This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Adaptive Deep Sparse Semantic Modeling Framework for High Spatial Resolution Image Scene Classification

Qiqi Zhu, Yanfei Zhong, *Senior Member, IEEE*, Liangpei Zhang, *Senior Member, IEEE*, and Deren Li, *Senior Member, IEEE*

***Abstract*— High spatial resolution (HSR) imagery scene classification, which involves labeling an HSR image with a specific semantic class according to the geographical properties, has received increased attention, and many algorithms have been proposed for this task. The employment of the probabilistic topic model to acquire latent topics and the convolutional neural networks (CNNs) to capture deep features for representing HSR images has been an effective ways to bridge the semantic gap. However, the midlevel topic features are usually local and significant, whereas the high-level deep features convey more global and detailed information. In this paper, to discover more discriminative semantics for HSR images, the adaptive deep sparse semantic modeling (ADSSM) framework combining sparse topics and deep features is proposed for HSR image scene classification. In ADSSM, the fully sparse topic model and a CNN are integrated. To exploit the multilevel semantics for HSR scenes, the sparse topic features and deep features are effectively fused at the semantic level. Based on the difference between the sparse topic features and the deep features, an adaptive feature normalization strategy is proposed to improve the fusion of the different features. The experimental results obtained with four HSR image classification data sets confirm that the proposed method significantly improves the performance when compared with the other state-of-the-art methods.**

***Index Terms*— Adaptive, convolutional neural network (CNN), fusion, normalization, probabilistic topic model (PTM), scene classification.**

I. Introduction

With the rapid development of earth observation technologies, many millions of high spatial resolution (HSR) images are now available for a wide range of applications, such as image and video retrieval, urban functional analysis, and environmental monitoring. However, the low interclass disparity and high intraclass variability in the HSR imagery poses a big challenge for image classification and have resulted in a surge of interest in the remote sensing field [1], [2]. In order to recognize and analyze the ground objects from HSR images, a lot of research has been undertaken over the years. However, HSR images usually contain different object classes with diverse spatial distributions, e.g., roads, trees, and buildings. The high-level semantic concepts, e.g., a residential scene or an industrial scene, are difficult to acquire because of the so-called semantic gap between the low-level features and the high-level semantic meanings. Scene classification plays an important role, and various scene classification methods have been proposed. The bag-of-visual-words (BoVW) model has attracted a great deal of attention due to its simplicity and effectiveness [4]–[7]. Based on the BoVW model, some classical variants, e.g., the probabilistic topic model (PTM) and feature coding, have been developed to improve the ability to describe complex HSR images [8]–[15].

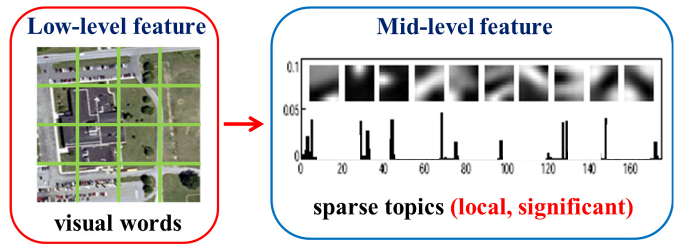


Fig. 1. Feature representation of the FSTM for HSR imagery.

The PTM, including the classical probabilistic latent semantic analysis (pLSA) and latent Dirichlet allocation (LDA) models, introduces latent variables to BoVW. For HSR image scene classification, single-feature-based PTM is usually inadequate to describe the HSR images, and multiple-feature-based PTM has been widely used to capture the scenes [11]–[15]. The LDA model treats the topic mixture parameters as variables drawn from a Dirichlet distribution, where the topic features are dense. This leads to feature redundancy and high time consumption. By representing the images with sparse and effective topic features, the fully sparse topic model (FSTM) solves these shortcomings and has been successfully applied to HSR scene classification [13]. However, as shown in Fig. 1, based on the low-level features of

Manuscript received December 7, 2017; revised March 18, 2018; accepted April 23, 2018. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0504202, in part by the National Natural Science Foundation of China under Grant 41622107 and Grant 41771385, and in part by the Natural Science Foundation of Hubei Province in China under Grant 2016CFA029. *(Corresponding author: Yanfei Zhong.)*

Q. Zhu is with the Faculty of Information Engineering, China University of Geosciences, Wuhan 430074, China (e-mail: zhuqiqi@whu.edu.cn).

Y. Zhong, L. Zhang, and D. Li are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China, and also with the Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China (e-mail: zhongyanfei@whu.edu.cn; zlp62@whu.edu.cn; drli@whu.edu.cn).

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

Digital Object Identifier 10.1109/TGRS.2018.2833293

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

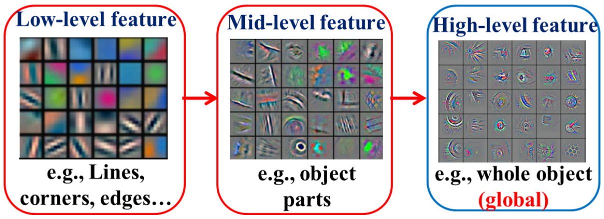


Fig. 2. CNN feature representation for HSR imagery.

the local visual words, the sparse topics of FSTM mainly focus on the local and significant information of HSR images. This results in the FSTM being unable to capture the features from the global perspective, and it loses the detailed information of the image. In addition, all these methods require a lot of prior information, as they are largely dependent on the selection of the handcrafted low-level features and the design of the midlevel feature representation. As such, these methods have limited transferability between different fields and data sets.

Deep learning is an automatic feature learning and representation framework, which extracts the features from the images in a joint spatial–spectral manner [16]–[18]. It has been successfully applied to objection detection [19], facial recognition [20], hyperspectral pixel classification and reconstruction [21]–[23], and video analysis [24]. Deep learning has also shown an impressive feature representation ability for HSR images, where the unsupervised feature learning method combining a sample selection strategy based on saliency was the first method to be introduced to HSR image scene classification [25]. Recently, by employing an end-to-end feature learning strategy to automatically extract intrinsic features without prior knowledge, deep convolutional neural network (CNN) architectures are quickly becoming prominent in remote sensing applications. A large amount of training samples is usually needed to train a CNN. However, it is difficult to obtain such a large HSR image data set. To address this issue, based on the generalizability of CNNs [26], [27], transfer learning, i.e., transferring the CNN model pretrained on a natural image data set, e.g., ImageNet [28], to the HSR image data set, has been proposed for HSR image scene classification [29]–[31].

CNNs discover features in multiple levels of representations, ranging from low-level information in the initial layers (e.g., lines, corners, and edges), to midlevel information in the intermediate layers (e.g., object parts), and high-level information (e.g., whole objects) in the final layers, as shown in Fig. 2. In general, the deep features can be extracted from different layers of the CNN and are then combined with a machine learning technique, such as the feature coding method and support vector machine (SVM), in the case of a classification setup [30]–[34]. For instance, Li *et al.* [31] combined the multilayer features from a pretrained CNN model with a multiscale improved fisher kernel (IFK) coding method. Hu *et al.* [30] encoded the CNN activations from the convolutional layer via different feature coding methods, such as the IFK and locality-constrained linear coding. Wang *et al.* [32] concatenated the convolutional features and the average pooled and normalized fully connected features as the final feature, where the convolutional features are encoded by the vector of locally aggregated descriptors and then reduced by the principal component analysis. CNN-based methods can be interpreted as extracting image descriptors (performed by the "convolutional layers") followed by pooling such features in a global image representation (performed by the "fully connected layers") [35]. In contrast to the PTM-based methods, such as the FSTM, the high-level features extracted from a CNN often convey more global information for HSR images and can discover the details of the objects. However, some critical and representative information cannot be highlighted to better discriminate the scenes.

In this paper, to solve the problem, the adaptive deep sparse semantic modeling (ADSSM) framework combining sparse topics and deep features for HSR image scene classification is proposed. In ADSSM, the FSTM and the CNN are first integrated to generate a more discriminative feature representation for HSR images. The FSTM is utilized to generate the midlevel handcrafted features, i.e., the mean and standard deviation (MSD)-based spectral feature, the wavelet-based texture feature, and the scale-invariant feature transform (SIFT) feature. The heterogeneous features are quantized separately to circumvent the hard assignment effect of k-means clustering. The sparse latent topics are then mined separately and concatenated as the final midlevel representation for the HSR images. To obtain the deep features, a pretrained CNN is employed, and the fully connected feature is extracted to be fused with the midlevel features. However, because of the difference between the midlevel and high-level features, directly concatenating the sparse topics and the deep features cannot improve the discriminative ability of the features. In ADSSM, an adaptive feature normalization and fusion strategy is proposed to transform the sparse latent topic-based midlevel features and the deep features. The transformed topic features are then effectively fused with the transformed deep features to acquire the final image representation. In this way, the scene labels of the images can be acquired based on the distinct information from the multilevel features.

The main contributions of this paper are as follows.

1) *ADSSM Framework:* The ADSSM framework is proposed to discover intrinsic and representative low-level features from the images. Based on the low-level visual information, three heterogeneous sparse topics and a deep scene feature are designed to describe the complex geometrical structures and spatial patterns of HSR images. The feature spaces of the different types of features are separately mined and adaptively fused to circumvent feature interruption in the feature learning process.
2) *Local and Global Feature Description by the FSTM and the CNN:* To capture discriminative semantics for the scenes, the midlevel-based FSTM and the high-level-based CNN scene classification methods are combined in ADSSM. The local and significant information is acquired from the FSTM, whereas the global and detailed information can be extracted from the CNN. The integration of the sparse topics and the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

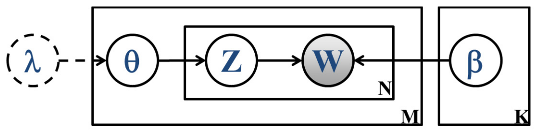


Fig. 3. Probabilistic graphical model of the FSTM.

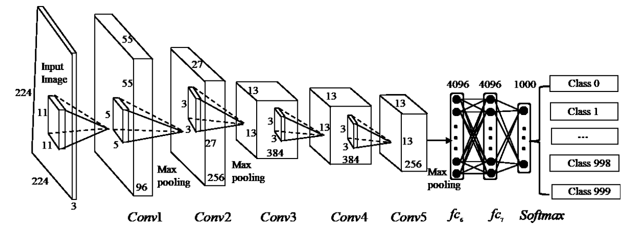


Fig. 4. Architecture of CaffeNet.

deep features provides a multilevel feature description for distinct scenes.

