IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 14, 2020, accepted November 27, 2020, date of publication December 1, 2020,
date of current version December 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3041645

Land Use Classification Using High-Resolution Remote Sensing Images Based on Structural Topic Model

HUA SHAO[1,2], YANG LI[2,3,4], YUAN DING[5], QIFENG ZHUANG[1], AND YICONG CHEN[1]

[1]College of Geomatics Science and Technology, Nanjing Tech University, Nanjing 211816, China
[2]Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing Normal University, Nanjing 210023, China
[3]Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Ministry of Education, Nanjing 210046, China
[4]School of Geographic Science, Nanjing Normal University, Nanjing 210046, China
[5]School of Earth Science and Engineering, Hohai University, Nanjing 211100, China

Corresponding author: Yang Li (li.yang@njnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 41501431 and Grant 41601449, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20190495.

ABSTRACT Remote sensing images are primary data sources for land use classification. High spatial resolution images enable more accurate analysis and identification of land cover types. However, a higher spatial resolution also brings new challenges to the existing classification methods. In the low-level feature spaces of remote sensing images, it is difficult to improve classification performance by modifying classifiers. Probabilistic topic models can connect low-level features and high-level semantics of remote sensing images. Latent Dirichlet allocation (LDA) models are representatives of probabilistic topic models. However, at present, probabilistic topic models are mainly adopted for scene classification and image retrieval in remote sensing image analysis only. In this study, multiscale segmentation was employed to construct bag-of-words (BoW) representations of high-resolution images. The segmented patches were then utilized as ''image documents.'' A structural topic model was used with an LDA model to import spatial information from the image documents at two levels: topical prevalence and topical content in the form of covariates. In this way, latent topic features in image documents can be more accurately deduced. The proposed method showed more satisfactory classification performance than standard LDA models and demonstrated a certain degree of robustness against the changes in the segmentation scale. Acknowledgement for the data support from ''Yangtze River Delta Science Data Center, National Earth System Science Data Center, National Science & Technology Infrastructure of China (http://nnu.geodata.cn:8008)''.

INDEX TERMS Bag-of-word model, latent topic, land cover, latent Dirichlet allocation, machine learning, probabilistic topic models.

I. INTRODUCTION

Automatic land use or land cover classification using remote sensing images has been receiving much attention from researchers and is expected to be studied extensively in the future. Due to the rapid development of electronics and information technology as well as sensor technology, high spatial resolution images are more available to researchers and practitioners. Hence, it becomes necessary to improve existing automatic classification methods based on remote sensing images. High spatial resolution images can provide abundant details of land surfaces; thus, the size, structure, and the spatial context of objects can be better characterized. Nevertheless, because of higher spatial resolution, there are greater spectral differences between the pixels of the same object type whereas the objects of different types may have similar spectral features. These lead to new challenges in the conventional classification methods [1].

In the past decades, owing to the rapid development of machine learning and artificial intelligence technology, various advanced supervised classification methods have been established such as support vector machines (SVM) [2] and random forests (RF) [3]. These methods generate relatively satisfactory results, and they are widely adopted. However, due to the extremely complex object compositions of high-resolution images, spectral responses of objects of the same type (e.g., construction land and roads) vary significantly. Hence, it is difficult to enhance the classification accuracy by

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

improving classifiers directly in the low-level feature spaces. Reference [4] pointed out that feature extraction from remote sensing images and training sample selection are more important than classifier improvement. Meanwhile, deep learning, which has been widely applied in remote sensing image analysis [5], can be viewed as a type of representation learning. It constructs models by simulating how the human brain observes and detects features in remote sensing images so that recognizable features at higher levels can be acquired. Regular rectangular input images are required for most deep learning models. Classification is performed by scenes or regular patches. As a result, few methods can achieve classification at the pixel or object levels [6]. However, for land use cover change analysis, classification for each pixel (or patch) is required.

Object types can be obtained from remote sensing images because land surfaces with different physical properties provide different spectral responses. In most feature selection or extraction methods, including deep learning methods, low-level features (e.g., spectra and textures) are modeled. However, in land cover classification, the surface of the objects of the same type may have rather complex physical properties. Objects consist of surface materials with different properties (these materials themselves may be unobservable latent factors) and the physical properties of these materials determine the low-level features. Thus, new methods are required to model the relationship between the latent factors determining the object types (middle-level features) and the observed low-level features and to classify these middle-level features to obtain higher classification accuracy.

Probabilistic topic models (PTMs) [7] use probability theory to acquire latent topic features. They have recently become popular for semantic analysis in natural language processing (NLP). They take text as a bag of words and view it as a collection of unordered words. They assume that each piece of text is generated by the mixed influence of different topics and subsequently use probabilistic graphical models to model the conditional probabilities between ''word,'' ''text,'' and ''topic.'' By estimating and inferencing model parameters, the topics that each piece of text and words within the text belong to are determined. The PTMs can convert text analysis from conventional word vector spaces (low-level feature spaces) to latent topic spaces (middle-level feature spaces). They can discover the synonyms and polysemous words commonly found in texts.

After many researchers have achieved successful results in NLP by extending PTMs such as probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA), the bag-of-words (BoW) model and PTMs have been adopted to solve problems related to image understanding. These include automatic image annotation, scene classification, and target recognition [8]–[11]. In this way, the gaps between low-level visual features and high-level semantics can be overcome. PTMs are successfully introduced in remote sensing image analysis and applied in scene annotation [12], target recognition [13], image fusion [14], and change detection [15]. Thus, useful high-level semantic information can be obtained by modeling latent topics in remote sensing images. However, previous studies investigated semantics at coarse scales, and thus there are no classifications at the pixel or target levels. Furthermore, standard LDA models assume that latent topics and text are independent from each other. Thus, much spatial information is lost during the modeling of high-resolution remote sensing images. When the existing LDA models are extended, much emphasis is placed on the processing of different text metadata, e.g., the author, year, publication time, and citation relationship of the text. These models cannot be used to import spatial information directly; thus, it is difficult to fuse the abundant spatial information into high-resolution images.

In this study, a structural topic model, which is an LDA-based model with the import of prior text knowledge, was adopted to propose an unsupervised latent topic feature extraction method applicable for high-resolution remote sensing images. This study examined how image documents, visual dictionaries, and visual words suitable for high-resolution remote sensing image analysis can be obtained. The construction of an appropriate BoW model was also investigated. It is likely that spatially close objects belong to similar types (latent topic compositions) and have similar low-level features (visual word distributions). Hence, in the spatial topic model, spatial information of image documents was imported as the prior knowledge for modelling the uncertainty in remote sensing images and to more accurately acquire the latent topic compositions of image documents.

II. RELATED WORK

A. LAND COVER CLASSIFICATION WITH REMOTE SENSING IMAGES

Land cover classification based on remote sensing images is performed to assign a pre-defined land cover type to each pixel in an image. In the past, classification was performed according to the spectrum of each pixel [16]. However, as the image resolution increases, intra-class spectral variations of pixels become more significant. This reduces the classification accuracy and leads to salt-and-pepper noise [17].

Object-based image classification (OBIC) has become a mainstream framework in high-resolution remote sensing, such as IKONOS, GeoEye, QuickBird and SPOT, and land-cover mapping over the last decade [18], [19]. However, this method has also been applied to medium-resolution Landsat images. The common supervised OBIC frameworks contain the following steps: image preprocessing, image segmentation, feature extraction, and supervised classification [20]. Apart from spectral features, texture, structural, and shape features can be included to enhance the distinguishability of classes in feature spaces. A previous study [19] provided an in-depth review on OBIC and examined the classification methods, sampling methods, feature selection methods, and accuracy. Another challenge to OBIC is the determination of appropriate segmentation scales. It is difficult to obtain a target that is

semantically complete. For most classification methods, only the superpixels obtained through segmentation are adopted as the basic classification units [19], [21].

In recent years, as deep learning has become more popular, many researchers have attempted to apply it to land cover classification. There are two main types of deep learning methods. In the first type, only remote sensing images of certain fixed sizes can be used as input. Deep neural network models are adopted as feature extractors. Classification is subsequently conducted using conventional classifiers (SVMs) [22]. Recently, deep learning methods of the second type have been more frequently used. For this type, end-to-end deep neural network models are constructed to more effectively perform classification pixel by pixel for remote sensing images without using additional classifiers [23]–[25]. However, because the resolution decreases as the receptive field increases, the scale has to be recovered with the help of max pooling indices [26] and dilated convolution [27]. Because the spaces in the models are quite large, both types of deep learning methods require numerous labelled data for training.

