Check for
updates

Multiscale principle of relevant information for hyperspectral image classification

Yantao Wei[1] · Shujian Yu[2] · Luis Sanchez Giraldo[3] · José C. Príncipe[2]

Abstract

This paper proposes a novel architecture, termed multiscale principle of relevant information (MPRI), to learn discriminative spectral-spatial features for hyperspectral image classification. MPRI inherits the merits of the principle of relevant information (PRI) to effectively extract multiscale information embedded in the given data, and also takes advantage of the multilayer structure to learn representations in a coarse-to-fine manner. Specifically, MPRI performs spectral-spatial pixel characterization (using PRI) and feature dimensionality reduction (using regularized linear discriminant analysis) iteratively and successively. Extensive experiments on three benchmark data sets demonstrate that MPRI outperforms existing state-of-the-art methods (including deep learning based ones) qualitatively and quantitatively, especially in the scenario of limited training samples. Code of MPRI is available at http://bit.ly/MPRI_HSI.

Keywords Hyperspectral image classification · Principle of relevant information · Spectral-spatial pixel characterization

---

✉ Shujian Yu
yusjlcy9011@ufl.edu

Yantao Wei
yantaowei@mail.ccnu.edu.cn

Luis Sanchez Giraldo
luis.sanchez@uky.edu

José C. Príncipe
principe@cnel.ufl.edu

[1] Hubei Research Center for Educational Informationization, Central China Normal University, Wuhan 430079, China

[2] Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA

[3] Department of Electrical and Computer Engineering, University of Kentucky, Lexington, KY 40508, USA

1 Introduction

With the rapid development of hyperspectral imaging techniques, current sensors always have high spectral and spatial resolution (He et al., 2018). For example, the ROSIS sensor can cover spectral resolution higher than 10 nm, reaching 1 m per pixel spatial resolution (Cao et al., 2018; Zare et al., 2018). The increased spectral and spatial resolution enables us to accurately discriminate diverse materials of interest. As a result, hyperspectral images (HSIs) have been widely used in many practical applications, such as precision agriculture, environmental management, mining and mineralogy (He et al., 2018). Among them, HSI classification, which aims to assign each pixel of HSI to a unique class label, has attracted increasing attention in recent years. However, the unfortunate combination of high-dimensional spectral features and the limited ground truth samples, as well as different atmospheric scattering conditions, make the HSI data inherently highly nonlinear and difficult to be categorized (Ghamisi et al., 2017).

Early HSI classification methods straightforwardly apply conventional dimensionality reduction techniques, such as the principal component analysis (PCA) and the linear discriminant analysis (LDA), on spectral domain to learn discriminative spectral features. Although these methods are conceptually simple and easy to implement, they neglect the spatial information, a complement to spectral behavior that has been demonstrated effective to augment HSI classification performance (He et al., 2018; Ghamisi et al., 2015). To address this limitation, Chen et al. 2011) proposed the joint sparse representation (JSR) to incorporate spatial neighborhood information of pixels. Soltani-Farani et al. (2015) designed spatial aware dictionary learning (SADL) by using a structured dictionary learning model to incorporate both spectral and spatial information. Kang et al., suggested using an edge-preserving filter (EPF) to improve the spatial structure of HSI (Kang et al., 2014) and also introduced PCA to encourage the separability of new representations (Kang et al., 2017). A similar idea appears in Pan et al. (2017), in which EPF is substituted with a hierarchical guidance filter. Although these methods perform well, the discriminative power of their extracted spectral-spatial features is far from satisfactory when being tested on challenging land covers.

A recent trend is to use deep neural networks (DNN), such as autoencoders (AE) (Ma et al., 2016) and convolutional neural networks (CNN) (Chen et al., 2016), to learn discriminative spectral-spatial features (Zhong et al., 2018). Although deep features always demonstrate superior discriminative power than hand-crafted features in different computer vision or image processing tasks, existing DNN based HSI classification methods either improve the performance marginally or require significantly more labeled data (Yang et al., 2018). On the other hand, collecting labeled data is always difficult and expensive in remote sensing community (Zare et al., 2018). Admittedly, transfer learning has the potential to alleviate the problem of limited labeled data, it still remains an open problem to construct a reliable relevance between the target domain and the source domain due to the large variations between HSIs obtained by different sensors with unmatched imaging bands and resolutions (Zhu et al., 2017).

Different from previous work, this paper presents a novel architecture, termed multiscale principle of relevant information (MPRI), to learn discriminative spectral-spatial features for HSI classification. MPRI inherits the merits of the principle of relevant information (PRI) (Chapter 8, Principe 2010) (Chapter 3, Rao 2008) to effectively extract multiscale information from given data, and also takes advantage of the

multilayer structure to learn representations in a coarse-to-fine manner. To summarize, the major contributions of this work are threefold.

- We demonstrate the capability of PRI, originated from the information theoretic learning (ITL) (Principe 2010), to characterize 3D pictorial structures in HSI data.
- We generalize PRI into a multilayer structure to extract hierarchical representations for HSI classification. A multiscale scheme is also incorporated to model both local and global structures.
- MPRI outperforms state-of-the-art HSI classification methods based on classical machine learning models (e.g., PCA-EPF Kang et al., 2017 and HIFI Pan et al., 2017) by a large margin. Using significantly fewer labeled data, MPRI also achieves almost the same classification accuracy compared to existing deep learning techniques (e.g., SAE-LR Chen et al., 2014 and 3D-CNN Li et al., 2017).

The remainder of this paper is organized as follows. Section 2 reviews the basic objective of PRI and formulates PRI under the ITL framework. The architecture and optimization of our proposed MPRI is elaborated in Sect. 3. Section 4 shows experimental results on three popular HSI data sets. Finally, Sect. 5 draws the conclusion.

2 Elements of Renyi's α-entropy and the principle of relevant information

Before presenting our method, we start with a brief review of the general idea and the objective of PRI, and then formulate this objective under the ITL framework.

2.1 PRI: the general idea and its objective

Suppose we are given a random variable $\mathbf{X}$ with a known probability density function (PDF) g, from which we want to learn a reduced statistical representation characterized by a random variable $\mathbf{Y}$ with PDF f. The PRI (Chapter 8, Principe, 2010) (Chapter 3, Rao, 2008) casts this problem as a trade-off between the entropy $H(f)$ of $\mathbf{Y}$ and its descriptive power about $\mathbf{X}$ in terms of their divergence $D(f\|g)$. Therefore, for a fixed PDF g, the objective of PRI is given by:

$$\underset{f}{\text{minimize}}\ H(f) + \beta D(f\|g), \tag{1}$$

where β is a hyper-parameter controlling the amount of relevant information that $\mathbf{Y}$ can extract from $\mathbf{X}$. Note that, the minimization of entropy can be viewed as a means of finding the statistical regularities in the outcomes of a process, whereas the minimization of information theoretic divergence, such as the Kullback-Leibler divergence (Kullback & Leibler, 1951) or the Chernoff divergence (Chernoff, 1952), ensuring that the regularities are closely related to $\mathbf{X}$. The PRI is similar in spirit to the Information Bottleneck (IB) method (Tishby et al., 2000), but the formulation is different because PRI does not require an observed relevant (or auxillary) variable and the optimization is done directly on the random variable $\mathbf{X}$, which provides a set of solutions that are related to the principal curves (Hastie & Stuetzle, 1989) of g, as will be demonstrated below.

2.2 Formulation of PRI using Renyi's entropy functional

In information theory, a natural extension of the well-known Shannon's entropy is the Renyi's α-entropy (Rényi et al., 1961). For a random variable $\mathbf{X}$ with PDF $f(x)$ in a finite set $\mathcal{X}$, the α-entropy of $H(\mathbf{X})$ is defined as:

$$H_\alpha(f) = \frac{1}{1-\alpha} \log \int_{\mathcal{X}} f^\alpha(x)dx. \tag{2}$$

On the other hand, motivated by the famed Cauchy–Schwarz (CS) inequality:

$$\left| \int f(x)g(x)dx \right|^2 \leq \int |f(x)|^2 \, dx \int |g(x)|^2 \, dx, \tag{3}$$

with equality if and only if $f(x)$ and $g(x)$ are linearly dependent (e.g., $f(x)$ is just a scaled version of $g(x)$), a measure of the "distance" between the PDFs can be defined, which was named the CS divergence (Jenssen et al., 2006), with:

$$\begin{aligned} D_{cs}(f\|g) &= -\log\left(\int fg\right)^2 + \log\left(\int f^2\right) + \log\left(\int g^2\right) \\ &= 2H_2(f;g) - H_2(f) - H_2(g), \end{aligned} \tag{4}$$

the term $H_2(f;g) = -\log \int f(x)g(x)dx$ is also called the quadratic cross entropy (Principe, 2010).

Combining Eqs. (2) and (4), the PRI under the 2-order Renyi's entropy can be formulated as:

$$\begin{aligned} f_{opt} &= \arg\min_f H_2(f) + \beta\big(2H_2(f;g) - H_2(f) - H_2(g)\big) \\ &\equiv \arg\min_f (1-\beta)H_2(f) + 2\beta H_2(f;g), \end{aligned} \tag{5}$$

the second equation holds because the extra term $\beta H_2(g)$ is a constant with respect to f.

