvement is obtained after applying the proposed SPWD method. As shown in Table VII, the classification results obtained by the SVM trained with the proposed SPWD method can reach higher OAs, AAs, and Kappa than those obtained by the SVM trained with other methods.

As shown in Table VIII, the classification accuracy of the SVM trained with the proposed SPWD method can be increased by 1.84%–5.24% under different training sets with noisy labels. In addition, the experimental classification results obtained for the Washington DC data set are shown in Table IX. Here, the classification performance obtained

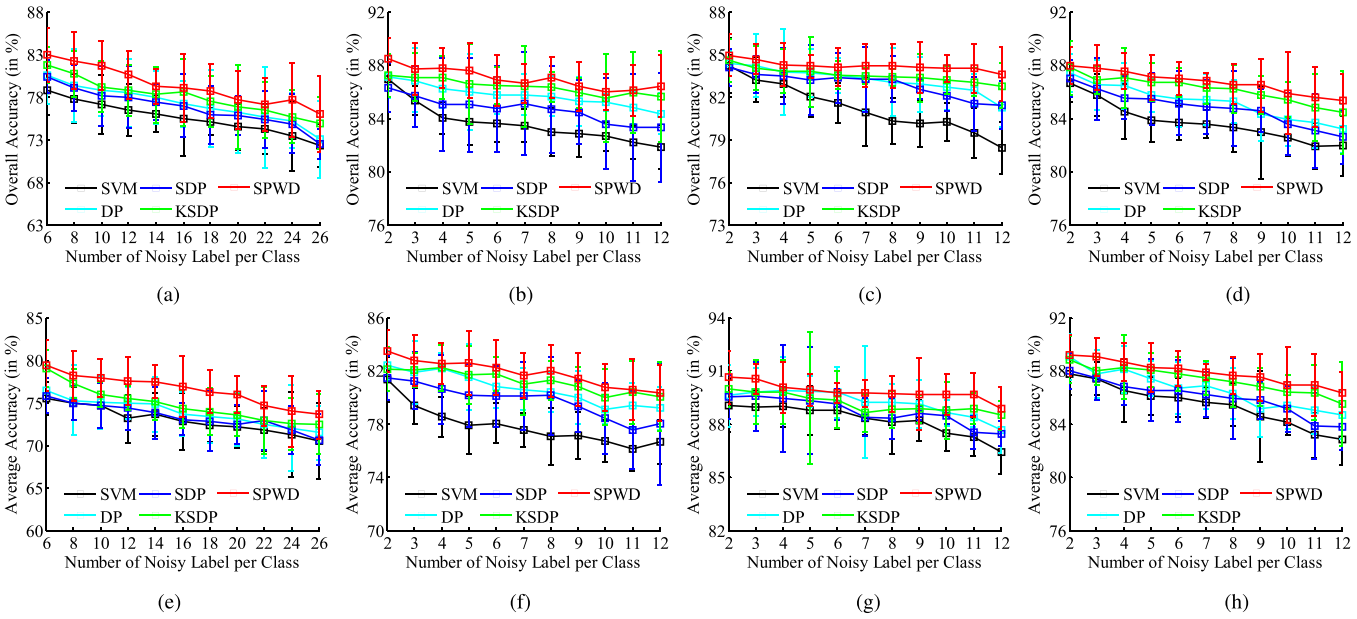


Fig. 14. (a) Performance comparison of the SVM (trained using the original training sets, and trained using the improved training sets obtained by the DP, SDP, KSDP, and SPWD methods) in terms of OA (first row) and AA (second row). (b) and (f) Experiments on the Kennedy Space Center data set with different numbers of mislabeled (varying from 2 to 12) and 20 true samples per class. (c) and (g) Experiments on the Salinas Valley data set with a different number of mislabeled (varying from 2 to 12) and 20 true samples per class. (d) and (h) Experiments on the Washington DC data set with different number of mislabeled (varying from 2 to 12) and 20 true samples per class.(e) Experiments on the University of Pavia data set with a different number of mislabeled (varying from 6 to 26) and 50 true samples per class.

