Zero-Shot Scene Classification for High Spatial Resolution Remote Sensing Images

Aoxue Li, Zhiwu Lu, Liwei Wang, Tao Xiang, and Ji-Rong Wen

***Abstract*—Due to the rapid technological development of various sensors, a huge volume of high spatial resolution (HSR) image data can now be acquired. How to efficiently recognize the scenes from such HSR image data has become a critical task. Conventional approaches to remote sensing scene classification only utilize information from HSR images. Therefore, they always need a large amount of labeled data and cannot recognize the images from an unseen scene class without any visual sample in the labeled data. To overcome this drawback, we propose a novel approach for recognizing images from unseen scene classes, i.e. zero-shot scene classification. In this approach, we first use the well-known natural language process model, word2vec, to map names of seen/unseen scene classes to semantic vectors. A semantic-directed graph is then constructed over the semantic vectors for describing the relationships between unseen classes and seen classes. To transfer knowledge from the images in seen classes to those in unseen classes, we make an initial label prediction on test images by an unsupervised domain adaptation model. With the semantic-directed graph and initial prediction, a label-propagation algorithm is then developed for zero-shot scene classification. By leveraging the visual similarity among images from the same scene class, a label refinement approach based on sparse learning is used to suppress the noise in the zero-shot classification results. Experimental results show that the proposed approach significantly outperforms the state-of-the-art approaches in zero-shot scene classification.**

***Index Terms*—Zero-shot learning, scene classification, high spatial resolution remote sensing images**

I. Introduction

WITH the development of modern sensor technologies, a large number of high spatial resolution (HSR) remote sensing images with abundant spatial and structural patterns are generated by various sensors everyday [1]–[5]. However, due to the huge volume and complex composition of remote sensing image data, it is difficult to directly access the HSR data that contains the scenes of interest. Therefore, how to efficiently recognize the scenes from HSR remote sensing images has become a challenging problem, which has drawn great interest in the remote sensing field [6]–[11].

In order to recognize and analyze the scenes from HSR remote sensing images, various scene classification approaches have been proposed in recent years. Zou *et al.* proposed a deep-belief-network-based feature selection strategy to construct discriminative features for scene classification [12]. Zhao *et al.* provided a concentric circle-structured multiscale bag-of-visual-words (BOVW) model using multiple features for land-use scene classification [13]. An unsupervised quaternion feature learning algorithm was proposed by Risojević *et al.* for remote sensing image scene classification, where quaternion representation was exploited to capture interrelationships between intensity and color information [14]. Zhong *et al.* proposed a semantic allocation level multifeature fusion strategy based on probabilistic topic model to effectively combine spectral and texture features for HSR remote sensing scene classification [15]. Li *et al.* proposed a multilayer feature learning approach to automatically learn simple edge features and complex corners/junctions features for satellite image scene classification [16]. Considering the importance of global features in interpreting the semantics in HSR remote sensing imagery, Zhu *et al.* improved the traditional BOVW model by introducing the shape-based invariant texture index as the global texture feature and then effectively combined the local BOVW and global features for HSR imagery scene classification [17]. In order to bridge the semantic gap between the low-level features and the high-level semantic concepts in HSR imagery scene classification, Zhao *et al.* proposed a Dirichlet-derived multiple topic model (DMTM) and then an efficient algorithm based on a variational expectation maximization framework was developed to infer the DMTM and estimate its parameters [1].

Although the aforementioned approaches have been shown to yield promising results in scene classification of HSR remote sensing images, they have two distinct drawbacks as follows:

- Firstly, these approaches need a certain number of labeled data for each scene class to train a good classifier for scene classification of HSR remote sensing images. However, some of scene classes are rare and collecting sufficient labeled training data for them may not be possible even if labeling cost is not a concern.
- Secondly, these approaches only utilize information from HSR remote sensing images for scene classification. Moreover, they cannot recognize the images from an unseen scene class that is not included in training data.

To overcome these two drawbacks, we introduce a new idea, termed zero-shot scene classification (ZSSC), to the remote sensing scene classification field. ZSSC is well-established in computer vision. However, to the best of our knowledge, it is still an unfamiliar paradigm in remote sensing. For humans, it

A. Li and L. Wang are with the Key Laboratory of Machine Perception (MOE), School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China.

Z. Lu and J.-R. Wen are with the Beijing Key Laboratory of Big Data Management and Analysis Methods, School of Information, Renmin University of China, Beijing 100872, China (email: zhiwu.lu@gmail.com).

T. Xiang is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom.

Fig. 1. Some class names and the corresponding examples from the Caltech-UCSD Birds 2011 data set.

is an easy task to recognize a new scene class even if they have not seen a single instance before. This is reasonable because a lot of knowledge is preserved and conveyed to humans via texts and nowadays online sources [18]–[21]. Therefore, combining seen instances and some auxiliary information (e.g. texts), humans can easily recognize new scene classes. Inspired by this phenomenon, researchers have proposed a new approach, zero-shot learning, which transfers knowledge from labeled data (from seen classes) to unlabeled data (from unseen classes) based upon some auxiliary information. Traditional zero-shot learning approaches are mainly developed for the tasks of recognizing natural images, such as bird image classification, human position estimation, and indoor object recognition. In these tasks, the classes usually have strong semantic correlations. For example, the Caltech-UCSD Birds 2011 data set [22] contains over 11,000 images from 200 types of birds and provides over 300 annotations per image. Fig. 1 provides some samples of class names and their corresponding visual examples in the Caltech-UCSD Birds 2011 data set. It can be seen that these class names are strongly semantically related, which is extremely important for recognizing images from unseen classes. However, for remote sensing scene classification, the names of typical scene classes are not so semantically related as those of the object classes in natural image recognition, which limits the use of the traditional zero-shot learning approaches in remote sensing.

In this paper, to overcome these limitations, we propose a novel zero-shot scene classification (ZSSC) approach for HSR remote sensing images. Concretely, the word2vec model [23], a well-known distributed word representation approach in natural language process, is firstly used to map names of scene classes (both seen and unseen) to semantic vectors. A semantic-directed graph is then constructed over the semantic vectors for describing the relationships between unseen classes and seen classes. Given that typical remote sensing scene classes have a limited amount of semantic relationships among each other, we adopt an unsupervised domain adaptation model [24] to provide an effective initial label prediction on test images. Here, this domain adaptation model can transfer knowledge from images in seen classes to those in unseen classes and thus helps to overcome the limitations of the traditional ZSSC approaches. Now with both the knowledge from images in seen classes and that from seen/unseen class semantic vectors residing on the same graph, a label-propagation algorithm [25] is developed to measure the distance between test images and each unseen class semantic vector for recognition of the unseen scene classes.