3) *Adaptive Distinct Feature Normalization Strategy:* To improve the fusion of the sparse topics and the deep features, an adaptive feature normalization strategy is proposed. In ADSSM, the mined sparse topics and deep features are adaptively and separately normalized to augment the significance of the representative features. Based on the adaptively fused feature description, the ADSSM framework can decrease the confusion of complex scenes.

Comprehensive evaluations on four distinct data sets, i.e., the 21-class UC Merced (UCM) data set, the 12-class Google data set of SIRI-WHU, the challenging 45-class NWPU-RESISC45 data set, and a new 20-class optical satellite remote sensing image (OSRSI20) data set, confirm the effectiveness of the ADSSM framework.

The rest of this paper is organized as follows. Section II discusses the related work. Section III provides details about the proposed ADSSM framework for HSR imagery scene classification. A description of the data sets and an analysis of the experimental results are presented in Section IV. Finally, the conclusions are drawn in Section V.

II. Background

The ADSSM framework is established based on the midlevel and high-level scene classification methods. In this section, we briefly introduce the classical scene classification models for HSR images: the FSTM and Caffe.

A. Fully Sparse Topic Model

The FSTM was first proposed for text analysis [36] and is a simplified variant of the classical PTMs, i.e., PLSA and LDA. It aims at solving the dense topic modeling problem of LDA, where the latent topics mined by LDA are often dense and lead to redundant information and high time and storage consumption. By removing the Dirichlet distribution of LDA, the FSTM employs a sparse prior and allows fewer topics to model an image.

Based on the similarities between the text analysis field and image processing, the FSTM can be applied to HSR image scene classification. Fig. 3 displays the probabilistic graphical model of the FSTM, which represents the relationship between the image, topic, and visual word. In Fig. 3, the nodes W, Z, θ, and β represent the visual word w, latent topic z, topic proportion θ, and visual word proportion β, respectively. For HSR images, certain features (e.g., the texture feature) are extracted from the local image patches or segmented region of an image. The features of the patches are then quantized by a clustering algorithm to acquire the set of cluster labels and their corresponding cluster centers, which form the visual dictionary. Each item of the visual dictionary is called a visual word. Modeled by the FSTM, the latent topic of an HSR image is characterized by a multinomial distribution over visual words and is a more representative representation of the HSR image, with a lower feature dimension. Hence, given an HSR image $\boldsymbol{H_{i\in M}}$, which is described by a set of N visual words w_j, K latent topics z_k can be mined from the image. Based on Fig. 3, the generative process for the HSR image can be constructed as follows.

1) A k-dimensional topic proportion θ is randomly chosen from each image.
2) For the jth visual word in the image $\boldsymbol{H_i}$, the latent topic z_k with probability $P(z_k|\boldsymbol{H_i}) = \theta_k$ is selected, and the visual word w_j with probability $P(w_j|z_k) = \beta_{kj}$ can be generated.

Based on the property of sparse inference, the FSTM uses an implicit sparse prior λ to model the topics, as shown in Fig. 3 with the node λ. The sparse prior in the FSTM conforms to the density function (1). In other words, the number of nonzero entries of latent topic proportion θ is denoted as $||\theta||_0$, and θ follows an implicit constraint $||\theta||_0 \leq L + 1$, where L is the iteration times. This makes the FSTM to converge at a linear rate to the optimal solution, and the FSTM is designed to generate heterogeneous sparse topics in the ADSSM framework

$$p(\theta|\lambda) \propto \exp(-\lambda \cdot ||\theta||_0). \tag{1}$$

B. CaffeNet

Caffe is a fully open-source framework that affords clear, modifiable, and easy implementations of deep architectures [37]. Caffe was developed by the Berkeley AI Research Lab, the Berkeley Vision and Learning Center, and community contributors. Based on the protocol buffer language, Caffe can easily create new architectures, with several functionalities, such as layer visualization, fine-tuning strategies, and feature extraction.

As a commonly used CNN, CaffeNet can directly extract features from the raw data without prior knowledge and has been widely used in many tasks, including HSR imagery scene classification. It is obtained using the same data set as the ILSVRC 2012 competition [38] and basically the same parameters as Krizhevsky's network. As shown in Fig. 4, CaffeNet usually consists of five convolutional layers, some of which are followed by max-pooling layers and three

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

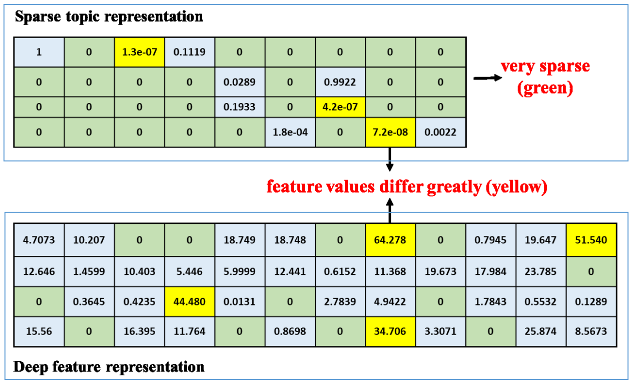


Fig. 5. Examples of midlevel and deep feature representations.

fully connected layers, with a final softmax layer. It has 60 million parameters and 650 000 neurons. Two important modifications for CaffeNet are: 1) training is undertaken without data augmentation and 2) the order of the pooling and normalization layers is exchanged, and pooling is done before normalization. CaffeNet allows feature extraction for any layer of the network and is used as a fixed feature extractor to generate the deep features in the ADSSM framework.

A couple of important questions need to be considered. Are the handcrafted features extracted from the midlevel scene classification methods, e.g., the FSTM-based scene classification method, still useful and complementary for the deep features? In addition, as big differences exist between deep features and handcrafted features, how do we effectively combine them? Examples of sparse topic-based midlevel feature representation and deep feature representation from the fully connected layer are shown in Fig. 5. The two kinds of features differ in both feature values and feature sparsity. The topic features are very sparse, and the zero values are colored in green in Fig. 5. The feature values of the topics range from 0 to 1 and may even be very small, such as 7.2e−08, which is colored in yellow. However, the deep features are denser, and most of the feature values are much bigger than the values of the topic features, which are colored in yellow in Fig. 5.

III. Scene Classification Based on the Adaptive Deep Sparse Semantic Modeling Framework

To effectively utilize the multilevel semantic information, the ADSSM framework is proposed for HSR image scene classification. Four tasks need to be addressed: 1) high-level deep feature learning; 2) midlevel sparse topic modeling; 3) adaptive feature normalization and fusion; and 4) scene label generation. The overall flowchart of scene classification based on the ADSSM framework for HSR images is shown in Fig. 6.

A. High-Level Deep Feature Learning

A pretrained CNN-based on CaffeNet, which is a biologically inspired multistage architecture, is employed in ADSSM. The flowchart of scene classification based on a CNN can be divided into three parts: 1) preprocessing; 2) feature extraction; and 3) feature classification. In the preprocessing phase, the input image is divided by the maximum pixel value (e.g., 255 in an RGB image), and then, a normalized image is obtained. After the preprocessing, the CNN is applied to extract features for scene classification. In the feature extraction process, given a normalized image I_n, convolutional computation is conducted by sliding a convolutional filter on the convolutional layer, and a feature map C_1 is obtained. For each channel of C_1, a pooling operation is applied to acquire the local maximum value or mean value of the region pixels. In this way, features, which are more robust to the noise and clutter and invariant to image transformation, are acquired [39]. After the pooling layer, the deep features are flattened into a feature vector, which is then fed into the fully connected layer to further extract information for the whole image. The output of the fully connected layer is finally classified by the softmax layer in the classification phase. As a generalization of logistic regression for a multiclass problem, the softmax layer gives the possibility of a sample belonging to each class. In the ADSSM framework, CaffeNet is used as a feature extractor network by extracting the features from the first fully connected layer, which provides the deep features for the HSR images. The features of the fully connected layer mainly describe the spatial layout information, with the k-dimensional feature D for each image.

B. Midlevel Sparse Topic Modeling

For the midlevel-based FSTM scene classification, three complementary features are designed to describe the complex HSR images. Before the feature descriptor extraction, the images are split into image patches employing the uniform grid sampling method.

1) By reflecting the attributes that constitute the ground components and structures, the spectral features are the fundamental characteristics for HSR images. The first-order statistics of the mean value and the second-order statistics of the standard deviation value in each spectral channel of each image patch are computed as the spectral feature, as shown in (2). Let I be the number of pixels in the sampled patch, and b_{ij} denotes the jth band value of the ith pixel in a patch. In this way, the mean (M_j) and the standard deviation (SD_j) of the spectral vector of the patch are then acquired

$$M_j = \frac{\sum_{i=1}^{I} b_{ij}}{I}, \quad \mathrm{SD}_j = \sqrt{\frac{\sum_{i=1}^{I} (b_{ij} - M_j)^2}{I}}. \quad (2)$$

2) In ADSSM, the structural features are extracted by employing the SIFT descriptor [40], which is able to overcome affine transformation, noise, and changes in illumination. Inspired by the work of Fei-Fei and Perona [41], gray dense SIFT is utilized as the patch descriptor. The image patches are divided into 4×4 nonoverlapping regions. For each region, a magnitude-weighted orientation histogram is computed. The gradient orientation histograms of eight directions are then counted, and $4 \times 4 \times 8 = 128$-dimension vectors are finally obtained to describe the keypoint descriptor.

