Expert Systems With Applications 129 (2019) 246–259

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Hyperspectral imagery classification based on semi-supervised 3-D deep neural network and adaptive band selection

Akrem Sellami[a,b,*], Mohamed Farah[a], Imed Riadh Farah[a,b], Basel Solaiman[b]

[a] RIADI Laboratory, National School of Computer Science, Manouba, Tunisia
[b] ITI Department, IMT Atlantique, Brest, France

ARTICLE INFO

Article history:
Received 10 October 2018
Revised 18 March 2019
Accepted 4 April 2019
Available online 5 April 2019

Keywords:
Hyperspectral imagery classification
Convolutional neural network (CNN)
Adaptive dimensionality reduction
Deep learning

ABSTRACT

This paper proposes a novel approach based on adaptive dimensionality reduction (ADR) and a semi-supervised 3-D convolutional neural network (3-D CNN) for the spectro-spatial classification of hyperspectral images (HSIs). It tackles the problem of curse of dimensionality and the limited number of training samples by selecting the most relevant spectral bands. The selected bands should be informative, discriminative and distinctive. They are fed into a semi-supervised 3-D CNN feature extractor, then a linear regression classifier to produce the classification map. In fact, the proposed semi-supervised 3-D CNN model seeks to extract the deep spectral and spatial features based on convolutional encoder-decoder to enhance the HSI classification. It uses several 3-D convolution and max-pooling layers to extract these features from the selected relevant bands. The main advantage of the proposed approach is to reduce the high dimensionality of HSI, preserve the relevant spectro-spatial information and enhance the classification using few labeled training samples. Experimental studies are carried out on three real HSI data sets: Indian Pines, Pavia University, and Salinas. The obtained results show that the proposed approach performs better than other deep learning-based methods including CNN-based methods, and significantly improves the classification accuracy of HSIs.

1. Introduction

Nowadays, hyperspectral remote sensing sensors can generate hundreds of contiguous and narrow spectral bands with high spectral resolution, i.e. they can provide rich spectral information for the discrimination and classification of materials in a scene (Fauvel, Tarabalka, Benediktsson, Chanussot, & Tilton, 2013). Therefore, hyperspectral images (HSIs) have been widely used in environmental management (Pan, Shi, An, Jiang, & Ma, 2017; Yousefi, Castanedo, Bédard, Beaudoin, & Maldague, 2018), spectral unmixing (Jiang, Gong, Li, Zhang, & Li, 2018; Lu, Wu, Yuan, Yan, & Li, 2013; Xu & Shi, 2017), anomaly detection (Matteoli, Diani, & Corsini, 2010), and many other applications. In fact, for all these applications, the identification of the class of each pixel in HSI is required (Jamshidpour, Safari, & Homayouni, 2017; Pan, Li, Wang, & Gao, 2018; Shi & Pun, 2018; Wu, Zhu, Huang, & Li, 2016). However, HSI classification still raises the following issues (Chen, Jiao, et al., 2017; Wen et al., 2016): (i) the high correlation between the spectral bands; (ii) the spatial variability of different spectral signatures; and (iii) the large number of spectral bands producing the Hughes phenomenon, i.e., the classification performance decreases when the number of bands is very high whereas the number of training samples is very limited (Feng, Jiao, Zhang, & Sun, 2014). Therefore, dimensionality reduction (DR) becomes necessary before HSI classification, since it allows reducing the number of spectral bands, as well as the time needed for classification.

The two main approaches for DR are feature extraction (spectral band projection) (Lu et al., 2013; Zhou, Peng, & Chen, 2015) and spectral band selection (Bai, Guo, Wang, Zhang, & Zhou, 2015; Verrelst et al., 2016). Feature extraction aims to project the original hyperspectral data into a reduced subspace with linear or nonlinear transformations of the original spectral bands, where the dimension of the reduced subspace is much smaller than the dimension of the original HSI. Band selection seeks to select a subset of relevant spectral bands, i.e., the selected spectral bands should be the most discriminative, informative, and distinctive with low

* Corresponding author at: ITI Department, IMT Atlantique, Brest, France.
E-mail addresses: akrem.sellami@imt-atlantique.fr (A. Sellami), mohamed.farah@riadi.rnu.tn (M. Farah), riadh.farah@ensi.rnu.tn (I. Riadh Farah), basel.solaiman@imt-atlantique.fr (B. Solaiman).

https://doi.org/10.1016/j.eswa.2019.04.006
0957-4174/

correlation and redundancy (Feng, Jiao, Liu, Sun, & Zhang, 2015; Jia, Ji, Qian, & Shen, Apr. 2012).

In this paper, in order to overcome the overfitting problem, i.e., the limited number of training samples and the high dimensionality of HSI, we propose a novel approach that allows reducing the number of spectral bands using a semi-supervised band selection. Also, a semi-supervised 3-D CNN is used to enhance HSI spatial-spectral classification. In fact, the CNN model has shown to produce high classification performances of HSI. However, to come up with the curse of dimensionality and the overfitting problems, dimensionality reduction becomes necessary. Feature extraction can extract useful information from HSI, but causes a loss of initial spectro-spatial features. For these reasons, band selection is used to select relevant spectral bands by preserving the physical meaning of HSI and the useful spectro-spatial features. It improves the classification accuracy by addressing the overfetting and the curse of dimensionality problems. Consequently, we propose in the first step to select relevant bands (informative, discriminative and distinctive) while preserving useful spectro-spatial features. Moreover, a semi-supervised 3-D CNN is performed for spectro-spatial features learning and classification.

The major contributions of our proposed approach can be summarized as follows:

- An adaptive DR approach is developed, aiming to solve the *curse of dimensionality* problem by finding the most informative, discriminative, and distinctive spectral bands with low redundancy, and preserving the physical meaning of HSI. It is a semi-supervised band selection approach, which does not need a large number of training samples to select spectral bands.
- A semi-supervised 3-D CNN approach based on convolutional encoder-decoder for HSI spatial-spectral classification is proposed to extract the spatial-spectral features of HSI using few training samples, which leads to high classification performances.
- The proposed approach is applied on three real HSIs. We compare the proposed approach with other traditional DL methods including CNN-based methods, using the following criteria: overall accuracy (OA), average accuracy (AA), kappa index (K), and processing time (t).

The remainder of the paper is organized as follows. In Section 2, we report state-of-the-art methods related to dimensionality reduction and deep learning for HSI classification. In Section 3, we present a description of the proposed approach, i.e., spatial-spectral classification based on adaptive DR and 3-D CNN. Section 4 describes the datasets and reports the obtained results and discussion. We conclude this paper with some future works in Section 5.

2. Related works

2.1. Dimensionality reduction (DR)

There are two main approaches for DR: feature extraction and band selection. They can be unsupervised, supervised, or semi-supervised.

Many unsupervised linear feature extraction approaches have been proposed, such as principal component analysis (PCA) (Agarwal, El-Ghazawi, El-Askary, & Le-Moigne, 2007; Yousefi et al., 2018), and locality preserving projection (LPP) (He & Niyogi, 2003). PCA seeks to find a projection matrix maximizing the data variance in the reduced subspace. LPP attempts to preserve the spatial information, i.e. the distance between the data points in the reduced subspace. The tensor model has also been used with linear feature extraction approaches to preserve the spatial information, such as Tensor-PCA (TPCA), Tensor-independent component analysis (TICA) (Renard & Bourennane, 2009), and Tensor-LPP (TLPP) (Sellami & Farah, 2016).

Moreover, supervised linear projection methods attempt to search the best projection matrix using a priori knowledge on classes. Fisher linear discriminant analysis (FLDA) finds the relevant features minimizing the within-class distance and maximizing the inter-class distance (Huang, Li, & Liu, 2012). Local discriminant embedding (LDE) aims to find the best projection matrix maximizing the inter-class scatter matrix by preserving the neighborhood of data points using a graph embedding model (Yan et al., 2007).

Many nonlinear feature extraction approaches have been developed for HSI classification. The main goal of these approaches is to take into account the non-linearity of the original data. Kernel PCA (KPCA) finds a projection matrix maximizing the data variance by using a kernel (Fauvel, Chanussot, & Benediktsson, 2006). Graph embedding (GE) (Zhou et al., 2015) seeks to find an optimal discriminative projection by minimizing a local spatial-spectral scatter (pixel neighborhood). Khan, Shafait, and Mian (2015) proposed a Joint Group Sparse PCA (JGSPCA) approach, which aims to find few groups of features that jointly capture most of the variation in the hyperspectral data. Karami, Yazdi, and Asli (2011) proposed a Genetic Kernel Tucker Decomposition (GKTD) algorithm, which exploits both spatial and spectral information of the HSIs. Hoyer (2004) used the Non-negative Matrix Factorization (NMF) method by incorporating the notion of sparseness in order to improve the decompositions.

Band selection approaches seek to find relevant bands and therefore keep a subset of the spectral information of the original HSI. Depending on the availability of prior information upon classes, band selection can be supervised, unsupervised or semi-supervised. Supervised band selection aims to find the most discriminative spectral bands using distance metrics (Feng et al., 2014; Yan et al., 2016), and information measures (Peng, Long, & Ding, 2005; Sotoca & Pla, 2010). However, supervised spectral band selection methods do not consider the correlation between bands. Besides, they cannot be used unless some labelled data are available.

Unsupervised band selection approaches (Du, Qi, Wang, Ramanath, & Snyder, 2003; Feng, Jiao, Liu, Sun, & Zhang, 2016; Jia et al., Apr. 2012) aim to find the most distinctive and informative spectral bands using spectral band ranking (Datta, Ghosh, & Ghosh, 2015; Kim & Rattakorn, 2011) or spectral band clustering (Wang, Zhang, Wang, Madani, & Sabourin, 2017; Xu, Li, Wu, & Plaza, 2018). However, unsupervised spectral band selection approaches do not consider the inter-bands redundancy.

Semi-supervised band selection approaches seek to find informative, discriminative, and distinctive spectral bands using clustering (Jiao, Feng, Liu, Sun, & Zhang, 2015; Sellami, Farah, Farah, & Solaiman, 2018; Su, Yang, Du, & Sheng, 2011), hypergraph model (HM) (Bai et al., 2015), or mutual information (MI) (Feng et al., 2015).

Typically, the main drawback of feature extraction methods is that the result after reducing dimensions no longer contains the original spectral information because the hyperspectral data are transformed, and some crucial spectral and spatial information may have been discarded or distorted (Chang & Wang, 2006). However, band selection approaches aim to preserve the physical meaning of HSI.