Given $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ and $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N$, both in $\mathbb{R}^p$, drawn i.i.d. from g and f, respectively. Using the Parzen-window density estimation (Parzen, 1962) with Gaussian kernel $G_\delta(\cdot) = \exp(-\frac{\|\cdot\|^2}{2\delta^2})$, Eq. (5) can be simplified as Rao (2008):

$$\begin{aligned} \mathbf{Y}_{opt} = \arg\min_{\mathbf{Y}} \Bigg[&-(1-\beta)\log\left(\frac{1}{N^2}\sum_{i,j=1}^N G_\delta(\mathbf{y}_i - \mathbf{y}_j)\right) \\ &-2\beta\log\left(\frac{1}{N^2}\sum_{i,j=1}^N G_\delta(\mathbf{y}_i - \mathbf{x}_j)\right) \Bigg]. \end{aligned} \tag{6}$$

It turns out that the value of β defines various levels of information reduction, ranging from data mean value ($\beta = 0$), clustering ($\beta = 1$), principal curves (Hastie and Stuetzle, 1989) extraction at different dimensions, and vector quantization obtaining back the initial data when $\beta \to \infty$ (Principe, 2010; Rao, 2008). Hence, the PRI achieves similar effects to a moment decomposition of the PDF controlled by a single parameter β, using a data driven optimization approach. See Fig. 1 for an example. From this figure we can see that the self organizing decomposition provides a set of hierarchical features of the input data beyond cluster centers, that may yield more robust features. Note that, despite its
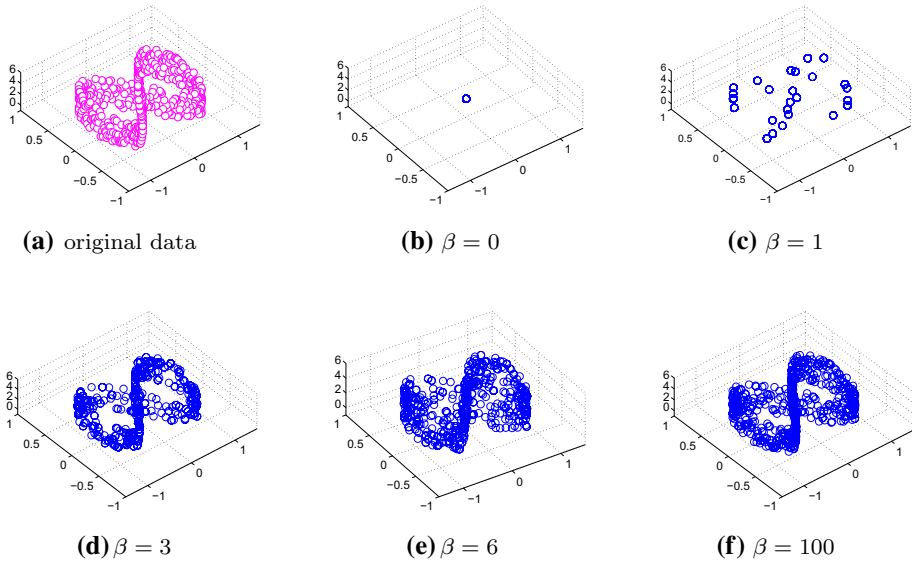
Fig. 1 Illustration of the structures revealed by the PRI for **a** Intersect data set. As the values of β increase the solution passes through **b** a single point, **c** modes, **d** and **e** principal curves at different dimensions, and in the extreme case of **f** $\beta \to \infty$ we get back the data themselves as the solution

strategic flexibility to find reduced structure of given data, the PRI is mostly unknown to practitioners.

3 Multiscale principle of relevant information (MPRI) for HSI classification

In this section, we present MPRI for HSI classification. MPRI stacks multiple spectral-spatial feature learning units, in which each unit consists of multiscale PRI and a regularized LDA (Bandos et al., 2009). The architecture of MPRI is shown in Fig. 2.
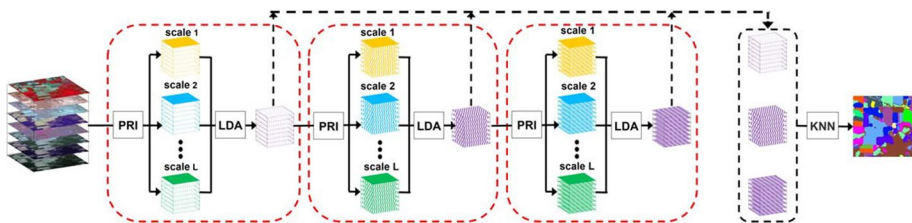

Fig. 2 The architecture of multiscale principle of relevant information (MPRI) for HSI classification. The spectral-spatial feature learning unit is marked with red dashed rectangle. The spectral-spatial features are extracted by performing PRI (in multiple scales) and LDA iteratively and successively on HSI data cube (after normalization). Finally, features from each unit are concatenated and fed into a *k*-nearest neighbors (KNN) classifier to predict pixel labels. This plot only demonstrates a 3-layer MPRI, but the number of layers can be increased or decreased flexibly

To the best of our knowledge, apart from performing band selection (e.g., Feng et al., 2015; Yu et al., 2019) or measuring spectral variability (e.g., Chang, 2000), information theoretic principles have seldom been investigated to learn discriminative spectral-spatial features for HSI classification. The most similar work to ours is Kamandar and Ghassemian (2013), in which the authors use the criterion of minimum redundancy maximum relevance (MRMR) (Peng et al., 2005) to extract linear features. However, owing to the poor approximation to estimate multivariate mutual information, the performance of Kamandar and Ghassemian (2013) is only slightly better than the basic linear discriminant analysis (LDA) (Du, 2007).

3.1 Spectral-spatial feature learning unit

Let $\mathbf{T} \in \mathbb{R}^{m \times n \times d}$ be the raw 3D HSI data cube, where m and n are the spatial dimensions, d is the number of spectral bands. For a target spectral vector $\mathbf{t}_\star \in \mathbb{R}^d$, we extract a local cube (denote $\hat{\mathbf{X}}$) from $\mathbf{T}$ using a sliding window of width $\hat{n}$ centered at $\mathbf{t}_\star$, i.e., $\hat{\mathbf{X}} = \{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \cdots, \hat{\mathbf{x}}_{\hat{N}}\} \in \mathbb{R}^{\hat{N} \times d}$, $\hat{n} \times \hat{n} = \hat{N}$, and $\mathbf{t}_\star = \hat{\mathbf{x}}_{\lfloor \hat{n}/2 \rfloor + 1, \lfloor \hat{n}/2 \rfloor + 1}$, where $\lfloor \cdot \rfloor$ is the nearest integer function. We obtain the spectral-spatial characterization $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \cdots, \hat{\mathbf{y}}_{\hat{N}}\} \in \mathbb{R}^{\hat{N} \times d}$ from $\hat{\mathbf{X}}$ using PRI via the following objective:

$$\text{minimize}_{\hat{\mathbf{Y}}} \left[-(1-\beta) \log \left(\frac{1}{\hat{N}^2} \sum_{i,j=1}^{\hat{N}} G_\delta(\hat{\mathbf{y}}_i - \hat{\mathbf{y}}_j) \right) - 2\beta \log \frac{1}{\hat{N}^2} \sum_{i,j=1}^{\hat{N}} G_\delta(\hat{\mathbf{y}}_i - \hat{\mathbf{x}}_j) \right]. \quad (7)$$

We finally use the center vector of $\hat{\mathbf{Y}}$, i.e., $\hat{\mathbf{y}}_{\lfloor \hat{n}/2 \rfloor + 1, \lfloor \hat{n}/2 \rfloor + 1}$, as the new representation of $\mathbf{t}_\star$. We scan the whole 3D cube with a sliding window of width $\hat{n}$ targeted at each pixel to get the new spectral-spatial representation. The procedure is depicted in Fig. 3.
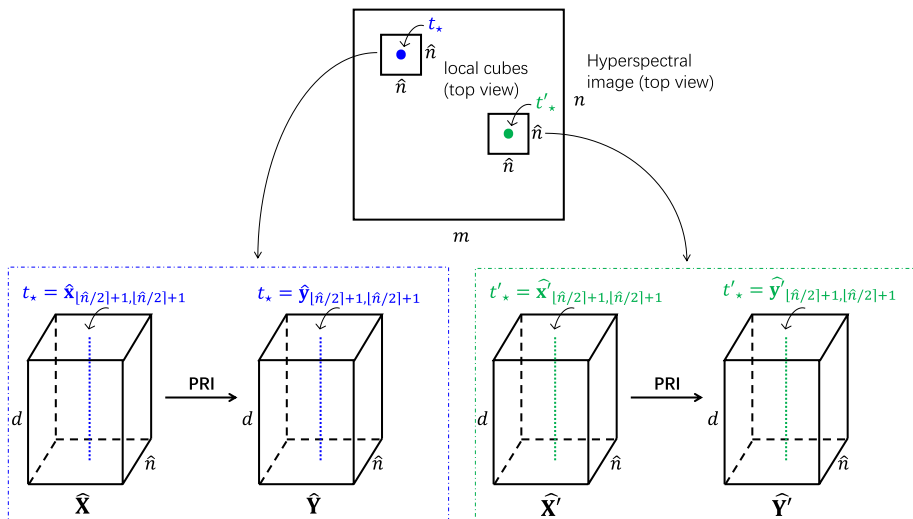

Fig. 3 For each target spectral vector (e.g., $\mathbf{t}_\star$ or $\mathbf{t}'_\star$) in the raw hyperspectral image, we obtain a new vector representation by performing PRI in its corresponding local data cube (e.g., $\hat{\mathbf{X}}$ or $\hat{\mathbf{X}}'$)

Equation (7) is updated iteratively. Specifically, denote $V(\hat{\mathbf{Y}}) = \frac{1}{\hat{N}^2} \sum_{i,j=1}^{\hat{N}} G_\delta(\hat{\mathbf{y}}_i - \hat{\mathbf{y}}_j)$ and $V(\hat{\mathbf{Y}};\hat{\mathbf{X}}) = \frac{1}{\hat{N}^2} \sum_{i,j=1}^{\hat{N}} G_\delta(\hat{\mathbf{y}}_j - \hat{\mathbf{x}}_i)$, taking the derivative of Eq. (7) with respect to $\hat{\mathbf{y}}_\star$ and equating to zero, we have:

$$\frac{1-\beta}{V(\hat{\mathbf{Y}})} \sum_{j=1}^{\hat{N}} G_\delta(\hat{\mathbf{y}}_\star - \hat{\mathbf{y}}_j) \left\{ \frac{\hat{\mathbf{y}}_j - \hat{\mathbf{y}}_\star}{\delta^2} \right\} + \frac{\beta}{V(\hat{\mathbf{Y}};\hat{\mathbf{X}})} \sum_{j=1}^{\hat{N}} G_\delta(\hat{\mathbf{y}}_\star - \hat{\mathbf{x}}_j) \left\{ \frac{\hat{\mathbf{x}}_j - \hat{\mathbf{y}}_\star}{\delta^2} \right\} = 0. \quad (8)$$

