Deep learning for remote sensing image classification

Li, Ying; Zhang, Haokui; Xue, Xizhe; Jiang, Yenan; Shen, Qiang

Published in:
WIREs Data Mining and Knowledge Discovery

DOI:
10.1002/widm.1264

Publication date:
2018

Citation for published version (APA):
Li, Y., Zhang, H., Xue, X., Jiang, Y., & Shen, Q. (2018). Deep learning for remote sensing image classification: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(6), [e1264]. https://doi.org/10.1002/widm.1264

Received: 20 December 2017 | Revised: 10 April 2018 | Accepted: 12 April 2018

DOI: 10.1002/widm.1264

WILEY WIREs DATA MINING AND KNOWLEDGE DISCOVERY

ADVANCED REVIEW

Deep learning for remote sensing image classification: A survey

Ying Li[1] | Haokui Zhang[1] | Xizhe Xue[1] | Yenan Jiang[1] | Qiang Shen[2]

[1]School of Computer Science, Northwestern Polytechnical University, Shaanxi, Xi'an, China

[2]Department of Computer Science, Institute of Mathematics, Physics and Computer Science, Aberystwyth University, Aberystwyth, UK

Correspondence
Qiang Shen, Department of Computer Science, Institute of Mathematics, Physics and Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, UK.
Email: qqs@aber.ac.uk

Funding information
National Key Research and Development Program of China, Grant/Award Number: 2016YFB0502502; Foundation Project for Advanced Research Field, Grant/Award Number: 614023804016HK03002; Shaanxi International Scientific and Technological Cooperation Project, Grant/Award Number: 2017KW-006

Remote sensing (RS) image classification plays an important role in the earth observation technology using RS data, having been widely exploited in both military and civil fields. However, due to the characteristics of RS data such as high dimensionality and relatively small amounts of labeled samples available, performing RS image classification faces great scientific and practical challenges. In recent years, as new deep learning (DL) techniques emerge, approaches to RS image classification with DL have achieved significant breakthroughs, offering novel opportunities for the research and development of RS image classification. In this paper, a brief overview of typical DL models is presented first. This is followed by a systematic review of pixel-wise and scene-wise RS image classification approaches that are based on the use of DL. A comparative analysis regarding the performances of typical DL-based RS methods is also provided. Finally, the challenges and potential directions for further research are discussed.

This article is categorized under:
Application Areas > Science and Technology
Technologies > Classification

KEYWORDS
convolutional neural network, deep belief network, deep learning, pixel-wise classification, remote sensing image, scene classification, stacked auto-encoder

1 | INTRODUCTION

Recently, deep learning (DL) has become the fastest-growing trend in big data analysis and has been widely and successfully applied to various fields, such as natural language processing (Ronan Collobert & Weston, 2008), image classification (Krizhevsky, Sutskever, & Hinton, 2012), speech enhancement (Xu, Du, Dai, & Lee, 2015), because of its outstanding performance compared with that of traditional learning algorithms. Such work is inspired by biology stating that for primate visual systems, the brain is organized in deep architecture and the perception is also represented at multiple levels of abstraction. DL architectures are characterized as artificial neural networks, involving usually more than two layers. As with their shallow counterpart, deep neural networks exploit feature representations learned exclusively from data. However, they do not require hand-crafted features that are mostly designed on the basis of domain-specific knowledge. This avoids the problem that hand-crafted features are highly dependent on domain knowledge. Besides, it is impractical to address the need of considering all of the details embedded in all forms of real data via the use of predesigned hand-crafted features. Instead of relying on shallow manually engineered features, DL techniques are able to automatically learn informative representations of raw input data with multiple levels of abstraction. Such learned features have achieved success by being used in many machine vision tasks. Representing an important and initial breakthrough in DL, deep belief networks (DBNs) (Hinton, Osindero, & Teh, 2006) were proposed through exploitation of restricted Boltzmann machines (RBMs) (Freund & Haussler, 1991). This was followed by the development focused on work that is based on Auto-encoder (Rumelhart & Mcclelland, 1988; Vincent, Larochelle,

Bengio, & Manzagol, 2008), which train the multiple intermediate levels of representation locally at each level. Recently, another DL architecture, of convolutional neural networks (CNNs) (Lecun, Bottou, Bengio, & Haffner, 1998), has achieved significant results in computer vision, attributing to the deep structure that facilitates the model to capture and generalize filtering mechanisms by performing convolutions in the image domain, leading to highly abstract and effective features.

Despite its great potential, in general, the use of DL in RS image classification brings forward significant new challenges. There are several reasons for this: First, many RS data, especially hyperspectral images (HSIs), contain hundreds of bands that can cause a small patch to involve a really large amount of data, which would demand a large number of neurons in a DL network (Berlin & Kay, 1969; Chen, Xiang, Liu, & Pan, 2013; Zhang et al., 2018). Apart from the visual geometrical patterns within each band, the spectral curve vectors across bands may also provide important information. However, how to utilize this information still requires further research. Second, the usually impressive performance of DL techniques relies on large numbers of labeled samples. Unfortunately, very few labeled samples are available in RS data. Third, compared with conventional natural scene images, RS images are more complex. The high spatial resolution RS images may involve various types of objects, which are also different in size, color, location and rotation. HSIs may be acquired using different sensors in the first place. The complexity of RS data makes it very difficult if not impossible to directly construct a DL network model for the classification of such images, assistance is required for DL to perform.

The aforementioned reasons make the application of DL in RS image classification rather specific, but challenging. Having recognized this, there have been a good number of approaches recently proposed to deal with such challenges. This paper presents a survey of such developments, focussing on two important aspects: one being on pixel-wise classification for HSIs and the other being scene classification for high-resolution aerial or satellite images. The former is concerned with identifying what category each pixel in a given RS image belongs to, and the later aims to automatically assign a semantic label to each RS scene image.

This survey is organized as follows. The second section outlines typical DL models which are used in RS image classification, including CNNs, stacked auto-encoders (SAEs), and DBNs. The third section reviews the pixel-wise and scene-wise RS image classification approaches that are based on DL. The classification performances of typical DL-based methods for RS images are also compared in this section. The fourth section summarizes the present work and discusses challenges ahead, pointing out potential directions for further research in RS image classification using DL techniques.

2 | TYPICAL DEEP NETWORK MODELS

In this section, we briefly review the following three typical deep neural network models that have been used for RS image classification. More details about DL architectures in machine learning can be found in (Bengio, 2009; Bengio, Courville, & Vincent, 2013).

2.1 | Convolutional neural networks

The leading model in DL is that of CNNs, which is adopted in a wide range of aspects in image processing, including image classification (He, Zhang, Ren, & Sun, 2014), object detection (Girshick, 2015; Girshick, Donahue, Darrell, & Malik, 2013), super-resolution restoration (Dong, Chen, He, & Tang, 2016), etc. Generally, a CNN mainly consists of three key parts: convolution layers, pooling layers, and fully connected layers. Different parts play different roles. An example of CNNs is shown in Figure 1.

In convolution layers, the input maps are convolved with learnable kernels and are subsequently put through the activation function to form the output feature maps. This process can be formulated as:

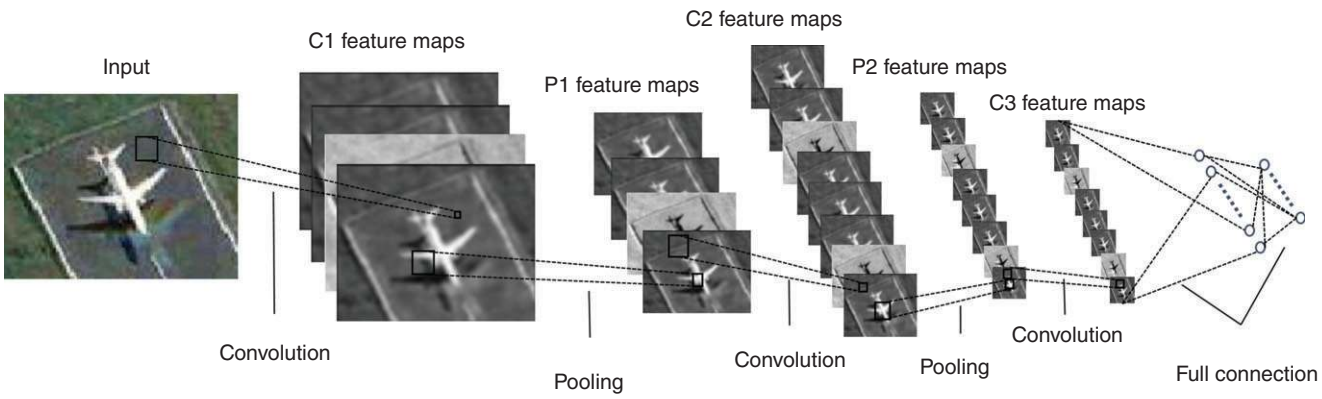

FIGURE 1 An example of convolutional neural networks

$$map_{l,j}^{x,y} = f\left(\sum_{m}\sum_{h=0}^{H_l-1}\sum_{w=0}^{W_l-1} k_{l,j.m}^{h,w} map_{(l-1),m}^{(x+h),(y+w)} + b_{l,j}\right) \quad (1)$$

where $k_{l,j.m}^{h,w}$ is the value at the position (h, w) of the kernel connected to the mth feature map in the $(l-1)$th layer, H_i and W_i are the height and width of the kernel, respectively, and $b_{l,j}$ is the bias of the jth feature map in the lth layer. Such convolution layers introduce weight sharing mechanism within same feature maps, which helps reduce significantly the number of parameters otherwise required. It can take two-dimensional (2D) images with any scale directly as input while reserving the location information of objects in the images. Due to the recognition of the inherent advantages of convolution operation, a significant amount of work has been focused on improving the ability of convolution layers in the literature. For instance Lin, Chen, and Yan (2013) proposed a network in a network, substituting the conventional convolution layer with a multilayer perceptron consisting of multiple fully connected layers. Long, Shelhamer, and Darrell (2017) replaced the fully connected layers in a CNN with a deconvolution layer to build a novel convolutional network.

