

Deep Learning for Hyperspectral Image Classification: An Overview

Shutao Li, *Fellow, IEEE*, Weiwei Song, *Student Member, IEEE*, Leyuan Fang, *Senior Member, IEEE*, Yushi Chen, *Member, IEEE*, Pedram Ghamisi, *Senior Member, IEEE*, and Jón Atli Benediktsson, *Fellow, IEEE*

Abstract—Hyperspectral image (HSI) classification has become a hot topic in the field of remote sensing. In general, the complex characteristics of hyperspectral data make the accurate classification of such data challenging for traditional machine learning methods. In addition, hyperspectral imaging often deals with an inherently nonlinear relation between the captured spectral information and the corresponding materials. In recent years, deep learning has been recognized as a powerful feature-extraction tool to effectively address nonlinear problems and widely used in a number of image processing tasks. Motivated by those successful applications, deep learning has also been introduced to classify HSIs and demonstrated good performance. This survey paper presents a systematic review of deep learning-based HSI classification literatures and compares several strategies for this topic. Specifically, we first summarize the main challenges of HSI classification which cannot be effectively overcome by traditional machine learning methods, and also introduce the advantages of deep learning to handle these problems. Then, we build a framework which divides the corresponding works into spectral-feature networks, spatial-feature networks, and spectral-spatial-feature networks to systematically review the recent achievements in deep learning-based HSI classification. In addition, considering the fact that available training samples in the remote sensing field are usually very limited and training deep networks require a large number of samples, we include some strategies to improve classification performance, which can provide some guidelines for future studies on this topic. Finally, several representative deep learning-based classification methods are conducted on real HSIs in our experiments.

Index Terms—Hyperspectral image, deep learning, classification, feature extraction.

I. INTRODUCTION

HYPERSPECTRAL imaging is an important technique in remote sensing, which collects the electromagnetic spectrum from the visible to the near-infrared wavelength

This work was supported by the National Natural Science Fund of China under Grant 61890962, 61520106001 and 61771192, the Science and Technology Plan Project Fund of Hunan Province under Grant CX2018B171, 2017RS3024 and 2018TP1013, the Science and Technology Talents Program of Hunan Association for Science and Technology under Grant 2017TJ-Q09, and the National Key RD Program of China under Grant 2018YFB1305200.

S. Li, W. Song, and L. Fang are with the College of Electrical and Information Engineering, Hunan University, Changsha, 410082, China, and also with the Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province, Changsha, 410082, China (e-mail: shutao_li@hnu.edu.cn; weiwei_song@hnu.edu.cn; fangleyuan@gmail.com).

Y. Chen is with the Department of Information Engineering, School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: chenyushi@hit.edu.cn).

P. Ghamisi is with the Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Helmholtz Institute Freiberg for Resource Technology (HIF), Exploration, D-09599 Freiberg, Germany (e-mail: p.ghamisi@gmail.com).

J. A. Benediktsson is with the Faculty of Electrical and Computer Engineering, University of Iceland, 101 Reykjavík, Iceland (e-mail: benedikt@hi.is).

ranges. Hyperspectral imaging sensors often provide hundreds of narrow spectral bands from the same area on the surface of the earth. In hyperspectral images (HSIs), each pixel can be regarded as a high-dimensional vector whose entries correspond to the spectral reflectance in a specific wavelength. With the advantage of distinguishing subtle spectral difference, HSIs have been widely applied in many fields [1]–[5].

Based on the recent studies published in [6], HSI classification (i.e., assigning each pixel to one certain class based on its spectral characteristics) is the most vibrant field of research in the hyperspectral community and has drawn broad attentions in the remote sensing field. In HSI classification tasks, there exist two main challenges: 1) the large spatial variability of spectral signatures and 2) the limited available training samples versus the high dimensionality of hyperspectral data. The first challenge is often brought by many factors such as changes in illumination, environmental, atmospheric, and temporal conditions. The second challenge will result in ill-posed problems for some methods and reduce the generalization ability of classifiers.

In the early stage of the study on HSI classification, most methods have focused on exploring the role of the spectral signatures of HSIs for the purpose of classification. Thus, numerous pixel-wise classification methods (e.g., neural networks [7], support vector machines (SVM) [8], multinomial logistic regression [9], [10], and dynamic or random subspace [11], [12]) have been proposed to classify HSIs. In addition, some other classification approaches have focused on designing an effective feature-extraction or dimension-reduction technique, such as principle component analysis (PCA) [13], [14], independent component analysis (ICA) [15], and linear discriminant analysis (LDA) [16]. However, the classification maps obtained by these pixel-wise classifiers are unsatisfactory since the spatial contexts are not considered. Recently, spatial features have been reported to be very useful in improving the representation of hyperspectral data and increasing the classification accuracies [17], [18]. More and more spectral-spatial features-based classification frameworks have been developed, which incorporates the spatial contextual information into pixel-wise classifiers. For example, in [19], [20], extended morphological profiles (EMPs) were used to exploit the spatial information via multiple morphological operations. Multiple kernel learning (e.g., composite kernel [21] and morphological kernel [22], [23]) was designed to explore the spectral-spatial information of HSIs. In [24], edge-preserving filtering was considered as a postprocessing technique to optimize the probabilistic results of an SVM. In

[25]–[27], the spatial information within a neighboring region was incorporated into a sparse representation model. These sparse representation methods are based on the observation that hyperspectral pixels can usually be represented by a linear combination of a few common pixels from the same class. Furthermore, spatial consistency was explored by segmenting HSIs into multiple superpixels based on the similarity of either intensity or texture [28]–[30]. Although most of spectral-spatial-based HSI classification methods have obtained good performance, they heavily depend on hand-crafted or shallow-based descriptors. However, most hand-crafted features are usually designed for a specific task and depend on expert knowledge in the parameter setup phase, which limits the applicability of those approaches in difficult scenarios. The ability of representation of hand-crafted features may not be enough to discriminate subtle variation between the different classes or a large variation between the same classes. Extracting more discriminative features is considered as a crucial procedure for HSI classification.

Recently, deep learning has become a growing trend in big data analysis and great breakthrough has been made with the approach in many computer vision tasks, e.g., image classification [31], [32], object detection [33], and natural language processing [34]. Motivated by those successful applications, deep learning has been introduced to classify HSIs and achieved good performance. Fig. 1 demonstrates the statistics for published papers related to deep learning-based HSI classification in the past five years according to the web of science¹. From this figure, we can conclude that the topic will be further explored with deep learning and more and more research works will be published in the next several years. Compared with traditional hand-crafted methods, deep learning techniques can extract informative features from the original data via a series of hierarchical layers. Specifically, earlier layers extract some simple features like texture and edge information. Furthermore, the deeper layers are able to represent more complicated features. The learning process is totally automatic, which makes deep learning more suitable for coping with the varieties of situations. In general, deep learning is recognized as an effective feature extraction approach during HSI classification while different networks focus on extracting different feature types.

In this paper, we focus on deep learning-based classification methodologies for HSIs and aim at providing a relatively general and comprehensive overview on the existing methods. The motivation of our work is based on two aspects. On one hand, this work is expected to clarify the mechanism behind existing deep learning-based classification methods. We systematically review a large number of relevant papers in the literature within the framework, where the papers are divided into spectral-feature networks, spatial-feature networks, and spectral-spatial-feature networks with respect to the types of features extracted by the adopted deep networks. On the other hand, we intend to include some strategies to handle the problem of limited available samples (i.e., the second challenge mentioned above), which is very important for

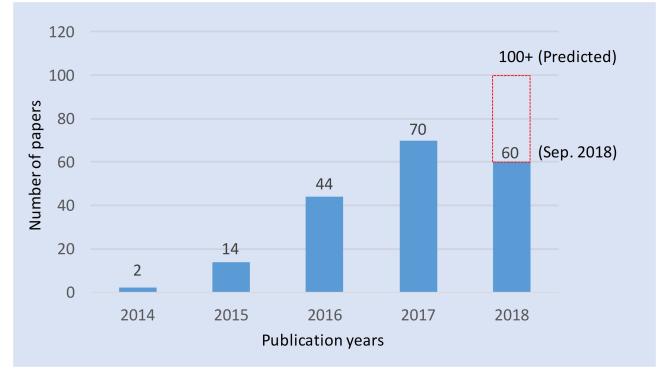


Fig. 1. The statistics for published papers related to deep learning in HSI classification according to the web of science.

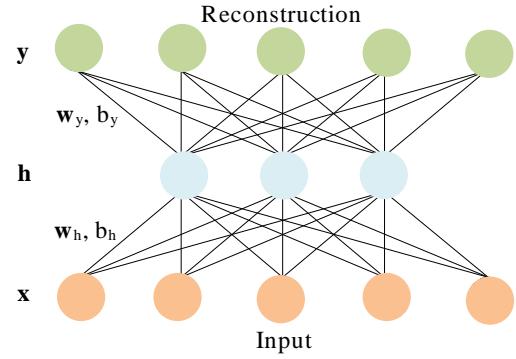


Fig. 2. A single hidden-layer auto-encoder. The model learns a hidden feature “ h ” by minimizing the error between the input “ x ” and the reconstructed output “ y ” [36].

designing a deep-learning method for HSI classification.

The remainder of this paper is organized as follows: Section II introduces several typical deep models. In Section III, we categorize the previous works into three aspects with respect to the types of features extracted by deep networks. Some strategies to cope with the problem of limited available samples are introduced in Section IV. In Section V, several representative deep learning-based classification methods are compared on real HSIs. Finally, conclusions and suggestions are provided in Section VI.

II. DEEP MODELS

In this section, we briefly introduce several deep network models that have been widely used in the HSI classification field. These deep networks include stacked auto-encoders (SAEs), deep belief networks (DBNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks (GANs).

A. SAEs

Auto-encoder (AE) is known as the main building block of the stacked AE (SAE) [35]. Fig. 2 shows an AE composing of one visible layer of d inputs, one hidden layer of L units, and one reconstruction layer of d units. The training procedure has two steps. First, $\mathbf{x} \in \mathbb{R}^d$ is mapped to $\mathbf{h} \in \mathbb{R}^L$ in the hidden

¹<http://www.webofknowledge.com/WOS>

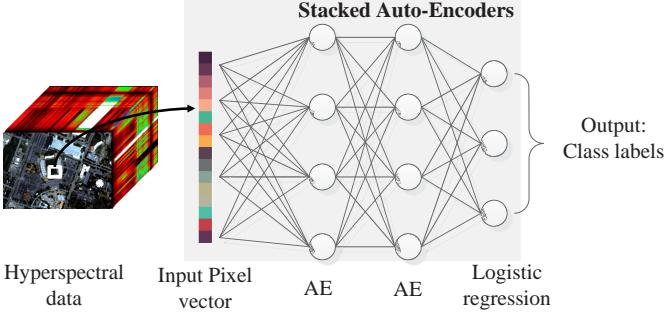


Fig. 3. A general illustration of an SAE connected with a subsequent logistic regression classifier [36].

layer (i.e., ‘‘encoder’’). Second, \mathbf{h} is mapped to $\mathbf{y} \in \mathbb{R}^d$ (i.e., ‘‘decoder’’). These two steps can also be defined as follows:

$$\mathbf{h} = f(\mathbf{w}_h \mathbf{x} + b_h)$$

$$\mathbf{y} = f(\mathbf{w}_y \mathbf{x} + b_y)$$

where \mathbf{w}_h and \mathbf{w}_y denote the input-to-hidden and the hidden-to-output weights, respectively. b_h and b_y represent the *bias* of the hidden and output units, and $f(\cdot)$ represents an activation function. The reconstruction error is estimated using the Euclidean distance between \mathbf{x} and \mathbf{y} to approximate the input data \mathbf{x} by minimizing $\|\mathbf{x} - \mathbf{y}\|_2^2$.

SAE can be constructed by stacking multiple layers of AEs that attaches the output of one layer to the input of the following layers. Fig. 3 illustrates a simple representation of an SAE connected with a subsequent logistic regression classifier. The SAE can be used as a spectral classifier where each pixel vector can be considered as input.

B. DBNs

Restricted Boltzmann machine (RBM) is a layer-wise training model and considered as the main building block of a DBN [37]. Fig. 4 illustrates an RBM, consisting of a two-layer network with ‘‘visible’’ units $\mathbf{v} = \{0, 1\}^d$ and ‘‘hidden’’ units $\mathbf{h} = \{0, 1\}^L$. Given these units, the energy of a joint configuration of the units can be defined by:

$$\begin{aligned} E(\mathbf{v}, \mathbf{h}; \theta) &= - \sum_{i=1}^d b_i v_i - \sum_{j=1}^L a_j h_j - \sum_{i=1}^d \sum_{j=1}^L w_{ij} v_i h_j \quad (1) \\ &= -\mathbf{b}^T \mathbf{v} - \mathbf{a}^T \mathbf{h} - \mathbf{v}^T \mathbf{w} \mathbf{h} \end{aligned}$$

where $\theta = \{b_i, a_j, w_{ij}\}$. Here, w_{ij} is the weight between the visible unit i and hidden unit j . b_i and a_j represent the *bias* of the visible and hidden units, respectively. The weight, w_{ij} , is learned using the divergence [37]. In RBM, the hidden units are conditionally independent given the visible states, and consequently, can obtain an unbiased sample from the posterior distribution when given a data vector.

