Scene Classification Based on the Sparse Homogeneous–Heterogeneous Topic Feature Model

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

***Abstract*— High spatial resolution (HSR) imagery scene classification has been the subject of increased interest in recent years, and has great potential for many applications, such as urban functional analysis. Rooted in natural information processing, the use of the probabilistic topic model (PTM) to capture latent topics to represent HSR images has been an effective way to bridge the semantic gap. However, how to effectively discover discriminative information to recognize the HSR scenes is a challenging task. In this paper, the sparse homogeneous–heterogeneous topic feature model (SHHTFM) is proposed for HSR image scene classification. Differing from the conventional PTM-based scene classification methods, which utilize only heterogeneous features, SHHTFM explores the effect of the homogeneous information. Based on the union of uniform grid sampling and simple linear iterative clustering superpixel sampling, SHHTFM exploits both the heterogeneous and homogeneous information. After separately mining different types of low-level features and latent topics, the sparse topic inference procedure of SHHTFM further improves the fusion of the sparse heterogeneous and homogeneous topics. In addition, multisource geographical data are effectively integrated, where the water and vegetation boundaries define a more accurate way to restrict the boundaries of different scenes, and are then combined with the road network data to further improve the scene annotation performance. This provides more reliable and applicable results for us to better understand the complex scenes. The experimental results obtained with two HSR image classification data sets and an HSR image annotation data set demonstrate that the proposed SHHTFM framework can solve the scene classification problem, with a high classification accuracy as well as a high time efficiency.**

***Index Terms*— Geographical, high spatial resolution (HSR) imagery, homogeneous topics, multisource, probabilistic topic model (PTM), scene classification, scene understanding.**

Manuscript received August 30, 2017; revised October 13, 2017; accepted December 4, 2017. Date of publication February 9, 2018; date of current version April 20, 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, Grant 41771385, and Grant 41371344, and in part by the Natural Science Foundation of Hubei Province in China under Grant 2016CFA029. (*Corresponding author: Yanfei Zhong.*)

The authors 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: zhuqiqi@whu.edu.cn; zhongyanfei@whu.edu.cn; qhywsqnjnu@163.com; 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.2017.2781712

I. Introduction

ENORMOUS numbers of high spatial resolution (HSR) remote sensing images are now available and can provide abundant spectral and spatial information for precise land-use and land-cover (LULC) investigation. In HSR imagery, the spectral homogeneity within a class and the spectral variation between classes are both reduced, which makes the pixel-based classification methods inadequate for such a case. Object-based and contextual-based methods have been widely applied for HSR image classification, and are usually based on segmenting the images into groups of local homogeneous regions [1]–[4]. However, an HSR image is usually composed of diverse land-cover types with a complex spatial distribution, e.g., roads, buildings, and trees. It is therefore difficult for these methods to acquire the semantic meaning of a scene image, e.g., a residential scene or an industrial scene. Scene classification is aimed at labeling an HSR image according to the geographical properties to obtain regions using semantic information, and has recently attracted increased attention in HSR image understanding.

Scene classification methods based on object recognition utilize a relevant model to define the spatial relationship between different objects, which is an approach that requires the prior information of the objects [7], [8]. For this approach, the object recognition needs to be well designed, and the spatial relationship is difficult to model. The deep learning-based methods have turned out to be good at discovering the intricate structures hidden in high-dimensional data, and have shown an impressive feature representation ability for HSR image scene classification [9]. On the other hand, these methods usually require a large amount of training samples [10], and a small volume of training data tends to magnify the overfitting problem for the convergence of the network. However, large amounts of training samples are unusual for most remote sensing problems, since the acquisition of training data comes with a high cost, both in terms of time and money [11]. To overcome these problems, transfer learning with pre-trained convolutional networks (ConvNets) has been proposed for HSR image scene classification [43]–[46]. There are two main transfer learning methods: one uses the pre-trained ConvNets to extract the high-level features, whereas the other method uses the pretrained ConvNets to partially initialize the transferred ConvNets. There is no ConvNet

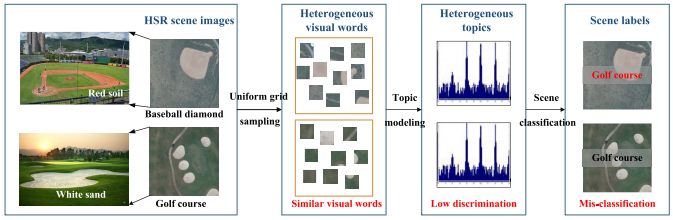


Fig. 1. Examples of baseball diamond and golf course scenes with similar visual words.

training process in the former xekmethod, which leads to a high speed. However, the separate training processes for the feature descriptor and the classifier weaken the classification performance. The latter method employs a fine-tuning strategy to improve the discriminative ability with a new set of HSR imagery, which yields better performance. On the other hand, this method is less efficient since it requires a little training over the parameters of the network [44]. In addition, transfer learning with pretrained ConvNets can be used only to process images with three channels, and this approach cannot be directly used for multispectral imagery scene classification. Due to its simplicity and effectiveness, the bag-of-visual-words (BOVW) model [12]–[14] is a popular method for representing HSR images with unordered local low-level descriptors. Based on the BOVW model, the probabilistic topic model (PTM) has been proposed and successfully applied to HSR scene classification [15]–[17], [42]. Among the PTMs, the fully sparse topic model (FSTM) [28] discovers sparse topics of the documents in linear time, which is a simple but efficient approach for complex LULC classification. However, the sparse topics mined by the FSTM may lose representative semantic information, and the FSTM does not perform well when directly applied to HSR scene classification.

To extract the representative features for HSR scene classification, most researchers have utilized a uniform grid sampling method as the first step [2], [12]–[16], [20]–[26]. The generated regions are usually heterogeneous, and heterogeneous visual words are then employed for the complex scenes. Fig. 1 shows an example of baseball diamond and golf course scenes. The baseball diamond and golf course scenes both contain road and grassland, which account for most areas of the scenes. The representative object in the baseball diamond scene is made up of red soil, whereas it is white sand for the golf course scene. For this type of scene, the homogeneous areas and the representative objects are critical for discriminating the scene from other scenes. However, after uniform grid sampling, the images are randomly split into a set of visual words, and many representative areas are mixed with others, which leads to heterogeneous visual words. In this way, the ratio of the representative visual words is very small, resulting in low discrimination.

Guided by this observation, the following questions are systematically considered. Can we design more adequate visual words to improve the discrimination? Can homogeneous information improve the scene classification performance? Superpixel segmentation methods, including graph-based and gradient-based approaches, are common ways to segment images into homogeneous areas. In this paper, a novel and effective framework that is called the sparse homogeneous–heterogeneous topic feature model (SHHTFM) for HSR image scene classification is proposed. In SHHTFM, the simple linear iterative clustering (SLIC) superpixel segmentation method [31] is first incorporated into the HSR scene classification framework as a region sampling strategy to efficiently generate homogeneous regions, where uniform grid sampling is simultaneously employed to generate heterogeneous regions. To exploit the diverse semantics in HSR imagery, three types of low-level features—the mean and standard deviation (MSD)-based spectral feature, the wavelet-based texture feature, and the scale-invariant feature transform (SIFT) features—are extracted from a set of regions. To circumvent the inadequate clustering capacity of the hard-assignment-based k-means clustering algorithm and the mutual interference between different types of topics, the different features are separately transferred to different topic spaces for the scenes. Moreover, the inference by SHHTFM improves the fusion of the heterogeneous and homogeneous topics (HHTs). In this way, the scene label of each image can be obtained, and a large HSR image can be annotated based on the prediction of the small images.

The scene annotation result based on pure remote sensing data can reflect only the natural properties of the land-cover objects, as the ambiguous, broken, and irregular boundaries may interfere with the identification of functional zones. However, an urban functional zone may be more concerned with the inner socioeconomic activities. In this way, scene understanding is a rich field, covering segmentation, object localization, classification, and annotation, to allow us to understand the remote sensing scene from local to global perspectives. The method of overlaying road network data has been implemented as a pretreatment for scene understanding [41], where the blocks derived by the road network data usually correspond to large areas. Most of these blocks are semantically composed of multiple types of scenes. Hence, in this paper, in order to obtain more accurate scene understanding results, the large-scale vegetation and water boundaries are extracted from the HSR imagery and are integrated with the road network data to complement the geographical data.

The main contributions of this paper are as follows.

1) To explore the effect of homogeneous information for HSR scenes, SLIC superpixel method is introduced as a novel region sampling strategy to segment the images into a set of homogeneous regions. Based on the semantics extracted from the homogeneous regions, the SHHTFM framework can decrease the confusion of the scenes with representative objects.
2) To capture the discriminative high-level semantics, the SHHTFM framework is proposed. In SHHTFM, the integration of the homogeneous visual words and heterogeneous visual words provides an adaptive feature description for distinct scenes. Based on separately mining the sparse topic spaces for the different types of features, the optimization-based inference task of SHHTFM improves the fusion of the HHTs.

These characteristics result in SHHTFM yielding efficient LULC scene classification results.

3) Based on the scene classification results of SHHTFM, multisource geographical data are effectively applied to obtain better scene understanding results. The water and vegetation boundaries define a more accurate way to restrict the boundaries of different scenes, and are integrated with the road network data to tackle the problems of LULC scene annotation.

Comprehensive evaluations on a challenging 21-class data set and a 12-class LULC data set and comparisons with the state-of-the-art approaches demonstrate the effectiveness and superiority of the SHHTFM framework. Scene annotation of a large LULC image also confirms the effectiveness of SHHTFM. In addition, the combination of multisource geographical data with the scene annotation results is experimentally verified to be both reliable and applicable for urban planning.