B. BAG OF VISUAL WORDS REPRESENTATION OF REMOTE SENSING IMAGE

The bag-of-words (BoW) methodology was first proposed for the text retrieval domain problem in text document analysis, and it was further adapted for computer vision applications [28]. BoW representations allow unstructured data to be expressed in word vector spaces. In addition, they are necessary steps for the application of PTMs to mine latent information. Images can be viewed as some collections of local features, which are not related to location information. These features are similar to words. Text documents can produce word-document co-occurrence matrices, which can similarly be acquired for images. Entries in images are often known as ''visual words'' whereas collections of these words are considered ''visual dictionaries.'' Documents correspond to regions in images. In word vector spaces, images can be subjected to image annotation, scene classification, and target recognition [28]. When an image is expressed using BoW representations, the following steps are required: (i) automatic detection of regions/points of interest, (ii) computation of local descriptors over those regions/points, (iii) quantization of the descriptors into words to form visual vocabulary, and (iv) determination of the occurrences in the image of each specific word in the vocabulary for constructing the BoW feature [29].

Remote sensing images can be expressed using BoW representations to obtain image documents and visual words. When the BoW models of scale invariant feature transform (SIFT) features were used, satisfactory results were obtained in remote sensing image retrieval [30]. In [31], K-means clustering performed using spectral and textural characteristics to generate a visual vocabulary list produced results superior to those obtained using SIFT features. In addition, SVM classifiers were applied to BoW representations of the aerial images. In [32], base images were selected for large collections of remote sensing images and similarity measurement methods in word vector spaces were improved to realize the retrieval of remote sensing images. Some researchers carried out superpixel segmentation of remote sensing images and used dense SIFT features in superpixels as low-level features to recognize clouds through SVM classification [33]. Uniform grid segmentation of the images was adopted in [34], in which a shape-based invariant texture index was designed to provide joint BoW representations for global texture features, local features, and dense SIFT features and to achieve scene classification for high-resolution remote sensing images. Reference [35] used spectral means, standard deviations, and SIFT features to construct a BoW model for scene classification and demonstrated that classification results obtained using both spectral and structural features show better performance than those obtained using only one type of feature. BoW models can also be constructed by extracting image features using trained convolutional neural networks to achieve scene classification [36]. In [37], image segmentation and uniform grid segmentation were adopted simultaneously to acquire low-level features to establish a multi-bag of visual words model to realize multi-label scene classification. In aforementioned studies, when BoW representations are formed, spectral, texture, and dense SIFT features facilitate the representations of images. Most studies focus on scene classification or image retrieval; land cover classification has not been sufficiently investigated. This study aims to examine land cover classification, where homogeneous image documents are required and features have to be extracted and classified in individual images.

C. REMOTE SENSING IMAGE FEATURE EXTRACTION BASED ON PROBABILISTIC TOPIC MODELS

Although BoW representations convert unstructured text or image information into word vector spaces, they are still representations of low-level features (spectral, texture, and dense SIFT). PTMs such as LDA can mine latent topic information, which is close to high-level semantic information, from low-level feature spaces [38]. When appropriate dictionary size and number of topics are selected, scene classification results for remote sensing images in latent topic spaces are superior to those obtained in word vector spaces [39]. In recent years, various LDA extension models have been developed to meet different application requirements. A comprehensive overview of these models can be found in [8].

After creating BoW representations of remote sensing images, many researchers applied PTMs to extract the latent features for scene classification, land cover classification, and image fusion. In [40], an LDA model was used for semantic annotation of remote sensing images whereas overlapping uniform grid segmentation was adopted to import the spatial relationships of image documents. In [41], uniform grids were adopted to segment documents into patches (image documents) whereas the latent topics with words were directly

projected into classes. Eventually, classification was carried out for each pixel (visual word) of Landsat images. Scene classification can also be performed using discriminant classifiers (e.g., SVM) after extracting latent topic features of remote sensing images [42]. In [43], the process of image mining was improved with a semantic annotation according to the spatial relationships between objects for scene-level analysis. In [44], a multi-scale latent Dirichlet allocation model was proposed to address the problem of semantic clustering of geo-objects in panchromatic images. In [45], over-segmented parts were used as image documents and gray-level values of pixels of panchromatic images as words. A semi-supervised latent Dirichlet allocation (ssLDA) was then employed for classification at the pixel level. Vocabulary can also be established separately by segmenting images with uniform grids and extracting spectral, texture, and SIFT features [46]. Subsequently, PLSA and LDA models can be employed separately to extract topics, which are fused to realize scene classification. In [47], uniform grid segmentation and an LDA model were used to analyze the scene-level land cover of time-series remote sensing images. In [48], uniform grid segmentation and super-pixel segmentation were used simultaneously to construct a BoW model, followed by an LDA model for scene classification. Some studies employed LDA models to detect changes in remote sensing images [15], to perform super-resolution reconstruction [49] andtoconduct image fusion [50] and target discrimination [13].

Most aforementioned studies adopted LDA for scene-level feature representation and mostly used standard LDA models. Pixel-level land use classification has not been possible. Since it is typically assumed that documents are independent from each other, it is difficult to represent spatial information in remote sensing images.

III. METHODOLOGY

To more effectively use PTMs to extract latent topic features from remote sensing images and thereby to realize pixel-level land use classification, this study proposed a framework consisting of three tasks: 1) BoW representations of high-resolution remote sensing images based on multiresolution segmentation; 2) latent topic feature extraction based on a structural topic model (STM); 3) supervised classification of image documents (Fig. 1).

A. BAG-OF-VISUAL-WORDS REPRESENTATION BASED ON MULTIRESOLUTION SEGMENTATION

Because there are no natural documents or words in remote sensing images, when a bag-of-visual-words (BOVW) model is adopted, an analogue of text-related terms in the image domain has to be built first. In scene classification, individual images are often treated as documents (basic units for classification) and images are subjected to uniform grid segmentation to obtain basic units for visual word extraction. The goal of this study is to perform pixel-wise land use classification. Hence, images were first segmented to obtain basic units for classification. Here, homogeneous basic units are required for classification, but complete object targets

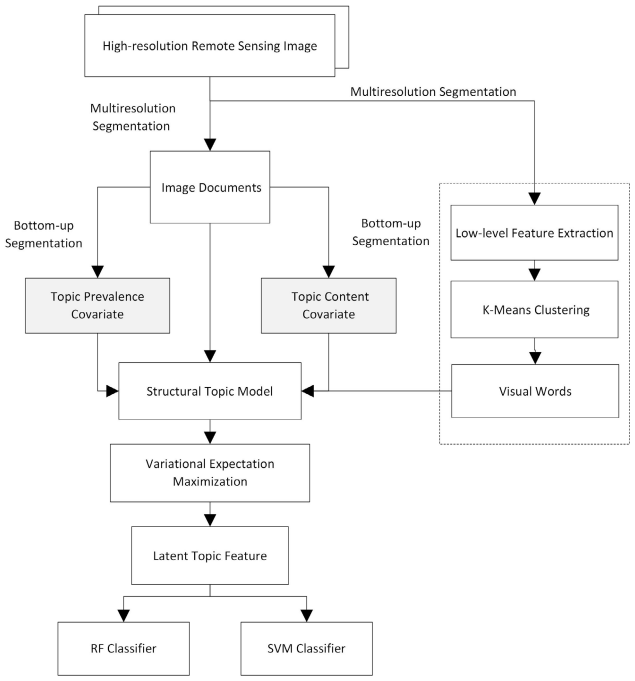

FIGURE 1. Flowchart of land use classification based on the structural topic model framework for high-resolution remote sensing imagery.

with semantic meanings, which are difficult to be segmented, are not required. This is different from what is required for conventional object-based classification methods. Thus, in this study, segmentation at smaller scales was performed to obtain homogeneous regions from images. This ensures that all pixels in a region belong to the same land cover type as much as possible. The segmented regions were treated as image documents whereas individual images were taken as corpuses.

