# Jointly Learning the Hybrid CRF and MLR Model for Simultaneous Denoising and Classification of Hyperspectral Imagery

Ping Zhong, *Member, IEEE*, and Runsheng Wang

***Abstract*— Despite much advance obtained in hyperspectral image sensors, they are still very sensitive to the noise, and thus cause the captured data to carry enough noise to degrade the classification results. The traditional approach first resorts to image denoising and then feeds the denoised image into a classifier. However, such a straightforward approach, treating denoising and classification separately, suffers greatly from neglecting their impacts on each other. This paper presents a new simultaneous denoising and classification method in the pursuit of cleanest image for optimal classification in the sense of given task evaluation measures. To obtain this objective, we develop a hybrid conditional random field (CRF) (for denoising) and multinomial logistic regression (MLR) (for classification) model at first, and then to train the proposed hybrid model, we propose a new joint learning method, which can effectively capture the impacts of denoising on classification, or vice versa, the effects of classification on denoising. Through the proposed joint learning method, the CRF and MLR, and thus the denoising and classification procedure, can be tightly combined. Moreover, the proposed joint learning method can directly optimize a large class of application specific performance measures including both the linear measures, such as the overall accuracy, and the nonlinear measures, such as kappa statistics. Meanwhile, the consistency between the criteria of model learning and model application has the potential to obtain the denoised image, which is at its best for optimal classification in the sense of the given measure. The extensive experiments of simultaneous denoising and classification tasks are conducted in both simulated and real noisy conditions to test our jointly learned model, which are shown to outperform the conventional methods of treating the two tasks independently.**

***Index Terms*— Classification, conditional random field (CRF), denoising, hyperspectral imagery, model learning, multinomial logistic regression (MLR).**

## I. Introduction

OVER the past decades, hyperspectral image sensors have experienced a significant success, and the hyperspectral image analysis is attracting a growing interest in real-world applications, such as in urban planning, mapping, agriculture, forestry, and disaster prevention and monitoring [1]–[5]. Many of these applications can be finally transformed into some classification tasks [6], which are equivalent to labeling imaged terrains or regions. However, despite the advance in hyperspectral image sensors, they are still very sensitive to the noise due to their nonlinear response in different spectral bands, and thus cause the captured data carry enough noise to degrade the classification results [7], [8].

A natural solution to this problem would first perform image denoising to obtain a hyperspectral image with better quality, and then feed the denoised result into a classification system. Much progress has been made on pure hyperspectral image denoising and classification, respectively. For the image denoising, the popular methods include principal component analysis (PCA) [9], [10], wavelet [11], [12], multidimensional analysis [13], [14], total variation [15], [16], Gaussian conditional random field (GCRF) [17], [18], neural networks [19], and so on. These methods have their own advantage and disadvantage [18]. Relatively, the newly defined GCRF models have the potentials to gain popularity since they can capture the short- and long-range constraints and statistics in both the noisy input and the latent clean image. The captured rich information could help the GCRF models to avoid the oversmoothing, ringing artifacts, and other problems usually appeared in the denoised results [17], [18].

For the hyperspectral image classification, many recent popular methods include support vector machines (SVMs) [20]–[22], multinomial logistic regression (MLR) [2], [23], [24], neural networks [25], [26], graph method [2], [27], [28], AdaBoost [29], Gaussian process approach [30], and so on. Compared with other methods, MLR model has its own advantages derived from several aspects. First, MLR model is relatively simple, and its simple closed formulation presents the feasibilities of further strict mathematical analysis. Second, the probabilistic nature of the MLR model also offers many practical advantages [31], such as the ability to set rejection thresholds, accommodate unequal relative class frequencies in the training set and in operation, and apply an appropriate loss matrix in making predictions that minimize the expected risk. Finally, as a discriminative classifier, the effectiveness of MLR can be obtained mainly from its generality: contrary to the conventional linear discriminant analysis [32], [33], MLR requires a fewer restrictive assumptions. In this case, the features need not be normally distributed and linearly related to the class.

Although much progress has been separately made on pure hyperspectral image denoising and hyperspectral image classification, only a few works have investigated their impacts

Manuscript received May 21, 2013; accepted November 23, 2013. Date of publication January 2, 2014; date of current version June 10, 2014. This work was supported in part by the Natural Science Foundation of China under Grant 61271439, in part by the Foundation for the Author of National Excellent Doctoral Dissertation of China under Grant 201243, and in part by the Program for New Century Excellent Talents in University under Grant NECT-13-xxxx.

The authors are with the ATR National Key Laboratory, School of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, China (e-mail: zhongping@nudt.edu.cn; rswang@nudt.edu.cn).

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

Digital Object Identifier 10.1109/TNNLS.2013.2293061

on each other [34]. There are close connections between the denoising and later high-level vision tasks, such as classification [34], [35]. Previous straightforward approach breaks down the close connections, and thus could have the problem that many denoising algorithms are designed for improving human visual perception only, rather than improving the classification results. Thus, there is no guarantee of the designed system to obtain the classification results as good as possible.

To deal with these problems, and meanwhile considering the previously mentioned advantages of GCRF (for denoising) and MLR (for classification), this paper aims to presents a new simultaneous denoising and classification method through a proposed hybrid GCRF and MLR model. The difficulty lies in the procedure to closely combine the denoising (by GCRF) and classification (by MLR). Corresponding to the hybrid GCRF and MLR model, we propose to combine the denoising and classification through a joint learning method, which can effectively capture the impacts of denoising on classification, or vice versa, the effects of classification on denoising. Its main objective is to pursue the cleanest image for optimal classification in the sense of given task evaluation measures.

However, the usual learning methods are performed to estimate the parameters, optimizing some predefined learning criteria, but not the task evaluation measures. For example, the usual learning criteria, such as maximum likelihood criterion, maximum *a posteriori* (MAP) criterion, large margin criterion, constrained empirical risk minimization framework [36], and so on [37], usually optimize the linear combination of average accuracies (AAs) or error rates, however, not the application specific performance measures, such as the kappa statistics usually used in the hyperspectral image classification. This inconsistency between the learning criterion and the task evaluation measure might produce a suboptimal result.

To overcome this inconsistency, this paper further develops a new joint learning method for the proposed hybrid GCRF and MLR model. The learning procedure can directly optimize the application specific performance measures. In the literature, some learning methods exist for directly optimizing linear application specific measures. However, for most of the application specific performance measures, one difficulty common to them is their nonlinear and multivariate nature. This results in decision theoretic risks that no longer decompose into expectations over individual examples [38]. To accommodate this problem, a few previous attempts have been made to the direct optimization of F-score for the logistic regression [39], SVM [38], [40], and conditional random field (CRF) [37]. In the same spirit as these studies and under the multivariate prediction framework [37], this paper proposes an approach that is fundamentally different from most of the conventional learning algorithms: instead of learning a univariate rule that predicts the label of a single example, we formulate the learning problem as a multivariate prediction of all examples in the training data set. With this proposed approach, GCRF and MLR in the hybrid model can be jointly learned to denoise the hyperspectral image for optimal classification in the sense of large classes of both potentially linear and nonlinear performance measures, in particular, the overall accuracy (OA) (linear) and kappa (nonlinear) usually used in hyperspectral image classification. Although it is designed for the hybrid GCRF and MLR model corresponding to the hyperspectral image analysis at hand, such an approach can be fruitfully extended to many other hybrid systems.

To sum up, the novelties of this paper derive from three aspects. First, the denoising and classification can be simultaneously implemented and thus the close connections between them can be of great benefit to each other. Second, the proposed new joint learning method can tightly combine the GCRF and MLR model, and thus the denoising and classification procedure, to make the given application specific performance measures to be optimal. Finally, because of the multivariate prediction framework, a novelty can be recognized as the flexibility of the proposed joint learning method to optimize both the linear and nonlinear performance measures.

The rest of this paper is arranged as follows. The hybrid GCRF and MLR model for simultaneous denoising and classification of hyperspectral images is proposed in Section II. Section III presents the joint learning algorithm of hybrid GCRF and MLR model with linear performance measures. With the theoretical background of Section III, Section IV proposes the approach for jointly learning GCRF and MLR model to optimize the nonlinear performance measures. Section V uses both the synthetic and real-world hyperspectral imagery to evaluate the proposed algorithms. Finally, our techniques are concluded in Section VI.

## II. Hybrid Framework for Simultaneous Denoising and Classification

In the context of hyperspectral image analysis, the observed data from an input hyperspectral image $\mathbf{y}$ are a 3-D data cube, which can be spatially denoted as a set of spectral vectors $\{\mathbf{y}_{1,:}, \mathbf{y}_{2,:}, \ldots, \mathbf{y}_{I,:}\}$ or spectrally denoted as a set of band images $\{\mathbf{y}_{:,1}, \mathbf{y}_{:,2}, \ldots, \mathbf{y}_{:,b}\}$, where $\mathbf{y}_{i,:} = [y_{i1}, y_{i2}, \ldots, y_{ib}]^T$ is a spectral vector associated with an image site $i \in S$ and $\mathbf{y}_{:,j} = [y_{1,j}, y_{2,j}, \ldots, y_{I,j}]^T$ is an image associated with the band $j \in \{1, 2, \ldots, b\}$. $S = \{1, 2, \ldots, I\}$ is the set of image sites and $b$ is the band number. For the given hyperspectral image $\mathbf{y}$, this paper focuses mainly on simultaneously obtaining the latent clean image $\mathbf{z} = \{\mathbf{z}_{1,:}, \mathbf{z}_{2,:}, \ldots, \mathbf{z}_{I,:}\}$ or $\mathbf{z} = \{\mathbf{z}_{:,1}, \mathbf{z}_{:,2}, \ldots, \mathbf{z}_{:,b}\}$ and the classification image $\mathbf{x} = \{x_1, x_2, \ldots, x_I\}$, which means the (class) labels assigned to image sites. Each $x_i, i = 1, \ldots, I$ in $\mathbf{x}$ takes value in the set $\chi = \{1, 2, \ldots, \mathrm{M}\}$, where M is the number of classes.

### A. GCRF Model for Hyperspectral Image Denoising

Real-world hyperspectral data cube usually contains simultaneously a significant number of noisy bands and many high-SNR spectral bands. The CRF model can be applied well to reduce the noise in the noisy bands because of its ability to capture both the spatial and spectral contextual information, which have been proved to be very useful to reduce the noise [18]. Thus, our previous work proposed a kind of GCRF named multiple-spectral-band CRF to recover the noisy band images using their neighboring high-SNR spectral bands.

The input hyperspectral data cube is $\mathbf{y} = \{\mathbf{y}_{:,1}, \mathbf{y}_{:,2}, \ldots, \mathbf{y}_{:,b}\}$. The sets of noisy junk bands and neighboring high-SNR

good bands are denoted as $\mathbf{y}^J = \{\mathbf{y}^J_{:,1}, \mathbf{y}^J_{:,2}, \ldots, \mathbf{y}^J_{:,m}\}$ and $\mathbf{y}^H = \{\mathbf{y}^H_{:,1}, \mathbf{y}^H_{:,2}, \ldots, \mathbf{y}^H_{:,n}\}$, respectively, where $m + n = b$. The set of latent clean band images corresponding to $\mathbf{y}^J$ is denoted as $\mathbf{z}^J = \{\mathbf{z}^J_{:,1}, \mathbf{z}^J_{:,2}, \ldots, \mathbf{z}^J_{:,m}\}$. Each latent clean band image $\mathbf{z}^J_{:,k}$ can be inserted into the set of good bands $\mathbf{y}^H$ with the same band order that they follow in the original hyperspectral data cube. This forms the latent clean data cube $\mathbf{z} = \{\mathbf{z}_{:,1}, \mathbf{z}_{:,2}, \ldots, \mathbf{z}_{:,b}\}$ of the original input data cube $\mathbf{y}$.

Thus, the main task is to recover the $\mathbf{z}^J$ from the given noisy data cube $\mathbf{y}$. From the view of hyperspectral image denoising, the GCRF framework considers the Markov property of latent clean band image $\mathbf{z}^J_{:,k}, k = 1, 2, \ldots, m$ conditioned on the input noisy hyperspectral image $\mathbf{y} = \{\mathbf{y}^J, \mathbf{y}^H\}$ and directly models the posterior as a Gibbs distribution [18]

$$p\left(\mathbf{z}^J_{:,k}|\mathbf{y}\right) = p\left(\mathbf{z}^J_{:,k}|\mathbf{y}^J, \mathbf{y}^H\right) = \frac{1}{Z\left(\Theta^k\right)} \cdot \exp\left\{-\sum_{q=1}^{N_f}\sum_{i,j} v_q\left(i, j, \tilde{k}; \mathbf{y}^J, \Theta^k\right) \left[\left(f_q * \mathbf{z}\right)\left(i, j, \tilde{k}\right) - r_q\left(i, j, \tilde{k}; \mathbf{y}\right)\right]^2\right\} \quad (1)$$

where $\tilde{k}$ is the index of $\mathbf{z}^J_{:,k}$ in the ordered latent clean band set $\mathbf{z}$, $\mathbf{f} = \{f_1, f_2, \ldots, f_{N_f}\}$ is the multiple-spectral-band filter bank containing 11 ($N_f = 11$) 3-D derivative filters, where each filter $f_q = f_q^p f_q^s$ composed of a 2-D spatial filter $f_q^s$ for each spectral band and a 1-D spectral filter $f_q^p$ across spectral bands [18]. $\left(f_q * \mathbf{z}\right)\left(i, j, \tilde{k}\right)$ denotes the value of convolving $\mathbf{z}$ with $f_q$ at location $\left(i, j, \tilde{k}\right)$. The function $r_q\left(i, j, \tilde{k}; \mathbf{y}\right)$ is used to estimate the value of $\left(f_q * \mathbf{z}\right)\left(i, j, \tilde{k}\right)$. For each filter $f_q$, the function $r_q\left(i, j, \tilde{k}; \mathbf{y}\right)$ uses the observed noisy hyperspectral image $\mathbf{y}$ to estimate the value of the filter response at each location. Using the same setting in [17], we set the first filter $f_1$ as an identity filter with $r_1(x, y, k; \mathbf{y}) = \mathbf{y}^J_k$ and the other estimated filter responses $r_i(.), i = 2, \ldots, N_f$ as zero to keep the linear system invertible.

