Deep Filter Banks for Land-Use Scene Classification

Hang Wu, Baozhen Liu, Weihua Su, *Member, IEEE*, Wenchang Zhang, and Jinggong Sun

***Abstract*— Land-use (LU) scene classification is one of the most challenging tasks in the field of remote sensing (RS) image processing due to its high intraclass variability and low interclass distance. Motivated by the challenge posed by this problem, we propose a novel hybrid architecture, deep filter banks, combining multicolumn stacked denoising sparse autoencoder (SDSAE) and Fisher vector (FV) to automatically learn the representative and discriminative features in a hierarchical manner for LU scene classification. SDSAE kernels describe local patches and a robust global feature of the RS image is built through the FV pooling layer. Unlike previous handcrafted features, we use machine-learning mechanisms to optimize our proposed feature extractor so that it can learn more suitable internal features from the RS data, boosting the final performance. Our approach achieves superior performance compared with the state-of-the-art methods, obtaining average classification accuracies of 92.7% and 90.4%, respectively, on the UC Merced and RSSCN7 data sets.**

***Index Terms*— Deep filter banks, Fisher vector (FV), land-use (LU) scene classification, stacked denoising sparse autoencoder (SDSAE).**

I. Introduction

With the rapid development of airborne or spaceborne imaging sensors, remote sensing (RS) images can provide a spatial resolution of up to 0.41 m [1]. A massive amount of high spatial resolution images has become available for precise land-use (LU) scene classification, which aims to assign a semantic label to an RS image according to its content. Therefore, it is necessary to develop effective and efficient scene classification methods to annotate the RS images.

High intraclass variability coupled with low interclass distance makes labeling RS images a challenging problem in the RS field. The same land cover and even the same objects may appear on RS images belonging to different LU classes [2]. LU scene classification calls for efficient and strong discrimination of features.

In recent research, the bag-of-visual-words (BOVW) model [3], [4] is a common and promising tool to solve the above problem. It represents RS images using the frequency of codewords that are constructed by quantifying local features, e.g., SIFT and HOG, with a clustering method such as K-means. The frequency vector as the final global representation is then fed into a pretrained classifier to obtain LU scene classification results. The traditional BOVW just considers the occurrences of visual words, neglecting information about the spatial distribution. Several improved variants of BOVW are proposed to make up for this deficiency. For example, spatial pyramid match kernel (SPMK) [3] and randomized spatial partition [5] have added absolute spatial information into the final representation. Further, spatial pyramid co-occurrence kernel (SPCK) [6] and pyramid of spatial relations [7] are designed to describe both the relative and absolute spatial arrangement of the codewords.

The multilayer model is another effective way to improve the scene classification performance. Hierarchical coding vector (HCV) [8] and two-layer sparse coding model [9] stack multiple BOVW coding layers to develop a hierarchical feature learning structure, acquiring a more powerful representation to describe the RS images. Some researchers choose to circumvent the BOVW model and directly design low-level global descriptors. Zhao *et al.* [2] propose a novel spectral feature, i.e., MeanStd, for LU scene classification and Chen *et al.* [10] construct a multiscale completed local binary patterns (MS-CLBP) descriptor to characterize the dominant texture features in the RS images. Although these methods have achieved good performance, they are essentially handcrafted descriptors, and it is difficult to achieve further enhancements in the LU scene classification performance due to the limited descriptive ability of these low- and mid-level features.