Note that the test images in the same class should have similar visual appearance. However, this is not directly considered in the above ZSSC approach, and thus there may exist strong noise in the zero-shot classification results. Therefore, we develop a label refinement approach based on sparse learning to obtain better results. Specifically, inspired by the successful use of L_1-optimization for noise reduction [26]–[29], we formulate the label refinement problem as noise reduction over the labels of test images, where the L_1-norm Laplacian regularization term is mainly used to reduce the noise in the labels. To solve the L_1-norm optimization problem efficiently, we limit the solution to the space spanned by the eigenvectors of the Laplacian matrix based upon the manifold structure of the data and solve this problem in a linear time complexity with respect to the number of test images. The framework of the proposed approach is illustrated in Fig. 2.

To verify the effectiveness of the proposed approach, we first conduct experiments on the UC Merced data set by randomly selecting a number of classes as seen classes and the other as unseen classes. To make the proposed approach more scalable in real-world applications, we also conduct experiments on a large HSR satellite image. Note that this HSR satellite image contains instances from seen/unseen scene classes, which are fully unlabeled in our experimental setting. In fact, since providing manual labels is expensive and time-consuming, we only use labeled remote sensing images from the RSSCN7 data set [12] for recognizing this satellite image, where both the RSSCN7 data set and the large satellite image are collected from Google Earth. Experimental results show that the proposed approach significantly outperforms the state-of-the-art zero-shot learning approaches [30]–[32].

The major contributions of this paper are as follows:

- This is the first work on scene classification of HSR remote sensing images without seeing any visual example in some scene classes (i.e. zero-shot scene classification). Our novelty mainly lies in that the need of labeled remote sensing images can be effectively reduced and the scalability of the traditional scene classification approaches can be improved for real-world applications.
- By leveraging the visual similarity among images from the same scene class, the proposed label refinement approach based on sparse learning can suppress the noise in the zero-shot classification results.

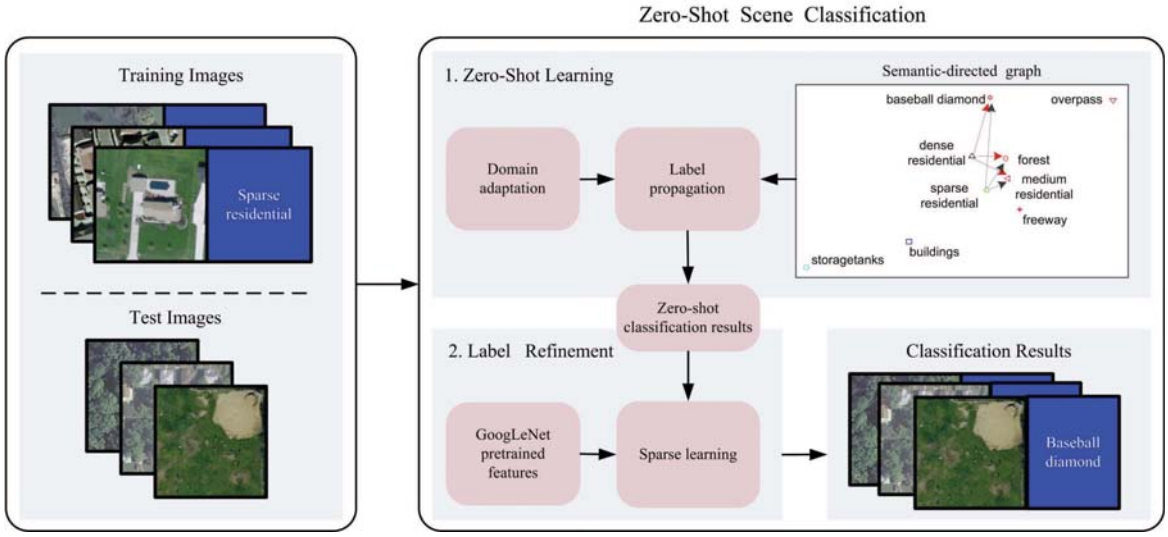

Fig. 2. The framework of the proposed approach to zero-shot scene classification.

- The proposed approach is shown to significantly outperform the state-of-the-art zero-shot learning approaches on both the UC Merced data set and the large HSR satellite image, which means that the proposed approach is more scalable in real-world applications.

The remainder of this paper is organized as follows. Section II provides a brief review of related works on zero-shot learning. Section III describes the details of the proposed approach for zero-shot scene classification of HSR remote sensing images. Section IV presents the experimental results to evaluate the performance of the proposed approach. Finally, the conclusions are drawn in Section V.

II. Related Works

Recently, many algorithms have been developed for zero-shot learning [30]–[36]. An attribute-based zero-shot learning approach was proposed by Lampert *et al.* to recognize different kinds of animals' images [34]. To describe the relationships between seen classes and unseen classes, Mensink *et al.* developed various metrics to leverage the co-occurrences of visual concepts in images, and then a regression approach was proposed to learn a weight for each related class [35]. Antol *et al.* proposed a visual-abstraction-based zero-shot learning approach to explore concepts related to people and their interactions with others, and achieved satisfactory results on human pose recognition [33]. Zhang *et al.* viewed test instances as arising from seen instances and attempted to express test instances as a mixture of seen class proportions [32]. To solve this problem, they proposed a semantic similarity embedding (SSE) approach for zero-shot learning. A general zero-shot learning framework which modeled the relationships between features, attributes, and classes as a two-linear-layers network was proposed by Paredes *et al.* to recognize animals and natural scenes [31]. Considering the manifold structure of semantic categories, Fu *et al.* provided a novel zero-shot learning approach by formulating a semantic manifold distance among test images and unseen classes [30]. Li *et al.* proposed a novel zero-shot learning approach that automatically learned label embeddings from the input data in a semi-supervised large-margin learning framework [36]. The above zero-shot learning approaches yielded promising results in the task of recognizing natural images. However, the semantic relationships among typical scene classes' names in remote sensing field are not so strong as those in natural image field, and thus these approaches have limited use for zero-shot scene classification of HSR remote sensing images.

III. Methodology

In this section, we provide the details of the proposed approach for zero-shot scene classification of HSR remote sensing images. Specifically, the proposed approach contains two main steps: 1) A zero-shot learning approach based on label propagation is developed for recognizing the HSR remote sensing images in unseen classes; 2) A label refinement approach based on sparse learning is used to suppress the noise in the zero-shot classification results.

A. Zero-Shot Learning Based on Label Propagation

Let $S = \{s_1, ..., s_p\}$ denote the set of seen classes and $U = \{u_1, ..., u_q\}$ denote the set of unseen classes, where p and q are the total numbers of seen classes and unseen classes, respectively. These two sets of classes are disjoint, i.e. $S \cap U = \phi$. We are given a set of labeled training images $D_s = \{(x_i, y_i) : i = 1, ..., M\}$, where x_i is the feature vector of the i-th image in the training set, $y_i \in S$ is the corresponding label, and M denotes the total number of labeled images. Let $D_u = \{(x_j, y_j) : j = 1, ..., N\}$ denote a set of unlabeled test images, where x_j is the feature vector of the j-th image in the test set, $y_j \in U$ is the corresponding unknown label, and N denotes the total number of unlabeled images. The main goal of zero-shot learning is to predict y_j by learning a classifier $f : X \to U$, where $X = \{x_j : j = 1, ..., N\}$.