TABLE VI

CLASSIFICATION PERFORMANCE OF THE SVM, DP, SDP, KSDP, AND SPWD METHODS FOR THE UNIVERSITY OF PAVIA DATA SET (WITH 52 TRUE SAMPLES AND DIFFERENT NUMBERS OF NOISY LABELS PER CLASS). THE NUMBER IN THE PARENTHESIS REPRESENTS THE STANDARD VARIATION OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Class	**SVM** 52(true)	**Number of true samples and mislabeled samples**														
		52(true)+8(noise)					52(true)+16(noise)					52(true)+24(noise)				
		SVM	DP	SDP	*K*SDP	SPWD	SVM	DP	SDP	*K*SDP	SPWD	SVM	DP	SDP	*K*SDP	SPWD
1	94.61	90.45	91.01	88.76	**92.54**	89.52	88.37	88.23	88.51	90.34	**90.75**	91.57	74.39	91.23	86.43	**92.21**
2	95.26	92.79	**94.67**	93.67	94.62	93.77	91.98	93.40	93.15	92.38	**93.48**	92.31	92.70	**94.74**	93.37	92.30
3	68.99	60.23	**68.72**	59.88	65.28	66.94	54.72	59.96	**61.47**	58.21	60.64	54.67	**59.67**	41.47	54.59	58.57
4	81.33	73.65	74.09	72.28	66.74	**73.95**	66.96	75.60	70.45	69.73	**83.56**	67.99	**82.69**	58.28	62.61	64.48
5	95.38	81.95	92.97	87.68	87.34	**93.06**	87.15	**90.74**	79.40	88.79	82.40	88.68	58.42	88.36	87.87	**96.79**
6	64.79	55.63	57.96	58.12	61.39	**65.79**	56.47	58.46	55.46	**61.14**	57.62	49.61	**57.99**	37.06	52.11	43.16
7	59.55	53.90	45.89	53.02	**59.20**	56.00	47.62	41.64	51.77	**49.30**	48.89	45.24	43.46	46.08	44.08	**45.99**
8	82.14	76.10	77.52	80.73	78.33	**81.28**	76.79	74.78	76.13	76.52	**82.24**	72.34	74.25	80.19	75.10	**80.89**
9	99.88	**90.15**	74.55	89.23	89.71	84.11	**84.97**	80.01	81.02	82.28	84.87	93.96	87.36	**100.0**	84.17	93.16
OA	84.80	77.80	79.36	79.07	80.69	**82.26**	75.49	77.20	76.97	78.59	**79.05**	73.39	75.21	74.84	76.66	**77.66**
(%)	(1.64)	(2.87)	(4.24)	(2.65)	(2.71)	**(3.36)**	(4.41)	(2.74)	(3.63)	(3.93)	**(3.95)**	(4.07)	(3.19)	(3.45)	(3.11)	**(4.33)**
AA	82.44	74.98	75.26	75.93	77.24	**78.27**	72.78	73.65	73.04	74.30	**76.05**	72.93	70.10	70.82	71.15	**74.06**
(%)	(1.25)	(1.95)	(4.12)	(2.04)	(1.70)	**(2.89)**	(3.35)	(2.60)	(1.88)	(2.22)	**(3.58)**	(4.98)	(5.03)	(2.78)	(3.04)	**(4.18)**
κ	0.803	0.719	0.739	0.734	0.754	**0.771**	0.691	0.711	0.709	0.726	**0.733**	0.668	0.688	0.684	0.706	**0.717**
	(0.02)	(0.03)	(0.04)	(0.03)	(0.03)	**(0.04)**	(0.05)	(0.03)	(0.04)	(0.04)	**(0.05)**	(0.04)	(0.03)	(0.04)	(0.03)	**(0.05)**

by the SVM trained with the improved training set provided by the proposed SPWD method can still obtain the highest accuracies for most of the classes. In particular, when the grass and shadow classes respectively contain 12 noisy labels, the classification accuracies achieve the highest OAs (69.11% and 97.30%, respectively). This suggests that the proposed SPWD method can effectively remove noisy labels and improve the classification performance of the SVM.

Finally, the running times (in seconds) for the different considered methods are reported in Table X. All of the codes have been implemented on a computer with an Intel®Core™ i7-7800X, CPU 3.50 GHz, and 32 GB of RAM, and the software platform is MATLAB R2014a. As shown in Table X, the time consumption of the SVM trained using the training set improved by the proposed SPWD method is lower than that of the SVM trained using the original noisy training set. This is a direct consequence of the fact that some noisy labels have been removed from the original training set. In addition to the time used for classification, the detection time of the different noisy label detection methods is given in Table X. It can be seen that