To more comprehensively represent the information in image documents, this study adopted pixels as basic units and used spectral, texture, and dense SIFT [51] features. A gray-level co-occurrence matrix (GLCM) describes the texture using related spatial characteristics of the gray level. It is one of the most commonly used texture methods producing the most satisfactory results in remote sensing classification [52]. Hence, five irrelevant statistics in the GLCM were selected in this study to describe the texture features of targets: the angular second moment and entropy, the dissimilarity and contrast, and correlation reflect the texture homogeneity, the texture smoothness, and the inter-correlation between a gray-level pair, respectively. SIFT features are often used in the scene classification of remote sensing images. Compared to SIFT, dense SIFT only extracts structural features from images at one scale. Thus, the 128-dimensional feature descriptor at any position can be obtained. This can prevent the problem of having sparsely distributed feature points in images due to few feature points extracted by the original SIFT method. Because the spectral, texture, and dense SIFT features are all continuous numerical features, there are no naturally existing dictionaries and words, which resemble those in the text analysis. Hence, K-means, which is a simple and effective

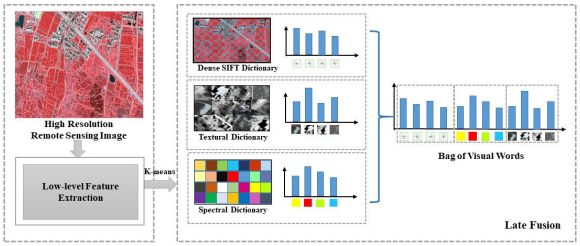

FIGURE 2. Process of constructing the bag-of-visual-words representation of high-resolution remote sensing images using the late fusion strategy.

clustering method, was adopted to quantize continuous features into discrete ones [34], [46]. During processing, there was only one input parameter (number of clusters). Unlike ordinary scene classification, all samples (instead of only the training samples) were subjected to clustering when forming visual words because image data had already been acquired.

The three low-level features that describe remote sensing images from different perspectives have different physical meanings, dimensions, and dimensionality. There are two commonly used feature fusion methods: early fusion and late fusion. For early fusion, low-level features of different types are stitched directly into a high-dimensional feature. Subsequently, visual words are acquired by clustering the resulting high-dimensional feature. This method is simple with a lower computational load. However, low-dimensional features can be submerged by high-dimensional ones. For late fusion, which is applied in the present study, visual words are first obtained separately within the spectral, texture, and dense SIFT feature spaces through clustering (Fig. 2). Therefore, an object can be depicted from different perspectives. This is similar to using synonyms in the text. Because late fusion can provide more perspectives to describe an image document and the over-segmented image documents are relatively homogeneous, this method can avoid the situation where there are insufficient types of words in image documents. In this method, the frequency of the visual words in each image document is obtained to present the BoW representation of the image. In addition, in the present framework, the classification accuracy is not sensitive to the segmentation scale, which avoids the difficult choice of segmentation scale.

B. LATENT TOPIC EXTRACTION BASED ON STRUCTURAL TOPIC MODEL

BOVW models allow objects to be represented using structured data after segmentation while most interior details of objects are preserved. Hence, supervised classification models can be adopted for classification. Nevertheless, images in the BOVW models are still represented by low-level features. Objects of identical land use types may have significantly different low-level features whereas objects of different land use types may have similar low-level features. These lead to considerable difficulties in classification. LDA models [38] were originally used to eliminate or reduce the gaps between high-level semantics and low-level words in the text represented by BoW models; thus, they assume that multiple topics

TABLE 1. Meanings of symbols in NATURAL LANGUAGE PROCESSING and in this paper.

Symbol	Definition in NLP	Definition in this paper
M	document set	image of the study area
D	number of documents in the document set	number of patches after image over-segmentation
K	number of topics in the document set	number of latent topics in the image
V	number of words in the dictionary	size of visual dictionaries generating from clustering
N_d	number of words in document d	number of visual words in image document d
$w_{d,n}$	the n th word in document d	the n th visual words in image document d
$z_{d,n}$	topic corresponding to the n th word in document d	latent topic corresponding to the n th word in image document d
α	K -dimensional vector, a hyperparameter representing the distribution of the topic mixing ratios of documents	K -dimensional vector, a hyperparameter representing the distribution of latent topic mixing ratios of image documents
β	scalar, a hyperparameter denoting the ratio of words of different topics	scalar, a hyperparameter denoting the ratio of visual words of different topics
θ_d	K -dimensional vector, topic mixing ratio of document d	K -dimensional vector, latent topic mixing ratio of image document d
φ_k	V -dimensional vector, proportion of words corresponding to topic k	V -dimensional vector, proportion of words corresponding to topic$k$
Θ	$M \times K$ matrix, topic mixing ratios of all M documents	$M \times K$ matrix, latent topic mixing ratios of all M documents
Φ	$K \times V$ matrix, proportions of words corresponding to all K topics	$K \times V$ matrix, proportions of words corresponding to all K latent topics

are mixed to present the document contents. Each topic is defined as a probability distribution of a word in the dictionary. Image documents of different land use types consist of several materials. Hence, image documents are assumed to be made up of multiple latent topics, each of which corresponds to the probability distribution of a low-level feature. In this way, different low-level features (spectral, texture, and dense SIFT) are generated. Using a PTM model, the uncertainty of remote sensing image can be modeled and observations can be made on documents and words to deduce the posterior probabilities of latent topics in each image document.

To more clearly illustrate the proposed method, Table 1 lists the definitions of parameters used in LDA models for text analysis in [38] and this study.

For LDA, a document set (image) is prepared as described below.

1) Select a K-dimensional Dirichlet random variable $\theta \sim \mathrm{Dir}(\alpha)$, where K is the number of the topics in the corpus.

2) For each of visual word w_n with $n \in$ Ê1, 2 . . . ĂĆÂN, perform the following actions:

a) Select a topic $z_n \sim Multinomial(\theta)$.

b) Select a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

The graphical illustration of the full data generation process for the LDA model is provided in Fig. 3. An LDA model

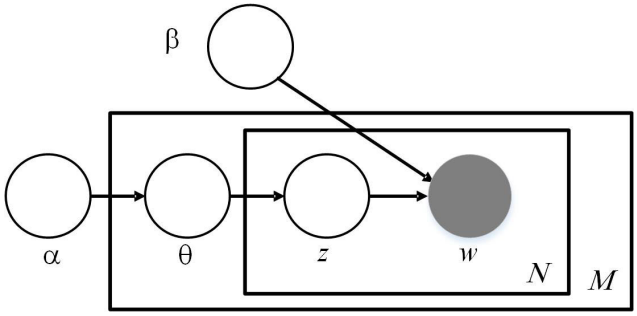

FIGURE 3. Graphical model representation of LDA [38].

is a three-layer Bayesian model with a complex structure; thus, an accurate calculation is extremely difficult. Owing to the advancement in computer software and hardware, parameter learning methods for the LDA model have been improved continuously in recent years. There are two types of parameter learning methods: Gibbs sampling and variational inference [7]. The former yields a higher accuracy whereas the latter has a higher speed.

The LDA model can be used to simulate the ''generation'' of remote sensing images. Each segment obtained by segmentation is composed of diverse materials with different physical properties (latent topic), and the observed low-level features (spectrum, texture, etc.) of the surface are precisely blended by the features (visual words) of the different materials. Furthermore, the segments classified into the same class may have different physical components, which is consistent with the fact that a document often contains several different topics and there are obvious differences in the distribution of words belonging to the same topic across different documents. In general, latent topic features are used to bridge the gap between low-level features and high-level semantic features.

The composition of topics of the text in an LDA model is determined by the Dirichlet distribution of hyperparameter α. The content of each piece of text is considered as the mixing of different topics. There are two assumptions concerning independence in an LDA model. First, topics are independent from each other (limited to a (K-1)-dimensional simplex). Second, documents are independent from each other. Unlike text, image documents are obtained via over-segmentation. Hence, it is expected that spatial information between patches can be included when image documents are modeled. For this purpose, an STM [53] was used to import document-level metadata for more accurately deducing the latent topics in image documents. The probabilistic graphical model of the STM is illustrated in Fig. 4.