For zero-shot learning, we need first estimate the semantic relationships between seen and unseen classes, which will be used for predicting the labels of images in unseen classes. In this paper, we adopt the word2vec model [23], which was trained with over 4,000,000 text documents from Wikipedia, to represent each class ($\in S \cup U$) by a semantic vector (empirically set as 400-dimensional). We further construct a semantic-directed graph $\mathcal{G} = \{V, E\}$ over all the classes, where V denotes the set of nodes (i.e. classes) in the graph and E denotes the set of directed edges between classes. The details of graph construction are given as follows:

- We first construct the edges among seen classes. For each seen class, the k-nearest-neighbors (k-NN) method is performed on the semantic vectors to find its k_1 nearest neighbors among seen classes. A directed edge is constructed between this class and each of its neighbors (from seen classes), and its edge weight is defined by applying Gaussian kernel (with the width=1) to the Euclidean distance between them.
- We further adopt the same strategy to construct the edges between seen classes and unseen classes. For each seen class, the k-NN method is performed on the semantic vectors to find its k_2 nearest neighbors among unseen classes. A directed edge is constructed between this class and each of its neighbors (from unseen classes), and its edge weight is defined by applying Gaussian kernel (with the width=1) to the Euclidean distance between them.
- Finally, for each unseen class, it has only one edge pointing to itself with a weight of 1.

By collecting the above edge weights up, we can denote the weight matrix $\mathcal{W}$ of the semantic-directed graph $\mathcal{G}$ as:

$$\mathcal{W} = \begin{bmatrix} \mathcal{R}_1 & \mathcal{R}_2 \\ 0 & I \end{bmatrix} \tag{1}$$

where $\mathcal{R}_1 \in \mathbb{R}^{p\times p}$ collects the edge weights among seen classes, $\mathcal{R}_2 \in \mathbb{R}^{p\times q}$ collects the edge weights between seen classes and unseen classes, and $I \in \mathbb{R}^{q\times q}$ is an identity matrix.

We further define a Markov chain process over $\mathcal{G}$ by constructing the transition matrix $\mathcal{T} = \mathcal{D}^{-1}\mathcal{W}$, where $\mathcal{D}$ is a $(p+q)\times(p+q)$ diagonal matrix with its i-th diagonal element being equal to the sum of the i-th row of $\mathcal{W}$. To guarantee that the Markov chain process has a unique stationary solution [25], we normalize the transition matrix $\mathcal{T}$ as follows:

$$P = \frac{\eta}{p+q-1}(1_{p+q} - I_{p+q}) + (1-\eta)\mathcal{T} \tag{2}$$

where η is a normalization parameter (which is empirically set as $\eta = 0.001$), and 1_{p+q} and I_{p+q} are the one matrix and identity matrix of the size $(p+q)\times(p+q)$, respectively.

Based on the normalized transition matrix $P = [p_{uv}] \in \mathbb{R}^{(p+q)\times(p+q)}$, we formulate zero-shot learning as a label propagation problem to propagate the labels from each unseen class semantic vector to a given test image and use the resultant propagation cost/probability as the distance for recognition:

$$\min_{F_{i\cdot}} \sum_{u,v} \pi(u)p_{uv}\left(\frac{F_{iu}}{\sqrt{\pi(u)}} - \frac{F_{iv}}{\sqrt{\pi(v)}}\right)^2 + \lambda\|F_{i\cdot} - Y_{i\cdot}\|_2^2 \tag{3}$$

where $F = [F_{iu}]_{N\times(p+q)}$ and $Y = [Y_{iu}]_{N\times(p+q)}$ collect the optimal and initial probabilities of the test images belonging to each category, respectively. Concretely, F_{iu} (or Y_{iu}) denotes the optimal (or initial) probability of the i-th test image belonging to the u-th category. Moreover, $F_{i\cdot}$ (or $Y_{i\cdot}$) denotes the i-th row of F (or Y). In addition, $\pi(u)$ is the sum of the u-th row of the transition matrix P (i.e. $\sum_v p_{uv}$), and λ is a positive regularization parameter.

The first term of the above objective function sums the weighted variation of $F_{i\cdot}$ on each edge of the directed graph $\mathcal{G}$, which aims to ensure that $F_{i\cdot}$ does not change too much between semantically similar classes for the i-th test image. The second term denotes an L_2-norm fitting constraint, which means that $F_{i\cdot}$ should not change too much from $Y_{i\cdot}$.

To solve the above label propagation problem, we adopt the technique introduced in [25] and define the operator Θ:

$$\Theta = (\Pi^{1/2}P\Pi^{-1/2} + \Pi^{-1/2}P\Pi^{1/2})/2 \tag{4}$$

where Π is a $(p+q)\times(p+q)$ diagonal matrix with its u-th diagonal element being equal to $\pi(u)$. According to [25], the optimal solution F^* of the problem in Equation (3) is:

$$F^* = Y(I - \alpha\Theta)^{-1} \tag{5}$$

where I is an identity matrix of the size $(p+q)\times(p+q)$ and $\alpha = 1/(1+\lambda) \in (0,1)$. This solution can be obtained at a linear time cost with respect to N.

For zero-shot learning, we need to provide Y in advance. Note that each row of Y consists of two parts: the probabilities of a test image belonging to seen classes, and the probabilities of a test image belonging to unseen classes. Given no labeled data in unseen classes, we directly set the probabilities belonging to unseen classes as 0. To compute the initial probabilities belonging to seen classes, we adopt an unsupervised domain adaptation model [24] which transfers the knowledge from labeled images in the seen classes to test images. This unsupervised domain adaptation model is developed based on a deep convolutional neural network which has three main components: a deep feature extractor, a label classifier, and a domain classifier. In this paper, we fine-tune a GoogLeNet model [37] to obtain the feature extractor and the label classifier, and also train a gradient reversal layer to connect the feature extractor and the domain classifier for transfer learning. For test images, the outputs of the softmax layer in the label classifier denote the initial probabilities belonging to seen classes. By integrating the knowledge from seen classes into the proposed ZSSC approach, we can better strengthen the relationships between seen classes and unseen classes, which makes the proposed ZSSC model more effective in recognizing images from unseen classes.

B. Label Refinement Based on Sparse Learning

Due to the limitations of word2vec in describing semantic relationships between seen and unseen scene classes, there still exists strong noise in the zero-shot classification results. Considering that images in the same scene class should have similar visual appearance, we thus propose a label refinement approach based on sparse learning to suppress the noise in the zero-shot classification results.