Rearrange Eq. (8), we have:

$$\left\{ \frac{\beta}{V(\hat{\mathbf{Y}};\hat{\mathbf{X}})} \sum_{j=1}^{\hat{N}} G_\delta(\hat{\mathbf{y}}_\star - \hat{\mathbf{x}}_j) \right\} \hat{\mathbf{y}}_\star = \frac{1-\beta}{V(\hat{\mathbf{Y}})} \sum_{j=1}^{\hat{N}} G_\delta(\hat{\mathbf{y}}_\star - \hat{\mathbf{y}}_j)\hat{\mathbf{y}}_j$$

$$- \left\{ \frac{1-\beta}{V(\hat{\mathbf{Y}})} \sum_{j=1}^{\hat{N}} G_\delta(\hat{\mathbf{y}}_\star - \hat{\mathbf{y}}_j) \right\} \hat{\mathbf{y}}_\star \qquad (9)$$

$$+ \frac{\beta}{V(\hat{\mathbf{Y}};\hat{\mathbf{X}})} \sum_{j=1}^{\hat{N}} G_\delta(\hat{\mathbf{y}}_\star - \hat{\mathbf{x}}_j)\hat{\mathbf{x}}_j.$$

Divide both sides of the Eq. (9) by

$$\frac{\beta}{V(\hat{\mathbf{Y}};\hat{\mathbf{X}})} \sum_{j=1}^{\hat{N}} G_\delta(\hat{\mathbf{y}}_\star - \hat{\mathbf{x}}_j), \qquad (10)$$

and let

$$c = V(\hat{\mathbf{Y}};\hat{\mathbf{X}})/V(\hat{\mathbf{Y}}), \qquad (11)$$

we obtain the fixed point update rule for $\hat{\mathbf{y}}_\star$:

$$\hat{\mathbf{y}}_\star^{\tau+1} = c\frac{1-\beta}{\beta} \frac{\sum_{j=1}^{\hat{N}} G_\delta\left(\hat{\mathbf{y}}_\star^\tau - \hat{\mathbf{y}}_j^\tau\right)\hat{\mathbf{y}}_j^\tau}{\sum_{j=1}^{\hat{N}} G_\delta\left(\hat{\mathbf{y}}_\star^\tau - \hat{\mathbf{x}}_j\right)} - c\frac{1-\beta}{\beta} \frac{\sum_{j=1}^{\hat{N}} G_\delta\left(\hat{\mathbf{y}}_\star^\tau - \hat{\mathbf{y}}_j^\tau\right)}{\sum_{j=1}^{\hat{N}} G_\delta(\hat{\mathbf{y}}_\star^\tau - \hat{\mathbf{x}}_j)}\hat{\mathbf{y}}_\star^\tau$$

$$+ \frac{\sum_{j=1}^{\hat{N}} G_\delta\left(\hat{\mathbf{y}}_\star^\tau - \hat{\mathbf{x}}_j\right)\hat{\mathbf{x}}_j}{\sum_{j=1}^{\hat{N}} G_\delta\left(\hat{\mathbf{y}}_\star^\tau - \hat{\mathbf{x}}_j\right)}, \qquad (12)$$

where τ is the iteration number. We move the sliding window pixel by pixel, and only update the representation of the center target pixel, as shown in Fig. 3.

We also introduce two modifications to increase the discriminative power of the new representation. First, different values of $\hat{n}$ (3, 5, 7, 9, 11, 13 in this work) are used to model both local and global structures. Second, to reduce the redundancy of raw features constructed by concatenating PRI representations in multiple scales, we further perform a regularized LDA (Bandos et al., 2009).

Note that, the hyper-parameter β and different values of $\hat{n}$ play different roles in MPRI. Specifically, β in PRI balances the trade-off between the regularity of extracted

representation and its discriminative power to the given data. Therefore, it should be set in a reasonable range to avoid over-smoothing effect of the resulting image and unsatisfactory classification performance. A deeper discussion is shown in Sect. 4.1.1. By contrast, $\hat{n}$ controls the spatial scale of the learned representation. The motivation is that the discriminative information of different categories may not be easily characterized by a sliding window of a fixed size (i.e., $\hat{n} \times \hat{n}$). Thus, it would be favorable if one can incorporate discriminative information from different scales, by changing the value of $\hat{n}$.

3.2 Stacking multiple units

In order to characterize spectral-spatial structures in a coarse-to-fine manner, we stack multiple spectral-spatial feature learning units described in Sect. 3.1 to constitute a multilayer structure and concatenate representations from each layer to form the final spectral-spatial representation. We finally feed this representation into a standard k-nearest neighbors (KNN) for classification.

Different from existing DNNs that are typically trained with error backpropagation or the combination of a greedy layer-wise pretraining and a fine-tuning stage, our multilayer structure is trained successively from bottom layer to top layer without error backpropagation. For the ith layer, the input of PRI is the representation learned from the previous layer (denoted T_{i-1}). We then learn new representation T_i by iteratively updates T_{i-1} with Eq. (12) and a dimensionality reduction step with LDA at the end of iteration. As for the multiscale PRI, it can be trained in parallel with respect to different sliding window sizes.

The interpretation of DNN as a way of creating successively better representations of the data has already been suggested and explored by many (e.g., Achille and Soatto, 2018). Most recently, Schwartz-Ziv and Tishby Shwartz-Ziv and Tishby (2017) put forth an interpretation of DNN as creating sufficient representations of the data that are increasingly minimal. For our deep architecture, in order to have an intuitive understanding to its inner mechanism, we plot the 2D projection (after 1000 t-SNE Maaten and Hinton, 2008 iterations) of features learned from different layers in Fig. 4. Similar to DNN, MPRI creates successively more faithful and separable representations in deeper layers. Moreover, the deeper features can discriminate the with-in class samples in different geography regions, even though we do not manually incorporate geographic information in the training.

4 Experimental results

We conduct three groups of experiments to demonstrate the effectiveness and superiority of the MPRI. Specifically, we first perform a simple test to determine a reliable range for the value of β in PRI and the number of layers in MPRI. Then, we implement MPRI and several of its degraded variants to analyze and evaluate component-wise contributions to performance gain. Finally, we evaluate MPRI against state-of-the-art methods on benchmark data sets using both visual and qualitative evaluations.

Three popular data sets, namely the Indian Pines (Baumgardner et al., 2015), the Pavia University and the Pavia Center, are selected in this work. We summarize the properties of each data set in Table 1.

1. The first image, displayed in Fig. 5a, is called Indian Pines. It was gathered by the airborne visible/infrared imaging spectrometer (AVIRIS) sensor over the agricultural
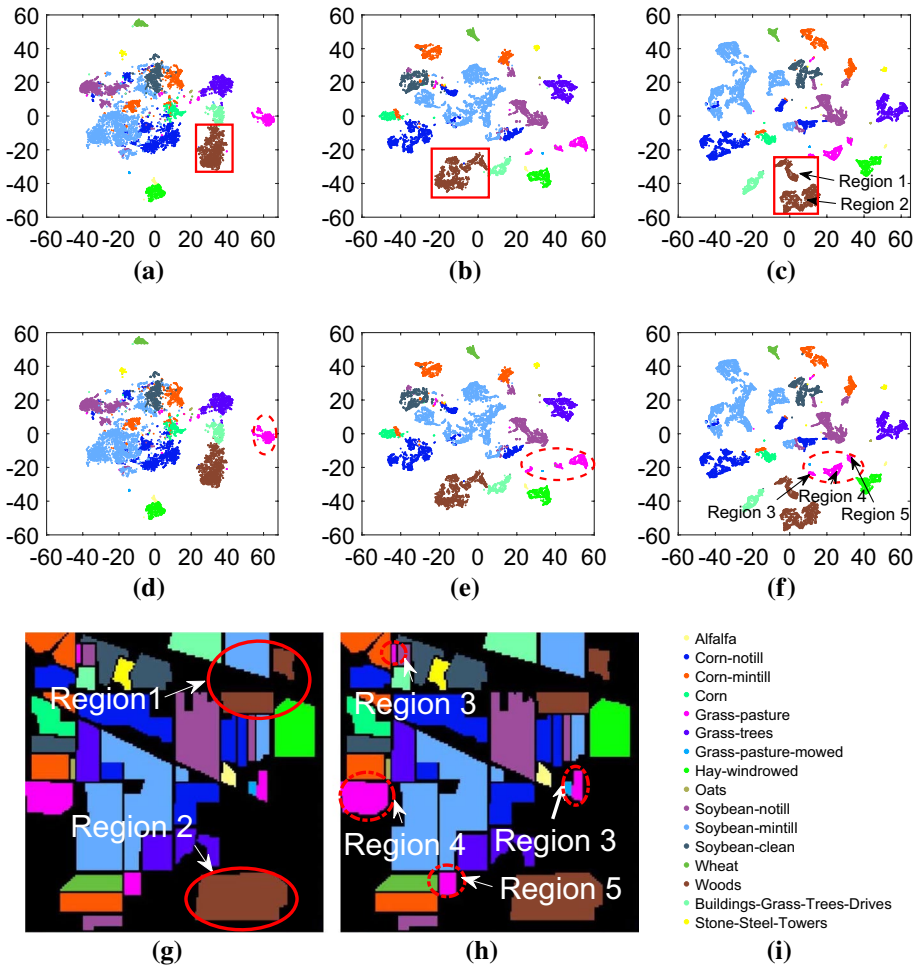
Fig. 4 2D projection of features learned by MPRI in different layers on Indian Pines data set. Features of "Woods" in the 1st layer, the 2nd layer, and the 3rd layer are marked with red rectangle in **a**–**c**. Similarly, features of "Grass-pasture" are marked with magenta ellipses in **d**–**f**. **g** The locations of "Region 1" and "Region 2". **h** shows the locations of "Region 3", "Region 4" and "Region 5". **i** shows class legend

Table 1 Details of data sets

Data set	Indian Pines	Pavia University	Pavia center
Sensor	AVIRIS	ROSIS	ROSIS-3
Spatial size	145×145	610×340	1096×492
♯ bands (used)	200	103	102
♯ classes	16	9	9