Based on the LDA model, an STM introduces additional covariates from two aspects: topical prevalence and topical content. The former describes the topic compositions of the documents whereas the latter provides the usage rates of words of different topics. For topical prevalence, the Dirichlet distribution that controls the proportion of words in a document attributable to different topics is replaced with a logistic-normal distribution with a mean vector parametrized as a

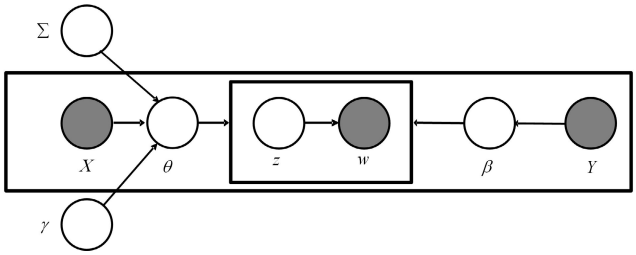

FIGURE 4. Graphical model representation of STM [53].

function of covariates. While all the documents in the LDA model use the same global parameter, the topical prevalence component can specify a document-level covariate in the model as prior knowledge affecting the latent topic compositions in the document. For topical content, the distribution is defined over the terms associated with different topics as an exponential family model, similar to multinomial logistic regression, parametrized as a function of the marginal frequency of occurrence deviations for each term, and of deviations from it that are specific to topics, covariates, and their interactions. The topical content establishes a base distribution for the words of each topic in the document set and parameterizes the deviation of the word distribution from the base distribution in the logarithmic space.

When the vocabulary size is V and the number of topics equals K, an STM can be created as illustrated below [54].

1) Draw the document-level attention to each topic from a logistic-normal generalized linear model based on a vector of document covariates X_d.

$$\vec{\theta}_d|X_{d\gamma}, \sum \sim \text{LogisticNormal}(\mu = X_{d\gamma}, \sum) \quad (1)$$

where X_d is a 1-by-p vector, γ is a p-by-$K-1$ matrix of coefficients and $\sum$ is $K-1$-by-$K-1$ covariance matrix.

2) Given a document-level content covariate yd, establish the document-specific distribution over words representing each topic (k) using the baseline word distribution (m), the topic specific deviation $\kappa_k^{(t)}$, the covariate group deviation $\kappa_{yd}^{(c)}$ and the interaction between the two $\kappa_{yd,k}^{(i)}$.

$$\beta_{d,k} \propto \exp(m + \kappa_k^{(t)} + \kappa_{yd}^{(c)} + \kappa_{yd,k}^{(i)}) \quad (2)$$

m, and each $\kappa_k^{(t)}$, $\kappa_{yd}^{(c)}$ and $\kappa_{yd,k}^{(i)}$ are V-length vectors containing one entry per word in the vocabulary. When no content covariate is present, β can be formed as $\beta_{d,k} \propto \exp(m+\kappa_k^{(t)})$ or simply point estimated (the latter approach is the default).

3) For each word in the document, ($n \in 1, \ldots, N_d$):

- Draw word's topic assignment based on the document-specific distribution over topics.

$$z_{d,n}|\vec{\theta}_d \sim \text{Multinomial}(\vec{\theta}_d) \quad (3)$$

- Based on the topic selected, draw an observed word from that topic.

$$w_{d,n}|z_{d,n}, \beta_{d,k=z_{d,n}} \sim \text{Multinomial}(\beta_{d,k=z_{d,n}}) \quad (4)$$

In the present study, topical prevalence in the STM was used to describe the phenomena that latent topics in image

documents may vary at different spatial locations. Because image documents are generated through over-segmentation, it is more likely for spatially close image documents to have similar latent topic compositions. Hence, they are attributed to the same land use type. This holds also true for the opposite. For image documents of some land use types (e.g., construction land and agricultural land), when they are spatially separated from each other, word compositions may differ considerably. On the contrary, when the image documents of the same type are spatially close to each other, their word compositions are relatively similar. The variations in word compositions with spatial location can be depicted by the topical content. Therefore, ''topical prevalence'' and ''topical content'' can describe the heterogeneity between image documents at the topic and word levels. However, the spatial distribution of the land uses in remote sensing images is not continuous. For example, patches adjacent to construction land may be agricultural land whereas those adjacent to water bodies may be forests. These patches are spatially adjacent to each other, but they have remarkably different latent topic compositions. Hence, it is not appropriate to directly use their spatial locations as the prior knowledge of their topic compositions. Thus, in this study, a multi-scale region merging algorithm in multi-scale segmentation was adopted to merge image documents. To minimize the heterogeneity of parent patches after merging, the following spectral and shape heterogeneity indicators were used [55].

1) The spectral heterogeneity indicator for an image patch is

$$h_{color} = \sum_c w_c \cdot \sigma_c \tag{5}$$

where w_c is the band weight, σ_c denotes the standard deviation of the spectral value for each band, and c is the number of bands;

2) The shape heterogeneity indicator for an image patch consists of smoothness and compactness.

$$h_{shape} = w_{smooth} \cdot h_{smooth} + w_{compact} \cdot h_{compact} \tag{6}$$

3) The overall heterogeneity indicator for an image patch is

$$h = w_{color} \cdot h_{color} + w_{shape} \cdot h_{shape}. \tag{7}$$

When multi-resolution merging is implemented, the influence of the spectra and shapes are controlled by w_{color} and w_{shape}, respectively. This paper focuses largely on the internal characteristics of image documents, and the shapes of the segments and the region is not critical; therefore, w_{color} is set to maximum and w_{shape} is set to minimum during multi-scale merging (performed in eCognition).

After merging the image documents with similar properties, the resulting parent patches were considered as the ''regions'' where the patch documents are located. In this study, the ''region'' attribute of a document was considered as a covariate. For topical prevalence, the covariate corresponding to document d is taken as a P-dimensional vector, where P denotes the number of regions. If document d belongs to the i th region, then the i th location of X_d is 1 whereas other locations are 0. For topical content, y_d is defined as the ID of the region where the patch is located.

In an STM, parameter inference can be performed using variational methods. Latent topics in image documents can thereby be obtained. However, this is different from text analysis. Latent topic compositions in image documents are not high-level features with semantic meanings, but middle-level features between low-level features and high-level semantics. Therefore, supervised classifiers are still required for the classification of image documents in topic spaces. In addition, it is worth pointing out that since the number of latent topics is usually much lower than the dictionary size, that is, the dimension of the latent topic feature space is much less than the BOW feature, which facilitates the avoidance of dimensional disaster and the achievement of superior classification performance.

IV. EXPERIMENT AND ANALYSIS

A. EXPERIMENTAL SETUP

To assess the performance of the proposed method for latent topic extraction from high-resolution remote sensing images, two datasets were adopted for land cover type classification experiments. For each dataset, the following features were used for classification: a. low-level features of segmented patches (spectral means and standard deviations and texture features) (Segment Low Level Feature, SLLF); b. word vector features based on the BoW model (Bag of Words, BoW); c. latent topic features extracted using a standard LDA model (Bag of Topics by LDA, BoT-LDA); d. latent topic features extracted using an STM by fusing spatial information (Bag of Topics by LDA, BoT-STM). For model hyperparameters (C and Gamma SVM; number of estimators and maximum depth for RF), the optimal values were determined by grid search. In addition, five-fold cross validation was adopted for model selection. For BoW and the proposed feature extraction strategies, there are two important parameters: visual dictionary size V and the number of latent topics K, will be tested with different values in the classification. Moreover, when fusing spatial information of image documents, the segmentation scale of multi-resolution merging adopted to obtain the regions where image documents are located may also affect the classification accuracy. Different values of the aforementioned parameters were adopted in the experiments and the results were compared. According to [19], SVMs and RFs demonstrate more satisfactory performance. Hence, these two supervised classifiers were used in this study, and 20% of the patches were arbitrarily selected as training samples whereas the remaining 80% were used as testing samples. The performance of the methods was evaluated using the overall accuracy measure and Kappa coefficients. Each classification method was tested 10 times. The average values were adopted as the results.

All experiments were conducted using a PC equipped with an i5-6500 CPU, a NVIDIA GT730 GPU, memory of 24 GB, and Windows 10 Pro. The models were implemented in

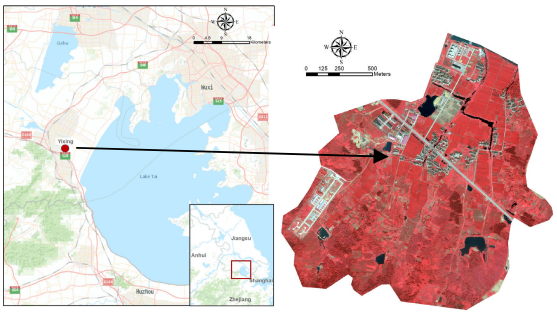

FIGURE 5. QuickBird Image dataset and its acquisition location.