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

II. Related Work

SHHTFM is closely related to the BOVW model and its classical PTM variants, e.g., probabilistic latent semantic analysis (pLSA), latent Dirichlet allocation (LDA), and the FSTM. The BOVW model is considered as an effective method for LULC HSR image scene classification. The BOVW model-based scene classification is dependent upon the local features of the images, which are then quantized into visual words. In this way, the BOVW model models the scene only by the simple statistics of the frequency of each visual word, whereas scene classification may require more information about the higher level of semantic information.

A. Scene Classification Based on the PTMs

The PTMs, including the classical pLSA [18] and LDA [19] models, are able to map the low-level features in the heterogeneous regions to high-level semantic concepts, and can reduce the dimensionality of HSR images. The strategies of using PTMs for HSR scene classification can be divided into several types. In general, a single feature is utilized to describe the visual words. The spectral feature (MSD) as the feature descriptor with LDA has also been proposed to describe HSR images [15], [20]. Besides LDA, SIFT with the Markov field topic model [21] and pLSA [22] have also been proposed. However, one single type of feature is always inadequate to capture the entire scene, since scene information is usually conveyed by multiple cues, e.g., color, shape, structural, or texture features. It is widely accepted that multiple features should be adaptively fused with the PTM to discriminate each class from the others. Multifeature-based LDA and pLSA have been proposed to incorporate different features to capture the various aspects of complex scenes and improve the performance of scene classification [16], [23]–[25].

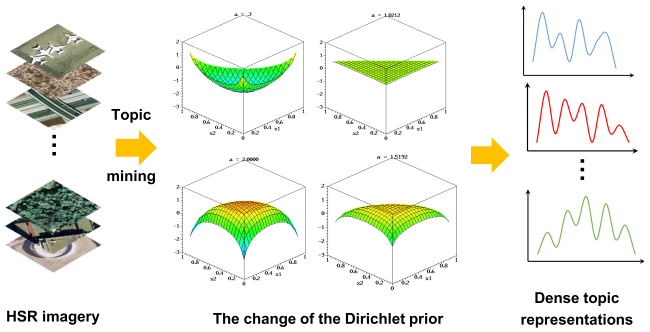


Fig. 2. Dirichlet distribution for LDA.

The commonly used PTM, i.e., LDA, treats the topic mixture parameters as variables drawn from a Dirichlet distribution. However, with the change of the Dirichlet parameter α, the topic variables acquired from the HSR imagery are always greater than zero (Fig. 2). The latent semantics mined from images by LDA are often dense, which results in a lot of useless information and requires a lot of storage space. The topic modeling is therefore complex and takes a lot of time. In an attempt to overcome the issue of dense topics for modeling the scenes, sparsity constraints have been imposed on the topics to change the objective function of the PTM [26], [27]. However, these models usually require model selection with many regularization term-based auxiliary parameters, which may be problematic with large-scale data sets.

Instead of the Dirichlet prior and sparsity constraints, the FSTM [28] was proposed, which models the documents with a sparse prior. In addition to the natural language processing domain, the FSTM has been successfully applied to content-based video retrieval [29] and abnormality detection in traffic videos [30]. FSTM is a hierarchical model, and introduces sparse topics to analyze the words in the documents from a corpus. When FSTM used in the text analysis field is applied to the image data set, an analogy between their respective terminologies is defined as follows.

1) A corpus is equivalent to an image data set.
2) A document corresponds to an image.
3) A word is equivalent to a patch or segmented region of an image, which is usually called a visual word.

Based on the bag-of-words assumption, the order of the visual words in an HSR image is ignored in FSTM. To reduce the dimension of representing the images, a k-means clustering is employed to construct the visual dictionary, where the number of visual words is the size of the vocabulary.

For a given HSR image G_i composed of a sequence of visual words, it can be represented by K topics. The generative process of FSTM for the HSR image is as follows.

1) For each image, choose a topic proportion θ.
2) For each visual word in an image G_i, select a latent topic t_k with probability $P(t_k|G_i) = \theta_k$, and choose a visual word v_j with probability $P(v_j|t_k) = \beta_{kj}$, where $j \in I_d$, and I_d is the set of term indices of image G_i.

The likelihood of an HSR image with FSTM is given by

$$\log P(G_i) = \sum_{j \in I_d} G_{ij} \log \sum_{k=1}^{K} \theta_k \beta_{kj}. \quad (1)$$

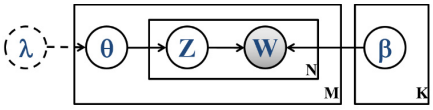


Fig. 3. Probabilistic graphical model of the FSTM.

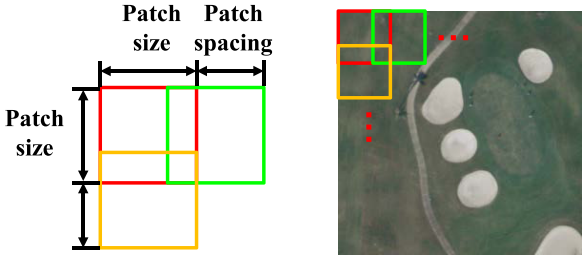


Fig. 4. Uniform grid-based region sampling method for HSR imagery.

Compared with LDA, there is no explicit distribution to model the topics from the HSR image. In fact, an implicit prior of the topic distribution does exist, having the density function of (2). In more detail, the latent topic proportion θ in the FSTM follows an implicit constraint $||\theta||_0 \leq L + 1$, where L is the iteration times. The number of nonzero entries of θ is denoted by $||\theta||_0$. This property of "implicit modeling" allows the FSTM to be able to converge at a linear rate to the optimal solution. The probabilistic graphical model of the FSTM is shown in Fig. 3. In this paper, the FSTM is chosen to model the HSR scenes with sparse topics

$$p(\theta|\lambda) \propto \exp(-\lambda \cdot ||\theta||_0). \tag{2}$$

B. Region Sampling Based on Uniform Grid Sampling

As a preprocessing step for the extraction of low-level features, region sampling is essential for HSR imagery scene classification. Region sampling is aimed at selecting a representative subset for the HSR image. Accordingly, it is crucial to appropriately sample the imagery to generate discriminative features for complex scenes. Previous researchers have divided the existing sampling methods for HSR imagery into random sampling and saliency-based sampling methods [35]. Here, the saliency-based sampling method consists of keypoint-based sampling and salient region-based sampling, and the uniform grid sampling is included as a specific type of random sampling method. The uniform grid sampling method is the most commonly used sampling method for HSR imagery, and it is chosen as one of the sampling strategies in this paper. As shown in Fig. 4, given an image, the proposed SHHTFM assumes that each image of a scene can be represented by a set of sampled patches from the image. Here, a patch is a local rectangular image region that can be used for feature description in the following step. SHHTFM employs the uniform sampling strategy with two parameters—patch size and patch spacing—to obtain the set of patches to represent the image. The patch spacing determines the frequency of the sampling. Compared with the original image, the sampled subset of patches is more compact and less complex.

C. SLIC Superpixel Sampling

Various superpixel classification methods have been proposed and have been proven to be effective in image segmentation for diverse images. In this paper, SLIC [31] is employed to aggregate nearby pixels into superpixels in the HSR image scenes, for its simplicity, memory efficiency, and excellent boundary adherence [34]. The only parameter of SLIC is the desired number of approximately equally sized superpixels, which is denoted by p. In SLIC, the HSR image with N pixels is converted into a 5-D vector in the CIELAB color space. Then p initial cluster centers are sampled based on a regular grid, which is spaced $S = \sqrt{N/p}$ pixels apart. These centers are moved to the positions with the lowest gradient in a 3×3 neighborhood, which avoids centering a superpixel on an edge. Differing from the conventional k-means clustering, each pixel in SLIC is associated only with the nearest cluster center whose search region overlaps its location, where the search region is set as $2S \times 2S$ around the superpixel center. The new cluster center is iteratively calculated for each pixel with a distance measure D, which is based on the color and spatial proximity. In addition, an update step adjusts the cluster centers to be the mean vector of all the pixels belonging to the cluster. The iteration continues until the residual error between the new cluster center and the previous ones converges. In this way, SLIC significantly reduces the number of distance calculations, and is very fast. Finally, the disjoint pixels are reassigned to nearby superpixels to enforce connectivity.

III. Scene Classification Based on the Sparse Homogeneous–Heterogeneous Topic Feature Model

To effectively employ the representative semantics, the SHHTFM framework is proposed for HSR image scene classification. Four tasks have to be addressed: 1) region sampling based on the union of uniform grid sampling and SLIC superpixel sampling; 2) heterogeneous and homogeneous visual word generation; 3) sparse topic representation; and 4) improved sparse topic fusion and classification. The scene classification results are then integrated with multisource geographical data to obtain an applicable scene understanding map. The overall flowchart of scene classification and understanding based on the SHHTFM framework is shown in Fig. 5.