Indian Pines test site in northwestern Indiana, United States. The size of this image is 145×145 pixels with spatial resolution of 20 m. The low spatial resolution leads to the presence of highly mixed pixels (Ghamisi et al., 2014). A three-band false color image
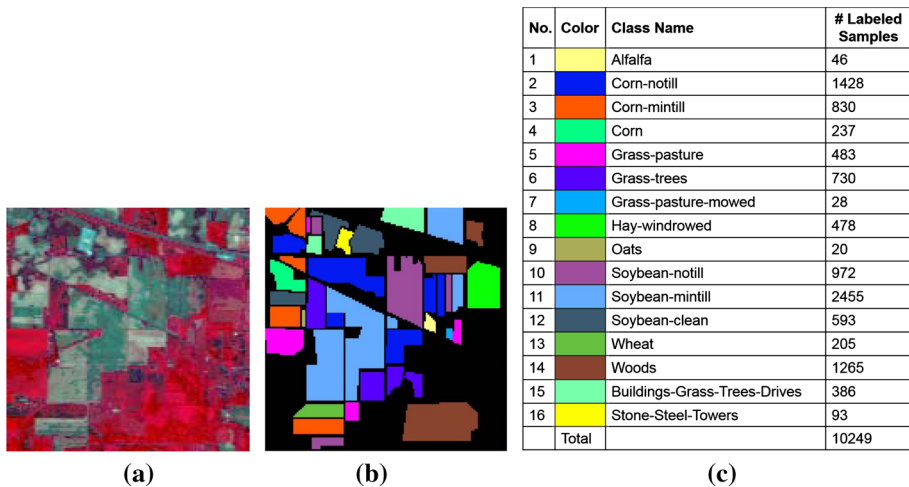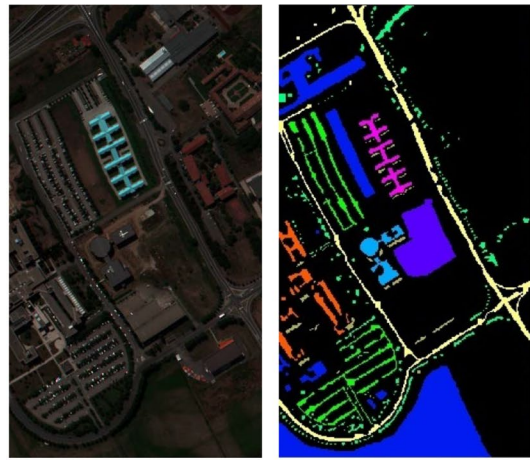
No.	Color	Class Name	# Labeled Samples
1		Alfalfa	46
2		Corn-notill	1428
3		Corn-mintill	830
4		Corn	237
5		Grass-pasture	483
6		Grass-trees	730
7		Grass-pasture-mowed	28
8		Hay-windrowed	478
9		Oats	20
10		Soybean-notill	972
11		Soybean-mintill	2455
12		Soybean-clean	593
13		Wheat	205
14		Woods	1265
15		Buildings-Grass-Trees-Drives	386
16		Stone-Steel-Towers	93
	Total		10249

(a)　　　　　**(b)**　　　　　**(c)**

Fig. 5 **a** False color composition of the AVIRIS Indian Pines scene. **b** Reference map containing 16 mutually exclusive land-cover classes. **c** The numbers of the labeled samples

and the ground-truth map are shown in Fig. 5a, b, where there are 16 classes of interest. And the name and quantity of each class are reported in Fig. 5c. The number of bands has been reduced to 200 by removing 20 bands covering the region of water absorption. This scene constitutes a challenging classification problem due to the significant presence of mixed pixels in all available classes and the unbalanced number of available labeled pixels per class (Li et al., 2013).

2. The second image is the Pavia University, which was recorded by the reflective optics spectrographic imaging system (ROSIS) sensor during a flight campaign over Pavia, northern Italy. This scene has 610×340 pixels with a spatial resolution of 1.3 m (covering the wavelength range from 0.4 to 0.9µ m). There are 9 ground-truth classes, including trees, asphalt, bitumen, gravel, metal sheet, shadow, bricks, meadow, and soil. In our experiments, 12 noisy bands have been removed and finally 103 out of the 115 bands were used. The class descriptions and sample distributions for this image are given in Fig. 6c. As can be seen, the total number of labeled samples in this image is 43,923. A three-band false color image and the ground-truth map are also shown in Fig. 6.

3. The third data set is Pavia Center. It was acquired by ROSIS-3 sensor in 2003, with a spatial resolution of 1.3m and 102 spectral bands (some bands have been removed due to noise). A three-band false color image and the ground-truth map are also shown in Fig. 7a, b. The number of ground truth classes is 9 (see Fig. 7) and it consists of 1096×492 pixels. The number of samples of each class ranges from 2108 to 65,278 (Fig. 7e). There are 5536 training samples and 98,015 testing samples (Fig. 7c, d). Note that these training samples are out of the testing samples.

Three metrics are used for quantitative evaluation (Cao et al., 2018): overall accuracy (OA), average accuracy (AA) and the kappa coefficient κ. OA is computed as the percentage of correctly classified test pixels, AA is the mean of the percentage of correctly classified pixels for each class, and κ involves both omission and commission errors and gives a good representation of the the overall performance of the classifier.

Fig. 6 **a** False color composition of the Pavia University scene. **b** Reference map containing 9 mutually exclusive land-cover classes. **c** The numbers of the labeled samples

(a) (b)

No.	Color	Class Name	# Labeled Samples
1		Asphalt	6631
2		Meadows	18649
3		Gravel	2099
4		Trees	3064
5		Painted metal sheets	1345
6		Bare Soil	5029
7		Bitumen	1330
8		Self-Blocking Bricks	3682
9		Shadows	947
	Total		42776

(c)

For our method, the values of the kernel width δ in PRI were tuned around the multivariate Silverman's rule-of-thumb (Silverman, 1986): $(\frac{4}{d+2})^{\frac{1}{d+4}} s^{\frac{-1}{4+d}} \sigma_1 \leq \delta \leq (\frac{4}{d+2})^{\frac{1}{d+4}} s^{\frac{-1}{4+d}} \sigma_2$, where s is the sample size, d is the variable dimensionality, σ_1 and σ_2 are respectively the smallest and the largest standard deviation among each dimension of the variable. For example, in Indian Pines data set, the estimated range in the 5th layer corresponds to [0.05, 0.51], and we set kernel width to 0.4. On the other hand, the PRI in each layer is optimized with $\tau = 3$ iterations, which has been observed to be sufficient to provide desirable performance.

4.1 Parameter analysis

4.1.1 Effects of parameter β in PRI

The parameter β in PRI balances the trade-off between the regularity of extracted representation and its discriminative power to the given data. We illustrate the values of OA, AA, and κ for MPRI with respect to different values of β in Fig. 8a. As can be seen, these quantitative values are initially stable, but decrease when $\beta \geq 3$. Moreover, the
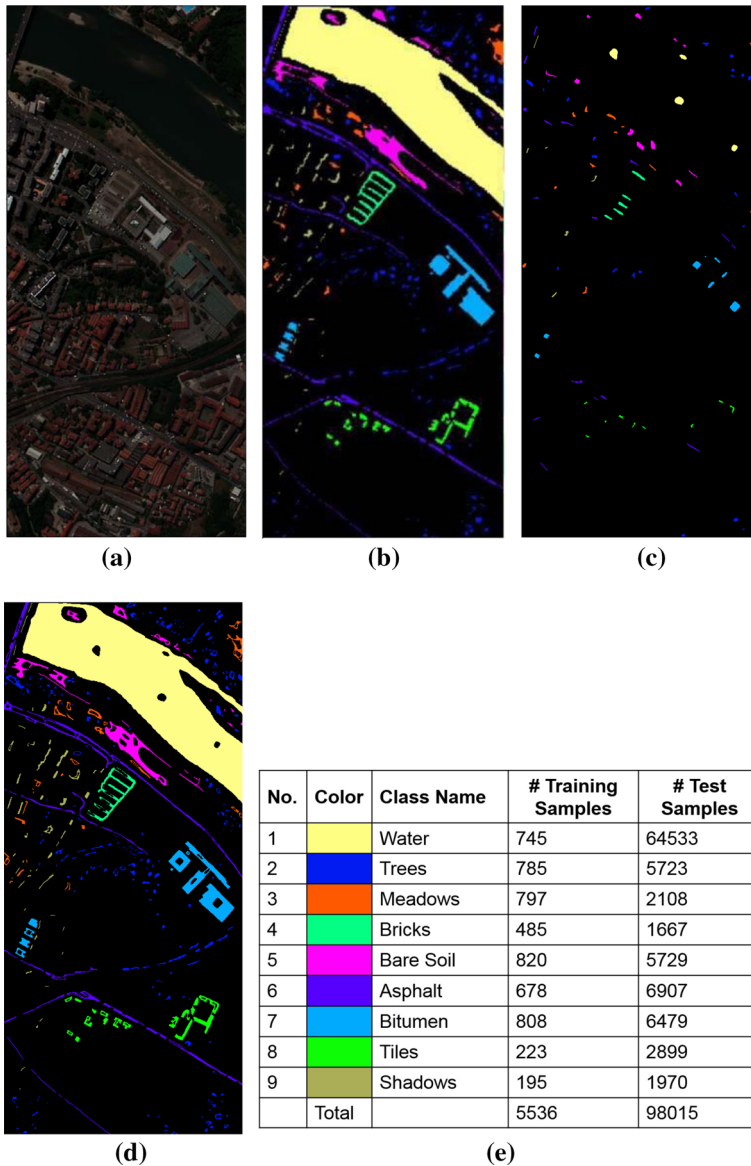
Fig. 7 **a** False color composition of the Pavia Center scene. **b** Reference map containing 9 mutually exclusive land-cover classes. **c** Training samples. **d** Testing samples. **e** The numbers of the labeled samples

No.	Color	Class Name	# Training Samples	# Test Samples
1		Water	745	64533
2		Trees	785	5723
3		Meadows	797	2108
4		Bricks	485	1667
5		Bare Soil	820	5729
6		Asphalt	678	6907
7		Bitumen	808	6479
8		Tiles	223	2899
9		Shadows	195	1970
	Total		5536	98015

value of AA drops more drastically than that of OA or κ. A likely interpretation is that when training samples are limited, many classes have only a few labeled samples (~ 1 for minority classes, such as Oats, Grass-pasture-mowed, and Alfalfa). An unreasonable value of β may severely influence the classification accuracy in these classes, hereby decreasing AA at first.