Python 3.7.6 and R 3.6.3 environments. Image segmentation was done using eCognition 8.7 Developer 64.

B. QUICKBIRD IMAGE DATASET

A high-resolution remote sensing image of the Yicheng subdistrict of Yixing in Wuxi taken by QuickBird was used for analysis (Fig. 5). The image was taken on August 16, 2005 at a width of 3350 pixels and a height of 3405 pixels. There are four bands: blue (450–520 nm), green (520–600 nm), red (630–690 nm), and infrared (760–900 nm). The fusion product has a spatial resolution of 0.6 m. This study focused on the Meilin catchment area, which is approximately 9 km from Taihu Lake. The study area contains various land use types. At higher altitudes, the hillside is covered by extensive pine and bamboo forests. In other regions, agricultural land dominates, covering five land cover types: paddy fields, dry land, wood land, urban land, and water areas. Basic pretreatment was performed for the image. The true values of the reference land use types were adopted from the data reported by the Yangtze River Delta Science Data Center, National Earth System Science Data Center, National Science & Technology Infrastructure of China; the data were acquired by interpreting human-computer interactions, with an accuracy of >92%.

eCognition Developer 8.7[1] was adopted for multiresolution segmentation to obtain image documents. The goal of segmentation is to obtain homogeneous patches, but not complete objects. In other words, the image segmentation in the present paper is actually a form of over-segmentation. Hence, through manual visual interpretation, the segmentation scale was set to 60 such that most image documents are homogeneous in terms of their land use types. Meanwhile, the color parameter is set at the default value, 0.5. A total of 7073 image documents were acquired. Image documents with strong ambiguity or land use types, which are difficult to determine, were discarded. In total, 5001 image documents were adopted for analysis. Different visual dictionary sizes (60, 90, 120, 150, 180, 210, 240, 270, and 300) were used to construct BoW representations of remote sensing images. Because the late fusion strategy has been used, spectral, text, and SIFT dictionaries were independent from each other. The sizes of these three types of dictionaries were set the same. Another parameter significantly affecting feature extraction

[1]http://www.ecognition.com/

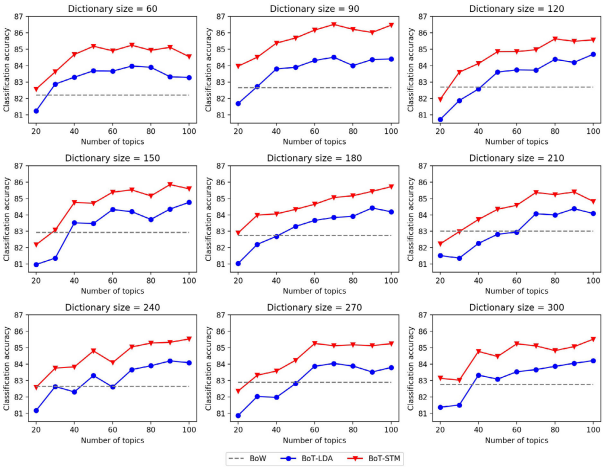

FIGURE 6. Overall classification accuracy using BoW, BoT-LDA, and BoT-STM models for various dictionary sizes using the QuickBird Image dataset.

TABLE 2. Kappa coefficients at visual dictionary sizes 180 and 300.

Dictionary Size	Number of Topics	BoW	LDA	STM
180	20	0.741	0.737	0.759
	30		0.758	0.775
	40		0.758	0.777
	50		0.764	0.782
	60		0.773	0.783
	70		0.776	0.789
	80		0.775	0.788
	90		0.782	0.796
	100		0.782	0.799
300	20	0.739	0.744	0.761
	30		0.744	0.762
	40		0.768	0.789
	50		0.767	0.779
	60		0.773	0.791
	70		0.772	0.788
	80		0.776	0.785
	90		0.778	0.790
	100		0.780	0.795

is the number of latent topics. It is often set to be substantially smaller than the dictionary size. The number of latent topics was set to be 20, 30, 40, 50, 60, 70, 80, 90, and 100 to compare the resulting classification accuracy. In eCognition, the minimum and maximum values that scales can be set at are 5 and 250, respectively. When constructing BoT-STM features, the largest segmentation scale (250) in eCognition was used to obtain regions where image documents are located. In total, 529 regions were formed. The classification results using SLLF, BoW, BoT-LDA, and BoT-STM features are presented in Fig. 6 and Table 2 shows the Kappa coefficient of classification results when the visual dictionary sizes are 180 and 300, which represent the ratio of error reduction between classification and completely random classification..

Because interior information of more image patches (documents) can be represented by using BoW models than by employing conventional object-based methods, the methods based on BoW and BoT (BoT-LDA and BoT-STM) features reveal significantly more accurate classification results than

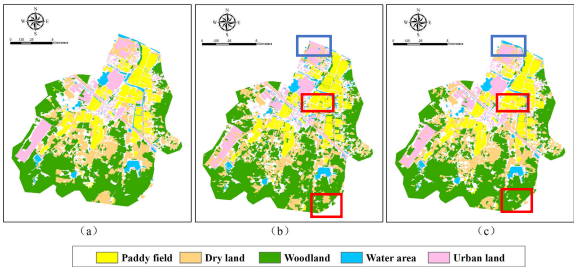

FIGURE 7. Classification results of QuickBird Image dataset. (a) Ground truth. (b) Results of BoT-LDA feature. (c) Results of BoT-STM feature. In these experiments, dictionary size = 300, topic number = 100, RF is applied to the topic latent feature.

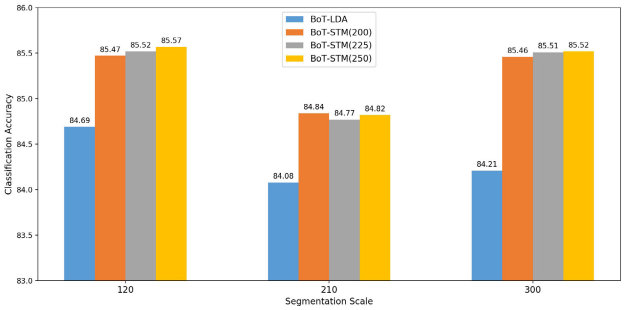

FIGURE 8. Overall classification accuracy using BoW, BoT-LDA, and BoT-STM models for dictionary size = 120, 210, 300 and region size = 200, 225, 250 using QuickBird Image dataset.

their counterparts based on SLLF features. Except for the conditions where there are few topics, for BoT feature-based classification, lower-dimensional features can be used to achieve higher accuracy. This is because document inference is performed using BoT features, and features inside image documents can be more efficiently represented using latent topics. In addition, BoT-STM-based method use the spatial regions where documents are located as prior knowledge and thus constraints for the classical LDA model. On the one hand, it is more likely that documents of the same region have the same topic composition. On the other hand, the same topics in the same region have the same probability distributions of visual words whereas the same topics in different regions may have different probability distributions of visual words. For all different visual dictionary sizes and topic numbers, BoT-STM-based classification results are more accurate than those based on BoT-LDA (Fig. 7). Furthermore, for all visual dictionary sizes, accuracy of both BoT-LDA- and BoT-STM-based classification is enhanced as the number of topics increases because more latent topics allow more comprehensive feature representations of the document content. However, when the number of latent topics increases to a certain number, the improvement of classification accuracy also decreases. Both classifiers (SVM and RF models) demonstrate satisfactory classification performances for high-dimensional feature data. The visual dictionary size insignificantly affects the accuracy of BoW-based classification. The accuracy is lower for the visual dictionary size of 60. However, for all other sizes, the classification accuracy based on BoW features is approximately 83%. BoT-LDA- and BoT-STM-based classification methods achieve the highest accuracy when the visual dictionary size is 90. This suggests that using latent topic features enables smaller visual dictionaries to be used to achieve higher accuracy.