The weights $\left\{v_q\left(i, j, \tilde{k}; \mathbf{y}^J, \Theta^k\right); q, i, j, \tilde{k}\right\}$ are added to the model to improve the model's ability to handle edges or textures, and can be computed as

$$v_q\left(i, j, \tilde{k}; \mathbf{y}^J, \Theta^k\right) = \exp\left(\sum_{l=1}^{72} \theta^k_{q,l} a_l\left(i, j, \tilde{k}\right)\right) \quad (2)$$

where $a_l\left(i, j, \tilde{k}\right)$ is the absolute response of $\mathbf{y}^J_{:,\tilde{k}}$ to the total 72 multiscale oriented edge and bar filters developed in [17], and $\theta^k_{q,l}$ represents the weight associated with response $a_l\left(i, j, \tilde{k}\right)$ for $l$th edge and bar filter. Thus, in the developed GCRF framework (1), both the filters in computation of $\left\{v_q\left(i, j, \tilde{k}; \mathbf{y}^J, \Theta^k\right); q, i, j, \tilde{k}\right\}$ and the filter bank $\mathbf{f}$ are fixed, and only the parameters $\Theta = \left\{\Theta^k, k = 1, 2, \ldots, m\right\} = \left\{\theta^k_{q,l}, q = 1, 2, \ldots, 11, l = 1, 2, \ldots, 72, k = 1, 2, \ldots, m\right\}$ need to be estimated. More details can be found in [18].

### B. MLR for Hyperspectral Image Classification

As analyzed in Section I, compared with other methods, MLR model has its own advantages derived from several aspects, such as simple closed formulation, ability to set rejection thresholds, generality, and so on. Thus, this paper selects MLR for the hyperspectral image classification. Under a MLR model, the probability that a spectral vector $\mathbf{z}_{i,:}$ belongs to class $c$ is written as

$$P\left(x_i = c|\mathbf{z}_{i,:}, \mathbf{w}\right) = \begin{cases} \frac{\exp\left(\mathbf{w}_c^T \mathbf{z}_{i,:}\right)}{1 + \sum_{l=1}^{\mathrm{M}-1} \exp\left(\mathbf{w}_l^T \mathbf{z}_{i,:}\right)} & \text{if } c < \mathrm{M} \\ \frac{1}{1 + \sum_{l=1}^{\mathrm{M}-1} \exp\left(\mathbf{w}_l^T \mathbf{z}_{i,:}\right)} & \text{if } c = \mathrm{M} \end{cases} \quad (3)$$

where $\mathbf{w}_c$ is the parameter vector $[w_{c1}, \ldots, w_{cb}]^T$ for $c$th class, and $\mathbf{w}$ denotes a $(b(\mathrm{M}-1))$-dimensional vector produced from concatenating the vectors $\{\mathbf{w}_c, c = 1, \ldots, \mathrm{M}-1\}$. For binary problems ($M = 2$), this is known as a logistic regression model; for $M > 2$, the usual designation is MLR (or soft-max in the neural networks literature).

### C. Formulation of the Hybrid Framework

This paper proposes a hybrid GCRF (for denoising) and MLR (for classification) model for simultaneous denoising and classification. The overview of the proposed framework is shown in Fig. 1, including two important steps: 1) learning and 2) inference. The learning procedure estimates the model parameters over the training samples, while the inference procedure infers the denoising and classification results using the learned model. The inference procedure is relatively simple: if the model parameters have been learned, the hybrid model infers the denoised result with GCRF model at first and then infers the classification map with MLR model on the whole image in a straightforward manner without alternating between the evaluation of denoised image and label inference. Thus, we focus mainly on the learning procedure, where both the parameters $\Theta$ in GCRF and $\mathbf{w}$ in MLR model need to be learned. The procedure can be generally implemented through two steps, i.e., training sample selection and joint model learning.

The training samples used to estimate $\mathbf{w}$ and $\Theta$ are selected at first. To estimate $\mathbf{w}$ in the MLR classification model, the usual procedure of image labeling task is learning classifier (i.e., $\mathbf{w}$) over the training samples at first, and then generalizing the learned classifier to completely unseen images (of the same category). However, for the hyperspectral image classification at hand, a practical work flow to get the final product is to select some samples from a given image for classifier learning, and then the learned classifier is used to classify all the pixels of the same given image [1]–[6]. Following the practical work flow, the training samples and the remaining test samples are from the same scene taken by the same sensor at the same imaging condition. This procedure could relieve the negative effects from the uncertainties from scenes, sensors, and imaging condition on the learning system, and also could effectively avoid the risk of overtraining. Therefore, considering these advantages and the practicability of our proposed method, we also adapt this practical work flow, and manually or randomly select and label several local patches as the training samples to estimate $\mathbf{w}$ in MLR.

For parameters $\Theta$ in GCRF model, the training samples are the noisy bands and their neighboring high-SNR bands [18],

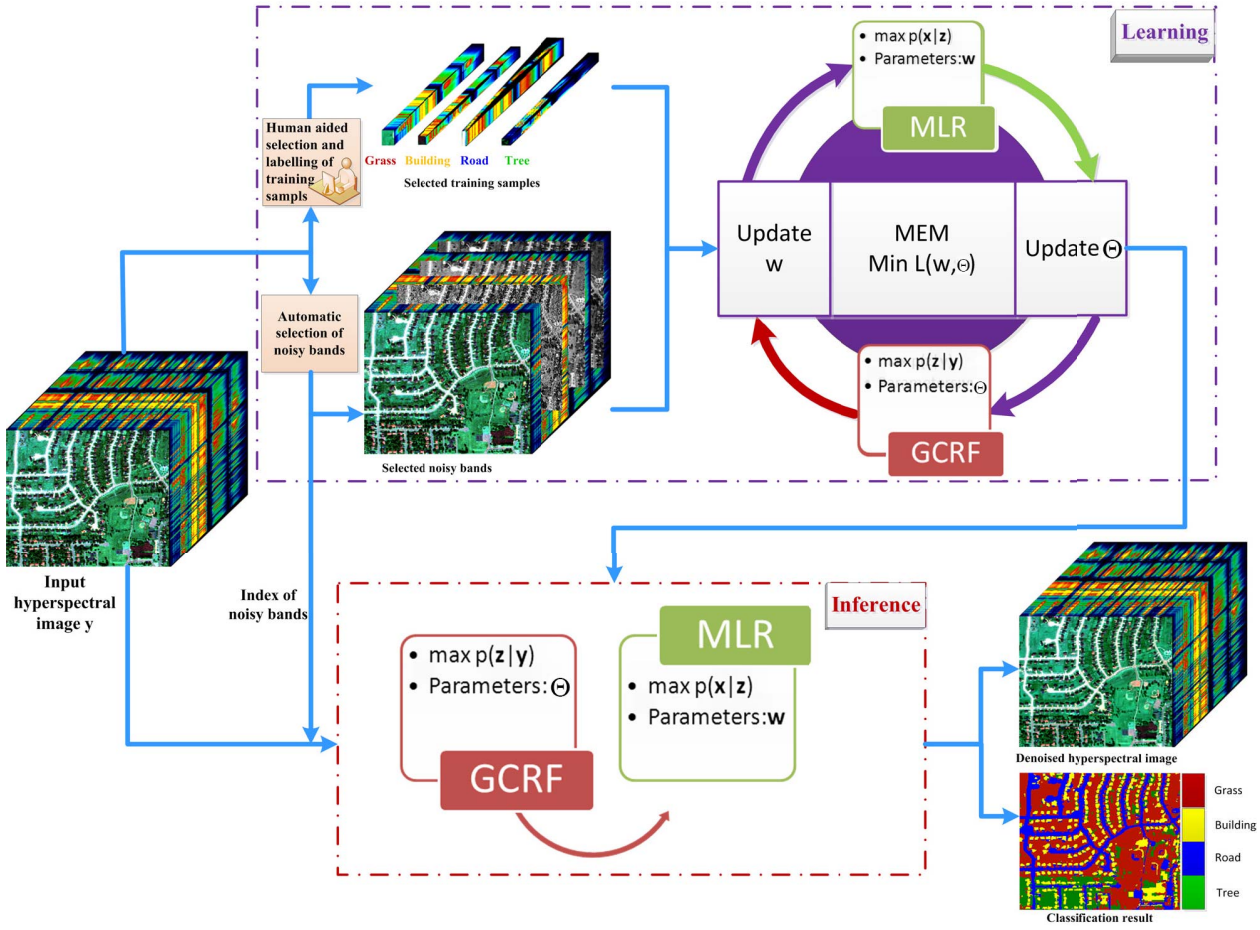


Fig. 1. Overview of the proposed hybrid framework learned with multivariate evaluation measures for simultaneous denoising and classification.

which can be automatically selected. We use the available wavelet-based method to sequentially estimate the noise variance of every band and set the noisy bands as that with noise variances higher than the average of all the variances [41]. Although this noise estimation method can obtain satisfactory results, we believe that more recently proposed noise estimation method could be used to further improve the estimation performance [43]. However, for the page limitation, we leave this topic for future work and focus on the formulation of learning algorithms. Then, corresponding to the graph structure of GCRF, both the noisy bands and their neighboring high-SNR bands are extracted and combined as the training samples to estimate $\Theta$ in GCRF.

After the selection of training samples, we propose a new joint learning method of the hybrid GCRF and MLR model to estimate the model parameters. The joint learning method is developed under the multivariate prediction framework, and could potentially optimize a wide range of multivariate evaluation measures (MEMs), including both the linear and nonlinear measures such as OA and kappa, respectively, in the hyperspectral image classification. Through the bridge provided by the task specific MEM, the GCRF and MLR are tightly combined in a unified framework by alternately updating the model parameters $\mathbf{w}$ (in MLR) and $\Theta$ (in GCRF). The tight connection between GCRF and MLR makes the hyperspectral image denoising and classification task can benefit greatly from each other. On the one hand, a better denoised image can be more accurate to represent the land cover classes, leading to a more discriminability between the samples of different classes, thus facilitating classification; on the other hand, a better classification result can give the denoising procedure more accurate class information rather than only the observed spectral information, leading to an extra semantic denoising, and thus may improve the denoising results.

In the remaining contents of this section, we will present the details about inference method related to the proposed hybrid framework. The joint learning methods with MEM will be presented in Sections III and IV.

### *D. Inference of the Hybrid Framework*

Let us now assume that we have learned the optimal GCRF model parameters $\Theta$ and the MLR model parameters $\mathbf{w}$. The proposed hybrid model does not involve complex computation in update $\Theta$ and $\mathbf{w}$, and makes it feasible to infer the denoised image and classification map in a straightforward manner without alternating between the evaluation of parameters and label inference. The flowchart of the inference of the hybrid framework is shown in Fig. 1. For the input

hyperspectral image $\mathbf{y}$, we compute its denoised image $\hat{\mathbf{z}}$ and classification map $\hat{\mathbf{x}}$ as following two steps.

*Infer the denoised images* $\hat{\mathbf{z}}(\mathbf{y}, \Theta)$ *by GCRF model:* When applying GCRF with the model parameters $\Theta^k$ and the input observation $\mathbf{y}$, the inferred images $\hat{\mathbf{z}}^J_{:,k}\left(\mathbf{y}, \Theta^k\right)$ corresponding to the automatically selected noisy bands $\mathbf{y}^J_{:,k}$ is derived by MAP estimation. The procedure to find the MAP estimation is equivalent to minimizing an energy function [18]

$$\hat{\mathbf{z}}^J_{:,k}\left(\mathbf{y}, \Theta^k\right) = \arg\min_{\mathbf{z}^J_{:,k}} \left\{\mathbf{V}\left(\mathbf{F}\mathbf{Z}_k - \mathbf{r}\right)^T \left(\mathbf{F}\mathbf{Z}_k - \mathbf{r}\right)\right\} \tag{4}$$

where for the convenience of computation, the GCRF in (1) is rewritten through matrix notation. $\mathbf{F}$ is a matrix by stacking the set of convolution matrices $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_{N_f}$ corresponding to the multiple-spectral-band filters $f_1, f_2, \ldots, f_{N_f}$. The diagonal matrix $\mathbf{V}$ is a block diagonal matrix constructed from the diagonal submatrices $\mathbf{V}_1\left(\mathbf{y}^J, \Theta^k\right)$, $\mathbf{V}_2\left(\mathbf{y}^J, \Theta^k\right), \ldots, \mathbf{V}_{N_f}\left(\mathbf{y}^J, \Theta^k\right)$ and each entry along the diagonal in the matrices $\mathbf{V}_q\left(\mathbf{y}^J, \Theta^k\right)$ is equal to the weight of a term at a particular pixel, i.e., $v_q\left(i, j, k; \mathbf{y}^J, \Theta^k\right)$ in (2). $\mathbf{Z}_k$ is the latent clean image vector. $\mathbf{r}$ is the vector by stacking the $r_i(.), i = 2, \ldots, N_f$. Because the objective function is quadratic, the minimum can be finally computed using pseudoinverse as

$$\hat{\mathbf{z}}^J_{:,k}\left(\mathbf{y}, \Theta^k\right) = \left(\mathbf{F}^T\ \mathbf{V}\mathbf{F}\right)^{-1}\mathbf{F}^T\ \mathbf{V}\mathbf{r}. \tag{5}$$

Putting all the $\hat{\mathbf{z}}^J_{:,k}, k = 1, 2, \ldots, m$ and the original high-SNR bands $\mathbf{y}^H$ together will obtain the denoised hyperspectral image $\hat{\mathbf{z}}(\mathbf{y}, \Theta)$.