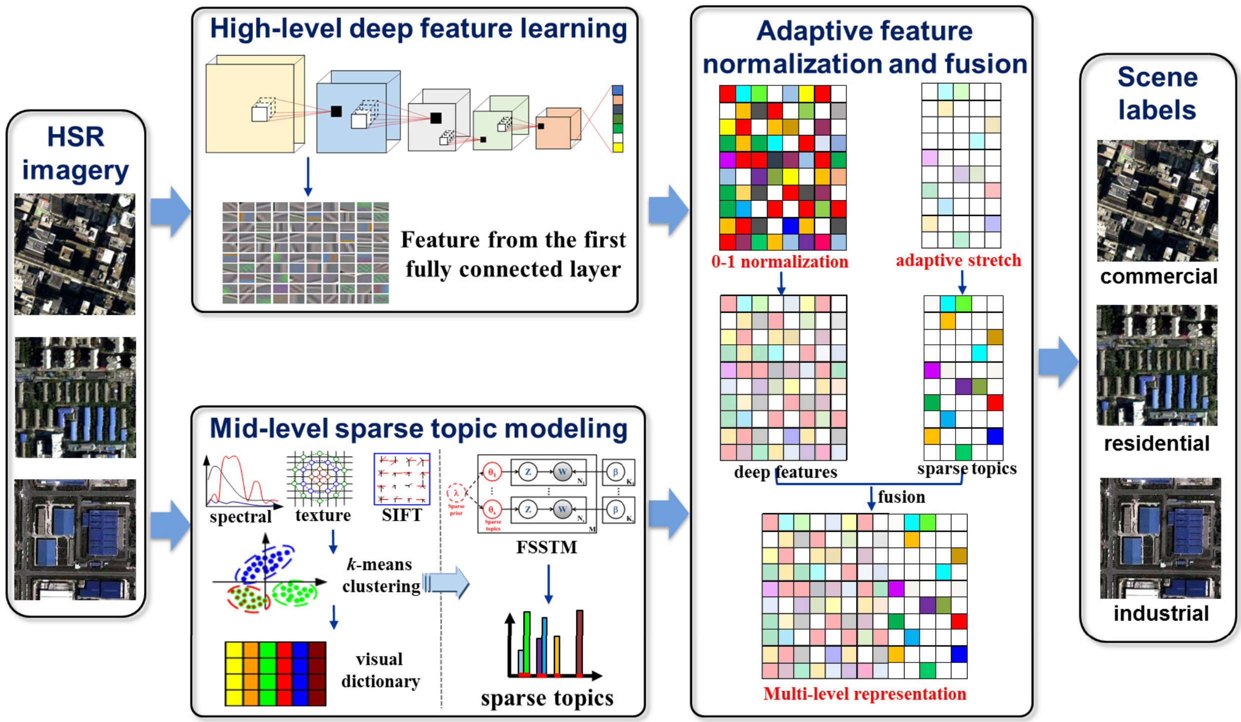


Fig. 6. Flowchart of scene classification based on the ADSSM framework for HSR images.

3) To compensate for the deficiency of the statistical spectral features and SIFT features, the wavelet feature is chosen as the texture feature in the ADSSM framework. By discovering information from an image about both the spatial and frequency contents, wavelets can be adopted to analyze texture for nonstationary or nonhomogeneous images, such as HSR remote sensing images [42]. Similar to the way the human visual system operates, wavelet transforms decompose the image into different frequency subbands and are suitable for image classification. In ADSSM, multilevel 2-D wavelet decomposition is utilized to extract the texture feature, and the level of the wavelet decomposition is optimally set to 3.

As a result of the influence of illumination, rotation, and scale variation, the same visual word in different images may be endowed with different feature values. Hence, the local low-level features, i.e., the spectral and SIFT features, are vector quantized by k-means clustering of ADSSM to make sure that the image patches with similar feature values correspond to the same visual word. Hence, given an image $\boldsymbol{H_i}$ described by a set of words w_j, where $w_j \in \{1, 2, \ldots, V\}$ and V is the number of visual words from the visual dictionary, the image can be represented as $\boldsymbol{H_i} = \{w_1, \ldots, w_j, \ldots, w_N\}$. By the statistical analysis of the frequency of each visual word, the corresponding visual dictionaries with a size of D_1, D_2, and D_3 for the spectral, SIFT, and texture features, respectively, can be obtained. Three 1-D histograms features are acquired, and K_1, K_2, K_3 topics can then be separately mined from the histograms. Specifically, for each type of feature, the set of term indices of image $\boldsymbol{H_i}$ is defined as I_{H_i}, and the log likelihood of $\boldsymbol{H_i}$ with K topics $\beta = (\beta_1, \ldots, \beta_K)$ can be defined as (3)

$$\log P(\boldsymbol{H_i}) = \sum_{j \in I_{H_i}} \boldsymbol{H}_{ij} \log \sum_{k=1}^{K} \theta_k \beta_{kj}$$
$$\sum\nolimits_k^{\theta_k} = 1, \quad \theta_k \geq 0 \tag{3}$$

To infer the latent topics, the goal is to search for θ_i of the ith image to maximize the likelihood of $\boldsymbol{H_i}$. The greedy approach-based Frank–Wolfe algorithm [43] is employed as the inference algorithm in ADSSM. In this way, the inference task of optimization over θ_i is reformulated as a concave maximization problem over the simplex $\Delta = \text{conv}(\beta_1, \ldots, \beta_K)$ of the topic. The topic probabilities are acquired at the lth iteration, as written in (4). Here, α' is defined by (5), where $i' := \arg\max_i \beta_i^t \nabla f(x_l)$, β_i denotes the standard unit vectors in Δ, and α can be solved by the gradient ascent approach. In addition, x_l is a convex combination of at most L+1 vertices of the simplex $\Delta = \text{conv}(\beta_1, \ldots, \beta_K)$ after L iterations, as written in (6). Hence, it can be implied that at most $l + 1$ out of the k-dimensional θ_l are nonzero in ADSSM, which provides us with a sparse solution. After acquiring the topic probabilities θ_i, the learning phase is to learn the topics $\beta = (\beta_1, \ldots, \beta_K)$ with the use of Jensen's inequality, as written in (7). By iterating the inference and learning task until convergence, the optimal sparse solutions of latent semantic topic proportions θ_1, θ_2, and θ_3 are separately obtained for the three types of features at a linear rate. Finally, the sparse topic features $\text{ST} = \{\theta_1^T, \theta_2^T, \theta_3^T\}^T$ can be obtained by fusing θ_1, θ_2, and θ_3 with $K_1 + K_2 + K_3$ dimensions. The pseudocode of the midlevel sparse topic modeling is shown in Algorithm 1. The parameters used in Algorithm 1 are

Algorithm 1 Pseudocode of Midlevel Sparse Topic Modeling

Input: image $\boldsymbol{H_i}$ and topics β
for $l = 0, \ldots, \infty$ **do**
$i' := \arg\max_i \beta_i^t \nabla f(x_l),\ x_l = \sum_{k=1}^{K} \theta_{lk}\beta_k$
Update α' using (5)
Compute log likelihood $\log P(\boldsymbol{H_i})$ using (3)
$x_{l+1} := \alpha'\beta_{i'} + (1-\alpha')x_l$
Update θ_{l+1} using (4)
end for
Update β_{kj} using (7)
Output: topic probability θ_i

as follows:

$$\theta_{l+1} := (1-\alpha')\theta_l \tag{4}$$

$$\alpha' := \arg\max_{\alpha\in[0,1]} f(\alpha\beta_{i'} + (1-\alpha)x_l) \tag{5}$$

$$x_l = \sum\nolimits_{k=1}^{K} \theta_{lk}\beta_k \tag{6}$$

$$\beta_{\mathrm{kj}} \propto \sum_{M\in D} d_j\theta_{dk}. \tag{7}$$

C. Adaptive Feature Normalization and Fusion

Compared with the deep features, the different types of topic features θ_1, θ_2, and θ_3 differ in two aspects: sparsity and feature values. The topic features are very sparse and representative, whereas the deep features are denser and more comprehensive. As written in (3), the feature values of the topic features range from 0 to 1, and some feature values may even be very small. However, the feature values of the deep features are much bigger. Directly concatenating the deep features and the topic features cannot make full use of the representative topic features. As such, the ADSSM framework employs 0–1 normalization for the deep features, as written in (8). ND represents the normalized deep features, ID denotes the minimum value of all the deep features, and AD denotes the maximum value of all the deep features. To circumvent the topic features being neglected by the deep features, the topic features are adaptively stretched in the ADSSM framework, as written in (9). ASST denotes the adaptively stretched sparse topic features, IST denotes the minimum value of all the sparse topic features, and AST denotes the maximum value of all the sparse topic features. In this way, the sparse topic features are adaptively stretched based on the maximum value of the deep features and are effective for discriminating the scenes when combined with ND. By fusing the adaptively stretched sparse topic features and normalized deep features at the semantic level, the final semantic representation $F = \{\mathrm{ASST,ND}\}$ for the HSR images can be obtained with $K + K_1 + K_2 + K_3$ dimensions

$$\mathrm{ND} = (D - \mathrm{ID})/(\mathrm{AD} - \mathrm{ID}) \tag{8}$$

$$\mathrm{ASST} = \mathrm{AD}/(\mathrm{AST} - \mathrm{IST}) * (\mathrm{ST} - \mathrm{IST}). \tag{9}$$

D. Scene Label Generation

In the scene label generation phase, the final semantic representation $F = \{\mathrm{ASST,ND}\}$ with discriminative characteristics is classified by the SVM classifier to predict the scene labels because of its excellent performance. The radial basis function (RBF) kernel-based SVM is able to handle the case where the relationship between the class labels and attributes is nonlinear [44], and can usually obtain better results than the linear kernel-based SVM. However, the parameters of the RBF kernel-based SVM are usually estimated by cross validation, which results in increased time consumption. By measuring the degree of similarity between two histograms, the histogram intersection kernel (HIK) deals with the scale change and has been applied to image classification using color histogram features [45]. For the proposed framework, the multilevel representations obtained by the fusion operation can be viewed as a generalized histogram. Hence, we use LibSVM [44] to train an SVM classifier with an HIK. Let $\widetilde{\boldsymbol{R}} = (\widetilde{r}_1, \widetilde{r}_2, \ldots, \widetilde{r}_M)$ be the ADSSM representation vectors of all the images, and the HIK is defined as shown in (10), where f is the index of the component of the feature vectors. Due to the simplicity of the HIK, the SVM classifier with HIK can be quickly trained, without cross validation of the parameters. Finally, the scene label of each image can be predicted

$$K(r, r^i) = \sum_f \min(r_f, r_f^i). \tag{10}$$

IV. Experiments and Analysis

A. Experimental Setup

In order to test the performance of ADSSM, the commonly used 21-class UCM data set, the 12-class Google data set of SIRI-WHU, the challenging large-scale NWPU-RESISC45 data set, and the OSRSI20 data set were evaluated in the experiments. In the experiments with uniform grid-based region sampling, the patch size and spacing for the spectral and texture features were optimally set to 8×8 pixels and 4×4 pixels, respectively. The patch size and spacing for the SIFT feature were optimally set to 16×16 pixels and 8×8 pixels, respectively, for the UCM data set and NWPU-RESISC45 data set, and 8×8 pixels and 4×4 pixels for the Google data set of SIRI-WHU and the OSRSI20 data set. The visual dictionary for BoVW with V visual words was constructed by employing Euclidean distance measurement-based k-means clustering over the image patches from the training data. The increase of the dictionary size leads to slight fluctuation of the accuracy and may result in much higher dimensional features for BoVW, which needs more time and space consumption [46]. Hence, the visual word number V was optimally set to 1000 for the spectral and SIFT features and 800 for the texture feature for the four data sets. The images in the four data sets were all resized to $227 \times 227_{\mathrm{pixels}}$ for CaffeNet, giving consideration to its predefined size requirement for the input image. CaffeNet was trained using stochastic gradient descent, with a batch size of 64, a momentum of 0.9, a weight decay of 0.00002, and a learning rate of 0.001. The experiments were performed on a personal computer equipped with dual Intel Xeon E5-2650 v2 processors, a single Tesla K20m GPU, and 64 GB of RAM, running Centos 6.6 with the CUDA 6.5 release. The different