Before giving problem formulation, we model all test images X as a graph $\mathscr{G} = \{X, W\}$ with its vertex set X and weight matrix $W = [w_{ij}]_{N\times N}$, where w_{ij} denotes the similarity between image feature vectors x_i and x_j. In this paper, we use pre-trained GoogLeNet features (extracted from the last layer of the GoogLeNet model [37] trained using 1.2M images from ImageNet [38]) as the feature vectors of test images. Note that the weight matrix W is usually assumed to be nonnegative and symmetrical. In this paper, we define the weight matrix W by applying Gaussian kernel (with the width=1) to the Euclidean distances between the GoogLeNet pretrained feature vectors of any two test images. The normalized Laplacian matrix $\mathcal{L}$ of the graph $\mathscr{G}$ can be computed by

$$\mathcal{L} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{6}$$

where I is an $N \times N$ identity matrix, and D is an $N \times N$ diagonal matrix with its i-th diagonal element being equal to the sum of the i-th row of W (i.e. $\sum_j w_{ij}$). Based on eigenvalue decomposition, the normalized Laplacian matrix $\mathcal{L}$ can be decomposed into the following symmetrical form:

$$\mathcal{L} = V\Sigma V^T = (\Sigma^{\frac{1}{2}} V^T)^T (\Sigma^{\frac{1}{2}} V^T) = B^T B \tag{7}$$

where V is an $N \times N$ orthonormal matrix with each column being an eigenvector of $\mathcal{L}$, and Σ is an $N \times N$ diagonal matrix with its diagonal element Σ_{ii} being an eigenvalue of $\mathcal{L}$ (sorted as $0 \leq \Sigma_{11} \leq ... \leq \Sigma_{NN}$).

We further use the new matrix $B = \Sigma^{\frac{1}{2}} V^T$ to define an L_1-norm smooth measure and refine the zero-shot classification results from the viewpoint of noise reduction over the labels predicted by the proposed zero-shot learning approach

$$\min_{\tilde{F}} \frac{1}{2} \|\tilde{F} - F^*\|_F^2 + \gamma \|B\tilde{F}\|_1 \tag{8}$$

where $\tilde{F}_{i.}$ (the i-th row of $\tilde{F} \in \mathbb{R}^{N\times q}$) denotes the optimal probability of the i-th test image belonging to unseen classes, $F^*_{i.}$ (the i-th row of $F^* \in \mathbb{R}^{N\times q}$) denotes the probabilities of the i-th test image belonging to unseen classes given by Equation (5), and γ denotes a positive regularization parameter. The first term denotes an L_2-norm fitting constraint, which means that $\tilde{F}$ should not change too much from F^*. The second term denotes an L_1-norm Laplacian regularization, which means that $\tilde{F}$ should not change too much between visual similar images. The good property of the second term in noise reduction has been given and proven in [39].

Note that directly solving Equation (8) is computationally intractable, considering that only the computation of B would incur too large time cost. Fortunately, we can use the dimension reduction technique to efficiently solve this problem. Concretely, we limit $\tilde{F}$ to the space spanned by a small set of eigenvectors of the normalized Laplacian matrix $\mathcal{L}$. To ensure the consistence of $\tilde{F}$, we choose m eigenvectors with the smallest eigenvalues as the base vectors. That is, $\tilde{F} = V_m A$, where V_m stores m eigenvectors with the smallest eigenvalues and A denotes linear combination coefficients. Equation (8) can now be reformulated as follows:

$$\begin{aligned}
& \arg\min_A \frac{1}{2}||V_m A - F^*||_F^2 + \gamma||BV_m A||_1 \\
= & \arg\min_A \sum_{j=1}^{q} \frac{1}{2}||V_m A_{.j} - F^*_{.j}||_2^2 + \gamma||\Sigma^{\frac{1}{2}} V^T V_m A_{.j}||_1 \\
= & \arg\min_A \sum_{j=1}^{q} \frac{1}{2}||V_m A_{.j} - F^*_{.j}||_2^2 + \gamma \sum_{i=1}^{m} \Sigma_{ii}^{\frac{1}{2}} |a_{ij}| \\
= & \arg\min_A \sum_{j=1}^{q} \sum_{i=1}^{m} \frac{1}{2} a_{ij}^2 - (V_{.i}^T F^*_{.j}) a_{ij} + \gamma \Sigma_{ii}^{\frac{1}{2}} |a_{ij}|
\end{aligned} \tag{9}$$

where $F^*_{.j}$ denotes the j-th column of F^*, and $V_{.i}$ denotes the i-th column of V_m.

It can be seen that the L_1-minimization problem in Equation (8) has been decomposed into $q * m$ independent quadratic optimization subproblems:

$$a^*_{ij} = \arg\min_{a_{ij}} \frac{1}{2} a_{ij}^2 - (V_{.i}^T F^*_{.j}) a_{ij} + \gamma \Sigma_{ii}^{\frac{1}{2}} |a_{ij}| \tag{10}$$

which has the following explicit solution:

$$a^*_{ij} = \begin{cases} 0, & |V_{.i}^T F^*_{.j}| \leq \gamma \Sigma_{ii}^{\frac{1}{2}} \\ V_{.i}^T F^*_{.j} + \gamma \Sigma_{ii}^{\frac{1}{2}}, & V_{.i}^T F^*_{.j} < -\gamma \Sigma_{ii}^{\frac{1}{2}} \\ V_{.i}^T F^*_{.j} - \gamma \Sigma_{ii}^{\frac{1}{2}}, & V_{.i}^T F^*_{.j} > \gamma \Sigma_{ii}^{\frac{1}{2}} \end{cases} \tag{11}$$

In this way, Equation (8) can be solved efficiently at a linear time cost with respect to N.

To sum up, by combining zero-shot learning and label refinement together, the full algorithm for zero-shot scene classification is outlined as Algorithm 1.

Algorithm 1 The Proposed ZSSC Algorithm

Input: the set of labeled training images D_s
the set of test images in unseen classes X
Zero-Shot Learning Based on Label Propagation:
1) Compute the initial probabilities of test images belonging to unseen classes Y with the domain adaptation model [24];
2) Construct the semantic-directed graph based on semantic vectors extracted by the word2vec model [23];
3) Compute the normalized transition matrix P according to Equations (1-2);
4) Find the solution F^* of the label propagation problem in Equation (3) according to Equations (4-5);
Label Refinement Based on Sparse Learning:
5) Construct a k-NN graph with its weight matrix W being defined over X;
6) Compute the normalized Laplacian matrix $\mathcal{L}$ according to Equation (6);
7) Find the m smallest eigenvectors of the normalized Laplacian matrix $\mathcal{L}$ and store them in V_m;
8) Find the solution A^* of the L_1-minimization problem in Equation (8) according to Equations (9-11);
9) Label each test image x_i with scene class $\arg\max_j \tilde{F}^*_{ij}$, where $\tilde{F}^* = V_m A^*$.
Output: the labels of test images in unseen classes.