2.2. HSI spatial-spectral classification

Usually, HSI classification is performed using low-dimensional features instead of the original set of high-dimensional features. This allows addressing the problems of the high-dimensionality of HSI and the unavailability of labeled samples. Moreover, in order

to improve HSI classification, many studies tried to consider at the same time spectral and spatial features. Extended morphological profiles (EMPs) aim to fuse together spatial and spectral features (Xia, Dalla Mura, Chanussot, Du, & He, 2015). Gabor filtering (GF) seeks to extract textures and edges as spatial features (Chen, Zhu, et al., 2017). Many techniques have been developed using the support vector machine (SVM) for HSI spectral-spatial classification (Li, Ge, & Gao, 2017; Zhong, Lin, & Zhang, 2014). However, these spatial feature extraction methods do not allow detecting rich spatial properties of the objects (Chen, Jiang, Li, Jia, & Ghamisi, 2016; Zhao & Du, 2016).

Recently, several deep learning (DL) models have been shown to produce high performances for HSI spectral-spatial classification (Ghamisi, Chen, & Zhu, 2016; Ma, Wang, & Wang, 2016; Shi & Pun, 2018; Zhao & Du, 2016). Some works have been proposed to combine DR and DL models for HSI classification. For instance, logistic regression (LR), stacked autoencoders (SAE) and PCA have been jointly used for HSI classification, where SAE is applied to extract the spatial-spectral features from HSI (Chen, Lin, Zhao, Wang, & Gu, 2014).

Deep belief network (DBN) has also been used for HSI classification (Chen, Zhao, & Jia, 2015; Zhong, Gong, Li, & Schönlieb, 2017). DBN is based on the generative model and the back-propagation algorithm in the training step. In Chen et al. (2015), authors have proposed a hybrid framework of PCA and DBN. Moreover, Zhou et al. (2017) proposed an approach which is based on band selection and DBN for HSI classification. However, DBN and SAE, which aim to extract the deep features, cannot extract the spatial features from HSI because the training samples have to be flattened into a 1-D vector before training. As a result, spatial information is lost by flattening training samples.

In recent years, convolutional neural networks (CNNs) have been used for HSI spatial-spectral classification. CNNs can extract spatial features from HSI without any flattening of the training samples, giving satisfactory results for the classification (Hu, Huang, Wei, Zhang, & Li, 2015). Moreover, an approach based on PCA, CNN and LR was employed in Zhao and Du (2016) for HSI classification. Also, Chen, Jiao, et al. (2017) have proposed a method for HSI classification, which combines Gabor filters (GF) with convolutional filters to mitigate the problem of overfitting. Also, 3-D CNN can extract the spatial and spectral features of HSI simultaneously. Furthermore, CNNs use shared weights, which reduce the number of parameters in the training stage compared to other DL architectures. Nevertheless, 3-D CNN needs a very large number of training samples in the training phase to obtain appropriate weights. Unfortunately, for HSI classification, the number of training samples is very limited, which degrades the performances of most supervised classification approaches.

3. Proposed approach

In our proposed approach, we first apply an adaptive dimensionality reduction method in order to reduce the high-dimensionality of HSI by selecting informative, discriminative and distinctive spectral bands for HSI classification with a limited number of training samples. We also use a semi-supervised 3-D CNN method which aims to use several 3-D convolution and max-pooling layers to extract the deep spatial and spectral features using labeled and unlabeled training samples. After that, these features can be flattened and be used to produce the final spectral-spatial classification.

The adaptive dimensionality reduction method allows tackling the problem of the curse of dimensionality Hughes (Jan. 1968). In fact, this problem occurs when the number of labeled training samples per class is very small compared to the dimensionality of the pixels, i.e, the ratio between the number of spectral bands (features) of the HSI and the number of samples used for training is very low (Passalis & Tefas, 2018). Indeed, the Hughes phenomenon often leads to a misclassification and therefore to poor performances. Moreover, when we increase the number of spectral bands in the input (the HSI) of the 3-D CNN classifier, the classification rate decreases. Usually, the curse of dimensionality can limit the generalization ability of the 3-D CNN classifier and gives rise to the over-fitting problem. Therefore, in order to overcome the curse of dimensionality and avoid the over-fitting in the semi-supervised 3-D CNN model, we propose a semi-supervised band selection approach, which aims to find the most informative, discriminative and distinctive spectral bands, while preserving the physical meaning of the original data in the HSI. Hence, instead of using the full set of spectral bands to train the semi-supervised 3-D CNN model, we only use a reduced subset of relevant bands. Thus, the trained model would have better classification accuracies. Moreover, the computational complexity is reduced during training since the convolutional operations would be reduced in the internal layers of the network.

Formally, we approximate a labelling function $f(x)$ from N labelled training samples $\{(x_i, f(x_i))\}_{i=1,\dots N}$, where $x_i = [x_i^1, x_i^2, \dots, x_i^D]$ is a pixel of a very high dimension D (D is the number of spectral bands), with a function $f'(\phi(x))$ where Φ is a feature selection operator, which seeks to select relevant features from X, i.e. $\phi(x)$ contains only $d \ll D$ dimensions.

Fig. 1 shows the general architecture of the proposed approach. More precisely, the input is an HSI $X = [x_1, \dots, x_N] \in \mathbb{R}^{D\times N}$, where x_i represents a spectral vector of a pixel. D is the number of spectral bands $B = \{B_1, B_2, \dots, B_D\}$ and N is the number of pixels. Therefore, the first phase, i.e, ADR aims to find a set relevant spectral features (discriminative, informative, non redundant, and distinctive), which can be denoted as $X' = [x_1, \dots, x_N] \in \mathbb{R}^{d\times N}$, where d is the number of selected bands ($d \ll D$). In the second phase, we define a spatial window of each pixel x_i with a size of $p \times p \times d$, i.e, x_i is represented as a 3-D cube. The aim is to extract spectro-spatial deep features and to learn a better representation of each pixel using a 3-D CNN encoder-decoder in order to improve the classification accuracy. Therefore, the model uses a 3-D convolution operation to preserve the spectro-spatial features of each pixel. Finally, we use the *softmax* function to achieve the spectro-spatial classification based on labelled and unlabelled samples.

3.1. Adaptive dimensionality reduction (ADR)

The main objective of the proposed adaptive dimensionality reduction (ADR) is to reduce the high dimensionality of HSIs by finding most relevant spectral bands. Therefore, we propose a semi-supervised approach which aims to seek the relevant bands with high Discrimination, high Information, and low Redundancy criterion (DIR) (Feng et al., 2015).

Formally, given an HSI denoted by $X = [x_1, \dots, x_N] \in \mathbb{R}^{D\times N}$, where x_i represents a spectral vector of a pixel. D is the number of spectral bands $B = \{B_1, B_2, \dots, B_D\}$ and N is the number of pixels. Each column X_j of X^t corresponds to a spectral band j. ADR seeks to select a reduced set of d spectral bands based on DIR criterion, where $d \ll D$. Let consider a set of p training samples, $p = u + l$, where u and l are the numbers of unlabeled and labeled training samples respectively. Let $X_j^l \subset X_j$ correspond to labeled samples and $X_j^u \subset X_j$ correspond to unlabeled samples.

ADR uses the DIR criterion to select most discriminative, informative and distinctive spectral bands from HSI. The first step aims to find the top-d discriminative spectral bands using only labeled samples as follows:

$$\max_{S\in R} \frac{1}{d} \sum_{j\in S} I(X_j^l; C) \tag{1}$$

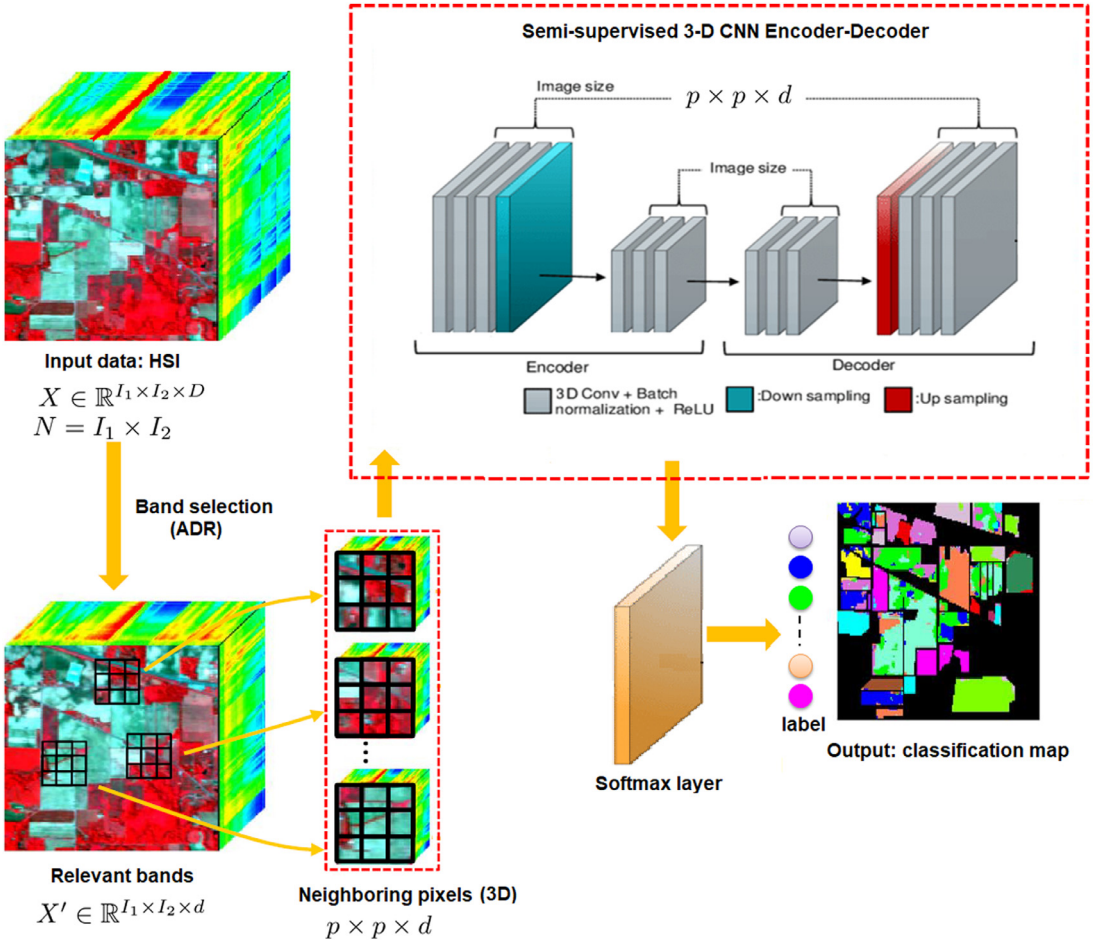

Fig. 1. Architecture of the proposed approach.

where R contains all the subsets of B with d elements, d is the number of selected bands, $I(.)$ is the mutual information (MI) between the labeled training sample X_j^l and the class label C, which seeks to measure the shared information between X_j^l and C. It is defined by:

$$I(X_j^l; C) = H(X_j^l) + H(C) - H(X_j^l; C) \tag{2}$$

where $H(.)$ is the measure of entropy.