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I

OVERALL CLASSIFICATION ACCURACY (%) COMPARISON WITH THE UCM DATA SET

pLSA [47]	89.51±1.31
LDA [48]	81.92±1.12
Cheriyadat [8]	81.67±1.23
DMTM [49]	92.92±1.23
IFK(VGG-VD16) [30]	98.49
Fine-tuned GoogLeNet [29]	97.78±0.97
Fine-tuned GoogLeNet descriptors [29]	99.47±0.50
VGG-VD16+AlexNet [31]	98.81±0.38
SPE-FSTM	78.33±1.42
WAV-FSTM	75.00±1.63
SIFT-FSTM	82.38±1.58
FSSTM [15]	95.71±1.01
CAFFE-FC6	95.89±0.74
DST	95.95±1.03
NDST	97.38±0.67
ND255ST	98.81±0.45
ADSSM	**99.76±0.24**

TABLE II

OVERALL CLASSIFICATION ACCURACY (%) COMPARISON WITH THE GOOGLE DATA SET OF SIRI-WHU

pLSA [47]	89.60±0.89
LDA [48]	60.32±1.20
LGFBOVW [7]	96.96±0.95
DMTM [49]	91.52±0.64
GCS_{10}-DSDM [52]	97.14±0.51
SPE-FSTM	83.33±1.06
WAV-FSTM	80.92±0.95
SIFT-FSTM	78.50±1.12
FSSTM [15]	97.83±0.93
CAFFE-FC6	91.79±0.75
DST	91.92±0.79
NDST	95.91±0.48
ND255ST	99.25±0.29
ADSSM	**99.75±0.15**

TABLE III

OVERALL CLASSIFICATION ACCURACY (%) COMPARISON WITH THE NWPU-RESISC45 DATA SET

Method	Training ratio	
	10%	20%
VGGNet-16 [50]	76.47±0.18	79.79±0.15
BoCF with VGGNet-16 [31]	82.65±0.31	84.32±0.17
Fine-tuned VGGNet-16 [50]	87.15±0.45	90.36±0.18
Triplet networks [51]		92.33 ± 0.20
SPE-FSTM	50.13±0.78	55.20±0.72
WAV-FSTM	35.21±1.02	39.32±1.13
SIFT-FSTM	38.49±1.05	42.46±0.79
FSSTM [15]	66.74±0.49	73.14±0.56
CAFFE-FC6	78.65±0.36	81.27±0.39
DST	78.79±0.92	81.57±0.78
NDST	82.10±0.88	86.07±0.69
ND255ST	84.34±0.73	88.69±0.65
ADSSM	**91.69±0.22**	**94.29±0.14**

TABLE IV

OVERALL CLASSIFICATION ACCURACY (%) COMPARISON WITH THE OSRSI20 DATA SET

Method	Training ratio	
	50%	80%
SPE-FSTM	64.2±0.98	64.33±1.02
WAV-FSTM	84.93±0.64	87.17±0.55
SIFT-FSTM	54.53±1.06	60.33±1.09
FSSTM [13]	88.80±0.47	91.33± 0.46
CAFFE-FC6	91.93±0.69	93.83±0.65
DST	92.73±0.73	94.83±0.57
NDST	95.67±0.39	96.83±0.42
ND255ST	94.80±0.19	97.17±0.25
ADSSM	**97.13%±0.18**	**99.83±0.12**

methods were implemented 20 times by randomly selecting the training samples to ensure that convincing results were obtained and the stability of the proposed ADSSM could be tested.

In Tables I–IV, SPE-FSTM, WAV-FSTM, and SIFT-FSTM denote the scene classification utilizing the MSD-based spectral feature, the wavelet-based texture feature, and the SIFT-based structural feature, respectively. Fully sparse semantic topic model (FSSTM) [15] denotes the midlevel sparse topic modeling-based scene classification utilizing the spectral, texture, and SIFT features. CAFFE-FC6 denotes the pretrained CaffeNet-based scene classification utilizing the first fully connected feature with the SVM classifier. Deep features and sparse topics (DST) and normalized DST (NDST) denote the scene classification combining the original deep features and sparse topics, and the normalized deep features and sparse topics, respectively. ND255ST denotes the scene classification combining the normalized deep features and

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

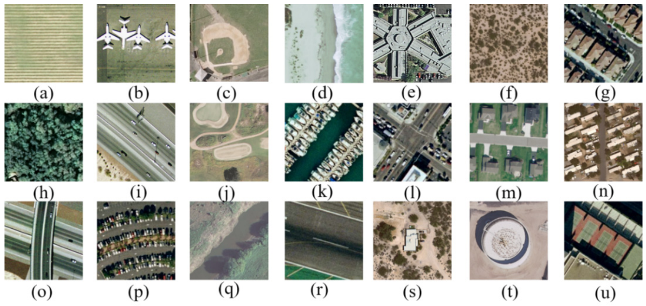


Fig. 7. UCM data set. (a)–(u) Agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts, respectively.

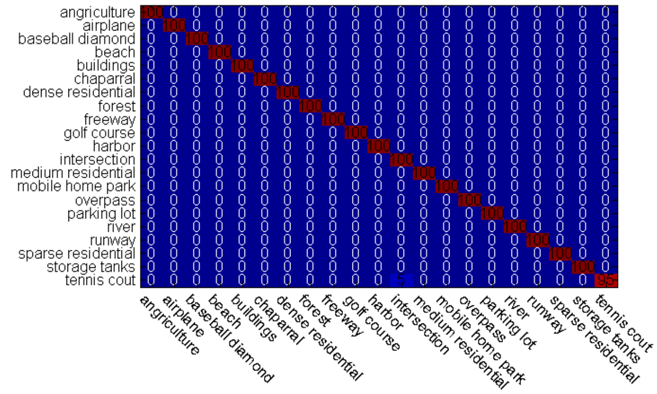


Fig. 8. Confusion matrix of ADSSM with the UCM data set.

255 stretched sparse topics, which was used as a comparison method to evaluate the effectiveness of the adaptive normalization strategy.

To further evaluate the performance of ADSSM, the experimental results obtained with the conventional pLSA [47] and LDA [48] are shown for comparison. We also provide the experimental results obtained with the UCM data set, as published in the latest papers by Cheriyadat [8], Nogueira *et al.* [29], Hu *et al.* [30], Li *et al.* [31], and Zhao *et al.* [49]. The experimental results obtained for the Google data set of SIRI-WHU by Zhu *et al.* [7], Zhao *et al.* [49], and Yuan *et al.* [52], and the results obtained for the NWPU-RESISC45 data set by Cheng *et al.* [33], [50] and Liu and Huang [51] are also provided for comparison with the proposed ADSSM framework.

B. Experiment 1: The UC Merced Image Data Set

The UCM data set[1] was downloaded from the USGS National Map Urban Area Imagery collection [5]. This data set consists of 21 land-use scenes (see Fig. 7), namely agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. Each class contains 100 images, measuring 256×256 pixels, with a 1-ft spatial resolution. Following the experimental setup published in Yang and Newsam [5], 80 samples were randomly selected per class from the UCM data set for training, and the rest were kept for testing.

The classification performances of the single-feature-based FSTM, the conventional pLSA, LDA, FSSTM, CAFFE-FC6, DST, NDST, ND255ST, the proposed ADSSM, and the comparison with the experimental results of the previous methods for the UCM data set are listed in Table I. As can be seen in Table I, the classification result of FSSTM, $95.71\% \pm 1.01\%$, is much better than the spectral, wavelet, and SIFT-based FSTMs, which confirms that FSSTM is an effective approach for HSR image scene classification. Compared with the pLSA [47], LDA [48], FSSTM, and CAFFE-FC6 methods, the classification accuracy for the proposed ADSSM, $99.76\% \pm 0.24\%$, is the best among all the different methods. This indicates that the combination of the midlevel features and the high-level features is effective. The classification accuracy of DST is almost the same as FSSTM and CAFFE-FC6, whereas ADSSM outperforms the different normalization and fusion strategies of DST, NDST, and ND255ST. This confirms the powerful representation ability based on the proposed adaptive normalization and fusion strategy. In Table I, compared with the other methods, i.e., the Cheriyadat's method [8], the Zhao *et al.*'s method [49], the Hu *et al.*'s method [30], the Li *et al.*'s method [31], and the Nogueira *et al.*'s method [29], the highest accuracy is acquired by the proposed ADSSM.

An overview of the performance of ADSSM is shown in the confusion matrix in Fig. 8. As can be seen in Fig. 8, all of the scene categories can be fully recognized by ADSSM, except for the tennis court scene, but there is some confusion between the tennis court and the intersection scenes. This may be because the two categories are composed of a mixture of vegetation cover, car, and bare ground.

To allow a better visual inspection, some of the classification results of DST and ADSSM are shown in Fig. 9.

C. Experiment 2: The Google Data Set of SIRI-WHU

The Google data set of SIRI-WHU[2] was acquired from Google Earth (Google Inc.), covering urban areas in China, and the scene image data set was designed by the Intelligent Data Extraction and Analysis of Remote Sensing Group of Wuhan University (SIRI-WHU) [7], [49]. The data set consists of 12 land-use classes, which are labeled as follows: agriculture, commercial, harbor, idle land, industrial, meadow, overpass, park, pond, residential, river, and water, as shown in Fig. 10. Each class separately contains 200 images, which are cropped to 200×200 pixels, with a spatial resolution of 2 m. In this experiment, 100 training samples were randomly selected per class from the Google data set of SIRI-WHU, and the remaining samples were retained for the testing.

The classification performances of the single-feature-based FSTM, FSSTM, CAFFE-FC6, DST, NDST, ND255ST,

[1]http://vision.ucmerced.edu/datasets/landuse.html.