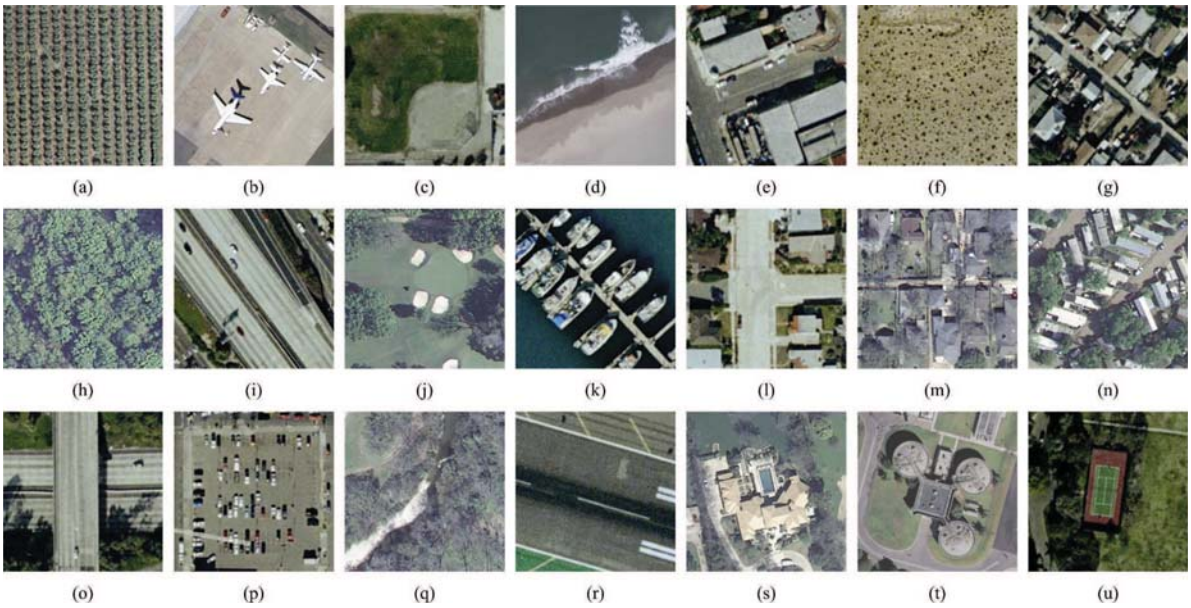

Fig. 3. Example images from 21 aerial scenes in the UC Merced data set. (a) Agricultural. (b) Airplane. (c) Baseballdiamond. (d) Beach. (e) Buildings. (f) Chaparral. (g) Denseresidential. (h) Forest. (i) Freeway. (j) Golfcourse. (k) Harbor. (l) Intersection. (m) Mediumresidential. (n) Mobilehomepark. (o) Overpass. (p) Parkinglot. (q) River. (r) Runway. (s) Sparseresidential. (t) Storagetanks. (u) Tenniscourt.

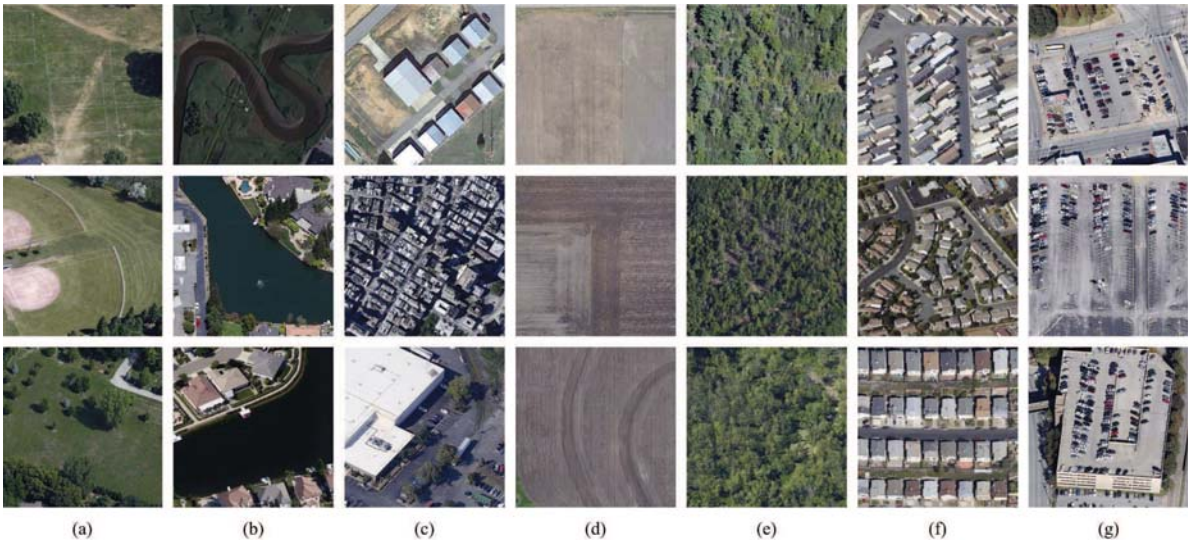

Fig. 4. Example images from 7 scene classes in the RSSCN7 data set. (a) Grass. (b) River. (c) Industrial. (d) Field. (e) Forest. (f) Residential. (g) Parking.

C. Computation Complexity Analysis

Note that both the label propagation problem in Equation (3) and the label refinement problem in Equation (8) can be solved efficiently at a linear time cost with respect to the total number of test images N. Hence, the proposed algorithm for zero-shot scene classification has a linear overall computational complexity, which is very important for the large scenes.

IV. Experimental Results

In this section, we provide the performance evaluation of the proposed ZSSC approach. We first describe the three datasets used in the experiments. We further report the results of zero-shot scene classification on the benchmark UC Merced data set [40]. In addition, we also evaluate the proposed approach in the task of recognizing a large HSR satellite image.

A. Description of the Data sets

The first data set used for performance evaluation is the UC Merced data set [40], which is the most widely used benchmark data set for remote sensing scene classification. This data set consists of 2,100 remote sensing images from 21 scene classes: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. Fig. 3 shows some

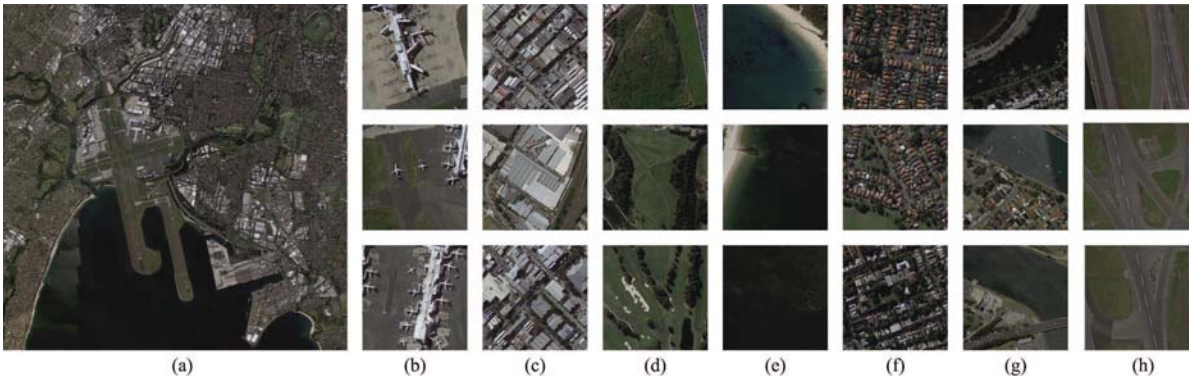

Fig. 5. The whole image and example subimages from 7 scene classes in the Sydney data set. (a) The whole image. (b) Airport. (c) Industrial. (d) Grass. (e) Ocean. (f) Residential. (g) River. (h) Runway.