In order to increase the feature representation capability of a single RBM, several RBMs can be stacked one after another which establishes a DBN, which can be learned to extract a deep hierarchical representation of the training data. Fig. 5 illustrates a DBN composing of several RBM layers. In order

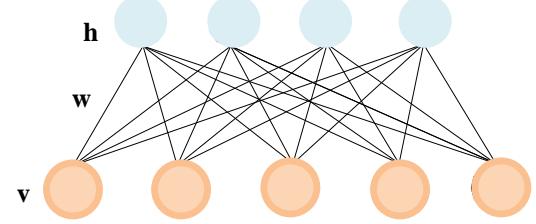


Fig. 4. Graphical illustration of a restricted Boltzmann machine. The top layer represents the hidden units and the bottom layer represents the visible units [36].

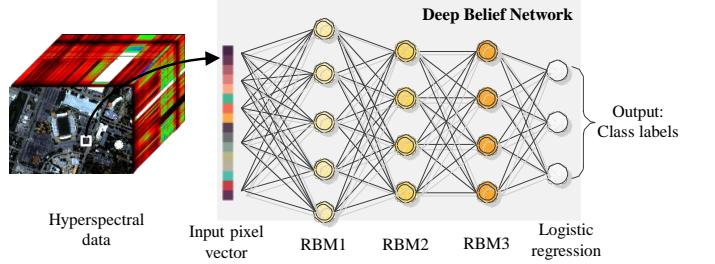


Fig. 5. A spectral classifier based on DBN. The classification scheme shown here has four layers: one input layer, two RBMs, and one logistic regression layer [36].

to use DBN as a classifier, one can add a logistic regression layer at the end of the network to form a spectral classifier.

C. CNNs

CNNs are inspired by the structure of the visual system. In contrast to fully connected networks, CNNs make use of local connections to extract the contextual 2-D spatial features of images. In addition, network parameters can be significantly reduced via the weight-share mechanism. A representative structure of CNNs is shown in Fig. 6, which mainly consists of a stack of alternating convolution layers and pooling layers with a number of fully connected layers. In the convolutional layers, the image patches with spatial context information are convolved with a set of kernels. Then, the pooling layers reduce the size of feature maps created by the convolutional layers to obtain more general and abstract features. Finally, these feature maps are further transformed into feature vectors via several fully connected layers. Here, each component is described as follows.

1) *Convolutional Layers*: In general, convolutional layers are the most important parts of CNNs. At each convolutional layer, the input cube is convoluted with multiple learnable filters, resulting in generating multiple feature maps. Specifically, let \mathbf{X} be the input cube and its size is $m \times n \times d$, where $m \times n$ refers to the spatial size of \mathbf{X} , d is the number of channels, and \mathbf{x}_i is the i -th feature map of \mathbf{X} . Supposing there are k filters at this convolutional layer and the j -th filter can be characterized by the weight \mathbf{w}_j and *bias* b_j . The j -th output of the convolutional layer can be represented as follows:

$$\mathbf{y}_j = \sum_{i=1}^d f(\mathbf{x}_i * \mathbf{w}_j + b_j), j = 1, 2, \dots, k \quad (2)$$

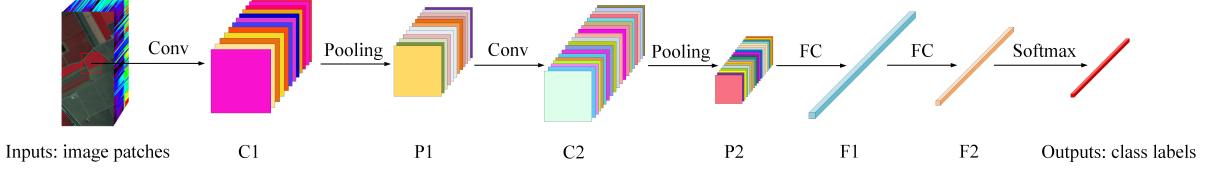


Fig. 6. The architecture of a conventional CNN, which consists of two convolutional-pooling layers and two fully connected layers.

where $*$ refers to the convolution operator and $f(\cdot)$ is an activation function which is utilized to improve the nonlinearity of the network. Recently, ReLU [31] has been the mostly-used activation function. ReLU mainly has two advantages: a fast convergence and robustness for gradient vanishing [38]. The formulation of ReLU is denoted as:

$$\sigma(\mathbf{x}) = \max(0, \mathbf{x}). \quad (3)$$

2) Pooling Layers: Due to the existence of redundant information in images, the pooling layers are periodically inserted after several convolutional layers in the CNNs. With the pooling operation, the spatial size of the feature maps is progressively reduced, at the same time, the amount of parameters and computation of the network are also decreased. Through the pooling operations, the sizes of feature maps tend to shrink, and the representation of extracted features becomes more abstract. Specifically, for a $p \times p$ window-size neighbor denoted as \mathbf{S} , the average pooling operation can be denoted as:

$$z = \frac{1}{F} \sum_{(i,j) \in \mathbf{S}} x_{ij} \quad (4)$$

where F is the number of elements in \mathbf{S} and x_{ij} is the activation value corresponding to the position (i, j) .

3) Fully Connected Layers: After the pooling layers, the feature maps of previous layer are flattened and fed to fully connected layers. In a traditional neural network, the fully connected layers are used to extract more deep and abstract features by reshaping feature maps into an n-dimension vector (e.g., $n=4096$ in AlexNet [31]). In general, the fully connected layer can be defined as:

$$\mathbf{Y}' = \sum_{i=1}^C f(\mathbf{W}\mathbf{X}' + \mathbf{b}) \quad (5)$$

where \mathbf{X}' , \mathbf{Y}' , \mathbf{W} , and \mathbf{b} refer to the input, output, weight, and *bias* of a fully connected layer, respectively.

D. RNNs

RNNs were originally introduced in [39], [40]. In contrast to a feedforward neural network, an RNN can recognize patterns in sequences of data and dynamic temporal characteristics by using a recurrent hidden state whose activation at each step depends on that of the previous steps.

Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)$ be sequential data, where \mathbf{x}_i is the data at the i -th time step. The recurrent hidden state $\mathbf{h}^{<t>}$ of the RNN at temporal sequence t can be updated by

$$\mathbf{h}^{<t>} = \begin{cases} 0 & \text{if } t = 0 \\ f_1(\mathbf{h}^{<t-1>}, \mathbf{x}^{<t>}) & \text{otherwise} \end{cases} \quad (6)$$

where f_1 is a nonlinear function (e.g., a logistic sigmoid function or hyperbolic tangent function). Conventionally, the update rule of the recurrent hidden state in (6) has been obtained by:

$$\mathbf{h}^{<t>} = f_1(\mathbf{w}\mathbf{x}^{<t>} + \mathbf{u}\mathbf{h}^{<t-1>} + \mathbf{b}_h) \quad (7)$$

where \mathbf{w} and \mathbf{u} represent the coefficient matrices for the input at the current step and for the activation of recurrent hidden units at the previous step, respectively, and \mathbf{b}_h shows the corresponding bias vector. Then, $\mathbf{h}^{<t>}$ will be used to predict $\mathbf{y}^{<t>}$ at time step t as follows:

$$\mathbf{y}^{<t>} = f_2(\mathbf{p}\mathbf{h}^{<t>} + \mathbf{b}_y) \quad (8)$$

where f_2 is a nonlinear function, \mathbf{p} is the coefficient matrix for the activation of recurrent hidden units at the current step, and \mathbf{b}_y is the corresponding bias vector. As can be noticed, one can use different nonlinearities, f_1 and f_2 , to estimate $\mathbf{h}^{<t>}$ and $\mathbf{y}^{<t>}$.

The performance of the conventional RNN can be downgraded when it deals with long-term sequential data due to the vanishing gradient or exploding gradient. To address this issue, long short-term memory (LSTM) [41], [42] and gated recurrent unit (GRU) [43] were introduced.

As discussed above, RNNs are able to recognize patterns in sequences of data. In this context, RNNs are well-suited for the analysis of HSI since each pixel vector in hyperspectral data can be considered as a set of orderly and continuing spectra sequences in the spectral space. In [44], the concept of RNN was adopted for pixel-wise spectral classification by developing a new activation function and a modified gated recurrent unit, which can effectively analyze hyperspectral pixels as sequential data and then assign a classification label via network reasoning.

E. GANs

In general, there are two typical classes of models in the machine learning community: generative approaches and discriminative approaches. Generative approaches try to learn the distribution parameters from data, and then they can generate new samples according to the learned models. Discriminative approaches attempt to model the dependence of labels (i.e.,

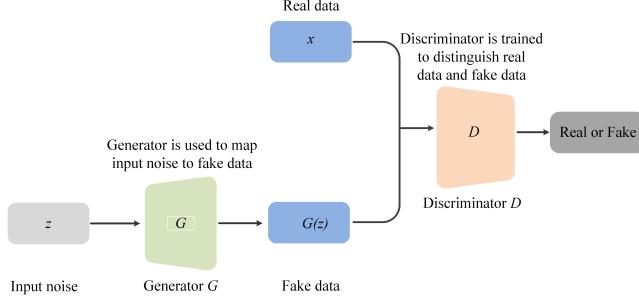


Fig. 7. A general illustration of a GAN [46].

y) on training data (i.e., x), which can be used for predicting y from x .

GANs are the relatively new kind of models, which contains a generative model G and a discriminative model D [45]. Models G and D are trained in an adversarial manner. Model G tries to generate fake inputs as real as possible, whereas model D tries to distinguish between real and fake inputs. Through the adversarial manner and competition of two models, the training process of the discriminator will proceed both continuously and effectively. The architecture of a GAN is shown in Fig. 7.

The generator G accepts random noise z as input and produces fake data $G(z)$. The discriminator D estimates the probability that x is a true sample. The discriminator D is trained to maximize $\log(D(x))$, which is the probability of assigning the right labels to the training samples; the generator G is trained to minimize $\log(1 - D(G(z)))$. So, the aim of the GAN is to solve the following minimax problem

$$\min_G \max_D V(D, G) = E_{x \sim p(x)} [\log(D(x))] + E_{z \sim p(z)} [\log(1 - D(G(z)))] \quad (9)$$

where E is the expectation operator. Due to the advantages of CNN, the deep convolutional generative adversarial network architecture, which uses deep convolution networks in G and D , was proposed in [47]. Although GAN is a promising technique, the original discriminator D only estimates whether the input sample is either real or fake. Therefore, it is not suitable for multi-class classification. In [48], Odena *et al.* proposed an auxiliary classifier GAN, which can be used for classification, in which D was modified to a softmax classifier that can output multi-class labels probabilities [48].

A proposed framework of GAN-based HSI classification is shown in Fig. 8. From Fig. 8, we can see that the generator G also accepts HSI class labels c in addition to the noise z and the output of G can be defined by $\mathbf{x}_{fake} = G(\mathbf{z}, c)$. The training samples with corresponding class labels and the fake data generated by G are regarded as the input of the discriminator D . The objective function contains two parts: the log-likelihood of the right source of input data L_s and the log-likelihood of the right class labels L_c :

$$L_s = E [\log P(s = real | \mathbf{x}_{real})] + E [\log P(s = fake | \mathbf{x}_{fake})] \quad (10)$$

$$L_c = E [\log P(c = real | \mathbf{c}_{real})] + E [\log P(c = fake | \mathbf{c}_{fake})]. \quad (11)$$

Therefore, D is optimized to maximize the $L_s + L_c$ while G is optimized to maximize the $L_c - L_s$.

III. DEEP NETWORKS-BASED HSI CLASSIFICATION

Recently, deep learning has become one of the most successful techniques and achieved impressive performance in the computer vision field [49]. Motivated by those great breakthroughs, deep learning has also been introduced to classify HSIs in the remote sensing field [50]–[52]. Compared with traditional hand-crafted features-based methods, deep learning can automatically learn high-level features from complex hyperspectral data. With these discriminative features, deep learning-based methods can effectively deal with the first problem mentioned above in Section I (i.e., the large spatial variability of spectral signature). Based on the fact, a large number of deep networks have been developed to extract features of HSIs and achieved good classification performance. However, the types of features extracted from deep networks may be different, e.g., spectral, spatial, and spectral-spatial features can be extracted by different deep networks. In this section, we systematically review the deep learning-based methods for HSI classification in a framework. In such framework, the deep networks used for HSI classification are divided into spectral-feature networks, spatial-feature networks, and spectral-spatial-feature networks. These different networks are expected to extract corresponding features for the subsequent classification. The following subsections will introduce the three kinds of networks in details.

A. Spectral-Feature Networks

Spectral information is the most important characteristic of HSIs and plays the vital role for the classification tasks [36]. However, the hyperspectral remote sensors usually provide hundreds of spectral bands which also contain redundant information. Therefore, the direct exploration of original spectral vectors not only results in high computational cost but also decreases the classification performance. Although some traditional spectral feature extraction methods (e.g., PCA [13], [14], ICA [15], and LDA [16]) can extract effective spectral features, the simple linear processing existed in these linear models is hard to handle the complex spectral property of HSIs. In this section, we introduce the deep learning-based framework, named spectral-feature networks, to extract deep spectral features.