Fig. 8 a Quantitative evaluation with different values of β. **b** Quantitative evaluation with different number of layers

(a)

(b)
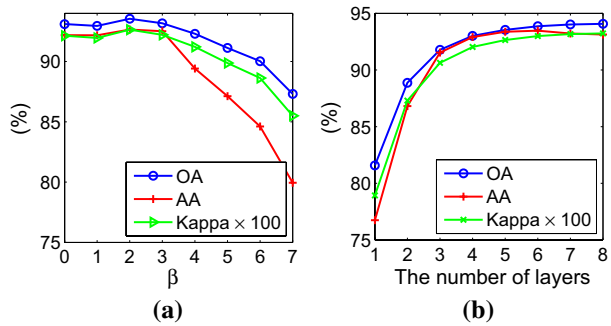
The corresponding classification maps are shown in Fig. 9. It is obviously that, the smaller the β, the more smooth results achieved by MPRI. This is because large β encourages a small divergence between the extracted representation and the original HSI data. For example, in the scenario of $\beta = 0$, PRI clusters both spectral and spatial structures into a single point (the data mean) that has no discriminative power. By contrast, in the scenario of $\beta \rightarrow \infty$, the extracted representation gets back to the HSI data itself (to minimize their divergence) such that PRI will fit all noisy and irregular structures.

From the above analysis, extremely large and small values of β are not interesting for classification of HSI. Moreover, the results also suggest that $\beta \in [2, 4]$ is able to balance a good trade-off between preserving relevant spatial information (such as edges in classification maps) and filtering out unnecessary one. Unless otherwise specified, the PRI mentioned in the following experiments uses three different values of β, i.e., $\beta = 2$, $\beta = 3$, and $\beta = 4$. The final representation of PRI is formed by concatenating representations obtained from each β.
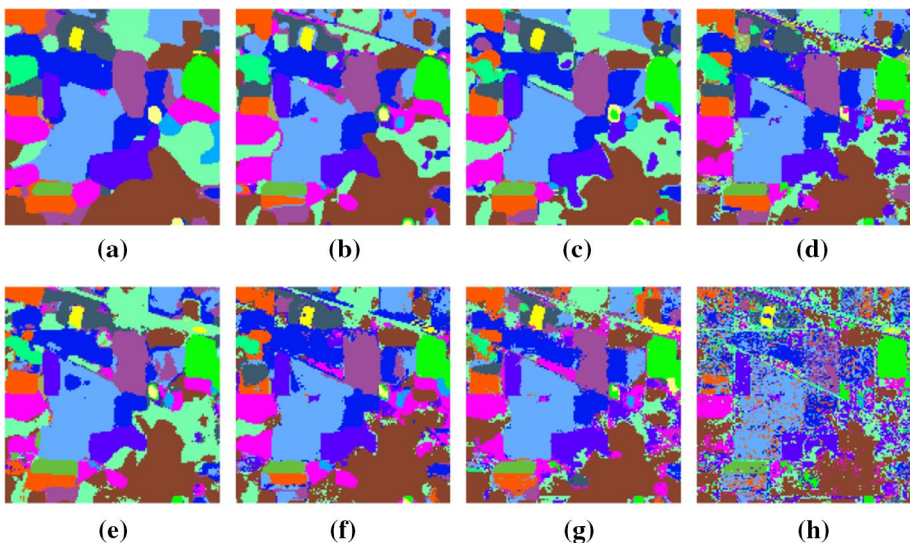

(a)   (b)   (c)   (d)

(e)   (f)   (g)   (h)

Fig. 9 Classification maps of MPRI with **a** $\beta = 0$; **b** $\beta = 1$; **c** $\beta = 2$; **d** $\beta = 3$; (e) $\beta = 4$; **f** $\beta = 5$; **g** $\beta = 6$; and **h** $\beta = 100$

4.1.2 Effects of the number of layers

We then illustrate the values of OA, AA and κ for MPRI with respect to different number of layers in Fig. 8b. The corresponding classification maps are shown in Fig. 10. Similar to existing deep architectures, stacking more layers (in a reasonable range) can increase performance. If we keep the input data size the same, more layers (beyond a certain layer number) will not increase the performance anymore and the classification maps become over-smooth. This work uses a 5-layer MPRI because it provides favorable visual and quantitative results.

4.1.3 Effects of the classifier

MPRI uses the basic KNN for classification and sets $k = 1$ throughout this work. The purpose is to validate the discriminative power of the spectral-spatial features extracted by multiple layers of PRI. To further confirm the superiority of our MPRI is independent to the used classifier, we also evaluate the performances of MPRI and other three feature extraction methods for HSI classification (EPF Kang et al., 2014, SADL Soltani-Farani et al., 2015, PCA-EPF Kang et al., 2017) and use a kernel SVM as the baseline classifier. The kernel size σ is tuned by cross-validation from the set $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$. The best performance is summarized in Table 2. As can be seen, KNN and SVM always lead to comparable results. Our MPRI is consistently better than other competitors, regardless of the used classifier.
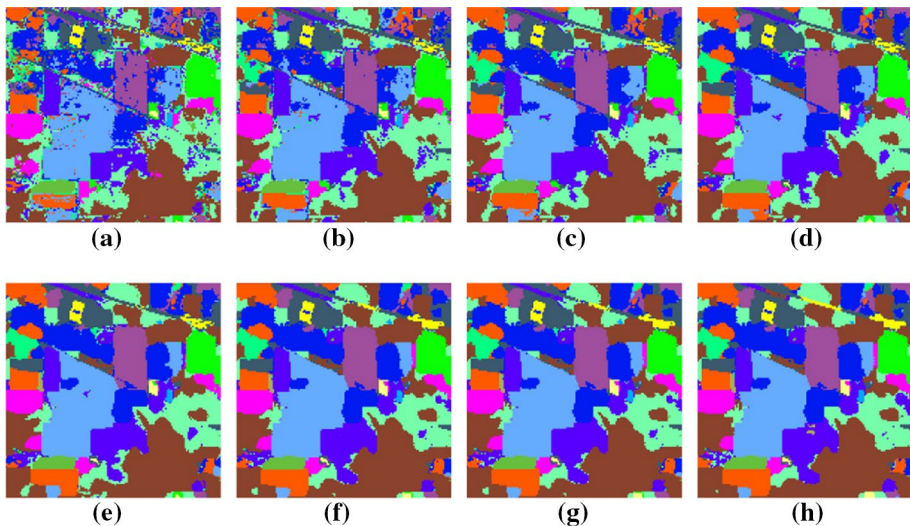

Fig. 10 Classification maps of MRPI with **a** 1 layer; **b** 2 layers; **c** 3 layers; **d** 4 layers; **e** 5 layers; **f** 6 layers; **g** 7 layers; and **h** 8 layers

Table 2 OAs (%) of different methods using KNN and SVM

Classifier	EPF	SADL	PCA-EPF	MPRI
KNN ($k = 1$)	77.02	87.48	90.03	**94.00**
SVM ($\sigma = 0.1$)	78.44	88.19	91.77	**93.11**

2% of labeled samples per class were randomly selected for training. The best performances are marked in bold

4.2 Evaluation on component-wise contributions

Before systematically evaluating the performance of MPRI, we first compare it with its degraded baseline variants to demonstrate the component-wise contributions to the performance gain. The results are summarized in Table 3. As can be seen, models that only consider one attribute (i.e., *multi-layer*, *multi-scale* and *multi-β*) improve the performance marginally. Moreover, it is interesting to find that *multi-layer* and *multi-scale* play more significant roles than *multi-β*. One possible reason is that the representations learned from different β contain redundant information with respect to class labels. However, either the combination of *multi-layer* and *multi-β* or the combination of *multi-scale* and *multi-β* can obtain remarkable improvements. Our MPRI performs the best as expected. This result indicates that *multi-layer*, *multi-scale* and *multi-β* are essentially important for the problem of HSI classification.

4.3 Comparison with state-of-the-art methods

Having illustrated component-wise contributions of MPRI, we compare it with several state-of-the-art methods, including EPF (Kang et al., 2014), MPM-LBP (Li et al., 2013), SADL (Soltani-Farani et al., 2015), MFL (Li et al., 2015), PCA-EPF (Kang et al., 2017), HIFI (Pan et al., 2017), hybrid spectral convolutional neural network (HybridSN) (Roy et al., 2020), similarity-preserving deep features (SPDF) (Fang et al., 2019), convolutional neural network with Markov random fields (CNN-MRF) (Cao et al., 2018), local covariance matrix representation (LCMR) (Fang et al., 2018), and random patches network (RPNet) (Xu et al., 2018).

Table 3 Quantitative evaluation of our MPRI (the last row) and its degraded baseline variants in terms of OA, AA, and κ

Multi-layer	Multi-scale	Multi-β	OA	AA	κ
			76.25	72.74	0.729
✓			80.86	79.33	0.7814
	✓		81.28	76.25	0.786
		✓	76.61	73.09	0.733
	✓	✓	83.30	77.89	0.809
✓		✓	86.94	85.35	0.851
✓	✓		<u>92.80</u>	<u>91.16</u>	<u>0.918</u>
✓	✓	✓	**94.00**	**91.20**	**0.932**

"✓" denotes the model contains the corresponding module. The best two performances are marked in bold and underlined, respectively

Tables 4, 5 and 6 summarized quantitative evaluation results of different methods. For each method, we report its classification accuracy in each land cover category as well as the overall OA, AA and κ values across all categories. To avoid biased evaluation, we average the results from 10 independent runs (except for the Pavia Center data set, in which the training and testing samples are fixed). Obviously, MPRI achieves the best or the second best performance in most of items. These results suggest that MPRI is able to learn more discriminative spectral-spatial features than its counterparts using classical machine learning models.

The classification maps of different methods in three data sets are demonstrated in Figs. 11, 12 and 13, which further corroborate the above quantitative evaluations. The performances of EPF and MPM-LBP are omitted due to their relatively lower quantitative evaluations. It is very easy to observe that our proposed MPRI improves the region uniformity (see the small region marked with dashed border) and the edge preservation (see the small region marked by solid line rectangles) significantly, both criteria are critical for evaluating classification maps (Kang et al., 2017). By contrast, other methods either fail to preserve local details (such as edges) of different classes (e.g., MFL) or generate noises in the uniform regions (e.g., SADL, PCA-EPF and HIFI).