Recently, Fisher vectors (FVs) and deep neural networks (DNNs) have attracted attention in the computer vision community as two great image classification pipelines with different strengths. DNNs have shown superior accuracy on a number of classification tasks, but FV classifiers are typically less costly to train and evaluate [11]. For DNNs, the typical architecture includes deep belief networks (DBNs) [12], stacked autoencoder (SAE) [13], and convolutional neural networks (CNNs) [14]. CNNs have achieved remarkable success in many computer vision benchmarks. However, the supervised deep model, like CNNs, requires a tremendous amount of labeled data, which is very expensive to obtain in the RS filed. This intrinsic characteristic limits its application for LU scene classification. SAE, a typical unsupervised feature learning method, is suitable for solving this dilemma. It learns complex semantic information from RS images

Manuscript received August 26, 2016; revised October 3, 2016; accepted October 9, 2016. Date of publication October 25, 2016; date of current version December 7, 2016. This work was supported by the Science and Technology Pillar Program, Tianjin, China, under Project 16YFZCSF00590.

H. Wu, B. Liu, W. Su, and J. Sun are with the Institute of Medical Equipment, Academy of Military Medical Science, Tianjin 300161, China (e-mail: 2008.wuhang@163.com; liubaozhen91@126.com; directorsu@126.com; sunjg@vip.sina.com).

W. Zhang is with the State Key Laboratory of Intelligent Technology and System, Computer Science and Technology School, Tsinghua University, Beijing 100084, China (e-mail: zwc0501@163.com).

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

Digital Object Identifier 10.1109/LGRS.2016.2616440

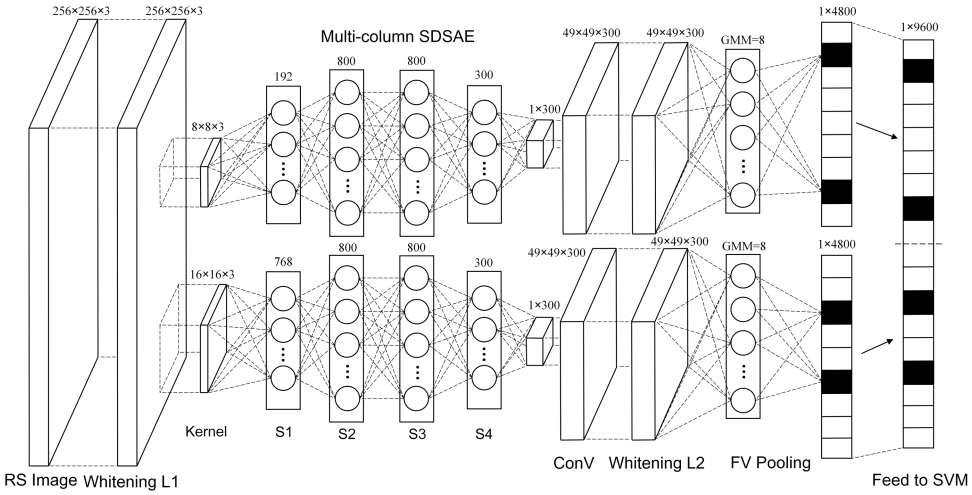

Fig. 1. Illustration of the proposed deep filter banks for LU scene classification.

by encoding vector-form input data and then reconstructing it [15], [16]. Stacked denoising sparse autoencoder (SDSAE) is an enhanced variant of SAE using sparsity and denoising criterions.

Inspired by the success of DNNs and FV in the computer vision community, we propose a hybrid architecture, i.e., *deep filter banks*, which employs unsupervised SDSAE and an FV pooling layer to automatically learn the abstract semantic representation for tackling the LU scene classification task. The SDSAE hierarchically refines deep semantic information from local patches in the RS images and subsequently feeds the information into an FV pooling layer to develop a robust global representation. The effectiveness of the proposed model is demonstrated on the UC Merced and RSSCN7 data sets. The major contributions of our work are the following three points.

1) We first combine multicolumn SDSAE and FV to construct deep filter banks. This structure forces our model to learn more robust and abstract semantic features for LU scene classification.
2) Unlike handcrafted feature representation-based methods, our deep filter banks use machine-learning mechanisms to optimize themselves to learn more suitable internal features from the RS data.
3) Superior performance is achieved by our deep filter banks compared with the state-of-the-art results on the UC Merced and RSSCN7 data sets.