*Infer the classification map* $\hat{\mathbf{x}}$ *by MLR model:* Using the denoised hyperspectral image $\hat{\mathbf{z}}(\mathbf{y}, \Theta)$ as the input of MLR model, the inferred class $\hat{x}_i\left(\hat{\mathbf{z}}, \mathbf{w}\right)$ for site $i$ is derived by maximization process

$$\hat{x}_i\left(\hat{\mathbf{z}}, \mathbf{w}\right) = \arg\max_{x_i} \left\{P\left(x_i | \hat{\mathbf{z}}_{i,:}(\mathbf{y}, \Theta), \mathbf{w}\right)\right\} \tag{6}$$

where the posteriori probability is defined as (3).

## III. Joint Learning of Hybrid GCRF and MLR Model With Linear Performance Measures

As mentioned in Section I, the model should be learned under the criteria, which are consistent with the application specific performance measures used in the test procedure. Thus, corresponding to the task at hand, we aim to jointly learn the GCRF and MLR model parameters to maximize a specific measure for the classification result evaluation. The usual and effective measures to evaluate the performance of hyperspectral image classification include the linear measures, such as OAs and AAs, and the nonlinear measures, such as the kappa statistics (kappa). Different measures focus on evaluating different aspect of the algorithms' performances, for example, AAs and kappa can measure the balance between the classification accuracies of different classes, while OAs tend to evaluate the accuracies over all samples, with no distinguishing classes. Thus, the selection of the measures should be task specific. This motivates that the proposed learning framework should optimize a wide range of evaluation measures. In this section, we will develop the joint learning method of the hybrid GCRF and MLR model to maximize the linear measure, i.e., OAs, and in the following section, we will further develop the joint learning method corresponding to the more complex nonlinear measure, i.e., kappa statistics.

### A. Problem Formalization

Let $\mathbf{y}^{tr} = \left\{\mathbf{y}^{tr}_{1,:}, \mathbf{y}^{tr}_{2,:}, \ldots, \mathbf{y}^{tr}_{N,:}\right\}$ be a collection of selected training spectral vector samples, $\mathbf{x}^{tr} = \left\{x^{tr}_1, x^{tr}_2, \ldots, x^{tr}_N\right\}$ be the corresponding class labels. We aim to jointly learn the hybrid GCRF and MLR model parameters to maximize OAs or equally to minimize classification error (MCE). Thus, inspired by [37], we first formulate the objective function of the MCE criterion.

The Bayes decision rule decides the most probable output $\hat{x}^{tr}_i$ for $\mathbf{y}^{tr}_{i,:}$ through maximization process

$$\hat{x}^{tr}_i = \arg\max_{x_i} \left\{P\left(x_i | \hat{\mathbf{z}}^{tr}_{i,:}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right)\right\}$$

where $\hat{\mathbf{z}}^{tr}_{i,:}\left(\mathbf{y}^{tr}, \Theta\right)$ is the latent clean spectral vector of $\mathbf{y}^{tr}_{i,:}$ and $P\left(x_i | \hat{\mathbf{z}}^{tr}_{i,:}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right)$ is defined as the MLR [see (3)] to label the spectral vector $\hat{\mathbf{z}}^{tr}_{i,:}\left(\mathbf{y}^{tr}, \Theta\right)$. In general, $P\left(x_i | \hat{\mathbf{z}}^{tr}_{i,:}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right)$ can be replaced by a more general discriminant function $g(\cdot)$, that is

$$\hat{x}^{tr}_i = \arg\max_{x_i} \left\{g\left(x_i | \hat{\mathbf{z}}^{tr}_{i,:}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right)\right\}.$$

Thus, a misclassification measure $d(\cdot)$ can be defined as

$$d\left(x^{tr}_i, \hat{\mathbf{z}}^{tr}_{i,:}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right) = -g\left(x^{tr}_i | \hat{\mathbf{z}}^{tr}_{i,:}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right) + \max_{x_i \in \chi \setminus x^{tr}_i} g\left(x_i | \hat{\mathbf{z}}^{tr}_{i,:}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right). \tag{7}$$

Since $x^{tr}_i$ is the correct output, $d(\cdot) > 0$ means misclassification. Then, the MCE can be written as the minimization of the sum of 0–1 losses of the given training data: $\lambda = \arg\min_\lambda L_\lambda$, where

$$L_\lambda = \frac{1}{N}\sum_{i=1}^{N} \delta\left(d\left(x^{tr}_i, \hat{\mathbf{z}}^{tr}_{i,:}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right)\right) \tag{8}$$

$\lambda = \{\Theta, \mathbf{w}\}$, $\delta(a)$ is a step function returning 0 if $a < 0$ and 1 otherwise, and thus $\sum_i \delta\left(d\left(x^{tr}_i, \hat{\mathbf{z}}^{tr}_{i,:}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right)\right)$ equals to the number of the misclassified training samples.

The above-stated optimization criterion is not easy to handle since the discontinuity introduced by the operator max in (7) and the step function $\delta$ in (8). To solve the computational problem from operator max, a continuous approximate is the soft-max, i.e., $\max_n a_n \approx \log\sum_n \exp(a_n)$. Thus, the misclassification measure $d(\cdot)$ in (7) can be rewritten as

$$d\left(x^{tr}_i, \hat{\mathbf{z}}^{tr}_{i,:}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right) = -g\left(x^{tr}_i | \hat{\mathbf{z}}^{tr}_{i,:}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right) + \log\left(\frac{1}{M-1}\sum_{x_i \in \chi \setminus x^{tr}_i} \exp\left(\psi g\left(x_i | \hat{\mathbf{z}}^{tr}_{i,:}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right)\right)\right) \tag{9}$$

where $\psi$ is a positive constant that represents $L_\psi$-norm. To solve the computational problem from step function $\delta$, the typical continuous and differentiable approximate is the sigmoid function $\tau^{\text{sig}}(\cdot)$

$$\delta(d(\cdot)) \approx \tau^{\text{sig}}(d(\cdot)) = \left(1 + \exp\left(-\alpha d(\cdot)\right) - \beta\right)^{-1}. \tag{10}$$

Then, the regularized objective function of the MCE in (8) can be finally written as

$$L_{d,g,\lambda}^{\mathrm{MCE}} = \frac{1}{N}\sum_{i=1}^{N} \tau^{\mathrm{sig}}\left(d\left(x_i^{tr}, \hat{\mathbf{z}}_{i,:}^{tr}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right)\right) + R_\lambda \quad (11)$$

where the regularization term $R_\lambda = \frac{\gamma_1}{\phi}\|\mathbf{w}\|^\phi + \frac{\gamma_2}{\phi}\|\Theta\|^\phi$ and $\gamma_1$ and $\gamma_2$ control the weights of term $\|\mathbf{w}\|^\phi/\phi$ and $\|\Theta\|^\phi/\phi$, respectively. The regularized objective function is equivalent to that of MAP criterion with the parameter prior distribution as the $L_\phi$-norm, which can make the estimated parameters smooth or sparse when select different $\phi$ [23]. In this paper, the $\phi$ is set as two and thus the prior is Gaussian, which makes the estimated parameters smooth. Optimal selection of the regularization weights $\gamma_1$ and $\gamma_2$ is an area of active research. Same as the work in [28] does, this paper selects the weights using cross validation. Then, the parameters $\Theta$ in GCRF and $\mathbf{w}$ in MLR can be jointly estimated through minimizing the objective function

$$\hat{\lambda} = \left\{\hat{\Theta}, \hat{\mathbf{w}}\right\} = \arg\min_{\Theta,\mathbf{w}}\left\{L_{d,g,\lambda}^{\mathrm{MCE}}\right\}. \quad (12)$$

### B. Optimization

This paper proposes a gradient descent algorithm to optimize the objective function in (12). The gradient of the objective with respect to $\lambda$ is denoted as $\nabla L_{d,g,\lambda}^{\mathrm{MCE}}$, which can be decomposed by the following chain rule:

$$\nabla L_{d,g,\lambda}^{\mathrm{MCE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial L_{d,g,\lambda}^{\mathrm{MCE}}}{\partial \tau^{\mathrm{sig}}}\frac{\partial \tau^{\mathrm{sig}}(d)}{\partial d} \times \frac{\partial d\left(x_i^{tr}, \hat{\mathbf{z}}_{i,:}^{tr}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right)}{\partial \lambda}\right) + \frac{dR_\lambda}{d\lambda} \quad (13)$$

where the derivatives of $\tau^{\mathrm{sig}}(d)$ given in (10) with respect to $d(\cdot)$ are computed as

$$\frac{\partial \tau^{\mathrm{sig}}}{\partial d} = \alpha \cdot \tau^{\mathrm{sig}} \cdot \left(1 - \tau^{\mathrm{sig}}\right). \quad (14)$$

*Compute* $\partial L_{d,g,\lambda}^{MCE}/\partial \mathbf{w}$*:* Corresponding to the parameters $\mathbf{w}$ in MLR, latent variable $\hat{\mathbf{z}}_{i,:}^{tr}\left(\mathbf{y}^{tr}, \Theta\right)$ are known, and thus the gradient with respect to $\mathbf{w}$ is written as

$$\frac{\partial L_{d,g,\lambda}^{\mathrm{MCE}}}{\partial \mathbf{w}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial L_{d,g,\lambda}^{\mathrm{MCE}}}{\partial \tau^{\mathrm{sig}}}\frac{\partial \tau^{\mathrm{sig}}(d)}{\partial d}\frac{\partial d\left(x_i^{tr}, \hat{\mathbf{z}}_{i,:}^{tr}\left(\mathbf{y}^{tr}, \Theta\right), \mathbf{w}\right)}{\partial \mathbf{w}}\right) + \gamma_1\|\mathbf{w}\|^{\phi-1}. \quad (15)$$

The component-wise gradient in (15) is finally written as (16) shown at the top of the next page.

*Compute* $\partial L_{d,g,\lambda}^{\mathrm{MCE}}/\partial \Theta$*:* Corresponding to the parameters $\Theta = \left\{\Theta^k, k = 1, 2, \ldots, m\right\}$, where $\Theta^k = \left\{\theta_{q,l}^k; q, l\right\}$ is the parameter set of GCRF for $k$th noisy bands, the latent variable $\hat{\mathbf{z}}_{i,:}^{tr}\left(\mathbf{y}^{tr}, \Theta\right)$ is the function of $\Theta$, and thus we further use the chain rule to compute the gradient of objective function with respect to $\Theta$ as

$$\frac{\partial L_{d,g,\lambda}^{\mathrm{MCE}}}{\partial \theta_{q,l}^k} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial L_{d,g,\lambda}^{\mathrm{MCE}}}{\partial \tau^{\mathrm{sig}}}\frac{\partial \tau^{\mathrm{sig}}(d)}{\partial d}\frac{\partial d\left(x_i^{tr}, \hat{\mathbf{z}}_{i,:}^{tr}\right)}{\partial \hat{\mathbf{z}}_{i,k}^{tr}} \times \frac{\partial \hat{\mathbf{z}}_{i,k}^{tr}\left(\mathbf{y}^{tr}, \Theta^k\right)}{\partial \theta_{q,l}^k}\right) + \gamma_2\left|\theta_{q,l}^k\right|^{\phi-1}. \quad (17)$$

It can be finally computed as (18) shown at the top of the next page.

*Algorithm:* The joint learning algorithm mainly optimizes the objective function through the gradient descent method over the parameters $\mathbf{w}$ in MLR, $\Theta$ in GCRF, latent images $\hat{\mathbf{z}}_{:,k}^{tr}\left(\mathbf{y}^{tr}, \Theta\right), k = 1, 2, \ldots, m$ for the selected noisy bands alternatively. Fig. 2 shows the procedures of our joint learning algorithm. Its corresponding flowchart is shown in Fig. 1.

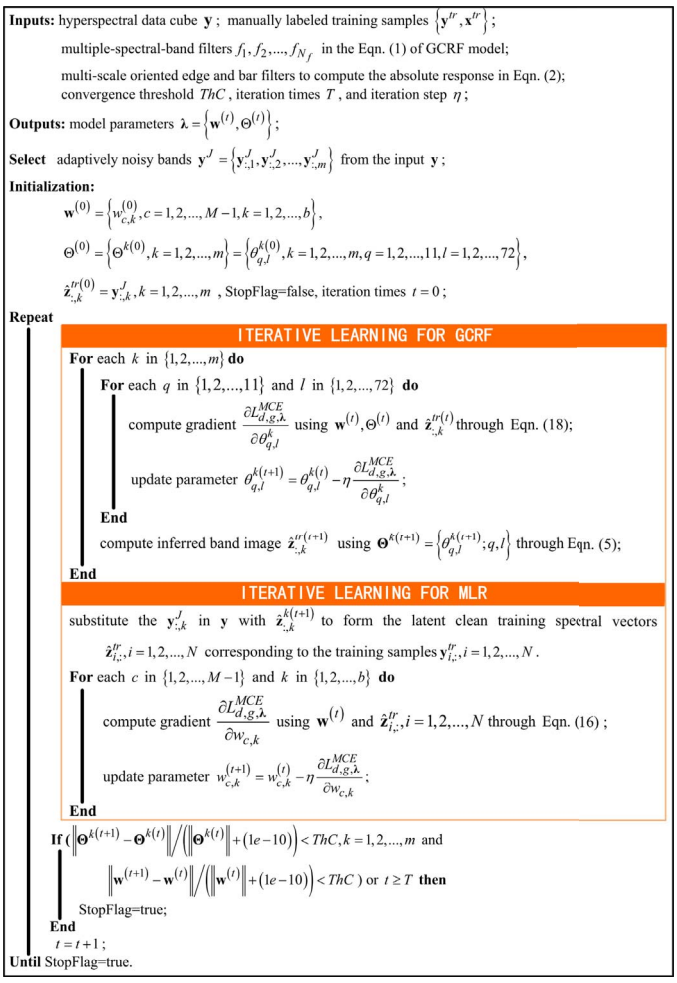


Fig. 2. Joint learning algorithm for the hybrid GCRF and MLR model with maximum linear performance measures.

