IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 22, 2019, accepted August 15, 2019, date of publication August 23, 2019, date of current version September 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2937188

Multi-Label Remote Sensing Scene Classification Using Multi-Bag Integration

XIN WANG[1], **XINGNAN XIONG**[1], **AND CHEN NING**[2]
[1]College of Computer and Information, Hohai University, Nanjing 211100, China
[2]School of Physics and Technology, Nanjing Normal University, Nanjing 210023, China

Corresponding author: Xin Wang (wang_xin@hhu.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2019B15314, in part by the National Natural Science Foundation of China under Grant 61603124, in part by the Six Talents Peak Project of Jiangsu Province under Grant XYDXX-007, in part by the Jiangsu Government Study Scholarship, and in part by the 333 High-Level Talent Training Program of Jiangsu Province.

ABSTRACT For remote sensing (RS) scene classification, most of the existing techniques annotate a scene image with merely a single semantic label. However, with the recent advance of remote sensing technology, more abundant information is contained in high-resolution scenes, making a scene image having multiple semantic meanings (i.e., multilabels). Since multi-label RS scene image annotation is a domain full of challenges due to the ambiguities between complicated scene contents and labels, it motivates us to present a novel algorithm which is based on multi-bag integration. First, to describe the semantic content of RS scene image, we propose to partition a scene image into image patches, defined by a regular grid, and extract the heterogeneous features within each. Second, two kinds of image instance bag, namely segmented instance bag (SIB) and layered instance bag (LIB), are designed to represent the scene image. Third, a Mahalanobis distance-based K-Medoids approach is applied to cluster SIB and LIB, respectively, to convert the multi-instance into single-instance, and then the obtained two single-instances are concatenated to generate more powerful scene-aware representation. At last, a multi-class classification technique is used to make predictions on the class labels. Experiments are performed on real remote sensing images and the results show that the proposed method is valid and can achieve superior performance to a number of state-of-the-art approaches.

INDEX TERMS Remote sensing, multi-label, scene classification, multi-bag.

I. INTRODUCTION

With the continuous advances in sensor technology, a great number of high spatial resolution (HSR) remote sensing (RS) images can now be available. These HSR images contain abundant spatial and structural information, making the perspective of traditional remote sensing image understanding change. In the traditional RS image understanding task, a scene image is classified into a certain category and assigned with only a unique semantic label. However, single labels may be insufficient for annotating more complex scenes with multiple semantic meanings or with ambiguous semantic contents. Hence, multi-label classification framework that can assign multiple labels to complex scenes becomes crucial for effective and comprehensive HSR remote sensing image annotation [1]–[3].

Although multi-label classification has been studied in the computer vision community in recent years, most works focus on natural scene images captured by ground-level sensors rather than remote sensing scene images captured by airborne or space-borne ones. For instance, in [4], a binary relevance (BR) strategy was proposed for multi-label natural scene classification. In [5], a multi-label text classification method via calibrated label ranking (CLR) was introduced. In [6], a classifier chains method based on ensembles of classifier chains (ECC) was addressed for classifying multi-label datasets from a variety of domains. In [7], Bayesian chain classifiers (BCC) were combined for multidimensional classification. In [8], a method called random k-labelsets (RAKEL) was proposed for multi-label classification. In [9], a multi-label learning algorithm based on label specific features (LIFT) was introduced. In [10], a distributed nearest neighbor classification method for large scale multi-label data on spark was proposed. In [11], a highly efficient parallel

The associate editor coordinating the review of this article and approving it for publication was Byung-Gyu Kim.

approach was presented for computing the multi-label k-Nearest Neighbor classifier on GPUs.

In [12], [13], a multi-instance multi-label learning (MIML) framework was proposed for multi-label classification. In MIML, the training samples are represented as bags [14], [15], each of which is described by multiple feature vectors named instances. A bag is labeled positively if at least one of its instances is positive, while it is defined negatively if all instances in it are negative. Compared to traditional multi-label learning frameworks, MIML is more convenient and reasonable for handling with multi-label problems, for it cautiously explores the inner causality between the training sample and its labels [12].

Although a number of MIML algorithms have been proposed for a variety of image classification problems, few promising methods have been developed for HSR remote sensing scene classification. In addition, most algorithms focus on the design of multi-label classifiers or the modeling of relationships between instances and labels, but make less research on how to construct full description of the semantics for the original training data set, how to build effective bags, as well as how to preserve the intrinsic information among instances. Nevertheless, the impact of these factors on multi-label classification performance is actually very large, especially for the multi-label RS scene classification tasks.

Thus, in this paper, we propose a novel MIML framework for multi-label RS scene classification, which takes into account all above factors. We demonstrate that the proposed framework achieves superior classification performance for RS scenes.

The main contributions of this paper are as follows.

- Given a RS scene image, proposing to partition it into a set of image patches, defined by a regular grid, and extracting the heterogeneous features within each patch.
- Designing two kinds of image instance bag, namely segmented instance bag (referred to as SIB) and layered instance bag (referred to as LIB), to represent the scene image. Specifically, to build the SIB, the scene image is first segmented into multiregions and then the SIB is defined as a bag of instances corresponding to the segmented regions in the image. To build the LIB, the scene image is partitioned into subregions via the idea of spatial pyramid, and then the LIB is composed of a set of instances corresponding to the layered subregions.
- Introducing a Mahalanobis distance-based K-Medoids approach to cluster the SIB and LIB, respectively, so as to convert the multi-instance into single-instance, and then the obtained two single-instances are concatenated in order to generate more powerful scene-aware representation.
- Adopting an efficient classification method for the automatic label prediction.
- Evalutating the superiority of the proposed framework on real remote sensing image data set.

The rest of this paper is organized as follows: Section II provides a necessary background in the area of multi-instance multi-label learning. Section III introduces the proposed multi-label remote sensing scene classification framework in detail. Section IV presents the experimental results and analysis. Finally, the conclusions are drawn in Section V.

II. BACKGROUND

In this section, we will give the necessary background information on multi-instance multi-label learning.

Before introducing the multi-instance multi-label learning, two learning frameworks, namely multi-label learning (MLL) and multi-instance learning (MIL), which are related to MIML are reviewed briefly.

A. MULTI-LABEL LEARNING

Multi-label learning, also referred to as single-instance multi-label learning, studies the problems in which an object is represented by a single instance while associated with a number of labels [10], [11].

Suppose $\mathcal{X}$ is the instance space and $\mathcal{Y}$ is the set of class labels. Given a dataset $\{(x_i, Y_i)\,|i = 1, 2, \ldots, N\}$, the goal of MLL is to learn a function $f : \mathcal{X} \rightarrow 2^{\mathcal{Y}}$ which maps an instance $x_i \in \mathcal{X}$ produced by an input image to a set of labels $Y_i = \left\{y_i^1, y_i^2, \ldots, y_i^{l_i}\right\} \subseteq \mathcal{Y}$ indicating which classes the image belongs to, where l_i represents the number of labels in Y_i.