II. Deep Filter Banks

A. Overall Architecture

The proposed deep filter banks form an unsupervised deep network that stacks multicolumn SDSAE and an FV pooling layer to generate high-level feature representation.

Fig. 1 illustrates the whole structure of our proposed model. An input RS image is first sphered by a whitening L1 layer to remove redundancy of raw data. Each kernel is a 3-D array with size $K_1 \times K_2 \times C$. $N \times N$ local patches are extracted from whitening L1 layer for each kernel using the sliding window technique. Each patch is subsequently transformed into a discriminating local feature vector $1 \times L$ through refining semantic information layer-by-layer. The local feature vectors developed by multicolumn SDSAE then construct robust and abstract feature maps, i.e., Conv layer $N \times N \times L$. Through whitening, the local abstract features are fed into the FV pooling layer to produce global deep representation. Different FVs from different columns are then concatenated into a final representation that can be input into a classifier such as a support vector machine (SVM) to obtain the LU classification result. Unlike handcrafted feature extractors, the SDSAE in our deep filter banks automatically learns parameters from those patches in training RS images using the layerwise pretraining approach and fine-tuning strategy.

Deep filter banks with more columns and deeper layers can learn more complicated abstract features, but this increases the complexity of our model. Considering the tradeoff between effectiveness and efficiency, two different kernels ($16 \times 16 \times 3$ and $8 \times 8 \times 3$) and three hidden layers (800-800-300) are used in this framework. The proposed deep filter banks can be generalized to more columns and hidden layers without difficulty. N and L are set as 49 and 300 in our work, respectively. It should be noted that our framework is independent of the size of the input images. The input as 256×256 pixels in Fig. 1 is used because it is the setting of the public data set.

B. Stacked Denoising Sparse Autoencoder

DNN is a computational model composed of multiple processing layers to learn representations of data with multiple levels of abstraction [17]. The SAE is a typical DNN that is composed of multiple layers of autoencoder (AE) [13]. SDSAE incorporates sparsity and denoising criterions on the basis of SAE. The AE learns features in an unsupervised manner by minimizing the reconstruction error between inputs at the encoding layer and reconstruction at the decoding layer [1]. During the encoding step, a nonlinear activation function $g(x)$transforms the input data $\boldsymbol{x} \in \mathbb{R}^K$ into a hidden representation $\boldsymbol{y} \in \mathbb{R}^M$

$$y = g(Wx + b) \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{M\times K}$ is a weight matrix, $\mathbf{b} \in \mathbb{R}^{K}$ is abias vector, and $g(x)$ is chosen to be a rectified linear unit $g(x) = \max(0, x)$ in our model. We perform the decoding of $\boldsymbol{y}$ using a separate linear decoding matrix

$$z = W' \cdot y + b' \tag{2}$$

where $\mathbf{W}' \in \mathbb{R}^{K\times M}$ is a weight matrix and $\mathbf{b}' \in \mathbb{R}^{M}$ is the decoding bias. Feature extractor is learned by minimizing the cost function

$$L(x, z) = \frac{1}{2}\sum\nolimits_{i=1}^{K} \|x_i - z_i\|^2 + \frac{\lambda}{2}\|W\|^2 \tag{3}$$

where $\boldsymbol{x}$ and $\boldsymbol{z}$ are the training and reconstruction data, respectively. The first term is the reconstruction error, and the second term is a weight decay term to relieve overfitting. λ is the weight of this term. Sparsity is considered to be an important attribute of the feature with strong discriminability [13]. Therefore, we add a sparsity constraint to the cost function

$$L(x, z) + \beta \sum\nolimits_{j=1}^{M} KL(\rho \parallel \rho_1) \tag{4}$$

$$KL(\rho \parallel \rho_1) = \rho \log\left(\frac{\rho}{\rho_1}\right) + (1-\rho)\log\left(\frac{1-\rho}{1-\rho_1}\right) \tag{5}$$

where β is the weight of the sparsity penalty, ρ_1 is the average activation of each hidden neuron, and ρ is the sparsity target that is typically a small value close to zero. $KL(\ )$ would increase monotonically as ρ_1 diverges from ρ. We train our model with noisy input to enhance the generalization performance. The denoising criterion forces the model to capture implicit invariances in the data, resulting in robust features.

