INTERNATIONAL JOURNAL OF REMOTE SENSING, 2016
VOL. 38, NO. 2, 514–536
http://dx.doi.org/10.1080/01431161.2016.1266059

Taylor & Francis
Taylor & Francis Group

Scene classification based on a hierarchical convolutional sparse auto-encoder for high spatial resolution imagery

Xiaobing Han, Yanfei Zhong, Bei Zhao and Liangpei Zhang

State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, P.R. China; State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing, P.R. China; Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, P.R. China

ABSTRACT

Efficiently representing and recognizing the semantic classes of the subregions of large-scale high spatial resolution (HSR) remote-sensing images are challenging and critical problems. Most of the existing scene classification methods concentrate on the feature coding approach with handcrafted low-level features or the low-level unsupervised feature learning approaches, which essentially prevent them from better recognizing the semantic categories of the scene due to their limited mid-level feature representation ability. In this article, to overcome the inadequate mid-level representation, a patch-based spatial-spectral hierarchical convolutional sparse auto-encoder (HCSAE) algorithm, based on deep learning, is proposed for HSR remote-sensing imagery scene classification. The HCSAE framework uses an unsupervised hierarchical network based on a sparse auto-encoder (SAE) model. In contrast to the single-level SAE, the HCSAE framework utilizes the significant features from the single-level algorithm in a feedforward and full connection approach to the maximum extent, which adequately represents the scene semantics in the high level of the HCSAE. To ensure robust feature learning and extraction during the SAE feature extraction procedure, a 'dropout' strategy is also introduced. The experimental results using the UC Merced data set with 21 classes and a Google Earth data set with 12 classes demonstrate that the proposed HCSAE framework can provide better accuracy than the traditional scene classification methods and the single-level convolutional sparse auto-encoder (CSAE) algorithm.

ARTICLE HISTORY
Received 28 January 2016
Accepted 17 November 2016

创新点

HCSAE框架最大限度地利用了单层算法中的重要特征，采用前馈和全连接方式，这充分地表示了HCSAE高层中的场景语义。

1. Introduction

In recent years, the rapid development of remote-sensing techniques has provided us with a huge volume of high spatial resolution (HSR) remote-sensing images, which provide more abundant detail and structural information compared with the traditional remote-sensing images (Myint et al. 2011; Batista and Haertel 2010). The traditional HSR image classification methods have developed from pixel-based to object-based

CONTACT Yanfei Zhong zhongyanfei@whu.edu.cn State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, P.R. China

methods, which can obtain satisfactory classification results using the low-level features, such as the spectral, textural, and geometrical features (Blaschke et al. 2014). These methods can successfully classify land objects, such as road, building, and vegetation, but they cannot obtain the semantic information or classes, e.g. residential and commercial areas, because of the so-called 'semantic gap' between the low-level objects and the high-level semantic meanings (Lienou, Maitre, and Datcu 2010). HSR remote-sensing imagery scene classification is a challenging problem because of the similar ground object categories with complicated compositions and distributions. For instance, residential areas and commercial areas both often contain buildings, roads, and trees, but they differ in their spatial layouts and structural organizations.

To efficiently and effectively represent and recognize complex HSR scenes, different scene classifiers for HSR imagery have been proposed to bridge the semantic gap (Cheriyadat 2013; Hu et al. 2015; Zhang and Du 2015a, 2015b; Zhao, Zhong, and Zhang 2013; Zhao, Zhong, Xia, et al. 2016; Zhao, Zhong, and Zhang 2016; Zhao, Zhong, and Zhang et al. 2016). However, most of the existing HSR scene classification methods extract the mid-level features based on the spectral and spatial low-level features. These methods include the bag-of-words (BoW) model (Cheng et al. 2013), the spatial pyramid matching (SPM) model (Grauman and Darrell 2005), the locality-constrained linear coding (LLC) model (Wang et al. 2010), probabilistic latent semantic analysis (PLSA) (Zhong et al. 2015), and the latent Dirichlet allocation (LDA) topic model (Blei, Ng, and Jordan 2003; Zhao et al. 2016a). The BoW model is an intermediate feature representation method, and the core idea of the BoW model is to represent the image as a histogram of visual word occurrences by mapping the local features (i.e. SIFT (scale invariant feature transformation)) to a visual vocabulary generated with a clustering method. The topic model is a hierarchical probabilistic model based on the visual word distribution of the topic allocations, and it considers each image as a finite mixture of an intermediate set of topics to summarize the scene information.

Although these scene classification methods can all obtain satisfactory results, they utilize *k*-means clustering to transform the patches of the scene images into visual words, which leads to high correlation of the codebook, and they do not consider the spatial relationships of the clustering centres. Meanwhile, all these methods import the attributes of the scene images (the spectral, spatial, and geometrical features) either separately or jointly before the *k*-means clustering, which may cause information loss when applied to complicated scenes.

To overcome the drawbacks of the traditional scene classification methods, and to fully and automatically utilize the spectral and spatial information of the image patches, we propose the patch-based spatial-spectral unsupervised hierarchical convolutional sparse auto-encoder (HCSAE) scene classifier based on deep learning for HSR remote-sensing imagery. Deep learning is an automatic feature extraction processing framework with superior discriminatory power for image representation, which simulates the hierarchical information processing mechanism of the human brain (Hinton and Salakhutdinov 2006; LeCun, Bengio, and Hinton 2015; Bengio 2009; Bengio, Courville, and Vincent 2013; Krizhevsky, Sutskever, and Hinton 2012; Simard, Steinkraus, and Platt 2003; Vincent et al. 2010). Convolutional neural networks (CNNs), stacked auto-encoder, and the sparse auto-encoder (SAE) are all efficient feature representation models. The SAE, as a representative feature extraction method in deep learning, has been applied in

several fields such as image processing and scene classification (Cao, Huang, and Sun 2016; Wang et al. 2016; Zhang, Du, and Zhang 2014; Penatti, Nogueira, and dos Santos 2015; Shin et al. 2013). The SAE is a reconstruction-oriented dimensionality reduction and feature extraction method, where the hidden units conserve enough significant features that can represent the input. The SAE local features contain enough orientation and structural information to represent the local part of the input images. However, the traditional convolutional sparse auto-encoder (CSAE) model is a single-level feature learning algorithm, which only extracts the local features and aggregates the information of the input images corresponding to the SAE features once, and the abundant information contained in the pooled feature maps of the single-level CSAE may not be fully utilized.

In the HCSAE, unlike the previous SAE algorithms, to better represent the input images, a convolution mechanism, namely the CSAE, is combined with the input images to better utilize the robust SAE local features. In addition, the HCSAE design features a two-layer CSAE, including a low-level CSAE and a high-level CSAE, where the SAEs in the two levels act as the feature extractors, and it is concatenated with a feedforward full connection approach. CNNs represent the supervised label information of the input image with the back-propagation mechanism, where the convolutional kernels are learnt from the input images. Unlike the supervised CNN information backpropagation mechanism, for HCSAE, the features utilized for convolution are extracted from the local patches from the original input images with the SAE, and the connection manner from the original input images and the latter-level pooled feature maps are carried out in a feedforward full connection manner, which can convey the information from the original input images to the final pooled feature maps to the maximum extent. Thus, this particular network design of the HCSAE can help generate a better classification performance. To further improve the calculation, and to reduce the overfitting of the SAE, a 'dropout' strategy is introduced to further improve the scene classification performance. The main contributions of this article are as follows.

(1) The unsupervised HCSAE model. Unlike the traditional SAE method, the CSAE appends the feature coding procedure via a convolution mechanism after the simple SAE feature learning and feature extraction procedures. Considering the limited feature description ability of the single-level CSAE, the HCSAE develops an unsupervised hierarchical feature learning and feature coding framework for the CSAE by extracting and utilizing the higher-level features of the original HSR remote-sensing images. 创新点
(2) The patch-based spatial-spectral feature learning algorithm. The proposed HCSAE considers the patch-based spatial-spectral features by opening a window on the image containing the spatial correlation information and by stretching all the pixels in the spatial window into a long vector. The long spatial-spectral vectors are then imported into the SAE to learn the spatial-spectral features automatically.
(3) The robust feature extraction dropout method. The 'dropout' technique is performed by stochastically 'dropping out' a number of input units or a number of hidden units to increase the stochastic properties of the SAE. This realizes the objective of reducing the overfitting of the SAE and results in robust features.

Compared with the traditional scene classification algorithms and the single-level CSAE algorithm, the experimental results using the 21-class UC Merced data set and a 12-class Google Earth data set demonstrate that the proposed HCSAE performs better at HSR remote-sensing image scene classification.

2. The SAE

The emergence of deep learning has been inspired by recent work in neuroscience (Lee et al. 1998). Deep learning is an efficient feature extraction framework that has achieved great success in natural image-processing areas, and has been used to help solve many problems in both industrial application areas and academic research areas (LeCun, Bengio, and Hinton 2015).

The SAE is a representative deep learning model for unsupervised feature extraction methods and unlabelled data sets, and consists of encoding and decoding stages (Ng 2010; Shin et al. 2013). During the encoding and decoding stages, the data sets are first encoded into hidden units and then decoded into the reconstructed data. Similar to the auto-encoder (AE) (Hinton and Salakhutdinov 2006), the SAE is a symmetric and reconstruction-oriented network. However, there are two major differences between the SAE and the AE. The first major difference is that the basic construction blocks of the AE are restricted Boltzmann machines (RBMs) with a stacked mechanism, and the basic construction blocks of the SAE are simply reconstruction-oriented symmetric networks. The second major difference is that the SAE has a sparse constraint on the hidden unit activations to make the network structure sparse. Therefore, the working mechanisms of the AE and the SAE differ according to the basic construction blocks and the different cost functions and solving procedures. The AE solves the reconstruction-oriented problems mostly with global features, whereas the SAE can address both global features and local features. By minimizing the reconstruction error and optimizing the network with the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm (Liu et al. 1989), the local representative features are extracted with the SAE.