B. MULTI-INSTANCE LEARNING

Multi-instance learning, also referred to as multi-instance single-label learning, studies the problems in which an object is described by a bag of instances while associated with a single label [14], [15].

Different from MLL, given a dataset $\{(X_i, y_i)|\ i = 1, 2, \ldots, N\}$, the goal of MIL is to learn a function $f : 2^{\mathcal{X}} \rightarrow \mathcal{Y}$ which maps a bag of instances $X_i = \{x_i^1, x_i^2, \ldots, x_i^{m_i}\} \subseteq \mathcal{X}$ produced by an input image to a label $y_i \in \mathcal{Y}$, where m_i represents the number of instances in X_i.

C. MULTI-INSTANCE MULTI-LABEL LEARNING

MIML, closely related to MLL and MIL, is actually a more general framework. It considers the ambiguities in both the instance and the label spaces, and thus is more natural and convenient to handle with tasks involving such objects [13].

Let $\{(X_i, Y_i)\,|i = 1, 2, \ldots, N\}$ be a given a dataset. $X_i = \left\{x_i^1, x_i^2, \ldots, x_i^{m_i}\right\} \subseteq \mathcal{X}$ is a bag of instances, and $Y_i = \left\{y_i^1, y_i^2, \ldots, y_i^{l_i}\right\} \subseteq \mathcal{Y}$ is a set of labels, where m_i and l_i represent the numbers of instances in X_i and labels in Y_i, respectively. The goal of MIML is to form a hypothesis $f : 2^{\mathcal{X}} \rightarrow 2^{\mathcal{Y}}$ which maps a bag of instances X_i produced by an input image to a set of labels Y_i.

Based on the MIML framework, two MIML algorithms, namely MIMLBOOST and MIMLSVM were designed for scene classification by using the degeneration strategy [13]. The former utilizes multi-instance learning as a bridge to transform the MIML data set into a multi-instance data set and then handles the problems by Boosting, while the

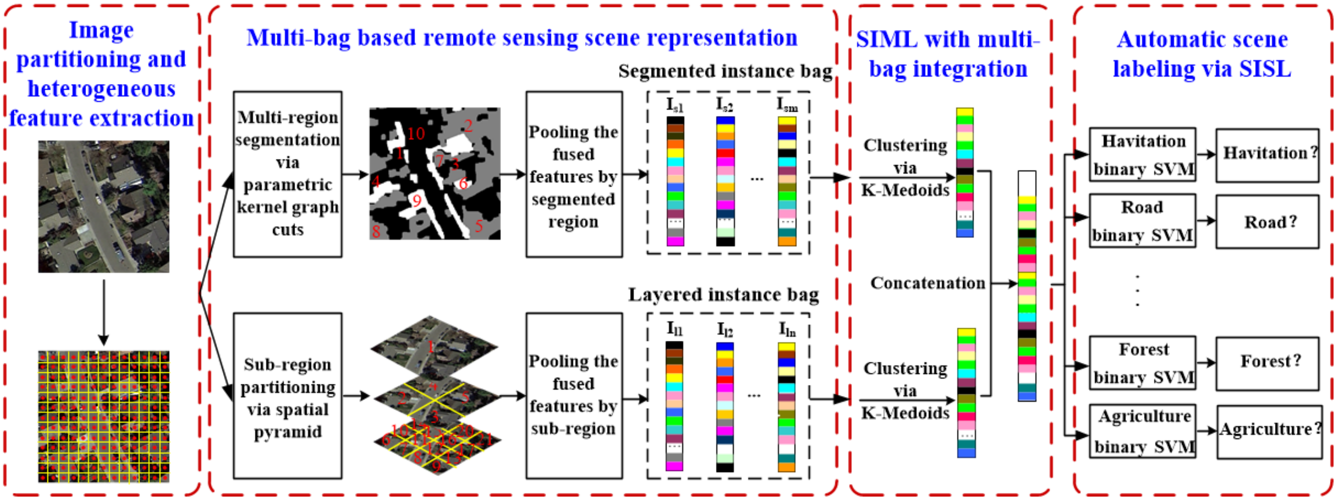


Image partitioning and heterogeneous feature extraction
Multi-bag based remote sensing scene representation
Multi-region segmentation via parametric kernel graph cuts
Pooling the fused features by segmented region
Segmented instance bag
I_s1
I_s2
I_sm
...
Sub-region partitioning via spatial pyramid
Pooling the fused features by sub-region
Layered instance bag
I_l1
I_l2
I_ln
SIML with multi-bag integration
Clustering via K-Medoids
Concatenation
Clustering via K-Medoids
Automatic scene labeling via SISL
Havitation binary SVM
Havitation?
Road binary SVM
Road?
Forest binary SVM
Forest?
Agriculture binary SVM
Agriculture?

FIGURE 1. Overall architecture of the proposed method.

latter uses multi-label learning as a bridge to convert the MIML example into a number of multi-label learning tasks and then solves them using support vector machine (SVM). In [16], a Bayesian MIML algorithm based on Gaussian process prior was presented. It can well exploit the connections between instances and labels as well as the correlations among labels for MIML learning. In [17], two MIML learning methods, MIMLSVM+ and E-MIMLSVM+ were addressed to solve gene expression pattern annotation problem. In MIMLSVM+, a degeneration scheme is designed to decompose the multi-label learning into a series of binary learning tasks. And E-MIMLSVM+ is the extension of MIMLSVM+. In [2], a hierarchical MIML algorithm using Gaussian process was presented for image semantic annotation. In [18], a MIML distance metric learning method was proposed for genome-wide protein function prediction. In [19], a fast MIML approach was studied for complicated labels learning.

III. PROPOSED METHOD

In this section, a novel multi-label remote sensing scene classification method based on multi-bag integration is presented in detail. Fig. 1 illustrates an overview of the proposed framework, which consists of four main modules, such as image partitioning and heterogeneous feature extraction, multi-bag based scene representation, SIML with multi-bag integration, and automatic scene labeling via SISL. Details of each module are given as below.

A. IMAGE PARTITIONING AND HETEROGENEOUS FEATURE EXTRACTION

In the previous MIML learning literatures, researchers usually focus on how to design multi-label classifiers or put emphasis on how to model the relationships between instances and labels, but make less research on the construct of full description of the semantics for the original training data, which is actually very important for multi-label RS scene classification. Therefore, in this paper, the first module of our proposed framework is aiming at extracting the meaningful features from each sample in the training set.

Suppose $Train = \{train_1, train_2, \ldots, train_N\}$ is the training set containing N remote sensing scene image samples. $Label = \{label_1, label_2, \ldots, label_N\}$ is the corresponding label set. Let $\mathcal{L} = \{1, 2, \ldots, |\mathcal{L}|\}$ be the set of all possible class labels associated with the images in the training set. Each training sample $train_i \in Train$ is associated with a vector $label_i = \left[l_i^1, l_i^2, \ldots, l_i^{|\mathcal{L}|}\right]$ of labels, where $l_i^c = 1$ if $train_i$ contains the class label $c \in \mathcal{L}$ and $l_i^c = 0$ otherwise.