A. Region Sampling Based on the Union of Uniform Grid Sampling and SLIC Superpixel Sampling

As the uniform grid-based region sampling method usually generates heterogeneous information for HSR imagery, can the homogeneous information also be mined from the images to improve the scene classification accuracy? Superpixel segmentation is the common way to mine the homogeneous information of images. Accordingly, in this paper, the simple and effective SLIC segmentation algorithm is used to construct a set of compact and homogeneous patches. The nearby pixels in the image are aggregated into superpixels by the SLIC-based sampling of SHHTFM. SHHTFM employs the SLIC-based sampling strategy with two parameters: region size and regularizer. The region size refers to the expected spatial extent of a superpixel. The regularizer is a constant that allows us to weight the relative importance between

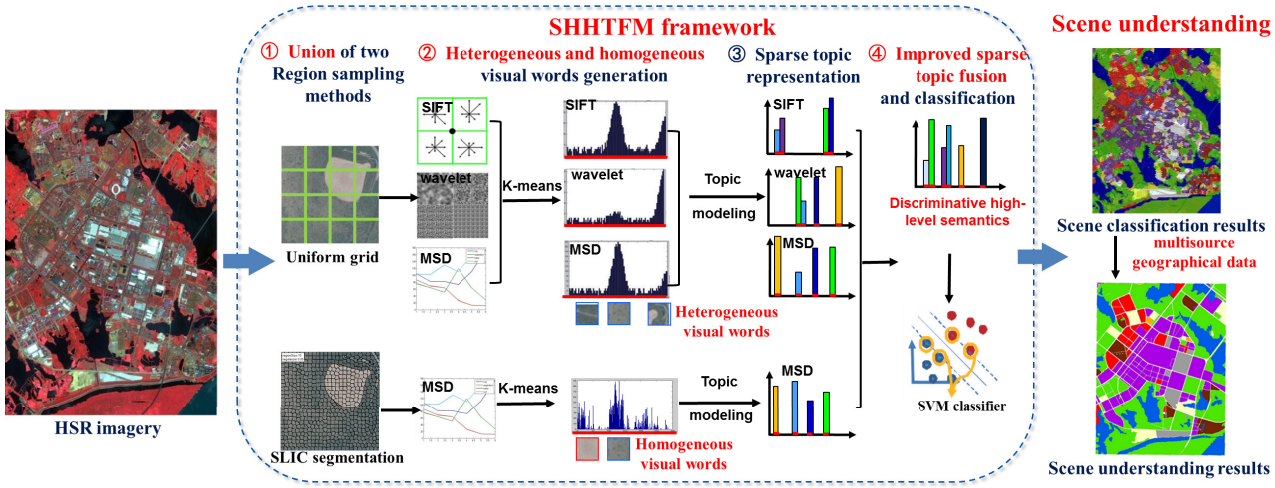


Fig. 5. Flowchart of scene classification and understanding based on the SHHTFM framework for HSR imagery, where the descriptions in red indicate the main contributions of the SHHTFM framework.

color similarity and spatial proximity. When the regularizer is large, the spatial proximity becomes more important, and the acquired superpixels are more compact and have a lower area to perimeter ratio. When the regularizer is small, the acquired superpixels adhere more tightly to image boundaries, but their size and shape are less regular. SHHTFM uses an adequate region sampling strategy to make up for the shortcomings of the traditional uniform grid sampling methods, and is more appropriate for complex HSR scene classification.

B. Heterogeneous and Homogeneous Visual Word Generation

After acquiring two sets of representative image patches, i.e., IP_{he} and IP_{ho}, by the combination of uniform grid sampling and SLIC superpixel sampling, SHHTFM utilizes a visual analog of a word, acquired by vector quantizing the region descriptors [18]. With diverse spatial configurations and complex details, HSR scenes are usually hard to recognize if not enough information is discovered. Two scenes that are made up of the same objects, e.g., buildings, trees, and roads, may result in different scene types since they have different spatial distributions. These diverse semantic concepts lead to different scene labels, e.g., commercial scene and residential scene (Fig. 6). In this way, the use of the FSTM to discover very sparse topics may result in the loss of some representative semantics. Hence, in this paper, SHHTFM is proposed to comprehensively capture the images using three feature descriptors, i.e., the MSD, wavelet, and SIFT features.

1) The MSD features refer to the first-order statistics of the mean value and the second-order statistics of the standard deviation value of the image patches. The MSD features are calculated in each spectral channel as the spectral feature, which reflects the attributes of the HSR images that constitute the ground components and structures.
2) To compensate for the deficiency of the statistical MSD features, the texture feature is utilized to describe the images. The texture feature contains information about the spatial distribution of the tonal variations within a band [36], which considers both the macroscopic properties and fine structure. Among the various texture features, wavelets discover information from an image about both the spatial and frequency content, and thus they can be adopted to analyze texture for nonstationary or nonhomogeneous images, such as HSR remote sensing images [37]. By decomposing the image into different frequency sub-bands, wavelet transforms are similar to the way the human visual system operates [38], which infers that they are suitable for image classification. SHHTFM employs multilevel 2-D wavelet decomposition to extract the texture feature, and the level of the wavelet decomposition is optimally set to three.
3) The SIFT feature [39] is able to overcome noise, affine transformation, and changes in illumination, and has been widely applied in image analysis. Gray dense SIFT is employed as the patch descriptor in SHHTFM, which was inspired by [40]. The image patches are divided into 4×4 neighborhood regions, where the gradient orientation histograms of eight directions are counted; $4 \times 4 \times 8 = 128$-D vectors are finally acquired to describe the keypoint descriptor.

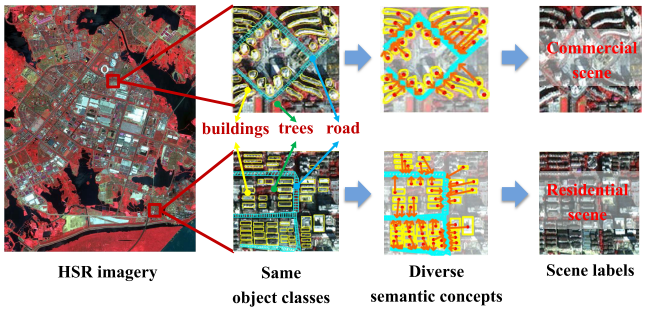


Fig. 6. Examples of two different scenes from HSR imagery.

The information in the image patches obtained by uniform grid sampling is usually heterogeneous. Hence, the distinct features, i.e., the MSD, wavelet, and SIFT features, are separately employed to describe the image patches. On the

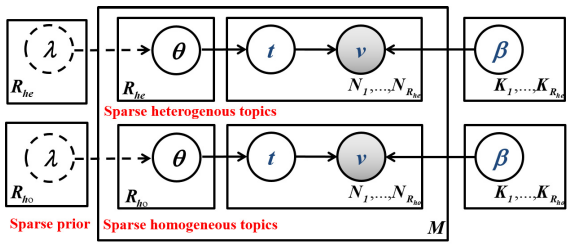


Fig. 7. Probabilistic graphical model of SHHTFM.

other hand, the information in the image patches obtained by SLIC segmentation is homogeneous, and the scenes with representative objects that are prone to confusion, e.g., baseball diamond with red soil and golf course with white sand, mainly differ in the spectral feature for the representative image patches. In addition, the spectral feature usually performs best for describing the HSR images according to empirical research. Hence, give consideration to both time efficiency and satisfactory performance and also to avoid feature redundancy, only the MSD feature is employed to describe the patches generated by SLIC sampling. In this way, the image patches acquired by region sampling of the HSR images are digitized by R types of features, and all the types of feature descriptors, $d_1, \ldots, d_R$, are obtained. The influence of illumination, rotation, and scale variation may lead to the same visual word in different images being endowed with different feature values. Accordingly, the feature descriptors are quantized by k-means clustering to generate R 1-D frequency histograms, where image patches with similar feature values correspond to the same visual word. By statistical analysis of the frequency of each visual word, the corresponding visual dictionary can be obtained.

C. Sparse Topic Representation

After the heterogeneous and homogeneous visual word generation, SHHTFM analyzes the different visual words by introducing probability theory, and the sparse topics are discovered to represent the semantic information. SHHTFM can greatly reduce the dimension of the feature vectors for the representation of HSR images. The probabilistic graphical model of SHHTFM is displayed in Fig. 7. Specifically, given M images G_i, they can each be described by N visual words v_j, where $G_i = \{v_1, \ldots, v_j, \ldots, v_N\}$. R_{he} and R_{ho} types of features are then extracted from the image sets IP_{he} and IP_{ho}, and separately quantized by the k-means clustering algorithm. In this way, the inadequate clustering capacity of the hard-assignment-based k-means algorithm is circumvented, and R_{he} and R_{ho} histograms are acquired and then transformed into word occurrence probability matrices. SHHTFM separately models the word occurrence probability matrices as random mixtures over the latent variable space. By choosing a K-dimensional latent variable θ, $K_1, \ldots, K_{R_{\text{he}}}$ and $K_1, \ldots, K_{R_{\text{ho}}}$ topics are selected from each histogram of the two image sets IP_{he} and IP_{ho}, respectively. This circumvents the mutual interference between the heterogeneous and homogeneous features, and allows the heterogeneous and homogeneous features to adequately describe the HSR images in the different latent topic spaces.