Generally, a pooling layer follows a convolutional layer and it is used to reduce the dimensionality of feature maps. There are two types of basic pooling operation which are the most commonly used: average pooling and max pooling, as shown in Figure 2. Detailed theoretical analysis of these is beyond the scope of this paper, but can be found in Scherer, Muller, and Behnke (2010). As the computation process of pooling operation takes neighboring pixels into account, a pooling layer is translation invariant. Apart from average and max pooling, there are several other pooling operations, including spatial pyramid pooling (He et al., 2014), stochastic pooling (Zeiler & Fergus, 2013) and def-pooling (Ouyang et al., 2014).

A fully connected layer is basically the same as one within a traditional neural network (such as Back Propagation network). The output maps of the last convolution layer or pooling layer are arranged into vectors, acting as the inputs to the first fully connected layer. The output of the final fully connected layer can be regarded as the learnt feature, forming the result of which is extracted from the input image by the convolutional network. The classification operation can be simply implemented by connecting this output to a learning classifier, such as Softmax (Krizhevsky et al., 2012).

Compared to shallow learning, the advantage of DL is that it introduces deep network architectures to learn more abstract and effective features. However, the large amount of parameters introduced in so doing may lead to overfitting. Numerous regularization methods have emerged in defense of potential overfitting, such as dropout (Krizhevsky et al., 2012) and batch normalization (Ioffe & Szegedy, 2015). The former randomly omits part of the feature detectors during each training case, and the later normalizes a certain part of the model architecture for each training mini-batch.

The learning and working process of CNN can be summarized into two stages: (a) networking training and (b) feature extraction and classification. There are two parts for the first stage: a forward part and a backward part. In the forward part, the input images are fed through the network to obtain an abstract representation, which will be used to compute the loss cost with regard to the given ground truth labels. Based on the loss cost, the backward part computes the gradients of each parameter of the network. Then all the parameters are updated in response to the gradients in preparation for the next forward computation cycle. After sufficient iterations of training, in the second stage, the trained network can be used to extract deep features and classify unknown images.

2.2 | Stacked autoencoders

A stacked autoencoder (SAE) is a deep network model consisting of multiple layers of autoencoders (AEs) in which the output of one layer is wired to the input of the successive layer as shown in Figure 3. An AE has one visible layer of d inputs and one hidden layer of h units with an activation function f. During training, it first maps the input $x \in R^d$ to the hidden layer and get the latent representation $y \in R^h$. Then y is mapped to an output layer that has the same size with input layer, which is called reconstruction. The reconstruction is denoted as $z \in R^d$. Mathematically, these procedures can be shown as:

$$\begin{aligned} y &= f(W_y x + b_y) \\ z &= f(W_z y + b_z) \end{aligned} \quad (2)$$

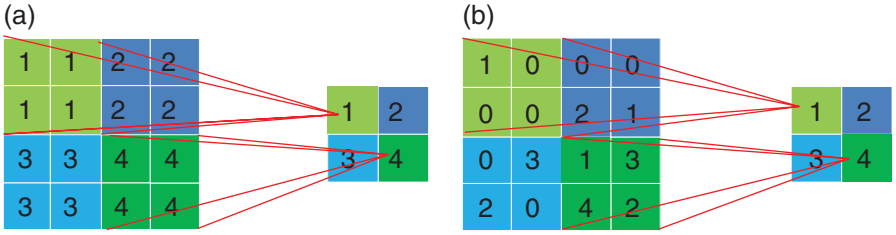

FIGURE 2 Two basic pooling operations. (a) Average pooling. (b) Max pooling

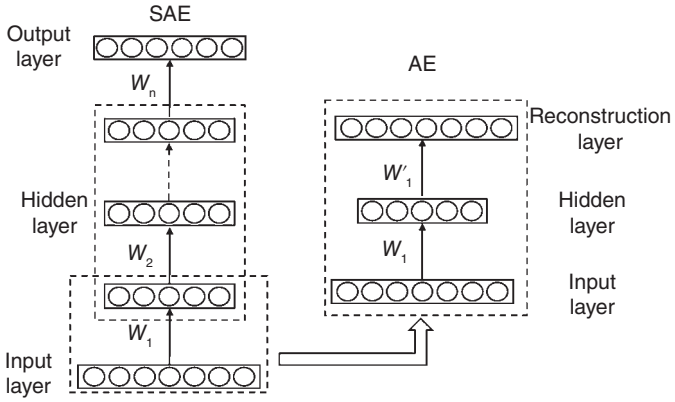

FIGURE 3 Stacked auto-encode and auto-encoder

where W_y,W_z denotes input-to-hidden and hidden-to-output weights, respectively, b_y and b_z denote the bias of hidden and output units, respectively, and $f(\cdot)$ denotes the activation function, which apply element-wise to its arguments. The loss function or energy function $J(\theta)$ measures the reconstruction z when given input x,

$$J(\theta) = \frac{1}{2M}\sum_{m=1}^{M}\left\|z^{(m)} - x^{(m)}\right\|_2^2 \tag{3}$$

where M denotes the number of training samples. The objective is finding the parameters $\theta = (W, b_y, b_z)$ which can minimize the difference between the output and the input over the whole training set $X = [x^{(1)}, x^{(2)}, \ldots, x^{(m)}, \ldots, x^{(M)}]$, and this can be efficiently implemented via the stochastic gradient descent algorithm (Johnson & Zhang, 2013).

There are two well-known variants of the AE, that is, denoising AE (Vincent et al., 2008) and sparse AE (Schlkopf, Platt, & Hofmann, 2006a). The former can recover the correct input from a corrupted version, thus forcing the model to capture the structure of the input distribution. The latter aims to extract sparse features from raw data where the objective is to minimize the reconstruction error with a sparsity constraint.

As indicated above, SAE consists of multiple layers of AEs, each of which is a special type of neural network used for efficient encodings. Instead of training the network to predict a certain target label given inputs, an AE is trained to reconstruct its own inputs. A single AE is not able to get the discriminative and representative features of raw input data. Multiple AEs are usually stacked with one other to form an SAE, which forwards the code learned from the previous AE to the next in order to accomplish a given task.

2.3 | Deep belief networks

An DBN model is constructed with a hierarchically arranged series of RBMs as shown in Figure 4. An RBM at l-layer in DBN is an energy-based generative model that consists of a layer with I binary visible units $v^l = \{v_1^l, v_2^l, \cdots, v_I^l\}$ and a layer with J binary hidden units $h^l = \{h_1^l, h_2^l, \cdots, h_J^l\}$. The energy of the joint configuration of the visible and hidden units (v^l, h^l) is

$$E(v^l, h^l|\theta^l) = -\sum_{i=1}^{I} a_i^l v_i^l - \sum_{j=1}^{J} b_j^l h_j^l - \sum_{i=1}^{I}\sum_{j=1}^{J} w_{ij}^l h_j^l v_i^l \tag{4}$$

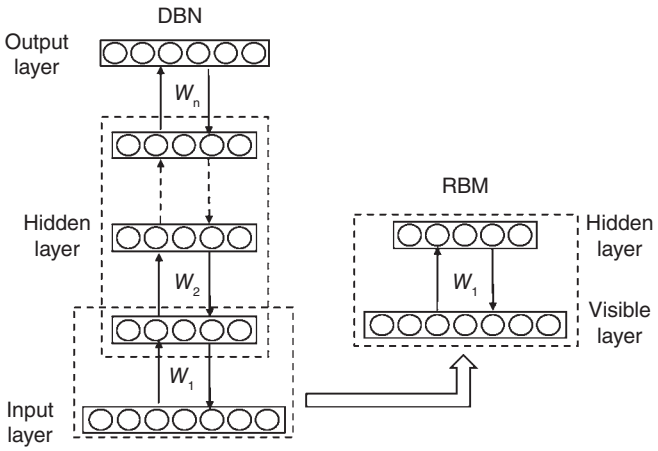

FIGURE 4 Deep belief network and restricted Boltzmann machine

where $\theta^l = \left\{ w_{ij}^l, a_i^l, b_j^l, i = 1,2,\cdots,I; j = 1,2\cdots,J \right\}$ forms the set of model parameters. An RBM defines a joint probability over the hidden units as

$$p\left(v^l, h^l | \theta^l\right) = \frac{\exp\left(-E\left(v^l, h^l | \theta^l\right)\right)}{Z\left(\theta^l\right)} \tag{5}$$

where Z is the so-called partition function,

$$Z\left(\theta^l\right) = \sum_{v^l} \sum_{h^l} \exp\left(-E\left(v^l, h^l | \theta^l\right)\right) \tag{6}$$

Then, the conditional distributions $p\left(h_j^l = 1 | v^l\right)$ and $p\left(v_i^l = 1 | h^l\right)$ can be readily computed. Figure 4 shows a typical DBN for deep feature learning from hyperspectral images. In DBN, the output of the preceding RBM is used as input data for the next RBM. Two adjacent layers have a full set of connections between them, but no two units in the same layer are connected. The input vector $\left(v_1^0, v_2^0, \cdots v_I^0\right)^T$ can be set to the spectral signature of each pixel or the contextual features from neighboring pixels. Every layer outputs a feature of its input data, and the further away from the network input a layer is, the more abstract the feature that is produced by it is.

DBN is a probabilistic generative model which provides a joint probability distribution over observable data and labels. A DBN first takes advantages of an efficient layer-by-layer greedy learning strategy to initialize the deep network, and then fine-tunes all of the weights jointly with the desired outputs.

3 | DL FOR REMOTE SENSING IMAGE CLASSIFICATION

Within the past decade, DL has emerged as of the most successful machine learning techniques and has achieved impressive performance in the field of computer vision and image processing, with applications such as image classification (He et al., 2014; Krizhevsky et al., 2012), object detection (Girshick, 2015; Girshick et al., 2013), and super-resolution restoration (Dong et al., 2016). DL is also taking off in remote sensing image classification most recently, and a growing number of relative papers are reported in the literature year by year. As a focus of this survey, in this section, we focus on pixel-wise and scene-wise remote sensing image classification approaches that are based on DL, supported with comparative experimental analyses.