However, the selected bands can be redundant. Therefore, the multivariable MI (MMI) (Matsuda, 2000) can be used to reduce the redundancy between the selected spectral bands, as follows:

$$\min_{S \in R} \frac{2}{(d^2 - d)} \sum_{j,m \in S; j<m} I(X_j^l; X_m^l; C) \tag{3}$$

The second step finds the informative spectral bands reducing the redundancy using unlabeled training samples as follows:

$$\max_{S \in R} \frac{1}{d} \sum_{j \in S} H(X_j^u) \tag{4}$$

$$\min_{S \in R} \frac{2}{(d^2 - d)} \sum_{j,m \in S; j<m} I(X_j^u; X_m^u) \tag{5}$$

Therefore, by combining Eqs. (1), (3), (4) and (5), we can select the top-d relevant bands with low redundancy. This can be achieved by maximizing the following expression:

$$\max_{S \in R} \left[\left(\frac{1}{d} \sum_{j \in S} I(X_j^l; C) - \frac{2}{(d^2 - d)} \sum_{j,m \in S; j<m} I(X_j^l; X_m^l; C) \right) + \lambda \left(\frac{1}{d} \sum_{j \in S} H(X_j^u) - \frac{2}{(d^2 - d)} \sum_{j,m \in S; j<m} I(X_j^u; X_m^u) \right) \right] \tag{6}$$

where λ is used as a parameter to adjust the relative importance of unlabeled and labeled training samples. Resolving Eq. (1) would give the best set of selected bands $SB_{(ADR)} = \{B_{<1>}, B_{<2>}, \ldots, B_{<d>}\}$ that will be considered as the input of the semi-supervised 3D-CNN model.

3.2. Spatial-spectral feature extraction based on 3-D CNN

In this section, we first introduce the 1-D CNN for spectral feature extraction, then the 2-D CNN-based spatial feature extraction. After that, we present the 3-D CNN-based spatial-spectral feature extraction.

3.2.1. 1-D CNN feature extraction

A CNN (Lawrence, Giles, Tsoi, & Back, 1997) is essentially based on three concepts: (1) local receptive field; (2) shared weights and biases; and (3) activation and pooling. In fact, only neurons within small regions of the input layer connect to neurons in the hidden layers. These regions are referred to as local receptive fields which are used to create feature maps and passed from the input layer to hidden layers using convolution operations (filters). Moreover, the weights and biases are the same for all the hidden neurons in a given layer. This means that all hidden neurons are detecting the same feature of different regions of the image. The output of each neuron is transformed using an activation function called

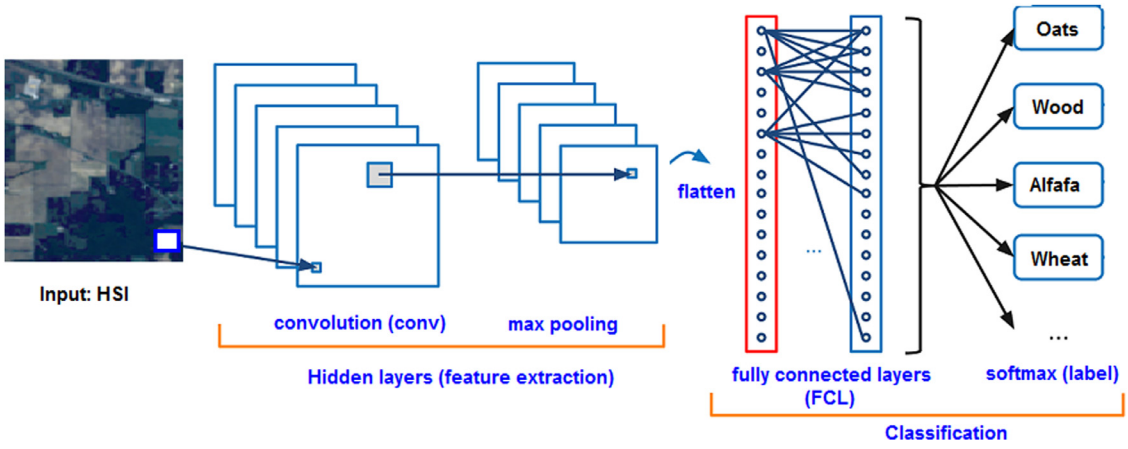

Fig. 2. The general process of CNN.

Rectified Linear unit (ReLu). Furthermore, a pooling phase is used to reduce the dimensionality of the feature maps by combining the outputs of neighboring neurons into a single output. This helps reducing the number of parameters to be learned. Fig. 2 shows a typical CNN architecture, which is alternately stacked by convolutional layers (C) and max-pooling layers (MP), and then followed by a set of fully connected layers (FC).

Formally, let consider a CNN that is composed of N layers. The output state of the ith layer is denoted by V^i, $i = 1, \ldots, N$. Moreover, the input data can be denoted by V^0. For each layer, we use the weight matrix W^i and the bias b^i as training parameters. In fact, the weight matrix W^i is used to connect the ith layer with its previous layer $(i-1)$. Therefore, the output state value of the jth feature map of the ith layer can be expressed as follows:

$$V_j^i = \sum_p \phi(V_p^{i-1} * w_{jp}^i + b_j^i) \tag{7}$$

where $V_p^{i-1} \in \mathbb{R}^{y \times z}$ denotes the feature map of the $(i-1)$th layer, and $*$ is the convolution operation. $w_{jp}^i \in \mathbb{R}^{w \times w}$ denotes the convolutional kernel for V_p^{i-1}. $\phi(.)$ is a non linear activation function. Therefore, the whole process of the CNN can be defined as follows:

$$V^i = mpool(\phi(V^{i-1} * W^i + b^i)) \tag{8}$$

where $mpool(.)$ denotes the max-pooling operation, which seeks the dominant features in the local windows for each feature map (Scherer, Müller, & Behnke, 2010).

In the CNN, several C layers and MP layers can be used to build the hierarchical feature extraction model. Afterwards, these features can be stacked into a 1-D feature vector in the FC layer. Therefore, the feature vector learned by the ith FC layer is obtained as follows:

$$V^i = \phi\left(V^{i-1} * W^i + b^i\right) \tag{9}$$

3.2.2. 2-D CNN-based spatial feature extraction

2-D CNN aims to extract spatial features directly from the raw input HSI. In fact, it requires a convolution operation for each input of the network, and the spectral bands of HSI can be considered as the input of the network. However, a large number of parameters is required due to a large number of spectral bands of HSI, which increases the computational costs. Fig. 3 shows an illustration of the 2D convolution operation.

Each 2D convolution layer contains one C layer and one MP layer, and the input hyperspectral data are convolved using 2D kernels. Then, the activation function can be applied to produce the feature maps. Formally, the 2D convolution operation can be formulated as:

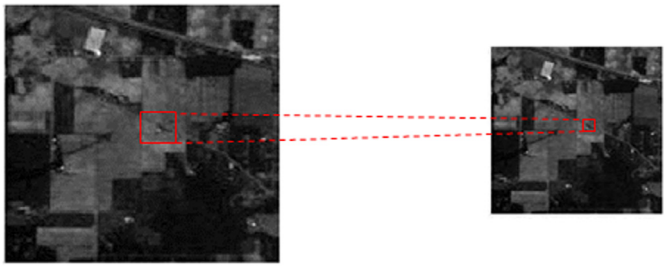

Fig. 3. Illustration of 2D convolution operation.

$$v_{ij}^{xy} = \phi\left(\sum_h \sum_{l=0}^{L_{i-1}} \sum_{k=0}^{K_{i-1}} v_{(i-1)h}^{(x+l)(y+k)} * w_{ijh}^{lk} + b_{ij}\right) \tag{10}$$

where v_{ij}^{xy} is the value of the neuron at position (x, y), h is the index of the feature map in the $(i-1)$th layer connected to the current jth feature map. w_{ijh}^{lk} is the weight of kernel at position (l, k) connected to the hth feature map, with K_i and L_i are the height and width of the convolution kernel.

3.2.3. 3-D CNN-based deep feature extraction

Generally, the 1-D convolution operation aims to extract spectral features while 2-D CNN extracts spatial features of each hyperspectral pixel. 3-D CNN seeks to extract both spatial and spectral features from the 3-D hyperspectral data. Fig. 4 shows the main principle of a 3-D convolution operation. Formally, the 3-D convolution operation can be defined as follows:

$$v_{ij}^{xyz} = \phi\left(\sum_h \sum_{l=0}^{L_{i-1}} \sum_{k=0}^{K_{i-1}} \sum_{m=0}^{R_{i-1}} v_{(i-1)h}^{(x+l)(y+k)(z+m)} * w_{ijh}^{lkm} + b_{ij}\right) \tag{11}$$

where v_{ij}^{xyz} is the value of the neuron at position (x, y, z), h is the index of the feature map in the $(i-1)$th layer connected to the current jth feature map. w_{ijh}^{lkm} is the weight of kernel at position (l, k, m) connected to the hth feature map, with K_i and L_i are the height and width of the convolution kernel. R_i is the size of the 3D kernel along the spectral dimension, and b_{ij} is the bias. We use the $ReLu()$ activation function (Krizhevsky, Sutskever, & Hinton, 2012), which can be formulated as follows:

$$ReLu(v) = \phi(v) = \max(0, v) \tag{12}$$

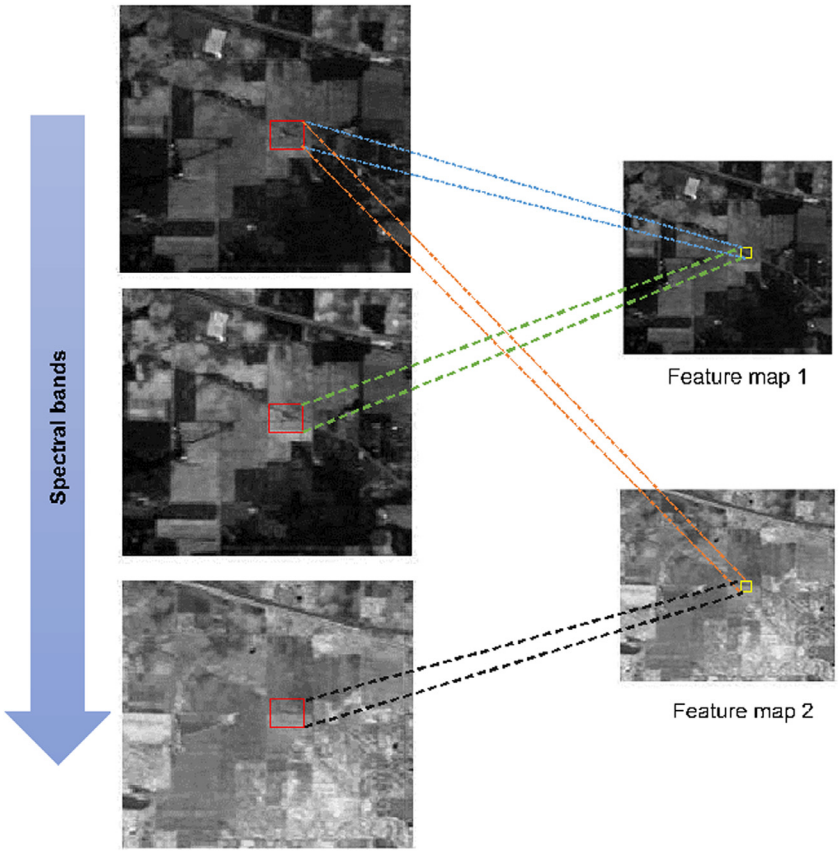

Fig. 4. Illustration of 3D convolution operation.

3.3. Proposed semi-supervised 3-D CNN for HSI classification

In this section, we present our proposed semi-supervised 3-D CNN for HSI spatial-spectral classification. HSI is represented as a 3-D cube, which contains two spatial dimensions (size of each spectral band) and one spectral dimension (number of spectral bands). Each pixel is considered along with its neighboring pixels, i.e., cubes of size ($p \times p \times d$). These cubes (images) are considered as input data of the semi-supervised 3-D CNN. $p \times p$ is the window size (spatial dimension), and d is the number of the selected relevant bands. In supervised learning, we use the *softmax* loss function (Simonyan & Zisserman, 2015) to train the model, with a number of neurons which is equal to the number of HSI classes to be classified. Let consider N labeled training samples $(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)$, where x_i is a pixel and t_i is its label. Each pixel x_i can be defined as $x_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$, where d is the length of the spectral vector. The *softmax* loss function can be formulated as:

$$C_s = \frac{-1}{N} \sum_{i=1}^{N} logP(y_i = t_i | x_i) \tag{13}$$

where N is the number of training samples, t_i is the label of the ith training sample, and y_i is a random variable corresponding to the label of a sample pixel. However, to be robust, this requires a lot of labeled training samples to train a good 3D CNN model. Therefore, we extend the convolutional encoder-decoder (Badrinarayanan, Kendall, & Cipolla, 2016) with supervised 3D CNN to perform a semi-supervised 3D CNN, which takes into account labeled and unlabeled samples by preserving the spatial features. Formally, let consider M unlabeled training samples $x_{N+1}, x_{N+2}, \ldots, x_{N+M}$. The convolutional encoder-decoder contains two main mappings: an encoder mapping f and a decoder mapping g. f seeks to adopt the feed-forward process of the CNN while g contains convolution operations and unsampling. Usually, a convolutional encoder-decoder aims to minimize the difference between x_i and the predicted input $\hat{x}_i$ (reconstructed decoder) as follows:

$$C_r^{(0)} = \frac{\lambda}{M} \sum_{i=N+1}^{N+M} \| \hat{x}_i - x_i \|_2^2 \tag{14}$$

where $\hat{x}_i$ is the reconstructed decoder input.

In our case, the semi-supervised 3D CNN based on convolutional encoder-decoder has three paths for the unlabeled and labeled training samples: clean encoding, noisy encoding, and decoding:

- **The clean encoding path:** Labeled and unlabeled training samples are processed through the clean encoding path to calculate hidden variables z_i^l, $l = 1, \ldots, L$. Therefore, this function can be expressed as follows:

$$z_i^{(1)}, \ldots, z_i^{(L)} = \text{Encoder}_{clean}(x_i) \tag{15}$$

- **The noisy encoding path:** Labeled and unlabeled training samples are corrupted using the Gaussian noise. Moreover, they are transformed to abstract representations $\check{z}_i^l$ by using the noisy encoder. Formally, we can define this step as:

$$\check{x}_i, \check{z}_i^{(1)}, \ldots, \check{z}_i^{(L)} = \text{Encoder}_{noisy}(x_i) \tag{16}$$

- **The decoder:** It seeks to reconstruct the predicted $\hat{x}_i$ such that they are as much close as possible to x_i. This can be expressed as follows:

$$\hat{x}_i = \text{Decoder}(\check{z}_i^{(1)}, \ldots, \check{z}_i^{(L)}) \tag{17}$$

Therefore, our proposed semi-supervised 3D CNN aims to use both cost functions, i.e. *softmax* function for the labeled training samples and convolutional encoder-decoder for the unlabeled

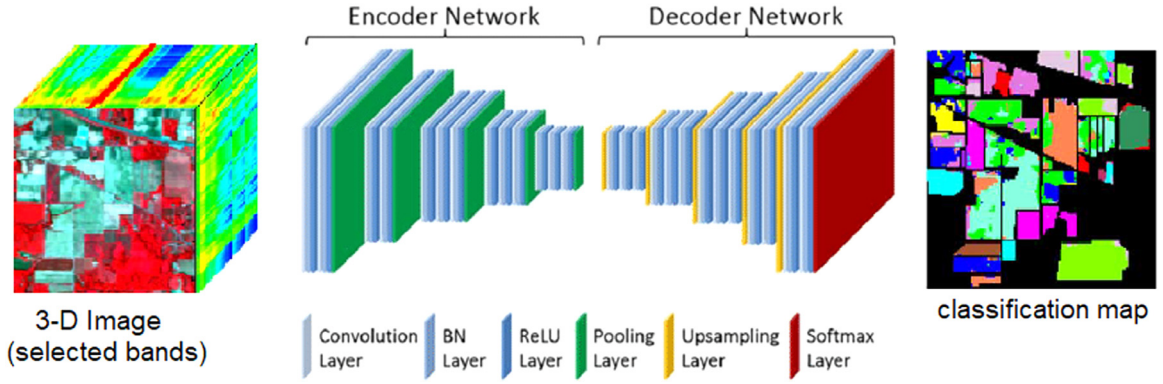

Fig. 5. Architecture of the semi-supervised 3-D CNN-based HSI classification.

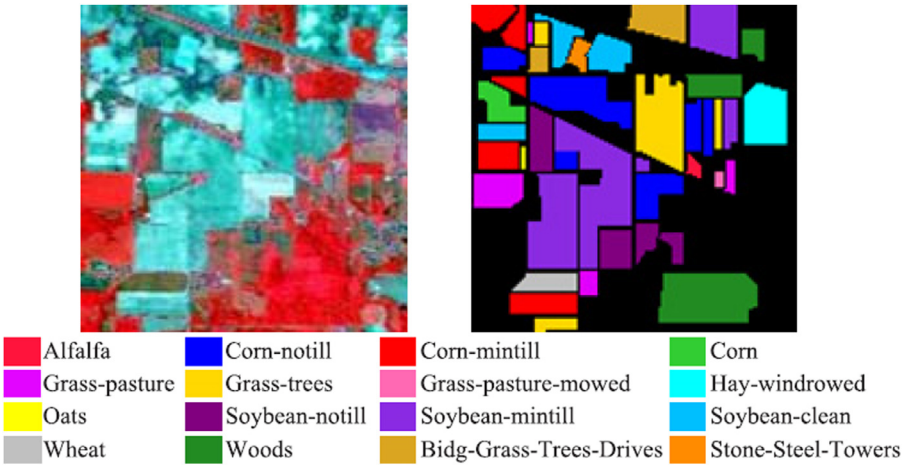

Fig. 6. Indian Pines scene. (a) Color composite image. (b) Ground truth classes.

training samples. Using Eqs. (13) and (14), we can obtain this formula:

$$C_e = C_s + C_r^{(0)} = \frac{-1}{N}\sum_{i=1}^{N} logP(y_i = t_i|x_i) + \frac{\lambda}{M}\sum_{i=N+1}^{N+M} \| \hat{x}_i - x_i \|_2^2 \tag{18}$$

Using this semi-supervised 3D CNN, we can learn the proposed network model and spectro-spatial features simultaneously from the HSI.

Fig. 5 shows the architecture of the proposed semi-supervised 3-D CNN for HSI spatial-spectral classification. In the encoder network of the 3-D CNN, we have the following operations: 3-D convolution, batch normalization and pooling. In the decoder network of the 3-D CNN, we have the following operations: 3-D convolution, batch normalization, and unpooling. The final layer is the *softmax* function.

4. Experiments and discussion

4.1. Hyperspectral data description

In this paper, three real HSIs are involved to evaluate the performance and effectiveness of the proposed approach. The first HSI is the Indian Pines HSI collected by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor, which represents the northwestern Indiana. It consists of 145×145 pixels with a spatial resolution of $20\,m$ per pixel and 220 spectral bands in the wavelength range from 0.4 to 2.5 µm. The ground truth contains 16 classes. Fig. 6(a) and (b) shows the color composite of the Indian Pines HSI and its ground truth.

The second HSI is the Pavia University data gathered by the Reflective Optics System Imaging Spectrometer (ROSIS-03) sensor, which consists of 610×610 pixels, and 115 spectral bands in the range from 0.43 to 0.86 µm, with a spatial resolution of 1.3 m. The ground truth contains 9 classes. Fig. 7(a) and (b) shows the color composite image of the Pavia University HSI and its ground truth.

The last HSI is the Salinas image collected by the AVIRIS sensor over Salinas, California, which consists of 512×217 pixels with a spatial resolution of $3.7\,m$ and 224 spectral bands. The ground truth contains 16 classes. Fig. 8(a) and (b) shows the color composite image of the Salinas data and its ground truth.