## IV. Joint Learning of Hybrid GCRF and MLR Model With MEMs

Previous section has presented the method of joint learning GCRF and MLR model to denoise hyperspectral image to minimize the classification error under the framework of MCE. The framework of MCE allows the embedding of not only a linear combination of error rates, but also any evaluation

$$\frac{\partial L_{d,g,\lambda}^{\mathrm{MCE}}}{\partial w_{c,k}} = \frac{1}{N}\sum_{i=1}^{N}\left(\left(\frac{\exp\left(\psi g^i|_{x_i=c}\right)}{\sum\limits_{x_i\in\chi\setminus x_i^{tr}}\exp\left(\psi g^i\right)}\hat{\mathbf{z}}_{i,k}^{tr}\delta\left(c\neq x_i^{tr}\right)-\hat{\mathbf{z}}_{i,k}^{tr}\delta\left(c=x_i^{tr}\right)\cdot\alpha\cdot\tau^{\mathrm{sig}}\cdot\left(1-\tau^{\mathrm{sig}}\right)\right)\right) + \gamma_1\left|w_{c,k}\right|^{\phi-1}, c=1,2,\ldots,M-1; k=1,2,\ldots,b \tag{16}$$

$$\frac{\partial L_{d,g,\lambda}^{\mathrm{MCE}}}{\partial \theta_{q,l}^{k}} = \frac{1}{N}\sum_{i=1}^{N}\left(\left(\left(G-w_{x_i^{tr},k}\right)\delta\left(x_i^{tr}=M\right)+G\cdot\delta\left(x_i^{tr}<M\right)\right)\cdot\alpha\cdot\tau^{\mathrm{sig}}\cdot\left(1-\tau^{\mathrm{sig}}\right)\Delta_\theta\hat{\mathbf{z}}_{i,k}^{tr}\right)+\gamma_2\left|\theta_{q,l}^{k}\right|^{\phi-1} \tag{18}$$

where

$$G=\frac{\sum\limits_{x_i\in\chi\setminus\{x_i^{tr},M\}}\left(\exp\left(\psi g^i\right)w_{x_i,k}\right)}{\sum\limits_{x_i\in\chi\setminus x_i^{tr}}\exp\left(\psi g^i\right)},\ \Delta_\theta\hat{\mathbf{z}}_{i,k}^{tr}=\frac{\partial\hat{\mathbf{z}}_{i,k}^{tr}\left(\mathbf{y}^{tr},\Theta^k\right)}{\partial\theta_{q,l}^{k}}=\left[\left(\mathbf{F}^T\ \mathbf{VF}\right)^{-1}\mathbf{F}^T\frac{\partial\mathbf{V}}{\partial\theta_{q,l}^{k}}\left(-\mathbf{FZ}_k+\mathbf{r}\right)\right]_i$$

and $[\mathbf{a}]_i$ is the operator to extract $i$th entry of the vector $\mathbf{a}$.

measure, including nonlinear measures. In this section, corresponding to the simultaneous denoising and classification of hyperspectral imagery, this paper focuses on developing the method to jointly learn the hybrid GCRF and MLR model to optimize the kappa, which is the usual nonlinear MEM in hyperspectral image classification.

### A. Problem Formalization

The kappa used as the evaluation measure for hyperspectral image classification is defined as

$$K=\frac{N\sum\limits_{c}^{N_{cc}}-\sum\limits_{c}^{N_{c+}N_{+c}}}{N^2-\sum\limits_{c}N_{c+}N_{+c}} \tag{19}$$

where $N$ is the number of total samples, $N_{c+}$ is the number of samples classified as the $c$th class, $N_{+j}$ is the number of samples belonging to $j$th class. If we define the $N_{cj}$ as the number of samples, which belongs to $j$th class but are classified as the $c$th class, $N_{c+}$ and $N_{+j}$ can be computed as

$$N_{c+}=\sum_{j=1}^{M}N_{cj}\quad\text{and}\quad N_{+j}=\sum_{c=1}^{M}N_{cj}.$$

Equation (19) shows that the kappa is a nonlinear MEM. It should be noted that $N_{+j}$ is the number of samples belonging to $j$th class and it is a constant if the sample set is given. Thus, to compute the kappa, two terms, i.e., $N_{cc}$ and $N_{c+}$ needed to be computed at first.

Equation (10) in the previous section shows that $\tau^{\mathrm{sig}}\left(d\left(x_i^{tr},\hat{\mathbf{z}}_{i,:}^{tr}\left(\mathbf{y}^{tr},\Theta\right),\mathbf{w}\right)\right)$ is the continuous count that the sample $\hat{\mathbf{z}}_{i,:}^{tr}\left(\mathbf{y}^{tr},\Theta\right)$ is misclassified. Thus, the number of the correctly classified samples, i.e., $\sum_c N_{cc}$, in (19) can be computed as

$$\sum_{c}N_{cc}=\sum_{i}\left(1-\tau^{\mathrm{sig}}\left(d\left(x_i^{tr},\hat{\mathbf{z}}_{i,:}^{tr}\left(\mathbf{y}^{tr},\Theta\right),\mathbf{w}\right)\right)\right). \tag{20}$$

To compute $N_{c+}$ in (19), we first define a new classification measure $\tilde{d}(\cdot)$ to evaluate the cost that the method classifies the sample $\hat{\mathbf{z}}_{i,:}^{tr}\left(\mathbf{y}^{tr},\Theta\right)$ as $c$th class

$$\begin{aligned}&\tilde{d}\left(\left(x_i^{tr},\hat{\mathbf{z}}_{i,:}^{tr}\left(\mathbf{y}^{tr},\Theta\right),\mathbf{w}\right),c\right)\\&\quad=-g\left(x_i^{tr}|\hat{\mathbf{z}}_{i,:}^{tr}\left(\mathbf{y}^{tr},\Theta\right),\mathbf{w}\right)+g\left(x_i=c|\hat{\mathbf{z}}_{i,:}^{tr}\left(\mathbf{y}^{tr},\Theta\right),\mathbf{w}\right)\\&\quad\triangleq-g^{*i}+g^{ci}\end{aligned} \tag{21}$$

where $g(\cdot)$ is the discriminant function defined in the previous section. Then the number of samples misclassified as $c$th class is $\sum_i\delta(\tilde{d})$. Using $\tau^{\mathrm{sig}}(\tilde{d})$ to approximate $\delta(\tilde{d})$, we can compute $N_{c+}$ as

$$N_{c+}=\sum_{i}\left(\tau^{\mathrm{sig}}(\tilde{d})\delta\left(x_i^{tr}\neq c\right)\right)+\sum_{i}\left(1-\tau^{\mathrm{sig}}(\tilde{d})\delta\left(x_i^{tr}=c\right)\right). \tag{22}$$

Using (19)–(22), the kappa is finally rewritten as

$$K=\frac{N\sum\limits_{c}^{N_{cc}}-\sum\limits_{c}^{N_{c+}N_{+c}}}{N^2-\sum\limits_{c}N_{c+}N_{+c}}\cong\frac{Z_N}{Z_D}$$

where

$$Z_D=N^2-\sum_{c}N_{+c}\left(\sum_{i}\left(\tau^{\mathrm{sig}}\left(\tilde{d}\right)\delta\left(x_i^{tr}\neq c\right)+\left(1-\tau^{\mathrm{sig}}\left(\tilde{d}\right)\right)\delta\left(x_i^{tr}=c\right)\right)\right)$$

and

$$Z_N=N\sum_{i}\left(1-\tau^{\mathrm{sig}}(d)\right)-\sum_{c}N_{+c}\times\left(\sum_{i}\left(\tau^{\mathrm{sig}}(\tilde{d})\delta\left(x_i^{tr}\neq c\right)+\left(1-\tau^{\mathrm{sig}}(\tilde{d})\right)\delta\left(x_i^{tr}=c\right)\right)\right)$$

where $d$ and $\tilde{d}$ are the abbreviates of $d\left(x_i^{tr},\hat{\mathbf{z}}_{i,:}^{tr}\left(\mathbf{y}^{tr},\Theta\right),\mathbf{w}\right)$ and $\tilde{d}\left(\left(x_i^{tr},\hat{\mathbf{z}}_{i,:}^{tr}\left(\mathbf{y}^{tr},\Theta\right),\mathbf{w}\right),c\right)$, respectively. Then, the regularized objective function to be minimized can be finally written as

$$L_{d,g,\lambda}^{\mathrm{MEM}}=\frac{Z_D}{Z_N}+R_\lambda \tag{23}$$

where the regularization term $R_\lambda$ is same as that in (11). The parameters $\Theta$ in GCRF and $\mathbf{w}$ in MLR can be jointly estimated to maximize the kappa

$$\hat{\lambda} = \left\{\hat{\Theta}, \hat{\mathbf{w}}\right\} = \arg\min_{\Theta,\mathbf{w}} \left\{L^{\text{MEM}}_{d,g,\lambda}\right\}. \quad (24)$$

### B. Optimization

This paper proposes a gradient descent algorithm to optimize the objective function in (23). The gradient of the objective with respect to $\lambda$ in (23) is computed as

$$\nabla L^{\text{MEM}}_{d,g,\lambda} = \frac{Z^{'}_D Z_N - Z_D Z^{'}_N}{Z_N^2} + \frac{dR_\lambda}{d\lambda}. \quad (25)$$

Thus, the main calculations lie in $Z^{'}_D$ and $Z^{'}_N$.

*Compute $Z^{'}_D$:* Corresponding to the parameters $\mathbf{w}$ in MLR, the component-wise gradient of $Z_D$ is finally written as (see supplementary)

$$\frac{\partial Z_D}{\partial w_{j,k}} = -\sum_c N_{+c} \sum_i \left(\alpha \cdot \tau^{\text{sig}} \cdot \left(1-\tau^{\text{sig}}\right) \cdot \kappa\left(j, c, x_i^{tr}\right)\right) \quad (26)$$

where

$$\kappa\left(j, c, x_i^{tr}\right) = \begin{cases} 0, & \text{if} \quad j = c = x_i^{tr} \\ -\hat{\mathbf{z}}^{tr}_{i,k}, & \text{if} \quad j = x_i^{tr}, j \neq c \\ \hat{\mathbf{z}}^{tr}_{i,k}, & \text{if} \quad j \neq x_i^{tr}, j = c \\ 0, & \text{if} \quad j \neq x_i^{tr}, j \neq c. \end{cases}$$

Corresponding to the parameters $\Theta$ in GCRF, the component-wise gradient of $Z_D$ is finally written as (see supplementary)

$$\frac{\partial Z_D}{\partial \theta^k_{q,l}} = -\sum_c N_{+c} \sum_i \left(\alpha \cdot \tau^{\text{sig}} \cdot \left(1-\tau^{\text{sig}}\right) \times \Delta_\theta \hat{\mathbf{z}}^{tr}_{i,k}\left(w_{c,k} - w_{x_i^{tr},k}\right)\right) \quad (27)$$

where $w_{c,k} = 0$ if $c = M$ and $w_{x_i^{tr},k} = 0$ if $x_i^{tr} = M$.

*Compute $Z^{'}_N$:* Corresponding to the parameters $\mathbf{w}$ in MLR, the component-wise gradient of $Z_N$ with respect to $\mathbf{w}$ is finally written as (see supplementary)

$$\frac{\partial Z_N}{\partial w_{j,k}} = \frac{\partial Z_D}{\partial w_{j,k}} - \alpha N \sum_i \tau^{\text{sig}} \cdot \left(1-\tau^{\text{sig}}\right) \cdot \xi\left(j, x_i^{tr}\right) \quad (28)$$

where

$$\xi\left(j, x_i^{tr}\right) = \begin{cases} -\hat{\mathbf{z}}^{tr}_{i,k}, & \text{if} \quad j = x_i^{tr} \\ \dfrac{\exp\left(\psi g^i|_{x_i=j}\right)}{\sum_{x_i \in \chi \setminus x_i^{tr}} \exp\left(\psi g^i\right)} \hat{\mathbf{z}}^{tr}_{i,k}, & \text{else.} \end{cases}$$

Corresponding to the parameters $\Theta$ in GCRF, the component-wise gradient of $Z_N$ is finally written as (see supplementary)

$$\frac{\partial Z_N}{\partial \theta^k_{q,l}} = -\sum_c N_{+c} \sum_i \left(\alpha \cdot \tau^{\text{sig}} \cdot \left(1-\tau^{\text{sig}}\right) \Delta_\theta \hat{\mathbf{z}}^{tr}_{i,k}\left(w_{c,k} - w_{x_i^{tr},k}\right)\right) - N \sum_i \left(\alpha \cdot \tau^{\text{sig}} \cdot \left(1-\tau^{\text{sig}}\right)\right) \cdot \Delta_\theta \hat{\mathbf{z}}^{tr}_{i,k} \cdot \left(-w_{x_i^{tr},k} + G\right). \quad (29)$$

*Compute $\nabla L^{MEM}_{d,g,\lambda}$:* Using the (25), (26), and (28), the component-wise gradient of $L^{\text{MEM}}_{d,g,\lambda}$ with respect to $w_{j,k}$ is written as

$$\frac{\partial L^{\text{MEM}}_{d,g,\lambda}}{\partial w_{j,k}} = \frac{1}{Z_N^2}\left(Z_N \frac{\partial Z_D}{\partial w_{j,k}} - Z_D \frac{\partial Z_N}{\partial w_{j,k}}\right) + \gamma_1 \left|w_{j,k}\right|^{\phi-1} \quad (30)$$

where $\partial Z_D / \partial w_{j,k}$ and $\partial Z_N / \partial w_{j,k}$ are computed as (26) and (28), respectively, while using the (25), (27), and (29), the component-wise gradient of $L^{\text{MEM}}_{d,g,\lambda}$ with respect to $\theta^k_{q,l}$ is written as

$$\frac{\partial L^{\text{MEM}}_{d,g,\lambda}}{\partial \theta^k_{q,l}} = \frac{1}{Z_N^2}\left(Z_N \frac{\partial Z_D}{\partial \theta^k_{q,l}} - Z_D \frac{\partial Z_N}{\partial \theta^k_{q,l}}\right) + \gamma_2 \left|\theta^k_{q,l}\right|^{\phi-1} \quad (31)$$

where $\partial Z_D / \partial \theta^k_{q,l}$ and $\partial Z_N / \partial \theta^k_{q,l}$ are computed as (27) and (29), respectively.