3.1 | Pixel-wise classification of HSIs

Hyperspectral remote sensors capture digital images in hundreds of continuous narrow spectral bands, producing three-dimensional (3D) hyperspectral imagery (HSI) which simultaneously involves spectral and spatial information. Also, with high-quality hyperspectral satellite data becoming available (e.g., via the launch of EnMAP, scheduled in 2020, and DESIS, originally planned for 2017, Zhu et al., 2017) more HSI data will become available. Such very rich spectral information is potentially very useful to help reveal any interesting unknown content contained within the images. In particular, HSI has been widely used in a range of practical applications, such as land cover, change detection and object identification. Indeed, applications of HSI data have been one of the most active research directions, with classification of individual pixels in an HSI image playing a crucial role in such applications.

Most recently, following the success of DL in natural image processing, DL methods have been introduced into HSI classification, achieving impressive results (Chen, Jiang, Li, Jia, & Ghamisi, 2016; Chen, Lin, Zhao, Wang, & Gu, 2017; Chen, Zhao, & Jia, 2015; Li, Zhang, & Shen, 2017). These approaches can be organized into three main categories, which will take advantage of spectral information, spatial information and spectral-spatial information, respectively. A review of each of these categories is given below, after the presentation of data sets and performance indicators used to conduct the comparative review.

3.2 | Data sets and performance indicators

Benchmark data sets of HSI are typically utilized for evaluating the performance of the relevant methods. In Table 1, we list several popular data sets for HSI classification. AVIRIS is a airborne visible/infrared imaging spectrometer which belongs to the Jet Propulsion Laboratory in the USA. ROSIS is a reflective optics system imaging spectrometer from the National Aeronautics and Space Agency of Germany. Taking the data set of Indian Pines for instance, it was acquired by the AVIRIS sensor in Indiana in June 1992. This data set covered 145 lines by 145 pixels, the geometric resolution of which is 20 m. The original

TABLE 1 Common data sets of HSI classification used

Data sets	Indian Pines	Salinas	Kennedy Space Center	Pavia Center	Pavia University	Botswana
Acquisition time	1992	1992	1996	2001	2001	2001
Location	Indiana	California	Florida	Northern Italy	Northern Italy	Okavango delta
Device	AVIRIS	AVIRIS	AVIRIS	ROSIS	ROSIS	Hyperion
Spectrum coverage (nm)	400–2,500	400–2,500	400–2,500	430–860	430–860	400–2,500
Data size (pixels)	145 × 145	512 × 217	512 × 614	1096 × 492	610 × 340	1476 × 256
Spectrum number(corrected)	224	224	224	115	115	242
Sample size	10,249	54,129	5,211	7,456	42,776	3,248
Category	16	16	13	9	9	14

data have 224 spectral in the wavelength range of 400–2,500 nm. The 24 bands covering the region of water absorption were removed to noise. The Indian Pines ground truth contains 16 classes with 10,249 pixels labeled in total.

As with the common literature, three performance indicators, overall accuracy (OA), average accuracy (AA), and kappa coefficient (K), are employed to evaluate the classification performance on benchmark data sets. OA equals to the number of properly classified samples divided by that of overall samples. AA is the averaged value of accuracies across all category. K is a relatively more comprehensive indicator and its formula is shown below:

$$K = \frac{N\sum_{i=1}^{n} C_{ii} - \sum_{i=1}^{n} C_{i+}C_{+i}}{N^2 - \sum_{i=1}^{n} C_{i+}C_{+i}} \tag{7}$$

where N is the number of overall samples, n is the number of categories and C_{ij} is the (i, j)th value of the confusion matrix C (Thompson & Walter, 1988), with C_{i+} and C_{+i}, respectively, denoting the sum of the ith row and that of the ith column of C.

3.3 | Spectral feature classification

An HSI image contains spectral and spatial information simultaneously. Compared with spatial resolution, spectral resolution is relatively higher. One spectral vector can be extracted from each pixel and used to identify the information content embedded in this spatial pixel. Traditional HSI classification approaches only make use of spectral information. Typical classifiers include those implemented on the basis of k-nearest-neighbors (Samaniego, Bardossy, & Schulz, 2008), distance measure (Du & Chang, 2001), logistic regression (Li, Bioucas-Dias, & Plaza, 2010), and maximum likelihood criterion (Ediriwickrema & Khorram, 1997). Classifying each pixel via the spectral vector directly is not reasonable and inefficient. With respect to the reported conventional HSI classification approaches, feature extraction methods are also adopted to support the classification. Traditional feature extraction operations adopt handcrafted features and "shallow" structures. However, the design of handcrafted features can be tedious and is typically suboptimal. In addition, HSI contains a variety of data and it is challenging to design robust, handcrafted features for HSI data (Zhang, Zhang, & Du, 2016).

DL models have ability to automatically extract deep features from raw input data and such deep features are high-level and abstract features, which are generally more robust and efficient than handcrafted features. In DL-based spectral feature classification methods, DL models take raw spectral vectors as inputs and output deep spectral features, which are then used for classification. Figure 5 illustrates a general framework of DL for spectral feature classification.

Conventional CNNs take a 2D image as input, such as AlexNet (Krizhevsky et al., 2012), VggNet (Simonyan & Zisserman, 2014), and GoogleNet (Szegedy et al., 2014). Directly applying 2D-CNNs to extract deep spectral feature is not feasible.

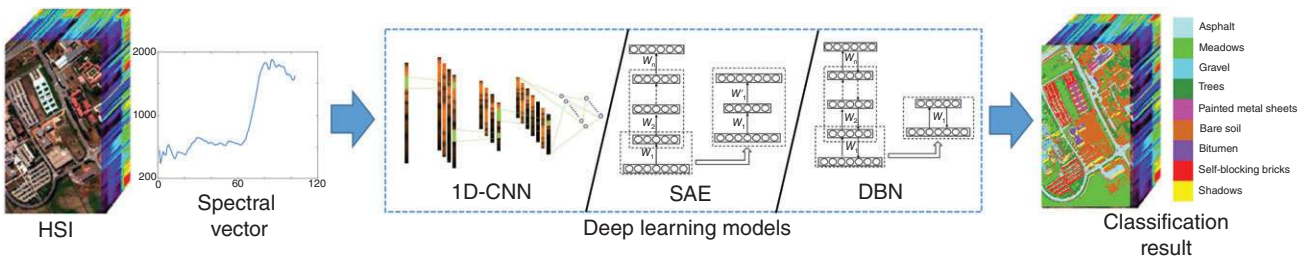

FIGURE 5 General framework of deep learning for spectral feature classification

Hu, Huang, Wei, Zhang, and Li (2015)) attempt to carry out HSI classification using 1D-CNN that contains four layers: one convolution layer followed by one pooling layer and two fully connected layers. The study of Mei et al. (2016) exploit a similar 1D-CNN to classify HSI, the difference is that the latter saves the pooling layer but relies on the exploitation of additional techniques like dropout (Krizhevsky et al., 2012) and batch normalization (Xu et al., 2015).

Among the three typical DL models (CNN, SAE, and DBN), SAE and DBN take the data that is represented in a vector form as input, thereby fitting the need to extract deep spectral feature from spectral vectors. In fact, as compared to CNN, SAE and DBN were introduced to perform HSI spectral feature classification earlier.

An initial attempt to work along this direction can be found in (Chen et al., 2017), where authors adopt an SAE to extract deep spectral feature. By exploiting the relationship between the input layer and the reconstruction layer in an AE, a constraint is introduced such that the weights associated with the connections between the hidden-to-output layer are the transposition of those of input-to-hidden. Following this original work, the use of a DBN instead of an SAE is reported in (Chen et al., 2015). Similarly, Ma, Wang, Geng, and Wang (2016) employ a SAE to learn effective feature and add a relative distance prior in the fine-tuning process, giving more effective guidance regarding the desirable features when there are not sufficient labeled samples. Xing, Ma, and Yang (2015) use stacked denoising AE (Vincent, Larochelle, Lajoie, Bengio, & Manzagol, 2010) to extract robust spectral features and complete the classification task.

A similar idea was adopted in (He, Li, Zhang, Zhang, & Wang, 2016), where HSI is classified via a novel model named deep stacking network (DSN). A DSN model stacks many simple modules, each of which contains an input layer, a hidden layer and an output layer, with the weights of input-to-hidden initialized randomly or via contrastive divergence (Hinton, 2002), and those of hidden-to-output initialized by computing pseudo-inverse (Golub & Kahan, 1965). Zhong, Gong, and Schnlieb (2016) added diversity promoting priors by incorporating the diversity promoting conditions into the optimization of training objective in the pretraining and fine-tuning processes of DBN, which helps improve the classification efficiency in HSI.

3.4 | Spatial feature classification

As stated previously, an HSI image contains spectral and spatial information simultaneously. Spectral feature classification methods just take account of spectral information and fail to make use of spatial information. Spatial information (also termed contextual information) helps reveal useful relationships between adjacent pixels. Making full use of such information and hence, the adjacent pixel relationships can improve the classification accuracy and efficiency significantly.

The general framework of DL-based spatial feature classification is shown in Figure 6. Such an approach takes the spectral vectors within the neighborhood region of a given pixel (situated at the center of the neighborhood), and outputs deep spatial feature of that pixel. Due to the complexity of there being hundreds of bands along the spectral dimension, the data in a neighborhood region of a certain pixel always involve high dimensionality, in the order of tens of thousands. Directly classifying HSI via all of such data may be computationally prohibitive. So, in the stage of data preprocessing, data dimensionality reduction is necessary. This is reflected in Figure 6.