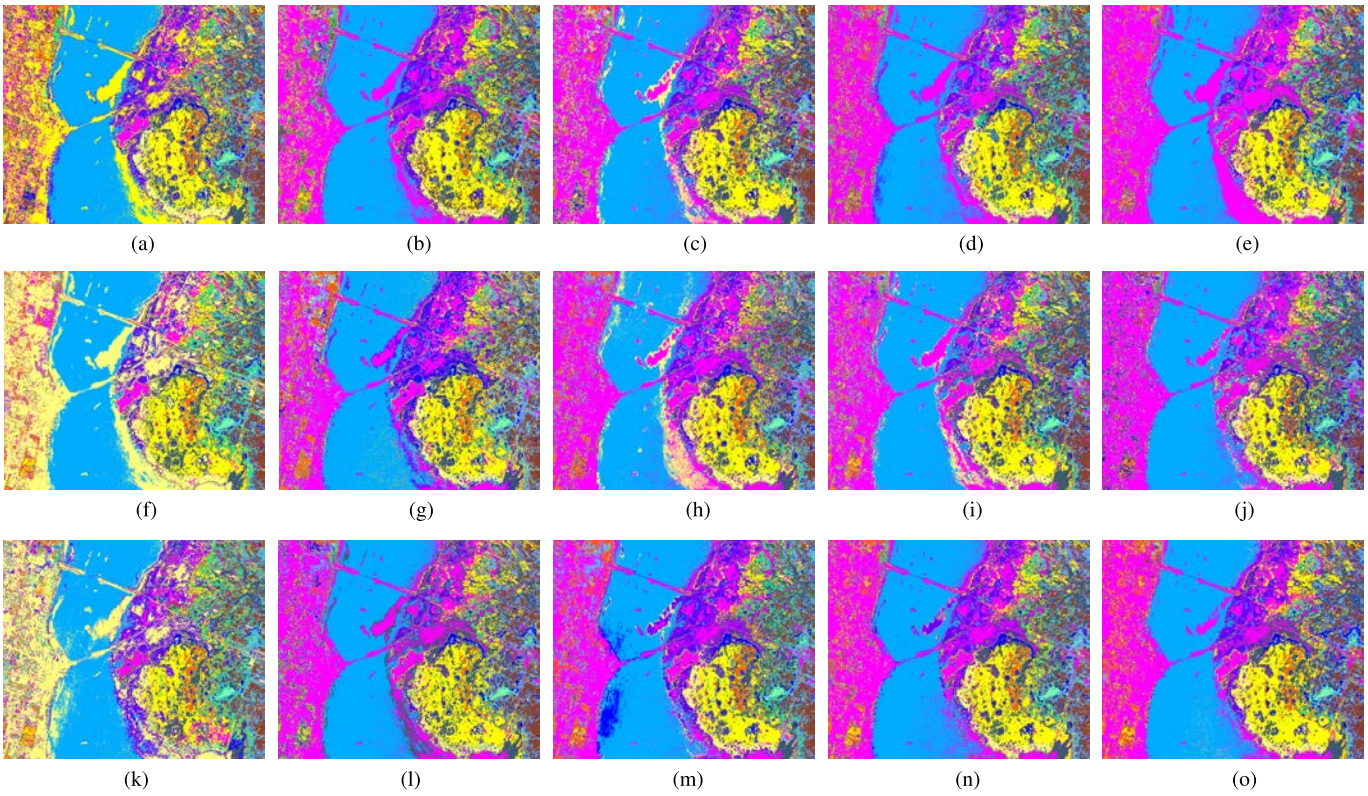


Fig. 15. Classification results (%) of various methods on the Kennedy Space Center data set. Classification maps obtained by the SVM (first column), the DP (second column), the SDP (third column), the KSDP (fourth column), and the proposed SPWD method (fifth column) trained with 24 true samples and different numbers of noisy labels per class. (a)–(e) Four noisy labels per class. (f)–(j) Eight noisy labels per class. (k)–(o) Twelve noisy labels per class. (a) OA = 83.42. (b) OA = 86.43. (c) OA = 85.02. (d) OA = 87.11. (e) OA = 88.56. (f) OA = 82.11. (g) OA = 85.64. (h) OA = 84.74. (i) OA = 86.25. (j) OA = 87.21. (k) OA = 80.92. (l) OA = 84.25. (m) OA = 83.02. (n) OA = 85.37. (o) OA = 86.22.

TABLE VII

CLASSIFICATION PERFORMANCE OF THE SVM, DP, SDP, KSDP, AND SPWD METHODS FOR THE KENNEDY SPACE CENTER DATA SET (WITH 24 TRUE SAMPLES AND DIFFERENT NUMBERS OF NOISY LABELS PER CLASS). THE NUMBER IN THE PARENTHESIS REPRESENTS THE STANDARD VARIATION OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Class	**SVM** 24(true)	**Number of true samples and mislabeled samples**														
		24(true)+4(noise)					24(true)+8(noise)					24(true)+12(noise)				
		SVM	DP	SDP	KSDP	SPWD	SVM	DP	SDP	KSDP	SPWD	SVM	DP	SDP	KSDP	SPWD
1	95.79	92.01	94.64	93.56	96.13	**95.67**	92.78	**96.12**	95.50	95.78	95.28	**96.36**	94.79	92.84	94.74	95.59
2	88.09	87.06	**88.68**	87.07	84.91	87.16	77.28	82.06	84.08	80.62	**87.03**	77.52	82.86	78.09	76.71	**83.30**
3	84.27	71.34	**91.27**	84.02	88.35	84.79	81.82	**87.76**	86.60	80.91	80.59	81.77	**86.79**	84.13	82.34	78.97
4	66.40	59.64	61.22	64.76	**66.03**	63.86	59.37	61.24	**66.18**	65.30	59.91	53.03	**64.06**	60.62	61.81	63.69
5	63.78	61.84	74.79	64.26	**66.09**	58.17	57.24	59.68	66.46	**68.87**	67.50	**66.80**	55.73	63.82	65.25	59.82
6	55.65	37.14	58.26	63.62	62.13	**63.76**	45.14	**61.20**	59.94	57.86	55.01	54.44	50.24	61.83	**63.30**	54.31
7	70.77	66.67	**70.79**	63.00	70.71	67.56	59.18	67.66	70.86	67.85	**72.00**	60.40	**68.94**	64.64	60.63	63.18
8	83.67	76.44	77.18	68.14	81.80	**85.07**	75.78	75.80	72.93	82.79	**82.81**	66.06	76.88	69.33	**79.33**	79.28
9	89.83	89.69	90.19	91.33	90.76	**91.89**	89.54	88.82	88.33	**93.60**	91.98	84.44	85.84	85.34	**90.61**	87.54
10	94.19	**95.58**	90.90	91.52	82.49	85.80	86.95	93.44	85.38	90.22	**96.69**	84.25	83.93	88.81	90.35	**92.38**
11	93.41	90.44	84.26	91.43	**93.98**	91.31	87.29	82.64	84.43	83.32	**89.42**	87.51	90.91	88.34	85.24	**94.20**
12	94.17	93.40	87.59	87.11	91.46	**93.50**	90.68	**90.89**	91.42	92.30	88.44	85.38	92.49	**93.33**	90.97	93.28
13	99.97	**99.89**	98.20	98.32	96.98	99.33	98.98	98.17	97.42	97.46	**99.10**	98.58	96.00	96.14	99.39	**98.83**
OA	88.38	84.02	86.24	85.03	87.04	**87.76**	82.96	85.65	84.67	86.34	**87.06**	81.84	84.34	83.32	85.62	**86.41**
(%)	(1.23)	(1.22)	(1.49)	(3.49)	(1.60)	**(1.49)**	(1.77)	(1.53)	(3.25)	(1.15)	**(1.53)**	(1.68)	(1.56)	(4.08)	(3.45)	**(2.35)**
AA	83.07	78.55	82.15	80.63	82.45	**82.14**	77.08	80.42	80.73	81.30	**81.98**	76.66	79.19	79.02	80.05	**80.34**
(%)	(1.77)	(1.52)	(1.17)	(2.58)	(1.72)	**(1.54)**	(2.11)	(1.51)	(2.93)	(1.47)	**(1.95)**	(1.66)	(1.62)	(4.59)	(2.58)	**(2.12)**
κ	0.871	0.822	0.847	0.833	0.855	**0.864**	0.810	0.840	0.829	0.848	**0.856**	0.798	0.825	0.814	0.840	**0.849**
	(0.01)	(0.01)	(0.02)	(0.04)	(0.02)	**(0.02)**	(0.02)	(0.02)	(0.04)	(0.01)	**(0.02)**	(0.02)	(0.02)	(0.04)	(0.04)	**(0.03)**