*Algorithm:* The joint learning algorithm mainly optimizes the objective function through the gradient descent method over the parameters $\mathbf{w}$ in MLR, $\Theta$ in GCRF, latent images $\hat{\mathbf{z}}^{tr}_{:,k}, k = 1, 2, \ldots, m$ for the selected noisy bands alternatively. Fig. 3 shows the procedures of our joint learning algorithm, and Fig. 1 shows its flowchart.

## V. Experimental Results

### A. Data Sets for Experiments

To validate the effectiveness of the proposed hybrid GCRF and MLR model for simultaneous denoising and classification of hyperspectral imagery, we perform the experiments on two data cubes, representing different environments in remote sensing.

1) *Indian Pine*: Airborne visible/infrared imaging spectrometer image was taken at the Indian Pine test site in Northwestern Indiana on 12 June, 1992 [1]. This image contains $145 \times 145$ pixels and 220 bands. Before the denoising and classification processing, the atmospheric and water absorption bands from bands 150–163 were removed from the original hyperspectral image. Therefore, there were only 206 bands used in the experiments. The adopted data cube is shown in Fig. 4(a). The major advantage in using this data set is the availability of a reference map [see Fig. 4(b) and (c)] prepared from the field surveys conducted at the time of image acquisition. Therefore, this data set has been extensively used to test various hyperspectral image analysis algorithms.

2) *Purdue Campus*: HyMap radiance data of the Purdue Campus was acquired by HyVista Corporation and Analytical Imaging and Geophysics LLC as a part of a mission flown during Fall 1999. The data cube covers an urban and mixed environment site, where the rich man-made structures present the ideal sources to evaluate the algorithms' performances to preserve image details. The whole image contains $1950 \times 512$ pixels and 126 spectral channels, covering 450–2500 nm. A Part of the image of a size $388 \times 219 \times 126$ is used to verify the performance of the proposed algorithms. In addition, to get an accurate estimate of the proposed method's classification performances, we manually generated detailed cover labels as the ground truth using the Multispec software [42]. Fig. 5(a) shows the adopted data cube, and

**Inputs:** hyperspectral data cube $\mathbf{y}$; manually labeled training samples $\left\{\mathbf{y}^{tr}, \mathbf{x}^{tr}\right\}$;
multiple-spectral-band filters $f_1, f_2, ..., f_{N_f}$ in the Eqn.(1) of GCRF model;
multi-scale oriented edge and bar filters to compute the absolute response in Eqn. (2);
convergence threshold $ThC$, iteration times $T$, and iteration step $\eta$;

**Outputs:** model parameters $\boldsymbol{\lambda} = \left\{\mathbf{w}^{(t)}, \Theta^{(t)}\right\}$;

**Select** adaptively noisy bands $\mathbf{y}^J = \left\{\mathbf{y}^J_{:,1}, \mathbf{y}^J_{:,2}, ..., \mathbf{y}^J_{:,m}\right\}$ from the input $\mathbf{y}$;

**Initialization:**

$\mathbf{w}^{(0)} = \left\{w^{(0)}_{j,k}, j = 1,2,...,M-1, k = 1,2,...,b\right\}$,

$\Theta^{(0)} = \left\{\Theta^{k(0)}, k = 1,2,...,m\right\} = \left\{\theta^{k(0)}_{q,l}, k = 1,2,...,m, q = 1,2,...,11, l = 1,2,...,72\right\}$,

$\hat{\mathbf{z}}^{tr(0)}_{:,k} = \mathbf{y}^J_{:,k}, k = 1,2,...,m$, StopFlag=false, iteration times $t = 0$;

**Repeat**

ITERATIVE LEARNING FOR GCRF

**For** each $k$ in $\{1,2,...,m\}$ **do**

**For** each $q$ in $\{1,2,...,11\}$ and $l$ in $\{1,2,...,72\}$ **do**

compute gradient $\frac{\partial Z_D}{\partial \theta^k_{q,l}}$ and $\frac{\partial Z_N}{\partial \theta^k_{q,l}}$ through Eqn. (27) and (29) respectively;

compute gradient $\frac{\partial L^{MCE}_{d,g,\boldsymbol{\lambda}}}{\partial \theta^k_{q,l}}$ using $\frac{\partial Z_D}{\partial \theta^k_{q,l}}$ and $\frac{\partial Z_N}{\partial \theta^k_{q,l}}$ through Eqn. (31);

update parameter $\theta^{k(t+1)}_{q,l} = \theta^{k(t)}_{q,l} - \eta \frac{\partial L^{MCE}_{d,g,\boldsymbol{\lambda}}}{\partial \theta^k_{q,l}}$;

**End**

compute inferred band image $\hat{\mathbf{z}}^{tr(t+1)}_{:,k}$ using $\boldsymbol{\Theta}^{k(t+1)} = \left\{\theta^{k(t+1)}_{q,l}; q, l\right\}$ through Eqn. (5);

**End**

ITERATIVE LEARNING FOR MLR

substitute the $\mathbf{y}^J_{:,k}$ in $\mathbf{y}$ with $\hat{\mathbf{z}}^{k(t+1)}_{:,k}$ to form the latent clean training spectral vectors $\hat{\mathbf{z}}^{tr}_{i,:}, i = 1,2,...,N$ corresponding to the training samples $\mathbf{y}^{tr}_{i,:}, i = 1,2,...,N$.

**For** each $j$ in $\{1,2,...,M-1\}$ and $k$ in $\{1,2,...,b\}$ **do**

compute gradient $\frac{\partial Z_D}{\partial w_{j,k}}$ and $\frac{\partial Z_N}{\partial w_{j,k}}$ through Eqn. (26) and (28) respectively;

compute gradient $\frac{\partial L^{MCE}_{d,g,\boldsymbol{\lambda}}}{\partial w_{j,k}}$ using $\frac{\partial Z_D}{\partial w_{j,k}}$ and $\frac{\partial Z_N}{\partial w_{j,k}}$ through Eqn. (30);

update parameter $w^{(t+1)}_{j,k} = w^{(t)}_{j,k} - \eta \frac{\partial L^{MCE}_{d,g,\boldsymbol{\lambda}}}{\partial w_{j,k}}$;

**End**

**If** $\left(\left\|\boldsymbol{\Theta}^{k(t+1)} - \boldsymbol{\Theta}^{k(t)}\right\| \Big/ \left(\left\|\boldsymbol{\Theta}^{k(t)}\right\| + (1e-10)\right) < ThC, k = 1,2,...,m$ and

$\left\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\right\| \Big/ \left(\left\|\mathbf{w}^{(t)}\right\| + (1e-10)\right) < ThC$) or $t \geq T$ **then**

StopFlag=true;

**End**

$t = t+1$;

**Until** StopFlag=true.

Fig. 3. Joint learning algorithm for the hybrid GCRF and MLR model with the multivariate evaluation measures.

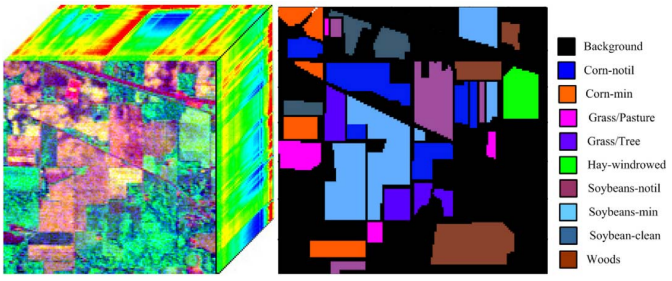


Fig. 4. Indian Pine data cube. (a) Hyperspectral 3-D cube. (b) Ground truth with nine classes. (c) Map color of ground truth.

the generated ground truth and label details are shown in Fig. 5(b) and (c).

### *B. Evaluation on Synthetic Noisy Data Cubes*

*1) Experimental Setup:* To quantitatively evaluate the algorithms' denoising performance, we compute the SNR (dB) for

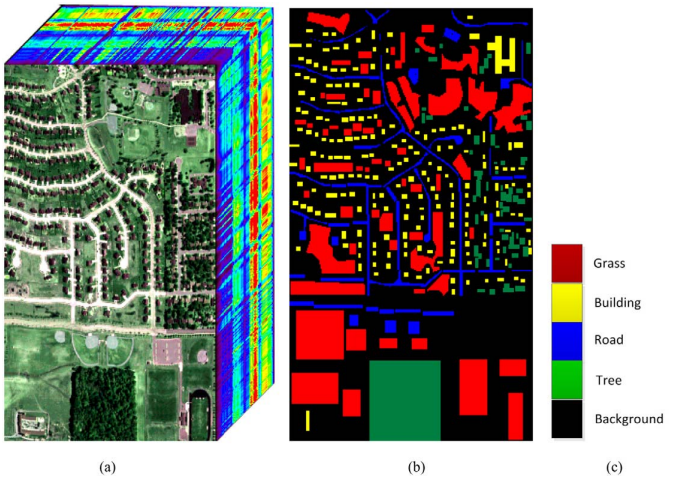


Fig. 5. Purdue Campus data cube. (a) Hyperspectral 3-D cube. (b) Ground truth with four classes. (c) Map color of ground truth.

hyperspectral data cubes as

$$\mathrm{SNR} = 10 \log \left( \frac{\sum_{i,j,m} \mathbf{I}^2(i,j,k)}{\sum_{i,j,k} \left[\tilde{\mathbf{I}}^2(i,j,k) - \mathbf{I}^2(i,j,k)\right]} \right) \tag{32}$$

where $\mathbf{I}$ is the reference data cube and $\tilde{\mathbf{I}}$ is the test noisy data cube or the denoised data cube corresponding to the SNR before or after denoising respectively. We can see from the above equation of SNR that we need both the noisy and reference data or both the denoised and reference data to compute the SNRs. However, they are usually not coexistent in the real world. Therefore, in this set of experiments, we extract the high-SNR bands as the reference data and add noise to them to synthesize the noisy data.

For sensors used in hyperspectral imagery, there are mainly two noise sources, i.e., the signal-independent noise, which can be assumed Gaussian distributed, and the signal-dependent (SD) photonic noise which is more likely Poisson-distributed noise [43]. In the simulated process, we followed the method in [43] and added the noise as a mixture of additive and SD components

$$y(i,j,k) = z(i,j,k) + n_G(i,j,k) + \sqrt{I(i,j,k)} n_P(i,j,k) \tag{33}$$

where $\mathbf{z}$ and $\mathbf{y}$ are reference noise free and noisy data, $\mathbf{n}_G$ and $\mathbf{n}_P$ are Gaussian- and Poisson-distributed noise, respectively. To synthesize the noisy bands considered mainly in this paper, 125 and 90 high-SNR bands from Indian Pine and Purdue Campus, respectively were firstly selected as the reference data cubes, and then half of the high-SNR bands were randomly selected to add noise.

Then, the proposed hybrid GCRF and MLR model were jointly learned over the synthetic noisy data cubes to denoise the synthetic data for the optimal classification in the sense of maximizing OA or kappa measure through the algorithms shown in Fig. 2 or 3. The training samples for the GCRF

TABLE I
NUMBER OF TRAINING AND TEST SAMPLES OF NINE CLASSES IN INDIAN PINE DATA CUBE

| CLASS | TRAINING | TEST |
|---|---|---|
| **C1-Corn-no till** | 430 | 1004 |
| **C2-Corn-min** | 250 | 584 |
| **C3-Grass/pasture** | 149 | 348 |
| **C4-Grass/trees** | 224 | 523 |
| **C5-Hay-windrowed** | 146 | 343 |
| **C6-Soybeans-no till** | 290 | 678 |
| **C7-Soybeans-min** | 740 | 1728 |
| **C8-Soybeans-clean till** | 184 | 430 |
| **C9-Woods** | 388 | 906 |
| **TOTAL** | **2803** | **6542** |

TABLE II
NUMBER OF TRAINING AND TEST SAMPLES OF FOUR CLASSES IN PURDUE CAMPUS DATA CUBE

| CLASS | TRAINING | TEST |
|---|---|---|
| **C1-Grass** | 735 | 13970 |
| **C2-Building** | 185 | 3518 |
| **C3-Road** | 212 | 4041 |
| **C4-Tree** | 304 | 5777 |
| **TOTAL** | **1437** | **27305** |

and MLR were selected through the methods mentioned in the flowchart of our proposed framework introduced in Section II. The number of training and test samples is presented in Tables I and II. To run our proposed learning algorithms (see Figs. 2 and 3), the convergence threshold ThC, iteration time $T$, and iteration step $\eta$ were set as 1e-8, 500, and 0.001 respectively. The trained hybrid models corresponding to the learning processes with OA and kappa measures are abbreviated as GCRF + MLR-O and GCRF + MLR-K, which were finally used to obtain simultaneously the denoising and classification results through the inference method presented in Section II-D.