In early research attempts, the hyperspectral pixel vector was intuitively fed into fully connected networks due to the requirement of the vector-based input in the network layers. Specifically, the original spectral vector was directly used to train the SAE or DBN in an unsupervised way [35], [37]. Subsequently, some improved methods were also developed for HSI classification. For example, Liu *et al.* [53] proposed an effective classification framework based on deep learning and active learning, where DBN was used to extract deep spectral features and an active learning algorithm was employed to

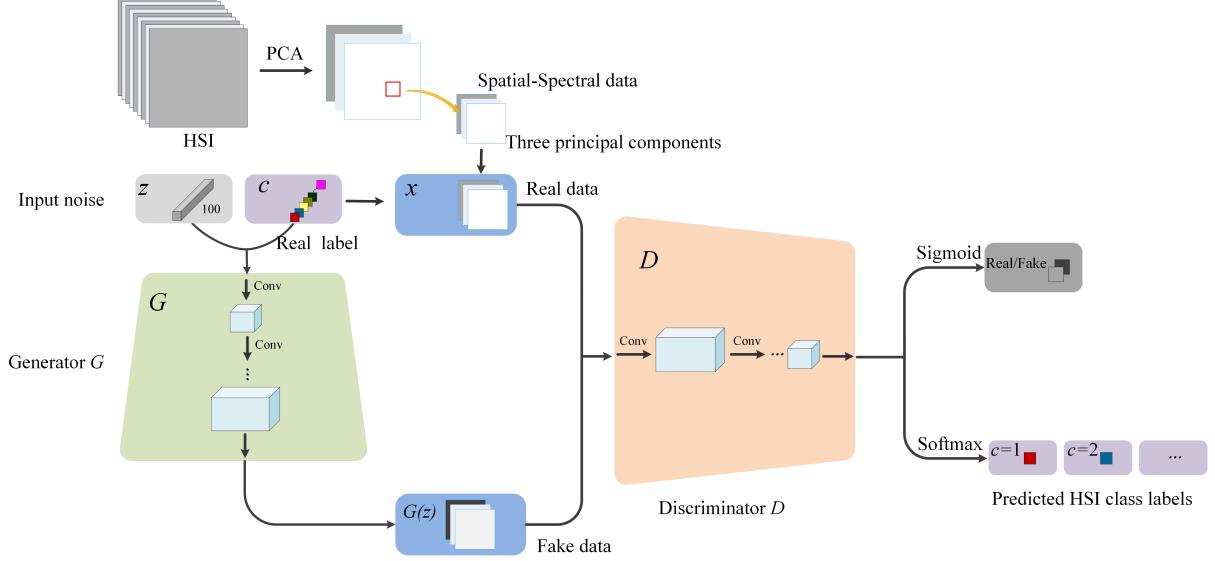


Fig. 8. Illustration of the GAN-based HSI classification approach developed in [46].

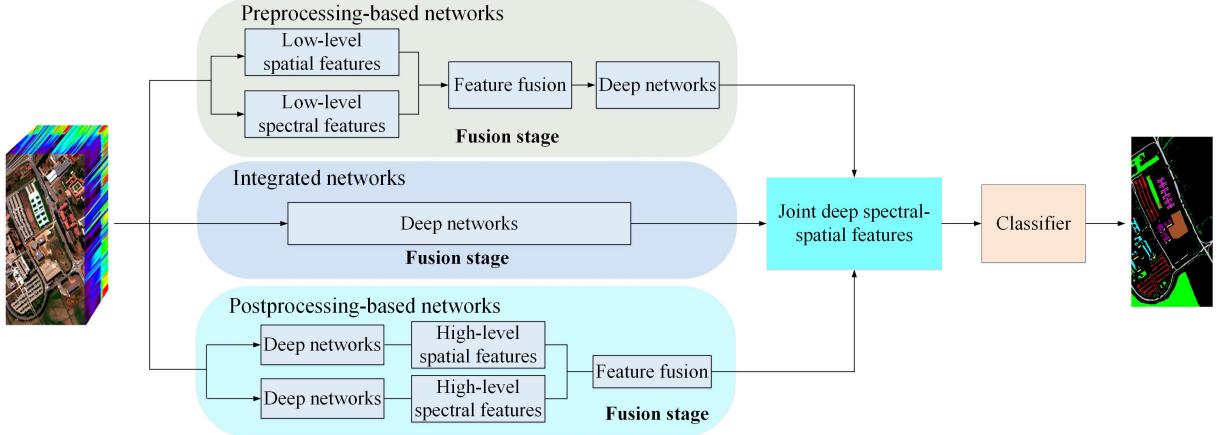


Fig. 9. Paradigms of spectral-spatial-feature networks, which can be further divided into three categories based on the stage of feature fusion, i.e., preprocessing-based networks, integrated networks, and postprocessing based networks.

iteratively select good-quality labeled samples as training samples. At the same time, Zhong *et al.* [54] developed an improved DBN model, named diversified DBN, to regularize the pretraining and fine-tuning procedures of DBN, which significantly improved the performance of DBN in terms of classification accuracies. In addition, 1-D CNN [55]–[58], 1-D GAN [46], [59], and RNN [44], [58], [60] were also used to extract spectral features for HSI classification. In [61], Li *et al.* used pixel-pair features extracted by CNN to explore correlation between hyperspectral pixels, where the convolution operation was mainly executed in the spectral domain. Furthermore, in [62], [63], the training of a deep network with the dictionary learning was reformulated. In such works, the multiple levels of dictionaries were learned in a robust fashion. This novel perspective obtained a better classification result than general deep networks.

B. Spatial-Feature Networks

Previous researches on HSI classification have proven that classification accuracies can be further improved by incorporating spatial features into classifiers [17], [64]. In this section, we will discuss the spatial-feature networks that exploit deep networks to extract spatial features of HSIs. For accurate HSI classification, the learned spatial features are subsequently fused with spectral features extracted by other feature extraction techniques.

In [56], [57], [65]–[67], PCA was first performed on the whole hyperspectral data to reduce the dimensionality of the original space, and then, the spatial information contained in the neighborhood region of the input hyperspectral pixel was exploited by a 2-D CNN. The above methods combined the PCA and CNN, which not only extracted the discriminative spatial features but also reduced the computational cost. More advanced, Liang *et al.* [68] introduced a sparse representation technique to encode the deep spatial features

extracted by a CNN into low-dimensional sparse features, which improved the ability of feature representation and the final classification accuracies. Chen *et al.* adopted the off-the-shell CNNs (e.g., AlexNet [31] and GoogLeNet [32]) to extract deep spatial feature [69]. In addition, Zhao *et al.* [70] proposed a spectral-spatial feature-based classification (SSFC) framework for HSI classification. In that framework, the spectral and spatial features were extracted by the balanced local discriminant embedding (BLDE) and a CNN, respectively. Then, the spectral and spatial features were fused to train a multiple-feature-based classifier. Instead of extracting spatial-spectral information by incorporating the pixels within a small spatial neighborhood into a classifier, a novel HSI classification framework, named the deep multiscale spatial-spectral feature extraction algorithm, was proposed in [71]. In more detail, the well pretrained FCN-8 [72] was first used to explore deep multiscale spatial structural information. Then, the weighted fusion mechanism was adopted to fuse the original spectral features and deep multiscale spatial features. Finally, the fused features were fed into a classifier to perform the classification operation.

C. Spectral-Spatial-Feature Networks

Instead of designing a deep network to extract either spectral features or spatial features, this kind of networks (i.e., spectral-spatial-feature networks) can extract joint deep spectral-spatial features for HSI classification. The joint deep spectral-spatial features are mainly obtained by the following three ways: 1) mapping the low-level spectral-spatial features to high-level spectral-spatial features via deep networks; 2) directly extracting deep features from original data or several principal components of the original data; 3) fusing two separate deep features (i.e., deep spectral features and deep spatial features). Based on this observation, the spectral-spatial-feature networks can be further divided into three categories: preprocessing-based networks, integrated networks, and postprocessing-based networks. Fig. 9 depicts the paradigm of three kinds of networks, which is actually divided in terms of the processing stage, where the spectral information and the spatial information are fused. Here, the three networks are discussed in the following parts.

1) *Preprocessing-based networks:* In this category, the spectral-spatial features are fused before feeding them into the subsequent deep network. In general, the whole classification process can be divided into three phases: 1) low-level spectral-spatial feature fusion; 2) high-level spectral-spatial feature extraction using deep networks; 3) joint deep spectral-spatial feature-based classification with simple classifiers (e.g., SVM, ELM, or multinomial logistic regression [36]). As mentioned above, the fully connected networks (e.g., DBN, SAE, and their variants) only deal with 1-D inputs. To extract joint spectral-spatial features, a common idea is to flatten the spatial neighboring region into a 1-D vector, and then, the obtained spatial vector and the original spectral vector are stacked and fed into fully connected networks. For those works, readers can refer to [35], [37], [73], [74]. In [75]–[77], a new spectral vector was first computed by averaging all spectral pixels within

a spatial neighboring. Finally, the averaged spectral vector which actually includes the spatial contextual information was processed by the following deep network. Furthermore, instead of directly exploiting the spatial information within a neighboring window, some different filtering methods (e.g., Gobar filtering [78], [79], attribute filtering [80], extinction filtering [81], and rolling guidance filtering [82]) were introduced to process the original hyperspectral data aiming to extract more effective spatial features. These filter-based works combine the deep learning techniques with other spatial-feature extraction methods and deliver more accurate classification results.

2) *Integrated networks:* Instead of separately acquiring spectral and spatial features and then processing them together, the joint deep spectral-spatial features were directly extracted by the original data via 2-D CNN [58], [83]–[91]. Actually, the hyperspectral data can be typically represented in the format of a 3-D cube. Thus, 3-D convolution in spectral and spatial dimensions can naturally offer a more effective method for simultaneously extracting the spectral-spatial features within such images. Based on this fact, 3-D CNN was used to effectively extract deep spectral-spatial-combined features for accurate HSI classification without relying on any preprocessing or postprocessing techniques [56]–[58], [92]–[95]. Apart from CNN, some powerful deep models were used to classify HSIs. Specifically, the fully convolutional network (FCN) [72], a very successful network in the semantic segmentation field, was utilized to reconstruct hyperspectral data, which can learn the deep features of HSIs in a supervised way [96] or an unsupervised way [97]. Besides, references [87], [93], [97]–[99] introduced residual learning [100] to build very deep and wide networks with the purpose of extracting more discriminative features for HSI classification. Fig. 10 shows a residual block, where the output is the sum of input and convoluted value of input. In [46], [101], a 3-D GAN [45] was utilized as a spectral-spatial classifier. In the GAN-based HSI classification framework, a CNN was first designed to discriminate the inputs, which is called the discriminative model. Then, another CNN was used to generate the so-called fake inputs as the generative model. Paoletti *et al.* refined the CapsNets [102] as the spectral-spatial capsules to extract spectral-spatial features of HSIs [103]. Furthermore, researchers were also focused on building hybrid deep networks to achieve the accurate classification of HSIs, which can make full use of different deep models. For instance, Kemker *et al.* [104] developed an unsupervised feature extraction framework to learn generalizable features from unlabeled hyperspectral pixels via a three-layer stacked convolutional autoencoder. Wu *et al.* [105] proposed a novel deep convolutional recurrent neural networks (CRNN), taking advantages of CNN and RNN models, to classify HSIs by using pseudo labels. In such a network, the convolutional layers were used to extract middle-level locally invariant features from the input hyperspectral sequence, and the recurrent layers can extract contextual information from the middle-level feature sequences generated by the previous convolutional layers. In addition, a spectral-spatial cascaded RNN model was proposed to extract spectral-spatial features for HSI classification. [106].

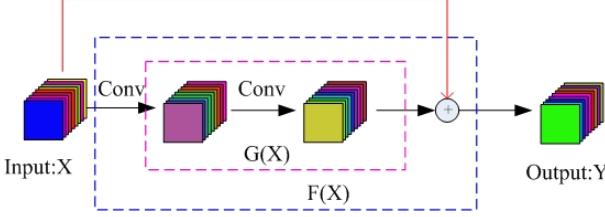


Fig. 10. Illustration of a residual block.