[2]http://www.lmars.whu.edu.cn/profweb/zhongyanfei/e-code.html.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

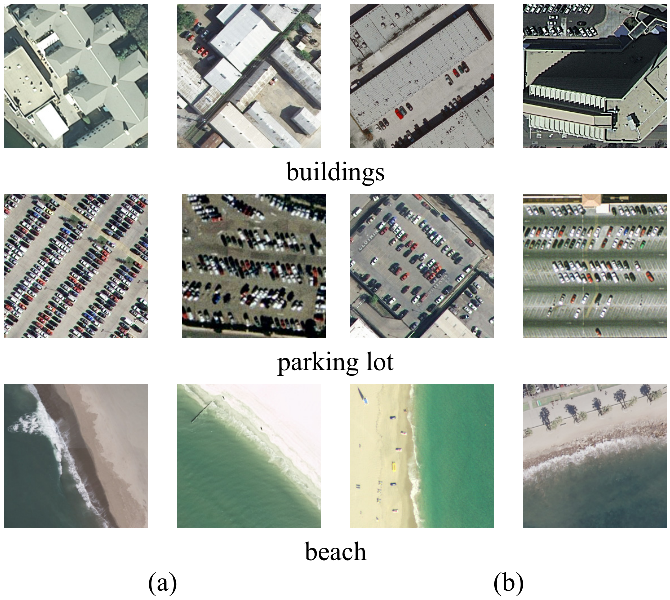


Fig. 9. Some of the classification results of DST and ADSSM. (a) Correctly classified images for the two methods. (b) Images correctly classified by ADSSM, but incorrectly classified by DST.

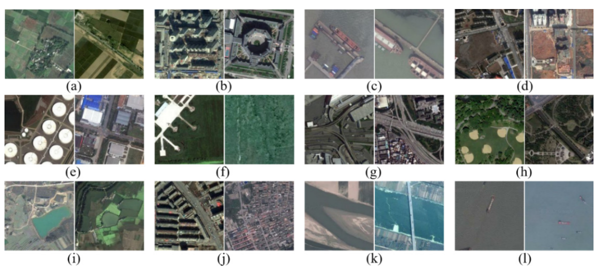


Fig. 10. Google data set of SIRI-WHU. (a)–(l) Agriculture, commercial, harbor, idle land, industrial, meadow, overpass, park, pond, residential, river, and water, respectively.

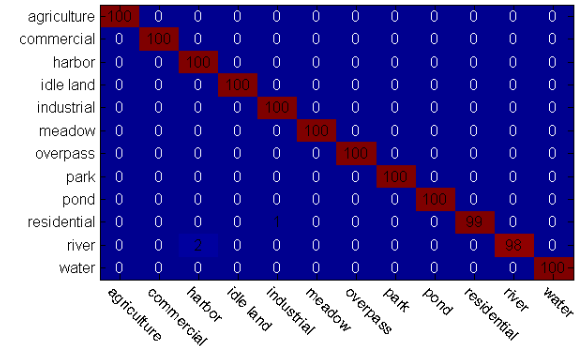


Fig. 11. Confusion matrix of ADSSM with the Google data set of SIRI-WHU.

SIRI-WHU.

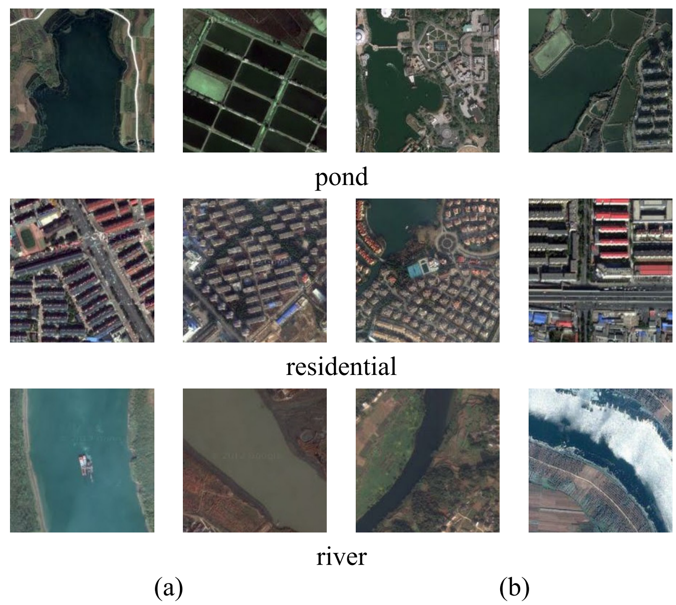


Fig. 12. Some of the classification results of DST and ADSSM. (a) Correctly classified images for the two methods. (b) Images correctly classified by ADSSM, but incorrectly classified by DST.

the proposed ADSSM, and the comparison with the experimental results of the previous methods for the Google data set of SIRI-WHU are listed in Table II. As can be seen in Table II, the classification results of FSSTM are better than the results of the single-feature-based FSTM, i.e., SPE-FSTM, WAV-FSTM, and SIFT-FSTM. This demonstrates that the fusion of multiple midlevel features is able to improve the classification performance. In addition, the classification results of FSSTM are better than the classification results of CAFFE-FC6, which indicates that the handcrafted features may sometimes be more effective than the deep features for some HSR image data sets, especially the small data sets. The classification accuracy for the proposed ADSSM, 99.75% $\pm$ 0.15%, is better than the results of pLSA [47], LDA [48], FSSTM, and CAFFE-FC6, which confirms that the ADSSM framework is an effective way to combine the distinct characteristics of midlevel and high-level features. The classification accuracy of DST is almost the same as CAFFE-FC6, but is much worse than FSSTM, whereas ADSSM outperforms DST, NDST, and ND255ST. This indicates that the adaptive normalization and fusion strategy in ADSSM improves the fusion of the different-level features. In addition, it can be seen that ADSSM performs better than the current state-of-the-art methods, i.e., the methods of Zhao *et al.* [49], Zhu *et al.* [7], and Yuan *et al.* [52].

Fig. 11 displays the confusion matrix of ADSSM for the Google data set of SIRI-WHU. On the whole, most of the scene classes achieve good classification performances with ADSSM. There is, however, some confusion between certain scenes. For instance, a scene belonging to river is classified as harbor. This can be explained by the fact that these classes have similar structural or spectral characteristics, such as both river and harbor featuring water and bank. The scene belonging to residential is classified as industrial, which is because the residential and industrial scenes have similar spatial distributions and the same objects, such as buildings and roads.

To allow a better visual inspection, some of the classification results of DST and ADSSM are shown in Fig. 12.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

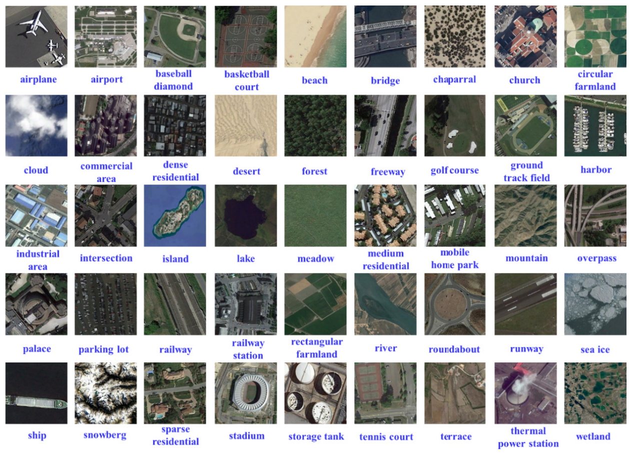


Fig. 13. Example images from the NWPU-RESISC45 data set.

D. Experiment 3: The NWPU-RESISC45 Data Set

The NWPU-RESISC45 data set[3] was acquired from Google Earth (Google Inc.), covering more than 100 countries and regions all over the world, including developing, transitional, and highly developed economies [50]. The NWPU-RESISC45 data set consists of 31 500 remote sensing images divided into 45 scene classes, namely airplane, airport, baseball diamond, basketball court, beach, bridge, chaparral, church, circular farmland, cloud, commercial area, residential, desert, forest, freeway, golf course, ground track field, harbor, industrial area, intersection, island, lake, meadow, medium residential, mobile home park, mountain, overpass, palace, parking lot, railway, railway station, rectangular farmland, river, roundabout, runway, sea ice, ship, snowberg, sparse residential, stadium, storage tank, tennis court, terrace, thermal power station, and wetland, as shown in Fig. 13. Each class separately contains 700 labeled images, which are cropped to 256×256 pixels. The spatial resolutions for most of the scene classes vary from about 30 to 0.2 m, except for the island, lake, mountain, and snowberg scene classes, which have lower spatial resolutions.

The classification performances of the single-feature-based FSTM, FSSTM, CAFFE-FC6, DST, NDST, ND255ST, the proposed ADSSM, and the comparison with the experimental results of the previous methods for the NWPU-RESISC45 data set are listed in Table III. As can be seen in Table III, under the training ratios of 10% and 20%, FSSTM can obtain better classification results than the single-feature-based FSTM, which proves its effectiveness. The classification results of the deep feature-based method, CAFFE-FC6, are much better than the handcrafted feature-based method, i.e., FSSTM, for the NWPU-RESISC45 data set. However, for the UCM data set and the Google data set of SIRI-WHU, the classification results of FSSTM are better than those of Caffe-FC6. This demonstrates that the handcrafted feature-based methods have limited transferability between different data sets, and the deep network is better able to capture the discriminative representations for a data set with a large amount of images. The classification accuracy of DST is very close to the classification accuracy of CAFFE-FC6, which infers that directly fusing the deep features and the sparse topics is unable to capture the representative information from the sparse topics. The classification accuracies for the proposed ADSSM, under the training ratios of 10% and 20%, respectively, are the best among the different methods, i.e., FSSTM, CAFFE-FC6, DST, NDST, ND255ST, and the experimental results published by Cheng *et al.* [33], [50] and Liu *et al.* [51]. This confirms the superiority of the ADSSM framework.

Fig. 14 displays the confusion matrix of ADSSM under the training ratios of 10% and 20% for the NWPU-RESISC45 data set. For ADSSM under the training ratio of 10%, the main confusion happens between the river and roundabout scenes, for the reason that they both are surrounded by vegetation or idle land, and some rivers run in a roundabout-type shape. The confusion also happens between the beach and bridge scenes, because they are all composed of vegetation, water, and bare ground. For ADSSM under the training ratio of 20%, the main confusion happens between the storage tank and railway station scenes, and the tennis court and intersection scenes, as they are similar in the spectral and structural characteristics. On the whole, most of the scenes show a satisfactory classification performance with ADSSM, which demonstrates the effectiveness of the ADSSM framework.

To allow a better visual inspection, some of the classification results of DST and ADSSM under the training ratios of 10% and 20% for the NWPU-RESISC45 data set are shown in Figs. 15 and 16.