For each type of feature, given K topics $\beta = (\beta_1, \ldots, \beta_K)$, the log likelihood of an image G_i can be decomposed as shown in (2), where G_{ij} is the frequency of visual word v_j in image G_i. In the inference procedure, SHHTFM sets $x_j = \sum_{k=1}^{K} \theta_k \beta_{kj}$ and $x = (x_1, \ldots, x_D)^t$, where D is the size of the visual dictionary. Differing from other PTMs, the latent variables are not directly inferred. SHHTFM treats the inference of optimizing the latent variables as a concave maximization problem over the simplex $\Delta = \text{conv}(\beta_1, \ldots, \beta_K)$. Following $\sum_k^{\theta_k} = 1$, where $\theta_k \geq 0$, x is a convex combination of the K topics $\beta = (\beta_1, \ldots, \beta_K)$. The Frank–Wolfe algorithm, which follows the greedy approach, is employed as the inference algorithm. The latent variable can be denoted by $\theta_{l+1} := (1 - \alpha')\theta_l$ after L iterations, where α' is defined in (3), and is solved by the gradient ascent approach. Here, $\beta_{y'}$ denotes the standard unit vectors in the simplex Δ, and x_l is a convex combination of at most $L + 1$ vertices of the simplex, which is defined in (4). This implies an implicit constraint $||\theta||_0 \leq L + 1$ in SHHTFM, which shows that at most $L + 1$ out of the latent variables are nonzero. By finding the $x \in \Delta$ that maximizes the objective function, the latent variable θ of image G_i can be inferred by converging at a linear rate to the optimal solution, which is a sparse solution

$$\alpha' := \arg \max_{\alpha \in [0,1]} f(\alpha \beta_{y'} + (1 - \alpha) x_l) \tag{3}$$

$$x_l = \sum_{k=1}^{K} \theta_{lk} \beta_k. \tag{4}$$

Based on the latent variable θ, an expectation–maximization scheme is executed to iteratively learn all the topics $\beta = (\beta_1, \ldots, \beta_K)$. In more detail, SHHTFM begins the E-step with the inference of the latent topic distributions, and undertakes the M-step to maximize the likelihood of the M images with regard to β. Accordingly, the lower bound of the log likelihood of the M images is maximized in the learning procedure. By taking into consideration of the Lagrangian multipliers, the topics $\beta = (\beta_1, \ldots, \beta_K)$ are obtained, as written in (5). In this way, SHHTFM models the topics from the HSR imagery by iterating the E-step and M-step until convergence. The sparse topic proportions $\theta_1, \ldots, \theta_{R_{\text{he}}}$ and $\theta_1, \ldots, \theta_{R_{\text{ho}}}$ are then obtained for all the types of features from the heterogeneous and homogeneous image patches, respectively

$$\beta_{kj} \propto \sum G_j \theta_k. \tag{5}$$

D. Improved Sparse Topic Fusion and Classification

In the task of semantic-level fusion, the MSD-based topic feature for the homogeneous regions, denoted by θ_1, and the MSD, wavelet, and SIFT-based topic features for the heterogeneous regions, denoted by $\theta_2, \theta_3, \theta_4$, are effectively fused at the semantic level. The inference task in SHHTFM reformulates the optimization of the latent topic distribution as a concave maximization problem, which leads to improved fusion of the distinct types of topics. In this way, the final sparse semantic representation can be denoted by $F = \{\theta_1^T, \theta_2^T, \theta_3^T, \theta_4^T\}^T$. In the task of HHT classification, the F with the discriminative semantics is classified by the support vector machine

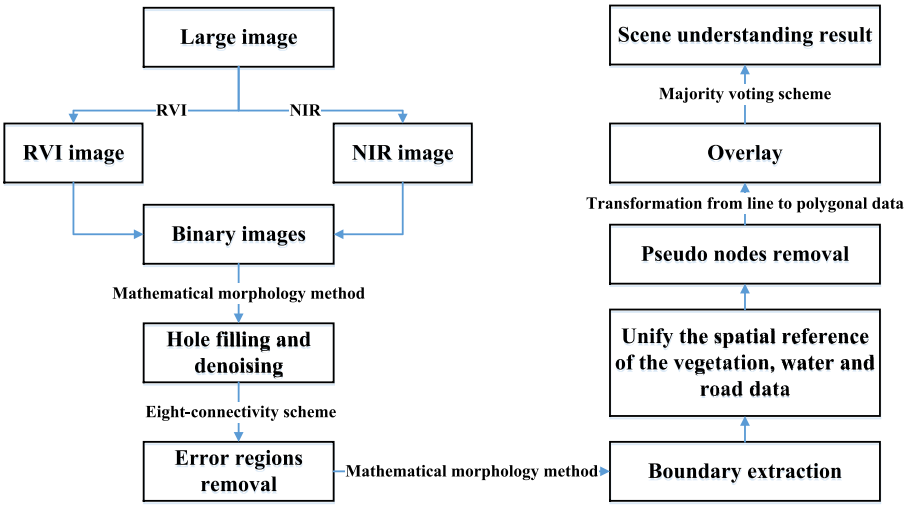


Fig. 8. Procedure of scene understanding combining multisource geographical data.

classifier with a histogram intersection kernel (HIK). By measuring the degree of similarity between two histograms, the HIK deals with the scale change, and has been applied to image classification using color histogram features. Assuming $\tilde{\mathbf{V}} = (\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \ldots, \tilde{v}_M)$ to be the SHHTFM representation vectors of M images, the HIK is calculated as shown in (6). Finally, the scene label of each image can be predicted

$$K(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_j) = \sum_k \min(\tilde{v}_{i,k}, \tilde{v}_{j,k}). \tag{6}$$

E. Scene Understanding Combining Multisource Geographical Data

To date, sensing technologies and large-scale computing infrastructures have produced a variety of big data in urban spaces (e.g., human mobility, air quality, traffic patterns, and geographical data) [32]. Among the big data, road network data, water boundaries, and vegetation boundaries are overlaid on the scene annotation results based on the scene labels predicted by SHHTFM. Specifically, the procedure of scene understanding combining the multisource geographical data is shown in Fig. 8. As can be seen in Fig. 8, the vegetation is acquired with the use of the ratio vegetation index, and the water is acquired by extracting the region with the use of the near-infrared band from the original large image. What we need is large connected regions of vegetation and water, instead of fragmented regions. Hence, a closing operation based on the mathematical morphology method is used to fill the small holes in the image, and an opening operation is used for denoising. However, the small regions of the urban landscape, which may belong to residential or commercial scenes, are also extracted. To avoid the interruption of these pixels, an eight-connectivity scheme is employed to label the connected regions. When the number of pixels in the connected region is smaller than 10 000, this region will be excluded and removed. The mathematical morphology method is then used to obtain the boundaries of the remaining regions. To overlay the raster data of the vegetation and water boundaries and the vector data of the road network data, their spatial references and data format are unified, and the raster data of the vegetation and water boundaries is transformed to vector data. In addition, some of the topological errors of the data are removed, such as removing pseudo nodes. Finally, the line data are transformed into polygonal data to form the final geographical blocks, which can be overlaid on the scene annotation result. The majority voting method is employed to justify which scene category the final geographical block belongs to. Combining scene annotation with geographical data is not only more reliable and meaningful in scene classification but also makes the results suitable for direct application, by urban planning departments and others.

IV. Experiments and Analysis

A. Experimental Setup

In order to test the performance of SHHTFM, the commonly used 21-class UC Merced data set and the 12-class Google data set of the scene image data set designed by the Intelligent Data Extraction and Analysis of Remote Sensing (RS_IDEA) Group in Wuhan University (SIRI-WHU) were evaluated in the experiments. In addition, an original large image of the Wuhan IKONOS data set was also used to test the scene annotation performance of SHHTFM. The multiple types of geographical data were applied to the scene annotation results for better scene understanding. In the experiments with uniform grid-based region sampling, the patch size and spacing were optimally set to 8×8 and 4×4 pixels, respectively. In the experiments with SLIC superpixel-based region sampling, the region size and regularizer were optimally set to 10 and 0.05 for the UC Merced data set, respectively; 10 and 0.01 for the Google data set of SIRI-WHU, respectively; and 15 and 0.05 for the Wuhan IKONOS data set, respectively. The visual dictionary with V visual words was constructed by employing Euclidean distance measurement-based k-means clustering over the image patches from the training data. The different methods were implemented 100 times by randomly selecting the training samples, to ensure that convincing results were obtained and the stability of the proposed SHHTFM could be tested. There are two free parameters in the proposed SHHTFM: the visual word number V and the topic

TABLE I

OPTIMAL K AND V VALUES FOR THE DIFFERENT METHODS WITH THE UC MERCED DATA SET

	SFSTM-HET	SFSTM-HOM	MFFSTM-HET	SHHTFM
V	1000	1000	2800	3800
K	240	800	840	1640

TABLE II

OPTIMAL K AND V VALUES FOR THE DIFFERENT METHODS WITH THE GOOGLE DATA SET OF SIRI-WHU

	SFSTM-HET	SFSTM-HOM	MFFSTM-HET	SHHTFM
V	1000	1000	2800	3800
K	240	700	870	1570

TABLE III

OVERALL CLASSIFICATION ACCURACY (%) COMPARISON WITH THE UC MERCED DATA SET

pLSA	89.51±1.31
LDA	81.92±1.12
Cheriyadat [6]	81.67±1.23
Zhao *et al*. [25]	92.92±1.23
Yao *et al*. [33]	93.57±1.02
Scenario (II) [45]	96.90±0.77
Fine-tuned GoogLeNet [44]	97.78±0.97
Fine-tuned GoogLeNet descriptors [44]	99.47±0.50
VGG-VD16+AlexNet [46]	98.81±0.38
SFSTM-HET	78.33±1.42
SFSTM-HOM	80.00±1.52
MFFSTM-HET	95.71±1.01
SHHTFM	**98.33±0.98**

number K, where V and K are determined according to the sensitivity analysis for the different data sets in Experiment 3. Taking the classification accuracy and the computational complexity into consideration, V and K were optimally set as shown in Tables I and II for the different feature strategies with the two data sets. In Tables I–IV, SFSTM-HET and SFSTM-HOM denote FSTM-based scene classification utilizing the MSD-based spectral features from heterogeneous and homogeneous image patches, respectively. MFFSTM-HET denotes FSTM-based scene classification utilizing multiple features, including the MSD-based spectral feature, the wavelet-based texture feature, and the SIFT features, from heterogeneous image patches. Different from MFFSTM-HET, the proposed method utilizes both the multiple features from heterogeneous patches and the spectral features extracted from homogeneous regions, which is generated by SLIC segmentation. Hence, by comparing the proposed method with MFFSTM-HET, we can analyze whether the homogeneous information is able to improve the scene classification performance.