3) *Postprocessing-based networks*: In this category, the whole classification procedure includes the following steps: 1) deep spectral features and deep spatial features are obtained via two deep networks; 2) the two kinds of features are fused in a fully connected layer to generate the joint deep spectral-spatial features; 3) joint deep spectral-spatial feature-based HSI classification with the subsequent classifier (e.g., SVM, ELM, or multinomial logistic regression [36]). It is worth mentioning that the two deep networks used to separately extract deep spectral and spatial features may share the same weights or may be totally different. For instance, in [107], [108], a deep CNN with two-branch architecture was proposed to extract the joint spectral-spatial features from HSIs. In such works, a 1-D CNN branch and a 2-D CNN branch were used to extract spectral features and spatial features, respectively. After that, the learned spectral features and spatial features were concatenated and fed to fully connected layers to extract joint spectral-spatial features for classification. In [109], Hao *et al.* developed a novel two-stream architecture for HSI classification, where one stream employed a stacked denoising autoencoder (SDAE) to encode the spectral values of each input pixel, and the other stream exploited a deep CNN to learn spatial features by processing the corresponding image patch. Finally, the prediction probabilities from two streams were fused by adaptive class-specific weights, which can be obtained by a fully connected layer. In addition, references [110], [111] used two similar network architectures (e.g., two principal component analysis-based networks [110], two stacked sparse AE [111]) to extract the spectral features and spatial features, respectively. Subsequently, these two kinds of features were fused in a fully connected layer and an SVM was further trained for classification. Furthermore, Santara *et al.* [112] adopted multiple CNNs to simultaneously process multiple spectral sets, where multiple CNNs shared the same network parameters. In fact, each CNN aimed to extract the corresponding spatial feature of several neighboring spectral bands, and the spectral-spatial features can be obtained by fusing these individual spatial features in a fully connected layer.

IV. STRATEGIES FOR LIMITED AVAILABLE SAMPLES

As a matter of fact, training a deep network requires a large number of training samples to learn the network parameters. However, in the remote sensing field, there is usually only a small amount of labeled data available since the collection of such labeled data is either expensive or time-demanding. This issue, also known as the imbalance between lots of

weights and limited availability of training samples, may result in poor classification performance. Recently, some effective methods have been proposed to cope with the problem to some extent. In this section, we include some strategies to improve deep learning-based HSI classification under the condition of limited available samples.

A. Data Augmentation

Data augmentation is considered as an intuitive way to effectively solve the above problem. It tries to create new training samples from given known samples. By investigating the literature, we pointed out two main strategies to generate the additional virtual samples: 1) transformation-based sample generation; 2) mixture-based sample generation. In the following part, we will discuss the two methods in details.

1) *Transformation-based sample generation*: Due to the complex situation of lighting in HSIs, objects of the same class in different locations may be affected by different radiations. Based on this fact, the virtual samples can be generated by transforming the current known samples. This method of augmenting data is widely used in [56], [80], [85], [98], [108]. Specifically, let \mathbf{x}_i be a known training sample, the new virtual sample \mathbf{y} can be obtained by

$$\mathbf{y} = f(\mathbf{x}_i) + \gamma \mathbf{n} \quad (12)$$

where f is a transforming function (e.g., performing the rotating, flipping, or mirroring operation), γ controls the weight of the random Gaussian noise \mathbf{n} , which may be produced via the interaction of neighboring pixels or imaging error. Finally, the newly generated virtual sample \mathbf{y} has the same class with \mathbf{x}_i and can be used to train a deep network.

2) *Mixture-based sample generation*: In general, the objects of the same class usually show similar spectral characteristics in a certain range. This phenomenon makes it possible to generate a virtual sample from two given samples of the same class. In [56], [78], the virtual sample \mathbf{y} is the linear combination of two training samples \mathbf{x}_i and \mathbf{x}_j , which can be represented by the following formulation

$$\mathbf{y} = \alpha_{ij} \mathbf{x}_i + (1 - \alpha_{ij}) \mathbf{x}_j \quad (13)$$

where α_{ij} represents the affinity between two training samples (i.e., \mathbf{x}_i and \mathbf{x}_j) coming from the same class. α_{ij} is usually defined as follows:

$$\alpha_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2). \quad (14)$$

Although there are many other techniques to generate the virtual samples, the above two methods are simple yet effective ways.

B. Transfer Learning

Transfer learning [113] is regarded as a set of techniques that introduces the useful information learned from the source data to the target data, which can significantly decrease the demand on training samples. In recent years, transfer learning has been successfully applied in many fields, especially in the remote sensing field [114], [115] where the available

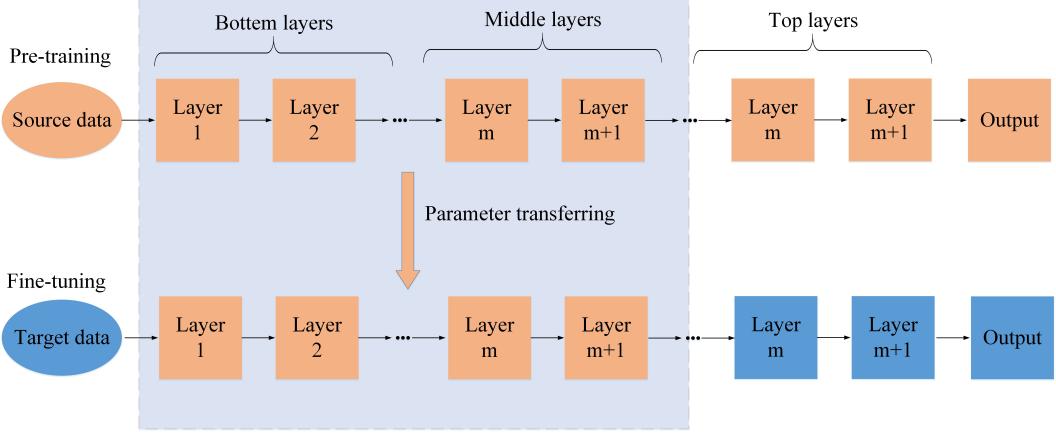


Fig. 11. The general framework of transfer learning based deep network training.

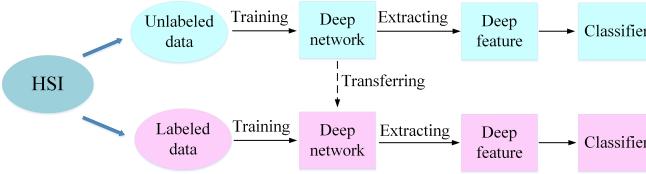


Fig. 12. Illustration of unsupervised/semisupervised feature learning associated with deep learning. The top part represents the unsupervised deep feature learning, and the bottom part illustrates the semisupervised way for feature learning.

training samples are not sufficient. In [77], [104], [105], [107], [108], transfer learning was employed to set the initial values of network parameters copied from other trained deep networks, which provides better classification performance compared with the random initialization-based methods. Fig. 11 shows the general framework of transfer learning for deep networks. Considering the fact that the low and middle layers usually capture generic features of input images which has high generalization for other images, it is feasible to directly transfer the network parameters of low and middle layers to a new network which has the same architecture as the previous network. Remarkably, the parameters of top layers are still initialized in a random way in order to deal with a specific task. Once the network parameters are transferred, the subsequent classification can be further divided into unsupervised and supervised methods. The former directly uses the deep features extracted from the transferred network to train the classifier. For the latter method, the network is further fine-tuned by using a small amount of training samples of the target data. In summary, by making full use of existing data sets, transfer learning can effectively solve the problem of degradation of network performance when the available training samples are limited.

C. Unsupervised/Semi-supervised Feature Learning

Though supervised feature learning has gained great breakthrough in the HSI classification field, there is still an urgent need to learn HSI features in an unsupervised or semisup-

ervised way. The main purpose of unsupervised/semisupervised feature learning is to extract useful features from an amount of unlabeled data. Recently, more and more research works [47], [97], [104], [105], [111], [116], [117] focus on designing a robust and effective unsupervised/semisupervised feature learning framework based on deep learning to classify HSIs. Fig. 12 illustrates unsupervised/semisupervised deep feature learning. From this figure, the top flowchart only uses unlabeled data to extract the informative features of HSIs, which is actually regarded as an unsupervised feature learning way. The deep network is elaborately designed as the encoder-decoder paradigm to learn networks without using label information. Furthermore, the classification performance can be improved via transferring the trained network and fine-tuning on the labeled data set, which is shown at the bottom of the flowchart in Fig. 12.

For example, research works [35], [37], [53], [54], [73], [104], [111], [116] utilized a fully connected network to classify HSIs, where the training of the network was divided into pre-training in an unsupervised manner and fine-tuning with labeled data. In the pre-training stage, the sample from unlabeled data set was first mapped into an intermediate feature. Then, the intermediate feature was subsequently reconstructed via a decoding operation. At last, the parameters of each layer can be learned by minimizing the error between original sample and reconstructed sample. However, the above greedy layer-wise training fashion may not be efficient when the network is deep. More advanced, an unsupervised end-to-end training framework was proposed in [97], where the convolutional and deconvolutional networks were regarded as encoder and decoder, respectively. In addition, the GAN [45] was also adopted to construct semisupervised feature learning framework for HSI classification [46], [59]. In such works, the generator created fake hyperspectral samples that were similar to the real data to train the GAN. Besides, Wu *et al.* used abundant unlabeled data with their pseudo labels obtained by a non-parametric Bayesian clustering algorithm to pre-train a CRNN for HSI classification [105]. The semisupervised learning mechanism can extract more representative spectral-spatial features and provide satisfactory classification result.

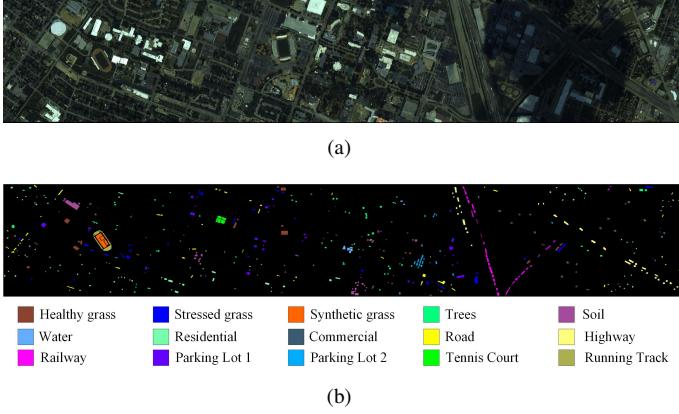


Fig. 13. The Houston data set. (a) The three-band false color composite. (b) Ground reference data and color code.

D. Network Optimization

The main purpose of network optimization is to further improve the network performance through adopting more efficient modules or functions. For instance, the current mainstream deep networks use ReLU [31] as nonlinear activation function followed by the batch normalization operation [118], which can effectively alleviate the overfitting during the training phase. In addition, research works [87], [88], [93], [97], [98] introduce residual learning [100] to build a very deep network with the purpose of extracting more discriminative features for HSI classification. By optimizing several convolutional layers as the identity mapping, residual learning can ease the training process. Fig. 10 shows the mechanism of residual learning. Let \mathbf{X} be the input of the first layer and $F(\mathbf{X})$ be the original underlying function to be learned by two stacked convolutional layers. The residual learning actually attempts to optimize the two convolutional layers as identity mapping, which is achieved by using skip connections (the red line in Fig. 10). By introducing residual function $G(\mathbf{X}) = F(\mathbf{X}) - \mathbf{X}$, the objective function from the original $F(\mathbf{X}) = \mathbf{X}$ equivalently converts to $G(\mathbf{X}) = 0$. As described above, $F(\mathbf{X})$ can be written by the following formulation:

$$F(\mathbf{X}) = G(\mathbf{X}) + \mathbf{X}. \quad (15)$$

As shown in Fig. 10, $G(\mathbf{X})$ can be computed by two convolutions with \mathbf{X} :

$$G(\mathbf{X}) = f(f(\mathbf{X} * \mathbf{W}_1 + \mathbf{b}_1) * \mathbf{W}_2 + \mathbf{b}_2) \quad (16)$$

where \mathbf{W}_1 and \mathbf{W}_2 are convolutional kernels, \mathbf{b}_1 and \mathbf{b}_2 are trainable *bias* parameters, and f refers to the ReLU [31] function. By stacking multiple residual blocks, the extracted features become more and more discriminative and the accuracy is gained from considerably increased depth. More details about theoretical and experimental proofs of the DRN can refer to [100].

Besides, other network optimization strategies are also exploited to boost the representation ability of features extracted by deep networks. For example, Zhong *et al.* [54] improved the performance of DBN via regularizing the pre-training and

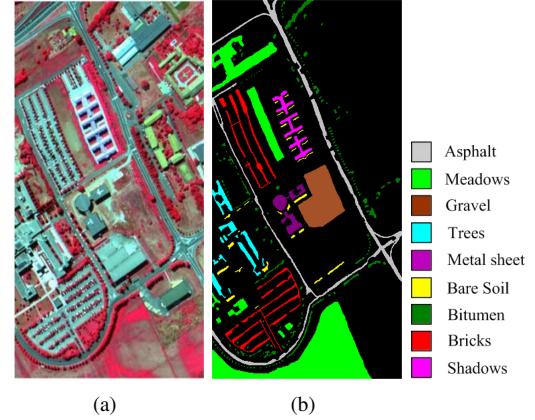


Fig. 14. The University of Pavia data set. (a) The three-band false color composite. (b) Ground reference data and color code.