To evaluate the robustness of our method with respect to the number of training samples, we demonstrate, in Fig. 14, the OA values of different methods in a range of the percentage of training samples per class. As can be expected, the more training samples, the better classification performance. However, MPRI is consistently superior to its counterparts, especially when the training samples are limited.

4.4 Computational complexity analysis

We finally investigate the computational complexity of different sliding window filtering based HSI classification methods. Note that, PRI can also be interpreted as a special kind of filtering, as the center pixel representation is determined by its surrounding pixels with a Gaussian weight [see Eq. (12)].

The computational complexity and the averaged running time on each pixel (in s) of different methods are summarized in Table 7. For PCA-EPF, $\tilde{d}$ is the dimension of averaged images, $\hat{S}$ is the number of different filter parameter settings, $\hat{T}$ is the number of iterations. For HIFI, d is the number of hyperspectral bands, $\hat{H}$ is the number of the hierarchies. For MPRI, L, S, and B are respectively the numbers of layers, scales and betas. Usually, $\tilde{d}$ is set to 16, $\hat{S}$ is set to 3, and $\hat{T}$ is set to 3, which makes PCA-EPF very fast.

According to Eq. (7), the computational complexity of PRI grows quadratically with data size (i.e., $\hat{N}$). Although one can simply apply rank deficient approximation to the Gram matrix for efficient computation of PRI, this strategy is preferable only when $l \ll \hat{N}$, where l is the square of number of subsamples used to approximate the original Gram matrix (Sánchez Giraldo and Príncipe, 2011). In our application, $\hat{N}$ is less than a few hundreds (~ 169 at most), whereas we always need to set $l \geq 25$ to guarantee a non-decreasing accuracy. From Table 7, the reduced computational power by Gram matrix approximation is marginal. However, as shown in Fig. 15, such an approximation method is prone to cause over-smooth effect.

Finally, one should note that, although MPRI takes more time than its sliding window filtering based counterparts, it is still much more timesaving than prevalent DNN based methods. For example, CNN-MRF takes more than 6, 000s (on a PC equipped with a

Table 4 Classification accuracies (%) of different methods on Indian Pines data set

Class	EPF	MPM-LBP	SADL	MFL	PCA-EPF	HIFI	Hybr-idSN	SPDF	CNN-MRF	RPNet	MPRI
1	10.00	51.73	75.77	66.92	**98.76**	87.33	17.78	40.44	46.67	14.67	74.44
2	74.41	80.98	83.89	80.58	88.87	88.34	70.29	81.66	70.33	86.90	**92.48**
3	89.35	68.46	81.57	76.41	88.15	89.00	71.87	73.01	67.75	69.43	**95.68**
4	79.01	44.76	64.15	26.46	**87.08**	81.85	70.69	53.88	43.36	26.21	72.67
5	**95.15**	80.86	90.12	80.76	94.78	83.97	83.30	77.02	78.94	83.62	89.98
6	73.93	**97.87**	97.34	91.20	92.81	96.90	97.76	85.94	85.69	90.69	96.91
7	50.00	29.20	**100.0**	67.20	80.30	96.30	48.15	40.74	36.30	70.74	96.67
8	86.57	98.37	99.65	96.41	99.83	98.57	99.57	88.63	92.46	84.81	**100.0**
9	10.00	40.53	93.16	32.11	76.45	**96.84**	0.000	40.53	69.47	41.58	81.05
10	77.92	75.79	85.79	82.59	90.10	88.80	79.22	70.41	63.03	78.68	**92.68**
11	71.76	90.25	91.55	90.50	92.65	90.19	91.77	90.89	87.88	**95.92**	94.90
12	74.80	79.53	66.27	70.70	86.78	87.30	54.73	69.98	59.52	48.43	**87.93**
13	98.02	**99.47**	98.55	96.28	96.40	99.30	78.11	80.50	71.05	89.10	98.40
14	91.21	96.36	96.34	96.47	97.72	98.24	96.53	92.15	95.05	92.17	**98.59**
15	78.84	63.17	84.78	74.11	**94.11**	81.83	63.76	72.09	57.27	55.63	93.89
16	87.15	50.54	90.32	73.01	89.67	**98.24**	75.82	46.15	80.77	29.67	92.97
OA	78.44	83.84	88.19	84.32	91.77	90.87	82.22	81.41	77.39	81.80	**94.00**
AA	71.76	71.74	87.45	75.11	90.90	**91.44**	68.71	69.00	69.10	66.14	91.20
κ	0.750	0.814	0.865	0.821	0.906	0.896	0.797	0.786	0.740	0.788	**0.932**

2% of labeled samples per class were randomly selected for training. The best two performances are marked in bold and underlined, respectively

Table 5 Classification accuracies (%) of different methods on Pavia University data set

Class	EPF	MPM-LBP	SADL	MFL	PCA-EPF	HIFI	Hybr-idSN	SPDF	CNN-MRF	RPNet	MPRI
1	93.05	97.38	92.66	**98.05**	96.56	92.37	90.78	96.10	93.79	91.40	97.75
2	95.37	99.45	98.92	99.61	98.67	98.54	99.84	99.25	98.51	97.94	**99.88**
3	**96.87**	79.02	74.65	74.35	87.43	80.40	71.80	92.98	65.04	86.48	89.52
4	**99.46**	90.62	93.50	89.78	98.34	81.61	73.95	83.54	95.95	91.23	94.05
5	98.09	97.52	99.38	98.45	99.44	96.39	98.05	96.06	96.66	95.51	98.87
6	98.52	93.37	94.57	95.04	99.00	90.49	**99.86**	95.07	93.15	93.83	96.26
7	99.54	82.95	77.26	94.12	94.29	87.67	**99.62**	83.30	64.00	81.66	99.02
8	87.63	91.52	77.73	93.11	92.22	89.71	92.24	**95.14**	80.21	87.40	91.69
9	96.88	98.94	99.14	91.54	98.36	**99.92**	59.23	55.39	93.55	96.03	97.18
OA	95.06	95.42	93.28	95.83	97.11	93.40	93.59	94.92	92.51	93.87	**97.31**
AA	96.16	92.31	89.76	92.67	**96.21**	90.78	87.26	88.54	86.76	91.28	96.03
κ	0.935	0.940	0.912	0.945	0.962	0.912	0.932	0.940	90.07	0.918	**0.965**

1% of labeled samples per class were randomly selected for training. The best two performances are marked in bold and underlined, respectively

Table 6 Classification accuracies (%) of different methods on Pavia Center data set with fixed training and testing split

Class	EPF	MPM-LBP	SADL	MFL	PCA-EPF	HIFI	Hybr-idSN	SPDF	CNN-MRF	RPNet	MPRI
1	**100.0**	99.10	98.79	99.43	**100.0**	99.57	97.57	97.45	95.84	94.40	**100.0**
2	**99.39**	94.57	92.66	87.02	97.21	92.63	90.21	89.15	98.03	90.58	97.76
3	86.78	95.83	96.63	95.83	88.49	96.16	99.34	98.29	86.01	99.15	**99.86**
4	98.28	81.94	98.32	99.46	97.74	93.70	97.78	59.87	99.22	80.50	**100.0**
5	96.64	97.75	97.91	99.43	**99.54**	98.32	98.22	92.15	98.22	99.09	98.55
6	77.48	95.21	95.51	95.05	93.93	98.39	99.48	97.47	92.24	96.24	**99.83**
7	95.29	94.51	98.24	96.39	**99.84**	94.00	**99.51**	87.16	98.69	90.83	97.85
8	**100.0**	99.62	98.76	99.34	98.61	99.72	99.97	84.20	99.76	88.58	**100.0**
9	**100.0**	**100.0**	**100.0**	97.23	**100.0**	99.95	97.97	92.23	99.29	94.06	97.31
OA	97.08	97.85	98.08	98.01	99.01	98.48	97.81	94.86	96.07	94.03	**99.57**
AA	94.87	95.39	97.42	96.47	97.26	96.94	97.82	88.66	96.36	92.60	**99.02**
κ	0.948	0.961	0.965	0.964	0.982	0.972	0.961	0.908	0.931	0.895	**0.992**

The best two performances are marked in bold and underlined, respectively
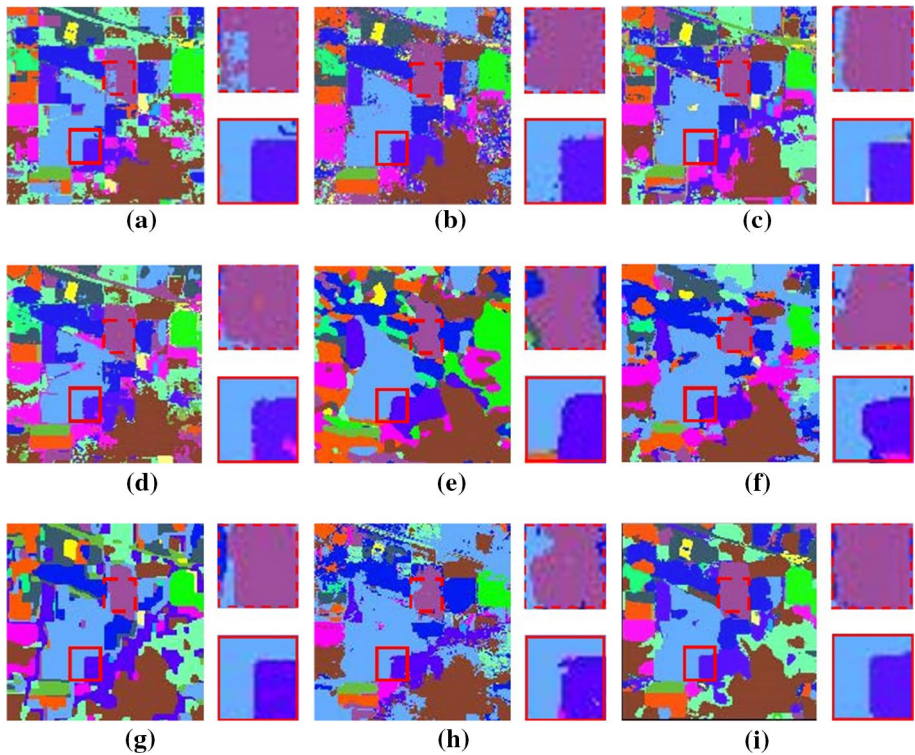
Fig. 11 Classification maps on Indian Pines data set. **a** SADL; **b** MFL; **c** PCA-EPF; **d** HIFI; **e** HybridSN; **f** SPDF; **g** CNN-MRF; **h** RPNet; **i** MPRI

single 1080 Ti GPU, i7 8700k CPU and 64 GB RAM) to train a CNN model using 2% labeled data on Indian pines data set with 10x data augmentation.