The most significant advantage of 2D-CNNs is that they are suitable for implementation to effectively learn features from raw 2D image. However, as HSI contains more than 100 spectral bands, directly applying a 2D-CNN to extract deep features from raw HSI require a mass of learnable kernels, which is hard to train while increasing computational overheads. Thus, Makantasis, Karantzalos, Doulamis, and Doulamis (2015) employed randomized principal components analysis to condense the spectral dimension of entire HSI to 10 or 30 dimensions (principal components, PCs) first. They then applied a 2D-CNN to extract deep features from the compressed HSI (with a small window size, 5×5 is used in the existing work), done prior to the classification task. In (Yue, Zhao, Mao, & Liu, 2015), the top three PCs are extracted from raw HSI, by conventional principal components analysis (PCA). Similar with the method last mentioned, condensed HSI is put though a 2D-CNN to extract

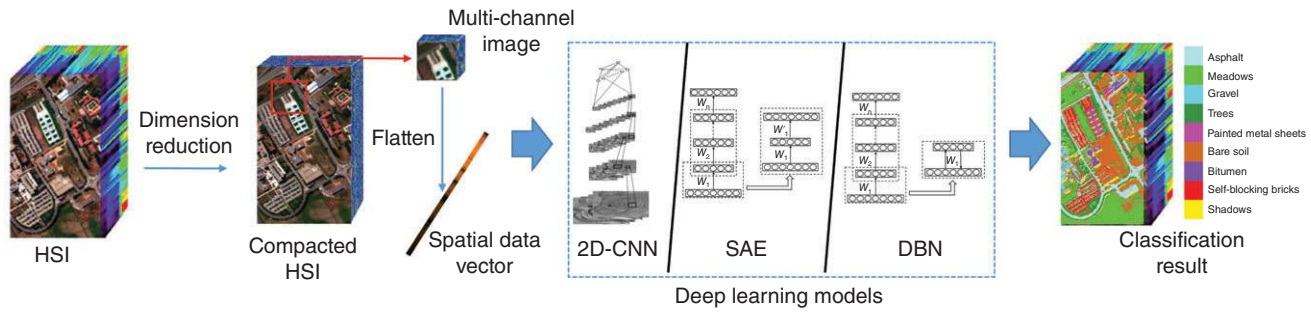

FIGURE 6 General framework of deep learning for spatial feature classification

spatial features (with a fairly large window size this time, of 42 × 42).Clearly, there is a trade-off between the number of PCs used and the window size to employ.

Both methods proposed in Makantasis et al. (2015) and Yue et al. (2015) are simple and intuitive, performing classification using 2D-CNN-based HSI spatial features. In addition to these, there are several alternative attempts. For instance, in order to capture multiscale spatial features, Zhao, Guo, Yue, Luo, and Luo (2015) used Laplacian pyramid transformation to extract multiscale data from condensed HSI, and then each scale data is fed to an independent 2D-CNN to extract deep spatial features. Aptoula, Ozdemir, and Yanikoglu (2016) preprocess raw HSI images with PCA and attribute profiles (Mura, Benediktsson, Waske, & Bruzzone, 2010) sequentially, followed by the use of a 2D-CNN that takes 42 × 42-sized HSI patch as input to accomplish the classification task. In Liang and Li (2016), the dimensionality of HSI images is reduced with PCA first, spatial features are next extracted using a 2D-CNN, the results are further processed with sparse coding and then, fed into a learning classifier for classification at last. Another novel approach has been proposed in Li, Xie, and Li (2016), where an HSI reconstruction model based on the use of a deep CNN is proposed to enhance spatial features, with the reconstructed image classified by the efficient extreme learning machine (Li, Chen, Su, & Du, 2015).

Similarly in the underlying ideas, in Chen et al. (2017) and Lin, Chen, Zhao, and Wang (2015), PCA is exploited to reduce the dimensionality of HSI images (particularly to four and three, respectively), the resultant data cubes are flattened or extracted from neighborhood regions. This is then followed by an SAE performing the actual classification. In the pretraining stage, the weights of hidden-to-output are restricted to the transposition of those of input-to-output with the cross entropy taken as the loss function to minimize. In the fine-tuning stage, softmax is taken as the activation function of the output layer of the SAE. Two similar frameworks which adopt DBN in their structure can be found in Chen et al. (2015) and Li, Zhang, and Zhang (2015) also.

3.5 | Spectral-spatial classification

Three-dimensional HSI simultaneously involves rich spectral information and spatial information. Neither spectral feature classification nor spatial feature classification framework outlined above takes full advantage of this important character of HSI. Compared with those methods described previously, DL-based spectral-spatial classification strategies are more suitable for HSI images and can lead to the improvement of classification accuracy. There are two main strategies to combine spectral and spatial information; one is to extract spectral and spatial features separately first, and then to combine spectral and spatial features; the other is to extract deep spectral-spatial features from sub-3D cubes of HSI directly. The former usually adopts 1D-CNN, 2D-CNN, SAE, and DBN, the latter typically uses 3D-CNN. The basic flow chart underlying these approaches is shown in Figure 7.

In Zhang, Li, Zhang, and Shen (2017), 1D-CNN and 2D-CNN are used to extract spectral features and spatial features, respectively, with their outputs of 1D-CNN and 2D-CNN jointly fed to softmax for classification. To address the potential problem of overfitting caused by limited training samples of HSIs, the samples are augmented with flip and rotation operations. Li et al. exploited a similar framework in Yang, Zhao, Chan, and Yi (2016). Also, Yue, Mao, and Li (2016) have proposed a joint framework where an SAE and a 2D-CNN are adopted, where spectral features are extracted by the SAE and

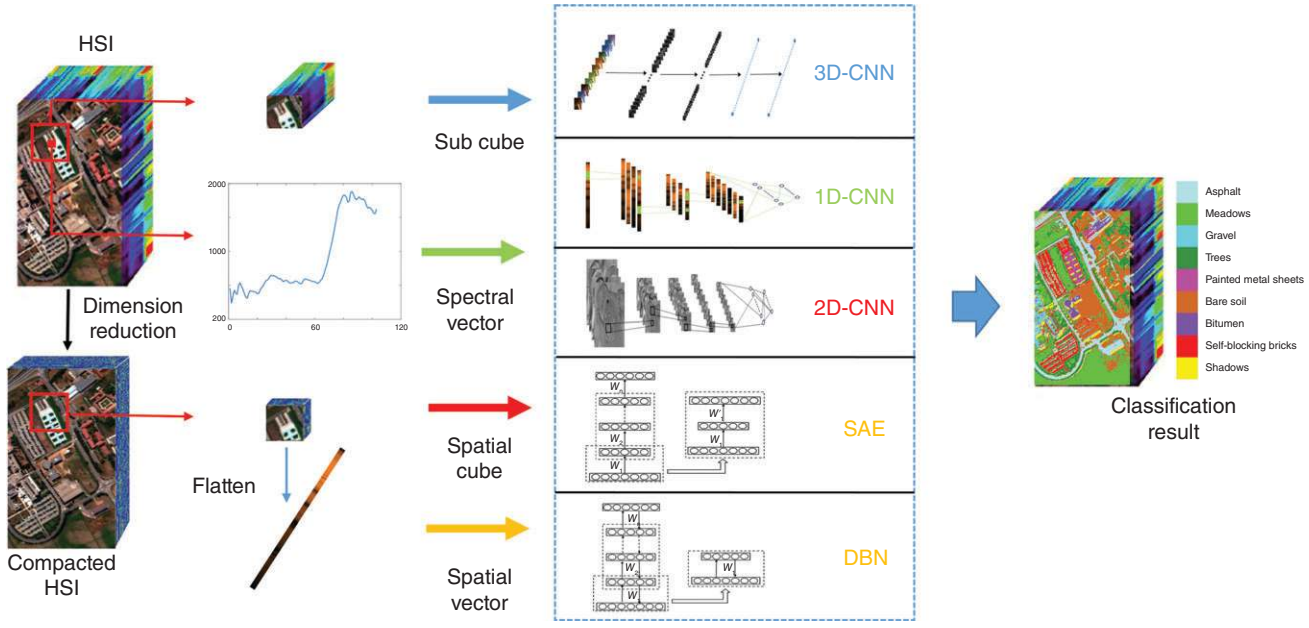

FIGURE 7 General framework of deep learning for spectral-spatial feature classification

spatial features are learned from compacted HSI via a 2D-CNN which implements spatial pyramid pooling (He et al. (2014). In Zhao and Du (2016), based on local discriminant embedding (LDE) (Chen, Chang, & Liu, 2005), a balanced LDE method was proposed and jointly used with a 2D-CNN to obtain the final classification result.

The main challenge facing 2D-CNN due to 3D HSI is the additional dimension. The CNN-based spectral-spatial classification methods in the literature deal with this challenge by either expanding 2D-CNN to 3D-CNN or rearranging 3D HSI to 2D HSI. Both in Chen et al. (2016) and Li et al. (2017), for example, a 3D-CNN is employed to learn deep spectral-spatial features. In particular, the former exploits a large-scale 3D-CNN which takes cubes of 27×27 in space size as input, while the latter uses a much more compact 3D-CNN with input cubes of 5×5 in size. In Lee and Kwon (2016)and Slavkovikj, Verstockt, Neve, Hoecke, and Walle (2015), a 2D-CNN is employed to learn spectral-spatial features. Particularly, Lee and Kwon (2016) presented an approach that convolves the 3D subcubes extracted from raw HSI images along the spatial dimension with 3×3-sized and 1×1-sized convolution kernels, and then reconstructs new 3D data using the convolved outputs jointly. The procedure employs a pointwise convolution layer at last to complete the classification. The work of Slavkovikj et al. (2015) extracts 3D cubes from raw HSI first, and then reshapes such cubes to 2D images.

Classification methods that rely on SAE- and DBN-based spectral-spatial features always extract spectral and spatial features separately and then joint them to form spectral-spatial features. Spectral information does not require any preprocessing, but spatial information has to be flattened to a 1D vector, as SAE and DBN can only handle 1D input. Following this general approach, Chen et al. applied SAE (Chen et al., 2017) and DBN (Chen et al., 2015) for spectral-spatial feature extraction and classification. Also, Ma, Wang, and Geng (2016) proposed a spatially updated deep AE for spectral-spatial feature extraction, by adding a sample similarity regularization mechanism and combining it with the collaborative representation-based classification to deal with the problem of small training sets. Tao, Pan, Li, and Zou (2015) adopted stacked sparse AE to extract high-level features from unlabeled data and then, to feed the leant features to SVM for classification. Li, Bruzzone, and Liu (2015) proposed a two-step framework, where HSI cubes are filtered by 3D Gabor wavelets first and then an SAE is trained using the outputs of the previous step via unsupervised pre-training, followed by fine-tuning over the entire network last. Ma, Geng, and Wang (2015) proposed a contextual DL algorithm which extracts spectral-spatial features through a deep SAE architecture. To extract spectral-spatial features efficiently, Han, Zhong, and Zhang (2016) in their study proposed unsupervised convolutional sparse AE (UCSAE) with window-in-window selection strategy. Again, to deal with the problem of having limited training samples, Ma, Wang, and Wang (2016) in their study proposed a novel semi-supervised classification framework based on the utilization of multidecision and deep features.