The performance of the proposed framework is evaluated using the overall accuracy (OA). The OA is calculated as the total number of correctly classified scene images divided by the total number of test images, which indicates how well

TABLE IV

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

pLSA	89.60±0.89
LDA	60.32±1.20
SAL-LDA	90.65±1.05
Zhao *et al*. [25]	91.52±0.64
SFSTM-HET	78.33±1.42
SFSTM-HOM	80.00±1.52
MFFSTM-HET	97.83±0.93
SHHTFM	**99.25±0.88**

the model predicts the actual data. In addition, the confusion matrices allow visualization of the performance of a framework [Figs. 12, 15, and 18(b)]. In these confusion matrices, each row represents the proportion of a scene in an actual class, while each column represents the proportion of a scene in a predicted class. This allows a more detailed analysis than the proportion of the overall classification accuracy. All the experiments were run on a personal computer with a single Intel core i3 CPU, an NVIDIA Quadro 600 GPU, and 8 GB of memory. The operating system was Windows 10, and the implementation environment was under MATLAB 2012a. The SHHTFM framework is divided into four steps in Section III, where the computational loads of the first and the fourth step are very small and can be ignored. The computational load of the second heterogeneous and homogeneous visual word generation step is about 25%–35% of the CPU and 800–1500 MB of the memory, and the third topic modeling step is about 22%–27% of the CPU and 14–17 MB of the memory. The whole process of the SHHTFM framework takes around 1 h to be completed.

To further evaluate the performance of SHHTFM, the experimental results obtained with the conventional MFFSTM-HET, pLSA [17], and LDA [15] are shown for comparison. We also provide the experimental results obtained with the UC Merced data set, as published in [6], [14], [25], [33], [44], [45], and [46]. In addition, the experimental results obtained with the Google data set of SIRI-WHU with the conventional MFFSTM-HET, pLSA [17], LDA [15], and SAL-LDA [23] are shown for comparison, along with the experimental results published in [25].

B. Experiment 1: The UC Merced Image Data Set

The UC Merced data set was downloaded from the USGS National Map Urban Area Imagery collection [14]. This data set consists of 21 land-use scenes (Fig. 9), 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 [14], 80 samples were randomly selected per class from the UC Merced data set for training, and the rest were kept for testing.

The classification performance of the single feature-based SFSTM-HET and SFSTM-HOM, the conventional

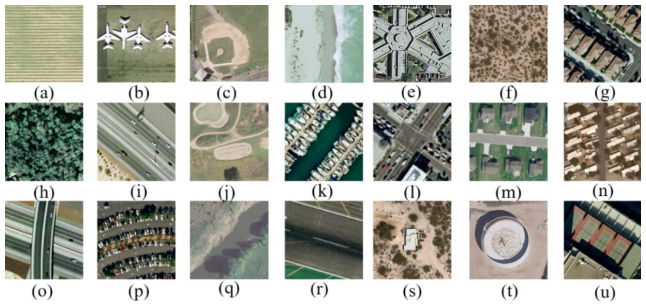


Fig. 9. UC Merced data set. (a) Agricultural. (b) Airplane. (c) Baseball diamond. (d) Beach. (e) Buildings. (f) Chaparral. (g) Dense residential. (h) Forest. (i) Freeway. (j) Golf course. (k) Harbor. (l) Intersection. (m) Medium residential. (n) Mobile home park. (o) Overpass. (p) Parking lot. (q) River. (r) Runway. (s) Sparse residential. (t) Storage tanks. (u) Tennis courts.

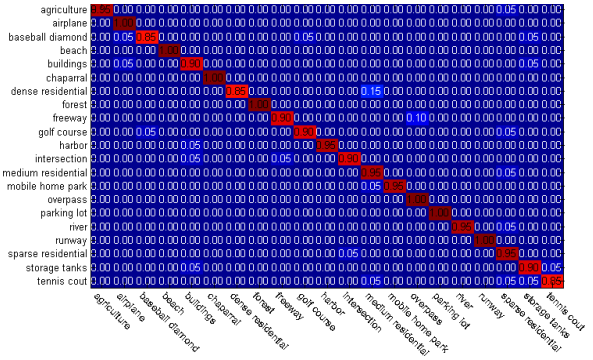


Fig. 10. Confusion matrix of MFFSTM-HET with the UC Merced data set.

MFFSTM-HET, the proposed SHHTFM, and the experimental results of previous methods for the UC Merced data set are reported in Table III. As can be seen in Table III, the classification results of the homogeneous feature-based SFSTM-HOM are slightly better than those of the heterogeneous feature-based SFSTM-HET, which indicates that the homogeneous topics are also valuable for HSR image scene classification. The classification results of MFFSTM-HET are better than those of the multiple heterogeneous feature-based LDA model, which shows the effectiveness of the sparse semantic representation for scene classification. However, the results of SFSTM-HOM, SFSTM-HOM, and MF-HEFST are all unsatisfactory. The classification accuracy for the proposed SHHTFM, $98.33\% \pm 0.98\%$, is the best among all the different methods. This indicates that the combination of the HHT description and the improved semantic fusion capacity is able to provide discriminative image representations for scene classification. In addition, it can be seen that SHHTFM performs better than the current state-of-the-art methods, i.e., the midlevel feature-based methods of pLSA and LDA, the methods in [6], [14], and [25], and the deep learning-based methods of the fine-tuned GoogLeNet approach [44], the Scenario (II) approach [45], and the method in [33]. The performance of the fine-tuned GoogLeNet approach is better than that of the Scenario (II) approach with no fine-tuning process, which confirms the superiority of the fine-tuned ConvNets. The proposed SHHTFM achieves better results than the fine-tuned GoogLeNet and Scenario (II) approaches, which indicates that SHHTFM performs better than the feature descriptor based or fine-tuning-based transfer learning with pretrained ConvNets. On the other hand, the proposed method performs slightly worse than the VGG-VD16 + AlexNet approach and the fine-tuned GoogLeNet descriptors approach [44], which can be explained. The VGG-VD16 + AlexNet approach [46] integrates many kinds of features, including the three scales of the image-based features extracted from the five convolutional and three fully connected layers of AlexNet and the three scales of the image-based features extracted from the last five convolutional and three fully connected layers of VGG-VD16, whereas SHHTFM fuses only the simple spectral, texture, and SIFT features. In addition, the computational requirements of the VGG-VD16 + AlexNet approach and the fine-tuned GoogLeNet descriptors approach are higher than that of SHHTFM, whereas the running time of SHHTFM is much shorter. This demonstrates that the homogeneous information can compensate for the heterogeneous semantics, and the proposed method is a highly efficient framework with a low computational cost.

An overview of the performance of MFFSTM-HET and SHHTFM is shown in the confusion matrices in Figs. 10 and 11, respectively. As can be seen in Fig. 11, most of the scene categories can be fully recognized by SHHTFM. Compared with the confusion matrix of MFFSTM-HET, the scene categories in the confusion matrix of SHHTFM obtain a better performance. For example, the baseball diamond, storage tanks, harbor, golf course, medium residential, mobile home park, river, sparse residential, and freeway scenes, which are confused in MFFSTM-HET, are fully recognized by SHHTFM. There is, however, some confusion between certain scenes. For instance, a scene belonging to dense residential is classified as medium residential. This can be explained by the fact that the two categories have the same objects and similar spatial distributions.

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

The Google data set of SIRI-WHU[1] was acquired from Google Earth (Google, Inc.), covering urban areas in China, and the scene image data set was designed by the RS_IDEA Group in Wuhan University [12], [25]. 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. 12. Each class separately contains 200 images, which were 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 performance of the single feature-based SFSTM-HET and SFSTM-HOM, the conventional MFFSTM-HET, the proposed SHHTFM, and the experimental results of previous methods for the Google data set of

[1]The Google data set of SIRI-WHU can be downloaded at http://www.lmars.whu.edu.cn/prof_web/zhongyanfei/e-code.html.

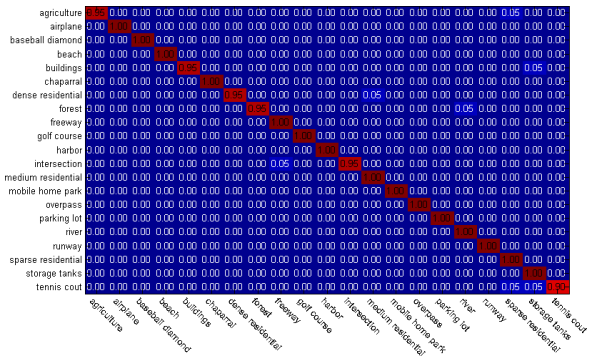


Fig. 11. Confusion matrix of SHHTFM with the UC Merced data set.