Fig. 7(b) and 7(c) provide the classification results based on BoT-LDA and BoT-STM features, respectively, when the visual dictionary size is 300 and the number of topics is 100. The true values are shown in Fig. 7(a). Misclassification is noted between dry land and woodland, between water areas and urban land, and between dry land and urban land. According to Fig. 7, satisfactory classification results were obtained using BoT-LDA and BoT-STM features. The overall classification accuracy was higher for the result based on BoT-STM features. As revealed by the results in red rectangles, because regional prior knowledge was introduced during BoT-STM feature construction, salt-and-pepper noise was only noted for the classification result by the BoT-LDA features. This is because topical prevalence was introduced for BoT-STM features. Hence, documents of the same region may have similar topic compositions. The STM model also adopts topical content as a covariate to describe the situation where the same topic may have different visual word distributions in different documents. As illustrated by the blue rectangles in Fig. 7, the BoT-LDA-based method incorrectly recognizes water areas as urban land whereas the BoT-STM-based method correctly recognize those areas. This is because water areas and urban land have been divided into different regions, but owing to the topical content, water areas can still be correctly recognized although their visual word compositions in the document may differ from those in the training samples.

Another parameter that may influence the classification accuracy is the segmentation scale used to obtain the regions where image documents are located. To evaluate the effects of this segmentation scale on the classification results, experiments were conducted by setting the number of topics as 100; the segmentation scales as 250 (forming 529 regions), 225 (650 regions), and 200 (799 regions); and the visual dictionary sizes as 120, 210, and 300. In this way, the BoT-STM-based classification accuracy under three different region sizes can be compared (Fig. 8). The classification accuracy based on BoT-STM features is not sensitive to the region size. Hence, the method based on BoT-STM features is fairly robust against the variations in region size, and its robustness is higher than that based on BoT-LDA features.

C. GAOFEN IMAGE DATASET

Being a new large land use and land cover (LULC) classification dataset [23], the Gaofen Image Dataset (GID) contains 150 high-quality Gaofen-2 (GF-2) images of more than 60 cities in China. These images cover a geographic area of more than 5×10^4 km^2 and have high intra-class diversity as well as low inter-class separability. GF-2 is the second satellite of the High-resolution Earth Observation System (HEOS) and generates panchromatic images with a spatial resolution of 1 m and multi-spectral images with a spatial resolution

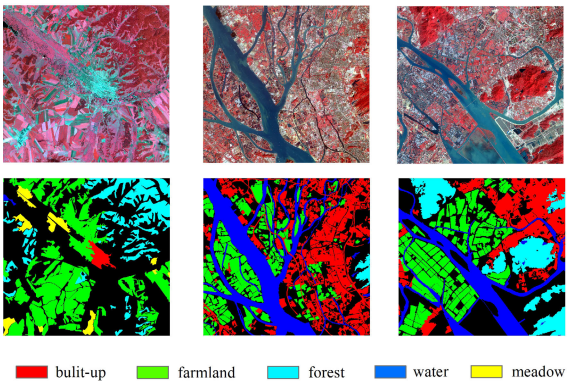

FIGURE 9. Three selected images from Gaofen Image dataset and their ground truth.

TABLE 3. The image documents construction information.

Image Name	Number of Objects from Segmentation	Number of Image Documents
GF2_PMS1__L1A0001680858-MSS1	30982	20079
GF2_PMS2__L1A0000607677-MSS2	41047	25859
GF2_PMS2__L1A0000607681-MSS2	49634	31822

of 4 m. The image size is 6908 × 7300 pixels. Multispectral images include images in the blue, green, red and infrared bands. Since its launch in 2014, the satellite has been adopted for important applications such as land surveys, environmental monitoring, crop yield estimation, and urban planning. The GID refers to Chinese Land Use Classification Criteria (GB/T21010-2017) to establish a hierarchical category system. In the large-scale classification set of the GID, five major categories are annotated: built-up, farmland, forest, meadow, and water. In this study, three GID images taken at different locations were selected to include all five land use types of interest. Feature extraction and classification experiments were then performed. The chosen images are GF2_PMS1__L1A0001680858-MSS1, GF2_PMS2__L1A0000607677-MSS2 and GF2_PMS2__L1A0000607681-MSS2 (Fig. 9). Table 3 shows the information of image document construction.

Fig. 10 illustrates the overall classification accuracy of the BoW, BoT-LDA, and BoT-STM models under different visual dictionary sizes. Because regions with class ambiguity have been removed from the ground truth data of the GID images, the resulting overall classification accuracy is higher although the spatial resolution of the images in this analysis is lower than that in section B. Except for the conditions with few topics (20 or 30), all the features represented by the latent topics exhibit more satisfactory classification performance than the BoW model. Because spatial information has been imported, classification based on BoT-STM features yields significantly more accurate results than its counterpart based

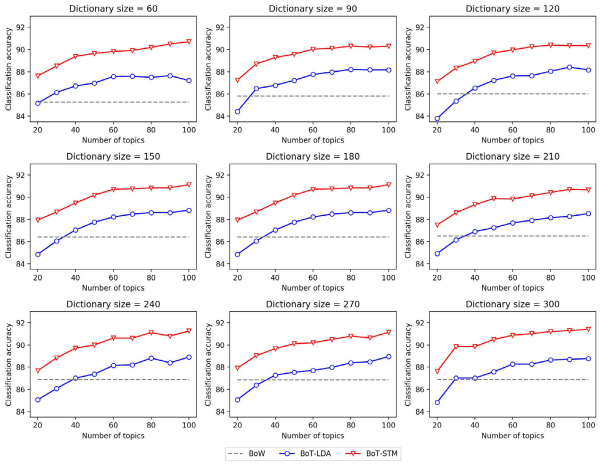

FIGURE 10. Overall classification accuracy using BoW, BoT-LDA, and BoT-STM models for various dictionary size using Gaofen image dataset.

on BoT-LDA features. In this analysis, another pattern is also noticeable: the classification accuracy using BoT-LDA and BoT-STM features is improved as the number of topics increases. This pattern is more distinct than that observed in section B because a larger number of image documents are involved in this analysis. The larger the corpus, the greater the number of topics required for semantic description.

In addition to the overall classification accuracy, which is adopted to evaluate the model performance, Blei [7] proposed that model perplexity can be used to assess the reliability of a PTM. Perplexity is defined as follows, where M_K denotes a topic model using K topics and The meanings of the other symbols are similar to those in Table 1:

$$Perplexity\,(M_K) = exp\left(\frac{-\sum_{d=1}^{D} logp\,(z_d)}{\sum_{d=1}^{D} N_d}\right) \tag{8}$$

Normally, when the perplexity obtained from (8) decreases as the number of topics K increases, the image document set fitted by the PTM becomes more similar to the true image document. Hence, the perplexity indicator describes the levels of similarity between the topic and visual word distributions of the images fitted by the model and the true distributions. It can also be used to evaluate the generalization ability of the model for the test data. Given the application scenarios of interest, the topic model was considered as a tool to mine information from the entire dataset. The training and testing sets were not separated from each other. Thus, perplexity was calculated directly from the original image document set. Fig. 11 shows the perplexity variations with the number of topics (with a constant visual dictionary size of 300) when BoT-LDA and BoT-STM models are employed separately. For both methods, perplexity decreases as the number of topics increases. For all different numbers of topics, the BoT-LDA-based method shows considerably lower perplexity than the BoT-LDA-based method. Moreover, when the number of topics increases, the perplexity of the BoT-STM-based method drops more rapidly because spatial information of image documents is included in the

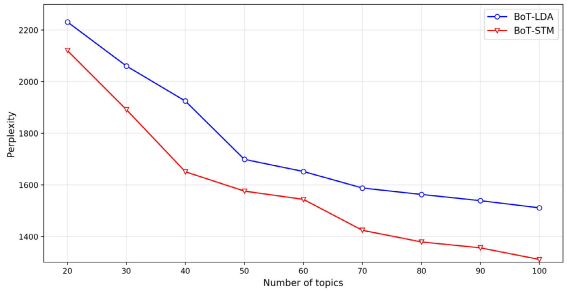

FIGURE 11. Perplexity value of LDA and STM model for various number of topics.