4.2. Comparison with other DR methods

In our proposed approach, ADR is developed to select relevant spectral bands while preserving useful spectral information for HSI classification. In order to evaluate the effectiveness and the performance of the proposed approach, we randomly selected $T =$ 10%, 20%, and 30% training samples from each HSI class. We used the linear SVM classifier in the experiments. The ADR approach is compared with three semi-supervised band selection methods: spectral-spatial hypergraph model (SSHM) (Bai et al., 2015), affinity propagation (SSAP) (Jiao et al., 2015), and band clustering (SSBC) (Su et al., 2011). Three supervised band selection methods are also considered: band weighting and interclass distance (SWID) (Yan et al., 2016), information measures (SIM) (Sotoca & Pla, 2010), and classification with SVM (BSVM) (Pal & Foody, 2010). We selected $d = 3, \ldots, 15$ spectral bands from the three real HSIs to in-

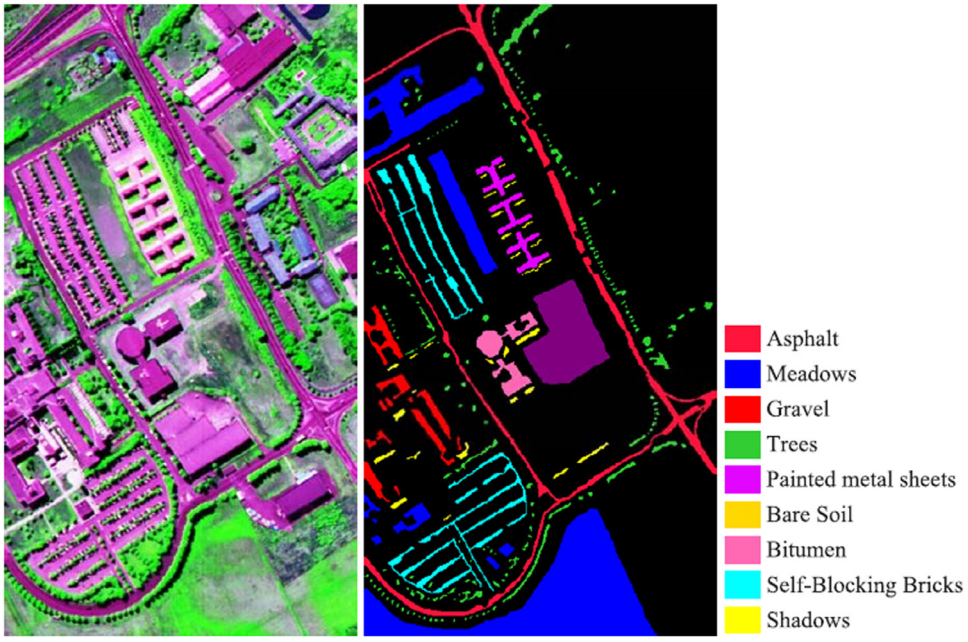

Fig. 7. Pavia University scene. (a) Color composite image. (b) Ground truth classes.

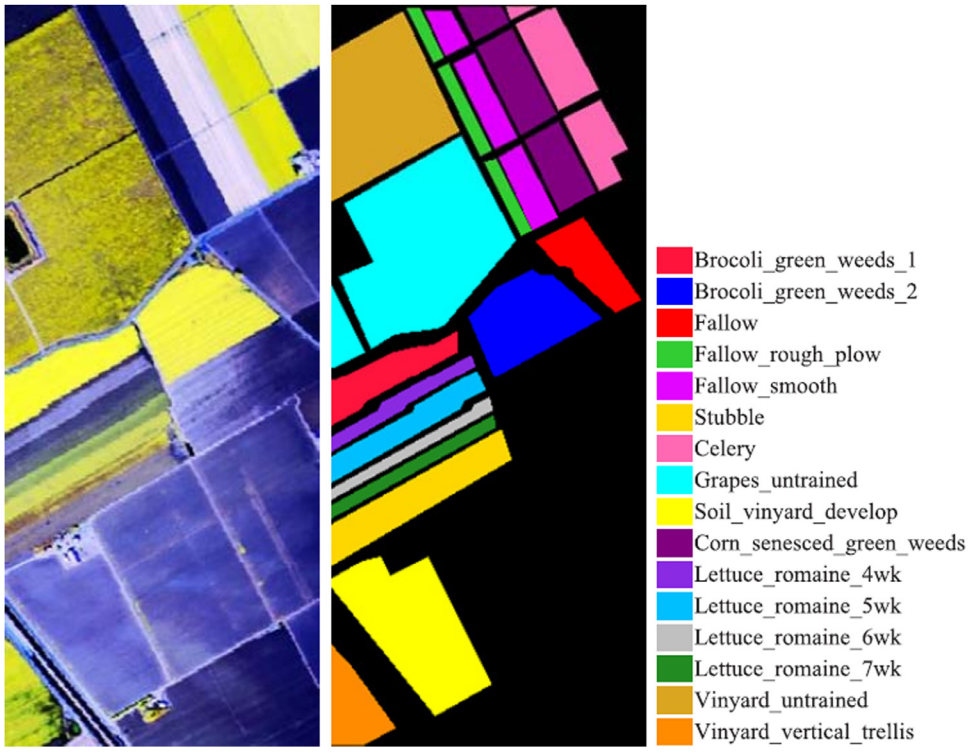

Fig. 8. Salinas scene. (a) Color composite image. (b) Ground truth classes.

vestigate the impact of varying the number of bands to be selected on the effectiveness of each method.

The performances of these DR techniques are reported in Fig. 9. The obtained OAs show that ADR outperforms SSHM, SSAB, SWID, SSBC, SIM, and BSVM. In fact, the OAs with ADR are 93.21% for Indian Pines, 94.57% for Pavia University, and 93.89% for Salinas, with $T = 30\%$.

Moreover, Fig. 10 reports the selected spectral bands produced by the different methods using the three HSIs. The color bars at the bottom of Fig. 10 give the correspondence between the spectral bands Ids and their corresponding wavelengths. These color bars help analysing the results produced by the various band selection methods. For all the datasets, we can see that the proposed approach ADR succeeds in selecting representative spectral bands throughout the spectrum, whereas the rest of the methods select almost spectral bands within reduced regions of the spectrum. For instance, for the Indian Pines, SSHM selects bands [69 ; 71 ; 73 ; 74 ; 75 ; 79 ; 80 ; 82 ; 83 ; 84 ; 85 ; 88], SSAP selects bands [105 ; 107 ; 109 ; 110 ; 112 ; 113 ; 115 ; 116 ; 119 ; 120 ; 122 ; 123] and BSVM selects bands [73 ; 75 ; 76 ; 77 ; 78 ; 81 ;

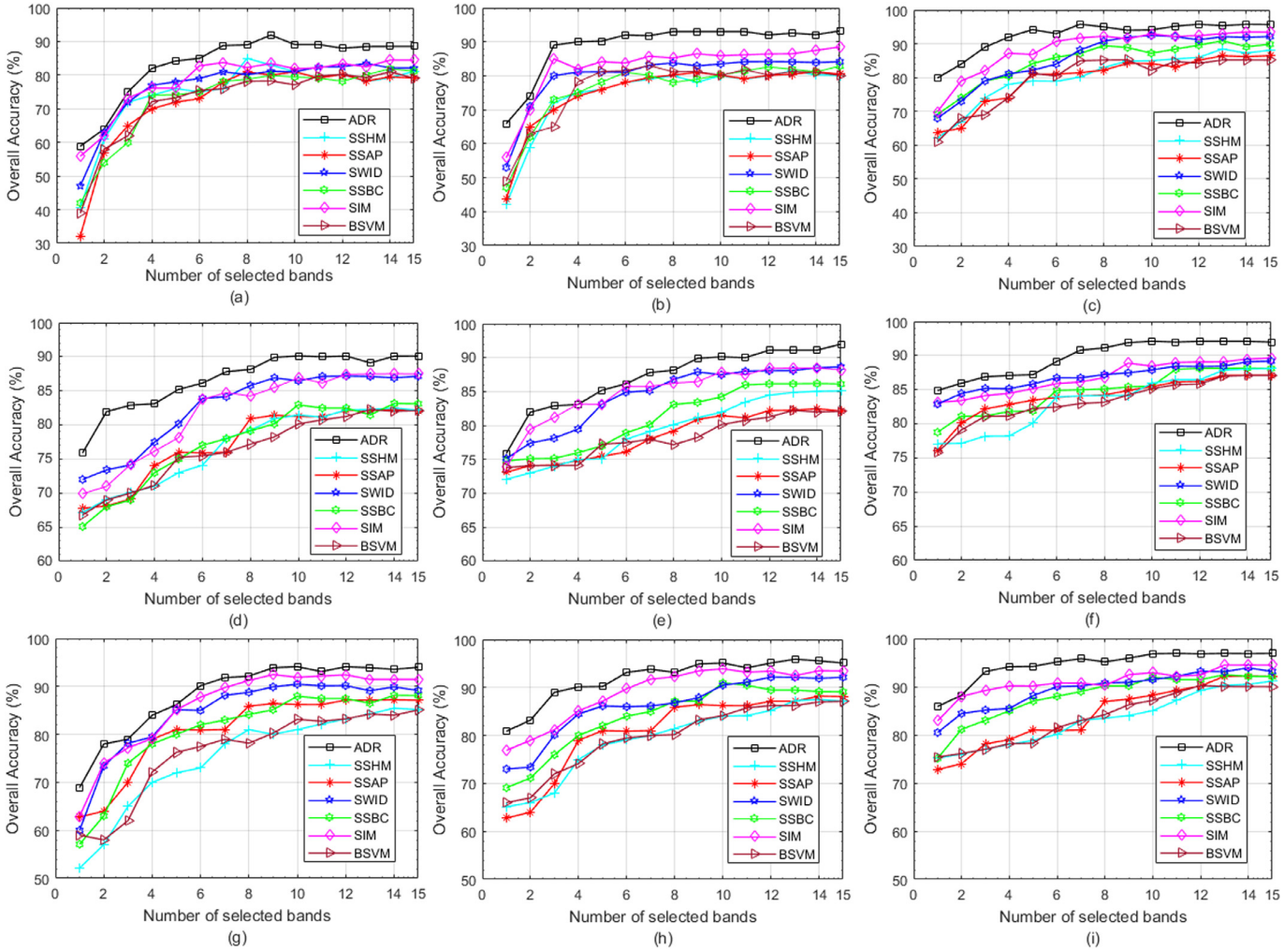

Fig. 9. Comparison of different DR techniques; from left to right, 10%, 20%, and 30% training samples are selected, respectively. (a)-(c) Classification results on Indian Pines. (d)-(f) Classification results on Pavia University. (g)-(i) Classification results on Salinas.