SAE是一种具有代表性的无监督特征提取方法和无标记数据集的深度学习模型，包括编码和解码阶段。

AE主要利用全局特征解决面向重构的问题，而SAE可以同时处理全局特征和局部特征。

Based on the local feature extraction properties of the SAE, and the stationary properties of HSR remote-sensing images, the features extracted by the SAE in one part of the image can be applied to other parts of the image. Therefore, the SAE adopts patches as the network input. The previous scene classification methods (e.g. BoW, SPM, and the LDA topic model) extract the image features (e.g. the spectral, structural, and textural features) either separately or jointly from the image patches, which may cause information loss. The SAE adopts a different approach by automatically taking the spectral and spatial information into consideration in an integral manner. Figure 1 shows the network structure of the SAE.

3. The HCSAE

To overcome the problem of the insufficient information utilization in the single-level CSAE, and to further improve the scene classification accuracy, the HCSAE is proposed in an unsupervised and hierarchical manner in this article. The HCSAE adds a convolutional step to enhance the SAE, namely the CSAE. The SAE is an efficient feature extraction algorithm that can mine the intrinsic spatial-spectral features from the original image. The

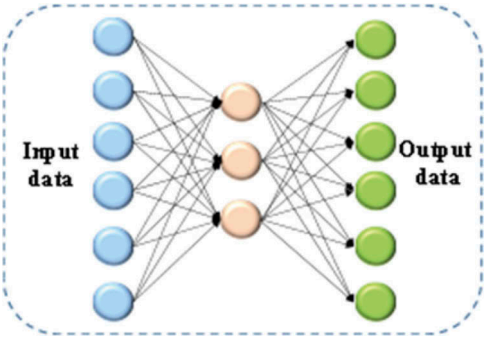

Figure 1. The network structure of the SAE.

CSAE is a combinatorial algorithm, which combines the advantages of the unsupervised SAE feature extraction ability with the efficient convolutional feature representation ability to realize the automatic representation of the scene image spatial-spectral information. After the convolution, the representative features extracted by the SAE are a response to the specific orientation and structural information of the original input image, and the convolved feature maps contain significant and desirable information for latter representation. To conserve the information of the original scene image, and considering the limited representative power of the low-level CSAE, the HCSAE makes a full connection between the features and the subsequent convolved feature maps and passes information from the low-level CSAE to the high-level CSAE in an unsupervised manner.

Considering that the pooled feature maps of the low-level CSAE contain abundant detail and structural information, the SAE feature extraction procedure is first performed on patches sampled from the pooled feature maps. Convolution between the features extracted by the SAE and the pooled feature maps from the low-level CSAE is then performed. The convolution in the high-level CSAE acts between the complicated SAE features and the pooled feature maps in the low-level CSAE, and thus the convolved feature maps in the high-level CSAE contain more representative and significant features than the low-level CSAE. The HCSAE can be divided into three stages: (1) the low-level CSAE; (2) the high-level CSAE; and (3) softmax classification. The low-level CSAE and the high-level CSAE can be subdivided into three substages: (1) SAE feature extraction; (2) convolution; and (3) pooling. The details of the HCSAE are described next.

HCSAE可分为三个阶段:
(1)低水平CSAE;
(2)高层CSAE;
(3) softmax分类

3.1. *The low-level CSAE*

The low-level CSAE is based on the local features extracted from the original scene images. In the complete procedure of the HCSAE, the low-level CSAE represents the original scene images only once, which results in information loss and coarse feature representation in the low-level CSAE.

低级CSAE从原始场景中提取局部特征

3.1.1. *Feature extraction of the low-level CSAE: the patch-based spatial-spectral feature learning approach*

In the low-level CSAE, the feature extraction procedure is executed by the SAE, which consists of two steps: encoding and decoding. Unlike the whole image-based process in

the AE, the SAE adopts a patch-based spatial-spectral feature learning approach. The patch-based spatial-spectral feature learning approach obtains the spatial-spectral information by opening a window in the image and concatenating the spectral information of each pixel in the window into a long vector. The advantage of the patch-based spatial-spectral feature learning approach is that the concatenated spatial-spectral information can be learnt by the SAE rather than extracted either separately or jointly, which can conserve the information of the original patches to a great extent.

During the encoding step, the input patches are first sampled stochastically from the original scene images. Normalization and zero-phase (ZCA) whitening are then consecutively performed on the patches. The input patch vector $x \in \mathrm{R}^N$ is processed by a nonlinear logistic sigmoid function $g(x)$.

$$\boldsymbol{a} = f(\boldsymbol{x}) = \mathrm{g}(\mathbf{W}_{\mathrm{enl}}\boldsymbol{x} + \boldsymbol{b}_{\mathrm{enl}}). \quad (1)$$

编码阶段公式

$$\mathrm{g}(\boldsymbol{x}) = (1 + \exp(-x))^{-1}. \quad (2)$$

During the decoding step, a linear activation function is utilized for the training to make the SAE model more robust.

$$\boldsymbol{z} = \mathbf{W}_{\mathrm{del}}\boldsymbol{a} + \boldsymbol{b}_{\mathrm{del}}, \quad (3)$$

解码

where $\mathbf{W}_{\mathrm{enl}} \in \mathrm{R}^{K\times N}$ and $\mathbf{W}_{\mathrm{del}} \in \mathrm{R}^{N\times K}$ are weight matrices with K features, and $\boldsymbol{b}_{\mathrm{del}} \in \mathrm{R}^K$ and $\boldsymbol{b}_{\mathrm{enl}} \in \mathrm{R}^N$ are the encoding and decoding biases, respectively. $\boldsymbol{a}$ represents the hidden layer neurons' activation value, which is utilized as the input of the decoding step. In the SAE, tied weight matrices $\mathbf{W}_{\mathrm{enl}} = \mathbf{W}_{\mathrm{del}}^{\mathrm{T}}$ ensure that the network is symmetric.

After the encoding and decoding steps, the input data and the reconstruction data are approximately equivalent. Thus, the feature extractors in the data set are learnt by minimizing the reconstruction error of the cost function in Equation (4) with the L-BFGS optimization algorithm (Larochelle et al. 2009). After the SAE feature extraction, the dimensionality of the features utilized for the feature coding is usually increased.

利用L-BFGS优化算法(Larochelle et al. 2009)最小化式(4)中代价函数的重构误差，学习数据集中的特征提取器

$$J_{\mathrm{sparse}}(\boldsymbol{X},\boldsymbol{Z}) = \frac{1}{2}\sum_{i=1}^{m}\left\|\boldsymbol{x}^i - \boldsymbol{z}^i\right\|^2 + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(\mathbf{W}_{ji}^{(l)}\right)^2 + \beta\sum_{j=1}^{s_2}(\mathrm{KL})(\rho\|\hat{\rho}_j), \quad (4)$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ represent the input data and reconstructed data vectors, respectively; m is the number of samples for training; λ is the weight decay parameter; β is the sparse constraint coefficient; n_l represents the layer number of the network; s_l represents the unit number of layer l, which does not include the basis number; s_{l+1} represents the unit number of layer $l-1$, which does not include the basis number; and s_2 represents the unit number in the hidden layer; J_{sparse} is the cost function of SAE. Supposing that $\hat{\rho}_j = \frac{1}{m}\sum_{i=1}^{m}\left[a_i(x_i)\right]$ is the average activation of the hidden unit i averaged over the training data set, then, to approximately keep the sparse constraint $\hat{\rho}_j = \rho$, the Kullback–Leibler(KL) divergence is utilized (Lee, Ekanadham, and Ng 2007), where ρ is the sparsity parameter, which is typically a small value close to zero.

In the SAE feature extraction step, a dropout strategy is added to further improve the computational efficiency and reduce the overfitting of the proposed algorithm. The

robust dropout technique works by stochastically leaving out both a number of input units and a number of hidden units at a certain probability in the feedforward procedure of the SAE to make the network more robust to stochastic noise, where the 'dropped out' units are regarded as non-existing zero units. The probabilities of the dropped-out input units and hidden units are set as 20% and 50%, respectively (Nitish 2013). After the SAE feature extraction, the local features will contain abundant detail and structural information.

3.1.2. *Convolutional feature representation of the low-level CSAE*

During the SAE feature extraction step, the input patches $\boldsymbol{x} \in \mathrm{R}^N$ are transformed into a new feature representation by $\boldsymbol{y} = f(\boldsymbol{x}) \in \mathrm{R}^K$, utilizing the feature extraction function $f : \mathrm{R}^N \rightarrow \mathrm{R}^K$. Given that the dimensionality of an image is $n \times n$ with three channels (i.e.R, G, B), and the patch size of an image patch is $w \times w$, after computing the representative features $\boldsymbol{y} \in \mathrm{R}^K$ for all the patches $w \times w \in \mathrm{R}^N$ to convolve the input images, we can obtain a convolved feature map $(n - w + 1) \times (n - w + 1)$ with K channels. In this article, the stride size for the convolution in the $n \times n$ image is set to 1.