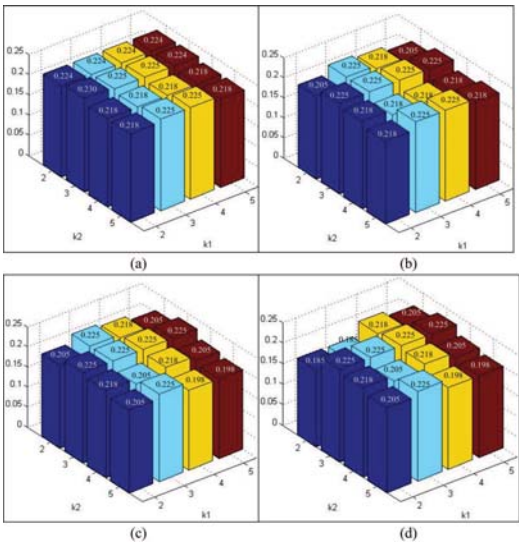

Fig. 6. Results of the label-propagation-based zero-shot learning method when tuning parameters k_1 and k_2 at different values of α. (a) α=0.1; (b) α=0.3; (c) α=0.5; (d) α=0.7.

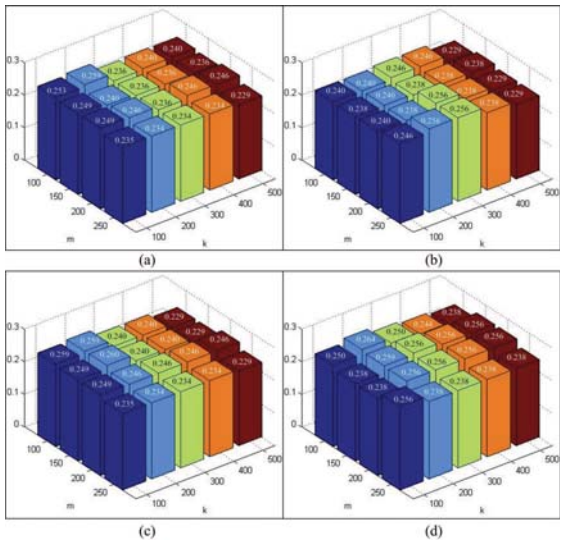

Fig. 7. Results of the label refinement method when tuning parameters k and m at different values of γ. (a) γ=0.3; (b) γ=0.5; (c) γ=0.7; (d) γ=0.9.

example images from the 21 aerial scenes. The images in this data set are manually extracted from large images from the USGS National Map Urban Area Imagery collection for various urban areas around the country. The pixel resolution of this public domain imagery is 1 foot. For each scene class, there are 100 images of the size 256×256 pixels.

The second data set is the RSSCN7 data set [12], which contains 2,800 remote sensing scene images collected from Google Earth. The images in this data set come from seven typical scene classes: grass, river, industrial, field, forest, residential, and parking. For each scene class, there are 400 images of the size 400×400 pixels. Some sample images from this data set are shown in Fig. 4.

The third data set is constructed from a large high-resolution satellite image which is acquired from Google Earth, for the city of Sydney, Australia. The spatial resolution of this large image is about 1 m. The large satellite image for Sydney is of $9,000 \times 9,000$ pixels, as shown in Fig. 5 (a). There exist seven scene classes within this large image: airport, industrial, grass, ocean, residential, river, and runway. Figs. 5(b)-(h) show some sample subimages of these scene classes. Specifically, the original large image is divided into 900 non-overlapping subimages of 300×300 pixels, where each subimage is supposed to only belong to a single scene class. In the following, the set of 900 subimages from the large satellite image is denoted as the Sydney data set.

B. Zero-Shot Scene Classification

1) Experimental Setup: To evaluate the effectiveness of the proposed approach, we conduct a group of experiments on the UC Merced data set by randomly selecting 16 of 21 classes as seen classes and the other 5 classes as unseen classes. Furthermore, we also test another 3 unseen/seen ratios to verify the effectiveness of the proposed approach in much weaker settings. For each unseen/seen ratio, the final classification results over unseen classes are averaged over 25 random seen/unseen splits. In this paper, we compare the proposed approach with the state-of-the-art zero-shot learning approaches [30]–[32].

TABLE I
PER-CLASS AND OVERALL CLASSIFICATION RATES (%) FOR DIFFERENT ZERO-SHOT LEARNING MODELS ON THE UC MERCED DATA SET. FOR COMPACTNESS, WE ONLY PROVIDE STANDARD DEVIATIONS (OVER 25 RANDOM SEEN/UNSEEN SPLITS) FOR THE OVERALL RATES.

Models	[30]	[32]	[31]	ZSL-LP	ZSSC
tenniscourt	31.6	15.3	46.0	43.0	60.0
storagetanks	4.1	51.4	13.1	36.4	42.7
runway	28.0	44.7	19.7	49.7	56.9
sparseresidential	11.5	19.0	6.5	23.7	26.8
parkinglot	41.0	70.0	7.0	46.4	55.4
river	5.8	37.5	16.7	24.0	25.3
overpass	97.9	7.6	9.3	99.6	99.6
mediumresidential	86.6	15.6	30.5	80.0	89.5
mobilehomepark	2.4	46.0	15.0	26.7	33.7
golfcourse	51.5	20.7	14.3	44.3	44.3
harbor	31.9	17.0	37.6	29.4	32.8
intersection	35.2	2.3	26.0	49.0	50.3
freeway	72.1	10.7	72.0	43.0	68.0
forest	12.0	28.3	18.0	24.3	21.3
buildings	44.8	18.0	7.4	47.0	79.0
chaparral	100.0	58.6	26.1	100.0	100.0
denseresidential	24.9	11.5	8.0	43.4	47.7
beach	54.8	1.2	44.9	42.7	44.3
baseballdiamond	77.0	50.2	45.5	67.5	72.8
airplane	43.0	16.0	5.1	72.0	76.3
agricultural	7.0	73.0	38.1	8.7	5.3
overall	47.2±1.7	32.5±1.1	25.4±0.7	53.9±1.3	58.7±0.9

2) Parameter Selection: The parameters of the proposed approach are selected in the unseen/seen ratio of 5/16. Concretely, we have randomly split the training set of the UC Merced data set into two halves and thus tuned the parameters in a two-fold cross-validation manner (i.e. images from 8 classes are used for training, and images from the other 8 classes are used for validation). For the label-propagation-based zero-shot learning method, we tune the parameters k_1 and k_2 at different values of α and the results are given in Fig. 6. We find that the label-propagation-based zero-shot learning method achieves the best result when $k_1 = 2$, $k_2 = 3$, and $\alpha = 0.1$. Therefore, we choose $k_1 = 2$, $k_2 = 3$, and $\alpha = 0.1$ for the proposed zero-shot learning model in the following experiments.