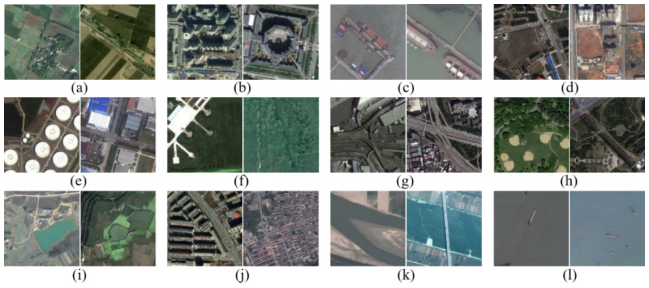


Fig. 12. Google data set of SIRI-WHU. (a) Agriculture. (b) Commercial. (c) Harbor. (d) Idle land. (e) Industrial. (f) Meadow. (g) Overpass. (h) Park. (i) Pond. (j) Residential. (k) River. (l) Water.

SIRI-WHU are reported in Table IV. As can be seen from Table IV, SFSTM-HOM obtains a slightly better performance than SFSTM-HET, which indicates that the MSD features extracted in the homogeneous regions are more discriminative for the HSR images. The classification result for the proposed SHHTFM, 99.25% $\pm$ 0.88%, is better than the results of the SFSTM-HOM, SFSTM-HOM, and MF-HEFST methods, which confirms that SHHTFM is an effective approach for HSR image scene classification. In Table IV, compared with the other methods, i.e., SAL-LDA [23], the LDA method proposed in [15], the pLSA method proposed in [17], and the experimental results published in [25], the highest accuracy is acquired by the proposed SHHTFM.

Figs. 13 and 14 display the confusion matrices of MFFSTM-HET and SHHTFM for the Google data set of SIRI-WHU, respectively. On the whole, most of the scene classes achieve good classification performances with SHHTFM. Compared with the confusion matrix of MFFSTM-HET, the performances of most of the scene categories, i.e., the harbor, industrial, overpass, and river scenes, are improved.

D. Experiment 3: Sensitive Analysis for Scene Classification

To study the sensitivity of MFFSTM-HET and SHHTFM in relation to the visual word number V, the values of the patch size and the patch spacing were kept constant at eight and four, respectively. The topic number K and the values of the region size and the regularizer for the SLIC sampling were optimally set to 1640, 10, and 0.05 for the UC Merced data set and 1570, 10, and 0.01 for the Google data set of SIRI-WHU. The visual word number V was then varied over the range of [2300, 2800, 3300, 3800, 4300] for the UC Merced data set and the Google

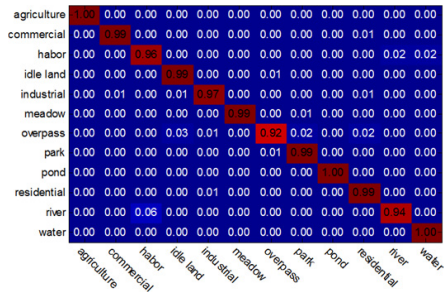


Fig. 13. Confusion matrix of MFFSTM-HET with the Google data set of SIRI-WHU.

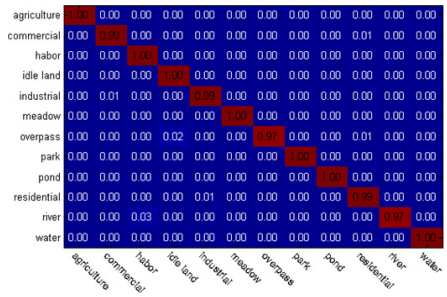


Fig. 14. Confusion matrix of SHHTFM with the Google data set of SIRI-WHU.

data set of SIRI-WHU. As shown in Fig. 15, with the increase in the visual word number V, the OA of SHHTFM is higher at the beginning and then tends to decline, whereas the OA curve of MF-HFFSTM displays fluctuation. It is notable that SHHTFM is superior to MF-HFFSTM over the entire range for the two data sets, which infers that the proposed SHHTFM can outperform the traditional heterogeneous feature-based methods.

To investigate the sensitivity of MFFSTM-HET and SHHTFM in relation to the topic number K, the values of the patch size, the patch spacing, and the visual word number V were kept constant at 8, 4, and 3800, respectively. The values of the region size and the regularizer for the SLIC sampling were optimally set to 10 and 0.05 for the UC Merced data set and 10 and 0.01 for the Google data set of SIRI-WHU. The topic number K was then varied over the range of [700, 1000, 1300, 1600, 1900] for the UC Merced data set and Google data set of SIRI-WHU. As can be seen in Fig. 16, SHHTFM obtains the best performance when K is 1600, while MFFSTM-HET demands fewer topics. This indicates that SHHTFM employing HHTs can effectively capture more semantic information for HSR scene classification. Comparing Figs. 15 and 16, it can be seen that the OA curves of SHHTFM and MFFSTM-HET display a smaller fluctuation in relation to the topic number K, and they are more sensitive to the visual word number V. Hence, the number of topics can be set as fixed at first to determine the optimal number of visual words. The topic number can then be varied to determine the optimal number of topics.

To investigate the time efficiency of the proposed SHHTFM and the conventional MFFSTM-HET and SAL-LDA, the values of the patch size, the patch spacing, the region size, the regularizer, and the visual word number V were kept

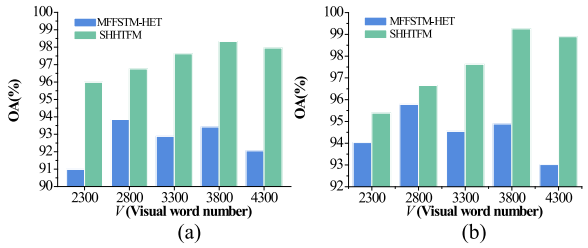


Fig. 15. Sensitivity analysis of MFFSTM-HET and SHHTFM in relation to the visual word number V. (a) UC Merced data set. (b) Google data set of SIRI-WHU.

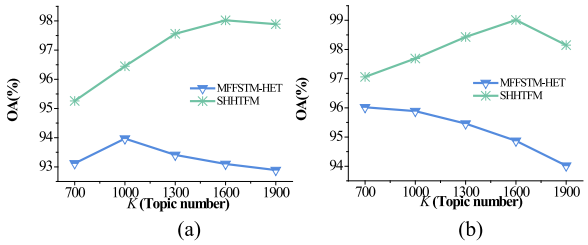


Fig. 16. Sensitivity analysis of MFFSTM-HET and SHHTFM in relation to the topic number K. (a) UC Merced data set. (b) Google data set of SIRI-WHU.

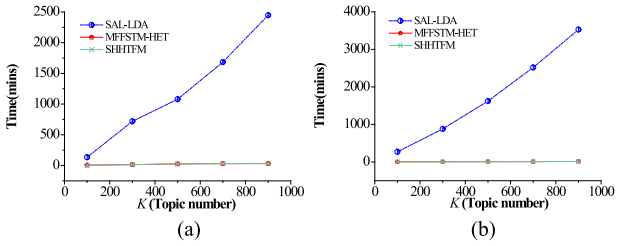


Fig. 17. Sensitivity analysis of the time efficiency for MFFSTM-HET and SHHTFM in relation to the topic number K. (a) UC Merced data set. (b) Google data set of SIRI-WHU.

at the optimal parameter settings, respectively. The topic number K was then varied over the range of [200, 400, 600, 800, 900] for the UC Merced data set and the Google data set of SIRI-WHU. As can be seen from Fig. 17, the topic modeling time of SAL-LDA far transcends the modeling time of MFFSTM-HET and SHHTFM. With the increase in the topic number K, the time curve of SAL-LDA displays linear growth, and the time curves of MFFSTM-HET and SHHTFM stay relatively smooth. In addition, even though SHHTFM requires more topics than MFFSTM-HET to reach the best performance, their time consumptions show little difference. This indicates that SHHTFM can give consideration to both time efficiency and satisfactory performance.

In order to study the sensitivity of MFFSTM-HET and SHHTFM in relation to the region size during SLIC superpixel sampling, the values of the patch size, the patch spacing, the regularizer, the topic number K, and the visual word number V were kept at the optimal parameter settings, respectively. The value of the region size was then varied over the range of [5, 10, 15, 20, 25] for the UC Merced data set and the Google data set of SIRI-WHU. The OAs obtained with different region sizes for the UC Merced data set and the Google data set of SIRI-WHU are reported in Fig. 18. As shown

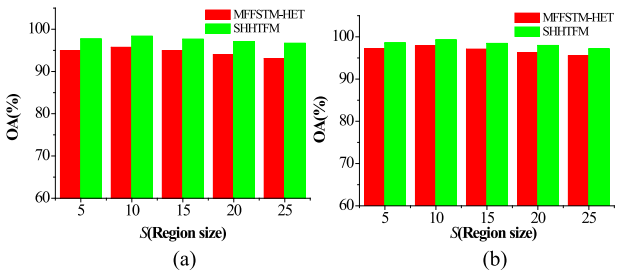


Fig. 18. Sensitivity analysis of MFFSTM-HET and SHHTFM in relation to the region size. (a) UC Merced data set. (b) Google data set of SIRI-WHU.