E. Experiment 4: The OSRSI20 Data Set

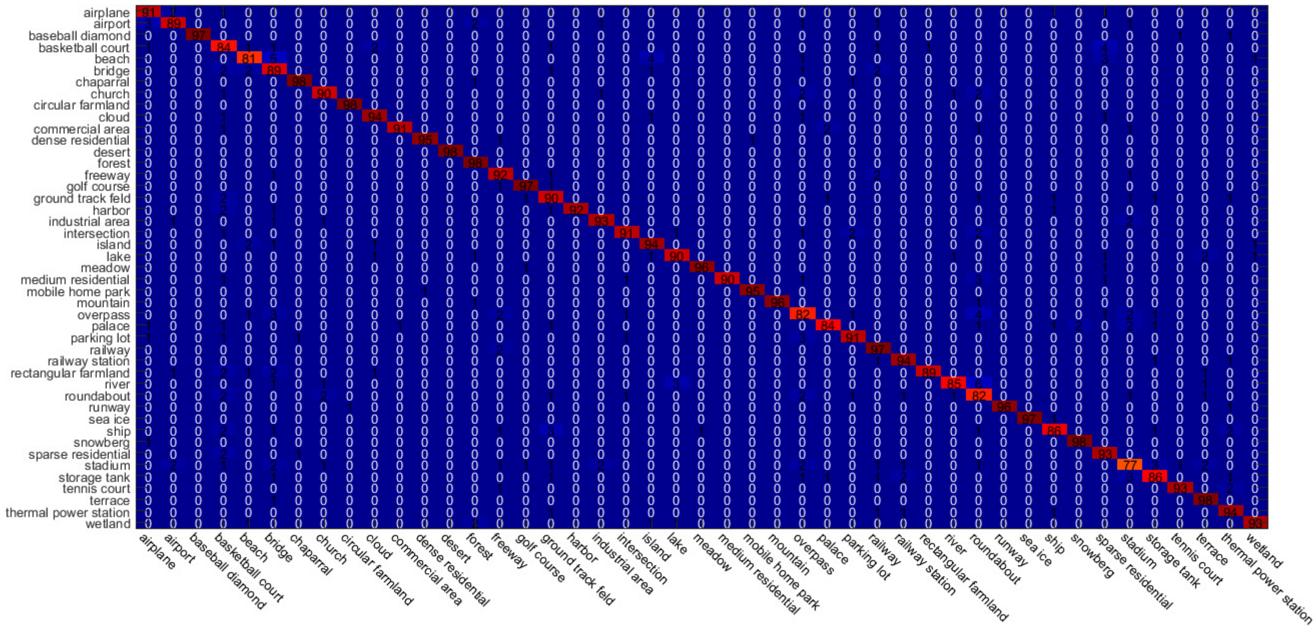
The OSRSI20 data set[4] is an optical satellite image data set. The individual images in the data set were downloaded from the DigitalGlobal Imagery Product Samples [53]. The OSRSI20 data set covers the regions of Boli (China), Brisbane (Australia), Buenos Aires (Argentina), Cape Town (South Africa), Hong Kong (China), Karamandere (Turkey), Miami (USA), Paris (France), San Clemente (USA), Santiago (Chile), and Shenzhen (China). The OSRSI20 data set consists of 3000 equally sized remote sensing images divided into 20 scene categories, namely agricultural, airplane, artificial lawn, beach, building, chaparral, cloud, container, dense residential, factory, forest, harbor, medium-density residential, ocean, parking lot, river, road, runway, sparse residential, and storage tanks, as shown in Fig. 17. Each class separately contains 150 labeled images selected from the satellite images with a ground sample distance of 50 cm, which have been cropped to 256×256 pixels with a spatial resolution of 2 m.

The classification performances of the single-feature-based FSTM, FSSTM, CAFFE-FC6, DST, NDST, ND255ST, and the proposed ADSSM are listed in Table IV. As can be seen in Table IV, under the training ratios of 50% and 80%, the classification result of FSSTM is better than that of the single-feature-based FSTM, which proves the effectiveness of the combination of multiple features. The classification

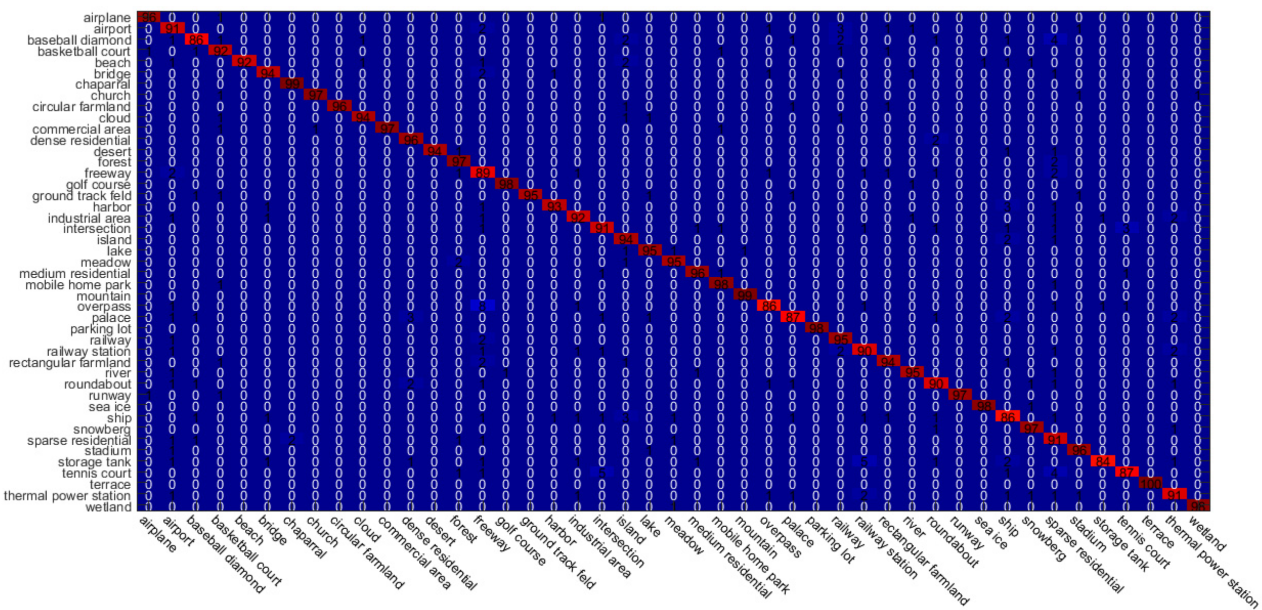
[3]http://www.escience.cn/people/JunweiHan/NWPU-RESISC45.html.

[4]https://sites.google.com/site/xutanghomepage/downloads.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



(a)



(b)

Fig. 14. Confusion matrix of ADSSM under the training ratios of (a) 10% and (b) 20% with the NWPU-RESISC45 data set.

accuracy of DST is slightly better than the classification accuracy of CAFFE-FC6. This infers that directly fusing the handcrafted features and the deep features is unable to combine the distinct characteristics of the midlevel and high-level features. The classification accuracies for the proposed ADSSM, 99.83% $\pm$ 0.12% under the training ratio of 50% and 97.13% $\pm$ 0.18% under the training ratio of 80%, are better than the results of FSSTM, CAFFE-FC6, DST, NDST, and ND255ST. This indicates that the ADSSM framework based on the adaptive normalization and fusion strategy is effective.

Fig. 18 displays the confusion matrix of ADSSM under the training ratios of 50% and 80% for the OSRSI20 data set. For ADSSM under the training ratio of 50%, the main confusion happens between the sparse residential, dense residential, and medium-density residential scenes, for the reason that they are composed of the same land-cover objects, and some images in the two scenes have very similar spatial distributions. For ADSSM under the training ratio of 80%, all of the scene categories can be fully recognized, except for the ocean scene. The ocean scene is misclassified as the beach scene, which is because the ocean is surrounded by the beach.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

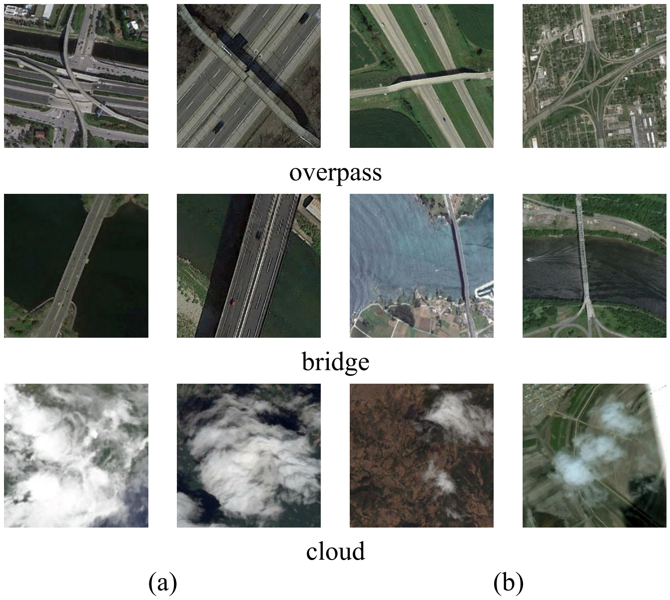


Fig. 15. Some of the classification results of DST and ADSSM under the training ratio of 10%. (a) Correctly classified images for the two methods. (b) Images correctly classified by ADSSM, but incorrectly classified by DST.

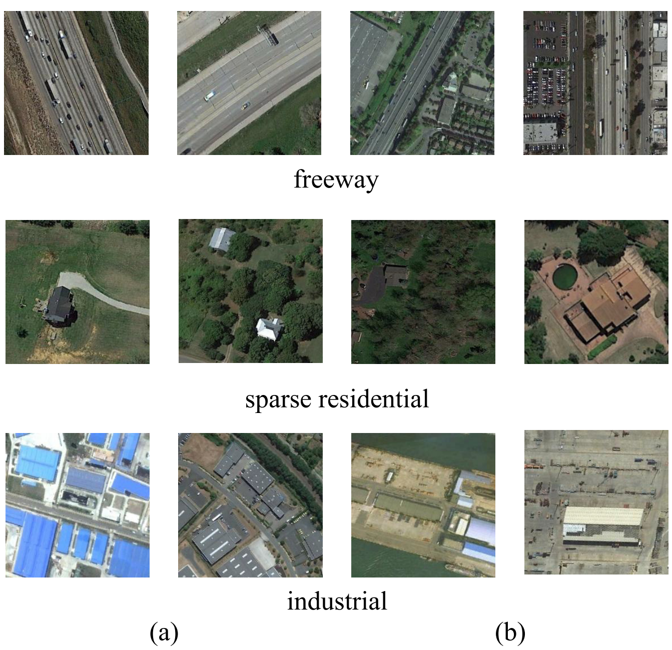


Fig. 16. Some of the classification results of DST and ADSSM under the training ratio of 20%. (a) Correctly classified images for the two methods. (b) Images correctly classified by ADSSM, but incorrectly classified by DST.