Moreover, for the label refinement method, we tune the parameters k, m, and γ in the same way. That is, the parameters k and m are tuned at different values of γ and the results are given in Fig. 7. We observe that the label refinement method achieves the best result when $k = 200$, $m = 100$, and $\gamma = 0.9$. Therefore, we choose $k = 200$, $m = 100$, and $\gamma = 0.9$ for the proposed label refinement method.

3) Comparison to the State-of-the-Art: Table I shows the comparison of the proposed ZSSC approach to the state-of-the-art zero-shot learning models [30]–[32] on the UC Merced data set. In this table, 'ZSL-LP' denotes the zero-shot learning method based on label propagation in Section III-A, while 'ZSSC' denotes the full ZSSC approach presented in Algorithm 1 (including label refinement). It can be seen that the proposed ZSSC approach not only significantly outperforms the state-of-the-art zero-shot learning models in terms of the overall rate, but also yields the best results over 10 scene classes with respect to the per-class rates. This observation can be explained as follows: 1) The models in [30]–[32] are proposed to cope with the traditional zero-shot learning tasks, where scene classes usually have strong semantic correlations. However, this is not the case for remote sensing scene classification. 2) The proposed ZSL-LP method not only exploits the semantic relationship between seen and unseen classes (also considered in the traditional zero-shot learning models [30]–[32]), but also uses a domain adaptation model to describe the relationships between the seen classes and images from unseen classes. This can strengthen the relationship between seen classes and unseen classes and thus effectively overcome the limitations of the traditional zero-shot learning models in remote sensing scene classification. 3) The label refinement method based on sparse learning can help to suppress the noise in the zero-shot learning results obtained by ZSL-LP (see ZSL-LP vs. ZSSC).

TABLE II
COMPARISON TO THE STATE-OF-THE-ART ZERO-SHOT LEARNING MODELS ON THE UC MERCED DATA SET WITH DIFFERENT UNSEEN/SEEN RATIOS. THE AVERAGE ACCURACIES ARE FOLLOWED BY STANDARD DEVIATIONS (OVER 25 RANDOM SEEN/UNSEEN SPLITS).

Models	Unseen/seen ratios			
	5 / 16	8 / 13	11 / 10	14 / 7
[30]	47.2 ± 1.7	21.2 ± 1.2	14.4 ± 0.7	12.1 ± 0.6
[32]	32.5 ± 1.1	18.1 ± 0.7	11.1 ± 0.5	7.5 ± 0.3
[31]	25.4 ± 0.7	15.2 ± 0.7	10.1 ± 0.2	7.3 ± 0.1
ZSL-LP	53.9 ± 1.3	28.4 ± 1.2	16.1 ± 0.7	12.4 ± 0.4
ZSSC	58.7 ± 0.9	35.4 ± 1.0	19.6 ± 0.5	15.1 ± 0.2

To further verify the effectiveness of the proposed approach, we also choose another three unseen/seen ratios (i.e. 8/13, 11/10 and 14/7) in the experiments. We randomly split the 21 scene classes in the UC Merced data set into seen and unseen classes according to the corresponding unseen/seen ratios. Table II provides the comparison of different zero-shot learning models on the UC Merced data set with different unseen/seen ratios. The accuracies are averaged over 25 random seen/unseen splits. It can be seen that performance of all the zero-shot learning models drops when the unseen/seen ratio increases, but our ZSSC approach outperforms the competing models in all cases. Note that the semantic correlation between unseen classes and seen classes becomes weaker with the increase of unseen/seen ratio. Since the model in [30] and the proposed ZSL-LP method both utilize the semantic correlation to infer the labels of test images, less semantic correlation between unseen classes and seen classes induces much more noise to the labels of test images (leading to the performance degradation). However, our label refinement method can utilize sparse learning to denoise the noisy labels, thus leading to significant improvement over the model in [30].

C. Large Satellite Image Recognition

1) Experimental Setup: To further evaluate the effectiveness of the proposed approach in large HSR satellite image recognition, we conduct another group of experiments as follows. Given that the RSSCN7 and Sydney data sets are both collected from Google Earth, we used the RSSCN7 data set as the training set and the Sydney data set as the test set. Under this setting, the set of seen scene classes contains: grass, river, industrial, field, forest, residential, and parking, while the set of unseen scene classes contains: ocean, runway,

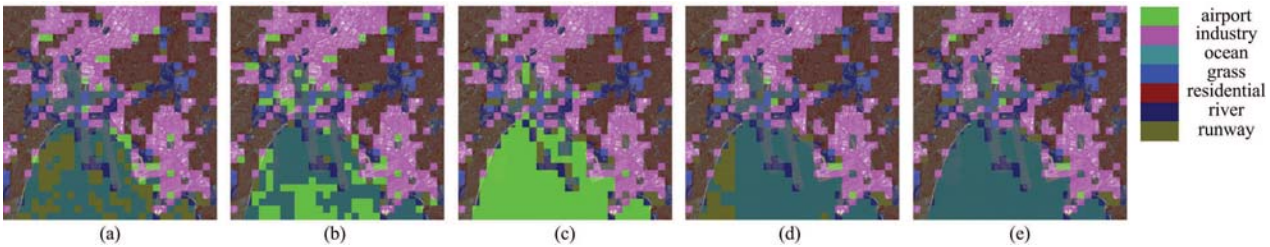

Fig. 8. Classification maps of the large satellite image with respect to seven scene classes, namely, airport, industrial, grass, ocean, residential, river, and runway. (a) Classification map by the approach in [30]. (b) Classification map by the approach in [32]. (c) Classification map by the approach in [31]. (d) Classification map by the proposed ZSL-LP approach. (e) Classification map by the proposed ZSSC approach.

TABLE III
PER-CLASS AND OVERALL CLASSIFICATION RATES (%) OF DIFFERENT ZERO-SHOT LEARNING MODELS ON THE SYDNEY DATA SET. THE OVERALL RATE IS COMPUTED OVER ALL THE IMAGES, AND NOT THE AVERAGE OF PER-CLASS RATES.