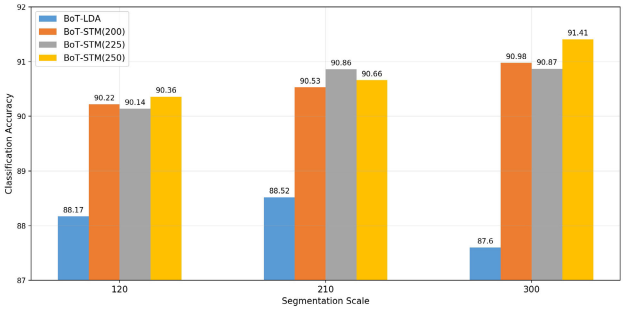

FIGURE 12. Overall classification accuracy using BoW, BoT-LDA, and BoT-STM models for dictionary size = 120, 210, 300 and region size = 200, 225, 250 using Gaofen Image dataset.

spatial LDA model. As a result, less topics are needed to represent useful information and to more accurately fit an approximate model to the true distributions.

To validate the effects on the classification accuracy by the segmentation scale used to obtain regions where image documents are located, three segmentation scales (200, 225, and 250) were adopted to acquire regions of image documents whereas three visual dictionary sizes (120, 210, and 300) and a constant number of topics (100) were used for classification. The overall classification accuracy of the methods based on BoT-LDA and BoT-STM features was compared (Fig. 12). The effect of the segmentation scale on the classification accuracy is not significant. The accuracy of the BoT-STM-based method is higher than that of the BoT-LDA-based method.

V. CONCLUSION

During land use classification based on remote sensing images, because objects of the same type may exhibit different spectral features and objects of different types may have similar spectral features in the low-level feature spaces, it is difficult to improve the classification performance by directly modifying the classifiers. Hence, a PTM-based classification framework was developed. First, BoW representations of remote sensing images were obtained. An LDA model was subsequently adopted for secondary mining of low-level features (spectral, texture, and dense SIFT features). Segmented image patches were used as documents while the low-level features inside were used as visual words. Given that an LDA model is a generative model, the document generation process includes the process in which low-level features are created according to the physical properties of land surfaces. Next, using an STM, a feature extraction method fusing spatial information of image documents was proposed. Two types of prior knowledge were introduced to the standard LDA model in the form of covariates to simulate the effects of the regions on the topic compositions of documents and word compositions of topics.

Two experiments were conducted on land use classification using high-resolution remote sensing images. The results demonstrate that using PTMs for secondary feature extraction from image documents yields more accurate classification results than adopting BoW models or using low-level features directly. Because spatial information of the documents is imported to the STM, latent topic information of image documents can be mined more accurately, effectively enhancing the classification accuracy. The proposed method is a supervised feature extraction method. Yet, it is also applicable for out-of-sample data to extract latent topic information of the new data.

In the future, further research is required to explore how to utilize label information to construct an end-to-end framework and a supervised or semi-supervised PTM so that discriminant models are no longer needed for classification. This can reduce the number of steps in the framework and enhance the classification performance. Secondly, how to determine the optimal number of latent topics and use more low-level features and systematically analyze the impacts of different low-level features in BoW and PTM on classification accuracy also needs to be resolved. In addition, the computational load can be lowered while increasing the computational efficiency. Finally, the semantic segmentation model based on deep learning can simultaneously segment and recognize remote sensing images. How to compare the mechanisms and experimental results between these methods and the methods in this paper should be explored in subsequent research. Furthermore, we will explore how to integrate PTM and deep learning models to achieve higher classification accuracy.

ACKNOWLEDGMENT

The authors would like to thank Editage (www.editage.cn) for English language editing.

REFERENCES

[1] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 65, no. 1, pp. 2–16, Jan. 2010, doi: 10.1016/j.isprsjprs.2009.06.004.

[2] U. Maulik and D. Chakraborty, "Remote sensing image classification: A survey of support-vector-machine-based advanced techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 33–52, Mar. 2017, doi: 10.1109/MGRS.2016.2641240.

[3] F. E. Fassnacht, H. Latifi, K. Stereńczak, A. Modzelewska, M. Lefsky, L. T. Waser, C. Straub, and A. Ghosh, "Review of studies on tree species classification from remotely sensed data," *Remote Sens. Environ.*, vol. 186, pp. 64–87, Dec. 2016, doi: 10.1016/j.rse.2016.08.013.

[4] A. E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine-learning classification in remote sensing: An applied review," *Int. J. Remote Sens.*, vol. 39, no. 9, pp. 2784–2817, May 2018, doi: 10.1080/01431161.2018.1433343.

[5] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019, doi: 10.1016/j.isprsjprs.2019.04.015.

[6] Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang, J. Gao, and L. Zhang, "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, May 2020, Art. no. 111716, doi: 10.1016/j.rse.2020.111716.

[7] D. Blei, "Probabilistic topic models," in *Proc. 17th ACM SIGKDD Int. Conf. Tuts.*, 2011, p. 1.

[8] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019, doi: 10.1007/s11042-018-6894-4.

[9] N. Rasiwasia and N. Vasconcelos, "Latent Dirichlet allocation models for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2665–2679, Nov. 2013, doi: 10.1109/TPAMI.2013.69.

[10] L. Laib, M. S. Allili, and S. Ait-Aoudia, "A probabilistic topic model for event-based image classification and multi-label annotation," *Signal Process., Image Commun.*, vol. 76, pp. 283–294, Aug. 2019, doi: 10.1016/j.image.2019.05.012.

[11] Z. Niu, G. Hua, L. Wang, and X. Gao, "Knowledge-based topic model for unsupervised object discovery and localization," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 50–63, Jan. 2018, doi: 10.1109/TIP.2017.2718667.

[12] M. Ushanandhini, S. Rajesh, and M. Rajakani, "Classification of high spatial resolution images using semantic allocation level-probabilistic topic model," in *Proc. Int. Conf. Recent Trends Inf. Technol. (ICRTIT)*, 2016, pp. 1–6.

[13] L. Du, Y. Wang, W. Xie, Z. Wang, and J. Chen, "A semisupervised infinite latent Dirichlet allocation model for target discrimination in SAR images with complex scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 666–679, Jan. 2020, doi: 10.1109/TGRS.2019.2939001.

[14] R. Fernandez-Beltran, J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "Multimodal probabilistic latent semantic analysis for sentinel-1 and sentinel-2 image fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 9, pp. 1347–1351, Sep. 2018, doi: 10.1109/LGRS.2018.2843886.

[15] B. Du, Y. Wang, C. Wu, and L. Zhang, "Unsupervised scene change detection via latent Dirichlet allocation and multivariate alteration detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4676–4689, Dec. 2018, doi: 10.1109/JSTARS.2018.2869549.

[16] P. Gong, D. Marceau and P. J. Howarth, "A comparison of spatial feature extraction algorithms for land-use classification with SPOT HRV data," *Remote Sens. Environ.*, vol. 40, no. 2, pp. 137–151, 1992, doi: 10.1016/0034-4257(92)90011-8.

[17] D. Phiri and J. Morgenroth, "Developments in Landsat land cover classification methods: A review," *Remote Sens.*, vol. 9, no. 9, p. 967, Sep. 2017, doi: 10.3390/rs9090967.

[18] Y. Qian, W. Zhou, J. Yan, W. Li, and L. Han, "Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery," *Remote Sens.*, vol. 7, no. 1, pp. 153–168, Dec. 2014, doi: 10.3390/rs70100153.

[19] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu, "A review of supervised object-based land-cover image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 277–293, Aug. 2017, doi: 10.1016/j.isprsjprs.2017.06.001.

[20] X. Zhang, Q. Wang, G. Chen, F. Dai, K. Zhu, Y. Gong, and Y. Xie, "An object-based supervised classification framework for very-high-resolution remote sensing images using convolutional neural networks," *Remote Sens. Lett.*, vol. 9, no. 4, pp. 373–382, Apr. 2018, doi: 10.1080/2150704X.2017.1422873.

[21] A. MacLachlan, G. Roberts, E. Biggs, and B. Boruff, "Subpixel land-cover classification for improved urban area estimates using Landsat," *Int. J. Remote Sens.*, vol. 38, no. 20, pp. 5763–5792, Oct. 2017, doi: 10.1080/01431161.2017.1346403.

[22] W. Zhao and S. Du, "Learning multiscale and deep representations for classifying remotely sensed imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 113, pp. 155–165, Mar. 2016, doi: 10.1016/j.isprsjprs.2016.01.004.

[23] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322, doi: 10.1016/j.rse.2019.111322.