the SPWD needs more time to execute the detection of noisy labels than the other tested methods, as a result of the inclusion of spatial information in the detection process. However, the increase in processing time is compensated by the higher accuracy obtained by the proposed method. As the proposed method is easy to parallelize, in future developments we will

TABLE VIII

CLASSIFICATION PERFORMANCE OF THE SVM, DP, SDP, KSDP, AND SPWD METHODS FOR THE SALINAS VALLEY DATA SET (WITH 24 TRUE SAMPLES AND DIFFERENT NUMBERS OF NOISY LABELS PER CLASS). THE NUMBER IN THE PARENTHESIS REPRESENTS THE STANDARD VARIATION OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Class	**SVM** 24(true)	**Number of true samples and mislabeled samples**														
		24(true)+4(noise)					24(true)+8(noise)					24(true)+12(noise)				
		SVM	DP	SDP	KSDP	SPWD	SVM	DP	SDP	KSDP	SPWD	SVM	DP	SDP	KSDP	SPWD
1	98.65	98.22	99.41	97.54	99.03	**99.49**	99.47	99.10	96.91	95.56	**99.58**	97.95	96.24	97.75	97.77	**99.13**
2	99.21	98.23	98.05	99.10	98.89	**99.35**	99.25	99.36	**98.83**	98.47	99.16	97.46	96.78	97.16	98.62	**99.02**
3	90.72	86.02	**89.97**	85.06	87.07	83.97	84.29	87.69	86.63	88.54	**89.37**	89.90	87.83	87.13	**90.71**	86.49
4	97.19	85.50	96.65	96.47	97.09	**97.53**	97.62	93.59	96.80	96.92	**97.73**	97.26	95.94	96.57	**97.43**	97.07
5	98.27	97.27	98.48	**98.83**	98.15	98.80	**99.02**	98.43	95.68	95.22	98.33	97.86	96.57	97.66	97.05	**98.90**
6	99.95	99.48	99.99	99.94	**100.0**	96.84	99.94	99.98	99.87	**100.0**	**100.0**	99.97	97.48	98.98	**100.0**	**100.0**
7	98.18	96.07	97.91	97.25	97.44	**98.21**	98.08	**98.20**	95.72	96.03	96.18	**97.63**	97.37	96.64	97.18	97.53
8	74.46	72.25	71.01	71.16	72.60	**75.20**	68.91	**74.27**	71.62	70.88	69.99	57.59	**70.42**	69.46	69.98	69.75
9	99.43	98.28	**99.41**	98.84	98.63	98.75	99.29	**99.30**	98.84	99.05	98.53	**97.21**	98.83	96.94	98.77	98.88
10	81.27	**85.17**	77.11	75.05	71.30	75.64	80.47	80.88	79.56	**82.45**	72.92	78.99	80.75	**82.14**	77.50	74.17
11	86.75	82.74	89.58	88.38	91.28	**93.25**	88.01	91.03	84.08	83.33	**93.22**	79.38	86.09	85.53	82.35	**94.38**
12	95.91	91.93	95.35	92.66	92.88	**95.37**	92.18	89.18	92.46	91.37	**93.64**	**97.02**	90.13	91.14	88.99	81.51
13	93.92	89.28	**94.24**	89.34	91.35	94.13	94.14	91.73	87.99	89.26	**93.45**	90.97	91.40	84.62	91.10	**92.57**
14	92.58	84.46	91.23	**94.15**	93.47	88.00	84.67	83.28	82.87	91.47	**95.78**	92.92	90.02	87.04	83.74	**94.60**
15	54.48	50.27	53.27	51.56	**53.63**	49.62	45.76	51.88	51.25	**53.00**	42.24	38.45	50.35	48.97	**52.92**	50.21
16	92.00	90.51	89.58	92.84	**96.46**	95.71	**96.11**	89.47	95.00	96.04	94.52	**91.44**	86.64	93.47	92.35	93.82
OA	84.77	82.35	83.75	83.16	84.02	**84.19**	81.16	83.03	82.88	83.81	**83.94**	78.41	82.86	82.47	83.50	**83.65**
(%)	(1.86)	(1.61)	(3.02)	(1.79)	(1.54)	**(1.05)**	(1.01)	(2.36)	(1.63)	(1.11)	**(1.46)**	(1.83)	(1.33)	(1.71)	(1.58)	**(1.58)**
AA	90.81	87.85	89.94	89.26	89.96	**90.07**	89.14	89.21	88.38	89.22	**89.67**	87.63	88.30	88.20	88.53	**89.26**
(%)	(0.93)	(1.80)	(1.85)	(2.16)	(1.82)	**(1.16)**	(1.64)	(1.06)	(1.15)	(0.92)	**(1.77)**	(1.26)	(1.28)	(0.70)	(0.81)	**(1.19)**
κ	0.831	0.805	0.820	0.813	0.823	**0.810**	0.792	0.813	0.810	0.820	**0.821**	0.762	0.810	0.806	0.817	**0.814**
	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	**(0.01)**	(0.01)	(0.03)	(0.02)	(0.01)	**(0.02)**	(0.02)	(0.01)	(0.02)	(0.02)	**(0.02)**