To thoroughly evaluate the performances of the proposed hybrid GCRF and MLR model for simultaneous denoising and classification, we ran several sets of experiments to compare the proposed methods with the recent methods from two aspects. First, to evaluate the algorithms' performances for denoising, the proposed hybrid method is compared with the channel-by-channel Wiener filtering method (CCWF), provided by MATLAB, PCA + wavelet-GSM (PCA + WGSM), and GCRF [18]. The PCA + WGSM method applied PCA at first to decorrelate the fine features of the data cube from the noise. After that the wavelet transform was performed in three decomposition levels. Then, the BLS-GSM method was used to generate the denoising results for each PCA output channel. The final denoising result was obtained through inverse PCA transformation. The details about the GCRF were presented in Section II. Second, to evaluate the algorithms' performances for classification, we compare the results of the proposed methods with that of the sequential methods, which perform image denoising using the previous comparative denoising methods at first, and then feed the denoised results into the MLR classification system. We denote the methods as CCWF + MLR, PCA + WGSM + MLR, and GCRF + MLR, respectively.

*2) Denoising and Classification Performance:* To obtain the statistically meaningful performance evaluation, we first alternate the human-aided training sample selection procedure shown in Fig. 1 as the usual random selection of training samples, and then the classification accuracies are calculated over the remaining test samples [1]–[6]. We computed the average values of SNRs, OAs, and kappa values from 10 runs of trainings and tests as the performance measures. Furthermore, we additionally calculated the AAs of all the classes for comprehensive evaluation. Tables III and IV show the denoising and classification performances of different methods. It is clearly observed from the results over original noisy data and the denoised data obtained by different methods that all the classification results have been improved after the denoising process. This demonstrates the necessity of denoising in the hyperspectral image analysis. In addition, the set of multidimetional methods, i.e., GCRF, CRF + MLR-O, and CRF + MLR-K, have obtained much better results than the set of channel-by-channel methods, i.e., CCWF. This demonstrates the importance of using multidimentional data and thus the spectral contextual information in hyperspectral image denoising. Fig. 6 further shows the SNR values of different denoising approaches in different bands of the synthetic noisy Indian Pine and Purdue Campus data cubes respectively It can be clearly observed that the proposed methods obtained better SNR performances than other methods at most of the bands.

Tables III and IV also show that the proposed GCRF + MLR-O and GCRF + MLR-K models, which were trained by the joint learning methods, obtained the results with SNR = 30.26 dB and SNR = 30.33 dB for Indian Pine data cube and SNR = 31.06 dB and SNR = 31.10 dB for Purdue Campus data cube, which are a little better than the SNR = 30.04 dB and SNR = 31.01 dB obtained by the denoising focus method, i.e., GCRF + MLR. However, from the view of classification evaluation, the GCRF + MLR-O and GCRF + MLR-K obtained much better results than GCRF + MLR at all the OA, AA, and kappa evaluation criteria. The results can be predicted since the proposed methods learn the hybrid GCRF and MLR model simultaneously, and thus the recovered results of GCRF and the class information from training samples of MLR can improve each other's performances, whereas the sequential method learns the GCRF and MLR separately, which makes the GCRF and MLR focus only on its own objective, neglecting their complementary actions. In addition, just as that can be the theoretically predicted, the GCRF + MLR-O obtained better OA than GCRF + MLR-K, while the kappa of GCRF + MLR-K is larger than GCRF + MLR-O. Therefore,

TABLE III

DENOISING AND CLASSIFICATION PERFORMANCES OF DIFFERENT METHODS OVER INDIAN PINE DATA CUBE. SEVERAL PERFORMANCE MEASURES ARE INCLUDED: SNR (dB), CLASS PERCENTAGE ACCURACY (%), OA (%), AA (%), AND KAPPA STATISTIC (K)

| METHOD | SNR(dB) | CLASS PERCENTAGE ACCURACY | | | | | | | | | OA | AA | K |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | | | |
| **NOISY+MLR** | 22.54 | 72.27 | 66.84 | 72.71 | 81.10 | 85.90 | 68.89 | 73.48 | 76.31 | 84.89 | 75.19 | 75.82 | 0.7264 |
| **CCWF+MLR** | 23.89 | 80.01 | 66.84 | 77.18 | 81.10 | 90.45 | 69.69 | 77.98 | 79.93 | 87.47 | 78.71 | 78.96 | 0.7585 |
| **PCA+WGSM+MLR** | 28.71 | 95.50 | 73.50 | 86.13 | 87.05 | 97.27 | 75.43 | 82.49 | 87.16 | 96.05 | 86.43 | 86.73 | 0.8452 |
| **GCRF+MLR** | 30.04 | 97.83 | 81.49 | 88.37 | 88.54 | 98.18 | 80.02 | 88.34 | 88.96 | 97.77 | 90.16 | 89.95 | 0.8895 |
| **GCRF+MLR-O** | 30.26 | **96.75** | 78.83 | 88.37 | **92.26** | 94.55 | 84.62 | **98.02** | 89.51 | **98.54** | **93.03** | 91.27 | 0.9169 |
| **GCRF+MLR-K** | **30.33** | 95.27 | **89.48** | **90.60** | 91.52 | **95.91** | **88.40** | 94.19 | **91.32** | 95.19 | 92.92 | **92.43** | **0.9180** |

TABLE IV

DENOISING AND CLASSIFICATION PERFORMANCES OF DIFFERENT METHODS OVER PURDUE CAMPUS DATA CUBE. SEVERAL PERFORMANCE MEASURES ARE INCLUDED: SNR (dB), CLASS PERCENTAGE ACCURACY (%), OA (%), AA(%), AND KAPPA STATISTIC (K)

| METHOD | SNR(dB) | CLASS PERCENTAGE ACCURACY | | | | OA | AA | K |
|---|---|---|---|---|---|---|---|---|
| | | C1 | C2 | C3 | C4 | | | |
| **NOISY+MLR** | 22.96 | 86.66 | 72.89 | 76.70 | 79.66 | 80.91 | 78.48 | 0.7325 |
| **CCWF+MLR** | 23.17 | 86.73 | 75.64 | 82.94 | 81.84 | 83.71 | 81.79 | 0.7738 |
| **PCA+WGSM+MLR** | 29.05 | 91.60 | 81.16 | 87.74 | 85.37 | 88.36 | 86.46 | 0.8265 |
| **GCRF+MLR** | 31.01 | 92.98 | 82.53 | 88.70 | 89.90 | 90.35 | 88.53 | 0.8517 |
| **GCRF+MLR-O** | 31.06 | **95.76** | 83.49 | 88.77 | 90.62 | **92.06** | 89.66 | 0.8763 |
| **GCRF+MLR-K** | **31.10** | 92.85 | **89.28** | **91.41** | **91.46** | 91.88 | **91.25** | **0.8773** |

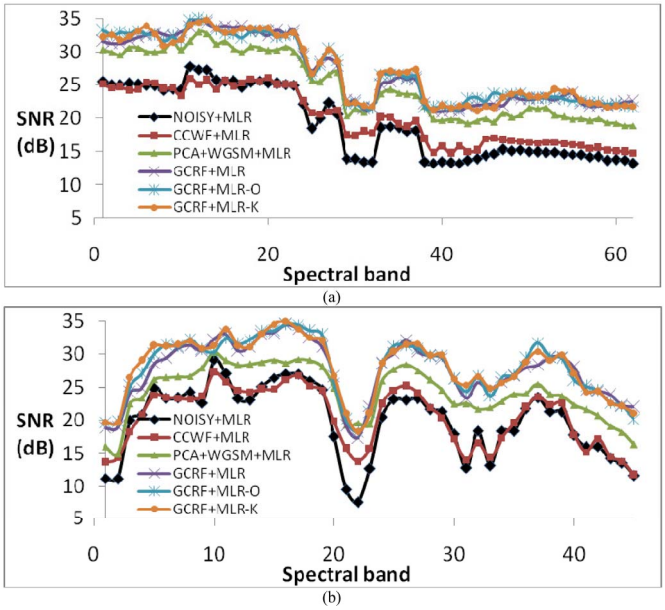


Fig. 6. SNR values of the different approaches in each noisy band of the synthetic noisy data cubes. (a) SNR values of Indian Pine data cube. (b) SNR values of Purdue Campus data cube.

we can select the appropriate methods for different tasks according to their preferences of performance measures. Moreover, in theory, the proposed methods can be fruitfully extended to other performance measures.

*3) Effect of Noise Degree on Denoising and Classification Performance:* In the experiments, we test the effects of noise degree on denoising and classification performance. Over Indian Pine data set, 10 different situations were analyzed in this setup, i.e., the values of $\sigma$ are 0.02, 0.025, 0.03, 0.04, 0.05, 0.06, 0.08, 0.1, 0.13, and 0.15 and the training and test samples for the evaluation of classification are fixed as Table I. Over Purdue Campus data set, we also analyzed 10 different situations with different values of $\sigma$: 0.02, 0.025, 0.03, 0.04, 0.05, 0.06, 0.08, 0.1, 0.13, and 0.15, and the training and test samples for the evaluation of classification are fixed as Table II. Fig. 7 shows that although all the measures of different methods over Indian Pine data set suffer from the increase of noise degree, our proposed methods show more stable changes, which could be derived from the useful class information from the manually labeled training samples. While over Purdue Campus data set (see Fig. 8), all the curves show similar decreasing behaviors with that in Fig. 7 but in relatively stable manners.

*4) Effect of Training Set Size on Denoising and Classification Performance:* At the end of this section, we test the effects of training set size on denoising and classification performance. Over Indian Pine data set, five different situations were analyzed in this setup: 10%, 20%, 30%, 40%, and 50% of the total samples were randomly selected to train MLR and the remaining was used for the test

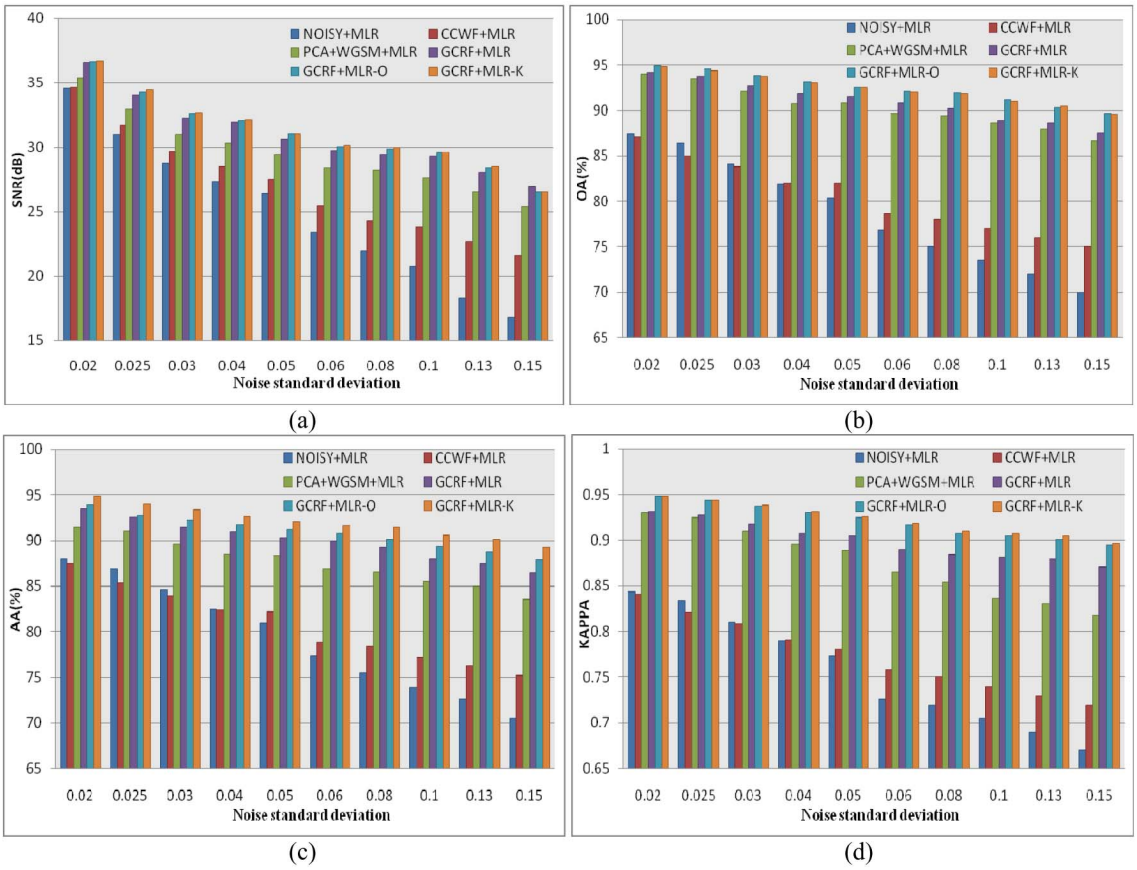


Fig. 7. Denoising and classification performances with different noise standard deviation over Indian Pine data cube. (a)–(d) The performances corresponding to the measure of SNR, OA, AA, and KAPPA.

procedure, while over Purdue Campus data set, we analyzed six different situations where training data sets have 1%, 3%, 5%, 10%, 15%, and 20% of the total samples. The noise degrees over Indian Pine and Purdue Campus data sets are fixed. Tables V and VI give the training and test data sizes and corresponding performance values. We can see from the tables that with the increase of training samples, the OA, AA, and kappa of all the methods increased significantly. In addition, the increase of training samples also improved the SNR values of the proposed methods although the training samples are mainly related to the classification. On the contrary, the GCRF + MLR, which learned the GCRF for denoising and MLR for classification separately, the increasing training samples for MLR have no any effect on the SNR. Then, the SNRs of GCRF + MLR always be 30.04 and 31.01 dB for Indian Pine and Purdue Campus, respectively. All the experiments demonstrate that our proposed joint learning methods of GCRF and MLR model can effectively learn and use the close connections between the denoising and classification, making the useful information can be transferred between them and thus making the denoising and classification presents positive effect on performance improvement for each other: denoising task presents a more clear and thus reliable input for the classification, while the classification task gives high-level class information for denoising.