Considering that, for RS scene images, both rough and smooth areas could contain important image information, and what's more, the semantic labels are related to regions rather than the whole images, we present to extract a number of dense regular patches from images, directly select the patch centers as a key feature points, use visual descriptors to describe the key feature points, and the results are used for the representations for the patches. The specific steps are as below.

First, each sample image $train_i$ is partitioned into n_i image patches $\left\{P_i^1, P_i^2, \ldots, P_i^{n_i}\right\}$ based on a regular grid, where P_i^k is the k-th patch of $train_i$.

Then, for each image patch P_i^k, as mentioned above, directly select the patch center as a key feature point. For this key point, extract its dense speeded up robust feature (referred to as D-SURF) D_i^k, Mean-Std feature (referred to as Mean-Std) M_i^k, as well as multiscale completed local binary pattern feature (referred to as MS-CLBP) C_i^k by using the heterogeneous feature extraction approaches presented in our previous work [20]. The effectiveness of these various features for remote sensing scene representation has been verified in [20]. The reader can be referred to the literature for detailed information about these heterogeneous features.

Third, in order to bridge the semantic gap between the above low-level features and high-level semantic meanings for RS scenes, a simple yet effective technique named locality-constrained linear coding (LLC) [21] is applied to code the heterogeneous features. Thus, the enhanced feature vectors α_i^k, β_i^k, υ_i^k for P_i^k are gotten, where α_i^k is the encoded result of D_i^k (referred to as D-SURF-LLC), β_i^k is the encoded result of M_i^k(referred to as Mean-Std-LLC), and υ_i^k is the encoded result of C_i^k (referred to as MS-CLBP-LLC).

At last, α_i^k, β_i^k, υ_i^k are concatenated to obtain the final fused feature vector for P_i^k:

$$f_i^k = \begin{bmatrix} \alpha_i^k \\ \beta_i^k \\ \upsilon_i^k \end{bmatrix} \tag{1}$$

Note that the main reason for using the simple stacking for the three types of features fusion is that these features are heterogeneous, so they might be uncorrelated. Therefore, the straightforward stacking of them not only would be an efficient way (and thus low computational complexity), but the fused feature would also be expected to be irredundant.

Fig. 2 shows an example of how image partitioning is done for a training image. And the red points are selected as the key feature points, for which the heterogeneous features are extracted.

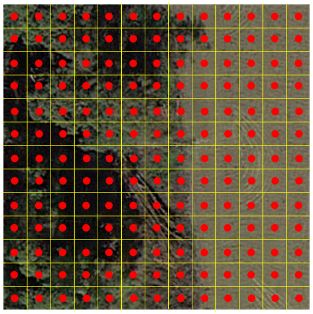


FIGURE 2. Example of image partitioning and heterogeneous feature extraction.

B. MULTI-BAG BASED SCENE REPRESENTATION

In the MIML framework, an image is usually represented as a bag consisting of multiple instances, and each instance corresponds to a segmented region in the image. The bag label is determined by the number of positive instances. For a specific label, an image is labeled positively if at least one segmented region has the corresponding semantic meaning and negatively otherwise. According such bag-instance setting, an image can be assigned with multiple semantic meanings, i.e., labels. It is noted that the construction of instances and bags is very important in MIML. To generate the instances, most MIML algorithms adopt various image segmentation techniques to obtain the regions at first, and after segmentation, features are extracted from each region to create a feature vector as an instance [22]–[25].

Although these algorithms are effective for a variety of situations, they have two defects:

(1) The small regions of the image are obtained using a certain image segmentation algorithm. Can it be ensured that each small region corresponds to an independent object with a specific semantic meaning in the image? Will there be a situation where several objects with different semantic meanings are included in a segmented region? Or will there be a situation where an object with a certain semantic meaning is divided into several segmented regions? Therefore, the accuracy of the segmentation algorithm affects the performance of the MIML learning significantly.

(2) After dividing the entire image by image segmentation, the relationship between different objects may be cut off. However, a large scene is often composed of multiple objects. The individual description of an object will lose the associated information between the objects.

Based on the above analysis, this paper proposes a multi-bag based scene representation algorithm, in which two different kinds of bags, namely segmented instance bag and layered instance bag, are constructed.

1) SEGMENTED INSTANCE BAG

Given a sample image $train_i$, an efficient segmentation algorithm [26] is first chosen to segment it into r_i semantically meaningful regions $\left\{S_i^1, S_i^2, \ldots, S_i^{r_i}\right\}$. For instance, as shown in Fig. 3, the scene image is segmented into $r_i = 4$ regions.

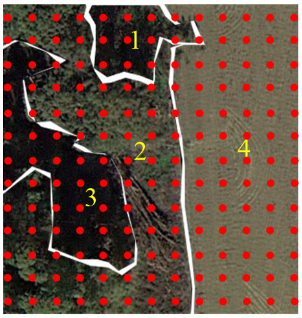


FIGURE 3. Example of the segmented instance bag.

Second, according to the locations of the key feature points which are obtained in Section III.A, each segmented region S_i^k may contain a number of key points. Suppose $\left\{f_i^{m_1}, f_i^{m_2}, \ldots, f_i^{m_t}\right\}$ are m_t fused feature vectors of these key points, we combine these vectors to form a matrix $B_i^k = \left[f_i^{m_1}, f_i^{m_2}, \ldots, f_i^{m_t}\right]$.

Third, for each segmented region S_i^k, a max pooling operation is applied to the matrix B_i^k, and thus a vector Ins_i^k can be obtained, which just represents the instance of S_i^k.

Fourth, integrate all the instances $\left\{Ins_i^1, Ins_i^2, \ldots, Ins_i^{r_i}\right\}$ together, and then a bag for the sample image $train_i$ can be obtained, which is called the segmented instance bag and represented as:

$$BagS_i = \left[Ins_i^1, Ins_i^2, \ldots, Ins_i^{r_i}\right] \tag{2}$$

At last, for the whole training set $Train = \{train_1, train_2, \ldots, train_N\}$, by following the above steps, we can get its segmented instance bag set:

$$TrBagS = [BagS_1, BagS_2, \ldots, BagS_N] \tag{3}$$

2) LAYERED INSTANCE BAG

Besides the segmented instance bag, to overcome defects mentioned at the beginning of Section III.B, in this subsection, we continue to propose another bag named the layered instance bag for RS scene representation inspired by the idea of spatial pyramid.

For a sample image $train_i$, partition it via the idea of spatial pyramid. Suppose the number of the spatial pyramid layers is P. For the j-th layer, the image is partitioned into $2^{j-1} \times 2^{j-1}$ sub-regions. Thus, we can obtain u_i sub-regions $\{L_i^1, L_i^2, \ldots, L_i^{u_i}\}$ from all layers. For example, as shown in Fig. 4, P is equal to 3, and the scene image is partitioned into $u_i = 21$ sub-regions.