TABLE IX

CLASSIFICATION PERFORMANCE OF THE SVM, DP, SDP, KSDP, AND SPWD METHODS FOR THE WASHINGTON DC DATA SET WITH 24 TRUE SAMPLES AND DIFFERENT NUMBER OF NOISY LABELS PER CLASS AS TRAINING SET. NUMBER IN THE PARENTHESIS REPRESENTS THE STANDARD VARIANCE OF THE ACCURACIES OBTAINED IN REPEAT EXPERIMENTS

Class	**SVM** 24(true)	**The number of true samples and mislabeled samples**														
		24(true)+4(noise)					24(true)+8(noise)					24(true)+12(noise)				
		SVM	DP	SDP	KSDP	SPWD	SVM	DP	SDP	KSDP	SPWD	SVM	DP	SDP	KSDP	SPWD
1	87.87	75.50	86.51	86.51	86.31	**88.09**	83.33	**87.99**	87.54	84.18	87.62	**90.51**	89.63	85.83	87.39	88.84
2	95.40	93.69	93.13	**96.07**	95.03	95.51	89.45	93.40	90.30	94.68	**94.91**	94.15	92.87	91.38	93.06	**95.01**
3	71.54	**76.87**	69.44	69.23	71.59	71.93	68.08	68.22	68.63	**72.67**	69.93	55.46	65.63	70.17	66.86	**66.98**
4	82.98	**83.64**	82.07	76.67	81.24	83.11	74.98	77.16	81.35	81.05	**85.97**	70.76	70.34	74.07	74.08	**74.15**
5	97.28	96.04	97.81	93.74	93.56	**96.38**	94.45	95.44	92.86	**95.94**	92.42	86.41	**94.34**	86.03	92.09	94.09
6	98.86	98.66	99.21	98.73	98.32	**98.33**	98.48	96.31	94.76	97.88	**97.93**	97.59	95.33	95.27	94.55	**97.65**
OA	87.71	84.28	86.14	85.53	86.56	**87.80**	82.84	85.40	84.72	85.99	**86.75**	80.94	83.21	82.59	83.80	**84.63**
(%)	(1.06)	(1.53)	(1.75)	(1.62)	(2.19)	**(1.35)**	(1.85)	(1.61)	(2.81)	(1.80)	**(1.50)**	(1.73)	(1.86)	(1.98)	(3.14)	**(2.03)**
AA	88.99	87.40	88.03	86.83	87.68	**88.89**	84.80	86.42	85.91	87.73	**88.12**	82.48	84.69	83.79	84.68	**86.11**
(%)	(1.15)	(1.25)	(1.78)	(1.35)	(2.39)	**(1.43)**	(1.52)	(1.88)	(3.04)	(1.91)	**(1.36)**	(1.69)	(1.83)	(1.72)	(3.15)	**(1.54)**
κ	0.848	0.803	0.829	0.822	0.834	**0.849**	0.789	0.821	0.812	0.826	**0.836**	0.770	0.796	0.787	0.802	**0.812**
	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)	**(0.02)**	(0.02)	(0.02)	(0.03)	(0.02)	**(0.02)**	(0.02)	(0.02)	(0.02)	(0.04)	**(0.02)**

accelerate its performance by resorting to implementations in high-performance computing architectures such as graphics processing units (GPUs).

F. Classification Performance Evaluation With Different Classifiers

In this section, the proposed SPWD method is incorporated to other classifiers (in addition to SVM), such as the SR classifier (SRC) [49], the extreme learning machine (ELM) [50], the basic thresholding classifier (BTC) [51], and the kernel BTC (KBTC) [52], to further demonstrate the effectiveness of the proposed noisy label detection method with other methods. The experiments are performed on the University of Pavia data with 52 true samples and 8 noisy labels. Table XI shows the classification results of the different spectral classifiers that were trained using the original noisy training set (marked as NLA) and the training set improved by the DP, SDP, KSDP, and SPWD methods. It can be observed that the spectral classifiers that use the training set improved by the proposed SPWD method always achieve better classification performance with respect to those trained with the original noisy training set and other improved training sets. For example, the OA value of the BTC, KBTC, SRC, and ELM classifiers are improved by nearly 6%–10%.

Moreover, the proposed method also has been performed with spatial–spatial classification methods. The experiments are conducted on the University of Pavia data set with 52 true training samples and 8 noisy labels. Table XI shows the classification results of different spectral–spatial classification

TABLE X

COMPARISON OF RUN TIME (SECONDS) BETWEEN DIFFERENT METHODS, WHERE THE VALUE IN THE PARENTHESIS REPRESENTS THE TIME CONSUMPTION IN THE DETECTION PROCESS