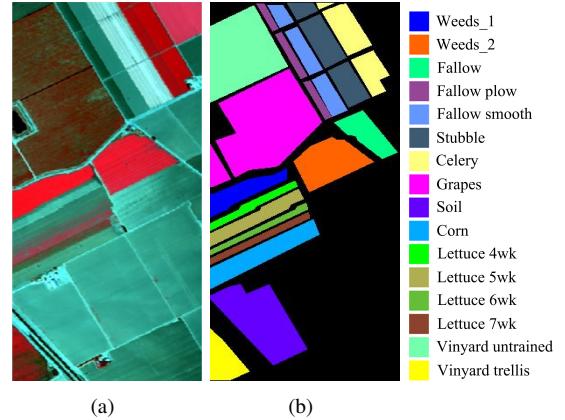


Fig. 15. The Salinas data set. (a) The three-band false color composite. (b) Ground reference data and color code.

fine-tuning procedure, which delivers a better classification accuracy than usual DBN models. In [75], the label consistency constraint was enforced into the training procedure of SAE. Moreover, the correlation between samples was considered in networks investigated in [61], [89].

In general, network optimization is still a challenging problem. In the future works, the unique characteristics of HSI should be considered as much as possible when designing an improved network for HSI classification.

V. EXPERIMENTS

In this section, we mainly conduct a comprehensive set of experiments from four aspects. Firstly, a series of experiments are designed to demonstrate the advantages of deep learning on HSI classification over traditional methods. Secondly, the classification performance of several recent state-of-the-art deep learning approaches is systematically compared. Thirdly, we visualize the learned deep features and network weights to further explore the “black box”. Finally, the effectiveness of strategies included in Section IV is further analyzed. To complete our experiments, three benchmark HSIs are used, i.e., the Houston, University of Pavia, and Salinas images. The three images are introduced in the following subsection.

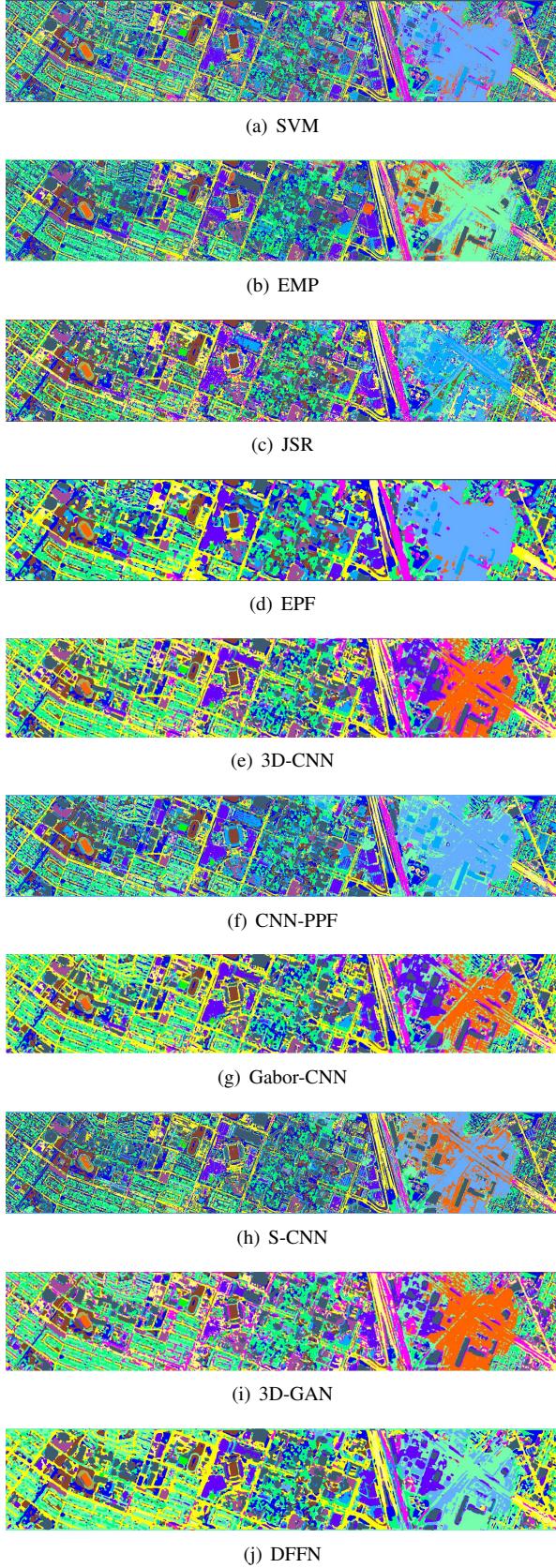


Fig. 16. Classification maps for the Houston data set obtained by (a) SVM [8], (b) EMP [19], (c) JSR [25], (d) EPF [24], (e) 3D-CNN [56], (f) CNN-PPF [61], (g) Gabor-CNN [79], (h) S-CNN [89], (i) 3D-GAN [46], and (j) DFFN [87].

TABLE I
NUMBER OF TRAINING AND TEST SAMPLES USED FOR THE HOUSTON DATA

Class name	Training samples	Test samples	Total samples
Healthy grass	198	1053	1251
Stressed grass	190	1064	1254
Synthetic grass	192	505	697
Trees	188	1056	1244
Soil	186	1056	1242
Water	182	143	325
Residential	196	1072	1268
Commercial	191	1053	1244
Road	193	1059	1252
Highway	191	1036	1227
Railway	181	1054	1235
Parking Lot 1	192	1042	1234
Parking Lot 2	184	285	469
Tennis Court	181	247	428
Running Track	187	473	660
Total	2832	12179	15011

TABLE II
NUMBER OF TRAINING AND TEST SAMPLES USED FOR THE UNIVERSITY OF PAVIA DATA SET

Class name	Training samples	Test samples	Total samples
Asphalt	200	6431	6631
Meadows	200	18449	18649
Gravel	200	1899	2099
Trees	200	2864	3064
Painted metal sheets	200	1145	1347
Bare Soil	200	4829	5029
Bitumen	200	1130	1330
Self-Blocking Bricks	200	3482	3682
Shadows	200	747	947
Total	1800	40976	42776

TABLE III
NUMBER OF TRAINING AND TEST SAMPLES USED FOR THE SALINAS DATA SET

Class name	Training samples	Test samples	Total samples
Brocoli_green_weeds_1	200	1809	2009
Brocoli_green_weeds_2	200	3526	3726
Fallow	200	1776	1976
Fallow_rough_plow	200	1194	1394
Fallow_smooth	200	2478	2678
Stubble	200	3759	3959
Celery	200	3379	3579
Grapes_untrained	200	11071	11271
Soil_vinyard Develop	200	6003	6203
Corn_senesced_green_weeds	200	3078	3278
Lettuce_romaine_4wk	200	868	1068
Lettuce_romaine_5wk	200	1727	1927
Lettuce_romaine_6wk	200	716	916
Lettuce_romaine_7wk	200	870	1070
Vinyard_untrained	200	7068	7268
Vinyard_vertical_trellis	200	1607	1807
Total	3200	50929	54129

A. Experimental Data Sets

The Houston data was distributed for the 2013 IEEE Geoscience and Remote Sensing Society (GRSS) data fusion contest. This scene was captured in 2012 by an airborne sensor over the area of University of Houston campus and the neighboring urban area. The size of the data is 349×1905 pixels with a spatial resolution of 2.5 m. This HSI consists of 144 spectral bands with wavelength ranging from 0.38 to 1.05 μm and includes 15 classes. Fig. 13 shows the false color composite of the Houston data set and the corresponding ground reference data, respectively. Table I shows the information on the number of training and test samples for the different classes of interests.

The University of Pavia data set, which captures an urban area surrounding the University of Pavia, Italy, was collected by the ROSIS-03 sensor in Northern Italy in 2001. This scene is of size $610 \times 340 \times 115$ with a spatial resolution of 1.3 m per pixel and spectral coverage ranging from 0.43 to 0.86 μm . This image includes 9 classes of interest and has 103 spectral bands after removing 12 very noisy bands. Fig. 14 shows the false color composite of the University of Pavia image and the corresponding ground reference map, respectively. Table II shows information on the number of training and test samples for the different classes of interests.

The Salinas data set was captured by the AVIRIS sensor over Salinas Valley, California. This image comprises of 512×217 pixels with a spatial resolution of 3.7 m and 204 bands after removing 20 water absorption bands. The available ground reference map covers 16 classes of interest. The false color composite of Salinas image and the corresponding ground reference data are shown in Fig. 15. Table III shows information on the number of training and test samples for the different classes of interests.

B. Compared Methods

In this review, we have investigated several recent state-of-the-art deep learning-based approaches, including 3D-CNN [56], Gabor-CNN [79], CNN with pixel-pair features (CNN-PPF) [61], siamese CNN (S-CNN) [89], 3D-GAN [46], and the deep feature fusion network (DFFN) [87], for HSI classification. Specifically, the 3D-CNN exploits 3-D convolutional filters to directly extract spectral-spatial features from the original hyperspectral cube. However, the network architecture adopted in 3D-CNN is relatively simple and the correlation between different layers is neglected. The DFFN adopts the DRN [100], which can be considered a more powerful network, to extract more discriminative features. In addition, the features from different-level layers are further fused to explore the correlation between layers. The main drawback of DFFN is that the optimal feature fusion mechanism depends on a hand-crafted setting with abundant experiments. In the Gabor-CNN, the Gabor filtering is first utilized as a preprocessing technique to extract spatial features of HSIs. Then, the filtered features are fed into a simple CNN-based classifier. Instead of considering a pixel-wised semantic information, CNN-PPF and S-CNN focus on exploring the correlation between samples. In more detail, CNN-PPF extracts the pixel-pair

features via CNN. However, the convolutional operation is mainly conducted in the spectral domain, and thus, the spatial information is not considered for CNN-PPF. By contrast, S-CNN adopts a two-branch CNN to simultaneously extract spectral-spatial features of HSIs. But, in such a method, the computational cost may be huge due to the high dimension vector in the Euclidean space. In addition, 3D-GAN utilizes adversarial training to improve the generalization capability of the discriminative CNN, which is very useful when the number of training samples is limited.

Apart from the above deep learning-based methods, some classical classification techniques, including the SVM [8], EMP [19], joint spare representation (JSR) [25], and edge-preserving filtering (EPF) [24] are also considered for comparison. As the spectral feature-based method, the SVM was implemented in the library for support vector machines (LIB-SVM) library [119], where the Gaussian kernel with fivefold cross validation was adopted for this classifier. The other three methods are all spectral-spatial feature-based methods. Specifically, for the EMP, the morphological profiles were constructed with the first three principal components. Furthermore, a circular structural element, a step size increment of two, and four openings and closings were performed for each principal component. For the JSR, the spatial information within a fixed-size local region was utilized by the joint sparse regularization. The EPF method was implemented by using the code which is available on Dr. Xudong Kang's homepage².

Among these investigated methods, the SVM and CNN-PPF have only utilized spectral features during the classification. In contrast, the rest of methods, including the EMP, JSR, EPF, 3D-CNN, Gabor-CNN, S-CNN, 3D-GAN, and DFFN, are belong to classification methods based on spectral-spatial features.

C. Classification Results

The first experiment was performed on the Houston data set. In this experiment, the training samples were given according to the 2013 GRSS data fusion contest. The amount of training and test samples per class is shown in Table I. Fig. 16 shows the classification maps obtained by different methods. From this figure, we can see that the classification maps obtained by the SVM and JSR methods are not very satisfactory since some noisy estimations are still visible. By contrast, other methods perform much better in removing "noisy pixels" and deliver a smoother appearance in their classification results. By comparing two filtering-based methods, i.e., EPF and Gabor-CNN, we can see that the classification map of EPF seems to be over-smoothing, but the Gabor-CNN preserves more details in edges. Apart from visual comparison, Table IV gives quantitative results of various methods on the image, where three metrics, i.e., overall accuracy (OA), average accuracy (AA), and Kappa coefficient, are adopted to evaluate the classification performance. All the classification accuracy values reported in Table IV are the average results over ten experiments. As can be seen, except for 3D-GAN, the deep learning-based methods obtain satisfactory classification

²<http://xudongkang.weebly.com>

TABLE IV

THE CLASSIFICATION ACCURACIES (IN PERCENTAGES) OBTAINED BY SVM [8], EMP [19], EPF [24], JSR [25], 3D-CNN [56], GABOR-CNN [79], CNN-PPF [61], S-CNN [89], 3D-GAN [46], AND DFFN [87] ON THE HOUSTON IMAGE. THE BEST ACCURACIES ARE MARKED IN BOLD.