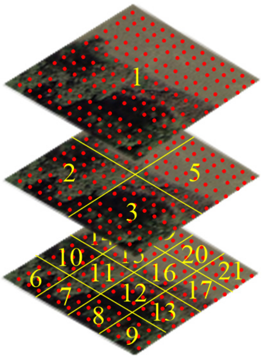


FIGURE 4. **Example of the layered instance bag.**

Second, according to the locations of the key feature points which are obtained in Section III.A, each partitioned region L_i^k may contain a number of key points. Suppose $\{f_i^{n_1}, f_i^{n_2}, \ldots, f_i^{n_t}\}$ are n_t fused feature vectors of these key points, combine these vectors to form a matrix $E_i^k = [f_i^{n_1}, f_i^{n_2}, \ldots, f_i^{n_t}]$.

Third, for each partitioned region L_i^k, a max pooling operation is applied to the matrix E_i^k, and thus a vector Inl_i^k can be obtained, which represents the instance of L_i^k.

Fourth, integrate all the instances $\{Inl_i^1, Inl_i^2, \ldots, Inl_i^{u_i}\}$ together, and then a bag for the sample image $train_i$ can be obtained, which is called the layered instance bag and represented as:

$$BagL_i = \left[Inl_i^1, Inl_i^2, \ldots, Inl_i^{u_i}\right] \tag{4}$$

At last, for the whole training set *Train*, by using the above steps, we can get its layered instance bag set:

$$TrBagL = [BagL_1, BagL_2, \ldots, BagL_N] \tag{5}$$

As can be seen, LIB is composed of a set of instances corresponding to the layered sub-regions. Compared with SIB, the advantage of LIB lies that it is based on the idea of spatial pyramid, so it can describe the scene images hierarchically. On the one hand, if there exist large objects in the scene image, a wide range of descriptions can be used for the entire large object and at the same time, it is also possible to describe some small objects with associations. On the other hand, as the range of descriptions is reduced, the local information of the scene image can be well represented, and it also makes the bag possess spatial structural information.

In all, for each image $train_i$, both SIB and LIB are constructed. The integration of them will be very good for the description of a complicated scene.

C. SIML WITH MULTI-BAG INTEGRATION

After obtaining the segmented instance bag as well as the layered instance bag, the next module aims to simplify the MIML learning problem into a single-instance multi-label (SIML) learning task by using a degeneration strategy. This is a commonly used way taken by the MIML algorithms, such as MIMLBOOST, MIMLSVM, MIMLSVM+ and E-MIMLSVM+ [13], [17]. In the degeneration process, how to measure the distances between instances is a vital step, and these algorithms always use Euclidean distance to measure the dissimilarity between instances. Then with the help of K-Medois clustering, the MIML learning task can be transformed into the SIML learning task. However, for multi-label RS scene classification, Euclidean distance can hardly capture the intrinsic dissimilarity in the feature and label spaces [27]. In addition, it is also unsuitable to maximize the distance between bags when minimizing the distance within each bag. Therefore, in this paper, we use Mahalanobis distance instead of Euclidean distance, for the Mahalanobis distance has been proven to be effective for preserving the intrinsic geometric information of both feature space and label space, and different distributions can be easily distinguished under Mahalanobis distance [19].

The specific procedures are given as below.

Given two instances x_i and x_j, suppose M is a positive semi-definite matrix. Mahalanobis distance between x_i and x_j is defined as:

$$d_{ij} = \sqrt{\left(x_i - x_j\right)^T M \left(x_i - x_j\right)} \tag{6}$$

Also, based on Mahalanobis distance, the distance between two bags X_i and X_j can be computed by:

$$D_{ij} = \sqrt{\left(\overline{X}_i - \overline{X}_j\right)^T M \left(\overline{X}_i - \overline{X}_j\right)} \tag{7}$$

where $\overline{X}_i, \overline{X}_j$ are the average values of all instances in X_i, X_j, respectively.

Under Mahalanobis distance, we apply K-Medoids clustering [18], [28] to the segmented instance bag set $TrBagS$ to seek for m cluster centers $CS = [CS_1, CS_2, \ldots, CS_m]$. Then, for each sample $train_i$, calculate the Mahalanobis distance IS_i^j from its segmented bag $BagS_i$ to each cluster center CS_j $(j = 1, 2, \ldots, m)$. Finally, integrate these distances $\{IS_i^1, IS_i^2, \ldots, IS_i^m\}$ together to form a vector

$IS_i = \left[IS_i^1, IS_i^2, \ldots, IS_i^m\right]$, which will be regarded as a single instance for $train_i$.

Similarly, for the layered instance bag set $TrBagL$, we also perform Mahalanobis distance-based K-Medoids clustering on it to find n cluster centers $CL = [CL_1, CL_2, \ldots, CL_n]$. Then, for each sample $train_i$, calculate the Mahalanobis distance IL_i^j from its segmented bag $BagL_i$ to each cluster center CL_j ($j = 1, 2, \ldots, n$). Finally, integrate these distances $\left\{IL_i^1, IL_i^2, \ldots, IL_i^n\right\}$ together to form a vector $IL_i = \left[IL_i^1, IL_i^2, \ldots, IL_i^n\right]$, which will be regarded as another single instance for $train_i$.

Subsequently, for each sample $train_i$, IS_i and IL_i which are derived from different kinds of bags are integrated in series to obtain a novel combined instance $I_i = [IS_i, IL_i]$.

At last, for the whole training set $Train$, we can get all combined instances $I = [I_1, I_2, \ldots, I_N]$. Thus, based on the multi-bag integration, the MIML learning task is successfully transformed into the SIML learning task.

D. AUTOMATIC SCENE LABELING VIA SISL

The last module of the proposed framework is the automatic scene labeling, which aims to assign appropriate labels to the scene images via single-instance single-label (SISL) learning [29]–[31]. In the third module, the MIML learning has been transformed into the SIML learning. Therefore, in this module, we adopt the following steps to change SIML to SISL for automatic scene labeling.

First, for each possible class label $c \in \{1, 2, \ldots, |\mathcal{L}|\}$, we train a binary SVM classifier. When training the classifier for the label c, all samples containing this label in the training set $Train$ are regarded as the positive samples, so that we can obtain the positive sample feature set $TrPos = \left\{I_{t_1}, I_{t_2}, \ldots, I_{t_p}\right\}$, where I_{t_i} is the combined instance of $train_{t_i}$, and the sample $train_{t_i}$ has the label c. Furthermore, we can also get the negative sample feature set $TrNeg = \left\{I_{z_1}, I_{z_2}, \ldots, I_{z_q}\right\}$, where I_{z_i} is the combined instance of $train_{z_i}$, and the sample $train_{z_i}$ does not have the label c. By using $TrPos$ and $TrNeg$, a binary SVM classifier can be trained.