跟CNN类似

After the convolution, the convolved feature maps are the combination of the convolution responses of each of the original scene images corresponding to each of the SAE features. In this way, each of the convolved feature maps represents the response to each state of the SAE features. The convolved feature maps therefore contain abundant information.

3.1.3. *Pooling of the low-level CSAE*

跟CNN中的pooling类似 关于CNN中的pooling表示：（n-f+1）*（n-f+1）

After the convolution, the convolved images may have a very high dimensionality, which results in low computational efficiency and a huge storage volume. Considering the stationary properties of the convolved images, to reduce the computational cost and decrease the storage volume, a max-pooling strategy is introduced to select the significant and salient features of each convolved feature map (Boureau et al. 2010). Supposing that the stride in the pooling stage is s, then the convolved feature map produces a $((n - w + 1)/s) \times ((n - w + 1)/s) \times K$ pooling representation.

3.2. *The high-level CSAE*

It is noted that the pooled feature maps contain abundant detail and structural information in the low-level CSAE. To utilize the information in the pooled feature maps of the low-level CSAE adequately, feature extraction and feature encoding on the pooled feature maps of the low-level CSAE are undertaken. The high-level CSAE is therefore made up of a finer feature representation based on the pooled feature maps from the low-level CSAE. Similar to the low-level CSAE, the high-level CSAE can be divided into three stages: 1) SAE feature extraction; 2) convolution; and 3) pooling.

3.2.1. *Feature extraction of the high-level CSAE*

Similar to the SAE feature extraction procedure in the low-level CSAE, the input patches of the SAE feature extraction procedure adopt a similar patch-based spatial-spectral approach. However, unlike the low-level CSAE, the inputs of the high-level CSAE are

HCSAE是从低级CSAE的汇集特征图中取样。

$\boldsymbol{x}_{\mathrm{h}} \in \mathrm{R}^{\mathrm{K}}$, which are sampled from the pooled feature maps from the low-level CSAE. Similar to the low-level SAE feature extraction procedures, the encoding and decoding stages of the high-level SAE feature extraction use Equations (1) and (3), respectively. However, unlike the low-level CSAE, the inputs of high-level SAE for Equations (1) and (2) are $\boldsymbol{x}_{\mathrm{h}} \in \mathrm{R}^{\mathrm{K}}$, and the weight and basis of Equations (1) and (3) are $\mathbf{W}_{\mathrm{enh}}$, $\boldsymbol{b}_{\mathrm{enh}}$ and $\mathbf{W}_{\mathrm{deh}}$, b_{deh}, respectively. Similar to the SAE cost function (Equation (4)) in the low-level CSAE, the cost function of the SAE in the high-level CSAE takes a similar approach, as is shown in Equation (5). Unlike Equation (4), the high-level SAE cost function is redefined in Equation (5).

$$J_{\mathrm{h\_sparse}}(\boldsymbol{X}_{\mathrm{h}}, \boldsymbol{Z}_{\mathrm{h}}) = \frac{1}{2}\sum\nolimits_{i_1=1}^{m_1} \left\| \boldsymbol{x}_{\mathrm{h}}{}^{i_1} - \boldsymbol{z}_{\mathrm{h}}{}^{i_1} \right\|^2 + \frac{\lambda_1}{2} \sum_{l_1=1}^{n_{l_1}-1} \sum_{i_1=1}^{s_{l_1}} \sum_{j_1=1}^{s_{l_1+1}} \left(\mathbf{W}_{\mathrm{h},j_1 i_1}^{(l_1)}\right)^2 + \beta_1 \sum_{j_1=1}^{s_{2_1}} (\mathrm{KL})(\rho_h || \hat{\rho}_{\mathrm{h},j_1}), \tag{5}$$

where $\boldsymbol{X}_{\mathrm{h}}$ and $\boldsymbol{Z}_{\mathrm{h}}$ represent the input data and reconstructed data, respectively. The corresponding parameters in Equations (4) and (5) have similar meanings. The dropout technique is performed on the SAE in the high-level CSAE in the same way as in the low-level CSAE.

在高级CSAE中的dropout方法与低级CSAE中类似

Unlike the SAE feature extraction in the low-level CSAE, the input patch of the high-level CSAE $\boldsymbol{x}_h \in \mathrm{R}^{\mathrm{K}}$ is transformed into a new feature representation by $f : \mathrm{R}^K \rightarrow \mathrm{R}^M$.

3.2.2. *Convolutional feature representation of the high-level CSAE*

高级CSAE的卷积特征表示

In the high-level CSAE feature extraction stage, the features $\mathbf{y} \in \mathrm{R}^{\mathrm{M}}$ extracted by the SAE are robust and representative. However, the pooled feature maps from the low-level CSAE cannot be directly represented by the local features extracted by the SAE in the high-level CSAE. To better represent the pooled feature maps from the low-level CSAE, convolution is utilized after the SAE feature extraction.

From the pooling stage in the low-level CSAE, the pooled feature maps are $((n-w+1)/s) \times ((n-w+1)/s)$, and they now work as the input images of the high-level CSAE. As mentioned before, the size of the local features for all the patches is $w_1 \times w_1$. Supposing that the convolution stride in the high-level CSAE is $s_1 = 1$, then, after the convolution, the convolved feature maps of the high-level CSAE are $((n-w+1)/s - w_1 + 1) \times ((n-w+1)/s - w_1 + 1)$ with M channels.

The convolved feature maps in the high-level CSAE are encoded twice from the original scene images by the different-level local features extracted by the SAE. Therefore, the convolved feature maps in the high-level CSAE contain more important and robust information than the convolved feature maps of the low-level CSAE.

3.2.3. *Pooling of the high-level CSAE*

To further improve the calculation efficiency of the final classification stage, a max-pooling strategy is conducted on the pooled feature maps in the high-level CSAE. Supposing that the pooling stride of the high-level CSAE is s_2, then the final pooled feature maps $(((n-w+1)/s - w_1 + 1)/s_2) \times (((n-w+1)/s - w_1 + 1)/s_2) \times M$ can

be obtained based on the convolved feature maps $((n-w+1)/s-w_1+1)\times((n-w+1)/s-w_1+1)\times M$.

3.3. *Softmax classification of the HCSAE*

In the final classification stage, the pooled feature maps $(((n-w+1)/s-w_1+1)/s_2)\times(((n-w+1)/s-w_1+1)/s_2)\times M$ from the high-level CSAE are imported into a softmax classifier [49]. The softmax logistic regression classifier is defined as follows:

$$J(\theta)=-\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k}\mathrm{I}(y^{(i)}=j)\ln\frac{e^{\theta_j^{\mathrm{T}}x^{(i)}}}{\sum_{l=1}^{k}e^{\theta_l^{\mathrm{T}}x^{(i)}}}\right]+\frac{\lambda}{2}\sum_{i=0}^{k}\sum_{j=0}^{n}\theta_{ij}^2, \tag{6}$$

where $I\{\cdot\}$ is an indicator function, and when the equation in $I\{\cdot\}$ is true, the value is 1, otherwise the value is 0. λ is the weight decay parameter used to balance these two terms, which is typically a positive parameter to make Equation (6) a convex function. The flow chart of the HCSAE is shown in Figure 2.

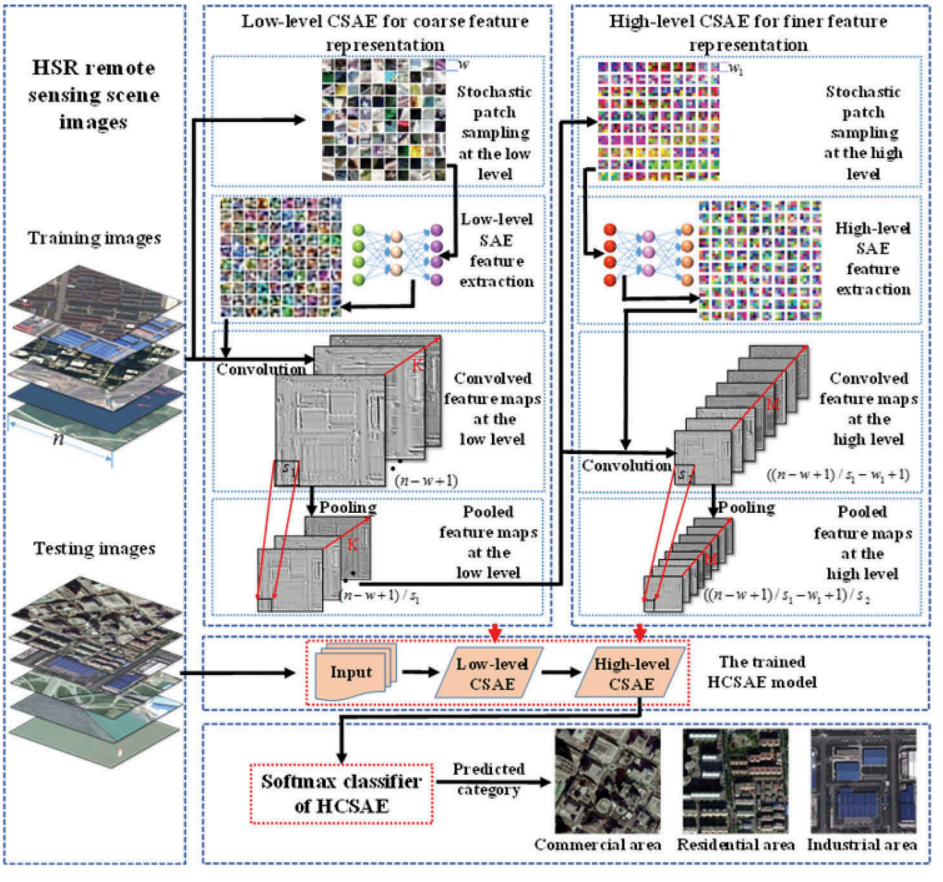

Figure 2. The flow chart of the HCSAE.