5 Conclusions

This paper proposes multiscale principle of relevant information (MPRI) for hyperspectral image (HSI) classification. MPRI uses PRI—an unsupervised information-theoretic learning principle that aims to perform mode decomposition of a random variable X with a known (and fixed) probability distribution g by a hyperparameter β—as the basic building block. It integrates multiple such blocks into a multiscale (by using sliding windows of different sizes) and multilayer (by stacking PRI successively) structure to extract spectral-spatial features of HSI data from a coarse-to-fine manner. Different from existing deep neural networks, MPRI can be efficiently trained greedy layer-wisely without error backpropagation. Empirical evidence indicates $\beta \in [2, 4]$ in PRI is able to balance the trade-off between the regularity of extracted representation and its discriminative power to HSI data. Comparative studies on three benchmark data sets demonstrate that MPRI is able to learn discriminative representations from 3D spatial-spectral
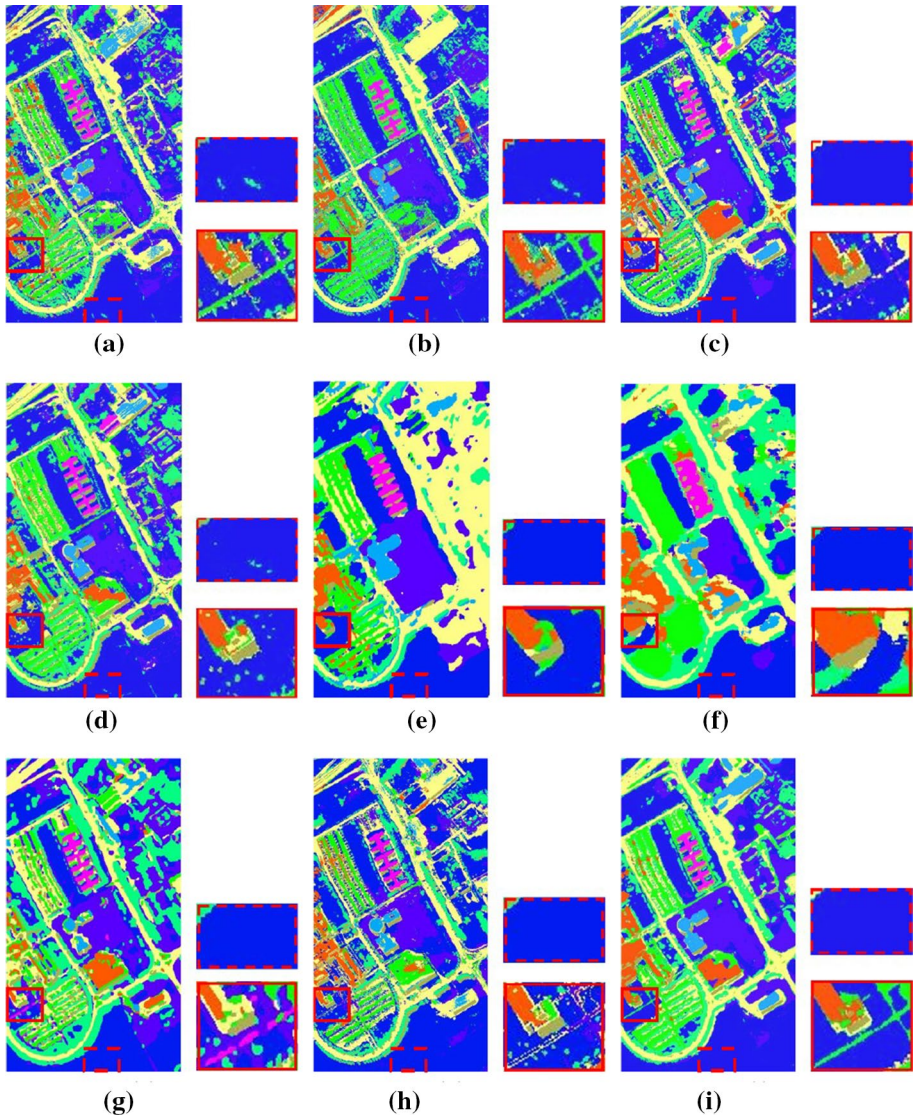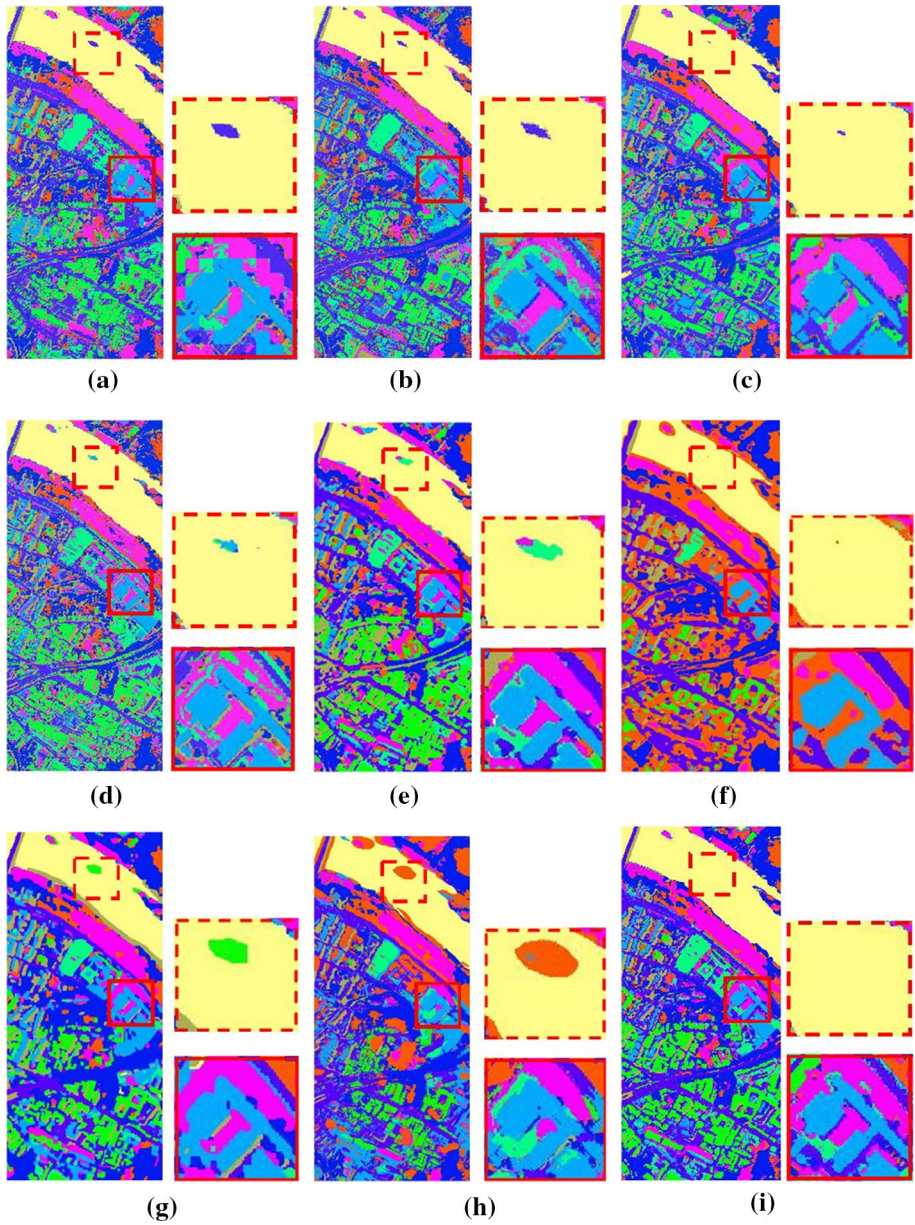
Fig. 12 Classification maps on University of Pavia data set. **a** SADL; **b** MFL; **c** PCA-EPF; **d** HIFI; **e** HybridSN; **f** SPDF; **g** CNN-MRF; **h** RPNet; **i** MPRI

data, with significantly fewer training samples. Moreover, MPRI enjoys an intuitive geometric interpretation, it also prompts the region uniformity and edge preservation of classification maps. In the future, we intend to speed up the optimization of PRI. In this line of research, the random fourier feature (Rahimi & Recht, 2008) seems to be a promising avenue.