For the multicolumn SDSAE in our deep filter banks, the parameters are updated using backpropagation with mini-batch stochastic gradient and a batch size of 100. A total of 25% of zero masking corrupt noise [13] is employed in the input to our proposed model. The λ, β, and ρ are set to 0.003, 3, and 0.05, respectively. To retain the spatial information, the local features $\mathbf{F}$ developed by SDSAE are augmented with their x- and y-coordinates before being fed into the FV pooling layer.

C. Fisher Vector Pooling Layer

The FV pooling layer aggregates local features $\mathbf{F} \in \mathbb{R}^{M\times \mathrm{T}}$ into the global representation $\mathbf{d} \in \mathbb{R}^{2MN}$, thus achieving greater invariance to image transformations and better robustness to noise and clutter. Local features were decorrelated using whitening technology before being fed into the FV pooling layer.

The FV method is based on fitting a parametric generative model [e.g., Gaussian mixture model (GMM)] to the input local features $\mathbf{F}$ and then encoding derivatives of the log-likelihood of the model with respect to its parameters. The GMMs with diagonal covariance are used in our deep filter banks framework, leading to a deep representation that captures the Gaussian mean (first) and variance (second) differences between the input local features $\mathbf{F}$ and each of the GMM centers

$$d_n^{(1)} = \frac{1}{T\sqrt{w_n}}\sum_{t=1}^{T}\alpha_t(n)\left(\frac{F_t - \mu_n}{\sigma_n}\right) \tag{6}$$

$$d_n^{(2)} = \frac{1}{T\sqrt{2w_n}}\sum_{t=1}^{T}\alpha_t(n)\left(\frac{(F_t - \mu_n)^2}{\sigma_n^2} - 1\right) \tag{7}$$

where $\{w_n, \mu_n, \sigma_n\}_n$ are the respective mixture weights, means, and diagonal covariance, respectively, of the GMM codebook $\mathbf{B} \in \mathbb{R}^{M\times N}$. This codebook is pregenerated in the training phase by GMM clustering. F_t is one local feature fed into the Fisher pooling layer and T is the number of the local features. $\alpha_t(n)$ is the soft assignment weight of the tth local features F_t to the nth Gaussian distribution. Finally, the global representation $\mathbf{d} \in \mathbb{R}^{M\times 2N}$ is obtained by stacking the first and second differences

$$d = \left[d_1^{(1)}, d_1^{(2)}, d_2^{(1)}, d_2^{(2)}, \ldots d_n^{(1)}, d_n^{(2)}, \ldots, d_N^{(1)}, d_N^{(2)}\right]. \tag{8}$$

The output vector $\mathbf{d}$ is subsequently normalized using the power $+ L_2$ scheme. Output vectors from different columns are then concatenated as the final scene representation $\mathbf{D}$ of our proposed deep filter banks. In our framework, we choose the size of the GMM codebook as 8. This value strikes a good compromise between efficiency gain and accuracy loss according to the experimental results.

D. Reducing Overfitting

To avoid overfitting, we use dropout in SDSAE of the proposed deep filter banks. This strategy randomly omits each neuron in the hidden layers with a given probability, forcing neurons to provide a more useful and robust contribution in combination with arbitrary active neuron combinations [14]. Different networks of SDSAE are trained in different periods. The changing training structure significantly reduces overfitting. The dropout rate is set as 50% for all the layers in the SDSAE of our proposed deep filter banks model.