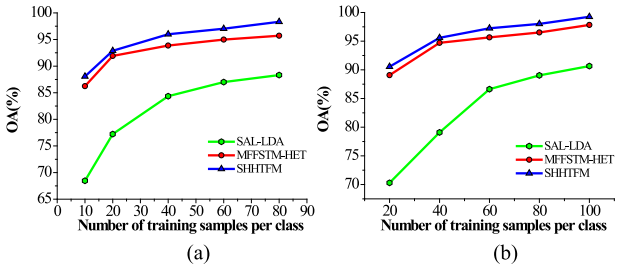


Fig. 19. Sensitivity analysis of MFFSTM-HET and SHHTFM in relation to the number of training samples. (a) UC Merced data set. (b) Google data set of SIRI-WHU.

in Fig. 18, the performance of the proposed SHHTFM is better than that of MFFSTM-HET. In addition, when the region size is changed from 5 to 25 with a step size of five, the OA of SHHTFM is relatively stable, which infers that the change of region size has little influence on the proposed framework.

In order to study the sensitivity of SAL-LDA, MFFSTM-HET, and SHHTFM in relation to the number of training samples, the values of the patch size, the patch spacing, the region size, the regularizer, the topic number K, and the visual word number V were kept at the optimal parameter settings, respectively. The number of training samples was then varied over the range of [80, 60, 40, 20, 10] for the UC Merced data set and [100, 80, 60, 40, 20] for the Google data set of SIRI-WHU. The curves of the OAs obtained by SAL-LDA, MFFSTM-HET, and SHHTFM for the UC Merced data set and the Google data set of SIRI-WHU are reported in Fig. 19. As shown in Fig. 19, the proposed SHHTFM 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 SAL-LDA and MFFSTM-HET.

E. Experiment 4: Semantic Annotation of the Wuhan IKONOS Image Data Set

The Wuhan IKONOS data set was acquired by the IKONOS sensor in June 2009, covering the Hanyang area of the city of Wuhan in China. All of the images in the Wuhan IKONOS data set were obtained by Gram–Schmidt pan-sharpening with ENVI 4.7 software. The spatial resolutions of the panchromatic images and the multispectral images are 1 and 4 m, respectively. The Wuhan IKONOS data set consists of eight land-use scenes, namely, dense residential, idle, industrial, medium residential, parking lot, commercial,

TABLE V
ANNOTATION ACCURACIES (%) WITH THE WUHAN IKONOS DATA SET FOR THE DIFFERENT SCENE CLASSIFICATION METHODS

Method	pLSA	LDA	Zhao *et al.* [25]	MFFSTM-HET	SHHTFM
Overall accuracy	77.34±6.23	84.38±7.24	88.96±3.95	95.83±1.74	**97.92±1.89**

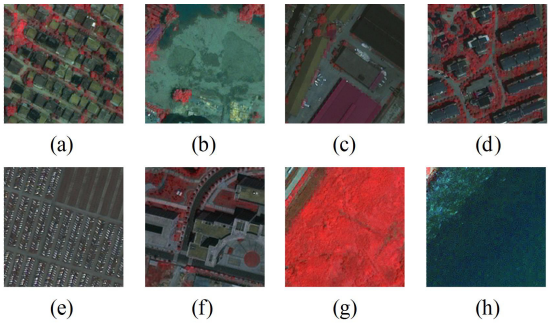


Fig. 20. Google data set of SIRI-WHU. (a) Agriculture. (b) Commercial. (c) Harbor. (d) Idle land. (e) Industrial. (f) Meadow. (g) Overpass. (h) Park. (i) Pond. (j) Residential. (k) River. (l) Water.

vegetation, and water, as shown in Fig. 20. Each class separately contains 30 labeled small images, which were cropped to 150 × 150 pixels, with a spatial resolution of 1 m. The size of the large image used for the annotation experiment was 6150 × 8250 pixels, as shown in Fig. 21.

In the annotation experiment, the large image was split into a set of small overlapping images of 150 × 150 pixels. The annotation experiment obtained good results when the overlap between two adjacent small images was set to 50 pixels. In this way, the spatial information lost during the large image sampling could be preserved. For the small images, the MSD, wavelet, and SIFT features extracted from the heterogeneous regions performed well when the patch size and overlap were set to 8 × 8 and 4 pixels, respectively. The region size and regularizer of the small images annotated based on the MSD feature extracted from the homogeneous regions were optimally set to 15 and 0.75, respectively. The final labels of the overlapping regions were decided according to the majority voting method.

To evaluate the performance of SHHTFM, the experimental results obtained with pLSA, LDA, MFFSTM-HET, and the experimental results published in [25] are shown for comparison. The different methods were evaluated using the evaluation method published in [15], where 80% of the labeled images were used as training images, and the remaining images were used for testing to evaluate the model. To annotate the large image, all the labeled images were used to train the model. The different methods were executed 20 times by random selection of training samples. To visually evaluate the large annotation maps, the annotation maps were overlaid on the original images with 50% transparency. From Table V, it can be seen that the accuracy of SHHTFM, 97.92% ± 1.89%, is the highest. This confirms the ability of SHHTFM to capture a discriminative and sparse semantic representation for HSR images.

The confusion matrices obtained by MFFSTM-HET and SHHTFM for the Wuhan IKONOS data set were selected from the results and are shown in Fig. 22. The misclassified image

Fig. 21. Large image in the Wuhan IKONOS data set for annotation.

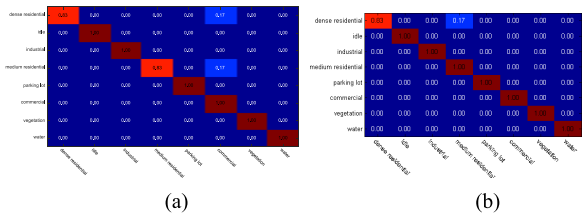


Fig. 22. Confusion matrices of (a) MFFSTM-HET and (b) SHHTFM with the Wuhan IKONOS data set, respectively.

of the medium residential scene in MFFSTM-HET is correctly classified by SHHTFM. In the confusion matrix of SHHTFM, all of the scenes can be recognized by SHHTFM, except for the dense residential scene. Only one image of the dense residential scene in SHHTFM is misclassified to the medium residential scene, which is mainly due to the composition of the similar LULC objects in these scenes, i.e., the trees, buildings, and roads.

A visual comparison of the performance of MFFSTM-HET and SHHTFM is displayed in Fig. 23(a) and (b), respectively, to further assess the semantic annotation results. On the whole, most of the scenes are annotated correctly. From the visual inspection, it can be seen that SHHTFM performs better than MFFSTM-HET. The annotation results obtained by SHHTFM are smoother, and some small confused regions in MFFSTM-HET are rectified with SHHTFM. Fig. 23(c) and (d) represents the same regions cut from Fig. 23(a) and (b), respectively, to allow a detailed evaluation. In Fig. 23(c), for MFFSTM-HET, there are many misclassifications with the other types of scene categories in the vegetation scenes, whereas SHHTFM [Fig. 23(d)] can fully recognize the vegetation scenes. However, some classes in the large image are undefined, i.e., the school, road, and gymnasium classes, which may lead to misclassification. In this way, the road may

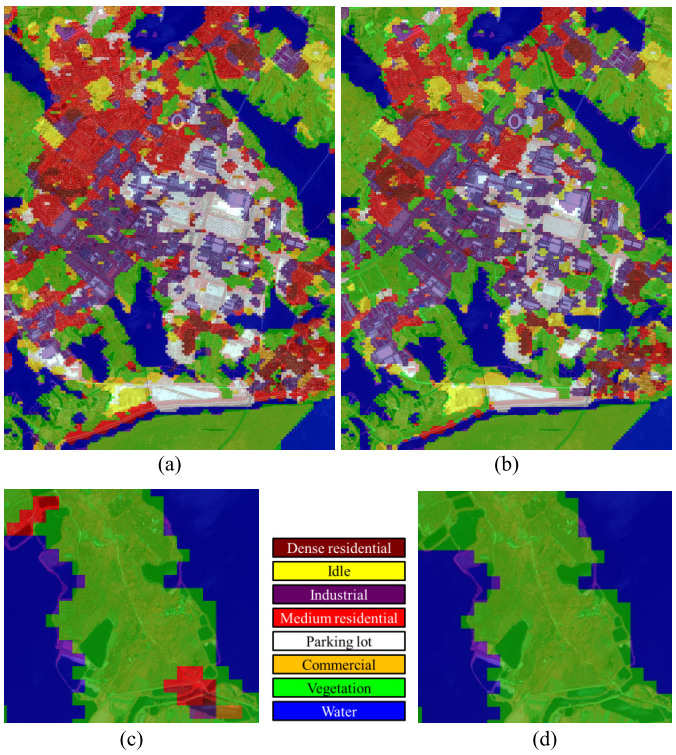


Fig. 23. Semantic annotation results obtained by the two methods. (a) MFFSTM-HET. (b) SHHTFM. (c) and (d) Detailed comparisons of the same regions from (a) and (b), respectively.

be classified to commercial, vegetation, or dense residential scenes.

Some misclassifications occur between the dense residential, medium residential, and commercial scenes. This is mainly due to two reasons. The first reason is that these scenes have similarity in both the MSD and texture characteristics. In addition, the large image sampling methods usually lead to ambiguous, br oken, and irregular boundaries between different scene categories, which may result in misclassification between the boundary regions of the dense residential, medium residential, and commercial scenes. To solve these problems, geographical data are employed in SHHTFM.