Apart from SAE, DBN, and CNN, these is one more popular DL model namely recurrent neural network (RNN), which is proposed for processing sequential data (such as speech data) and it has also been introduced into HSI classification. Compared with the amount of HSI classification methods based SAE, DBN, and CNN, that of RNN is relatively less, but the time that RNN-based HSI classification methods are proposed is more recent. The first attempt can be found in Mou, Ghamisi, and Zhu (2017), where Mou et al. used an RNN to capture the sequential property of a hyperspectral pixel vector to perform classification tasks. They also used parametric rectified tanh (PRetanh) in their network to avoid the risk of divergence during the training procedure. Wu and Prasad (2017) proposed convolutional recurrent neural network (CRNN), in which a few convolution layers are followed by recurrent layers. Middle-level and locally invariant features are extracted from raw HSI and spectrally contextual features are then extracted from the features generated by convolution layers.

3.6 | Experimental results and analysis

In this section, we firstly compare three typical DL models, that is, CNN, SAE, and DBN, for HSI classification that use spectral feature, spatial feature, and spectral-spatial features, respectively. Taking the Pavia University scene image data set for example, we compare nine DL-based HSI classification methods. For CNN- and DBN-based classification, 10% and 50% labeled samples are randomly selected as training data, respectively. For SAE-based classification, tagged samples are split into three sets for: training, validation, and testing data, with a split ratio 6:2:2. The classification results are shown in Table 2.

As can be seen in Table 2, in terms of classification accuracy, spectral-spatial feature-based classification methods obtained the best performance, followed by spatial feature based as the second best, leaving spectral feature-based classification approaches as the last. In HSI there are certain pixels belonging to different objects with the same spectral character, there are also pixels belonging to the same object but showing different spectral characters. It is difficult to classify such pixels via spectral information only. In addition, approaches taking only spectral information into account usually fail to capture the spatial distribution information, leading to poor performance. Spatial feature-based classification methods take advantage of spatial information contained within HSI. However, during the dimensionality reduction process, part spectral information may be lost. Spectral-spatial feature-based classification approaches exploit both spectral information and spatial information, making the most of HSI and therefore, achieving the best performance.

TABLE 2 Classification results of Pavia University

	Spectral feature			Spatial feature			Spectral-spatial feature		
Models evaluation system	**OA(%)**	**AA(%)**	***K*(×100)**	**OA(%)**	**AA(%)**	***K*(×100)**	**OA(%)**	**AA(%)**	***K*(×100)**
CNN (Y. Chen et al., 2016)	92.28	92.55	90.37	94.04	97.52	92.43	**99.54**	**99.77**	**99.56**
SAE (Y. Chen et al., 2017)	95.14	94.01	93.70	98.12	97.32	97.55	**98.52**	**97.82**	**98.07**
DBN (X. X. Zhu et al., 2017)	96.42	95.09	95.30	98.62	97.95	98.19	**99.05**	**98.48**	**98.75**

CNN = convolutional neural networks; SAE = stacked auto-encoders; DBN = deep belief network. The bold values are the best results.

TABLE 3 Classification results of Indian Pines

Models	**SAE**	**DBN**	**2D-CNN**	**3D-CNN**	**DC-CNN**
OA (%)	93.98	95.91	95.97	99.07	**99.92**
AA (%)	93.81	94.20	93.23	98.66	**99.57**
K(×100)	93.13	95.34	95.40	98.93	**99.91**

CNN = convolutional neural networks; SAE = stacked auto-encoders; DBN = deep belief network. The bold values are the best results.

Now, taking the Indian Pines data set for another example. By splitting the labeled samples into training data and testing data with a split ratio 1:1, we compare five spectral-spatial feature-based classification frameworks, which are based on SAE, DBN, 2D-CNN, 3D-CNN, and dual channel convolutional neural network (DC-CNN), respectively. The experimental results are listed in Table 3 and the visual classification results are shown in Figure 8.

As shown in Figure 8, CNN-based HSI spectral-spatial feature classification approaches, including 2D-CNN, 3D-CNN, and DC-CNN can achieve better performance than SAE- and DBN-based methods. It is interesting to note that historically, works on HSI classification methods based on SAEs and DBNs were developed earlier than those based on CNNs. However, statistically, it shows that recently, the number of papers regarding the use of CNNs for HSI classification grows fastest and the performance of CNNs is generally better.

3.7 | Scene classification of aerial/satellite images

Distinguished from the aforementioned pixel-level image classifications which interprets HSIs with a bottom-up manner, scene classification aims to automatically assign a semantic label to each scene image (Hu et al., 2017; Hu, Xia, Hu, Zhong, & Xu, 2016; Xia et al., 2016; Yang, Yin, & Xia, 2015; Zhao, Zhong, Xia, & Zhang, 2016). Here, a scene image usually refers to a local image patch manually extracted from large-scale high-resolution aerial or satellite images that contain explicit semantic classes (e.g., residential area, commercial area, etc.). Due to high resolutions in such data, different scene images may contain the same types of objects or share similar spatial arrangements between objects. For example, both residential area and commercial area may contain buildings, roads and trees, but they belong to two different categories of build-up areas. Indeed, great

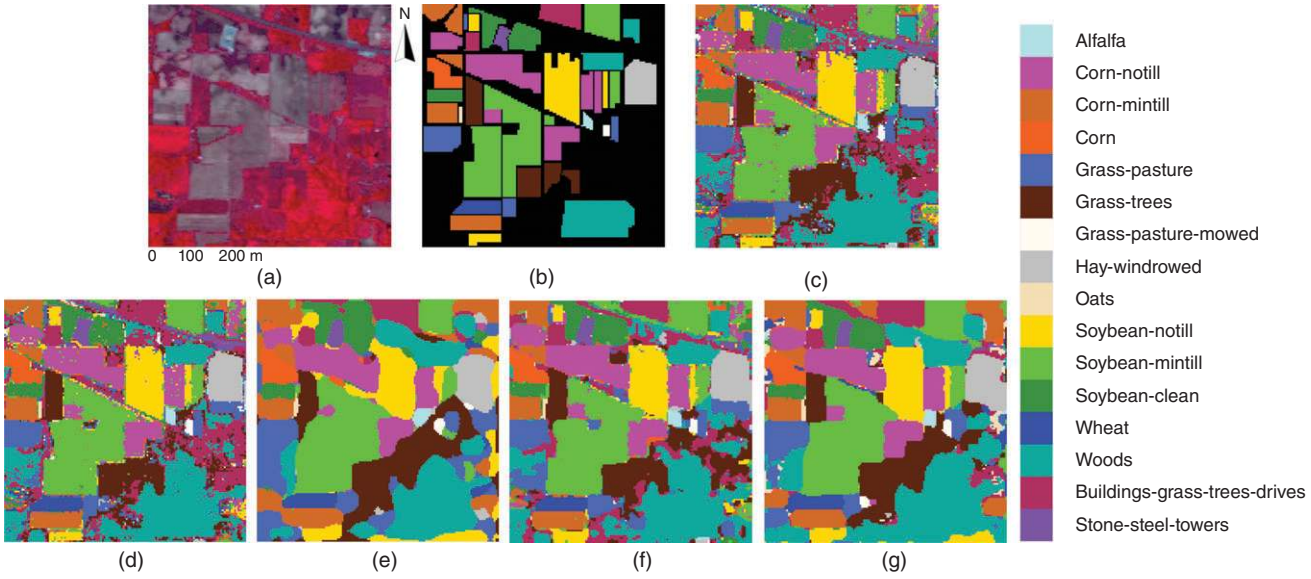

FIGURE 8 Classification of Indian pines. (a) false-color composite; (b) ground truth; (c) SAE, OA = 93.98%; (d) DBN, OA = 95.91; (e) 2D-CNN, OA = 95.97%; (f) 3D-CNN, OA = 99.07%; (g) DC-CNN, OA = 99.92%

variations potentially existing in scene images in the spatial arrangements and structural patterns make scene classification a considerably challenging task (Zhu et al., 2017).

To construct a high-powered scene classification method, the use of efficient and effective feature representations is very important. The early works for scene classification are mainly based on handcrafted features. These methods generally focus on the use of a considerable amount of domain-specific properties to design various low-level visual features or on middle-level feature representations by encoding low-level local features. The former include properties such as color (typically the color histograms, CH; Swain & Ballard, 1991), texture (typically the local binary patterns, LBP; Ojala, Pietikinen, & Menp, 2000), and structure (typically the scale invariant feature transforms, SIFT; Kim, Madden, & Warner, 2009), and the latter include representations such as bag of visual words (BoVW; Sivic & Zisserman, 2003), spatial pyramid matching (SPM; Lazebnik, Schmid, & Ponce, 2006), locality-constrained linear coding (LLC; Wang et al., 2010), and improved Fisher kernel (IFK; Perronnin & Mensink, 2010). The potential for improvement over such traditional approaches is limited by the ability of experts to design the feature extractors and the expensive encoding power. In recent years, learned high-level deep features have been reported to achieve state-of-the-art performance on aerial image classification (Hu et al., 2016, 2017; Xia et al., 2016; Yang et al., 2015; Zhao & Du, 2016).

3.8 | Data sets and performance indicators

Here, we summarize eight publicly available data sets for scene classification from aerial images as presented in Table 4. Taking the AID data set (Xia et al., 2016) for instance, it is made up of the 30 aerial scene types: airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks, and viaduct. The numbers of sample images vary a lot with respect to different aerial scene types, from 220 up to 420. In all, the AID data set contains a total of 10,000 images within the 30 classes. The pixel resolution changes from about 8 m to about half a meter, and the size of each aerial image is fixed to be 600 × 600 pixels to cover a scene with various resolutions. The images in AID are actually acquired from multiple sources, as Google Earth images are obtained from different remote imaging sensors. Such multisource problems bring more challenges for scene classification than single-source images. Moreover, all the sample images per class in AID are chosen from different countries and regions around the world, mainly in China, the United States, England, France, Italy, Japan, and Germany, and they are extracted at different times and seasons under different imaging conditions. All these factors increase the intraclass diversities of the data.