classes	SVM	EMP	JSR	EPF	3D-CNN	CNN-PPF	Gabor-CNN	S-CNN	3D-GAN	DFFN
1	82.05	81.32	94.49	82.72	82.16	82.66	82.30	83.00	81.58	77.41
2	80.55	82.65	83.46	82.71	85.19	84.02	84.24	83.27	79.74	81.39
3	100	98.67	98.42	99.78	90.69	100	95.61	98.66	97.42	94.59
4	92.52	84.26	96.59	93.09	93.35	92.87	92.39	93.50	93.36	87.82
5	98.11	99.50	99.53	100	99.61	99.91	99.80	96.91	99.71	96.05
6	95.10	95.10	99.30	95.80	97.89	97.90	96.47	94.12	95.08	96.15
7	75.00	92.35	72.85	80.88	91.25	84.70	84.58	79.95	89.90	80.25
8	40.17	53.41	43.02	44.54	75.92	46.59	73.09	68.19	70.52	77.78
9	74.88	79.51	71.01	84.32	83.20	78.41	78.14	76.39	54.89	84.66
10	51.64	53.57	47.20	49.61	48.80	53.15	58.01	48.10	49.89	64.63
11	78.37	66.22	76.66	83.59	69.21	78.91	73.35	74.64	77.36	88.61
12	68.40	84.44	68.78	75.12	85.55	89.24	86.74	85.49	60.46	98.57
13	69.47	68.41	43.16	60.70	90.85	79.65	91.16	88.58	81.71	83.09
14	100	100	88.66	100	90.66	100	100	99.19	95.14	99.72
15	98.10	99.58	99.79	99.48	76.83	98.59	77.73	96.40	65.35	81.90
OA	76.88	79.58	77.12	79.98	82.23	81.38	83.42	82.34	78.16	84.56
AA	80.29	80.78	78.86	82.21	84.08	84.44	84.17	84.43	79.98	86.18
Kappa	0.7513	0.7815	0.7528	0.7824	0.8013	0.7991	0.8114	0.8052	0.7616	0.8328

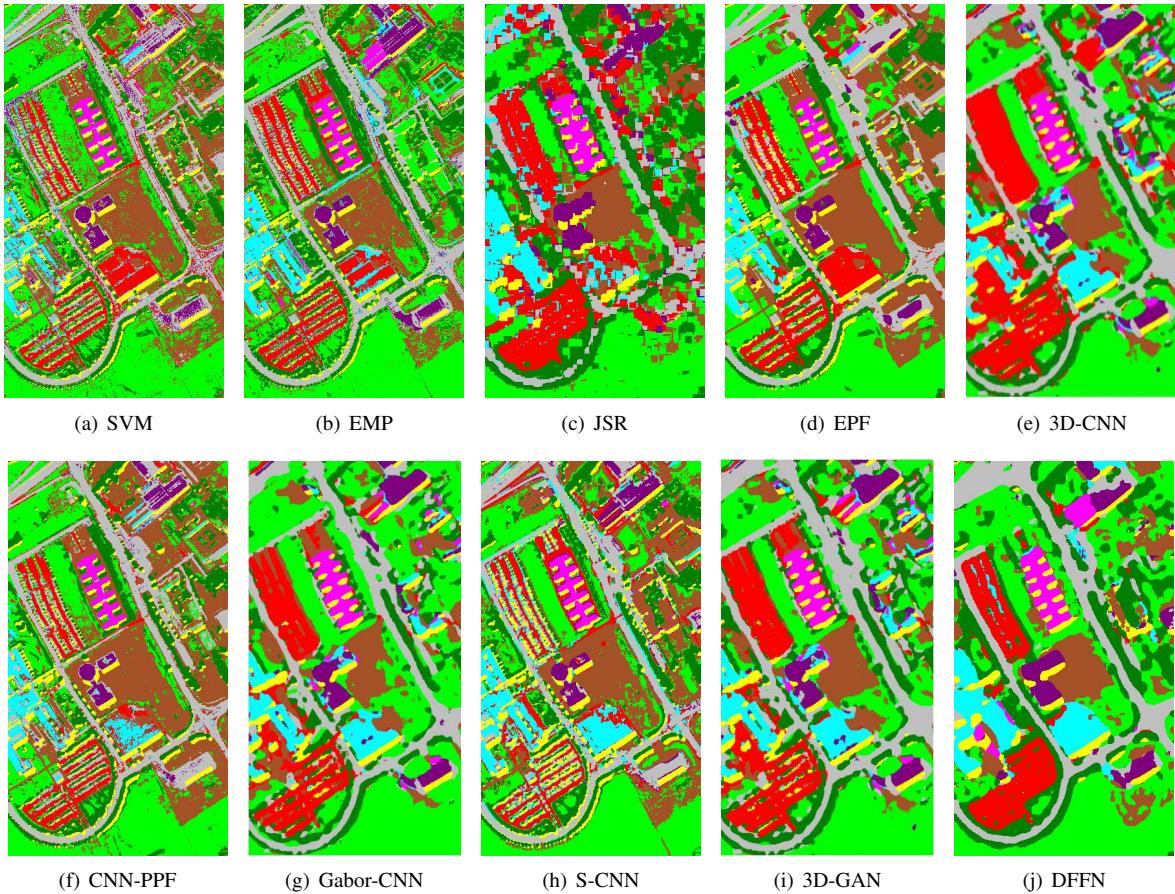


Fig. 17. Classification maps for the University of Pavia data set obtained by (a) SVM [8], (b) EMP [19], (c) JSR [25], (d) EPF [24], (e) 3D-CNN [56], (f) CNN-PPF [61], (g) Gabor-CNN [79], (h) S-CNN [89], (i) 3D-GAN [46], and (j) DFFN [87].

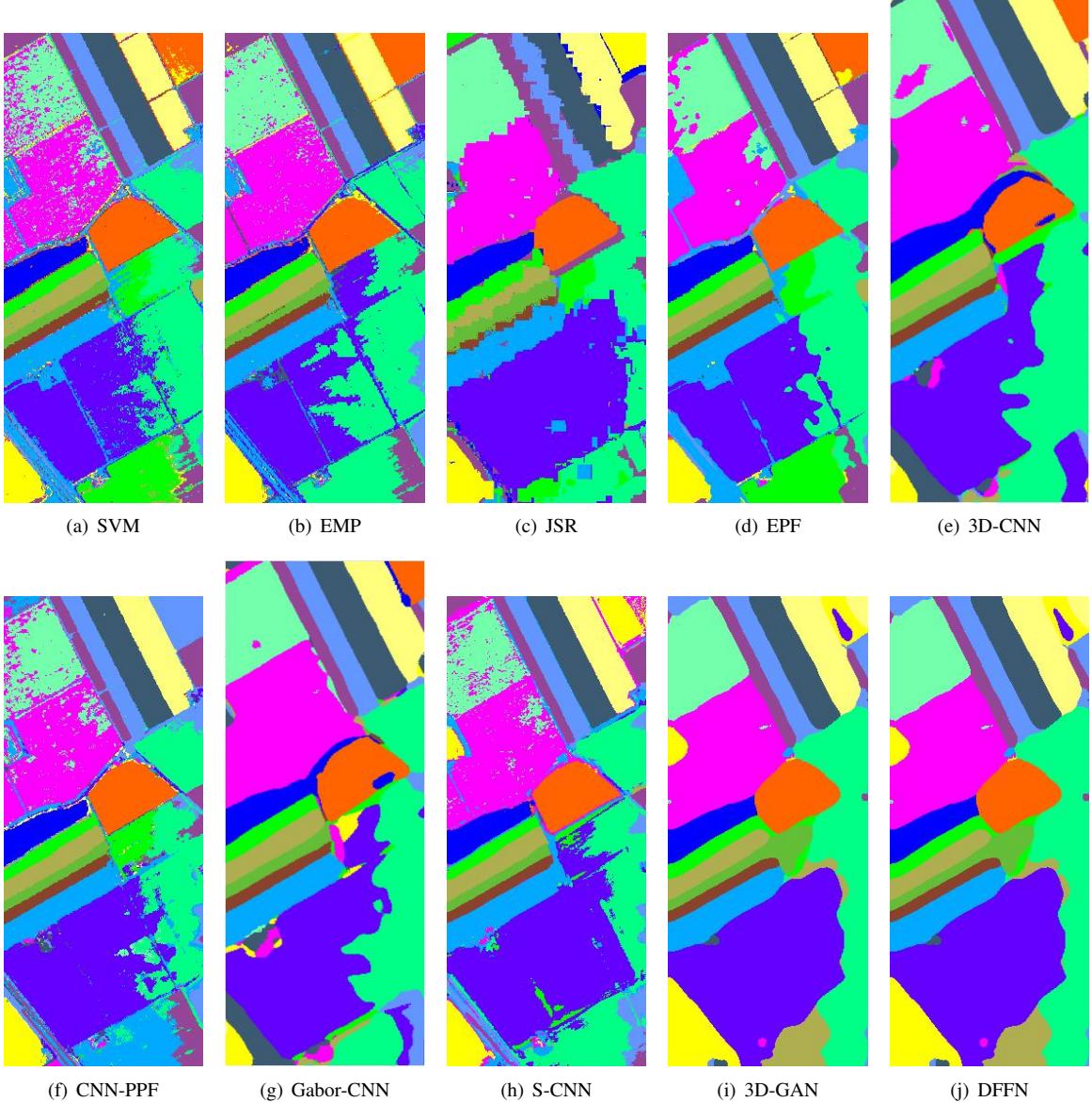


Fig. 18. Classification maps for the Salinas data set obtained by (a) SVM [8], (b) EMP [19], (c) JSR [25], (d) EPF [24], (e) 3D-CNN [56], (f) CNN-PPF [61], (g) Gabor-CNN [79], (h) S-CNN [89], (i) 3D-GAN [46], and (j) DFFN [87].

TABLE V
THE CLASSIFICATION ACCURACIES (IN PERCENTAGES) OBTAINED BY SVM [8], EMP [19], EPF [24], JSR [25], 3D-CNN [56], GABOR-CNN [79], CNN-PPF [61], S-CNN [89], 3D-GAN [46], AND DFFN [87] ON THE UNIVERSITY OF PAVIA IMAGE. THE BEST ACCURACIES ARE MARKED IN BOLD.

classes	SVM	EMP	JSR	EPF	3D-CNN	CNN-PPF	Gabor-CNN	S-CNN	3D-GAN	DFFN
1	85.49	98.70	83.80	96.08	99.03	97.23	99.53	95.47	99.18	99.53
2	92.12	98.92	96.31	98.24	98.11	95.27	98.21	98.71	98.86	97.71
3	85.77	94.75	98.63	95.25	88.56	95.13	89.74	97.32	94.94	99.89
4	96.41	96.66	93.89	98.27	83.51	96.89	93.02	97.72	90.15	97.88
5	98.60	98.87	98.38	100	99.49	99.99	99.42	100	99.49	99.48
6	92.52	84.52	99.54	98.16	95.33	98.55	98.77	97.67	98.56	99.69
7	93.79	87.97	98.50	98.79	96.31	96.56	98.82	98.36	92.74	100
8	86.56	98.72	96.51	96.81	97.58	94.43	94.12	95.65	97.18	98.59
9	97.97	99.87	75.88	97.91	96.25	99.39	97.91	100	98.51	99.61
OA	90.78	96.16	94.43	97.58	96.37	97.63	97.33	97.93	97.81	98.57
AA	92.14	95.08	92.57	97.72	94.82	97.04	96.62	97.88	96.65	99.16
Kappa	0.8813	0.9504	0.9349	0.9688	0.9502	0.969	0.9662	0.9743	0.9697	0.9808

TABLE VI
THE CLASSIFICATION ACCURACIES (IN PERCENTAGES) OBTAINED BY SVM [8], EMP [19], EPF [24], JSR [25], 3D-CNN [56], GABOR-CNN [79], CNN-PPF [61], S-CNN [89], 3D-GAN [46], AND DFFN [87] ON THE SALINAS IMAGE. THE BEST ACCURACIES ARE MARKED IN BOLD.

classes	SVM	EMP	JSR	EPF	3D-CNN	CNN-PPF	Gabor-CNN	S-CNN	3D-GAN	DFFN
1	99.61	98.77	100	100	99.94	99.84	100	100	75.35	99.99
2	99.69	98.27	99.83	98.98	85.45	99.77	88.06	97.75	98.49	99.94
3	99.56	99.74	99.92	99.99	100	98.11	99.25	98.88	100	100
4	99.41	98.98	94.77	100	99.77	99.57	100	100	99.76	100
5	98.72	97.33	83.31	99.49	99.96	98.54	98.88	99.93	100	99.11
6	99.77	99.46	92.34	100	100	99.92	100	89.48	99.97	99.95
7	99.52	99.11	95.92	99.93	99.60	99.96	98.94	99.30	99.45	99.43
8	76.74	93.79	96.97	92.55	99.31	89.11	99.48	98.69	98.30	99.56
9	99.37	99.73	99.99	99.99	99.97	99.69	97.27	99.34	99.95	100
10	95.13	98.38	92.42	99.69	99.41	97.78	99.21	100	99.79	99.79
11	99.32	98.69	93.86	100	100	99.33	100	99.93	100	99.48
12	99.70	96.52	83.94	99.53	100	100	100	99.89	100	99.84
13	99.15	99.37	89.25	99.99	100	99.67	100	88.62	100	99.96
14	98.38	98.77	70.06	99.94	100	98.75	99.79	88.72	100	99.95
15	75.56	92.75	96.74	92.71	90.36	89.99	94.31	90.62	99.52	99.45
16	99.23	97.68	99.81	99.96	85.93	99.07	93.38	99.96	90.25	99.96
OA	90.92	96.89	95.33	97.32	97.28	94.87	97.63	97.62	98.22	99.71
AA	96.18	97.95	94.78	98.88	97.48	98.07	98.04	96.94	97.75	99.78
Kappa	0.8985	0.9707	0.9307	0.9700	0.9695	0.9404	0.9734	0.9510	0.9793	0.9967

accuracies (e.g., all OAs are above 80%), which is overall higher than traditional methods, including the SVM, EMP, JSR and EPF. Among these deep learning-based methods, CNN-PPF which only considers spectral information of HSIs obtains the inferior classification results compared with other spectral-spatial features methods, e.g., 3D-CNN, Gabor-CNN, S-CNN, and DFFN. Moreover, the DFFN which combines the residual learning with feature fusion in the a deep CNN framework delivers the best classification accuracies in terms of OA, AA, and Kappa.