III. Experiments and Analysis

A. Experimental Design

To evaluate the effectiveness of the proposed deep filter banks, we conducted LU scene classification experiments using UC Merced and RSSCN7 data sets. The one versus rest linear SVM classifier is employed and the average classification accuracy (mean $\pm$ SD) is set as the evaluation index.

B. Experiment 1: UC Merced Image Data Set

The UC Merced image data set [3] is one of the first publicly available LU geographical image data sets with ground truth (http://vision.ucmerced.edu/data sets.html). The data set consists of 21 LU classes, and each class contains 100 images of the same size (i.e., 256×256 pixels). The pixel resolutions of all the images are 30 cm per pixel. Sample images of each LU class are shown in Fig. 2. Following [3], the data set was randomly partitioned into five equal subsets. Each subset contained 20 images from each LU category. Four subsets were used for training, and the remaining subset was used for testing.

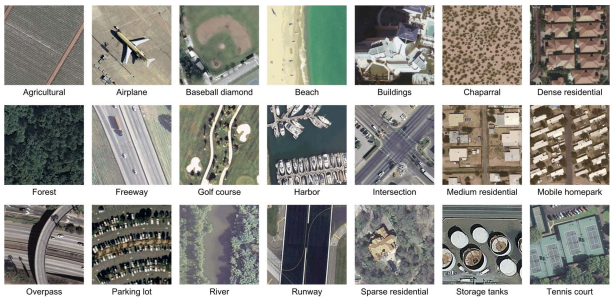

Fig. 2. Sample image from the 21 categories in the UC Merced data set.

TABLE I
COMPARISON OF OUR DEEP FILTER BANKS WITH THE STATE-OF-THE-ART PERFORMANCE REPORTED IN THE LITERATURE ON THE UC MERCED DATA SET

Method	Accuracy (%)
BOVW [3]	76.8
SPMK [3]	75.3
Color Gabor [3]	80.5
SPCK + SPM [6]	77.4
Wavelet BOVW [4]	87.4 ± 1.3
Unsupervised feature learning [18]	81.1 ± 1.2
Saliency-guided feature learning [19]	82.7 ± 1.2
Concentric circle-structured BOVW [10]	86.6 ± 0.8
Multifeature concatenation [20]	89.5 ± 0.8
Pyramid-of-spatial-relations [7]	89.1
CLBP [10]	85.5 ± 1.9
MS-CLBP [10]	90.6 ± 1.4
RDF-CNN [21]	85.8
RDSG-CNN [21]	89.9
UFL-SC [22]	90.3 ± 1.5
HCV [8]	90.5 ± 1.1
SSBFC [2]	90.9 ± 0.9
Deep filter banks	**92.7 ± 0.8**

The classification performances of different methods with the UC Merced image data set are shown in Table I. As shown in Table I, our deep filter banks outperformed the current state-of-the-art results on this data set. The result of our method (92.7 ± 0.8) is the best among all, which demonstrates the effectiveness of the proposed deep filter banks for LU scene classification. Furthermore, the statistical z-test is used to test the validity of the improvement. The result $p \leq 0.05$ ensures that the performance improvement is meaningful.

C. Experiment 2: RSSCN7 Data Set

The RSSCN7 data set [23] is a public data set (https://sites.google.com/site/qinzoucn/documents) released in 2015. It contains 2800 RS scene images that are from seven typical LU scene categories. There are 400 images with sizes of 400×400 pixels for each class. Each scene category is of four different scales (1:700, 1:1300, 1:2600, and 1:5200) with 100 images per scale. The experimental setup in [23] is used. Half of the images in each category were fixed for training and the rest for testing.