Fig. 13 Classification maps on Pavia Center data set. **a** SADL; **b** MFL; **c** PCA-EPF; **d** HIFI; **e** HybridSN; **f** SPDF; **g** CNN-MRF; **h** RPNet; **i** MPRI

Fig. 14 OA values of different methods with respect to different percentages of training samples per class on **a** Indian Pines; and **b** Pavia University. The results on Pavia Center is omitted, because the training and testing samples are fixed
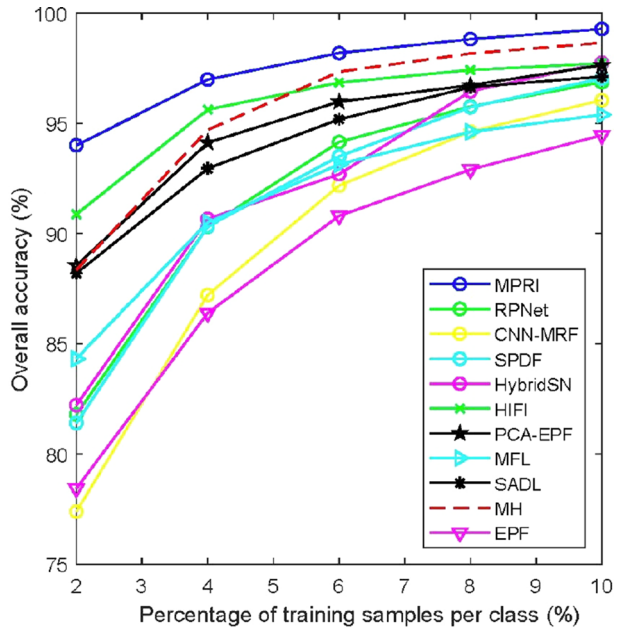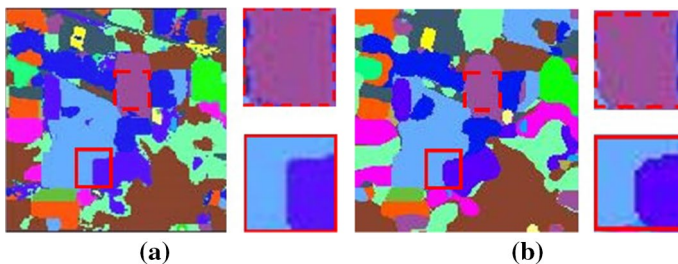

(a)
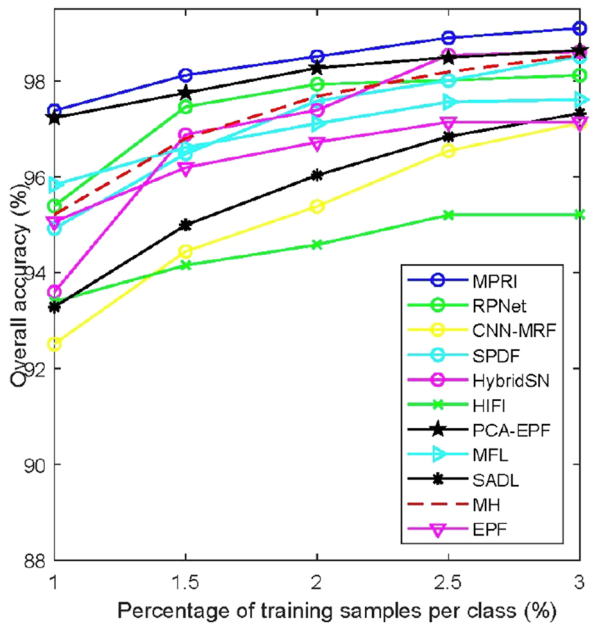
(b)

Table 7 Computation complexity and running time (per pixel) of different methods

Method	Computational complexity	Running time per pixel (s)
EPF	$\mathcal{O}(1)$	2.9×10^{-6}
PCA-EPF	$\mathcal{O}(\tilde{d}\hat{S}\hat{T})$	2.6×10^{-5}
HIFI	$\mathcal{O}(d\hat{H})$	1.0×10^{-3}
MPRI	$\mathcal{O}(d\hat{N}^2\tau LSB)$	3.2×10^{-1}
MPRI with Nyström-KECA	$\mathcal{O}(d\hat{N}l\tau LSB)$	2.6×10^{-1}

(a)   **(b)**

Fig. 15 Classification maps of **a** MPRI and **b** MPRI with Nyström-KECA. The Gram matrix approximation is prone to cause over-smooth effect

References

Achille, A., & Soatto, S. (2018). Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(12), 2897–2905.

Bandos, T., Bruzzone, L., & Camps-Valls, G. (2009). Classification of hyperspectral images with regularized linear discriminant analysis. *IEEE Transactions on Geoscience and Remote Sensing, 47*(3), 862–873.

Baumgardner, M. F., Biehl, L. L., & Landgrebe, D. A (2015). 220 band aviris hyperspectral image data set: June 12, 1992 Indian pine test site 3. https://doi.org/10.4231/R7RX991C

Cao, X., Zhou, F., Xu, L., Meng, D., Xu, Z., & Paisley, J. (2018). Hyperspectral image classification with Markov random fields and a convolutional neural network. *IEEE Transactions on Image Processing, 27*(5), 2354–2367.

Chang, C. I. (2000). An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis. *IEEE Transactions on Information Theory, 46*(5), 1927–1932.

Chen, Y., Jiang, H., Li, C., Jia, X., & Ghamisi, P. (2016). Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing, 54*(10), 6232–6251.

Chen, Y., Lin, Z., Zhao, X., Wang, G., & Gu, Y. (2014). Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7*(6), 2094–2107.

Chen, Y., Nasrabadi, N. M., & Tran, T. D. (2011). Hyperspectral image classification using dictionary-based sparse representation. *IEEE Transactions on Geoscience and Remote Sensing, 49*(10), 3973–3985.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics, 23*(4), 493–507.

Du, Q. (2007). Modified fisher's linear discriminant analysis for hyperspectral imagery. *IEEE Geoscience and Remote Sensing Letters, 4*(4), 503–507.

Fang, L., He, N., Li, S., Plaza, A. J., & Plaza, J. (2018). A new spatial-spectral feature extraction method for hyperspectral images using local covariance matrix representation. *IEEE Transactions on Geoscience and Remote Sensing, 56*(6), 3534–3546.

Fang, L., Liu, Z., & Song, W. (2019). Deep hashing neural networks for hyperspectral image feature extraction. *IEEE Geoscience and Remote Sensing Letters, 16*(9), 1412–1416.

Feng, J., Jiao, L., Liu, F., Sun, T., & Zhang, X. (2015). Mutual-information-based semi-supervised hyperspectral band selection with high discrimination, high information, and low redundancy. *IEEE Transactions on Geoscience and Remote Sensing, 53*(5), 2956–2969.

Ghamisi, P., Benediktsson, J. A., Cavallaro, G., & Plaza, A. (2014). Automatic framework for spectral-spatial classification based on supervised feature extraction and morphological attribute profiles. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7*(6), 2147–2160.

Ghamisi, P., Dalla Mura, M., & Benediktsson, J. A. (2015). A survey on spectral-spatial classification techniques based on attribute profiles. *IEEE Transactions on Geoscience and Remote Sensing, 53*(5), 2335–2353.

Ghamisi, P., Yokoya, N., Li, J., Liao, W., Liu, S., Plaza, J., et al. 2017). Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art. *IEEE Geoscience and Remote Sensing Magazine, 5*(4), 37–78.

Hastie, T., & Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association, 84*(406), 502–516.

He, L., Li, J., Liu, C., & Li, S. (2018). Recent advances on spectral-spatial hyperspectral image classification: An overview and new guidelines. *IEEE Transactions on Geoscience and Remote Sensing, 56*(3), 1579–1597.

Jenssen, R., Principe, J. C., Erdogmus, D., & Eltoft, T. (2006). The Cauchy-Schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels. *Journal of The Franklin Institute, 343*(6), 614–629.

Kamandar, M., & Ghassemian, H. (2013). Linear feature extraction for hyperspectral images based on information theoretic learning. *IEEE Geoscience and Remote Sensing Letters, 10*(4), 702–706.

Kang, X., Li, S., & Benediktsson, J. A. (2014). Spectral-spatial hyperspectral image classification with edge-preserving filtering. *IEEE Transactions on Geoscience and Remote Sensing, 52*(5), 2666–2677.

Kang, X., Xiang, X., Li, S., & Benediktsson, J. A. (2017). Pca-based edge-preserving features for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 55*(12), 7140–7151.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86.

Li, J., Bioucas-Dias, J. M., & Plaza, A. (2013). Spectral-spatial classification of hyperspectral data using loopy belief propagation and active learning. *IEEE Transactions on Geoscience and Remote Sensing, 51*(2), 844–856.

Li, J., Huang, X., Gamba, P., Bioucas-Dias, J. M., Zhang, L., Benediktsson, J. A., & Plaza, A. (2015). Multiple feature learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 53*(3), 1592–1606.

Li, Y., Zhang, H., & Shen, Q. (2017). Spectral-spatial classification of hyperspectral imagery with 3d convolutional neural network. *Remote Sensing, 9*(1), 67.

Ma, X., Wang, H., & Geng, J. (2016). Spectral-spatial classification of hyperspectral image based on deep auto-encoder. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9*(9), 4073–4085.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research, 9,* 2579–2605.

Pan, B., Shi, Z., & Xu, X. (2017). Hierarchical guidance filtering-based ensemble classification for hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing, 55*(7), 4177–4189.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics, 33*(3), 1065–1076.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(8), 1226–1238.

Principe, J. C. (2010). *Information theoretic learning: Renyi's entropy and kernel perspectives*. New York: Springer.

Rahimi, A., & Recht, B (2008). Random features for large-scale kernel machines. In: Advances in neural information processing systems, pp. 1177–1184

Rao, S. M. (2008). *Unsupervised learning: An information theoretic framework*. Ph.D. thesis, University of Florida

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: Contributions to the theory of statistics*. The Regents of the University of California.

Roy, S. K., Krishna, G., Dubey, S. R., & Chaudhuri, B. B. (2020). Hybridsn: Exploring 3-d-c2-d cnn feature hierarchy for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters, 17*(2), 277–281.

Sánchez Giraldo, L. G., & Príncipe, J. C. (2011). An efficient rank-deficient computation of the principle of relevant information. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2176–2179). IEEE.

Shwartz-Ziv, R., & Tishby, N (2017). Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Soltani-Farani, A., Rabiee, H. R., & Hosseini, S. A. (2015). Spatial-aware dictionary learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 53*(1), 527–541.

Tishby N., et al (2000). The information bottleneck method. arXiv preprint arXiv:physics/0004057

Xu, Y., Du, B., Zhang, F., & Zhang, L. (2018). Hyperspectral image classification via a random patches network. *ISPRS Journal of Photogrammetry and Remote Sensing, 142,* 344–357.

Yang, X., Ye, Y., Li, X., Lau, R. Y., Zhang, X., & Huang, X. (2018). Hyperspectral image classification with deep learning models. *IEEE Transactions on Geoscience and Remote Sensing, 56*(9), 5408–5423.

Yu, S., Sanchez Giraldo, L. G., Jenssen, R., & Principe, J. C. (2019). Multivariate extension of matrix-based Renyi's α-order entropy functional. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(11), 2960–2966.

Zare, A., Jiao, C., & Glenn, T. (2018). Discriminative multiple instance hyperspectral target characterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(10), 2342–2354.

Zhong, Z., Li, J., Luo, Z., & Chapman, M. (2018). Spectral-spatial residual network for hyperspectral image classification: A 3-d deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing, 56*(2), 847–858.

Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: a comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine, 5*(4), 8–36.