4. Experiments and analysis

To evaluate the proposed HCSAE for the HSR imagery scene classification, the widely utilized 21-class UC Merced data set and a 12-class Google data set were utilized for the final HSR imagery scene classification assessment. To demonstrate the effectiveness of the HCSAE, the single-level CSAE and the previous scene classification methods of BoW, SPM, LLC, LDA, and CSAE with saliency were compared. In addition, for the UC Merced data set and the Google Earth data set, the BoW model adopted a Support Vector Machine with radial basis function (SVM-RBF classifier with three-fold cross-validation to obtain parameters C and $)\gamma$. For the SPM model, an SVM-linear kernel classifier with $C = 300$ was adopted for the comparison. For the LLC and LDA models, the SVM-linear classifier was adopted with parameter $C = 300$.

4.1. *The experimental setups for the HCSAE-based scene classification*

To import the input images into the SAE, the input images were normalized between 0 and 1. To ensure the diversity of the patches on each input image, the patches were sampled stochastically. The patch size of the low level of the HCSAE was set as 10×10 pixels with three channels (i.e.R, G, B). For the low level of the HCSAE, the number of 10×10 patches sampled on the original input images was set as 1000 for each image. The dimensionality of the pooled feature maps in the low level of the HCSAE was $4 \times 4 \times K$, where K represents the channel number of the pooled feature maps. For the high level of the HCSAE, the number of 3×3 patches sampled on the pooled feature maps from the low level of the HCSAE was set as 4 for each pooled feature map.

4.2. *Experiment 1: 21-class UC Merced data set*

The first scene data set is the widely utilized 21-class UC Merced data set, which was manually extracted from large images from the USGS National Map Urban Area Imagery collection for various urban areas around the country (Yang and Newsam 2010). Some representative images for each category are shown in Figure 3. The UC Merced data set contains 21 scene categories with 100 scene images per category, and each scene image consists of 256×256 pixels with a spatial resolution of one foot per pixel. For this data set, we stochastically selected 80% of the scene images as the training set, and the remaining images were set as the test set. The scene classification results of the 21-class UC Merced data set are shown in Table 1.

Table 1 shows that the HCSAE performs extremely well with the UC Merced data set, and produces better results than the traditional scene classification methods. Compared with the single-level CSAE and the single-level CSAE with saliency, the HCSAE achieves an accuracy increase of 16% and 26%, respectively. The reason why the HCSAE can achieve such an increase in accuracy is that the unsupervised hierarchical approach preserves the significant information from the single-level CSAE, which can be utilized in the final classification. The HCSAE helps mine deeper-level feature representations on the foundation of the low level of the HCSAE. Figure 4 shows an overview of the confusion matrix for the 21-class UC Merced data set, where A1 represents Agriculture, A2 represents Airplane, B1 represents Baseball diamond, B2 represents Beach, B3

Figure 3. Representative images of the 21 land-use categories in the UC Merced data set: (*a*) Agriculture, (*b*) Airplane, (*c*) Baseball diamond, (*d*) Beach, (*e*) Buildings, (*f*) Chaparral, (*g*) Dense residential, (*h*) Forest, (*i*) Freeway, (*j*) Golf course, (*k*) Harbour, (*l*) Intersection, (*m*) Medium residential, (*n*) Mobile home park, (*o*) Overpass, (*p*) Parking lot, (*q*) River, (*r*) Runway, (*s*) Sparse residential, (*t*) Storage tanks, and (*u*) Tennis court.

Table 1. Scene classification results for the 21-class UC Merced data set.

Scene classification method	Classification accuracy (%)
BoW	72.05 ± 1.41
SPM (Zhao, Zhong, Xia, et al. 2016)	82.30 ± 1.48
LLC (Zhao, Zhong, and Zhang 2016)	82.85 ± 1.54
LDA (Zhao, Zhong, Xia, et al. 2016)	81.92 ± 1.12
Single-level CSAE with saliency (Zhang, Du, and Zhang 2014)	82.72 ± 1.18
Single-level CSAE	71.85 ± 0.52
GBRCN (Zhang, Du, and Zhang 2015)	94.53
Multiview deep learning (Lu et al. 2015)	93.48 ± 0.82
CNN with Overfeat feature (Marmanis et al. 2016)	92.4
CaffeNet (fine-tuning) (Castelluccio et al. arXiv 2015)	95.48
GoogLeNet (fine-tuning) (Castelluccio et al. arXiv 2015)	97.10
VGG-M (IFK) (Hu et al. 2015)	96.90
HCSAE	**97.14 ± 1.19**

represents Building, C represents Chaparral, D represents Dense residential, F1 represents Forest, F2 represents Freeway, G represents Golf course, H represents Harbour, I represents Intersection, M1 represents Medium residential, M2 represents Mobile home-park, O represents Overpass, P represents Parking lot, R1 represents River, R2 represents Runway, S1 represents Sparse residential, S2 represents Storage tanks, and T represents Tennis court. From Figure 4, it can be seen that certain categories are all correctly

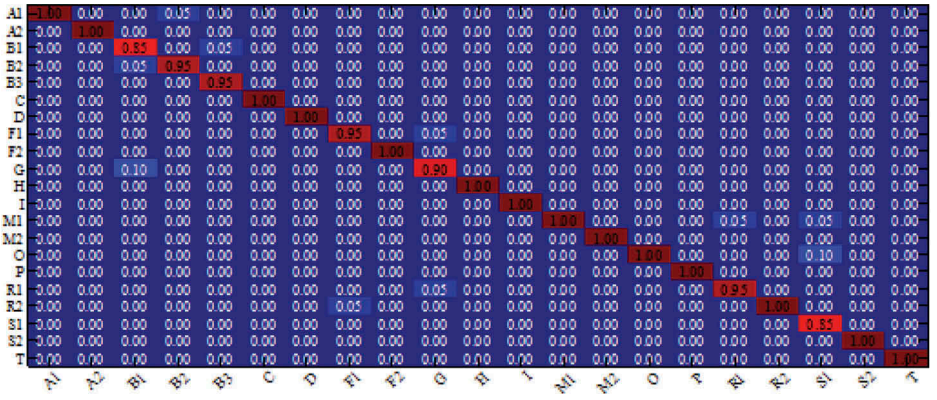

Figure 4. Confusion matrix for the HCSAE with the UC Merced data set, where A1 represents Agriculture, A2 represents Airplane, B1 represents Baseball diamond, B2 represents Beach, B3 represents Building, C represents Chaparral, D represents Dense residential, F1 represents Forest, F2 represents Freeway, G represents Golf course, H represents Harbour, I represents Intersection, M1 represents Medium residential, M2 represents Mobile homepark, O represents Overpass, P represents Parking lot, R1 represents River, R2 represents Runway, S1 represents Sparse residential, S2 represents Storage tanks, and T represents Tennis court.

classified, such as Airplane, Dense residential, Harbour, Intersection, and Overpass. However, some categories have a low classification accuracy, such as Agriculture misclassified as Forest because of the similar ground object composition.

By comparing the scene classification results of the different methods in Table 1, the differences between the HCSAE and the traditional deep learning-based methods become apparent. These differences can be explained from the following aspects, namely the HCSAE and CNNs, the HCSAE and the SAE, and the HCSAE and the stacked AE. 1) The difference between the HCSAE and CNNs can be summarized from two aspects. The first aspect is the network structure. The HCSAE is an unsupervised feature learning and representation algorithm that mainly adopts a feedforward and full connection approach between convolutional kernels and the subsequent convolved feature maps, which can conserve the information from the input image to the final pooled feature maps to the maximum extent. CNNs are supervised feature learning, extraction, and representation algorithms, which adopt the local connection and weight-sharing strategy from the input image and the final pooled feature maps, and adopt the back-propagation mechanism to adjust the network structures. The second aspect is the convolutional kernel extraction manner. The HCSAE learns the convolutional kernels from the local patches with the SAE, whereas CNNs extract the convolutional kernels from the whole input image by tuning the entire network structure. 2) The difference between the HCSAE and the SAE can be summarized as follows. The SAE is a reconstruction-oriented feature extractor and is constructed with the encoding and decoding stages with a sparse constraint. The SAE can be applied to both local features and the whole image, and it functions as one part of the HCSAE algorithm. The HCSAE is constructed with a two-level CSAE. Each level of the CSAE consists of SAE feature extraction, convolution, and pooling stages. 3) The difference between the HCSAE and the stacked AE can be summarized as follows. First, the HCSAE is constructed with two-

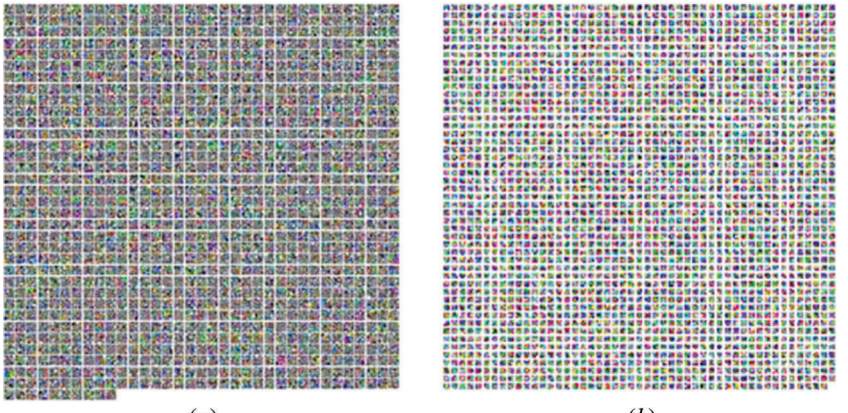

Figure 5. (*a*) and (*b*), respectively, represent the weight feature maps of the low level of the HCSAE and the high level of the HCSAE for the UC Merced data set.

level CSAE. Each level of the CSAE consists of SAE feature extraction, convolution, and pooling stages. In contrast, the stacked AE is constructed of AE, which usually adopts the pretraining and a fine-tuning mechanism. Second, the HCSAE usually represents an image with the convolution mechanism, whereas the stacked AE represents an image with the reconstruction-oriented feature representation procedures.