To compare the performances of different approaches, two measures are commonly adopted as performance indicators in scene classification. One is the OA of the image being classified, which is used here. The other indicator is confusion matrix (Thompson & Walter, 1988), each column and row of which represent the number or percentage of those instances in a predicted class or an actual class. Such a matrix reflects the statisticed properties of the classification performance in terms of the true positive, true negative, fake positive, and fake negative.

3.9 | Supervised deep feature extraction

The most typical supervised DL methods for image feature extraction are those using CNNs. The DL models implemented with CNNs can hierarchically extract more abstract sematic features from the input images or patches, as compared to traditional shallow models. The existing CNN-based supervised feature extraction methods can be summarized into the following three categories (Nogueira, Penatti, & Santos, 2016).

TABLE 4 Publicly available data sets for aerial scene classification

Data sets	Source	Images per class	Scene classes	Total images	Spatial resolution (m)	Image sizes	Year
AID (Xia et al., 2016)	Google Earth	220,420	30	10,000	8	600 × 600	2017
ORNL-I (Cheriyadat, 2013)	USDA&NAIP,etc	About 17	5	850	1	99.57	2016
WHU-RS19 (Sheng, Yang, Xu, & Sun, 2012)	Google Earth	50	19	1,005	Up to 0.5	600 × 600	2012
Brazilian coffee scenes (Penatti, Nogueira, & Santos, 2015)	SPOT satellite	1,438	2	2,876	—	64 × 64	2015
RSC11 (L. Zhao, Tang, & Huo, 2016)	Google Earth	About 100	11	1,232	0.2	512 × 512	2016
RSSCN7 (Zou, Ni, Zhang, & Wang, 2015)	Google Earth	400	7	2,800	—	400 × 400	2015
SIRI-WHU (B. Zhang, Zhang, & Du, 2016)	Google Earth	200	12	2,400	2	200 × 200	2016
UC Merced land-use (Xia et al., 2010)	USGS	100	21	2,100	0.3	256 × 256	2010

3.9.1 | Category 1: Directly using pretrained networks

A pretrained DL network can be used as a feature extractor for any type of image, since the features learned by the network are less dependent on the final application and could be used in a myriad of tasks (Nogueira et al., 2016). Pretrained CNN models on the natural image data set such as ImageNet (Deng, Dong, Socher, & Li, 2009) have led to remarkable results on scene classification of RS images (Chen et al., 2017; Firat et al., 2014; Zhang, Du, & Zhang, 2014). Deep features can be directly extracted from the intermediate layers of a freely available CNN architecture, such as AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan & Zisserman, 2014), and GoogLeNet (Szegedy et al., 2014), to construct global feature representations for a given application. For example, work in Castelluccio, Poggi, Sansone, and Verdoliva (2015), Penatti et al. (2015), and Xia et al. (2016) directly makes use of features extracted from the fully connected layers as the input of the classifier. An alternative strategy is to use multiscale input for multiview DL to improve the performance of aerial scene classification as reported in Luus, Salmon, Bergh, and Maharaj (2015). Furthermore, we can use a pretrained CNN as the local feature extractor and combine it with feature coding to generate the final image representation. An example of this is what is done in (Hu, Xia, Hu, and Zhang (2015), where multiscale dense CNN activations from the final convolutional layer are used as local feature descriptors and are subsequently further coded using feature encoding methods (e.g., BoVW; Sivic & Zisserman, 2003 and IFK Perronnin & Mensink, 2010). For all the deep CNN architectures mentioned in the previous discussions, either global or local features are obtained from the networks pretrained on the natural image data sets and are then, directly used for aerial image classification.

3.9.2 | Category 2: Involving fine-turning pretrained networks

When the new data set is reasonably large, but not sufficient to fully train a new network, fine-tuning is a positive option to reach the maximum effectiveness possible from pretrained deep CNNs. This is because it can significantly improve the performance of the final classifier as empirically proven (Nogueira et al., 2016). We can pretrain a network on the natural image data set first, and then fine-tune the network using a smaller group of labeled aerial images. Specifically, fine-tuning realizes adjustment in the parameters of a pretrained network by resuming the training of the network from a current setting of parameters while considering a new data set. It is not difficult to understand that such measures can help to exploit the intrinsic characteristic of RS images. However, unlike the natural image data set that consists of more than 10 millions of samples, the scales of publically available aerial image data sets (namely, UC-Merced data set [Xia et al., 2010], RSSCN7 data set [Zou et al., 2015], and WHU-RS19 data set [Sheng et al., 2012]) are comparatively fairly small. We cannot fine-tune the entire CNNs to make them more adaptive to such RS images. Nogueira et al. (2016) proposed a method that takes the UC-Merced data set (Xia et al., 2010) to fine-tune certain high-level layers of the GoogLeNet (Szegedy et al., 2014), achieving an impressive result. There have been different approaches proposed in the literature, where smaller networks are used to better fit the RS images (Luus et al., 2015; Volpi & Tuia, 2016; Zhang, Du, & Zhang, 2016; Zou et al., 2015), which may also help improve the performance regarding the overfitting and local minimum problems.

3.9.3 | Category 3: Using full-trained networks

Training a CNN model from scratch with a random initialization of the network parameters is useful when the data set is sufficiently large to ensure network converge. It has several advantages over the other two outlined above, such as generating more accurate features and gaining fully control of the network (Nogueira et al., 2016). However, large-scale networks usually contain millions of parameters to be learned from the training data. Thus, fully training for them usually needs millions of training images. Yet, the existing aerial data sets only contain hundreds or thousands of images, it will therefore, get easily trapped in local optimum and become overfitted. Indeed, as shown in Nogueira et al. (2016), using the existing aerial scene data sets (e.g., UC-Merced dataset and WHURS19 data set) to fully train popular CNN architectures, for example, AlexNet (Hu, Xia, et al., 2015), CaffeNet (Jia et al., 2014) or GoogLeNet (Szegedy et al., 2014), may well lead to a drop in accuracies as compared to pretraining networks as global feature extractors or fine-tuning pretrained networks. Thus, to better fit problems involving aerial data sets, simpler networks (Zhang, Du, & Zhang, 2016) are trained to perform classification. As an example, in (Zhang, Du, & Zhang, 2016), a boosting random CNN framework is proposed for scene classification with only two convolutional layers. However, the generalization ability of such a small network is often lower than that of large-scale networks (Nogueira et al., 2016). In theory, when practically feasible, it is highly recommended to train a large-scale network with a large number of annotated aerial images for the implementation of effective deep feature extractors.

3.10 | Unsupervised deep feature extraction

Although supervised DL methods such as CNNs and their variants can achieve impressive performance in image classification, there are limitations since their performance generally relies on a large number of labeled training samples. Yet, in the RS image domain, high-fidelity images with labels are rather limited. It is therefore, necessary to learn the features with

unlabeled images. Unsupervised feature extraction methods can learn feature representations from the images or patches with no prior labels. Traditional unsupervised feature extraction methods include RBMs, AEs, sparse coding, and *k*-means clustering. For instance, Risojevic and Babic (2014) proposed an approach combining quaternion PCA and *k*-means for unsupervised feature learning that makes joint encoding of the intensity and color information possible. Cheriyadat (2013) introduced a sparse coding-based method where the dense low-level features were extracted and encoded in terms of the basis functions in an effort to generate a new sparse representation. All of these feature-learning models are shadow and can be stacked to form deep unsupervised models, several of which have been successfully applied to RS image scene classification (Zhang et al., 2014).For instance, Zhang et al. (2014) proposed an unsupervised feature learning framework for scene classification using the saliency-guided sparse AE model and a new dropout technique (Hinton, Srivastava, Krizhevsky, Sutskever, & Salakhutdinov, 2012). In addition, CNNs can also be trained in unsupervised fashion, by means of greedy layerwise pretraining (Lee, Grosse, Ranganath, & Ng, 2009; Masci, Meier, Dan, & Schmidhuber, 2011; Schlkopf, Platt, & Hofmann, 2006b). For example, the use of deep CNNs for RS image classification was introduced in Romero, Gatta, and Camps-Valls (2016) with the network being trained by an unsupervised method that seeks sparse feature representation.

3.11 | Experimental results and analysis

The work reported in Xia et al. (2016) comprehensively compared those existing scene classification methods which use low-level visual features, mid-level visual representations, or high-level vision information, on the UC-Merced land-use data set, WHU-RS19 data set, RSSCN7 data set, and AID data set, respectively. Here, we selected representative experimental results from Xia et al. (2016) to evaluate different scene classification methods as shown in Table 5. Specifically, our selections include three low-level features (LBP [Ojala et al., 2000], SIFT [Kim et al., 2009], CH features [Swain & Ballard, 1991]), four middle-level feature coding methods (BoVW [Sivic & Zisserman, 2003], SPM [Lazebnik et al., 2006], LLC [Wang et al., 2010], and IFK [Perronnin & Mensink, 2010]), and three pretrained CNN models (CaffeNet [Jia et al., 2014], VGG-VD-16 [Simonyan & Zisserman, 2014], and GoogleNet [Szegedy et al., 2014]) for high-level scene feature extraction.

Table 5 clearly demonstrates that CNN-based methods can significantly improve the performance attainable by traditional handcrafted feature-based techniques. This may well reflect the general observation deep features are normally more robust than hand-crafted features. Hand-created features are predesigned and it is difficult to design features that can extract all required information. Different from this, deep features learned from input images by the model itself directly. In the experiments reported, CaffeNet uses only eight layers, and VGG-VD-16 and GoogLeNet use 16 and 22 layers, respectively. However, CaffeNet performs similarly as VGG-Net on all the data sets, while outperforming GoogLeNet. Note that all the networks are pretrained using the natural image data set ILSVRC 2012 (Russakovsky et al., 2015) and are directly used as feature extractors in the experiments. Therefore, the deeper the network, the more strongly the learned features facilitate the natural image processing task, although such networks may lead to worse performance in classifying aerial scenes (Xia et al., 2016).