82 ; 84 ;85 ; 88 ; 90 ; 91] which are all located within reduced regions of the spectrum. On the contrary, the proposed approach ADR selects bands [7 ; 17 ; 26 ; 42 ; 59 ; 86 ; 121 ; 147 ; 165 ; 181 ; 193 ; 205] which cover all the spectrum.

4.3. Architecture of the proposed semi-supervised 3-D CNN model

The proposed semi-supervised 3-D CNN seeks to learn the deep spatial (pixels neighborhoods) and spectral features by itself. These learned deep features were used in HSI classification. In order to define the parameters of the semi-supervised 3-D CNN model, i.e., the number of convolution layers, the kernel size, the learning rate, and the number of nodes in the different hidden layers, we used the trial and error technique (Sosna, Trevinyo-Rodríguez, & Velamuri, 2010). The selected bands $SB_{(ADR)}$ are used as the length of pixel vectors, which are supposed to preserve the most important spectral-spatial information of the whole HSI. After many runs, we fixed $d = 12$ spectral bands. Table 1 reports the obtained architectures of the semi-supervised 3-D CNN for each HSI, containing three convolution layers and two pooling layers. We fixed $p = 25$ neighboring pixels for the three data sets, i.e. $25 \times 25 \times 12$ ($p \times p \times d$) neighbors of each pixel is used as the input data.

Table 1
Parameter setting of the architecture of the semi-supervised 3-D CNN model.

HSI	Hidden layer	Convolution	ReLu	Pooling
Indian pines	1	$4 \times 4 \times 32 \times 128$	Yes	2×2
	2	$5 \times 5 \times 32 \times 192$	Yes	2×2
	3	$4 \times 4 \times 32 \times 256$	Yes	
Pavia University	1	$4 \times 4 \times 32 \times 32$	Yes	2×2
	2	$5 \times 5 \times 32 \times 64$	Yes	2×2
	3	$4 \times 4 \times 32 \times 128$	Yes	
Salinas	1	$4 \times 4 \times 32 \times 32$	Yes	2×2
	2	$5 \times 5 \times 32 \times 64$	Yes	2×2
	3	$4 \times 4 \times 32 \times 128$	Yes	

For the linear regression with *softmax* function, the learning rate is fixed to 5.10^{-3} for Indian Pines, 10^{-2} for Pavia University, and 10^{-3} for Salinas. After several tests, we fixed the training epoch to 150, 200 and 300 for Indian Pines, Pavia University and Salinas respectively.

4.4. Classification performance analysis

In this section, we aim to evaluate our proposed approach for HSI spatial-spectral classification. We first used different band selection methods and the semi-supervised 3-D CNN model,

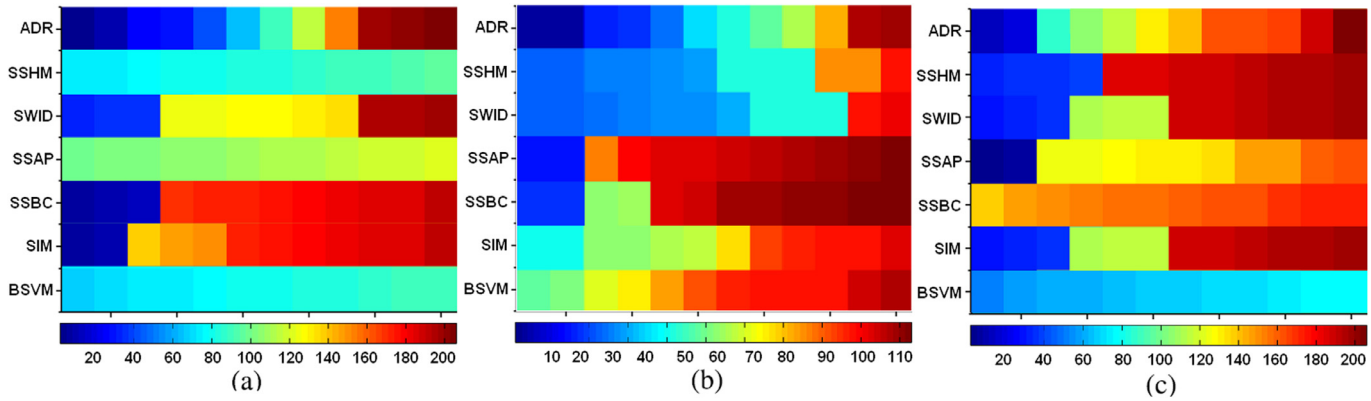

Fig. 10. Selected spectral bands with different band selection methods. (a) Indian Pines, (b) Pavia University, (c) Salinas ($d = 12$ selected bands). The blue color denotes the lower spectral band indexes and the red color indicates higher spectral band indexes.

Table 2
Classification Accuracy obtained by the semi-supervised 3-D CNN using different band selection methods (Indian Pines).

Class	Samples		Band selection methods						
	Train	Test	BSVM	SIM	SSHM	SSBC	SSAP	SWID	ADR
Alfalfa	5	49	95.92	91.84	93.88	87.76	91.84	93.88	97.96
Building-G	38	342	95.03	92.69	92.98	90.64	92.69	90.35	96.49
Corn	23	211	93.84	97.63	95.73	96.68	95.26	98.10	99.53
Corn-M	83	751	92.41	96.14	93.21	95.87	93.08	96.27	97.47
Corn-N	143	1291	92.95	94.89	94.11	95.66	93.73	96.67	97.21
Grass-P-W	5	21	95.24	90.48	76.19	85.71	90.48	90.48	95.24
Grass-P	50	447	93.96	94.41	93.74	93.96	93.29	94.63	98.88
Grass-T	75	672	96.73	97.02	97.02	95.54	92.56	95.54	98.51
Hay-W	49	440	96.59	97.27	96.59	96.62	95.45	97.05	97.73
Oats	2	18	94.44	88.89	94.44	83.33	88.89	88.89	94.44
Soybeans-C	62	552	94.20	97.64	96.74	97.10	96.56	97.64	97.83
Soybeans-M	247	2221	93.65	97.25	96.80	97.03	96.76	97.07	97.70
Soybeans-N	97	871	99.31	96.10	93.34	93.80	93.69	93.80	97.24
Stone-S	10	85	94.12	92.94	95.29	94.12	94.12	98.82	96.47
Wheat	21	191	95.29	97.38	95.29	97.38	95.81	97.91	95.81
Woods	130	1164	96.22	98.11	97.25	98.11	96.99	98.20	99.83
OA (%)	–	–	94.82	96.45	95.45	95.94	94.96	96.35	**97.89**
AA (%)	–	–	94.99	95.04	93.91	93.70	93.82	95.33	**97.39**
$k \times 100$	–	–	94.87	96.12	94.78	94.96	94.82	96.12	**98.72**
Time (s)	–	–	428	385	329	336	417	327	371

to assess the effectiveness of the proposed ADR approach (Section 4.4.1). Then, we compare the proposed method with other CNN-based methods in Section 4.4.2.

For all HSIs, we randomly choose 10% of pixels as training samples and 90% as testing samples. Therefore, we used 90% from the training samples to learn the CNN model parameters, i.e., weights W and biases b. The remaining 10% of the training samples are used to check if there is any overfitting.

In order to evaluate the classification results and to compare the effectiveness of the proposed model, three metrics have been used, including overall accuracy (OA), average accuracy (AA), and kappa coefficient (K). Moreover, the experiments were executed fifteen times with different random training samples. After that, the overage OA, AA, and K are reported.

4.4.1. HSI Classification based on semi-supervised 3-D CNN model and different DR methods

In this section, we compare the proposed method with other DR methods using the semi-supervised 3-D CNN, including, BSVM, SIM, SSHM, SSBC, SSAP, and SWID.

Table 2 shows the obtained OAs for the Indian Pines HSI. Therefore, we can observe that the proposed approach ADR gives better classification rates, compared to other DR methods. In fact, the obtained OA is 97.89%, AA is 97.39%, and k is 98.72%. However, the SWID method gives the best classification performance for the following two classes: 'Stone-S' and 'Wheat', with OAs of 98.82% and 97.91%, respectively. For the remaining 14 classes of Indian Pines, the proposed method ADR gives better classification rates.

Table 3 reports the obtained classification results for the Pavia University data set. From this table, we can see that the OA with ADR is 98.45%, AA is 98.60%, and k is 98.53%. Moreover, the ADR approach gives better classification rates for 8 out of 9 classes of Pavia University data set. Also, most classification rates are greater than 98%. For the class 'Shadows', the SWID performs slightly better than ADR, which a classification rate of 98.59%.

For the Salinas data set, ADR gives also better classification results than other DR methods (Table 4). In fact, the OA of ADR is 98.29%, AA is 98.28%, and k is 98.16%. Moreover, ADR gives better classification results for 14 out of 16 classes. The two other classes, 'Brocoli-G-W-1' and 'Fallow', they are better classified with SIM, with a classification rate of 97.84% and 97.18%, respectively. More-

Table 3
Classification Accuracy obtained by the semi-supervised 3-D CNN using different band selection methods (Pavia University).

Class	Samples		Band selection methods						
	Train	Test	BSVM	SIM	SSHM	SSBC	SSAP	SWID	ADR
Asphalt	132	6499	90.78	97.09	95.39	96.16	96.93	97.24	98.01
Bare-S	100	4929	93.93	94.74	93.52	95.33	95.35	97.07	99.41
Bitumen	26	1304	95.85	97.39	98.92	96.62	95.09	97.54	98.92
Gravel	41	2058	95.67	97.18	96.20	96.79	96.93	97.18	98.15
Meadows	373	18321	96.82	98.24	96.66	96.06	96.02	97.00	98.20
Painted-M	26	1319	95.52	94.84	97.72	98.10	95.52	96.13	99.31
Self-B	73	3609	96.42	98.36	96.97	98.50	93.65	94.90	98.08
Shadows	19	928	94.82	96.98	95.79	97.62	91.59	98.59	98.06
Tree	61	3003	93.24	98.56	94.13	97.06	96.96	97.40	99.23
OA (%)	–	–	95,08	97.47	96.00	96.42	95.85	96.93	**98.45**
AA (%)	–	–	94.78	97.04	96.15	96.92	95.34	97.01	**98.60**
$k \times 100$	–	–	94.91	97.12	96.06	96.63	95.41	96.95	**98.53**
Time (s)	–	–	298	195	149	189	201	175	265

Table 4
Classification Accuracy obtained by the semi-supervised 3-D CNN using different band selection methods (Salinas).