To allow a better visual inspection, some of the classification results of DST and ADSSM under the training ratios of 50% and 80% for the OSRSI20 data set are shown in Fig. 19.

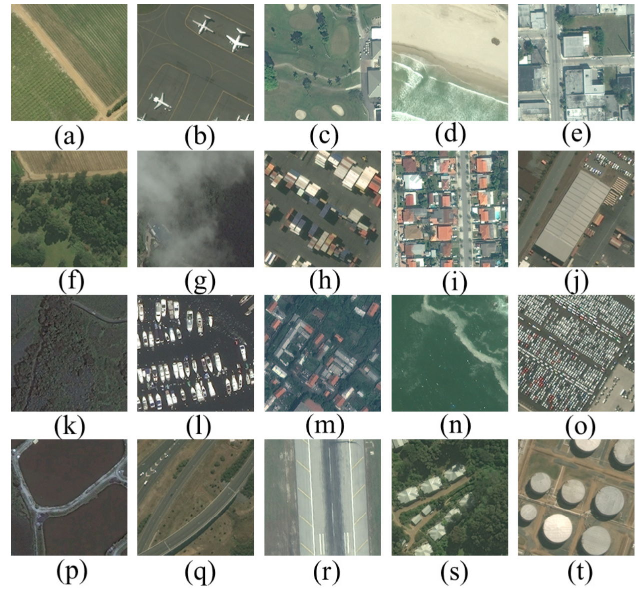


Fig. 17. Example images from the OSRSI20 data set. (a)–(t) Agricultural, airplane, artificial lawn, beach, building, chaparral, cloud, container, dense residential, factory, forest, harbor, medium-density residential, ocean, parking lot, river, road, runway, sparse residential, and storage tanks, respectively.

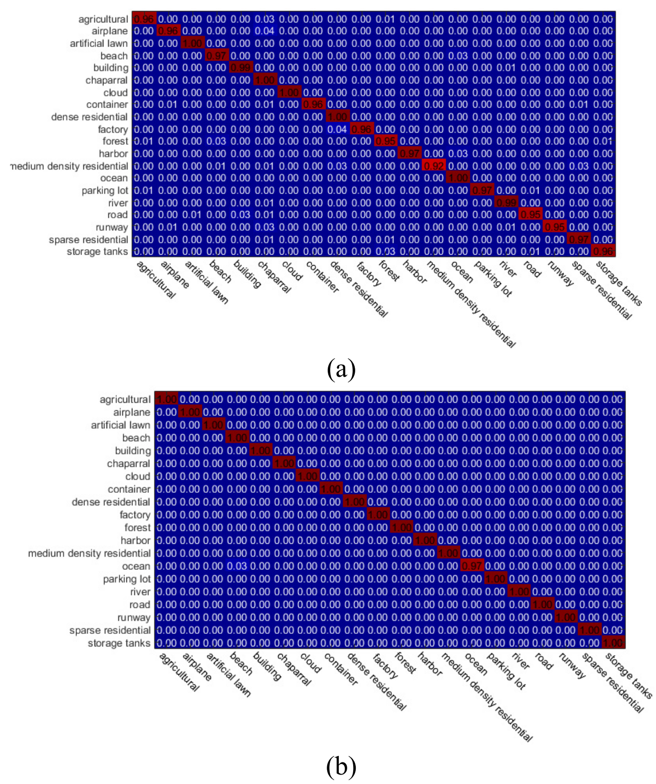


Fig. 18. Confusion matrix of ADSSM under the training ratios of (a) 50% and (b) 80% with the OSRSI20 data set.

V. Sensitivity Analysis

A. Sensitivity Analysis in Relation to the Patch Size for the Midlevel Feature Extraction

To investigate the sensitivity of the different midlevel features in relation to the patch size, the parameter settings were kept at the optimal, as described in the experimental setup. The patch size was then varied over the range of [4, 8, 12, 16, 20] for the UCM data set and the Google data set of SIRI-WHU, with the patch spacing as 50% of the size. From Fig. 20, it can be seen that the overall accuracy (OA)

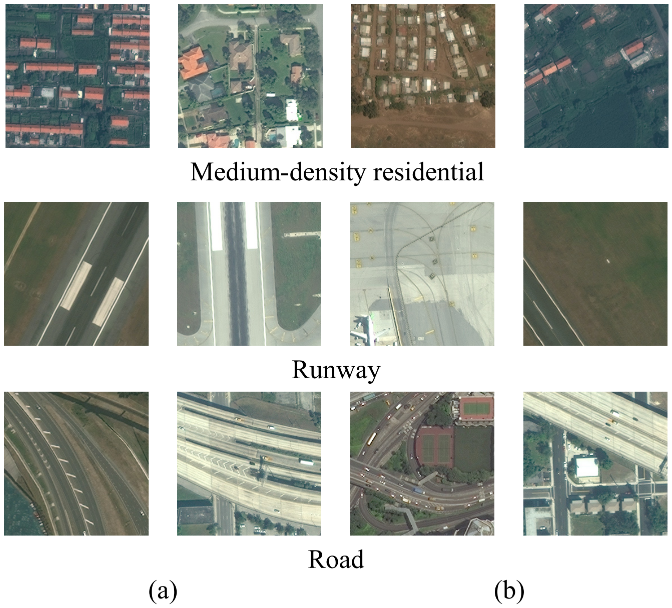


Fig. 19. Some of the classification results of DST and ADSSM under the training ratio of 50%. (a) Correctly classified images for the two methods. (b) Images correctly classified by ADSSM, but incorrectly classified by DST.

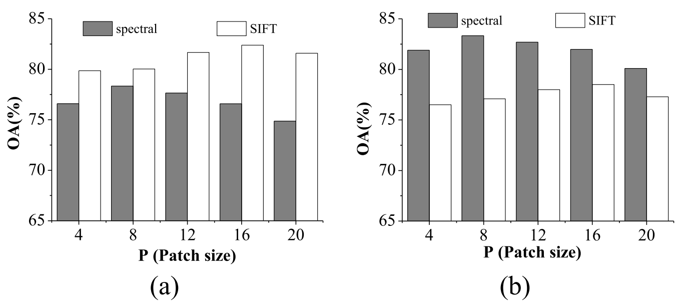


Fig. 20. Sensitivity analysis of FSSTM and ADSSM in relation to the patch size P. (a) UCM data set. (b) Google data set of SIRI-WHU.

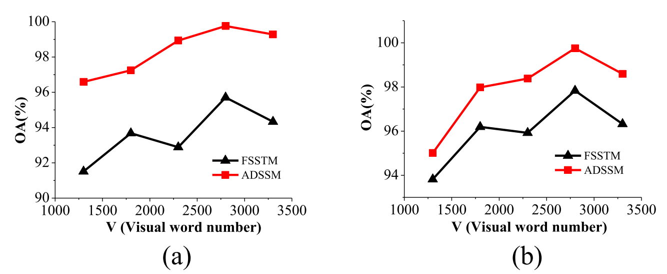


Fig. 21. Sensitivity analysis of FSSTM and ADSSM in relation to the visual word number V. (a) UCM data set. (b) Google data set of SIRI-WHU.

of the spectral and SIFT features is high at the beginning and then tends to decline with the increase of the patch size. The spectral feature obtains the highest accuracy when the patch size is 8 for the two data sets, whereas the SIFT feature demands a patch size of 16.

B. Sensitivity Analysis in Relation to the Visual Word Number

To study the sensitivity of the FSSTM and ADSSM frameworks in relation to the visual word number V, the parameter settings were kept at the optimal, as described in the experimental setup. The visual word number V was then varied over the range of [1300, 1800, 2300, 2800, 3300] for the UCM data set and the Google data set of SIRI-WHU. As shown in Fig. 21, with the increase of the visual word number, the OA curves of FSSTM and FSSTM display fluctuations. It is notable that ADSSM is superior to FSSTM over the entire range, which infers that the proposed ADSSM can outperform the other midlevel-based scene classification methods.

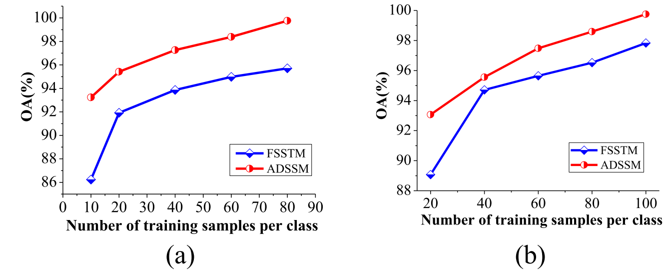


Fig. 22. Sensitivity analysis of FSSTM and ADSSM in relation to the ratio of training samples. (a) UCM data set. (b) Google data set of SIRI-WHU.

C. Sensitivity Analysis in Relation to the Ratio of Training Samples

To study the sensitivity of the FSSTM and ADSSM frameworks in relation to the number of training samples, the parameter settings were kept at the optimal, as described in the experimental setup. The number of training samples was then varied over the range of [80, 60, 40, 20, 10] for the UCM data set and [100, 80, 60, 40, 20] for the Google data set of SIRI-WHU. The curves of the OAs obtained by FSSTM and ADSSM for the UCM data set and the Google data set of SIRI-WHU are reported in Fig. 22. As shown in Fig. 22, the proposed ADSSM performs the best and is relatively stable with the decrease in the number of training samples per class for the two data sets, compared with FSSTM.

D. Sensitivity Analysis in Relation to the Different Convolutional Layers

To study the sensitivity of the ADSSM framework in relation to the different convolutional layers for the four HSR data sets, the parameter settings were kept at the optimal, as described earlier. The numbers of training samples were set as 20% and 80% for the NWPU-RESISC45 and OSRSI20 data sets, respectively. The first, third, and fifth convolutional layers and the first and second fully connected layers from the pretrained CaffeNet were selected for comparison. As shown in Fig. 23, features from the first fully connected layer, i.e., Fc6, outperform the features from the other layers. For the convolutional layers, as the depth of the layer increases, the OA also tends to increase. For the fully connected layer, Fc6 performs better than the features from the second fully connected layer (Fc7). Compared with Fc7, the fifth convolutional layer, i.e., Conv5, performs slightly worse in distinguishing the images. Hence, the Fc6 layer is chosen as the deep feature in the ADSSM framework.