4 | CONCLUSIONS AND FURTHER RESEARCH

In this literature survey, we have briefly introduced a number of typical DL models that may be used to perform RS image classification, including: CNNs, SAEs and DBNs. Following the introduction, from two main perspectives, pixel-wise image classification and scene-wise image classification, we have systematically reviewed the state-of-the-art DL approaches for RS image classification. In particular, classification methods based on spectral features, spatial features and joint spectral and spatial features have been discussed having both supervised and unsupervised feature extraction methods using DL. We have also

TABLE 5 Results of scene classification methods

Feature level	Method	UC-Merced (50%)	WHU-RS19 (60%)	RSSCN7 (50%)	AID (50%)
Low	SIFT	28.92	27.21	32.76	16.67
	LBP	34.57	44.08	60.38	29.99
	CH	42.09	51.87	60.54	37.28
Middle	BoVW(SIFT)	71.9	80.13	81.34	67.65
	SPM(SIFT)	56.5	55.82	68.45	45.52
	LLC(SIFT)	77.08	80.71	83.34	75.01
	IFK(SIFT)	77.09	86.95	84.41	77.33
High	CaffeNet	93.98	96.24	88.35	86.86
	VGGNet	94.14	96.05	87.18	86.59
	GoogleNet	92.7	94.71	85.84	83.44

compared and analyzed the performances of such typical methods. The performance of DL-based RS classification techniques have shown their effectiveness in solving real-world problems, although such performance does not reflect the full potential of DL yet. In the upcoming years, rapid advancement of DL in remote sensing image classification is expected, owing to the increased availability of RS data and computational resources. Nevertheless, there is still a long way to go in order to realize full potential while coping with many unanswered challenges. We discuss several important open issues and point out the corresponding possible future directions in addressing such issues as follows.

1. Limited labeled samples: Although DL models can learn high-level abstract features from raw images with excellent performance in dealing with a wide range of problems, we have to pay attention to the observation that such performance heavily relies on large amounts of training samples. In RS images, the available labeled samples are rather limited, thereby restricting the DL-based RS images classification approaches to obtain better performance. How to build an efficient network and train it with a small number of training samples is both challenging and interesting. Investigating into novel models that can exploit unlabeled samples is clearly a desirable direction for further work.
2. Transfer between data sets: For natural image classification, the common practice is to pretrain a DL model using a data set with a large number of labeled samples, such as ImageNet, and then to fine-tune the model using a data set which contains limited training samples. However, RS data are more complex than natural image, parts of them are typically even acquired by the use of different remote sensors. How to introduce transfer learning to RS image classification therefore, presents a major challenge, which needs significant further research.
3. DL model architecture: Recently, an increasing number of novel deep networks have been proposed. These networks can often achieve excellent performance in performing their dedicated task. For instance, U-net (Ronneberger, Fischer, & Brox, 2015) can obtain an impressive performance in segmentation, ResNet (He, Zhang, Ren, & Sun, 2015) can have an outstanding accuracy in applicable image classification and object detection. However, almost of such networks are aimed at coping with natural image processing. As we mentioned previously, RS images are generally different from natural images. Exploring appropriate network structures for a given RS image classification problem is still an open topic.

In addition, there are other HSI classification techniques that have been proposed in recent years. These techniques may also lead to good performance and therefore should be worth paying attention to. For instance, Zhu, Hu, Jia, and Li (2018) proposed a multiple 3D feature fusion framework to extract spectral-spatial features via 3D morpholigical profiles, 3D LBPs and 3D gabor surface features. Also, Fang, He, Li, Ghamisi, and Benediktsson (2018) proposed a novel fusion framework termed extinction profile fusion to exploit the information contained within and among EPs for HSI classification. Such recent developments are not examined in detail in this work, but may establish themselves in future practical applications.

ACKNOWLEDGMENTS

This research has received funding from the National Key Research and Development Program of China (Grant No. 2016YFB0502502), Foundation Project for Advanced Research Field (614023804016HK03002), and Shaanxi International Scientific and Technological Cooperation Project (2017KW-006).

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

RELATED WIREs ARTICLES

A survey on graphic processing unit computing for large-scale data mining

ORCID

Qiang Shen http://orcid.org/0000-0001-9333-4605

REFERENCES

Aptoula, E., Ozdemir, M. C., & Yanikoglu, B. (2016). Deep learning with attribute profiles for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, *13*(12), 1970–1974.

Bengio, Y. (2009). Learning deep architectures for ai. *Foundations & Trends in Machine Learning*, *2*(1), 1–127.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

Berlin, B., & Kay, P. (1969). *Basic color terms : Their universality and evolution*. University of California Press.

Castelluccio, M., Poggi, G., Sansone, C., & Verdoliva, L. (2015). Land use classification in remote sensing images by convolutional neural networks. *Acta Ecologica Sinica*, *28*(2), 627–635.

Chen, H. T., Chang, H. W., & Liu, T. L. (2005). *Local discriminant embedding and its variants.* Paper presented at IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 846–853.

Chen, X., Xiang, S., Liu, C. L., & Pan, C. H. (2013). *Aircraft detection by deep belief nets.* Paper presented at 2013 2nd IAPR Asian Conference on Pattern Recognition (ACPR), Naha, Japan. https://doi.org/10.1109/ACPR.2013.5

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. (2017). 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). Spectralspatial 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.

Cheriyadat, A. M. (2013). Unsupervised feature learning for aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(1), 439–451.

Deng, J., Dong, W., Socher, R., & Li, L. J. (2009). *Imagenet: A large-scale hierarchical image database.* Paper presented at IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 248–255.

Dong, C., Chen, C. L., He, K., & Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 295–307.

Du, Q., & Chang, C. I. (2001). A linear constrained distance-based discriminant analysis for hyperspectral image classification. *Pattern Recognition*, *34*(2), 361–373.

Ediriwickrema, J., & Khorram, S. (1997). Hierarchical maximum-likelihood classification for improved accuracies. *IEEE Transactions on Geoscience and Remote Sensing*, *35*(4), 810–816.

Fang, L., He, N., Li, S., Ghamisi, P., & Benediktsson, J. A. (2018). Extinction profiles fusion for hyperspectral images classification. *IEEE Transactions on Geoscience and Remote Sensing*, *PP*(99), 1–13.

Firat, O., Can, G., Vural, F. T. Y., Firat, O., Can, G., & Vural, F. T. Y. (2014). *Representation learning for contextual object and region detection in remote sensing.* Paper presented at International Conference on Pattern Recognition, Stockholm, Sweden, 3708–3713.

Freund, Y., & Haussler, D. (1991). Unsupervised learning of distributions on binary vectors using two layer networks. *Advances in Neural Information Processing Systems*, *4*, 912–919.

Girshick, R. (2015). Fast r-cnn. *Computer Science.* arXiv:1504.08083.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2013). *Rich feature hierarchies for accurate object detection and semantic segmentation.* Paper presented at Proceedings of the 2014 I.E. Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 580–587.

Golub, G., & Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics*, *2*(2), 205–224.

Han, X., Zhong, Y., & Zhang, L. (2016). Spatial-spectral classification based on the unsupervised convolutional sparse auto-encoder for hyperspectral remote sensing imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *III-7*, 25–31.

He, K., Zhang, X., Ren, S., & Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1904–1916.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition* (pp. 770–778).

He, M., Li, X., Zhang, Y., Zhang, J., & Wang, W. (2016). Hyperspectral image classification based on deep stacking network. Paper presented at Geoscience and Remote Sensing Symposium, Beijing, China, 3286–3289.

Hinton, G. E. (2002). *Training products of experts by minimizing contrastive divergence*. Cambridge, MA: MIT Press.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *Computer Science*, *3*(4), 212–223.

Hu, F., Xia, G. S., Hu, J., & Zhang, L. (2015). Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing*, *7*(11), 14680–14707.

Hu, F., Xia, G. S., Hu, J., Zhong, Y., & Xu, K. (2016). Fast binary coding for the scene classification of high-resolution remote sensing imagery. *Remote Sensing*, *8*(7), 555.

Hu, F., Xia, G. S., Wang, Z., Huang, X., Zhang, L., & Sun, H. (2017). Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(5), 2015–2030.

Hu, W., Huang, Y., Wei, L., Zhang, F., & Li, H. (2015). Deep convolutional neural networks for hyperspectral image classification. *Journal of Sensors*, *2015*(2), 1–12.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. eprint arXiv:1502.03167, 448–456.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). Caffe: Convolutional architecture for fast feature embedding, arXiv preprint arXiv: 1408.5093, 675–678.

Johnson, R., & Zhang, T. (2013). *Accelerating stochastic gradient descent using predictive variance reduction.* Paper presented at International Conference on Neural Information Processing Systems, Daegu, South Korea, 315–323.

Kim, M. H., Madden, M., & Warner, T. A. (2009). Forest type mapping using object-specific texture measures from multispectral ikonos imagery: Segmentation quality and image classification issues. *Photogrammetric Engineering & Remote Sensing*, *75*(7), 819–829.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks.* Paper presented at International Conference on Neural Information Processing Systems, Doha, Qatar, 1097–1105.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories.* Paper presented at IEEE Computer Society Conference on Computer Vision & Pattern Recognition, New York, NY, USA, 2169–2178

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). *Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations.* Paper presented at International Conference on Machine Learning, Montréal, Canada, 609–616.

Lee, H., & Kwon, H. (2016). *Contextual deep cnn based hyperspectral classification.* Paper presented at Geoscience and Remote Sensing Symposium, Beijing, China, 1–1.

Li, J., Bioucas-Dias, J. M., & Plaza, A. (2010). Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Transactions on Geoscience and Remote Sensing*, *48*(11), 4085–4098.

Li, J., Bruzzone, L., & Liu, S. (2015). *Deep feature representation for hyperspectral image classification.* Paper presented at Geoscience and Remote Sensing Symposium, Milan, Italy, 4951–4954.

Li, T., Zhang, J., & Zhang, Y. (2015). Classification of hyperspectral image based on deep belief networks. In *IEEE international conference on image processing* (p. 5132–5136).