### C. Evaluation on Real-World Noisy Data Cubes

*1) Experimental Setup:* Previous experiments on synthetic noisy data cubes demonstrated the performance improvement of our proposed methods compared with other popular methods. In this section, using the real-world noisy data cubes, we will further validate their performances of removing noise and classification. The real-world noisy data cubes are shown in Section V-A. The parameters in the proposed joint learning methods, i.e., convergence threshold ThC, iteration time $T$, and iteration step $\eta$, were set as 1e-8, 500, and 0.001. The number of training and test samples is presented in Tables I and II.

*2) Denoising and Classification Performance:* In our experiments, the proposed methods automatically selected 81 and 46 noisy bands for the Indian Pine and *Purdue Campus* data cube, respectively. Figs. 9(a) and 10(a) show a noisy band of Indian Pine and *Purdue Campus* data cube, respectively. Figs. 9(e) and 10(e) show the recovered images of GCRF + MLR-O, whereas the results of GCRF + MLR-K are shown in Figs. 9(f) and 10(f). Figs 9(b)–(d) and 10(b)–(d) show the denoising results of other methods to give comparisons.

The results produced by our algorithms show that our techniques are able to smooth out flat regions, preserve sharp edges, as well as keep subtle texture details. In the nearly uniform regions, such as the natural land covers in Fig. 9 and

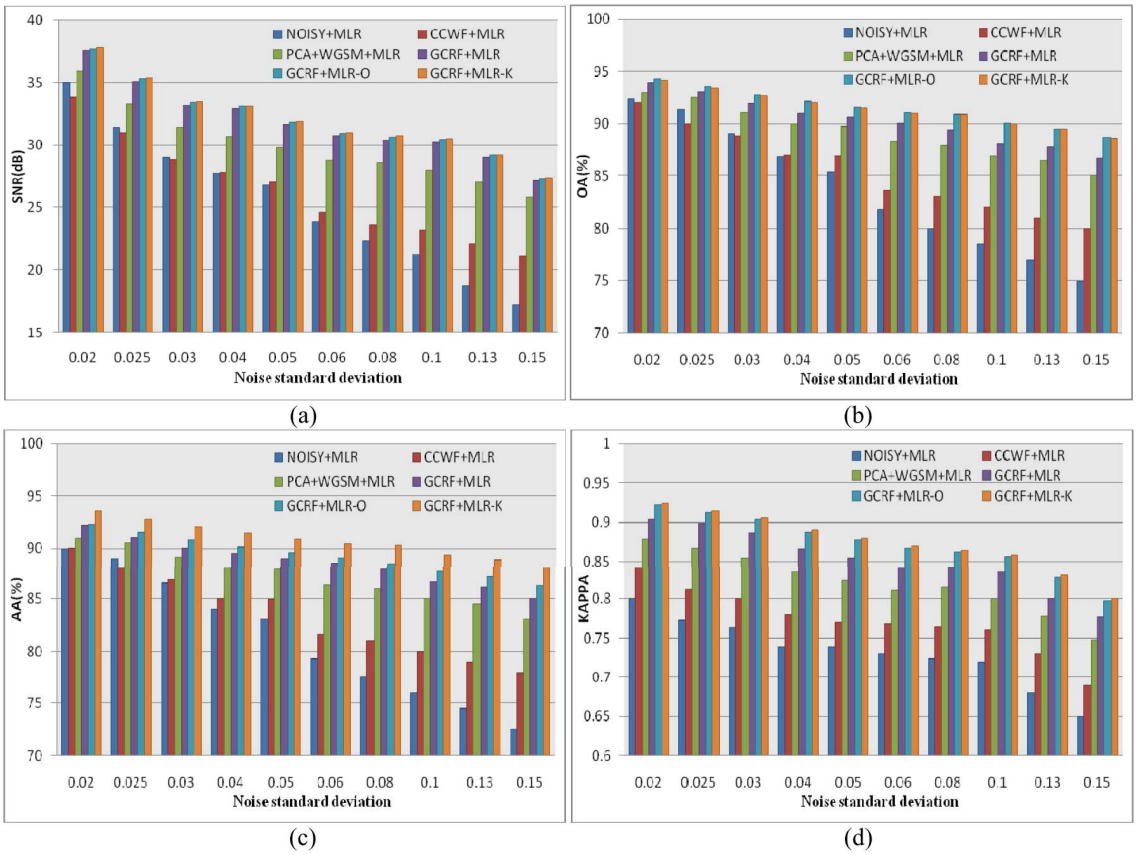


Fig. 8. Denoising and classification performances with different noise standard deviation over Purdue Campus data cube. (a)–(d) The performances corresponding to the measure of SNR, OA, AA, and KAPPA.

TABLE V

DENOISING AND CLASSIFICATION PERFORMANCES OF GCRF + MLR, GCRF + MLR-O, AND GCRF + MLR-K WITH DIFFERENT TRAINING AND TEST DATA SIZES (RATE OF TOTAL SAMPLES) OVER THE SYNTHETIC NOISY INDIAN PINE DATA CUBE

| *TRAINING SIZE* | *TEST SIZE* | *GCRF+MLR* | | | | *GCRF+MLR-O* | | | | *GCRF+MLR-K* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *SNR* | *OA* | *AA* | *K* | *SNR* | *OA* | *AA* | *K* | *SNR* | *OA* | *AA* | *K* |
| 934(10%) | 8411(90%) | 30.04 | 88.32 | 88.02 | 0.8763 | 30.05 | 90.18 | 89.27 | 0.8869 | 30.08 | 90.07 | 89.73 | 0.8879 |
| 1869(20%) | 7476(80%) | 30.04 | 89.27 | 88.95 | 0.8876 | 30.13 | 91.73 | 90.51 | 0.8957 | 30.18 | 91.65 | 91.25 | 0.8968 |
| 2803(30%) | 6542(70%) | 30.04 | 90.16 | 89.95 | 0.8895 | 30.26 | 93.03 | 91.27 | 0.9169 | 30.33 | 92.92 | 92.43 | 0.9180 |
| 3738(40%) | 5607(60%) | 30.04 | 92.93 | 92.46 | 0.9109 | 30.45 | 94.07 | 92.83 | 0.9305 | 30.57 | 93.96 | 93.51 | 0.9308 |
| 3738(50%) | 5607(50%) | 30.04 | 93.47 | 93.09 | 0.9213 | 30.79 | 94.62 | 93.39 | 0.9376 | 30.86 | 94.53 | 94.09 | 0.9386 |

grass in Fig. 10, our methods output almost uniform patches. In the significantly fluctuated regions such as fine man-made structures in Fig. 9 and buildings in Fig. 10, the recovered boundaries are sharp, whereas in the textural regions, such as trees in Fig. 10, many subtle texture details are preserved. In addition, our methods effectively avoid the appearance of structural artifacts. However, the denoising result using the CCWF appears oversmoothed, and most of the edge information is lost. The PCA + WGSM a global method, deals with the whole data cube automatically. It effectively reduced the noise, but the restorations are usually impaired by a whole blur and thus are relatively insufficient in details. Even compared with the GCRF trained mainly for denoising, our proposed methods obtained comparative or better denoising results.

Figs. 11 and 12 further show the classification results for the performance evaluation. It can be clearly observed that the classification results have been greatly improved after the denoising process. The classification results of the original noisy data cubes appear fragmentary, because of the effect of the strong noise information in most of the bands. CCWF + MLR, which usually oversmooth the denoised results or introduce additional structures, led to the introduction of artifacts in the spectral signatures and

TABLE VI

DENOISING AND CLASSIFICATION PERFORMANCES OF GCRF + MLR, GCRF + MLR-O, AND GCRF + MLR-K WITH DIFFERENT TRAINING AND TEST DATA SIZES (RATE OF TOTAL SAMPLES) OVER THE SYNTHETIC NOISY PURDUE CAMPUS DATA CUBE

| *TRAINING SIZE* | *TEST SIZE* | *GCRF+MLR* | | | | *GCRF+MLR-O* | | | | *GCRF+MLR-K* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *SNR* | *OA* | *AA* | *K* | *SNR* | *OA* | *AA* | *K* | *SNR* | *OA* | *AA* | *K* |
| 287(1%) | 28455(99%) | 31.01 | 89.96 | 89.35 | 0.8405 | 31.02 | 91.65 | 90.29 | 0.8678 | 31.03 | 91.48 | 90.89 | 0.8687 |
| 862(3%) | 27880(97%) | 31.01 | 90.17 | 89.58 | 0.8487 | 31.04 | 91.88 | 90.47 | 0.8721 | 31.06 | 91.66 | 91.07 | 0.8732 |
| 1437(5%) | 27305(95%) | 31.01 | 90.35 | 89.83 | 0.8517 | 31.06 | 92.06 | 90.86 | 0.8763 | 31.10 | 91.88 | 91.25 | 0.8773 |
| 2874(10%) | 25868(90%) | 31.01 | 90.84 | 90.24 | 0.8602 | 31.20 | 92.57 | 91.15 | 0.8812 | 31.26 | 92.38 | 91.76 | 0.8822 |
| 4311(15%) | 24431(85%) | 31.01 | 91.22 | 90.68 | 0.8681 | 31.32 | 93.01 | 91.71 | 0.8927 | 31.40 | 92.89 | 92.21 | 0.8936 |
| 5748(20%) | 22994(80%) | 31.01 | 91.76 | 91.12 | 0.8769 | 31.59 | 93.32 | 92.07 | 0.9006 | 31.67 | 93.26 | 92.65 | 0.9015 |

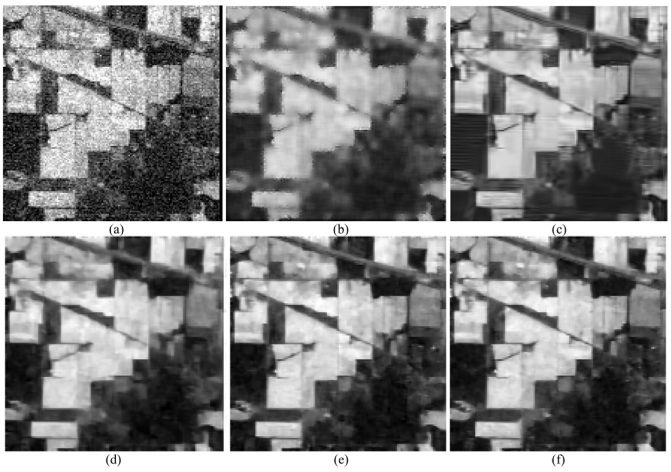


Fig. 9. Denoising results of the real-world Indian Pine data cube. (a) Noisy band. (b)–(f) Denoising results of (a) by CCWF, PCA + WGSM, GCRF, GCRF + MLR-O, and GCRF + MLR-K.

TABLE VII

CLASSIFICATION PERFORMANCES OF DIFFERENT METHODS OVER THE REAL-WORLD INDIAN PINE DATA CUBE

| METHODS / PERFORMANCE | NOISY +MLR | CCWF +MLR | PCA+WGSM +MLR | GCRF +MLR | GCRF +MLR-O | GCRF +MLR-K |
|---|---|---|---|---|---|---|
| OA | 77.46% | 79.85% | 88.75% | 92.29% | 95.07% | 94.85% |
| KAPPA | 0.7315 | 0.7653 | 0.8851 | 0.9146 | 0.9197 | 0.9211 |

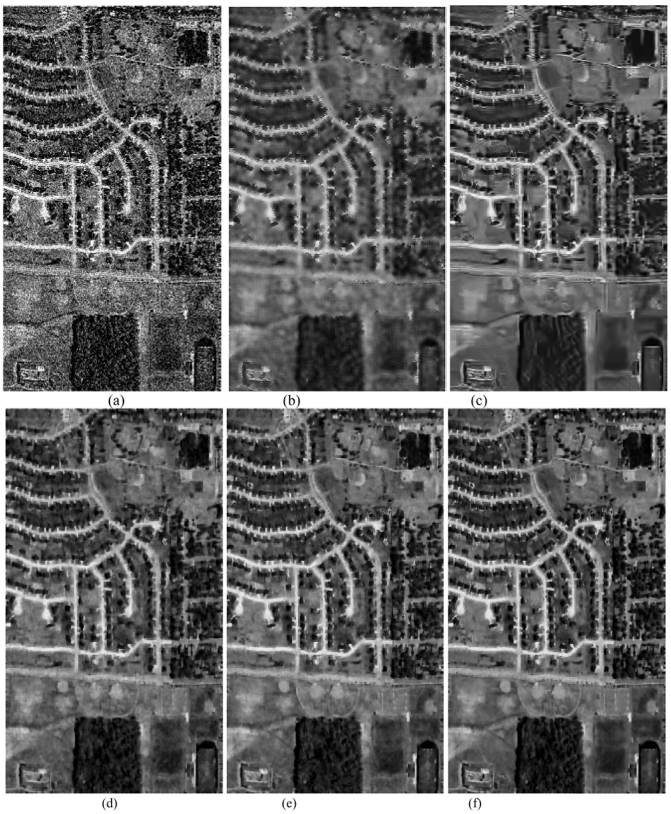


Fig. 10. Denoising results of the real-world Purdue Campus data cube. (a) Noisy band. (b)–(f) Denoising results of (a) by CCWF, PCA + WGSM, GCRF, GCRF + MLR-O, and GCRF + MLR-K.

thus the relative noisy and poor classification results. The PCA + WGSM + MLR, which deals with the whole data cube automatically, obtained much better results. It can effectively smooth the classification of homogenous regions, but show some insufficiencies to deal with image details. In contrast, it can be obviously noted that our methods, the GCRF + MLR-O, and GCRF + MLR-K, obtained more smooth classification results at the homogenous regions, while preserve much more image details.