Table II shows the classification accuracies of different methods for the RSSCN7 data set. It can be observed that our method achieves an accuracy of 90.4 ± 0.6, improving

TABLE II
COMPARISON OF OUR DEEP FILTER BANKS WITH THE STATE-OF-THE-ART PERFORMANCE REPORTED IN THE LITERATURE ON THE RSSCN7 DATA SET

Method	Accuracy (%)
GIST [8]	69.5 ± 0.9
Color histogram [8]	70.9 ± 0.8
LBP [8]	75.3 ± 1.0
DBN-based feature selection [23]	77.0
HCV [8]	84.7 ± 0.7
Deep filter banks	**90.4 ± 0.6**

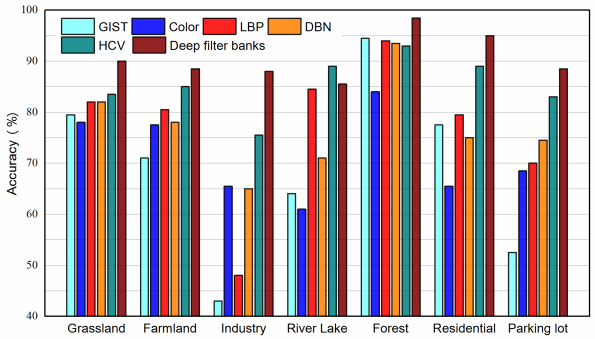

Fig. 3. Comparison of deep filter banks with GIST, color histogram, LBP, DBN, and HCV based on categorywise performance.

the classification performance significantly with a noticeable margin on this data set. To further investigate the performance of deep filter banks, we illustrate per-class accuracies of the RSSCN7 data set in Fig. 3.

From Fig. 3, we observe that the proposed deep filter banks is effective for almost all the geographical classes on the RSSCN7 data set. Except for the *River and Lake*, the deep filter banks achieve better performance in all the other categories compared with other methods. The performance improvement is especially profound over the *Industry* and *Parking lot* categories, which need stronger semantic understanding. It should be noted that our deep filter banks used the one versus rest linear SVM classifier in the two data sets. The linear classifier makes the framework simpler and more conducive to practical application. The classification performance of our method should be improved further with a sophisticated classifier, e.g., nonlinear SVM kernel or extreme learning machine.

D. Experiment 3: Framework Analysis

We study the performance of different frameworks in our proposed deep filter banks. We evaluated four different depths of SDSAE (corresponding hidden layers: 300, 800-300, 800-800-300, and 800-800-800-300) and compared the FV-pooling layer to the traditional average-pooling approach. Results on the UC Merced and RSSCN7 data sets are shown in Fig. 4. In addition, we show the performance using FV-pooling constructed from the typical handcrafted feature, i.e., dense SIFT.

As shown in Fig. 4, the classification accuracy increases with the deeper framework of more hidden layers at the

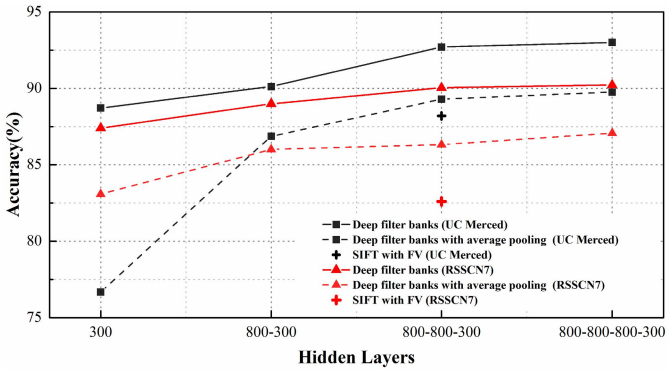

Fig. 4. Performance of the deep filter banks with different frameworks. (The unmarked accuracies of SIFT with average pooling are 42.2% and 50.4% for UC Merced and RSSCN7, respectively.)

slowing growth rate. The deep filter banks significantly outperform SIFT with the same pooling approach, demonstrating the strong discriminability of the local features developed by SDSAE. Furthermore, it can be observed that the FV-pooling layer can produce a much better global representation compared with the traditional average-pooling approach.