[24] W. Zhao, S. Du, and W. J. Emery, "Object-based convolutional neural network for high-resolution imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 7, pp. 3386–3396, Jul. 2017, doi: 10.1109/JSTARS.2017.2680324.

[25] C. Zhang, I. Sargent, X. Pan, H. Li, A. Gardiner, J. Hare, and P. M. Atkinson, "Joint deep learning for land cover and land use classification," *Remote Sens. Environ.*, vol. 221, pp. 173–187, Feb. 2019, doi: 10.1016/j.rse.2018.11.014.

[26] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[28] C.-F. Tsai, "Bag-of-words representation in image annotation: A review," *ISRN Artif. Intell.*, vol. 2012, pp. 1–19, Nov. 2012, doi: 10.5402/2012/376804.

[29] A. Bosch, X. Muñoz, and R. Martí, "Which is the best way to organize/classify images by content?" *Image Vis. Comput.*, vol. 25, no. 6, pp. 778–791, Jun. 2007, doi: 10.1016/j.imavis.2006.07.015.

[30] S. Newsam and Y. Yang, "Comparing global and interest point descriptors for similarity retrieval in remote sensed imagery," in *Proc. 15th Annu. ACM Int. Symp. Adv. Geograph. Inf. Syst.*, 2007, p. 9.

[31] S. Xu, T. Fang, D. Li, and S. Wang, "Object classification of aerial images with bag-of-visual words," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 2, pp. 366–370, Apr. 2010, doi: 10.1109/LGRS.2009.2035644.

[32] J. Yang, J. Liu, and Q. Dai, "An improved bag-of-words framework for remote sensing image retrieval in large-scale image databases," *Int. J. Digit. Earth*, vol. 8, no. 4, pp. 273–292, Apr. 2015, doi: 10.1080/17538947.2014.882420.

[33] Y. Yuan and X. Hu, "Bag-of-words and object-based classification for cloud extraction from satellite imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 8, pp. 4197–4205, Aug. 2015, doi: 10.1109/JSTARS.2015.2431676.

[34] 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, doi: 10.1109/LGRS.2015.2513443.

[35] B. Zhao, Y. Zhong, and L. Zhang, "A spectral–structural bag-of-features scene classifier for very high spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 116, pp. 73–85, Jun. 2016, doi: 10.1016/j.isprsjprs.2016.03.004.

[36] 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, doi: 10.1109/LGRS.2017.2731997.

[37] X. Wang, X. Xiong, and C. Ning, "Multi-label remote sensing scene classification using multi-bag integration," *IEEE Access*, vol. 7, pp. 120399–120410, 2019, doi: 10.1109/ACCESS.2019.2937188.

[38] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003, doi: 10.1109/ICDM.2008.75.

[39] 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, doi: 10.1109/LGRS.2015.2402391.

[40] 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, doi: 10.1109/LGRS.2009.2023536.

[41] 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, doi: 10.1109/JSTARS.2010.2081349.

[42] 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, Dec. 2013, doi: 10.1080/2150704X.2013.858843.

[43] C. Vaduva, I. Gavat, and M. Datcu, "Latent Dirichlet allocation for spatial analysis of satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2770–2786, May 2013, doi: 10.1109/TGRS.2012.2219314.

[44] H. Tang, L. Shen, Y. Qi, Y. Chen, Y. Shu, J. Li, and D. A. Clausi, "A multiscale latent Dirichlet allocation model for object-oriented clustering of VHR panchromatic satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1680–1692, Mar. 2013, doi: 10.1109/TGRS.2012.2205579.
[45] L. Shen, H. Tang, Y. Chen, A. Gong, J. Li, and W. Yi, "A semisupervised latent Dirichlet allocation model for object-based classification of VHR panchromatic satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 4, pp. 863–867, Apr. 2014, doi: 10.1109/LGRS.2013.2280298.
[46] 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, doi: 10.1109/TGRS.2015.2435801.
[47] D. Espinoza-Molina, R. Bahmanyar, M. Datcu, R. Diaz-Delgado and J. Bustamante, "Land-cover evolution class analysis in image time series of Landsat and sentinel-2 based on latent Dirichlet allocation," in *Proc. 9th Int. Workshop Anal. Multitemporal Remote Sens. Images (MultiTemp)*, 2017, pp. 1–4.
[48] Q. Zhu, Y. Zhong, S. Wu, L. Zhang, and D. Li, "Scene classification based on the sparse Homogeneous–Heterogeneous topic feature model," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2689–2703, May 2018, doi: 10.1109/TGRS.2017.2781712.
[49] R. Fernandez-Beltran, P. Latorre-Carmona, and F. Pla, "Latent topic-based super-resolution for remote sensing," *Remote Sens. Lett.*, vol. 8, no. 6, pp. 498–507, Jun. 2017, doi: 10.1080/2150704X.2017.1287974.
[50] R. Fernandez-Beltran, J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "Remote sensing image fusion using hierarchical multimodal probabilistic latent semantic analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4982–4993, Dec. 2018, doi: 10.1109/JSTARS.2018.2881342.
[51] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011, doi: 10.1109/TPAMI.2010.147.
[52] F. R. de Siqueira, W. R. Schwartz, and H. Pedrini, "Multi-scale gray level co-occurrence matrices for texture description," *Neurocomputing*, vol. 120, pp. 336–345, Nov. 2013, doi: 10.1016/j.neucom.2012.09.042.
[53] M. E. Roberts, B. M. Stewart, and E. M. Airoldi, "A model of text for experimentation in the social sciences," *J. Amer. Stat. Assoc.*, vol. 111, no. 515, pp. 988–1003, Jul. 2016, doi: 10.1080/01621459.2016.1141684.
[54] M. E. Roberts, B. M. Stewart, and D. Tingley, "Stm: An r package for structural topic models," *J. Stat. Softw.*, vol. 91, no. 2, pp. 1–40, 2019.
[55] U. C. Benz, P. Hofmann, G. Willhauck, I. Lingenfelder, and M. Heynen, "Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information," *ISPRS J. Photogramm. Remote Sens.*, vol. 58, nos. 3–4, pp. 239–258, Jan. 2004, doi: 10.1016/j.isprsjprs.2003.10.002.

HUA SHAO was born in Yangzhou, Jiangsu, China, in 1981. He received the B.S. degree in information and computing science from Nanjing University in 2004, the M.S. degree in geographic information system from the Nanjing Institute of Geography and Limnology, Chinese Academy of Science, in 2007, and the Ph.D. degree in geographic information science from Nanjing Normal University in 2014.

From 2007 to 2011, he was a Software Engineer with Motorola and Gameloft. Since 2014, he has been a Lecturer with the College of Geomatics Science and Technology, Nanjing Tech University, China. His research interests include remote sensing image analysis and spatial data mining.

YANG LI was born in Suzhou, Jiangsu, China, in 1984. She received the B.S. and M.S. degrees in geographic information system from Nanjing Forest University in 2007 and 2010, respectively. She is currently pursuing the Ph.D. degree with Nanjing Normal University.

Since 2010, she has been a Lecturer with the Key Laboratory for Virtual Geographic Environment, School of Geography, Nanjing Normal University. Her research interests include remote sensing applications in land resources and scientific data sharing.

YUAN DING was born in Nanjing, China. He received the master's and Ph.D. degrees from Nanjing Normal University, Nanjing, in 2012 and 2018, respectively.

From 2012 to 2014, he had an industrial career in Nanjing Guotu Information Industry Company, Ltd., Nanjing. From 2018 to 2020, he was a Postdoctoral Fellow with Hohai University, Nanjing, where he has been a Lecturer Fellow since 2020. His research interests include 3D cadastral systems, spatial data modeling, and optimization algorithms for GIS.

QIFENG ZHUANG was born in Yixing, Jiangsu, China, in 1988. He received the B.S. degree from Nanjing Normal University in 2011 and the Ph.D. degree in geographic information science from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, in 2016.

Since 2017, he has been a Lecturer with the College of Geomatics Science and Technology, Nanjing Tech University, China. His research interest includes remote sensing of agriculture and water resource.

YICONG CHEN was born in Nanjing, Jiangsu, China, in 1998. He received the B.S. degree in surveying engineering from Nanjing Tech University in 2020, where he is currently pursuing the M.S. degree.

His research interests include remote sensing image processing and land cover classification.