The weight feature maps of the low level of the HCSAE and the high level of the HCSAE are shown in Figure 5 to demonstrate what the SAE learns at the low level of the HCSAE and the high level of the HCSAE. From Figure 5, we can see that each square of the weight feature maps of the low level and the high level of the HCSAE represents a certain kind of edge and textural information. As the HCSAE is an unsupervised feature learning algorithm, the SAE extracts the features from all of the local patches together and each weight feature of the SAE algorithm is a combinatorial response, which reflects a certain kind of structural information.

Figure 6 shows some representative scene images correctly classified by the HCSAE algorithm whereas the single-level CSAE with saliency shows worse classification performance. By analysing Figures 4 and 6, it can be seen that the HCSAE algorithm shows a better classification performance for the scene classes of Airplane, Freeway, Golf course, Intersection, and Medium residential. For the Medium residential category, the scene images with more trees can be correctly classified whereas the single-level CSAE with saliency shows a better classification performance for only the Medium residential category. All these examples show that the HCSAE algorithm demonstrates a better classification ability for scene images with more complicated ground object compositions.

4.3. *Experiment 2: 12-class Google Earth data set*

The second data set was acquired from Google Earth, and mainly covers urban areas in China, with a spatial resolution of 2 m. The data set consists of 12 land-use classes, with

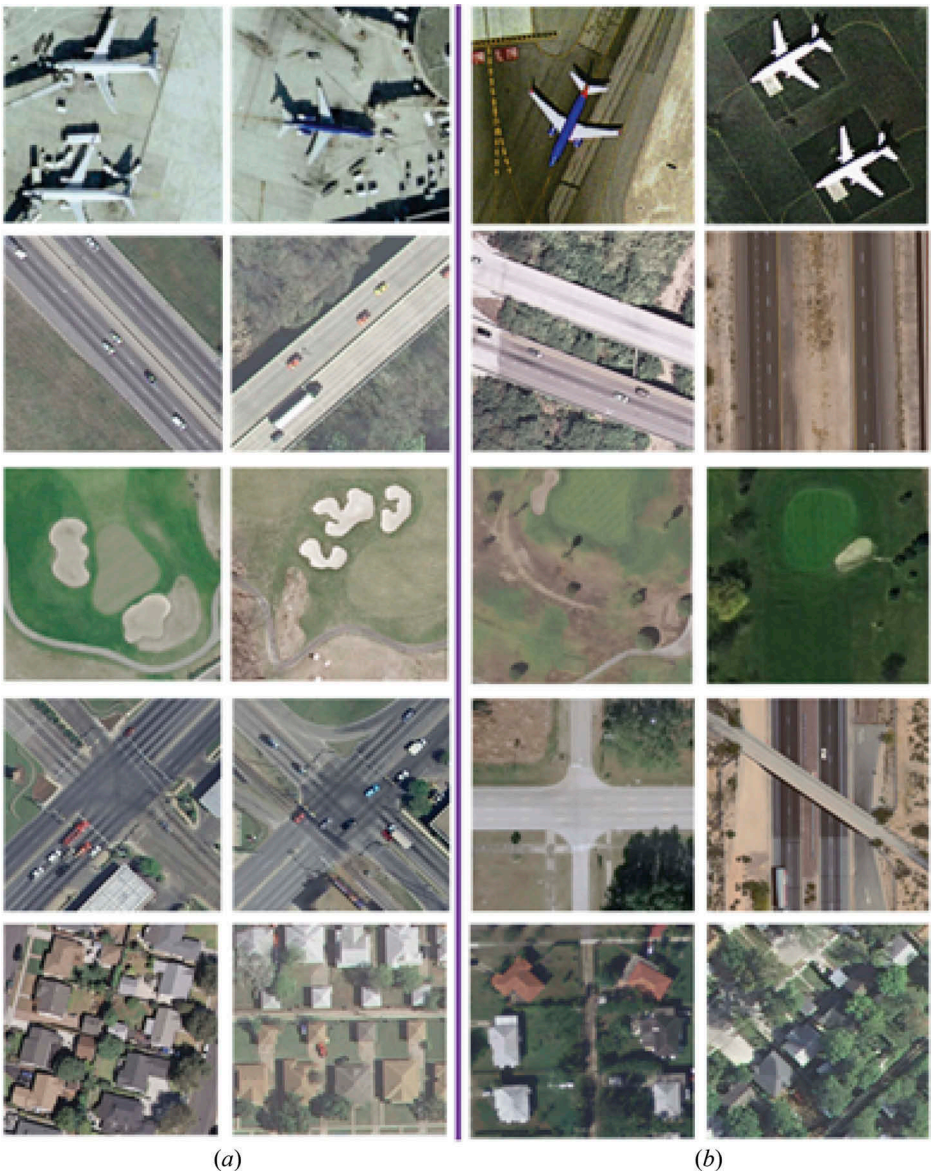

Figure 6. Some of the classification results of the single-level CSAE with saliency and the HCSAE. The first to fifth rows correspond to the scene classes of Airplane, Freeway, Golf course, Intersection, and Medium residential, respectively. (*a*) Correctly classified images for all of the strategies. (*b*) Images classified correctly by the HCSAE, but incorrectly classified by the single-level CSAE with saliency.

200 scene images per category. Each scene image has 200 × 200 pixels. In the experiments with this data set, we stochastically sampled 80% of the scene images as the training data set, and the rest were used as the test data set. Figure 7 demonstrates some representative images for each scene category.

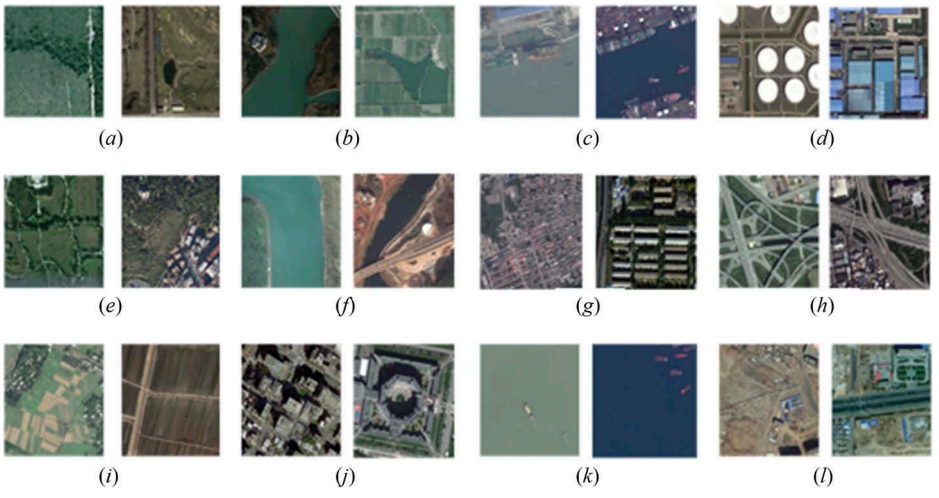

Figure 7. Representative images of the Google Earth data set: (*a*) Meadow, (*b*) Pond, (*c*) Harbour, (*d*) Industrial, (*e*) Park, (*f*) River, (*g*) Residential, (*h*) Overpass, (*i*) Agriculture, (*j*) Commercial, (*k*) Water, and (*l*) Idle land.

Table 2. Scene classification results for the 12-class Google Earth data set.

Scene classification method	Classification accuracy (%)
BoW	73.93 ± 1.41
SPM	80.26 ± 1.86
LLC	70.89 ± 1.49
LDA	66.85 ± 2.12
TF-CNN (Zhong, Fei, and Zhang 2016)	82.81
Single-level CSAE with saliency	74.84 ± 1.49
Single-level CSAE	66.13 ± 1.83
HCSAE	**86.32 ± 1.36**

The scene classification results for the Google Earth data set are shown in Table 2. From Table 2, it can be seen that the HCSAE achieves a better scene classification result than the traditional scene classification methods. Compared with the single-level CSAE with saliency and the single-level CSAE, the HCSAE achieves an increase in accuracy of 15% and 20%, respectively. The reason why the HCSAE can perform better than the single-level CSAE is that the HCSAE can mine finer feature representations based on the pooled feature maps from the single-level CSAE. Figure 8 gives an overview of the confusion matrix for the 12-class Google Earth data set, where A represents Agriculture, C represents Commercial, H represents Harbour, I1 represents Idle land, I2 represents Industrial, M represents Meadow, O represents Overpass, P1 represents Park, P2 represents Pond, R1 represents Residential, R2 represents River, and W represents Water. From Figure 8, it can be seen that the Agriculture and Commercial classes achieve a better classification accuracy. However, the Overpass, River, and Park classes are mainly misclassified as Idle land, Pond, and Meadow due to the similar ground object compositions. This can be explained by the fact that the scene images of these categories share similar ground objects.