Second, for complicated remote sensing scene images, the proportion of positive and negative samples under each label is usually unbalanced. Generally, the number of positive samples is small, while the number of negative samples is larger. Traditional classification methods may not perform well when encountering such an unbalanced data set. In order to solve this problem, we propose to set the weight of the positive sample to w ($w > 1$) and the weight of the negative sample to 1 when training each binary SVM classifier. In this way, the accuracy of the classifier can be effectively improved.

After training all classifiers for various labels, multi-label prediction can be performed on the test scene sample. Since a binary classifier is trained for each label, the test sample is predicted by all the classifiers at the same time. If the test image is judged to be a positive sample under a label, it is considered to have such a label and do not have this label otherwise. Finally, after multiple classifier predictions, multiple labels can be generated for the test image. For instance, for a certain test sample, the results of automatic scene labeling are shown in Eq. (8):

$$\text{if} \begin{cases} \mathrm{SVM}_1\,(TeI) = 1 \\ \mathrm{SVM}_2\,(TeI) = 0 \\ \ldots \\ \mathrm{SVM}_{|\mathcal{L}|}\,(TeI) = 1 \end{cases} \Rightarrow LabelTe = [1, 0, \ldots, 1] \tag{8}$$

where TeI is the combined instance of the test sample $test$. SVM_c is the trained binary SVM classifier for the label c. $LabelTe$ is the final label vector for $test$, in which if the element is equal to 1, it means the test image is assigned the corresponding label; if the element is equal to 0, it means the test image is not assigned the corresponding label.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

1) DATASET DESCRIPTION

Since this paper makes research on the multi-label remote sensing scene classification problem, the publicly available RS datasets with single labels, such as UC Merced Land Use dataset [32], SIRI-WHU [33], and WHU-RS [34], are no longer appropriate. Therefore, we manually extract 637 multi-label images from Google Earth to construct the multi-label RS scene dataset. Each image contains at least one class of scenes, including agriculture, forest, habitation, road, river, and sparse building. The size of each image is 320 × 320 pixels. Over 91% images in the dataset contain more than two categories of labels. The specific label distribution of our multi-label dataset is shown in Table 1, and some examples are illustrated in Fig. 5.

A well-known concern about machine learning is general over fitting. Therefore, to avoid over-fitting, among these 637 images, 60% of the images under per category are randomly selected for training, and the other 40% of images are used for testing.

2) EVALUATION METRICS

According to [35]–[39], five commonly used evaluation metrics, including Hamming Loss, Coverage, One Error, Ranking Loss, and Average Precision, are adopted to quantitatively evaluate the multi-label scene classification performance. The definition of these metrics is given as below. Let $S = \{(x_i, Y_i)\,|1 \leq i \leq p\}$ be the test set. $h(\cdot)$ denotes the multi-label classifier. Δ represents the symmetric difference between two label sets: one is the set of computed labels, and the other is the true label set. q denotes the size of label space. $rank_f\,(x, y)$ returns the rank of y in the label space via descending order. $f\,(x, y)$ returns the confidence of y being proper label of x [35]–[39]. $\overline{Y}$ is the complementary set of Y. $\langle\zeta\rangle$ returns 1 if ζ holds, and 0 otherwise. $\psi\,(\cdot, \cdot)$ returns the number of positive samples which are wrongly predicted.

TABLE 1. Label distribution of the multi-label dataset (f: forest, a: agriculture, h: habitation, ro: road, ri: river, sb: sparse building).

Labels	Number	Labels	Number
f	25	f+ri	112
a	26	f+ro	87
a+f	97	a+sb	106
f+h	69	f+h+ro	2
h+ro	99	a+f+ro	6
a+ro	8		

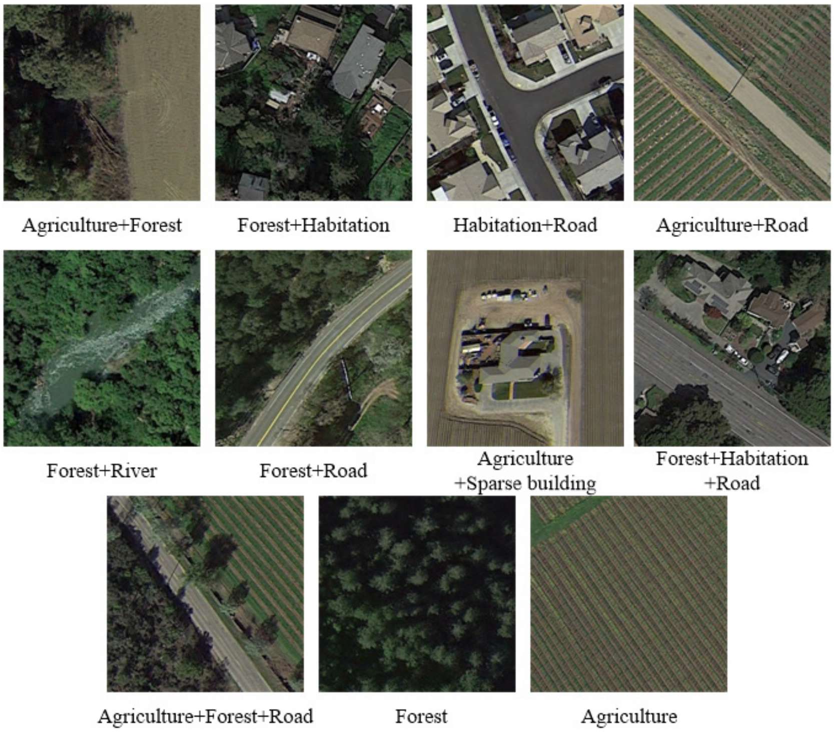


FIGURE 5. Examples of the multi-label remote sensing scene dataset.

(1) Hamming Loss: indicates the fraction of wrong labels to the total number of labels. Since it is a loss function, the lower the value of it is, the better the classification performance is.

$$HammingLoss\left(h\right)=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{q}\left|h\left(x_i\right)\Delta Y_i\right| \tag{9}$$

(2) Coverage: indicates the average depth to cover all true labels. Also, the smaller the value of it is, the better the classification performance is.

$$Coverage\left(f\right)=\frac{1}{p}\sum_{i=1}^{p}\max_{y\in Y_i} rank_f\left(x_i,y\right)-1 \tag{10}$$

(3) One Error: judges whether the top ranked label is not in set of true labels. The lower the value of it is, the better the classification performance is.

$$OneError\left(f\right)=\frac{1}{p}\sum_{i=1}^{p}\left\langle\left[\arg\max_{y\in Y} f\left(x_i,y\right)\right]\notin Y_i\right\rangle \tag{11}$$

(4) Ranking Loss: evaluates the average fraction of miss-ordered label pairs. Obviously, the lower the value of it is, the better the classification performance is.

$$RankingLoss\left(f\right)=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{\left|Y_i\right|\left|\overline{Y}_i\right|}\left|\left\{\left(y',y''\right)\right|f\left(x_i,y'\right)\le f\left(x_i,y''\right),\left(y',y''\right)\in Y_i\times\overline{Y}\right\}\right| \tag{12}$$

(5) Average Precision: assesses the average fraction of labels ranked above a specific label. Different from the above

four metrics, for average precision, the higher the value of it is, the better the classification performance is.

$$AveragePrecison\,(f) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i|}\sum_{y\in Y_i} \times \frac{\left|\left\{y'\,\middle|rank_f\left(x_i,y'\right)\leq rank_f\left(x_i,y\right),y'\in Y_i\right\}\right|}{rank_f\left(x_i,y\right)} \tag{13}$$

Besides the above five metrics, to more intuitively reflect the multi-label classification results, in this work, we propose three novel metrics, i.e., All Accuracy, All Error, and Correct One as supplement. They are defined as follows.