E. Computational Complexity

Many approaches with a nonlinear classifier have to pay a penalty for computational complexity $O(n^2)$ or $O(n^3)$ in the train phase and $O(n)$ in the testing phase, where n is the training size. It implies poor scalability for the real application. Our deep filter banks, using a simple linear SVM, reduce the training complexity to $O(n)$ and obtain constant complexity in testing while still achieving superior performance. Finally, we evaluated the computation complexity of our proposed deep filter banks and used the UC Merced image data set to obtain the processing time. All the codes of our proposed model are implemented in MATLAB 2014a and run on a computer with an Intel Xeon CPU E5-2620 v2 at 2.1 GHz and 32-GB RAM in a 64-b Win7 operation system. As observed from our experiment, the training phase takes approximately 6.25 h, and the average processing time for a test RS image (size of 256×256 pixels) is 0.26 ± 0.01 s. The training phase of our method is slightly time-consuming compared with the handcrafted features, which are highly dependent on the prior knowledge. After completing the training phase, our method is of good efficiency.

IV. Conclusion

In this letter, we proposed hybrid architecture, deep filter banks for LU scene classification that combines the benefits of FV and DNNs pipelines. The proposed model stack multicolumn SDSAE and FV pooling layer to learn robust and abstract hierarchical semantic feature representations from raw RS data. The experimental results validated the effectiveness of our method and showed that it outperforms the current state-of-the-art methods on the challenging UC Merced and RSSCN7 data sets.

References

[1] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[2] B. Zhao, Y. Zhong, and L. Zhang, "A spectral–structural bag-of-features scene classifier for very high spatial resolution remote sensing imagery," *ISPRS J. Photogram. Remote Sens.*, vol. 116, pp. 73–85, Jun. 2016.

[3] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM SIGSPATIAL GIS*, Nov. 2010, pp. 270–279.

[4] L. Zhao, P. Tang, and L. Huo, "A 2-D wavelet decomposition-based bag-of-visual-words model for land-use scene classification," *Int. J. Remote Sens.*, vol. 35, no. 6, pp. 2296–2310, Mar. 2014.

[5] Y. Jiang, J. Yuan, and G. Yu, "Randomized spatial partition for scene recognition," in *Proc. Comput. Vis.*, vol. 7573. 2012, pp. 730–743.

[6] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1465–1472.

[7] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1947–1957, Apr. 2015.

[8] H. Wu, B. Liu, W. Su, W. Zhang, and J. Sun, "Hierarchical coding vectors for scene level land-use classification," *Remote Sens.*, vol. 8, no. 5, p. 436, May 2016.

[9] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 173–176, Jan. 2011.

[10] C. Chen, B. Zhang, H. Su, W. Li, and L. Wang, "Land-use scene classification using multi-scale completed local binary patterns," *Signal, Image Video Process.*, vol. 10, no. 4, pp. 745–752, Apr. 2016.

[11] F. Perronnin and D. Larlus, "Fisher vectors meet neural networks: A hybrid classification architecture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3743–3752.

[12] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[13] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, Dec. 2010.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2012, pp. 1097–1105.

[15] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1279–1289, Jun. 2016.

[16] Y. Yuan, J. Lin, and Q. Wang, "Dual-clustering-based hyperspectral band selection by contextual analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1431–1445, Mar. 2016.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[18] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2014.

[19] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[20] W. Shao, W. Yang, G. S. Xia, and G. Liu, "A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization," in *Computer Vision Systems*. Lecture Notes in Computer Science, vol. 7963. Berlin, Germany: Springer-Verlag, 2013, pp. 324–333.

[21] Y. Zhong, F. Fei, and L. Zhang, "Large patch convolutional neural networks for the scene classification of high spatial resolution imagery," *J. Appl. Remote Sens.*, vol. 10, no. 2, art. no. 025006, Apr. 2016.

[22] F. Hu, G. S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2015–2030, May 2015.

[23] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.