Class	Samples		Band selection methods						
	Train	Test	BSVM	SIM	SSHM	SSBC	SSAP	SWID	ADR
Brocoli-G-W-1	200	1809	93.97	97.84	95.63	96.18	96.73	97.29	96.73
Brocoli-G-W-2	372	3354	96.72	95.52	97.01	95.82	96.12	96.42	98.50
Fallow	197	1779	90.44	97.18	92.69	94.37	94.94	95.50	96.06
Fallow-R-P	139	1255	94.42	95.61	94.42	95.21	96.01	95.61	98.80
Fallow-S	267	2411	95.81	97.88	95.81	96.22	97.05	97.88	97.88
Stubble	395	3564	95.23	98.59	94.66	95.23	95.51	95.79	98.87
Celery	357	3222	95.43	96.52	96.05	96.05	96.30	96.52	96.58
Grapes-U	1127	10144	96.55	98.58	96.62	96.68	97.47	97.67	98.61
Soil-V-D	620	5583	95.34	97.13	95.70	96.41	96.59	96.95	98.92
Corn-S-G-W	327	2951	94.84	94.91	96.27	6.61	97.28	97.96	98.30
Lettuce-R-4wk	106	962	95.84	96.88	97.92	93.76	94.80	96.88	98.96
Lettuce-R-5wk	192	1735	92.21	96.82	92.79	94.52	95.67	96.25	99.71
Lettuce-R-6wk	91	825	87.63	97.33	91.27	92.48	94.90	96.12	98.78
Lettuce-R-7wk	107	963	93.35	95.84	91.27	93.35	95.01	96.05	98.96
Vineyard-U	726	6542	98.01	95.13	96.05	96.33	97.09	97.40	98.01
Vineyard-V-T	180	1627	86.47	93.85	92.62	94.46	95.08	95.69	98.77
OA (%)	–	–	94.83	97.25	95.55	95.88	96.53	96.94	**98.29**
AA (%)	–	–	93.71	96.78	94.80	95.23	96.03	96.62	**98.28**
$k \times 100$	–	–	94.25	96.90	95.48	95.61	96.35	96.84	**98.16**
Time (s)	–	–	321	301	195	210	245	223	289

over, we can see that the DR methods SWID and SAAP are very close in terms of classification rates.

According to the obtained results, we can affirm that the proposed approach ADR is more effective than many other state-of-the-art DR methods. Also, ADR can select relevant spectral bands of HSI while preserving the physical meaning of data, i.e., the spectral and spatial information simultaneously, and it also provides good spectral-spatial classification by exploiting the semi-supervised 3-D CNN model.

Table 5
Obtained classification results using different DL-based methods (Indian Pinces).

	DL-based methods				
	PCA-DBN	BS-DBN	PCA-CNN	GF-CNN	SS-3DCNN
OA (%)	95.21	96.01	96.41	96.85	**97.89**
AA (%)	95.52	96.22	96.12	96.54	**97.39**
$k \times 100$	95.39	96.13	96.23	96.61	**97.72**
Time(s)	362	510	425	522	371

4.4.2. Comparative study with DL-based methods

In this section, we compare our proposed method (SS-3DCNN) with other DL-based methods, including: PCA-DBN (Chen et al., 2015), band selection and DBN (BS-DBN) (Zhou et al., 2017), PCA-CNN (Zhao & Du, 2016), and gabor filters (GF) with convolutional filters (GF-CNN) (Chen, Jiao, et al., 2017).

Table 5 reports the obtained OA, AA and K for PCA-DBN, BS-DBN, PCA-CNN, GF-CNN, and the proposed SS-3DCNN on the Indian Pines. Based on the obtained results, we can notice that the classification accuracy values of the proposed approach SS-3DCNN in terms of classification rates OA, AA, and kappa coefficient k are higher than those of the other DL-based methods. In fact, the OA is 97.89% for SS-3DCNN, while it is 96.85% for GF-CNN, and 96.41% for PCA-CNN. The obtained OA of BS-DBN method is 96.01%, which is close to the OA of PCA-CNN. The lowest classification rates are those of PCA-DBN, with an OA of 95.21%.

Moreover, Table 6 gives the obtained classification rates for Pavia University. In fact, the proposed SS-3DCNN gives also better results compared to GF-CNN, PCA-CNN, BS-DBN, and PCA-DBN, with an OA of 98.45%. The lowest classification rates are obtained for the BS-DBN method, which has an OA of 96.24%.

Furthermore, Table 7 shows the classification results of the Salinas data set. We can see also that the proposed approach is more effective than the rest of DL-based methods, with respect to all the criteria OA, AA and k. In fact, OA of SS-3DCNN is 98.29%, AA is 98.28%, and k is 98.16%. The obtained OA with GF-CNN is 97.23%, which is close to SS-3DCNN.

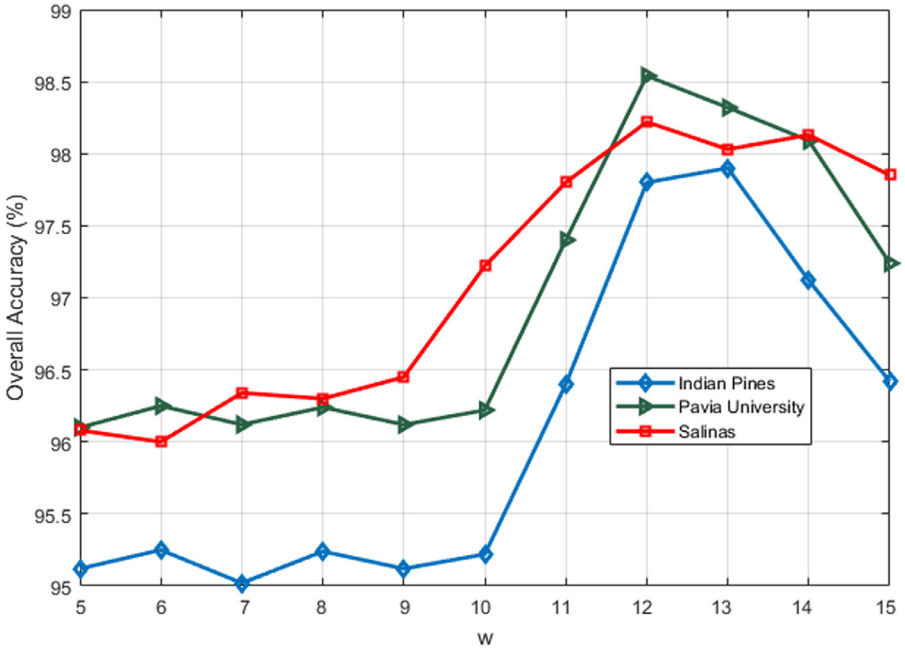

Fig. 11. Sensitivity of the semi-supervised 3-D CNN with respect to the size of the window.

Table 6
Obtained classification results using different DL-based methods (Pavia University).

	DL-based methods				
	PCA-DBN	BS-DBN	PCA-CNN	GF-CNN	SS-3DCNN
OA (%)	96.98	96.24	97.35	97.21	**98.45**
AA (%)	97.02	96.45	97.54	97.36	**98.60**
$k \times 100$	97.00	96.38	97.42	97.30	**98.53**
Time(*s*)	279	302	298	325	286

Table 7
Obtained classification results using different DL-based methods (Salinas).

	DL-based methods				
	PCA-DBN	BS-DBN	PCA-CNN	GF-CNN	SS-3DCNN
OA (%)	96.32	97.01	96.99	97.23	**98.29**
AA (%)	96.50	97.34	96.81	97.45	**98.28**
$k \times 100$	96.47	97.25	96.87	97.35	**98.16**
Time(*s*)	248	310	276	295	289

According to the comparative study, we can affirm that the proposed method gives good results and has very satisfactory classification performances compared to the other spectro-spatial classification methods. Indeed, it makes it possible to reduce the massive size of hyperspectral data while preserving the spectral and spatial information in order to improve the precision of classification. In terms of computation time, the proposed method is quite fast compared to other methods, since the training of the 3-D CNN requires much less time.

4.5. Influence of the spatial windows

The size of the input data (3-D) is an important parameter of the 3-D CNN model. In this subsection, we fix the spectral dimension, i.e. the number of selected bands d, while varying the size of the spatial dimension. Therefore, we run a set of experiments in order to get a suitable window size of the input data.

Fig. 11 shows the obtained classification performances (OA) for all the three data sets using different values W of the half-width size of the spatial window, ranging from 5 to 15. The full-width size of the spatial window is therefore $2W + 1$. According to the obtained results, we can see that for Indian Pines HSI, the classification rate is the best (97, 90%) for $W = 13$. For Pavia University and Salinas HSIs, the obtained results show that the best OAs are obtained for $W = 12$, with OA$= 98,45\%$ for Pavia University, and 98.29% for Salinas. Therefore, we can affirm that when we use the spatial window 25×25, we can obtain better spatial-spectral classification.

5. Conclusion and future works

In this paper, a novel approach for HSI classification based on adaptive dimensionality reduction (ADR) and semi-supervised 3-D convolutional neural network (3-D CNN) has been proposed. The proposed approach seeks to find the most informative, distinctive, and discriminative spectral bands using labeled and unlabeled training samples. Furthermore, a semi-supervised 3-D CNN model is proposed to extract the deep spatial and spectral features from the selected bands using few training samples. The main advantage of this semi-supervised 3-D CNN model is to jointly preserve the spatial and spectral information in order to improve the classification performances of HSI. Finally, these features have been used for classification based on linear regression classifier. Experimental results have shown that the proposed approach is more effective compared to state-of-the-art DL-based classification methods, including CNN-based methods. As future work, we plan to extend this approach for other fields of applications.

Credit authorship contribution statement

Akrem Sellami: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Mohamed Farah:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision. **Imed Riadh Farah:** Writing - review & editing, Supervision. **Basel Solaiman:** Writing - review & editing, Supervision.

References

Agarwal, A., El-Ghazawi, T., El-Askary, H., & Le-Moigne, J. (2007). Efficient hierarchical-pca dimension reduction for hyperspectral imagery. In *Signal processing and information technology, IEEE international symposium on* (pp. 353–356). IEEE.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2016). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*, 2481–2495.

Bai, X., Guo, Z., Wang, Y., Zhang, Z., & Zhou, J. (2015). Semisupervised hyperspectral band selection via spectral–spatial hypergraph model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8*(6), 2774–2783.

Chang, C.-I., & Wang, S. (2006). Constrained band selection for hyperspectral imagery. *Geoscience and Remote Sensing, IEEE Transactions on, 44*(6), 1575–1585.

Chen, P., Jiao, L., Liu, F., Gou, S., Zhao, J., & Zhao, Z. (2017a). Dimensionality reduction of hyperspectral imagery using sparse graph learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10*(3), 1165–1181.