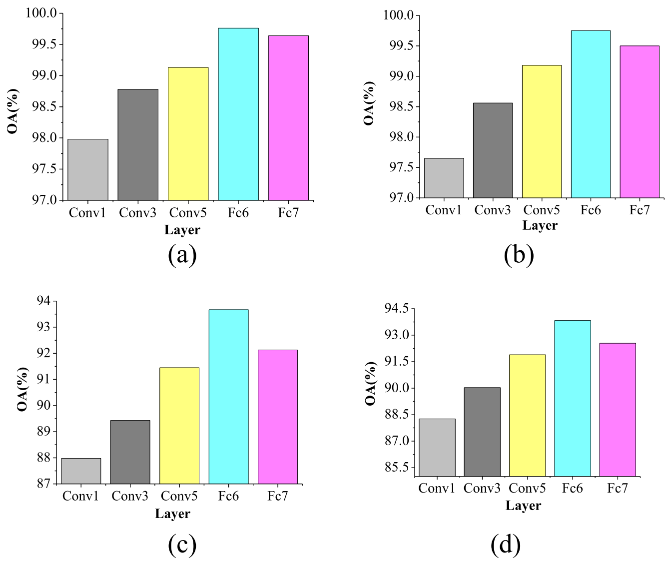


Fig. 23. Sensitivity analysis of ADSSM in relation to the different convolutional layers. (a) UCM data set. (b) Google data set of SIRI-WHU. (c) NWPU-RESISC45 data set. (d) OSRSI20 data set.

VI. Conclusion

In this paper, the ADSSM framework combining sparse topics and deep features has been proposed for the HSR remote sensing imagery scene classification. In ADSSM, the FSTM-based midlevel scene classification method and the CNN-based high-level scene classification method are employed for modeling the images. The combination of local spectral, texture, and SIFT features with deep features provides a comprehensive description for complex HSR images. The local spectral, texture, and SIFT features are fused at the semantic level, which circumvents the feature interruption of distinct image characteristics. Based on the adaptive feature normalization strategy, the midlevel features and the deep features are better fused, and thus, a discriminative multilevel description is obtained for distinguishing the scenes. The classification experiments undertaken in this paper showed that the proposed ADSSM framework performs better than the conventional midlevel feature-based methods, the deep feature-based scene classification methods, and the conventional feature fusion strategies in discovering high-quality semantics from HSR images.

In our future research, we plan to use more social media data, e.g., volunteered geographic information data, point of interest data, and OpenStreetMap data, to combine the scene classification results with practical use. In addition, multitemporal HSR images will be considered to further analyze the change information in the scenes.

References

[1] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[2] T. Blaschke *et al.*, "Geographic object-based image analysis—Towards a new paradigm," *ISPRS J. Photogram. Remote Sens.*, vol. 87, pp. 180–191, 2014.

[3] D. Bratasanu, I. Nedelcu, and M. Datcu, "Bridging the semantic gap for satellite image annotation and automatic mapping applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 1, pp. 193–204, Mar. 2011.

[4] 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.

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

[6] L.-J. Zhao, P. Tang, and L.-Z. Huo, "Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 12, pp. 4620–4631, Dec. 2014.

[7] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, Jun. 2016.

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

[9] J. Fan, T. Chen, and S. Lu, "Unsupervised feature learning for land-use scene recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2250–2261, Apr. 2017.

[10] J. Wang, C. Luo, H. Huang, H. Zhao, and S. Wang, "Transferring pre-trained deep CNNs for remote scene classification with general features learned from linear PCA network," *Remote Sens.*, vol. 9, no. 3, p. 225, 2017.

[11] Y. Zhong, M. Cui, Q. Zhu, and L. Zhang, "Scene classification based on multifeature probabilistic latent semantic analysis for high spatial resolution remote sensing images," *J. Appl. Remote Sens.*, vol. 9, no. 1, p. 095064, 2015.

[12] Y. Zhong, Q. Zhu, and L. Zhang, "Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6207–6222, Nov. 2015.

[13] Q. Zhu, Y. Zhong, L. Zhang, and D. Li, "Scene classification based on the fully sparse semantic topic model," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5525–5538, Oct. 2017.

[14] W. Luo, H. Li, G. Liu, and L. Zeng, "Semantic annotation of satellite images using author–genre–topic model," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1356–1368, Feb. 2014.

[15] Q. Zhu, Y. Zhong, S. Wu, L. Zhang, and D. Li, "Scene classification based on the sparse homogeneous–heterogeneous topic feature model," *IEEE Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2689–2703, May 2018.

[16] 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.

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

[18] 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.

[19] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[20] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[21] Y. Li, W. Xie, and H. Li, "Hyperspectral image reconstruction by deep convolutional neural network for classification," *Pattern Recognit.*, vol. 63, pp. 371–383, Mar. 2017.

[22] 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.

[23] Y. Zhong, A. Ma, Y. S. Ong, Z. Zhu, and L. Zhang, "Computational Intelligence in optical remote sensing image recognition," *Appl. Soft Comput.*, vol. 64, no. 3, pp. 75–93, Mar. 2018.

[24] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[25] 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.

[26] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.

[27] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2014, pp. 3320–3328.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[29] K. Nogueira, O. A. B. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, Jan. 2017.

[30] 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, 2015.

[31] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.

[32] G. Wang, B. Fan, S. Xiang, and C. Pan, "Aggregating rich hierarchical features for scene classification in remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4104–4115, Sep. 2017.

[33] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.

[34] A. Qayyum *et al.*, "Scene classification for aerial images based on CNN using sparse coding technique," *Int. J. Remote Sens.*, vol. 38, nos. 8–10, pp. 2662–2685, 2017.

[35] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi, "Deep filter banks for texture recognition, description, and segmentation," *Int. J. Comput. Vis.*, vol. 118, no. 1, pp. 65–94, May 2016.

[36] K. Than and T. B. Ho, "Fully sparse topic models," in *Proc. Eur. Conf. Mach. Learn. Principles Pract. Knowl. Discovery Databases (ECML PKDD)*, vol. 7523. Bristol, U.K., 2012, pp. 490–505.

[37] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, Nov. 2014, pp. 675–678.

[38] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[39] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proc. IEEE Conf. Mach. Learn.*, Nov. 2010, pp. 111–118.

[40] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[41] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 524–531.

[42] G.-S. Xia, J. Delon, and Y. Gousseau, "Shape-based invariant texture indexing," *Int. J. Comput. Vis.*, vol. 88, no. 3, pp. 382–403, Jul. 2010.

[43] K. L. Clarkson, "Coresets, sparse greedy approximation, and the Frank–Wolfe algorithm," *ACM Trans. Algorithms*, vol. 6, no. 4, 2010, Art. no. 63.

[44] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.

[45] A. Barla, F. Odone, and A. Verri, "Histogram intersection kernel for image classification," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2003, pp. III-513–III-516.

[46] G.-S. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[47] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 712–727, Apr. 2008.

[48] M. Lienou, H. Maitre, and M. Datcu, "Semantic annotation of satellite images using latent Dirichlet allocation," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 28–32, Jan. 2010.

[49] B. Zhao, Y. Zhong, G.-S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.

[50] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[51] Y. Liu and C. Huang, "Scene classification via triplet networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 1, pp. 220–237, Jan. 2018.

[52] L.-X. Yuan, S.-J. Li, and Y. P. Jiang, "Remote sensing scene classification using a preclassification strategy and an improved structural feature," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4094–4103, Sep. 2017.

[53] X. Tang, L. Jiao, W. J. Emery, F. Liu, and D. Zhang, "Two-stage reranking for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5798–5817, Oct. 2017.

Qiqi Zhu received the B.S. degree from the China University of Mining and Technology, Xuzhou, China, in 2013. She is currently pursuing the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China.

Her research interests include scene analysis for high spatial resolution remote sensing imagery and topic modeling.

Yanfei Zhong (M'11–SM'15) received the B.S. degree in information engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

He is currently the Dean of the Remote Sensing Division, State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, where he has been a Full Professor since 2010. He has published more than 150 research papers, including more than 80 peer-reviewed articles in international journals, such as the *ISPRS Journal of Photogrammetry and Remote Sensing*, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and *Pattern Recognition*. His research interests include hyperspectral remote sensing information processing, high-resolution remote sensing image understanding, and geoscience interpretation for multisource remote sensing data and applications.

Dr. Zhong was a recipient of the Excellent Young Scientist Foundation selected by the National Natural Science Foundation of China and the National Excellent Doctoral Dissertation Award of China. He was a recipient of the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). He was also a referee of more than 30 international journals. He is serving as an Associate Editor or an Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, the *International Journal of Remote Sensing*, and *Remote Sensing*.

Liangpei Zhang (M'06–SM'08) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He is a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. He was a Principal Scientist for the China State Key Basic Research Project from 2011 to 2016 appointed by the Ministry of National Science and Technology of China, Beijing, China, to lead the remote sensing program in China. He is currently a Chang-Jiang Scholar Chair Professor appointed by the Ministry of Education of China, Beijing. He has authored more than 450 research papers and five books. He holds 15 patents. He has edited several conference proceedings, issues, and geoinformatics symposiums. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a fellow of the Institution of Engineering and Technology. He received the 2010 Best Paper Boeing Award and the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing. He regularly serves as a Co-Chair of the series SPIE conferences on multispectral image processing and pattern recognition, conference on Asia remote sensing, and many other conferences. He also serves as an Associate Editor for the *International Journal of Ambient Computing and Intelligence*, the *International Journal of Image and Graphics*, the *International Journal of Digital Multimedia Broadcasting*, the *Journal of Geospatial Information Science*, and the *Journal of Remote Sensing*, and a Guest Editor of the *Journal of Applied Remote Sensing*and the *Journal of Sensors*. He is currently serving as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Deren Li (M'02–SM'03) received the M.Sc. degree in photogrammetry and remote sensing from the Wuhan Technical University of Surveying and Mapping, Wuhan University, Wuhan, China, in 1981, and the Dr. Eng. degree in photogrammetry and remote sensing from Stuttgart University, Stuttgart, Germany, in 1985.

He was elected as an Academician of the Chinese Academy of Sciences, Beijing, China, in 1991, and the Chinese Academy of Engineering, Beijing, and the Euro-Asia Academy of Sciences, Beijing, in 1995. He is currently the Academic Committee Chairman of the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. His research interests include spatial information science and technology, such as remote sensing, GPS and geographic information system (GIS), and their integration.

Dr. Li was the President of the International Society for Photogrammetry and Remote Sensing Commissions III and VI and the first President of the Asia GIS Association from 2002 to 2006.