The second and third experiments were conducted on the University of Pavia and Salinas images. As mentioned above, 200 samples per class were randomly selected as the training samples and the rest of samples as the test samples (see Tables II and III). The selected training samples account for about 4% and 6% of the whole labeled reference data for the University of Pavia and Salinas images, respectively, which provide a challenging test set. Figs. 17 and 18 show the classification maps obtained by different methods. Tables V and VI present the quantitative results (averaged over ten experiments) of different methods.

From the above experimental results, the deep learning-based methods show great advantages over other traditional methods in terms of visual classification maps and quantitative results. For instance, considering that in comparison of several SVM-based classifiers, including SVM, EMP, EPF, and S-CNN, S-CNN performs the best on three hyperspectral data sets, which can be used to verify the effectiveness of deep features compared with hand-crafted features. In addition, for two filter-based methods, i.e., EPF and Gobor-CNN, the OA of Gabor-CNN is about 3.5% higher than that of EPF on the Houston image, which demonstrates that combining the filter technique with deep leaning can give good classification results. Although CNN-PPF only utilizes spectral information, it still delivers a better performance than what was obtained by the other four traditional-based methods on the Houston and University of Pavia images.

D. Deep Feature Visualization

In general, deep learning can be regarded as a black box in most applications, where the inside information of network is often unclear. Actually, exploring the inside features is very useful for analyzing the network performance and further designing the deep architecture. In this section, we used Salinas data set as an example to visualize the deep features. The weights of different convolutional kernels in the first convolutional layer are shown in Fig. 19. The Fig. 19 (a) is random initial weights, and Fig. 19 (b) demonstrates the learned weights. From this figure, we can see that the distribution of the weights for each filter become more regular and present evident textural features after training. For example, the intensities of the learned weights in the first kernel are low on the left side and high on the right side, *i.e.*, the weights reveal a trend of gradual decrease, which is different from the structure of randomly initialized weights significantly.

In addition, Fig. 20 shows the features after the convolutional layer, ReLU layer, and pooling layer. Fig. 21 shows the extracted features after three convolutional layers of Fallow and Fallow_roughg_plow samples. From Figs. 20 and 21, one can see that the earlier convolution layers can extract some simple features like texture and edge information, and those lower level features can be composed to become more complicated high-level features though the deeper convolution layers. The learning process is totally automatic, which makes CNNs more suitable for coping with varieties of situations.

E. Effectiveness Analysis of Strategies for Limited Samples

In this section, we will validate the effectiveness of some strategies proposed in Section IV for the problem of limited training samples. Specifically, we build a simple CNN and adopt three strategies, *i.e.*, data augmentation, transfer learning and residual learning, to improve the classification accuracy. These methods, *i.e.*, original CNN, CNN with data augmentation, CNN with transfer learning, and CNN with residual

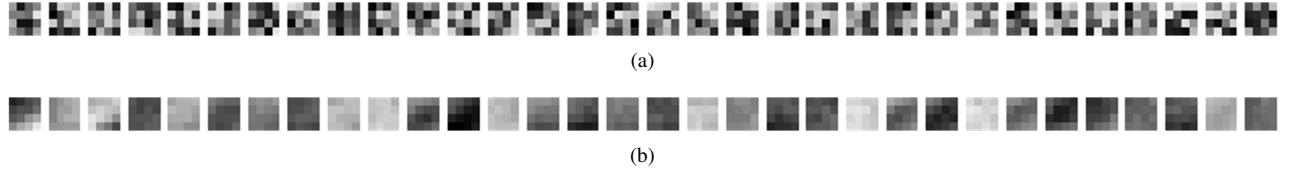


Fig. 19. Weights of the first convolutional layer. The size of each convolutional kernel is 4×4 . Each kernel of 32 kernels in the first convolutional layer is shown in the forms of a tiny image. (a) Randomly initialized weights of the first convolution layer on the Salinas data set. (b) Learned weights of the first convolution layer on the Salinas data set.

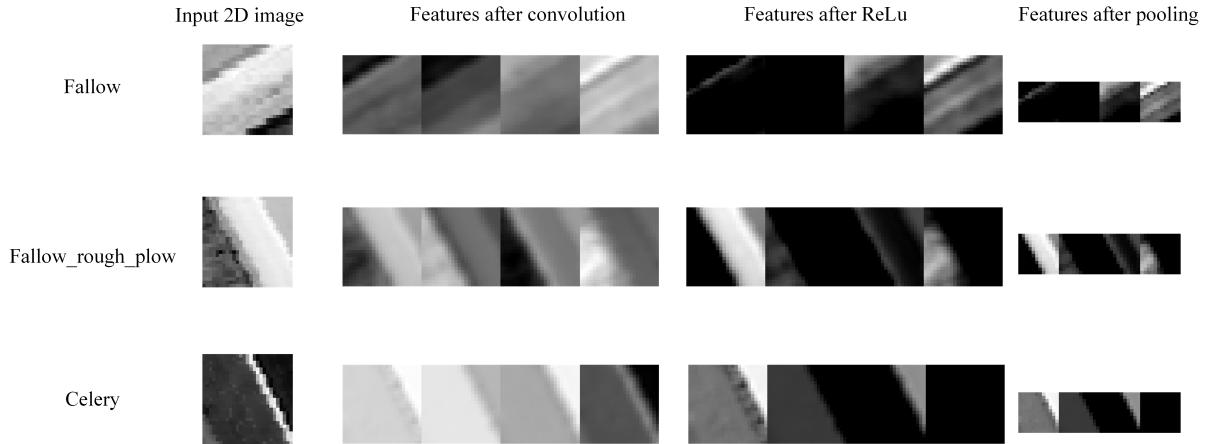


Fig. 20. Extracted features of the Salinas data set. Each row of the images represents one class. There are four columns in the figure. The first column represents the input images. The second column shows the four feature maps after the first convolution. The third column shows the four feature maps after the first ReLU operation. The last column shows the four features of the first pooling operation.

learning, are denoted as CNN-Original, CNN-DA, CNN-TL, and CNN-RL, respectively. The following experiments were conducted in the Salinas data set, where the training set consists of 5, 10, 15, 20, 25, and 30 labeled samples per class, respectively, and the remaining available sample were regarded as test data sets. All the used CNNs have the same architectures which consist of seven convolutional layers followed by the batch normalization operation, one global pooling layer, and two fully connected layers. The detailed structural information is presented in Table VII. For the CNN-DA, new samples were generated according to the equation (14). For the CNN-TL, the CNN was pretrained on the Indian Pines data set because the Indian Pines and Salinas images were collected by the same sensor, i.e., the Airborne Visible/Infrared Imaging Spectrometer sensor. For CNN-RL, the residual learning technique was incorporated into the CNN to optimize the network.

Fig. 22 shows the classification accuracies obtained by different methods in terms of OA values. From this figure, one can see that OA values of CNN-DA, CNN-TL, and CNN-RL are higher than OA values of CNN-Original for all training situations, which demonstrates that these strategies, indeed, improve the network performance to some extent when few training samples are available. The improvement almost approaches 1% by comparing CNN-TL with CNN-Original when the training set consists of 5 samples per class. In addition, CNN-RL performs better than other methods for most situations, which demonstrates that the residual learning

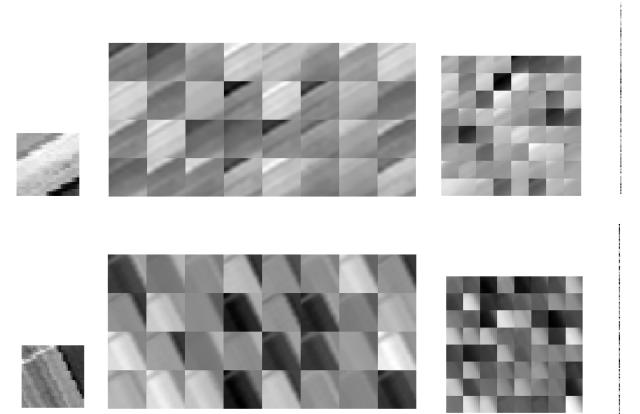


Fig. 21. Extracted features after three convolutional layers on Fallow and Fallow_rough_plow samples. The first column represents the 27×27 input images; the second column shows 32 feature maps after the first convolution layer, for which the size of each feature map is 24×24 ; the third column shows 64 feature maps after the second convolution layer, for which the size of each feature map is 8×8 ; and the number of feature maps after the third convolution layer shown in the last column is 128, and the size of each feature map is 1×1 .

is a very useful network optimization method for HSI classification.

VI. CONCLUSION

Recently, deep learning-based HSI classification has drawn a significant attention in the remote sensing field and obtained

TABLE VII
THE CONFIGURATION OF CNN MODEL

Layer No.	Type	Size	Feature maps
1-3	Convolution	3×3	16
4-5	Convolution	3×3	32
6-7	Convolution	3×3	64
8	Pooling	3×3	64
9	Fully connected	-	200
10	Fully connected	-	16

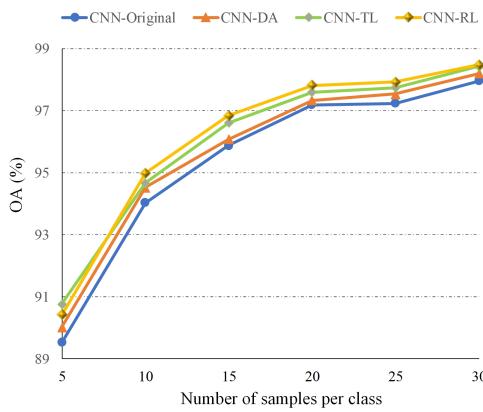


Fig. 22. The OA values obtained by different methods versus number of samples.

good performance. In contrast to traditional hand-crafted feature-based classification methods, deep learning can automatically learn complex features of HSIs with a large number of hierarchical layers. In this literature survey, we briefly introduced several deep models that are often used to classify HSIs, including SAE, DBN, CNN, RNN, and GAN. Then, we focused on deep learning-based classification methodologies for HSIs and provided a general and comprehensive overview on the existing methods in a unified framework. Specifically, these deep networks used in the HSI classification were divided into spectral-feature networks, spatial-feature networks, and spectral-spatial-feature networks, where each category extracts the corresponding feature. Through this framework, we can easily see that deep networks make full use of different feature types for classification. We have also compared and analyzed the performances of various HSI classification methods, including four traditional machine learning-based methods and six deep learning-based methods. The classification accuracies obtained by different methods demonstrate that deep learning-based methods overall outperform the non-deep-learning-based methods and the DFFN which combines the residual learning and feature fusion achieves best classification performance. Furthermore, deep features and network weights were visualized, which is useful for analyzing the network performance and further designing the deep architecture. In addition, considering the fact that available training samples in remote sensing are usually very limited and training deep network requires a large number of samples, we also included some strategies to improve classification performance. We have also conducted experiments to verify and compared the

effectiveness of these strategies. The final results show that the residual learning obtains the highest improvement among all approaches. This experimental result may provide some guidelines for the future study on this topic.

REFERENCES

- [1] L. Liang, L. Di, L. Zhang, M. Deng, Z. Qin, S. Zhao, and H. Lin, "Estimation of crop lai using hyperspectral vegetation indices and a hybrid inversion method," *Remote Sens. Environ.*, vol. 165, pp. 123–134, Aug. 2015.
- [2] X. Yang and Y. Yu, "Estimating soil salinity under various moisture conditions: An experimental study," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2525–2533, May 2017.
- [3] N. Yokoya, J. C.-W. Chan, and K. Segl, "Potential of resolution-enhanced hyperspectral data for mineral mapping using simulated enmap and sentinel-2 images," *Remote Sens.*, vol. 8, no. 3, pp. 172–189, Feb. 2016.
- [4] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4118–4130, Aug. 2018.
- [5] S. Zhang, J. Li, Z. Wu, and A. Plaza, "Spatial discontinuity-weighted sparse unmixing of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5767–5779, Oct. 2018.
- [6] P. Ghamisi, N. Yokoya, J. Li, W. Liao, S. Liu, J. Plaza, B. Rasti, and A. Plaza, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, Dec. 2017.
- [7] Y. Zhong and L. Zhang, "An adaptive artificial immune network for supervised classification of multi-/hyperspectral remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 894–909, Mar. 2012.
- [8] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.
- [9] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4085–4098, Nov. 2010.
- [10] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 809–823, Mar. 2012.
- [11] B. Du and L. Zhang, "Random-selection-based anomaly detector for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 5, pp. 1578–1589, May 2011.
- [12] B. Du and L. Zhang, "Target detection based on a dynamic subspace," *Pattern Recog.*, vol. 47, no. 1, pp. 344–358, Jan. 2014.
- [13] G. Licciardi, P. R. Marpu, J. Chanussot, and J. A. Benediktsson, "Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 447–451, May 2012.
- [14] S. Prasad and L. M. Bruce, "Limitations of principal components analysis for hyperspectral target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 625–629, Oct. 2008.
- [15] A. Villa, J. A. Benediktsson, J. Chanussot, and C. Jutten, "Hyperspectral image classification with independent component discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 4865–4876, Dec. 2011.
- [16] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, Mar. 2009.
- [17] P. Ghamisi *et al.*, "New frontiers in spectral-spatial hyperspectral image classification: The latest advances based on mathematical morphology, markov random fields, segmentation, sparse representation, and deep learning," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 3, pp. 10–43, Sep. 2018.
- [18] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral-spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.
- [19] J. A. Benediktsson, J. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.