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., Zhao, X., & Jia, X. (2015). Spectral–spatial classification of hyperspectral data based on deep belief network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8*(6), 2381–2392.

Chen, Y., Zhu, L., Ghamisi, P., Jia, X., Li, G., & Tang, L. (2017b). Hyperspectral images classification with Gabor filtering and convolutional neural network. *IEEE Geoscience and Remote Sensing Letters, 14*(12), 2355–2359.

Datta, A., Ghosh, S., & Ghosh, A. (2015). Combination of clustering and ranking techniques for unsupervised band selection of hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8*(6), 2814–2823.

Du, H., Qi, H., Wang, X., Ramanath, R., & Snyder, W. E. (2003). Band selection using independent component analysis for hyperspectral image processing. In *Applied imagery pattern recognition workshop, 2003. Proceedings. 32nd* (pp. 93–98). IEEE.

Fauvel, M., Chanussot, J., & Benediktsson, J. A. (2006). Kernel principal component analysis for feature reduction in hyperspectrale images analysis. In *Signal processing symposium, 2006. NORSIG 2006. Proceedings of the 7th nordic* (pp. 238–241). IEEE.

Fauvel, M., Tarabalka, Y., Benediktsson, J. A., Chanussot, J., & Tilton, J. C. (2013). Advances in spectral-spatial classification of hyperspectral images. *Proceedings of the IEEE, 101*(3), 652–675.

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.

Feng, J., Jiao, L., Liu, F., Sun, T., & Zhang, X. (2016). Unsupervised feature selection based on maximum information and minimum redundancy for hyperspectral images. *Pattern Recognition, 51*, 295–309.

Feng, J., Jiao, L., Zhang, X., & Sun, T. (2014). Hyperspectral band selection based on trivariate mutual information and clonal selection. *IEEE Transactions on Geoscience and Remote Sensing, 52*(7), 4092–4105.

Ghamisi, P., Chen, Y., & Zhu, X. X. (2016). A self-improving convolution neural network for the classification of hyperspectral data. *IEEE Geoscience and Remote Sensing Letters, 13*(10), 1537–1541.

He, X., & Niyogi, P. (2003). Locality preserving projections. In *Advances in neural information processing systems 16* (pp. 153–160). MIT Press.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research, 5*(Nov), 1457–1469.

Hu, W., Huang, Y., Wei, L., Zhang, F., & Li, H. (2015). Deep convolutional neural networks for hyperspectral image classification. *Journal of Sensors, 2015*.

Huang, H., Li, J., & Liu, J. (2012). Enhanced semi-supervised local fisher discriminant analysis for face recognition. *Future Generation Computer Systems, 28*(1), 244–253.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory, 14*(1), 55–63.

Jamshidpour, N., Safari, A., & Homayouni, S. (2017). Spectral–spatial semisupervised hyperspectral classification using adaptive neighborhood. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10*(9), 4183–4197.

Jia, S., Ji, Z., Qian, Y., & Shen, L. (2012). Unsupervised band selection for hyperspectral imagery classification without manual band removal. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5*(2), 531–543.

Jiang, X., Gong, M., Li, H., Zhang, M., & Li, J. (2018). A two-phase multiobjective sparse unmixing approach for hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing, 56*(1), 508–523.

Jiao, L., Feng, J., Liu, F., Sun, T., & Zhang, X. (2015). Semisupervised affinity propagation based on normalized trivariable mutual information for hyperspectral band selection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8*(6), 2760–2773.

Karami, A., Yazdi, M., & Asli, A. Z. (2011). Noise reduction of hyperspectral images using kernel non-negative tucker decomposition. *IEEE Journal of Selected Topics in Signal Processing, 5*(3), 487–493.

Khan, Z., Shafait, F., & Mian, A. (2015). Joint group sparse pca for compressed hyperspectral imaging. *IEEE Transactions on Image Processing, 24*(12), 4934–4942.

Kim, S. B., & Rattakorn, P. (2011). Unsupervised feature selection using weighted principal components. *Expert Systems with Applications, 38*(5), 5704–5710.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks, 8*(1), 98–113.

Li, L., Ge, H., & Gao, J. (2017). A spectral-spatial kernel-based method for hyperspectral imagery classification. *Advances in Space Research, 59*(4), 954–967.

Lu, X., Wu, H., Yuan, Y., Yan, P., & Li, X. (2013). Manifold regularized sparse nmf for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing, 51*(5), 2815–2826.

Ma, X., Wang, H., & Wang, J. (2016). Semisupervised classification for hyperspectral image based on multi-decision labeling and deep feature learning. *ISPRS Journal of Photogrammetry and Remote Sensing, 120*, 99–107.

Matsuda, H. (2000). Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Physical Review E, 62*(3), 3096.

Matteoli, S., Diani, M., & Corsini, G. (2010). A tutorial overview of anomaly detection in hyperspectral images. *IEEE Aerospace and Electronic Systems Magazine, 25*(7), 5–28.

Pal, M., & Foody, G. M. (2010). Feature selection for classification of hyperspectral data by svm. *IEEE Transactions on Geoscience and Remote Sensing, 48*(5), 2297–2307.

Pan, B., Shi, Z., An, Z., Jiang, Z., & Ma, Y. (2017). A novel spectral-unmixing-based green algae area estimation method for goci data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10*(2), 437–449.

Pan, C., Li, J., Wang, Y., & Gao, X. (2018). Collaborative learning for hyperspectral image classification. *Neurocomputing, 275*, 2512–2524.

Passalis, N., & Tefas, A. (2018). Dimensionality reduction using similarity-induced embeddings. *IEEE Transactions on Neural Networks and Learning Systems, 29*(8), 3429–3441.

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.

Renard, N., & Bourennane, S. (2009). Dimensionality reduction based on tensor modeling for classification methods. *IEEE Transactions on Geoscience and Remote Sensing, 47*(4), 1123–1131.

Scherer, D., Müller, A., & Behnke, S. (2010). Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks* (pp. 92–101). Springer.

Sellami, A., & Farah, I. R. (2016). High-level hyperspectral image classification based on spectro-spatial dimensionality reduction. *Spatial Statistics, 16*, 103–117.

Sellami, A., Farah, M., Farah, I. R., & Solaiman, B. (2018). Hyperspectral imagery semantic interpretation based on adaptive constrained band selection and knowledge extraction techniques. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11*(4), 1337–1347.

Shi, C., & Pun, C.-M. (2018). Superpixel-based 3d deep neural networks for hyperspectral image classification. *Pattern Recognition, 74*, 600–616.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR*. San Diego, CA, USA. arXiv: 1409.1556

Sosna, M., Trevinyo-Rodríguez, R. N., & Velamuri, S. R. (2010). Business model innovation through trial-and-error learning: The naturhouse case. *Long range planning, 43*(2-3), 383–407.

Sotoca, J. M., & Pla, F. (2010). Supervised feature selection by clustering using conditional mutual information-based distances. *Pattern Recognition, 43*(6), 2068–2081.

Su, H., Yang, H., Du, Q., & Sheng, Y. (2011). Semisupervised band clustering for dimensionality reduction of hyperspectral imagery. *IEEE Geoscience and Remote Sensing Letters, 8*(6), 1135–1139.

Verrelst, J., Rivera, J. P., Gitelson, A., Delegido, J., Moreno, J., & Camps-Valls, G. (2016). Spectral band selection for vegetation properties retrieval using gaussian processes regression. *International Journal of Applied Earth Observation and Geoinformation, 52*, 554–567.

Wang, J., Zhang, K., Wang, P., Madani, K., & Sabourin, C. (2017). Unsupervised band selection using block-diagonal sparsity for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters, 14*(11), 2062–2066.

Wen, J., Fowler, J. E., He, M., Zhao, Y.-Q., Deng, C., & Menon, V. (2016). Orthogonal nonnegative matrix factorization combining multiple features for spectral–spatial dimensionality reduction of hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing, 54*(7), 4272–4286.

Wu, B., Zhu, Y., Huang, X., & Li, J. (2016). Generalization of spectral fidelity with flexible measures for the sparse representation classification of hyperspectral images. *International Journal of Applied Earth Observation and Geoinformation, 52*, 275–283.

Xia, J., Dalla Mura, M., Chanussot, J., Du, P., & He, X. (2015). Random subspace ensembles for hyperspectral image classification with extended morphological attribute profiles. *IEEE Transactions on Geoscience and Remote Sensing, 53*(9), 4768–4786.

Xu, X., Li, J., Wu, C., & Plaza, A. (2018). Regional clustering-based spatial preprocessing for hyperspectral unmixing. *Remote Sensing of Environment, 204*, 333–346.

Xu, X., & Shi, Z. (2017). Multi-objective based spectral unmixing for hyperspectral images. *ISPRS Journal of Photogrammetry and Remote Sensing, 124*, 54–69.

Yan, C., Bai, X., Ren, P., Bai, L., Tang, W., & Zhou, J. (2016). Band weighting via maximizing interclass distance for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters, 13*(7), 922–925.

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., & Lin, S. (2007). Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(1).

Yousefi, B., Castanedo, C. I., Bédard, É., Beaudoin, G., & Maldague, X. P. (2018a). Mineral identification in lwir hyperspectral imagery applying sparse-based clustering. *Quantitative InfraRed Thermography Journal*, 1–16.

Yousefi, B., Sojasi, S., Castanedo, C. I., Maldague, X. P., Beaudoin, G., & Chamberland, M. (2018b). Comparison assessment of low rank sparse-pca based-clustering/classification for automatic mineral identification in long wave infrared hyperspectral imagery. *Infrared Physics & Technology, 93*, 103–111.

Zhao, W., & Du, S. (2016). Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. *IEEE Transactions on Geoscience and Remote Sensing, 54*(8), 4544–4554.

Zhong, P., Gong, Z., Li, S., & Schönlieb, C.-B. (2017). Learning to diversify deep belief networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 55*(6), 3516–3530.

Zhong, Y., Lin, X., & Zhang, L. (2014). A support vector conditional random fields classifier with a mahalanobis distance boundary constraint for high spatial resolution remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7*(4), 1314–1330.

Zhou, X., Li, S., Tang, F., Qin, K., Hu, S., & Liu, S. (2017). Deep learning with grouped features for spatial spectral classification of hyperspectral images. *IEEE Geoscience and Remote Sensing Letters, 14*(1), 97–101.

Zhou, Y., Peng, J., & Chen, C. P. (2015). Dimension reduction using spatial and spectral regularized local discriminant embedding for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 53*(2), 1082–1095.