To comprehensively evaluate our methods, we further present the quantitative evaluation through the values of OA and kappa. The classification performances are shown in Tables VII and VIII. We noted that the methods over all the denoised data cubes presented better results than that over the original data cubes. These improvements demonstrate the importance of hyperspectral image denoising. In addition, the proposed GCRF + MLR-O and GCRF + MLR-K obtained better results than other method from both the OA and

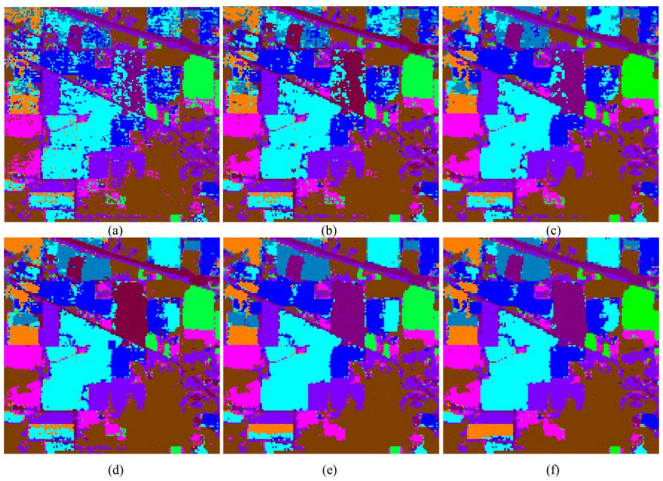


Fig. 11. Classification results over the real-world Indian Pine data cube. (a) Classification over the original data cube. (b)–(f) classification results by CCWF + MLR, PCA + WGSM + MLR, GCRF + MLR, GCRF + MLR-O, and GCRF + MLR-K.

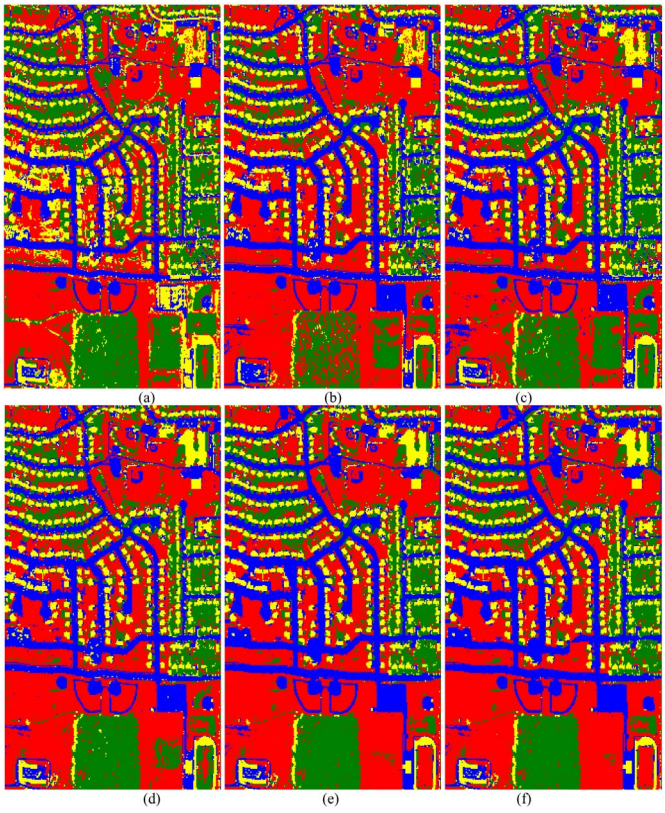


Fig. 12. Classification results over the real-world Purdue Campus data cube. (a) Classification over the original data cube. (b)–(f) classification results by CCWF + MLR, PCA + WGSM + MLR, GCRF + MLR, GCRF + MLR-O, and GCRF + MLR-K.

kappa measures. Moreover, as we can expect, the OA of GCRF + MLR-O is higher than that of GCRF + MLR-K, whereas the GCRF + MLR-K obtained better results than GCRF + MLR-O in the kappa measure.

TABLE VIII
CLASSIFICATION PERFORMANCES OF DIFFERENT METHODS OVER THE REAL-WORLD PURDUE CAMPUS DATA CUBE

| METHODS / PERFORMANCE | NOISY +MLR | CCWF +MLR | PCA+WGSM +MLR | GCRF +MLR | GCRF +MLR-O | GCRF +MLR-K |
|---|---|---|---|---|---|---|
| OA | 82.06% | 84.82% | 89.93% | 91.75% | 93.58% | 93.16% |
| KAPPA | 0.7625 | 0.7831 | 0.8502 | 0.8651 | 0.8846 | 0.8872 |

## VI. CONCLUSION

This paper has presented a new simultaneous denoising and classification method through joint learning method for the hybrid GCRF and MLR model developed under the multivariate prediction framework. With the hybrid GCRF and MLR model and developed joint learning method, this paper further proposed two simultaneous denoising and classification algorithms to optimize the linear measure, i.e., OA, and nonlinear measure, i.e., kappa statistics, respectively. The experiments on denoising and classification tasks were performed with simulated and real noise on two real-world hyperspectral data cubes, representing different environments in remote sensing. The results demonstrated that by combining the two interactive tasks, our algorithms obtained improvements over that of treating them separately.

In the current method, we investigate the relatively simple but effective MLR for the classification task. Our proposed approach can be fruitfully extended to many other hybrid systems. For future work, more complex classifiers, such as SVM, support vector regression [44], and network-based method [45] can be incorporated into our framework if only the classifiers have closed formulations. Moreover, in addition to the image denoising, the most common ability of CRF model is contextual classification. Therefore, developing a single CRF model for simultaneous denoising and classification is another interesting future work.

## ACKNOWLEDGMENT

The authors would like to thank Prof. Landgrebe for presenting the Indian Pine and Purdue Campus data cubes.

## REFERENCES

[1] W. Li, S. Prasad, J. E. Fowler, and L. M. Bruce, "Locality-preserving dimensionality reduction and classification for hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1185–1198, Apr. 2012.

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

[3] L. Zhang, L. Zhang, D. Tao, and X. Huang, "On combining multiple features for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 879–893, Mar. 2012.

[4] 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.

[5] Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Multiple spectral–spatial classification approach for hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4122–4132, Nov. 2010.

[6] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. New York, NY, USA: Wiley, 2003.
[7] M. Farzam and S. Beheshti, "Simultaneous denoising and intrinsic order selection in hyperspectral imaging," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 9, pp. 3423–3436, Sep. 2011.
[8] N. Acito, M. Diani, and G. Corsini, "Subspace-based striping noise reduction in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 4, pp. 1325–1342, Apr. 2011.
[9] G. Chen and S. Qian, "Denoising of hyperspectral imagery using principal component analysis and wavelet shrinkage," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 973–980, Mar. 2011.
[10] K. W. Jørgensen and L. K. Hansen, "Model selection for Gaussian kernel PCA denoising," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 1, pp. 163–168, Jan. 2012.
[11] P. Scheunders, "Wavelet thresholding of multivalued images," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 475–483, Apr. 2004.
[12] H. Othman and S. Qian, "Noise reduction of hyperspectral imagery using hybrid spatial-spectral derivative-domain wavelet Shrinkage," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 2, pp. 397–408, Feb. 2006.
[13] D. Letexier and S. Bourennane, "Noise removal from hyperspectral images by multidimensional filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 7, pp. 2061–2068, Jul. 2008.
[14] A. Karami, M. Yazdi, and A. Z. Asli, "Noise reduction of hyperspectral images using kernel non-negative Tucker decomposition," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 487–493, Jun. 2011.
[15] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D*, vol. 60, pp. 259–268, Nov. 1992.
[16] Q. Yuan, L. Zhang, and H. Shen, "Hyperspectral image denoising employing a spectral–spatial adaptive total variation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3660–3677, Oct. 2012.
[17] M. F. Tappen, C. Liu, E. H. Adelson, and W. T. Freeman, "Learning Gaussian conditional random fields for low-level vision," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–8.
[18] P. Zhong and R. Wang, "Multiple-spectral-band CRFs for denoising junk bands of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 2260–2275, Apr. 2013.
[19] Y. Xia, C. Sun, and W. Zheng, "Discrete-time neural network for fast solving large linear $L_1$ estimation problems and its application to image restoration," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 5, pp. 812–820, May 2012.
[20] 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.
[21] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.
[22] M. Chi and L. Bruzzone, "Semisupervised classification of hyperspectral images by SVMs optimized in the primal," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1870–1880, Jun. 2007.
[23] P. Zhong, P. Zhang, and R. Wang, "Dynamic learning of SMLR for feature selection and classification of hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 2, pp. 280–284, Apr. 2008.
[24] Q. Cheng, P. K. Varshney, and M. K. Arora, "Logistic regression for feature selection and soft classification of remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 4, pp. 491–494, Oct. 2006.
[25] D. Ashish, R. W. McClendon, and G. Hoogenboom, "Land-use classification of multispectral aerial images using artificial neural networks," *Int. J. Remote Sens.*, vol. 30, no. 8, pp. 1989–2004, 2009.
[26] 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.
[27] G. Camps-Valls, T. V. B. Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, Oct. 2007.
[28] P. Zhong and R. Wang, "Learning conditional random fields for classification of hyperspectral images," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1890–1907, Jul. 2010.
[29] S. Kawaguchi and R. Nishii, "Hyperspectral image classification by bootstrap AdaBoost with random decision stumps," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 11, pp. 3845–3851, Nov. 2007.
[30] Y. Bazi and F. Melgani, "Gaussian process approach to remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 1, pp. 186–197, Jan. 2010.
[31] G. C. Cawley, N. Talbot, and M. Girolami, "Sparse multinomial logistic regression via Bayesian $L_1$ Regularisation," in *Proc. NIPS*, 2006, pp. 209–216.
[32] J. D. F. Habbema and J. Hermans, "Selection of variables in discriminant analysis by F-statistic and error rate," *Technometrics*, vol. 19, no. 4, pp. 487–493, Nov. 1977.
[33] F. Dufrenois and J. C. Noyer, "Formulating robust linear regression estimation as a one-class LDA criterion: Discriminative hat matrix," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 2, pp. 262–273, Feb. 2013.
[34] A. Castrodad, "Graph-based denoising and classification of hyperspectral imagery using nonlocal operators," *Proc. SPIE*, vol. 7334, pp. 73340E-1–73340E-12, Apr. 2009.
[35] H. Zhang, J. Yang, Y. Zhang, N. M. Nasrabadi, and T. S. Huang, "Close the loop: Joint blind image restoration and recognition with sparse representation prior," in *Proc. ICCV*, 2011, pp. 770–777.
[36] W. Bian and D. Tao, "Constrained empirical risk minimization framework for distance metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1194–1205, Aug. 2012.
[37] J. Suzuki, E. McDermott, and H. Isozaki, "Training conditional random fields with multivariate evaluation measures," in *Proc. 21st ICCL/44th AMACL*, 2006, pp. 217–224.
[38] T. Joachims, "A support vector method for multivariate performance measures," in *Proc. 22nd ICML*, 2005, pp. 377–384.
[39] M. Jansche, "Maximum expected F-measure training of logistic regression models," in *Proc. HLT/EMNLP*, 2005, pp. 692–699.
[40] D. Musicant, V. Kumar, and A. Ozgur, "Optimizing F-measure with support vector machines," in *Proc. FLAIRS*, 2003, pp. 356–360.
[41] L. Sendur and I. W. Selesnick, "Bivariate shrinkage with local variance estimation," *IEEE Signal Process. Lett.*, vol. 9, no. 12, pp. 438–441, Dec. 2002.
[42] D. Landgrebe and L. Biehl. (2011, Sep.). *Multispec Software* [Online]. Available: http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/
[43] M. L. Uss, B. Vozel, V. V. Lukin, and K. Chehdi, "Local signal-dependent noise variance estimation from hyperspectral textural images," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 469–486, Jun. 2011.
[44] F. Bellocchio, S. Ferrari, V. Piuri, and N. A. Borghese, "Hierarchical approach for multiscale support vector regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 9, pp. 1448–1460, Sep. 2012.
[45] T. C. Silva and L. Zhao, "Network-based high level data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 6, pp. 954–970, Jun. 2012.

**Ping Zhong** (M'09) received the M.S. degree in applied mathematics and the Ph.D. degree in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2003 and 2008, respectively.

He is currently an Associate Professor with the ATR National Laboratory, National University of Defense Technology. He has published more than ten peer-reviewed articles in international journals, such as the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. His current research interests include computer vision, machine learning, and pattern recognition.

Dr. Zhong was a recipient of the National Excellent Doctoral Dissertation Award of China in 2011 and the New Century Excellent Talents in University of China in 2013. He is a Referee of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.

**Runsheng Wang** received the Degree from the Harbin Institute of Technology, Harbin, China, in 1964.

He was a Visiting Scholar with the Department of Computer Science, University of Massachusetts, Amherst, MA, USA, from 1984 to 1986 and from 1992 to 1993. He has taught at the Harbin Institute of Technology and Changsha Institute of Technology, Changsha, China. He is currently a Professor with the ATR National Laboratory, National University of Defense Technology, Changsha. His current research interests include image analysis and understanding, pattern recognition, and information fusion.