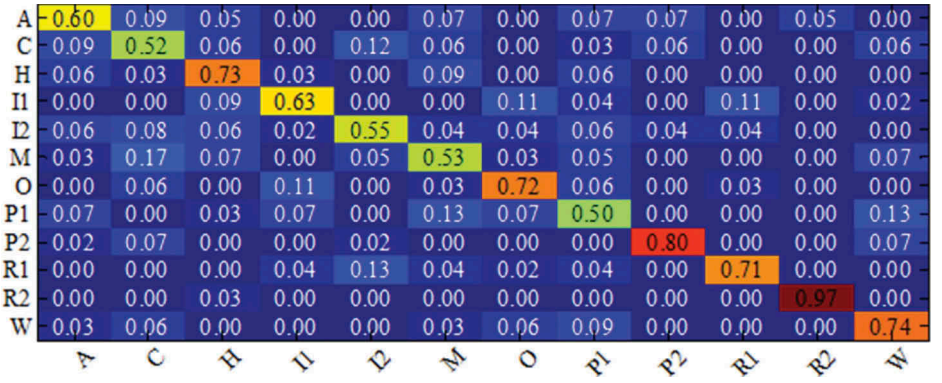

Figure 8. Confusion matrix for the HCSAE with the Google Earth data set, where A represents Agriculture, C represents Commercial, H represents Harbour, I1 represents Idle land, I2 represents Industrial, M represents Meadow, O represents Overpass, P1 represents Park, P2 represents Pond, R1 represents Residential, R2 represents River, and W represents Water.

Figure 9. (*a*) and (*b*), respectively, represent the weight feature maps of the low level of the HCSAE and the high level of the HCSAE for the Google Earth data set.

The weight feature maps of the low level of the HCSAE and the high level of the HCSAE for the Google Earth data set are shown in Figure 9 to demonstrate what the SAE learns at the low level of the HCSAE and the high level of the HCSAE. From Figure 9, we can see that each square of the weight feature maps of the low level and the high level of the HCSAE represents a certain kind of orientation information.

Figure 10 shows that some representative scene images on the left are classified correctly by the HCSAE and the single-level CSAE with saliency. The right column shows some results that are correctly classified by the HCSAE and misclassified by the single-level CSAE. Through a careful examination of Figure 10(b), it can be seen that the HCSAE algorithm can recognize the square roof shapes of the Industrial category, whereas CSAE with saliency shows a worse recognition ability. The HCSAE algorithm

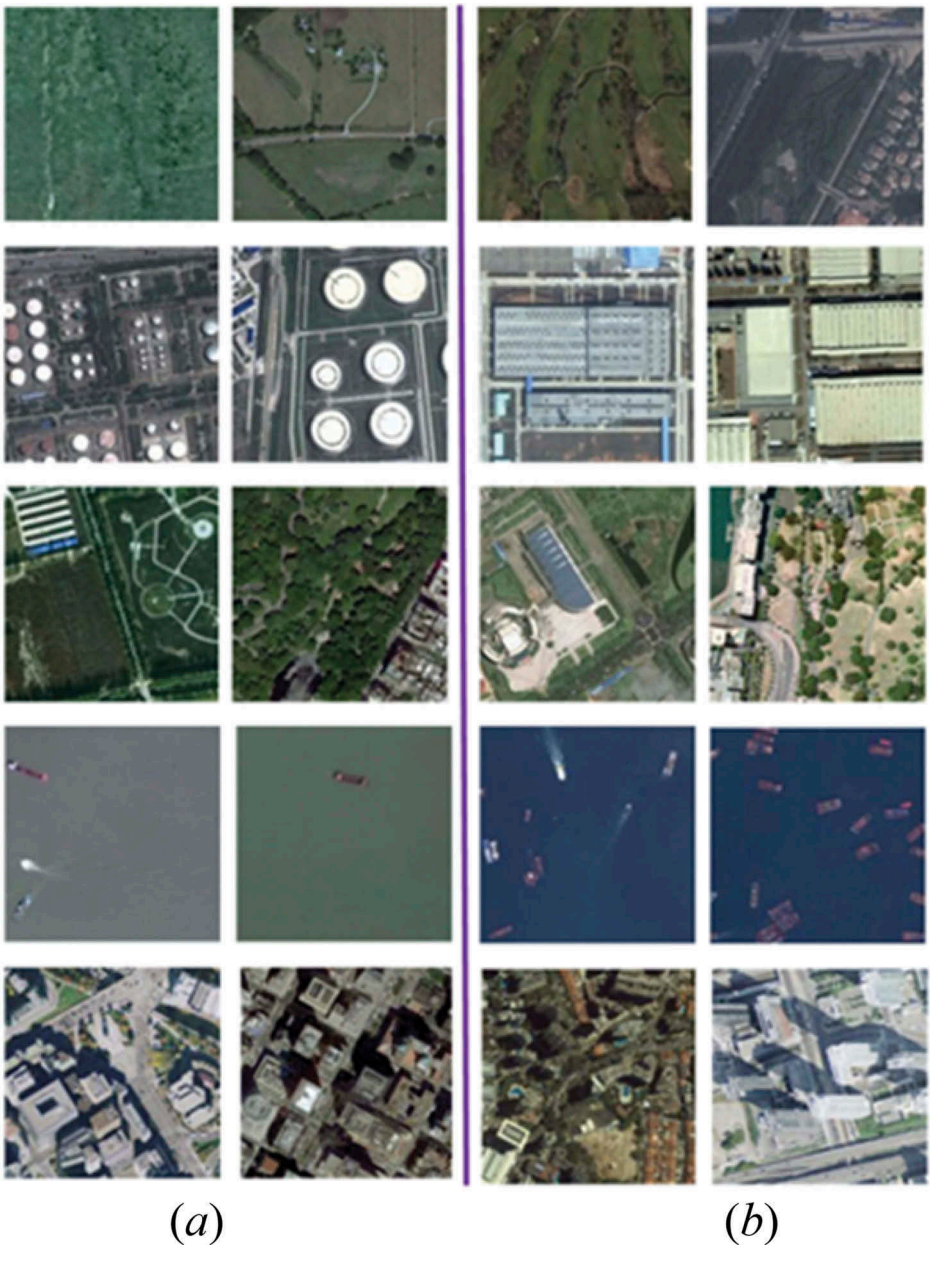

Figure 10. Some of the classification results of the single-level CSAE with saliency and the HCSAE with the Google Earth data set. The first to fifth rows correspond to the scene classes of Meadow, Industrial, Park, Water, and Commercial, respectively. (*a*) Correctly classified images for all of the strategies. (*b*) Images classified correctly by the HCSAE, but incorrectly classified by the single-level CSAE with saliency.

also shows a better classification performance on the Water category and the Commercial category. All these facts show that the HCSAE algorithm has a better classification performance than the single-level CSAE when faced with the complicated ground object composition, which can be ascribed to the finer feature representation ability of the HCSAE algorithm.

5. Sensitivity analysis

敏感性分析

5.1. *Sensitivity analysis of the SAE for the UC Merced data set*

隐藏层的单元数和SAE网络的稀疏性在HSR图像场景分类中扮演重要角色

The hidden unit number (HUN) and sparsity (SP) of the SAE network play a significant role in HSR imagery scene classification. According to previous work (Coates, Ng, and Lee 2011), the HUN is the main factor deciding the hidden network structure and influencing the properties of the SAE network. The SP is also a significant factor influencing the classification performance of the HCSAE. For the HCSAE algorithm, the SAE stages in both low-level CSAE and high-level CSAE adopt the same parameter setting approach. For the initial values of HUN and SP, please refer to the previous work (Zhang, Du, and Zhang 2014). For the optimal HUN and SP value, the parameter search is conducted in the two-dimensional range space and an element-wise traversing method is adopted. Based on the specific data structures and categories of the UC Merced data set, when HUN and SP value were set as 1200 and 0.5, respectively, the low-level CSAE achieved a satisfactory scene classification result. Based on fixed parameters in the low-level CSAE, when the HUN and SP value were set as 2400 and 0.7, respectively, the HCSAE achieved the best classification performance.