(6) All Accuracy: denotes the proportion of correctly predicted samples to the total number of samples. The larger the value of it is, the better the classification performance is.

$$AllAccuracy\,(h) = \frac{1}{p}\sum_{i=1}^{p}\langle h\left(x_i\right) == Y_i\rangle \tag{14}$$

(7) All Error: indicates the proportion of positive samples which are wrongly predicted to the total number of samples. The lower the value of it is, the better the classification performance is.

$$AllError\,(h) = \frac{1}{p}\sum_{i=1}^{p}\psi\left(h\left(x_i\right),Y_i\right) \tag{15}$$

(8) Correct One: reflects the proportion of correctly predicted samples as well as the samples in which only one positive or negative label is predicted wrongly to the total number of samples. The higher the value of it is, the better the classification performance is.

$$CorrectOne\,(h) = \frac{1}{p}\left(\sum_{i=1}^{p}\langle h\left(x_i\right) == Y_i\rangle + \sum_{i=1}^{p}\langle\left(h\left(x_i\right)\Delta Y_i\right) == 1\rangle\right) \tag{16}$$

3) COMPARED METHODS

First, we evaluate the effects of heterogeneous feature extraction on multi-label RS scene classification. Since in our proposed framework, three different features, D-SURF-LLC, Mean-Std-LLC, and MS-CLBP-LLC, are adopted. We compare them as well as their fused result with a commonly used feature extraction approach (named SBN) in [13].

Second, in our framework, two bags, namely segmented instance bag and layered instance bag, are designed and then integrated for complex scene representation. Therefore, we separately utilize these three different kinds of bags (i.e., the segmented instance bag, the layered instance bag, and the multi-bag integration) to verify the effectiveness of our proposed multi-bag integration scheme.

Third, to verify the overall performance of our whole framework, we compare it with a number of widely used MIML algorithms, including MIMLBOOST, MIMLSVM, MIMLSVM+ and E-MIMLSVM+. The first two methods compared in our experiments are the classical MIML methods, which are proposed by Zhou et al. [13] for scene classification based on a simple degeneration scheme. MIMLSVM+ and E-MIMLSVM+ are two improved MIML approaches and have been popularly adopted in existing literatures.

For our method, it involves some parameters, for instance, the region number r_i by the region segmentation, the sub-region number u_i by the sub-region partition, and the class number m and n by the K-Medoids clustering for the segmented instance bag set and layered instance bag set. Thereinto, for each sample image, the parameter r_i is adaptively chosen by using the strategy adopted in [26]. For the parameter u_i, its value depends on the number of the spatial pyramid layers P ($P \geq 2$). When the value of P become larger, the subsequent computational cost may continuously increase, and more importantly, an object with a certain semantic meaning may be partitioned into several sub-regions. Therefore, P is chosen as 3 by our tests on training sets. And thus, u_i is equal to 21. In addition, in our experiment, we empirically set the class number m and n of the K-Medoids clustering to be 20% of the number of training bags. Actually, it has been verified that the setting of this class number does not significantly affect the performance of MIML [12], [13], [20]. At last, to make a fair comparison, these algorithms are set to the best parameters which are reported in the papers.

All the experiments are performed on our multi-label dataset with ten trials of randomly partition of this dataset. The average performances of different methods under various metrics are calculated.

B. RESULTS AND DISCUSSION

1) EFFECTS OF HETEROGENEOUS FEATURE EXTRACTION

In this subsection, we evaluate the effects of different feature extraction methods on multi-label classification. In order to make the comparisons as fair and simple as possible, after obtaining the features, we only use the segmented instance bag for scene representation, and then adopt the same automatic labeling prediction approach introduced in our work for classification.

The comparison results are shown in Table 2. As can be seen, our proposed fused feature method achieves the best performance. Also, the results of each heterogeneous feature, D-SURF-LLC, Mean-Std-LLC, and MS-CLBP-LLC, are better than those of SBN. This reflects that feature extraction is a vital step for multi-label classification.

Specifically, the Hamming Loss result for our method indicates that our method achieves the lowest value of the fraction of wrong labels to the total number of labels. Besides, our method also achieves the lowest Coverage, One Error, Ranking Loss, All Error values, indicating the better performance against the other four algorithms. Furthermore, Table 2 also shows the Average Precision, All Accuracy, Correct One results of different approaches, the higher the values of which are, the better the performance is. Thus, it can be easily seen

TABLE 2. Performance comparison based on various features ('↓' indicates that smaller is better, '↑' indicates that larger is better).

Evaluation Metrics	SBN	D-SURF-LLC	Mean-Std-LLC	MS-CLBP-LLC	Our fused feature
Hamming Loss (↓)	0.273±0.008	0.155±0.010	0.115±0.009	0.185±0.006	0.077±0.009
Coverage (↓)	3.075±0.186	1.773±0.056	1.367±0.045	1.806±0.126	1.240±0.055
One Error (↓)	0.348±0.024	0.126±0.021	0.118±0.010	0.198±0.015	0.061±0.022
Ranking Loss (↓)	0.379±0.108	0.132±0.012	0.082±0.007	0.157±0.017	0.048±0.009
Average Precision (↑)	0.674±0.026	0.856±0.012	0.901±0.008	0.828±0.012	0.941±0.011
All Accuracy (↑)	0.089±0.027	0.413±0.035	0.625±0.024	0.427±0.029	0.704±0.024
All Error (↓)	0.338±0.027	0.094±0.022	0.085±0.012	0.152±0.019	0.029±0.014
Correct One (↑)	0.612±0.025	0.748±0.025	0.805±0.028	0.695±0.015	0.876±0.021

TABLE 3. Performance comparison of different kinds of bags (' ↓' indicates that smaller is better, '↑' indicates that larger is better).