- [20] J. Li, P. R. Marpu, A. Plaza, J. M. Bioucas-Dias, and J. A. Benediktsson, "Generalized composite kernel framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4816–4829, Sep. 2013.
- [21] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Trans. Geosci. Remote Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.
- [22] Ma. Fauvel, J. Chanussot, and J. A. Benediktsson, "A spatial-spectral kernel-based approach for the classification of remote-sensing images," *Pattern Recog.*, vol. 45, no. 1, pp. 381–392, Jan. 2012.
- [23] L. Fang, S. Li, W. Duan, J. Ren, and J. A. Benediktsson, "Classification of hyperspectral images by exploiting spectral-spatial information of superpixel via multiple kernels," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6663–6674, Dec. 2015.
- [24] X. Kang, S. Li, and J. A. Benediktsson, "Spectral-spatial hyperspectral image classification with edge-preserving filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2666–2677, May 2014.
- [25] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.
- [26] L. Fang, S. Li, X. Kang, and J. A. Benediktsson, "Spectral-spatial hyperspectral image classification via multiscale adaptive sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7738–7749, Dec. 2014.
- [27] L. Fang, C. Wang, S. Li, and J. A. Benediktsson, "Hyperspectral image classification via multiple-feature-based adaptive sparse representation," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 7, pp. 1646–1657, Jul. 2017.
- [28] S. Li, T. Lu, L. Fang, X. Jia, and J. A. Benediktsson, "Probabilistic fusion of pixel-level and superpixel-level hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7416–7430, Dec. 2016.
- [29] T. Lu, S. Li, L. Fang, X. Jia, and J. A. Benediktsson, "From subpixel to superpixel: A novel fusion framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4398–4411, Aug. 2017.
- [30] L. Fang, N. He, S. Li, P. Ghamisi, and J. A. Benediktsson, "Extinction profiles fusion for hyperspectral images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1803–1815, Mar. 2018.
- [31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. R., D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1–9.
- [33] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 580–587.
- [34] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, "Joint learning of words and meaning representations for open-text semantic parsing," in *Proc. Int. Conf. Art. Intell. Stat.*, 2012, pp. 127–135.
- [35] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.
- [36] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Trans. Geosci. Remote Sens. Maga.*, vol. 5, no. 1, pp. 8–32, Mar. 2017.
- [37] Y. Chen, X. Zhao, and X. Jia, "Spectral-spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015.
- [38] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.
- [39] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Comput.*, vol. 1, no. 2, pp. 270–280, 1989.
- [40] P. Rodriguez, J. Wiles, and J. L. Elman, "A recurrent neural network that learns to count," *Connection Sci.*, vol. 11, no. 1, pp. 5–40, 1999.
- [41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [42] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint, arXiv:1308.0850*, 2013.
- [43] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.
- [44] L. Mou, P. Ghamisi, and X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.
- [45] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [46] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.
- [47] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv: 1511.06434v2*, 2016.
- [48] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," *arXiv preprint arXiv: 1610.09585*, 2016.
- [49] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436, 2015.
- [50] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Trans. Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.
- [51] X. Zhu, D. Tuia, L. Mou, G. Song Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Trans. Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.
- [52] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: A survey," *WIREs Data Mining Knowl. Discov.*, , no. 12, pp. 1264–1280, May 2018.
- [53] P. Liu, H. Zhang, and K. B. Eom, "Active deep learning for classification of hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 2, pp. 712–724, Feb. 2017.
- [54] P. Zhong, Z. Gong, S. Li, and C. B. Schönlieb, "Learning to diversify deep belief networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3516–3530, Jun. 2017.
- [55] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, Jan. 2015, Art. no. 258619.
- [56] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.
- [57] J. M. Haut, M. E. Paoletti, J. Plaza, J. Li, and A. Plaza, "Active learning with convolutional neural networks for hyperspectral image classification using a new bayesian approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6440–6461, Nov. 2018.
- [58] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018.
- [59] Y. Zhan, D. Hu, Y. Wang, and X. Yu, "Semisupervised hyperspectral image classification based on generative adversarial networks," *IEEE Geosci. Remote Lett.*, vol. 15, no. 2, pp. 212–216, Feb. 2018.
- [60] H. Wu and S. Prasad, "Convolutional recurrent neural networks for hyperspectral data classification," *Remote Sens.*, vol. 9, no. 3, pp. 298–317, Mar. 2017.
- [61] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.
- [62] S. Tariyal, H. K. Aggarwal, and A. Majumdar, "Greedy deep dictionary learning for hyperspectral image classification," in *Proc. IEEE GRSS Workshop Hyperspectral Image Signal Process.: Evol. Remote Sens.*, 2016, pp. 21–24.
- [63] V. Singhal, H. K. Aggarwal, S. Tariyal, and A. Majumdar, "Discriminative robust deep dictionary learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5274–5283, Sep. 2017.
- [64] J. A. Benediktsson and P. Ghamisi, *Spectral-spatial classification of hyperspectral remote sensing images*, Artech House, INC, Boston, USA, 2015.
- [65] J. Yue, W. Zhao, S. Mao, and H. Liu, "Spectral-spatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sens. Lett.*, vol. 6, no. 6, pp. 468–477, May 2015.
- [66] K. Makantasis, K. Karantzalos, A. Doula, and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2015, pp. 4959–4962.

- [67] L. Fang, Z. Liu, and W. Song, "Deep hashing neural networks for hyperspectral image feature extraction," *IEEE Geosci. Remote Sens. Lett.*, DOI: 10.1109/LGRS.2019.2899823, to be published, 2019.
- [68] H. Liang and Q. Li, "Hyperspectral imagery classification using sparse representations of convolutional neural network features," *Remote Sens.*, vol. 8, no. 2, pp. 99–114, Jan. 2016.
- [69] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6712–6722, Nov. 2018.
- [70] W. Zhao and S. Du, "Spectral-spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, Aug. 2016.
- [71] L. Jiao, M. Liang, H. Chen, S. Yang, H. Liu, and X. Cao, "Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5585–5599, Oct. 2017.
- [72] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3431–3440.
- [73] X. Ma, H. Wang, and J. Geng, "Spectral-spatial classification of hyperspectral image based on deep auto-encoder," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 9, pp. 4073–4085, Sep. 2016.
- [74] X. Zhou, S. Li, F. Tang, K. Qin, S. Hu, and S. Liu, "Deep learning with grouped features for spatial spectral classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 1, pp. 97–101, Jan. 2017.
- [75] X. Sun, F. Zhou, J. Dong, F. Gao, Q. Mu, and X. Wang, "Encoding spectral and spatial context information for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2250–2254, Dec. 2017.
- [76] Q. Lv, X. Niu, Y. Dou, J. Xu, and Y. Lei, "Classification of hyperspectral remote sensing image using hierarchical local-receptive-field-based extreme learning machine," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 434–438, Mar. 2016.
- [77] S. Mei, J. Ji, J. Hou, X. Li, and Q. Du, "Learning sensor-specific spatial-spectral features of hyperspectral images via convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4520–4533, Aug. 2017.
- [78] X. Kang, C. Li, S. Li, and H. Lin, "Classification of hyperspectral images by gabor filtering based deep network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1166–1178, Apr. 2018.
- [79] Y. Chen, L. Zhu, P. Ghamisi, X. Jia, G. Li, and L. Tang, "Hyperspectral images classification with gabor filtering and convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2355–2359, Dec. 2017.
- [80] E. Aptoula, M. Can Ozdemir, and B. Yanikoglu, "Deep learning with attribute profiles for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1970–1974, Dec. 2016.
- [81] P. Ghamisi, B. Höfle, and X. Zhu, "Hyperspectral and lidar data fusion using extinction profiles and deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 3011–3024, Jun. 2017.
- [82] B. Pan, Z. Shi, and X. Xu, "R-VCANet: A new deep-learning-based hyperspectral image classification method," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1975–1986, May 2017.
- [83] N. He, M. E. Paoletti, Haut J. M., L. Fang, S. Li, A. Plaza, and J. Plaza, "Feature extraction with multiscale covariance maps for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 755–769, Feb. 2019.
- [84] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, "Deep fusion of remote sensing data for accurate classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1253–1257, Aug. 2017.
- [85] M. Zhang, W. Li, and Q. Du, "Diverse region-based CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2623, Jun. 2018.
- [86] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu, and J. Paisley, "Hyperspectral image classification with markov random fields and a convolutional neural network," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2354–2367, May 2018.
- [87] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018.
- [88] W. Song, S. Li, and Y. Li, "Hyperspectral images classification with hybrid deep residual network," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2017, pp. 2235–2238.
- [89] B. Liu, X. Yu, P. Zhang, A. Yu, Q. Fu, and X. Wei, "Supervised deep feature extraction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 1909–1921, Apr. 2018.
- [90] J. Zhu, L. Fang, and P. Ghamisi, "Deformable convolutional neural networks for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 8, pp. 1254–1258, Aug. 2018.
- [91] L. Fang, G. Liu, S. Li, P. Ghamisi, and J. A. Benediktsson, "Hyperspectral image ilassification with squeeze multibias network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1291–1301, Mar. 2019.
- [92] Y. Li, H. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens.*, vol. 9, no. 1, pp. 67–87, Jan. 2017.
- [93] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral-spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–856, Feb. 2018.
- [94] A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018.
- [95] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "A new deep convolutional neural network for fast hyperspectral image classification," *ISPRS J. of Photogramm. and Remote Sens.*, vol. 145, pp. 120–147, 2018.
- [96] J. Li, X. Zhao, Y. Li, Q. Du, B. Xi, and J. Hu, "Classification of hyperspectral imagery using a new fully convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 292–296, Feb. 2018.
- [97] L. Mou, P. Ghamisi, and X. Zhu, "Unsupervised spectral-spatial feature learning via deep residual conv-deconv network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 391–406, Jan. 2018.
- [98] H. Lee and H. Kwon, "Going deeper with contextual cnn for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.
- [99] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral-spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.
- [100] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.
- [101] M. Zhang, M. Gong, Y. Mao, J. Li, and Y. Wu, "Unsupervised feature extraction in hyperspectral images based on wasserstein generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, DOI: 10.1109/TGRS.2018.2876123, to be published, 2019.
- [102] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3856–3866.
- [103] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, J. Li, and F. Pla, "Capsule networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, DOI: 10.1109/TGRS.2018.2871782, to be published, 2019.
- [104] R. Kemker and C. Kanan, "Self-taught feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2639–2705, May 2017.
- [105] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1259–1270, Mar. 2018.
- [106] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, DOI: 10.1109/TGRS.2019.2899129, to be published, 2019.
- [107] J. Yang, Yong Q. Zhao, and C.-W. Chan, "Learning and transferring deep joint spectral-spatial features for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4729–4742, Aug. 2017.
- [108] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.
- [109] S. Hao, W. Wang, Y. Ye, T. Nie, and L. Bruzzone, "Two-stream deep architecture for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2349–2361, Apr. 2018.

- [110] B. Pan, Z. Shi, N. Zhang, and S. Xie, “Hyperspectral image classification based on nonlinear spectral-spatial network,” *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1782–1786, Dec. 2016.
- [111] C. Tao, H. Pan, Y. Li, and Z. Zou, “Unsupervised spectral-spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2438–2442, Dec. 2015.
- [112] A. Santara, K. Mani, P. Hatwar, A. Singh, A. Garg, K. Padia, and P. Mitra, “BASS Net: Band-adaptive spectral-spatial feature learning neural network for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5293–5301, Sep. 2017.
- [113] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.
- [114] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, “Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery,” *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, Nov. 2015.
- [115] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, “Deep learning earth observation classification using imagenet pretrained networks,” *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.
- [116] X. Zhang, Y. Liang, C. Li, H. Ning, L. Jiao, and H. Zhou, “Recursive autoencoders-based unsupervised feature learning for hyperspectral image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 1928–1932, Nov. 2017.
- [117] A. Romero, C. Gatta, and G. Camps-Valls, “Unsupervised deep feature extraction for remote sensing image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1349–1362, Mar. 2016.
- [118] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv preprint arXiv: 1502.03167*, pp. 448–456, 2015.
- [119] C.-C. Chang and C.-J Lin, “Libsvm: A library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, Jul. 2011.