Dataset	24(true)+4(noise)				24(true)+8(noise)				24(true)+12(noise)			
	SVM	SDP	KSDP	SPWD	SVM	SDP	KSDP	SPWD	SVM	SDP	KSDP	SPWD
KSC	31.75	34.82 (0.61)	24.59 (0.55)	29.51 (2.83)	39.52	35.62 (0.67)	33.52 (0.65)	32.67 (3.66)	47.39	36.61 (0.74)	36.44 (0.76)	42.21 (4.48)
Salinas	33.33	29.93 (1.73)	30.88 (1.36)	26.12 (4.45)	41.61	36.23 (1.97)	40.08 (1.65)	30.41 (5.59)	50.39	44.35 (2.33)	47.90 (1.75)	39.11 (6.97)
Washington DC	5.50	3.96 (0.35)	4.27 (4.17)	3.30 (1.29)	6.65	5.23 (0.40)	5.55 (4.84)	5.18 (1.57)	7.83	6.33 (0.44)	6.66 (6.16)	6.10 (1.90)
Dataset	52(true)+8(noise)				52(true)+16(noise)				52(true)+24(noise)			
	SVM	SDP	KSDP	SPWD	SVM	SDP	KSDP	SPWD	SVM	SDP	KSDP	SPWD
University of Pavia	20.71	19.31 (1.34)	18.01 (1.32)	19.94 (8.31)	27.22	26.01 (1.55)	23.19 (1.53)	24.82 (10.5)	34.02	31.65 (1.76)	29.61 (1.77)	31.16 (13.0)

TABLE XI

CLASSIFICATION ACCURACIES OBTAINED BY DIFFERENT SPECTRAL CLASSIFIERS AND SPECTRAL–SPATIAL CLASSIFICATION METHODS USING THE ORIGINAL TRAINING SET (MARKED AS NLA) AND THE IMPROVED TRAINING SET OBTAINED BY DP, SDP, KSDP, AND SPWD METHODS FOR THE UNIVERSITY OF PAVIA DATA SET (WITH 52 TRUE SAMPLES AND 8 NOISY LABELS PER CLASS). THE NUMBER IN THE PARENTHESIS REPRESENTS THE STANDARD VARIATION OF THE ACCURACIES OBTAINED IN REPEATED EXPERIMENTS

Training condition	Classifiers	OA(%)					AA(%)					κ				
		NLA	DP	SDP	KSDP	SPWD	NLA	DP	SDP	KSDP	SPWD	NLA	DP	SDP	KSDP	SPWD
52(true)+8(noise)	BTC	66.61 (1.91)	71.29 (2.39)	69.95 (1.29)	72.20 (2.19)	**73.00 (1.40)**	65.03 (1.29)	70.37 (2.66)	67.99 (1.41)	70.54 (1.77)	**71.28 (1.87)**	0.587 (1.98)	0.638 (2.55)	0.623 (0.01)	0.646 (0.02)	**0.656 (0.02)**
	KBTC	63.47 (2.02)	70.20 (1.27)	67.62 (1.95)	72.88 (1.77)	**73.99 (1.39)**	57.81 (1.56)	65.54 (2.45)	61.64 (2.08)	68.38 (2.23)	**69.11 (1.28)**	0.556 (0.02)	0.628 (0.02)	0.602 (0.02)	0.659 (0.02)	**0.659 (0.02)**
	SRC	55.72 (1.75)	60.41 (1.93)	58.71 (2.05)	61.22 (1.75)	**62.71 (1.63)**	68.28 (1.08)	72.75 (0.99)	71.36 (0.82)	71.66 (0.78)	**72.27 (0.87)**	0.466 (0.02)	0.517 (0.02)	0.497 (0.02)	0.523 (0.02)	**0.536 (0.02)**
	ELM	78.80 (1.46)	79.99 (2.25)	79.66 (1.60)	80.54 (1.74)	**81.94 (1.34)**	77.58 (0.86)	78.21 (1.87)	78.12 (1.28)	78.42 (0.80)	**79.35 (1.28)**	0.732 (0.02)	0.746 (0.03)	0.742 (0.02)	0.752 (0.02)	**0.768 (0.02)**
Training condition	Methods	OA(%)					AA(%)					κ				
		NLA	DP	SDP	KSDP	SPWD	NLA	DP	SDP	KSDP	SPWD	NLA	DP	SDP	KSDP	SPWD
52(true)+8(noise)	EMP	85.96 (2.26)	89.06 (2.59)	87.68 (2.12)	89.69 (2.09)	**91.40 (3.10)**	93.90 (0.86)	95.11 (1.09)	94.84 (0.74)	95.24 (0.59)	**95.62 (0.93)**	0.822 (0.03)	0.859 (0.03)	0.843 (0.03)	0.867 (0.03)	**0.888 (0.04)**
	LMLL	87.86 (3.73)	90.49 (2.07)	89.60 (2.00)	91.80 (2.75)	**92.28 (1.75)**	92.61 (1.27)	94.08 (0.97)	93.33 (0.91)	94.07 (0.93)	**94.55 (0.78)**	0.845 (0.05)	0.878 (0.03)	0.866 (0.02)	0.894 (0.03)	**0.899 (0.02)**
	IFRF	88.67 (1.52)	90.59 (4.41)	89.86 (2.45)	91.65 (2.73)	**92.43 (1.74)**	82.92 (2.29)	85.00 (5.06)	83.87 (3.39)	87.15 (3.00)	**88.05 (2.88)**	0.903 (0.05)	0.878 (0.05)	0.868 (0.03)	0.891 (0.03)	**0.909 (0.02)**
	EPF	89.23 (1.77)	92.60 (4.01)	90.13 (2.49)	93.27 (4.90)	**93.55 (4.47)**	87.88 (1.87)	91.31 (3.17)	88.26 (3.58)	92.25 (4.34)	**91.76 (4.25)**	0.861 (0.02)	0.903 (0.05)	0.872 (0.03)	0.911 (0.06)	**0.915 (0.06)**

methods trained using the original noisy training set and different improved training sets. Here, four classical spectral–spatial classification methods are adopted: the extended morphological profiles (EMPs) [53], the logistic regression and multilevel logistic (LMLL) [54], the image fusion recursive filtering (IFRF) [55], and the edge-preserving filtering (EPF) [56]. It can be observed from Table XI that the methods trained with the improved training sets achieve better classification accuracy when compared with the methods trained the original (noisy) training set. The performance improvements obtained by the methods trained with the training set provided by the proposed SPWD method are more noticeable.