F. Experiment 4: Scene Understanding With the Combination of Multisource Geographic Data

In this experiment, the road network data were acquired from the Wuhan Land Resources and Planning Bureau for 2009, reflecting the road network information of the city of Wuhan in China. The scene annotation results were directly combined with the road network data, and the corresponding scene understanding results are shown in Fig. 24(a). As can be seen in Fig. 24(a), the derived scene blocks are mixed and coarse, and the results are unpractical for urban land-use planning. The multisource geographical data, i.e., the integration of the road network data and the boundaries of vegetation and water, were then overlaid on the scene annotation results. Based on the majority voting strategy, the final scene understanding results were obtained and are shown in Fig. 24(b). Compared with the scene annotation results and the results in Fig. 24(a), the results obtained with

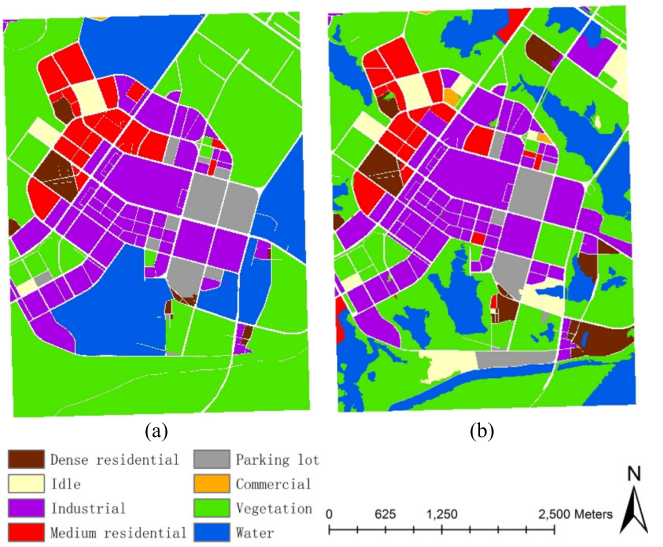


Fig. 24. Scene understanding results. (a) Scene understanding results combining only road network data. (b) Scene understanding results combining multisource geographical data.

the multisource geographical data are better and are more suited to practical use. Specifically, in Fig. 24(b), the road network and distribution of the vegetation and water areas are accurately defined, which could provide a reference for urban planning. The distributions and land occupation of the dense residential, medium residential, commercial, industrial, idle, and parking lot classes are easily identified, which could help the government to study the urbanization of Wuhan and rationally exploit the unutilized land. On the whole, it can be seen that the industrial and residential scenes occupy more area than the other scenes, whereas the commercial scene occupies the least area in the Hanyang region. This indicates that the Hanyang region is the industrial center of the city of Wuhan, and is undergoing the process of urban expansion.

V. Conclusion

In this paper, the efficient SHHTFM framework has been proposed for HSR remote sensing imagery scene classification. In SHHTFM, an effective region sampling strategy based on the union of uniform grid sampling and SLIC superpixel sampling is employed, and thus both the homogeneous information and heterogeneous information are simultaneously exploited for modeling the images. The sparse inference task of SHHTFM further improves the fusion of the HHTs, and thus a discriminative semantic description is obtained for distinguishing the scenes. The classification and annotation experiments undertaken in this paper showed that the proposed SHHTFM method performs better than the conventional PTM in discovering high-quality semantics from HSR images, with high time efficiency. In addition, the combination of multisource geographical data with the scene annotation results provides more reliable and applicable scene understanding results.

In our future research, we plan to use more social media data, e.g., point of interest data, volunteered geographic information data, and OpenStreetMap data, to further improve the scene classification results. On the other hand, to further analyze the scenes, multitemporal HSR images and images

with different resolutions from diverse remote sensing sensors will also be considered.

REFERENCES

[1] T. Blaschke and J. Strobl, "What's wrong with pixels? Some recent developments interfacing remote sensing and GIS," *Zeitschrift Geoinformationssysteme*, vol. 14, no. 6, pp. 12–17, 2001.

[2] J. C. Tilton, Y. Tarabalka, P. M. Montesano, and E. Gofman, "Best merge region-growing segmentation with integrated nonadjacent region object aggregation," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4454–4467, Nov. 2012.

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

[4] G. J. Hay *et al.*, "A comparison of three image-object methods for the multiscale analysis of landscape structure," *ISPRS J. Photogramm. Remote Sens.*, vol. 57, nos. 5–6, pp. 327–345, 2003.

[5] 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.

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

[7] S. Aksoy, K. Koperski, C. Tusk, G. Marchisio, and J. C. Tilton, "Learning Bayesian classifiers for scene classification with a visual grammar," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 581–589, Mar. 2005.

[8] D. Wang and X. Liu, "Scene analysis by integrating primitive segmentation and associative memory," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 32, no. 3, pp. 254–268, Jun. 2002.

[9] 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.

[10] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[11] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 606–617, Jun. 2011.

[12] 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.

[13] 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.

[14] 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, 2010, pp. 270–279.

[15] 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.

[16] R. Bahmanyar, S. Cui, and M. Datcu, "A comparative study of bag-of-words and bag-of-topics models of EO image patches," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 6, pp. 1357–1361, Jun. 2015.

[17] 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.

[18] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1, pp. 177–196, Jan. 2001.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[20] B. Zhao, Y. Zhong, and L. Zhang, "Scene classification via latent Dirichlet allocation using a hybrid generative/discriminative strategy for high spatial resolution remote sensing imagery," *Remote Sens. Lett.*, vol. 4, no. 12, pp. 1204–1213, 2013.

[21] K. Xu, W. Yang, G. Liu, and H. Sun, "Unsupervised satellite image classification using Markov field topic model," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 1, pp. 130–134, Jan. 2013.

[22] G. Cheng, L. Guo, T. Zhao, J. Han, H. Li, and J. Fang, "Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA," *Int. J. Remote Sens.*, vol. 34, no. 1, pp. 45–59, 2013.

[23] 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.

[24] 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.

[25] 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.

[26] M. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete latent variable decomposition of counts data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1313–1320.

[27] J. Zhu and E. P. Xing, "Sparse topical coding," in *Proc. Uncertainty Artif. Intell. (UAI)*, 2011, pp. 831–838.

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

[29] R. Fernandez-Beltran and F. Pla, "Incremental probabilistic Latent Semantic Analysis for video retrieval," *Image Vis. Comput.*, vol. 38, pp. 1–12, Jun. 2015.

[30] R. Kaviani, P. Ahmadi, and I. Gholampour, "Incorporating fully sparse topic models for abnormality detection in traffic videos," in *Proc. 4th Int. eConf. Comput. Knowl. Eng. (ICCKE)*, Mashhad, Iran, Oct. 2014, pp. 586–591.

[31] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[32] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, 2014, Art. no. 38.

[33] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.

[34] J. Cheng *et al.*, "Superpixel classification based optic disc and optic cup segmentation for glaucoma screening," *IEEE Trans. Med. Imag.*, vol. 32, no. 6, pp. 1019–1032, Jun. 2013.

[35] J. Hu, G.-S. Xia, F. Hu, and L. Zhang, "A comparative study of sampling analysis in the scene classification of optical high-spatial resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14988–15013, 2015.

[36] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[37] S. D. Newsam and C. Kamath, "Retrieval using texture features in high-resolution multispectral satellite imagery," *Proc. SPIE*, vol. 5433, pp. 21–32, Apr. 2004.

[38] K. Huang and S. Aviyente, "Wavelet feature selection for image classification," *IEEE Trans. Image Process.*, vol. 17, no. 9, pp. 1709–1720, Sep. 2008.

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

[40] 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.

[41] X. Zhang and S. Du, "A linear Dirichlet mixture model for decomposing scenes: Application to analyzing urban functional zonings," *Remote Sens. Environ.*, vol. 169, pp. 37–49, Nov. 2015.

[42] 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.

[43] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva. (2015). "Land use classification in remote sensing images by convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1508.00092

[44] 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.

[45] 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.

[46] 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.

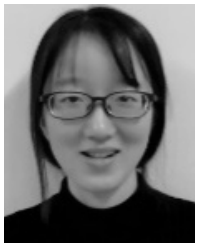
Qiqi Zhu received the B.S. degree from the China University of Mining and Technology, Xuzhou, China, in 2013. She is currently pursuing the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, 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 has been a Teacher with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, since 2007, and was promoted to Associate Professor and Full Professor in 2008 and 2010, respectively. He has authored more than 140 research papers, including more than 70 peer-reviewed articles in international journals such as *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, the National Excellent Doctoral Dissertation Award of China, and the New Century Excellent Talents in the University of China selected by the Ministry of Education of China. He was also a recipient of the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics. He was also a Referee of more than 30 international journals. He is an Associate Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, *International Journal of Remote Sensing*, and *Remote Sensing*.



Siqi Wu received the B.S. degree from Nanjing Normal University, Nanjing, China, in 2016. She is currently pursuing the M.S. 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 classification and scene understanding.

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 currently the Head with the Remote Sensing Division, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. He is also a Chang-Jiang Scholar Chair Professor appointed by the Ministry of Education of China. He is currently a Principal Scientist for the China state key basic research project (2011–2016) appointed by the Ministry of National Science and Technology of China to lead the remote sensing program in China. He has authored more than 450 research papers and five books, and holds 15 patents. 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 was a recipient of the 2010 Best Paper Boeing Award, the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing, and the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics. He is the Co-Chair of the series SPIE conferences on multispectral image processing and pattern recognition, conference on Asia remote sensing, and many other conferences. He is the Editor of several conference proceedings, issues, and geoinformatics symposiums. 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 Geo-Spatial Information Science*, and the *Journal of Remote Sensing*, and a Guest Editor for 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 in 1991, and the Chinese Academy of Engineering and Euro-Asia Academy of Sciences in 1995. He is currently the Academic Committee Chairman with 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 GIS, and their integration.

Dr. Li was the President of ISPRS Commissions III and VI, and the first President of the Asia GIS Association from 2002 to 2006.