Figures 11(*a*) and (*b*) shows how the HUN and SP, respectively, influence the scene classification accuracy of the low-level CSAE. Figures 11(*a*)–(*d*) demonstrate how the hidden unit and SP, respectively, influence the scene classification accuracy of the high-

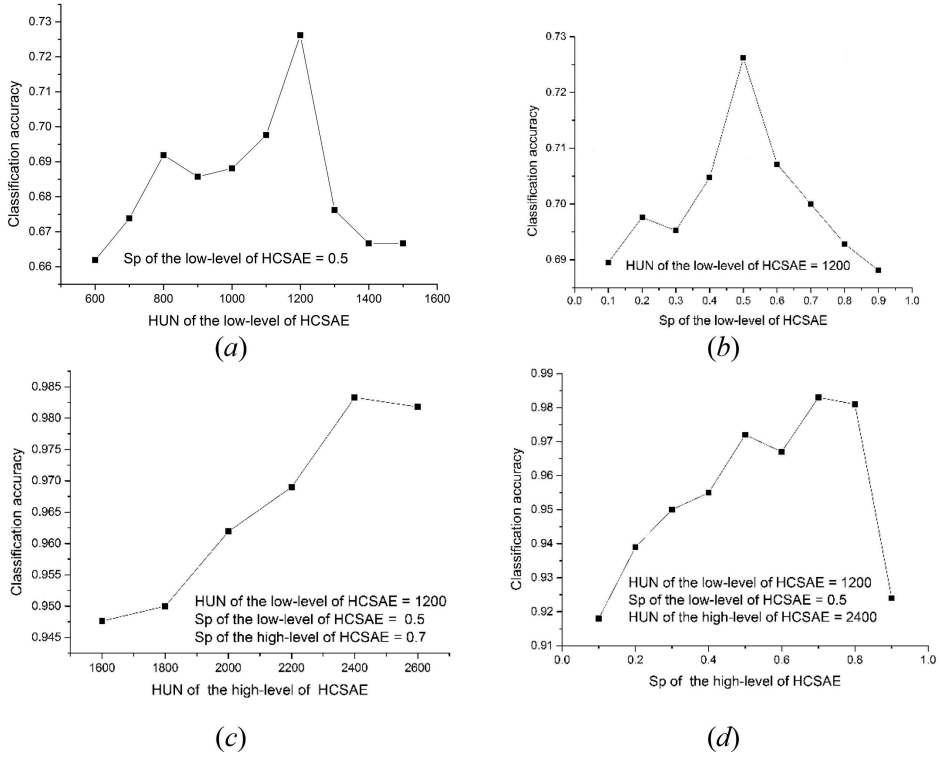

Figure 11. Parameter analysis for the hidden unit number and sparsity with the UC Merced data set, where HUN and Sp represent the hidden unit number and sparsity, respectively.

level CSAE. Using the above parameter sensitivity analysis, it can be seen that the optimum HUN in the high-level CSAE is about twice that of the low-level CSAE.

也就是说 high-level CSAE 的最优 要比low-level CSAE大概高2倍

5.2. *Sensitivity analysis of the SAE for the Google Earth data set*

The Google Earth data set has fewer categories than the UC Merced data set, but more samples in each category. Based on Coates, Ng, and Lee (2011), when the HUN and the SP value in the low-level CSAE were set as 900 and 0.7, respectively, the low-level CSAE obtained a satisfactory classification result. Based on fixed parameter settings in the low-level CSAE, the HCSAE achieved the best classification performance when the HUN and the SP value of the high-level CSAE were set as 2000 and 0.8, respectively.

Figures 12(*a*) and (*b*) shows how the HUN and SP, respectively, influence the scene classification accuracy of the low-level CSAE. By analysing the parameters of the low level and high level of the HCSAE in Figure 12, it can be seen that the optimum HUN in the high-level CSAE is about twice that of the low-level CSAE. Different values of the optimum HUNs in different levels of the HCSAE indicate that the SAEs in different levels have different feature extraction performances.

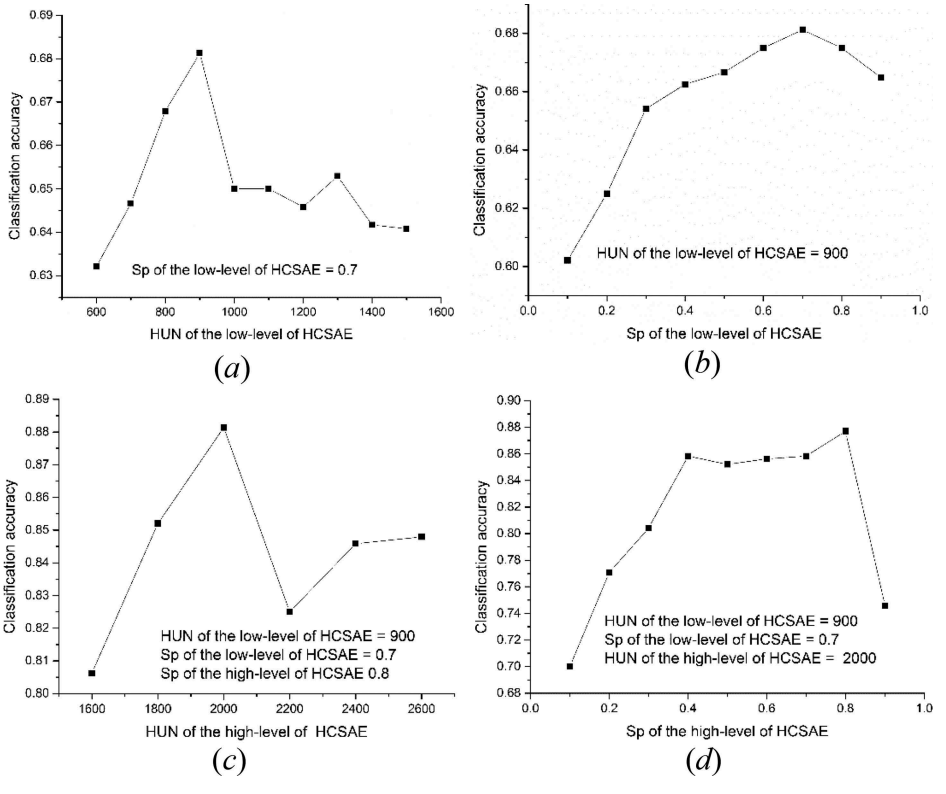

Figure 12. Parameter analysis for the hidden unit number and sparsity with the Google Earth data set, where HUN and Sp represent the hidden unit number and sparsity, respectively.

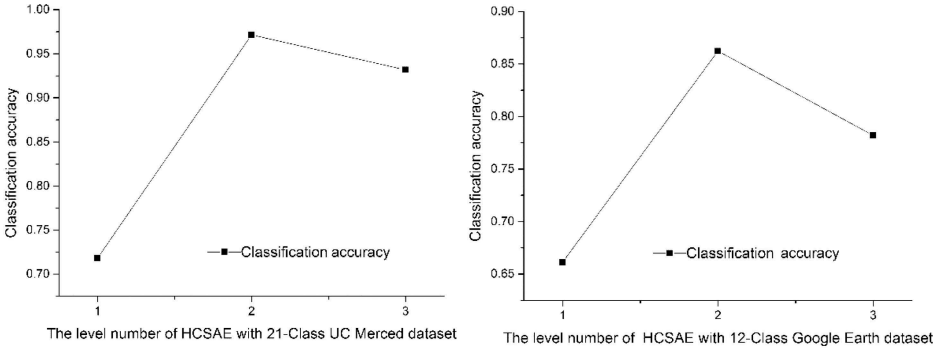

Figure 13. Parameter analysis for the level number of the HCSAE with the UC Merced data set and the Google Earth data set.

5.3. *Sensitivity analysis of the level number of the HCSAE for the UC Merced data set and the Google Earth data set*

UC Merced data set和谷歌Earth data set的HCSAE level number的敏感性分析

The influence of the number of levels in the HCSAE is shown in Figure 13. Figures 13(*a*) and (*b*) demonstrates the 21-class UC Merced data set and the 12-class Google Earth data set, respectively.

As shown in Figures 13(*a*) and (*b*), it can be seen that when the level number of the HCSAE is equal to 2, the best classification performance can be obtained for the 21-class UC Merced data set and the 12-class Google Earth data set. However, with the increase of the level number, the classification performances of these two data sets decrease, but this phenomenon is acceptable. For neural networks, with the increase of layer number, the classification performance may not always increase. For instance, the classification performance of a CNN may not always improve with the increase of the layer number, and more layers may cause gradient dispersion and error accumulation. For the HCSAE algorithm, the SAE feature extraction is a vital procedure, which directly influences the quality of the convolutional kernels. However, with the increase of level numbers, the dimensionality of the convolutional kernels becomes huge, which may introduce both abundant useful information and redundant information that influences the quality of the extracted SAE features. Thus, when the level number of the HCSAE algorithm is set to three, the decrease of the classification performances can still be accepted. From Figures 13(*a*) and (*b*), it can be seen that although the classification performance of the HCSAE decreases when the level number is equal to three, the classification accuracy of the three-level CSAE is still higher than the single-level CSAE.

6. Conclusions

This article has proposed a new hierarchical feature extraction and representation framework: the HCSAE for HSR remote-sensing imagery scene classification. The single-level CSAE extracts the features of the input data by the SAE only once, which may not sufficiently utilize the information in the input data. To solve the problem of insufficient utilization of information and feature representation in the single-level CSAE, the proposed HCSAE can solve the

problem by extracting higher-level features of the input data. The HCSAE extracts higher-level features by setting the output of the single-level CSAE as the input of the latter-level CSAE. Unlike the traditional approaches that utilize handcrafted low-level feature descriptors or unsupervised features, the HCSAE can extract the features of the patches automatically and integrally in an unsupervised feature extraction and feature representation manner. In addition, the dropout technique in the SAE feature extraction procedure is designed to increase the robustness of the features. The experimental results with two HSR remote-sensing imagery data sets showed that by processing the features of the input data automatically, the HCSAE has the potential to extract and learn high-level spectral and spatial features. In our future work, multi-scale features such as the convolutional kernels will be considered to extract the information of the input data for different scene images with different-scale features.

方法：通过设置单层CSAE的输出作为后一层CSAE的输入

dropout层为了增加SAE特征提取的程序的健壮性

在今后的工作中，我们将考虑多尺度特征，如卷积核，提取不同尺度特征的不同场景图像的输入数据信息。

第二篇就是从多尺度特征考虑的 原文为：Large patch convolutional neural networks for the scene classification of high spatial resolution imagery

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Natural Science Foundation of China [Grant Numbers: 41622107 and 41371344]; Natural Science Foundation of Hubei Province [Grant Number: 2016-29]; State Key Laboratory of Earth Surface Processes and Resource Ecology [Grant Number: 2015-KF-02].