V. CONCLUSION

In this article, a new algorithm for noisy label detection is presented. The proposed approach combines a new superpixel-to-pixel weighting distance and the DP clustering algorithm to detect and remove noisy labels from the training set, thus improving the performance of spectral and spatial–spectral HSI classifiers. A key aspect of our approach is that it can deal with two weak assumptions when exploiting the spectral–spatial information contained in the HSI: 1) all the pixels in a superpixel belong to the same class and 2) close pixels in spectral space have the same label. In order to overcome the first weak assumption, we use K nearest neighbors to obtain the closest neighborhoods of pixels around each superpixel, and a Gaussian weight is employed to mitigate the second weak assumption by adapting the original distance information. Our experimental results, conducted on several real HSI data sets, demonstrate that the method can effectively improve the performance of a variety of classifiers trained with noisy training sets.

In the future, we will do our utmost to prove whether SPWD can effectively address other related HSI applications. We will also explore other strategies to obtain the superpixel segmentation map, including techniques that not only use the

first few principal components to perform this task. Although we anticipate that the processing of RGB images by means of the proposed approach can be successful, there would be additional work to do in terms of RGB-based model reconstruction and parameter settings. Moreover, we will investigate new strategies that are able to correct mislabeled samples instead of removing them.

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] M. A. Lee, Y. Huang, H. Yao, S. J. Thomson, and L. M. Bruce, "Determining the effects of storage on cotton and soybean leaf samples for hyperspectral analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2562–2570, Jun. 2014.

[2] J. L. Boggs, T. D. Tsegaye, T. L. Coleman, K. C. Reddy, and A. Fahsi, "Relationship between hyperspectral reflectance, soil nitrate-nitrogen, cotton leaf chlorophyll, and cotton yield: A step toward precision agriculture," *J. Sustain. Agricult.*, vol. 22, no. 3, pp. 5–16, Jan. 2003.

[3] J. Pontius, M. Martin, L. Plourde, and R. Hallett, "Ash decline assessment in emerald ash borer-infested regions: A test of tree-level, hyperspectral technologies," *Remote Sens. Environ.*, vol. 112, no. 5, pp. 2665–2676, May 2008.

[4] B. Zhang, D. Wu, L. Zhang, Q. Jiao, and Q. Li, "Application of hyperspectral remote sensing for environment monitoring in mining areas," *Environ. Earth Sci.*, vol. 65, no. 3, pp. 649–658, Feb. 2012.

[5] M. Dalponte, H. O. Orka, T. Gobakken, D. Gianelle, and E. Naesset, "Tree species classification in boreal forests with hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2632–2645, May 2013.

[6] J. Chi and M. M. Crawford, "Spectral unmixing-based crop residue estimation using hyperspectral remote sensing data: A case study at Purdue University," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2531–2539, Jun. 2014.

[7] Y. Yuan, Q. Wang, and G. Zhu, "Fast hyperspectral anomaly detection via high-order 2-D crossing filter," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 620–630, Feb. 2015.

[8] 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.

[9] L. Fang, S. Li, X. Kang, and J. Benediktsson, "Spectral–spatial hyperspectral image classification via multiscale adaptive sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7738–7749, Dec. 2014.

[10] L. He, Y. Li, X. Li, and W. Wu, "Spectral–spatial classification of hyperspectral images via spatial translation-invariant wavelet-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2696–2712, May 2015.

[11] C. Li, Y. Ma, X. Mei, J. Ma, and C. Liu, "Hyperspectral image classification with robust sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 13, no. 5, pp. 641–645, Mar. 2016.

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

[13] J. Xia, P. Ghamisi, N. Yokoya, and A. Iwasaki, "Random forest ensembles and extended multiextinction profiles for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 202–216, Jan. 2018.

[14] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[15] J. Li, P. R. Marpu, A. Plaza, J. M. Bioucas-Dias, and J. A. Benediktsson, "Generalized composite kernel framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4816–4829, Sep. 2013.

[16] J. Wang, L. Jiao, H. Liu, S. Yang, and F. Liu, "Hyperspectral image classification by spatial–spectral derivative-aided kernel joint sparse representation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2485–2500, Jun. 2015.

[17] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018.

[18] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.

[19] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 486–500, Mar. 2017.

[20] Z. W. 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. Intell.*, vol. 39, no. 3, pp. 486–500, Mar. 2017.

[21] Y. Yao *et al.*, "Learning latent stable patterns for image understanding with weak and noisy labels," *IEEE Trans. Cybern.*, vol. 49, no. 12, pp. 4243–4252, Dec. 2019.

[22] V. Mnih and G. E. Hinton, "Learning to label aerial images from noisy data," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 567–574.

[23] G. M. Foody, "The effect of mis-labeled training data on the accuracy of supervised image classification by SVM," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 4987–4990.

[24] Y. Song, C. Wang, M. Zhang, H. Sun, and Q. Yang, "Spectral label refinement for noisy and missing text labels," in *Proc. AAAI*, Jan. 2015, pp. 2972–2978.

[25] B. Hou, Q. Wu, Z. Wen, and L. Jiao, "Robust semisupervised classification for PolSAR image with noisy labels," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6440–6455, Nov. 2017.