Evaluation Metrics	Segmented instance bag	Layered instance bag	Multi-bag integration
Hamming Loss (↓)	0.077±0.009	0.038±0.007	0.027±0.005
Coverage (↓)	1.240±0.055	1.073±0.029	1.026±0.025
One Error (↓)	0.061±0.022	0.012±0.009	0.008±0.007
Ranking Loss (↓)	0.048±0.009	0.019±0.004	0.012±0.004
Average Precision (↑)	0.941±0.011	0.978±0.006	0.985±0.005
All Accuracy (↑)	0.704±0.024	0.837±0.025	0.879±0.017
All Error (↓)	0.029±0.014	0.005±0.006	0.004±0.004
Correct One (↑)	0.876±0.021	0.940±0.015	0.964±0.012

that our method gets the highest values for these three metrics, supporting our our proposed fused feature method as a competitive feature extraction method for multi-label RS scene classification. By designing suitable features and combining them together, the feature discriminative capability can be effectively enhanced.

2) EFFECTS OF MULTI-BAG INTEGRATION

In this subsection, we examine the ability of the proposed multi-bag integration technique. It is also compared with the separate bags, i.e., the segmented instance bag and the layered instance bag.

Table 3 summarizes the experimental results by using different kinds of bags. According to the experimental results, the following observations can be obtained: the proposed layered instance bag achieves a more satisfied performance than the segmented instance bag in terms of all the metrics. Specifically, the Hamming Loss, Coverage, One Error, Ranking Loss, as well as All Error values of the layered bag are much lower than those of the segmented bag, while the Average Precision, All Accuracy, and Correct One values of LIB are obviously higher than those of SIB. Furthermore, the multi-bag integration approach further enhances the classification power and obtains the best performance.

This behavior emphasizes the importance of integrating these two different kinds of bags. Based on the segmented instance bag, various objects with different semantic meanings may be divided into different segmented regions; while based on the layered instance bag, both of the large and small objects could be well described for LIB's spatial pyramid idea. As a result, by combining these two different kinds of bags together, the remote sensing scene can be well represented.

3) COMPARISON WITH OTHER ALGORITHMS

In this subsection, a comparison between the proposed framework of this paper and other algorithms-MIMLBOOST, MIMLSVM, MIMLSVM+ and E-MIMLSVM+ is made.

TABLE 4. Performance comparison of different MIML algorithms (' ↓' indicates that smaller is better, ' ↑ ' indicates that larger is better).

Evaluation Metrics	MIMLSVM	MIMLBoost	MIMLSVM+	E-MIMLSVM+	Proposed method
Hamming Loss (↓)	0.232±0.006	0.280±0.010	0.145±0.013	0.131±0.011	0.027±0.005
Coverage (↓)	2.000±0.040	2.565±0.167	2.365±0.100	2.298±0.081	1.026±0.025
One Error (↓)	0.284±0.016	0.357±0.044	0.031±0.017	0.045±0.015	0.008±0.007
Ranking Loss (↓)	0.189±0.008	0.286±0.025	0.310±0.019	0.293±0.014	0.012±0.004
Average Precision (↑)	0.685±0.010	0.576±0.014	0.814±0.022	0.819±0.016	0.985±0.005
All Accuracy (↑)	0.234±0.020	0.044±0.015	0.397±0.024	0.439±0.023	0.879±0.017
All Error (↓)	0.237±0.018	0.506±0.045	0.031±0.017	0.045±0.015	0.004±0.004
Correct One (↑)	0.627±0.030	0.470±0.052	0.806±0.048	0.844±0.037	0.964±0.012

The results of a quantitative comparison using eight evaluation metrics are shown in Table 4. As expected, it can be found that our proposed method outperforms existing state-of-the-arts for multi-label remote sensing scene classification.

For instance, the proposed method obtains around 30%, 41%, 17%, and 17% improvements compared to MIMLSVM, MIMLBoost, MIMLSVM+, and E-MIMLSVM+ about the metric of Average Precision, respectively. And the value of Hamming Loss for the proposed method is about 20%, 25%, 12%, and 10% less than those of MIMLSVM, MIMLBoost, MIMLSVM+, and E-MIMLSVM+, respectively. The reasons mainly lie in three-fold: image partitioning and heterogeneous feature extraction, multi-bag based scene representation, and SIML with multi-bag integration.

4) COMPUTATIONAL TIME EVALUATION

All experiments are run on a PC with Intel Core 2.3 GHz processor and 4.00 GB RAM. The implementation environment was under MATLAB 2010a.

For each training or testing image with the size of 320 × 320, the average computational time for heterogeneous features extraction and multi-bag construction is about 35 s. Then, the training time of the MIML classifier is about 612 s. Finally, the testing time for all test samples is about 2 s. As can be seen, the efficiency of the heterogeneous features extraction and multi-bag construction is not very high, since our method adopts multiple types of features as well as multiple kinds of bags, which brings a heavy burden on the methods. However, just because of this, our method achieves good performance for multi-label RS scene classification.

In fact, to reduce the overall execution time of the proposed method, an efficient C/C++ implement or even a parallel architecture could be used. Besides, as indicated in [10], [11], [14], the multi-label RS scene classification problem can be handled using parallel computing in a distributed environment, for it can considerably improve the performance by offering all shared computational and memory resources. Therefore, we will make deep research on this topic in the near future.

V. CONCLUSION

This paper focuses on the multi-label classification problem for remote sensing scene images. We have proposed a novel framework based on multi-bag integration for the problem.

This framework was divided into four main parts: image partitioning and heterogeneous feature extraction, multi-bag based scene representation, SIML with multi-bag integration, and automatic scene labeling via SISL. Through extracting heterogeneous features for scene images, the semantic contents can be well exploited. We also propose two kinds of bags: SIB and LIB. By integrating them together, the scenes can be effectively represented, which is significantly beneficial to multi-label classification. Also, the Mahalanobis distance-based K-Medoids approach is applied for SIML learning, and automatic scene labeling is done via SISL.

Experimental study was conducted on a real multi-label remote sensing scene dataset. Results showed that the proposed method significantly outperforms the existing ones. This was especially reflected in the average comparison, shown in Tables 2, 3, and 4, over Hamming Loss, Coverage, One Error, Ranking Loss, All Error, Average Precision, All Accuracy, and Correct One, where the proposed method was best in rank.

Obtained results encourage us to pursue future work in the area of applying the proposed method to a large set of datasets from different sources & problems. In addition, we plan to utilize the convolutional neural network to construct the image bags. At last, given the high computational complexity of MIML, we will also do research on the application of parallel and distributed strategies to our proposed method.

REFERENCES

[1] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.

[2] K. Chen, P. Jian, Z. Zhou, J. Guo, and D. Zhang, "Semantic annotation of high-resolution remote sensing images via Gaussian process multi-instance multilabel learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 6, pp. 1285–1289, Nov. 2013.

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

[4] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, pp. 1757–1771, Sep. 2004.

[5] J. Fürnkranz, E. Hüllermeier, E. L. Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *J. Mach. Learn.*, vol. 73, no. 2, pp. 133–153, Nov. 2008.