Li, W., Chen, C., Su, H., & Du, Q. (2015). Local binary patterns and extreme learning machine for hyperspectral imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(7), 3681–3693.

Li, Y., Xie, W., & Li, H. (2016). Hyperspectral image reconstruction by deep convolutional neural network for classification. *Pattern Recognition*, *63*, 371–383.

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

Liang, H., & Li, Q. (2016). Hyperspectral imagery classification using sparse representations of convolutional neural network features. *Remote Sensing*, *8*(2), 99.

Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *Computer Science.*

Lin, Z., Chen, Y., Zhao, X., & Wang, G. (2015). Spectral-spatial classification of hyperspectral image using autoencoders. Paper pretended at International Conference on Information, Communications and Signal Processing,Tainan, Taiwan, 1–5.

Long, J., Shelhamer, E., & Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 640–651.

Luus, F. P. S., Salmon, B. P., Bergh, F. V. D., & Maharaj, B. T. J. (2015). Multiview deep learning for land-use classification. *IEEE Geoscience and Remote Sensing Letters*, *12*(12), 2448–2452.

Ma, X., Geng, J., & Wang, H. (2015). Hyperspectral image classification via contextual deep learning. *Eurasip Journal on Image & Video Processing*, *2015*(1), 20.

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

Ma, X., Wang, H., Geng, J., & Wang, J. (2016). *Hyperspectral image classification with small training set by deep network and relative distance prior.* Paper presented at Geoscience and Remote Sensing Symposium, Beijing, China, 3282–3285.

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 & Remote Sensing*, *120*, 99–107.

Makantasis, K., Karantzalos, K., Doulamis, A., & Doulamis, N. (2015). *Deep supervised learning for hyperspectral data classification through convolutional neural networks.* Paper presented at Geoscience and Remote Sensing Symposium, Milan, Italy, 4959–4962.

Masci, J., Meier, U., Dan, C., & Schmidhuber, J. (2011). *Stacked convolutional autoencoders for hierarchical feature extraction.* Paper presented at International Conference on Artificial Neural Networks, Espoo, Finland, 52–59.

Mei, S., Ji, J., Bi, Q., Hou, J., Du, Q., & Li, W. (2016). *Integrating spectral and spatial information into deep convolutional neural networks for hyperspectral classification.* Paper presented at Geoscience and Remote Sensing Symposium, Beijing, China, 5067–5070.

Mou, L., Ghamisi, P., & Zhu, X. (2017). Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(7), 3639–3655.

Mura, M. D., Benediktsson, J. A., Waske, B., & Bruzzone, L. (2010). Morphological attribute profiles for the analysis of very high resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, *48*(10), 3747–3762.

Nogueira, K., Penatti, O. A. B., & Santos, J. A. D. (2016). Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, *61*, 539–556.

Ojala, T., Pietikinen, M., & Menp, T. (2000). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987.

Ouyang, W., Luo, P., Zeng, X., Qiu, S., Tian, Y., Li, H., … Tang, X. (2014). Deepid-net: Multi-stage and deformable deep convolutional neural networks for object detection. *Eprint Arxiv.*

Penatti, O. A. B., Nogueira, K., & Santos, J. A. D. (2015). *Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?* Paper presented at Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 44–51.

Perronnin, F., & Mensink, T. (2010). *Improving the fisher kernel for large-scale image classification.* Paper presented at European Conference on Computer Vision, Hersonissos, Heraklion, Crete, Greece, 143–156.

Risojevic, V., & Babic, Z. (2014). *Unsupervised learning of quaternion features for image classification.* Paper presented at International Conference on Telecommunication in Modern Satellite, Cable and Broadcasting Services, Nis, Serbia, 345–348.

Romero, A., Gatta, C., & Camps-Valls, G. (2016). Unsupervised deep feature extraction for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(3), 1349–1362.

Ronan Collobert, J. W. & Weston, J. (2008). *A unified architecture for natural language processing: Deep neural networks with multitask.* Paper presented at Proceedings of the 25th International Conference on Machine, New York, USA, 160–167.

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-net: Convolutional networks for biomedical image segmentation*. Springer International Publishing.

Rumelhart, D., & Mcclelland, J. (1988). *Learning internal representations by error propagation*. Cambridge, MA: MIT Press.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Samaniego, L., Bardossy, A., & Schulz, K. (2008). Supervised classification of remotely sensed imagery using a modified k-nn technique. *IEEE Transactions on Geoscience and Remote Sensing*, *46*(7), 2112–2125.

Scherer, D., Muller, A., & Behnke, S. (2010). *Evaluation of pooling operations in convolutional architectures for object recognition.* Paper presented at International Conference on Artificial Neural Networks, Thessaloniki, Greece, 92–101.

Schlkopf, B., Platt, J., & Hofmann, T. (2006a). *Efficient learning of sparse representations with an energy-based model.* Paper presented at Advances in Neural Information Processing Systems, Vancouver, B.C., Canada, 1137–1144.

Schlkopf, B., Platt, J., & Hofmann, T. (2006b). *Greedy layer-wise training of deep networks.* Paper presented at International Conference on Neural Information Processing Systems, Vancouver, B.C., Canada, 153–160.

Sheng, G., Yang, W., Xu, T., & Sun, H. (2012). High-resolution satellite scene classification using a sparse coding based multiple feature combination. *International Journal of Remote Sensing*, *33*(8), 2395–2412.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Computer Science.*

Sivic, J., & Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. Paper presented at IEEE International Conference on Computer Vision, Nice, France, 1470.

Slavkovikj, V., Verstockt, S., Neve, W. D., Hoecke, S. V., & Walle, R. V. D. (2015). *Hyperspectral image classification with convolutional neural networks.* Paper presented at ACM International Conference on Multimedia, 1159–1162.

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11–32.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2014). Going deeper with convolutions, arXiv:1409.4842, 1–9.

Tao, C., Pan, H., Li, Y., & Zou, Z. (2015). Unsupervised spectralspatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification. *IEEE Geoscience and Remote Sensing Letters*, *12*(12), 2438–2442.

Thompson, W. D., & Walter, S. D. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology*, *41*(10), 949–958.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008). *Extracting and composing robust features with denoising autoencoders.* Paper presented at International Conference on Machine Learning, Kunming, China, 1096–1103.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, *11*(12), 3371–3408.

Volpi, M., & Tuia, D. (2016). Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, *PP*(99), 1–13.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., & Gong, Y. (2010). *Locality-constrained linear coding for image classification.* Paper presented at Computer vision and pattern recognition, San Francisco, CA, USA, 3360–3367.

Wu, H., & Prasad, S. (2017). Convolutional recurrent neural networks for hyperspectral data classification. *Remote Sensing*, *9*(1), 298.

Xia, G. S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., & Zhang, L. (2016). Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *PP*(99), 1–17.

Xia, G. S., Yang, W., Delon, J., Gousseau, Y., Sun, H., & Matre, H. (2010). *Structural highresolution satellite image indexing.* Paper presented at ISPRS TC VII Symposium - 100 Years ISPRS, XXXVIII, Vienna, Austria, 298–303.

Xing, C., Ma, L., & Yang, X. (2015, 2016). Stacked denoise autoencoder based feature extraction and classification for hyperspectral images. *Journal of Sensors*, *2016*, 1–10.

Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (2015). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *23*(1), 7–19.

Yang, J., Zhao, Y., Chan, C. W., & Yi, C. (2016). *Hyperspectral image classification using two-channel deep convolutional neural network.* Paper presented at Geoscience and Remote Sensing Symposium, Beijing, China, 5079–5082.

Yang, W., Yin, X., & Xia, G. S. (2015). Learning high-level features for satellite image classification with limited labeled samples. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(8), 4472–4482.

Yue, J., Mao, S., & Li, M. (2016). A deep learning framework for hyperspectral image classification using spatial pyramid pooling. *Remote Sensing Letters*, *7*(9), 875–884.

Yue, J., Zhao, W., Mao, S., & Liu, H. (2015). Spectralspatial classification of hyperspectral images using deep convolutional neural networks. *Remote Sensing Letters*, *6*(6), 468–477.

Zeiler, M. D., & Fergus, R. (2013). Stochastic pooling for regularization of deep convolutional neural networks. *Eprint Arxiv.* arXiv:1301.3557

Zhang, F., Du, B., & Zhang, L. (2014). Saliency-guided unsupervised feature learning for scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(4), 2175–2184.

Zhang, F., Du, B., & Zhang, L. (2016). Scene classification via a gradient boosting random convolutional network framework. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(3), 1793–1802.

Zhang, H., Li, Y., Zhang, Y., & Shen, Q. (2017). Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network. *Remote Sensing Letters*, *8*(5), 438–447.

Zhang, L., Wei, W., Zhang, Y., Shen, C., van den Hengel, A., & Shi, Q. (2018). Cluster sparsity field: An internal hyperspectral imagery prior for reconstruction. *International Journal of Computer Vision*, 1–25.

Zhang, L., Zhang, L., & Du, B. (2016). Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, *4*(2), 22–40.

Zhao, B., Zhong, Y., Xia, G. S., & Zhang, L. (2016). Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(4), 2108–2123.

Zhao, L., Tang, P., & Huo, L. (2016). Feature significance-based multibag-of-visual-words model for remote sensing image scene classification. *Journal of Applied Remote Sensing*, *10*(3), 035004.

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

Zhao, W., Guo, Z., Yue, J., Luo, L., & Luo, L. (2015). On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery. *International Journal of Remote Sensing*, *36*(13), 3368–3379.

Zhong, P., Gong, Z. Q., & Schnlieb, C. (2016). A diversified deep belief network for hyperspectral image classification. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLI-B7*, 443–449.

Zhu, J., Hu, J., Jia, S., & Li, Q. (2018). Multiple 3-d feature fusion framework for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *PP*(99), 1–14.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A review. arXiv:1710.03959

Zou, Q., Ni, L., Zhang, T., & Wang, Q. (2015). Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, *12*(11), 2321–2325.

How to cite this article: Li Y, Zhang H, Xue X, Jiang Y, Shen Q. Deep learning for remote sensing image classification: A survey. *WIREs Data Mining Knowl Discov*. 2018;e1264. https://doi.org/10.1002/widm.1264