References

Batista, M. H., and V. Haertel. 2010. "On the Classification of Remote Sensing High Spatial Resolution Image Data." *International Journal of Remote Sensing* 31 (20): 5533–5548. doi:10.1080/01431160903485786.

Bengio, Y. 2009. "Learning Deep Architectures for AI." *Foundations and Trends® in Machine Learning* 2: 1–127. doi:10.1561/2200000006.

Bengio, Y., A. Courville, and P. Vincent. 2013. "Representation Learning: A Review and New Perspectives." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35: 1798–1828. doi:10.1109/TPAMI.2013.50.

Blaschke, T., G. J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R. Queiroz Feitosa, et al. 2014. "Geographic Object-Based Image Analysis – Towards a New Paradigm." *ISPRS Journal of Photogrammetry and Remote Sensing* 87: 180–191. doi:10.1016/j.isprsjprs.2013.09.014.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

Boureau, Y.-L., F. Bach, Y. LeCun, and J. Ponce. 2010. "Learning Mid-Level Features for Recognition." *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)* 2559–2566.

Cao, L., W. Huang, and F. Sun. 2016. "Building Feature Space of Extreme Learning Machine with Sparse Denoising Stacked-Autoencoder." *Neurocomputing* 174: 60–71. doi:10.1016/j.neucom.2015.02.096.

Castelluccio, M., G. Poggi, C. Sansone, and L. Verdoliva. 2015. "Land Use Classification in Remote Sensing Images by Convolutional Neural Networks." Accessed 14 August 2015. http://arxiv.org/abs/1508.00092.

Cheng, G., L. Guo, T. Zhao, J. Han, H. Li, and J. Fang. 2013. "Automatic Landslide Detection from Remote Sensing Imagery Using a Scene Classification Method Based on Bovw and Plsa." *International Journal of Remote Sensing* 34: 45–59. doi:10.1080/01431161.2012.705443.

Cheriyadat, A. M. 2013. "Unsupervised Feature Learning for Aerial Scene Classification." *IEEE Transactions on Geoscience and Remote Sensing* 52: 439–451. doi:10.1109/TGRS.2013.2241444.

Coates, A., A. Y. Ng, and H. Lee. 2011. "An Analysis Of Single-layer Networks In Unsupervised Feature Learning." *Journal of Machine Learning Research: Workshop and Conference Proceedings* 15: 215–223.

Grauman, K., and T. Darrell. 2005. "The Pyramid Match Kernel Discriminative Classification with Sets of Image Features." *IEEE International Conference on on Computer Vision* 2: 458–1465.

Hinton, G. E., and R. R. Salakhutdinov. 2006. "Reducing the Dimensionality of Data with Neural Networks." *Science* 313: 504–507. doi:10.1126/science.1127647.

Hu, F., G. Xia, J. Hu, and L. Zhang. 2015. "Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery." *Remote Sensing* 7: 14680–14707. doi:10.3390/rs71114680.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. "Imagenet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems (NIPS)* 25: 1097–1105.

Larochelle, H., Y. Bengio, J. Louradour, and P. Lamblin. 2009. "Exploring Strategies for Training Deep Neural Networks." *Journal of Machine Learning Research* 10: 1–40.

LeCun, Y., Y. Bengio, and G. E. Hinton. 2015. "Deep Learning." *Nature* 521: 436–444. doi:10.1038/nature14539.

Lee, H., C. Ekanadham, and A. Y. Ng. 2007. "Sparse Deep Belief Net Model for Visual Area V2." *Advances in Neural Information Processing Systems* 20: 873–880.

Lee, T. S., D. Mumford, R. Romero, and V. A. F. Lamme. 1998. "The Role of the Primary Visual Cortex in Higher Level Vision." *Vision Research* 38: 2429–2454. doi:10.1016/S0042-6989(97)00464-1.

Lienou, M., H. Maitre, and M. Datcu. 2010. "Semantic Annotation of Satellite Images Using Latent Dirichlet Allocation." *IEEE Geoscience and Remote Sensing Letters* 7: 28–32. doi:10.1109/LGRS.2009.2023536.

Liu, D. C., and J. Nocedal. 1989. "On the Limited Memory BFGS Method for Large Scale Optimization." *Mathematical Programming* 45: 503–528. doi:10.1007/BF01589116.

Lu, F. P. S., B. P. Salmon, F. Van Den Bergh, and B. T. J. Maharaj. 2015. "Multiview Deep Learning for Land-Use Classification." *IEEE Geoscience Remote Sensing Letters* 12 (12): 2448–2452. doi:10.1109/LGRS.2015.2483680.

Marmanis, D., M. Datcu, T. Esch, and U. Stilla. 2016. "Deep Learning Earth Observation Classification Using Imagenet Pretrained Networks." *IEEE Geoscience and Remote Sensing Letters* 13 (1): 105–109. doi:10.1109/LGRS.2015.2499239.

Myint, S. W., P. Gober, A. Brazel, S. Grossman-Clarke, and Q. Weng. 2011. "Per-Pixel vs. Object-Based Classification of Urban Land Cover Extraction Using High Spatial Resolution Imagery." *Remote Sensing of Environment* 115 (5): 1145–1161. doi:10.1016/j.rse.2010.12.017.

Ng, A. 2010. "Sparse Autoencoder." CS294A Lecture Notes. Stanford University. http://www.stanford.edu/class/archive/cs/cs294a/cs294a.1104/sparseAutoencoder.pdf.

Nitish, S. 2013. "Improving neural networks with dropout." Ph.D. thesis, Univ. Toronto, Canada.

Penatti, O. A., K. Nogueira, and J. A. dos Santos.2015. "Do Deep Features Generalize from Everyday Objects to Remote Sensing and Aerial Scenes Domains?" In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, June 12, 44–51.

Shin, H., M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach. 2013. "Stacked Autoencoders for Unsupervised Feature Learning and Multiple Organ Detection in a Pilot Study Using 4D Patient Data." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (8): 1930–1943. doi:10.1109/TPAMI.2012.277.

Simard, P. Y., D. Steinkraus, and J. C. Platt. 2003. "Best Practices For Convolutional Neural Networks Applied To Visual Document Analysis." In *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, 958–962. Washington, DC: IEEE Computer Society.

Vincent, P., H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol. 2010. "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion." *Journal of Machine Learning Research* 11: 3371–3408.

Wang, J., J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. 2010. "Locality Constrained Linear Coding for Image Classification." 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 13–18, 3360–3367.

Wang, Y., Z. Xie, K. Xu, Y. Dou, and Y. Lei. 2016. "An Efficient and Effective Convolutional Auto-Encoder Extreme Learning Machine Network for 3-D Feature Learning." *Neurocomputing* 174: 988–998. doi:10.1016/j.neucom.2015.10.035.

Yang, Y., and S. Newsam. 2010. "Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification." Proceedings of 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, November 2–5, 270–279.

Zhang, F., B. Du, and L. Zhang. 2014. "Saliency-Guided Unsupervised Feature Learning for Scene Classification." *IEEE Transactions on Geoscience and Remote Sensing* 53: 2175–2184. doi:10.1109/TGRS.2014.2357078.

Zhang, F., B. Du, and L. Zhang. 2015. "Scene Classification via a Gradient Boosting Random Convolutional Network Framework." *IEEE Transactions on Geoscience and Remote Sensing* 54 (3): 1–10.

Zhang, X., and S. Du. 2015a. "A Linear Dirichlet Mixture Model for Decomposing Scenes: Application to Analyzing Urban Functional Zonings." *Remote Sensing of Environment* 169: 37–49. doi:10.1016/j.rse.2015.07.017.

Zhang, X., and S. Du. 2015b. "Semantic Classification of Heterogeneous Urban Scenes Using Intra-Scene Feature Similarity and Inter-Scene Semantic Dependency." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8 (5): 2005–2014. doi:10.1109/JSTARS.2015.2414178.

Zhao, B., Y. Zhong, G. Xia, and L. Zhang. 2016. "Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery." *IEEE Transactions on Geoscience and Remote Sensing* 54 (4): 2108–2123. doi:10.1109/TGRS.2015.2496185.

Zhao, B., Y. Zhong, and L. Zhang. 2013. "Scene Classification via Latent Dirichlet Allocation Using a Hybrid Generative/Discriminative Strategy for High Spatial Resolution Remote Sensing Imagery." *Remote Sensing Letters* 4: 1204–1213. doi:10.1080/2150704X.2013.858843.

Zhao, B., Y. Zhong, and L. Zhang. 2016. "A Spectral–Structural Bag-Of-Features Scene Classifier for Very High Spatial Resolution Remote Sensing Imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 116: 73–85. doi:10.1016/j.isprsjprs.2016.03.004.

Zhao, B., Y. Zhong, L. Zhang, and B. Huang. 2016. "The Fisher Kernel Coding Framework for High Spatial Resolution Scene Classification." *Remote Sensing*. doi:10.3390/rs8020157.

Zhong, Y., M. Cui, Q. Zhu, and L. Zhang. 2015. "Scene Classification Based on Multifeature Probabilistic Latent Semantic Analysis for High Spatial Resolution Remote Sensing Images." *Journal of Applied Remote Sensing* 9 (1): 095064. doi:10.1117/1.JRS.9.095064.

Zhong, Y., F. Fei, and L. Zhang. 2016. "Large Patch Convolutional Neural Networks for the Scene Classification of High Spatial Resolution Imagery." *Journal of Applied Remote Sensing* 10 (2). doi:10.1117/1.JRS.10.025006.