[26] 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.

[27] 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. 52, no. 2, pp. 851–865, Feb. 2019.

[28] 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.

[29] 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.

[30] Q. Leng, H. Yang, and J. Jiang, "Label noise cleansing with sparse graph for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 9, p. 1116, May 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] 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.

[33] B. Tu, X. Yang, N. Li, X. Ou, and W. He, "Hyperspectral image classification via superpixel correlation coefficient representation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4113–4127, Nov. 2018.

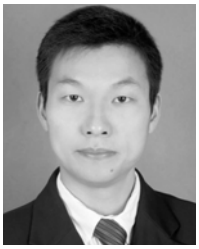
[34] B. Tu, S. Huang, L. Fang, G. Zhang, J. Wang, and B. Zheng, "Hyperspectral image classification via weighted joint nearest neighbor and sparse representation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4063–4075, Nov. 2018.

[35] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2281, Nov. 2012.

[36] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy-rate clustering: Cluster analysis via maximizing a submodular function subject to a matroid constraint," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 99–112, Jan. 2014.

[37] G. Nemhauser, L. Wolsey, and M. Fisher, "An analysis of approximations for maximizing submodular set functions," *Math. Program.*, vol. 14, no. 1, pp. 265–294, Dec. 1978.

[38] G. N. Yesilyurt, A. Ertürk, and S. Ertürk, "Metric and transform performance analysis for hyperspectral superpixel segmentation," in *Proc. 25th Signal Process. Commun. Appl. Conf. (SIU)*, May 2017, pp. 1–4.
[39] J.-K. Lee, I.-S. Jeon, and C.-H. Lee, "Command-shaping guidance law based on a Gaussian weighting function," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 50, no. 1, pp. 772–777, Jan. 2014.
[40] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.
[41] K. Sun, X. Geng, and L. Ji, "Exemplar component analysis: A fast band selection method for hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 998–1002, May 2015.
[42] 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.
[43] X. Luo, R. Xue, and J. Yin, "Information-assisted density peak index for hyperspectral band selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1870–1874, Oct. 2017.
[44] 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, 2011.
[45] M. Cui and S. Prasad, "Class-dependent sparse representation classifier for robust hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2683–2695, May 2015.
[46] 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.
[47] 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.
[48] N. Audebert, B. L. Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, Jun. 2019.
[49] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification via kernel sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 217–231, Jan. 2013.
[50] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.
[51] M. A. Toksöz and İ. Ulusoy, "Hyperspectral image classification via basic thresholding classifier," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 4039–4051, Jul. 2016.
[52] M. A. Toksöz 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.
[53] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.
[54] 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.
[55] X. Kang, S. Li, and J. A. Benediktsson, "Feature extraction of hyperspectral images with image fusion and recursive filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 6, pp. 3742–3752, Jun. 2014.
[56] 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.



Bing Tu (Member, IEEE) received the M.S. degree from the Guilin University of Technology, Guilin, China, in 2009, and the Ph.D. degree 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 was 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 (Student Member, IEEE) received the B.S. degree from the Hunan Institute of Science and Technology, Yueyang, China, in 2019, where he is currently pursuing the M.S. degree.

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

Danbing He (Student Member, IEEE) is currently pursuing the B.S. degree with the School of Information Science and Technology, Hunan Institute of Science and Technology, Yueyang, China.

Her research interest includes hyperspectral image analysis.

Siyuan Huang (Student Member, IEEE) received the B.S. degree from the Hunan Institute of Science and Technology, Yueyang, China, in 2016, where he is currently pursuing the M.S. degree.

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

Antonio Plaza (Fellow, IEEE) received the M.Sc. and Ph.D. degrees in computer engineering from the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, University of Extremadura, Cáceres, Spain, in 1999 and 2002, respectively.

He is currently the Head of the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, University of Extremadura. He has authored more than 600 publications, including over 200 JCR journal articles (over 160 in IEEE journals), 23 book chapters, and around 300 peer-reviewed conference proceeding papers. His research interests include hyperspectral data processing and parallel computing of remote sensing data.

Dr. Plaza was a member of the Editorial Board of the *IEEE Geoscience and Remote Sensing Newsletter* from 2011 to 2012 and the *IEEE Geoscience and Remote Sensing Magazine* in 2013. He was also a member of the Steering Committee of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS). He is a fellow of the IEEE for contributions to hyperspectral data processing and parallel computing of earth observation data. He was a recipient of the recognition of Best Reviewers of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2009 and the Best Reviewer of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 2010, for which he has served as an Associate Editor from 2007 to 2012. He was a recipient of the Best Column Award of the *IEEE Signal Processing Magazine* in 2015, the 2013 Best Paper Award of JSTARS, and the Most Highly Cited Paper (2005–2010) in the *Journal of Parallel and Distributed Computing*. He received the Best Paper Awards at the IEEE International Conference on Space Technology and the IEEE Symposium on Signal Processing and Information Technology. He has served as the Director of Education Activities for the IEEE Geoscience and Remote Sensing Society (GRSS) from 2011 to 2012 and as the President of the Spanish Chapter of the IEEE GRSS from 2012 to 2016. He has reviewed more than 500 manuscripts for over 50 different journals. He has also served as the Editor-in-Chief of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING from 2013 to 2017. He has Guest Edited ten special issues on hyperspectral remote sensing for different journals. He is also an Associate Editor of IEEE ACCESS (receiving the recognition as an Outstanding Associate Editor of the journal in 2017). Additional information: http://www.umbc.edu/rssipl/people/aplaza