[6] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, pp. 333–359, Dec. 2011.

[7] J. C. Zaragoza, E. Sucar, E. Morales, C. Bielza, and P. Larrañaga, "Bayesian chain classifiers for multidimensional classification," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 2192–2197.

[8] G. Tsoumakas, I. Katakis, and L. Vlahavas, "Random k-labelsets for multilabel classification," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 7, pp. 1079–1089, Jul. 2011.

[9] M.-L. Zhang and L. Wu, "LIFT: Multi-label learning with label-specific features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 107–120, Jan. 2015.

[10] J. Gonzalez-Lopez, S. Ventura, and A. Cano, "Distributed nearest neighbor classification for large-scale multi-label data on spark," *Future Gener. Comput. Syst.*, vol. 87, pp. 66–82, Oct. 2018.

[11] P. Skryjomski, B. Krawczyk, and A. Cano, "Speeding up k-nearest neighbors classifier for large-scale multi-label learning on GPUs," *Neurocomputing*, vol. 354, pp. 10–19, Aug. 2019.

[12] Z.-H. Zhou and M.-L. Zhang, "Multi-instance multi-label learning with application to scene classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1609–1616.

[13] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artif. Intell.*, vol. 176, pp. 2291–2320, Jan. 2012.

[14] A. Cano, A. Zafra, and S. Ventura, "Speeding up multiple instance learning classification rules on GPUs," *Knowl. Inf. Syst.*, vol. 44, no. 1, pp. 127–145, 2015.

[15] G. Melki, A. Cano, and S. Ventura, "MIRSVM: Multi-instance support vector machine with bag representatives," *Pattern Recognit.*, vol. 79, pp. 228–241, Jul. 2018.

[16] J. He, H. Gu, and Z. Wang, "Bayesian multi-instance multi-label learning using Gaussian process prior," *Mach. Learn.*, vol. 88, nos. 1–2, pp. 273–295, 2012.

[17] Y.-X. Li, S. Ji, S. Kumar, J. Ye, and Z.-H. Zhou, "Drosophila gene expression pattern annotation through multi-instance multi-label learning," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 9, no. 1, pp. 98–112, Jan./Feb. 2012.

[18] Y. Xu, H. Min, H. Song, and Q. Wu, "Multi-instance multi-label distance metric learning for genome-wide protein function prediction," *Comput. Biol. Chem.*, vol. 63, pp. 30–40, Aug. 2016.

[19] S.-J. Huang, W. Gao, and Z.-H. Zhou, "Fast multi-instance multi-label learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published. doi: 10.1109/TPAMI.2018.2861732.

[20] X. Wang, X. Xiong, C. Ning, A. Shi, and G. Lv, "Integration of heterogeneous features for remote sensing scene classification," *Proc. SPIE*, vol. 12, no. 1, 2018, Art. no. 015023.

[21] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained Linear Coding for image classification," in *Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.

[22] Z.-F. He, M. Yang, Y. Gao, H.-D. Liu, and Y. Yin, "Joint multi-label classification and label correlations with missing labels and feature selection," *Knowl.-Based Syst.*, vol. 163, pp. 145–158, Jan. 2019.

[23] Z. Shao, K. Yang, and W. Zhou, "Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset," *Remote Sens.*, vol. 10, no. 6, p. 964, 2018.

[24] H. Zhang, W. Wu, and D. Wang, "Multi-instance multi-label learning of natural scene images: Via sparse coding and multi-layer neural network," *IET Comput. Vis.*, vol. 12, no. 3, pp. 305–311, Apr. 2018.

[25] A. T. Pham, R. Raich, and X. Z. Fern, "Dynamic programming for instance annotation in multi-instance multi-label learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2381–2394, Dec. 2017.

[26] M. B. Salah, A. Mitiche, and I. B. Ayed, "Multiregion image segmentation by parametric kernel graph cuts," *IEEE Trans. Image Process.*, vol. 20, no. 2, pp. 545–557, Feb. 2011.

[27] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. Neural Inf. Process. Syst. Conf.*, 2005, pp. 1473–1480.

[28] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, 2009.

[29] N. Nguyen, "A new SVM approach to multi-instance multi-label learning," in *Proc. 10th IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 384–392.

[30] A. T. Pham, R. Raich, and X. Z. Fern, "Multi-instance multi-label learning in the presence of novel class instances," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2427–2435.

[31] Q. Da, Y. Yu, and Z.-H. Zhou, "Learning with augmented class by exploiting unlabeled data," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1760–1766.

[32] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.

[33] 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.

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

[35] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[36] K. Brinker, J. Fürnkranz, and E. Hüllermeier, "A unified model for multilabel classification and ranking," in *Proc. 17th Eur. Conf. Artif. Intell.*, 2006, pp. 489–493.

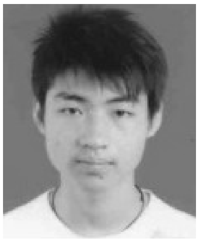
[37] K. Karalas, G. Tsagkatakis, M. Zervakis, and P. Tsakalides, "Deep learning for multi-label land cover classification," *Proc. SPIE*, vol. 9643, Oct. 2015, Art. no. 96430Q.

[38] N. Ghamrawi and A. McCallum, "Collective multi-label classification," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manage.*, 2005, pp. 195–200.

[39] O. E. Dai, B. Demir, B. Sankur, and L. Bruzzone, "A novel system for content-based retrieval of single and multi-label high-dimensional remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2473–2490, Jul. 2018.

XIN WANG was born in Fuyang, Anhui, China, in 1981. She received the B.S. and M.S. degrees in signal and information processing from Anhui University, Hefei, in 2006, and the Ph.D. degree in computer application technology from the Nanjing University of Science and Technology, Nanjing, in 2010.

From 2010 to 2013, she was an Assistant Professor with the College of Computer and Information, Hohai University, where she has been an Associate Professor, since 2014. She is the author of three books, more than 80 articles, and more than 50 patents. Her research interests include image processing and analysis, computer vision, pattern recognition, and computer vision. She was a recipient of the Science and Technology Progress Award, in 2015.



XINGNAN XIONG was born in Jiujiang, Jiangxi, China, in 1993. He received the B.S. degree in communication engineering from Jiangxi Normal University, Jiangxi, in 2016. He is currently pursuing the M.S. degree in signal and information processing with the College of Computer and Information, Hohai University, Nanjing, China.

His research interests include remote sensing image processing and analysis, computer vision, and pattern recognition.

CHEN NING was born in Fuyang, Anhui, China, in 1978. He received the B.S. degree in communication engineering from Anhui University, Hefei, in 2000, and the M.S. degree in signal and information processing from the University of Science and Technology of China, Hefei, in 2003.

Since 2010, he has been an Assistant Professor with the School of Physics and Technology, Nanjing Normal University. He is the author of 20 articles, and more than 10 patents. His research interests include image processing and analysis, computer vision, pattern recognition, and deep learning.