Models	[30]	[32]	[31]	ZSL-LP	ZSSC
airport (unseen)	7.4	12.8	0.0	3.7	3.7
industrial (seen)	88.6	88.6	88.6	88.6	88.6
grass (seen)	16.2	16.2	16.2	16.2	16.2
ocean (unseen)	51.8	58.1	0.0	71.2	93.7
residential (seen)	88.3	88.3	88.3	88.3	88.3
river (seen)	57.5	57.5	57.5	57.5	57.5
runway (unseen)	0.0	1.3	11.7	3.9	3.9
overall	61.3	63.0	51.1	65.7	70.4

TABLE IV
COMPARISON OF DIFFERENT ZERO-SHOT LEARNING MODELS IN TERMS OF TIME COST ON THE SYDNEY DATA SET.

Models	[30]	[32]	[31]	ZSL-LP	ZSSC
Time (sec.)	7.8	31.8	17.4	10.8	12.1

and airport. Note that the test images from the Sydney data set not only come from unseen classes (i.e. ocean, runway, and airport) but also from seen classes (i.e. grass, river, industrial, and residential). We thus need make novelty detection to determine whether a test image comes from a seen class or not. Concretely, for each seen class, a one-class support vector machine (SVM) classifier is trained with all the images from this class in the RSSCN7 data set, where the pre-trained GoogLeNet features are exacted for the training images. If all the trained one-class SVMs decide that a test image does not belong to any of the four seen classes (i.e. grass, river, industrial, and residential), we regard this image as an image from unseen classes and adopt the proposed ZSSC approach to predict its label; otherwise, only the domain adaptation approach [24] is used to predict its label. The parameters in the one-class SVM for novelty detection are tuned up on the images in the seen classes in the RSSCN7 data set.

2) Comparison to the State-of-the-Art: Note that there exist two main steps in large HSR satellite image recognition, i.e., novelty detection and zero-shot learning. The experimental results show that both of the two steps are effective on the Sydney data set. Specifically, the novelty detection method achieves an accuracy of 89.2% and the proposed ZSSC approach achieves an accuracy of 70.4%. Moreover, we made fair comparison to the state-of-the-art zero-shot learning models [30]–[32], by replacing the proposed ZSSC approach with these models and keeping the other settings unchanged. The per-class and overall classification rates of different zero-shot learning models are reported in Table III. It can be seen that the proposed approach not only significantly outperforms the state-of-the-art models in terms of the overall performance, but also yields the best results on the largest unseen class with respect to the per-class rate. In addition, the classification maps of all the models are also illustrated in Fig. 8. The proposed approach is still shown to perform the best. These observations demonstrate that the proposed approach is more scalable for real-world applications in remote sensing.

We also make comparison to the state-of-the-art zero-shot learning models on the Sydney data set in terms of time cost, which is shown in Table IV. All the experiments are conducted on a computer with 3.9 GHz CPU and 32GB RAM. From this table, we can observe that the time consuming of our approach on the large scenes is acceptable.

V. CONCLUSIONS

This paper proposes a novel scene classification approach for HSR remote sensing images to recognize images from unseen classes without any visual sample in the training set. In this approach, we first use the word2vec model to map names of scene classes to semantic vectors. A semantic-directed graph is then constructed over the semantic vectors for describing the relationships between unseen classes and seen classes. With the semantic-directed graph and knowledge transferred from images in the seen classes by an unsupervised domain adaptation model, a label-propagation algorithm is developed to measure the distance between test images and each unseen class for recognition of the unseen scene classes. To further suppress the noise in the zero-shot classification results, a label refinement approach is developed based on sparse learning. Experimental results show that the proposed approach significantly outperforms the state-of-the-art zero-shot learning models in scene classification for HSR remote sensing image. This means that the proposed approach can provide an effective way for remote sensing scene classification in the shortage of labeled data. In the future work, we will make further improvements in two aspects: 1) The word2vec model is trained only with the documents on geoscience and remote sensing to obtain better semantic relationships among scene classes; 2) Deep learning is used to directly formulate the zero-shot scene classification problem.

ACKNOWLEDGEMENTS

This work was partially supported by National Natural Science Foundation of China (61573363 and 61573026), 973 Program of China (2014CB340403 and 2015CB352502), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (15XNLQ01), and European Research Council FP7 Project SUNNY (313243).

REFERENCES

[1] 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 Transactions on Geoscience and Remote Sensing*, vol. 54, no. 4, pp. 2108–2123, 2016.

[2] F. Zhang, B. Do, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, pp. 1793–1802, 2016.

[3] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2175–2184, 2015.

[4] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using vhr remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 53, no. 8, pp. 4238–4242, 2015.

[5] M. L. Mekhalfi, F. Melgani, Y. Bazi, and N. Alajlan, "Land-use classification with compressive sensing multifeature fusion," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 10, pp. 2155–2159, 2015.

[6] K. Qi, H. Wu, C. Shen, and J. Gong, "Land-use scene classification in high-resolution remote sensing images using improved correlatons," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 12, pp. 2403–2407, 2015.

[7] G. Cheng, P. Zhou, J. Ha, L. Guo, and J. Han, "Auto-encoder-based shared mid-level visual dictionary learning for scene classification using very high resolution remote sensing images," *IET Computer Vision*, vol. 9, no. 5, pp. 639–647, 2015.

[8] J. Hu, T. Jiang, X. Tong, G.-S. Xia, and L. Zhang, "A benchmark for scene classification of high spatial resolution remote sensing imagery," in *IEEE International Geoscience and Remote Sensing Symposium*, 2015, pp. 5003–5006.

[9] B. Zhao, Y. Zhong, and L. Zhang, "Hybrid generative/discriminative scene classification strategy based on latent dirichlet allocation for high spatial resolution remote sensing imagery," in *IEEE International Geoscience and Remote Sensing Symposium*, 2013, pp. 196–199.

[10] F. P. S. Luus, B. P. Salmon, F. van den Bergh, and B. T. J. Maharaj, "Multiview deep learning for land-use classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 12, pp. 2448–2452, 2015.

[11] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 1947–1957, 2015.

[12] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2321–2325, 2015.

[13] L.-J. Zhao, P. Tang, and L.-Z. Huo, "Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model," *IEEE Journal of Seleteced Topics in Applied Eearth Observations and Remote Sensing*, vol. 7, no. 12, pp. 4620–4631, 2014.

[14] V. Risojevic and Z. Babic, "Unsupervised quaternion feature learning for remote sensing image classification," *IEEE Journal of Seleteced Topics in Applied Eearth Observations and Remote Sensing*, vol. 9, no. 4, pp. 1521–1531, 2016.

[15] 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 Transactions on Geoscience and Remote Sensing*, vol. 53, no. 11, pp. 6207–6222, 2015.

[16] Y. Li, C. Tao, Y. Tan, K. Shang, and J. Tian, "Unsupervised multilayer feature learning for satellite image scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 2, pp. 157–161, 2016.

[17] Y. Z. Q. Zhu, 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 Geoscience and Remote Sensing Letters*, vol. 13, no. 6, pp. 747–751, 2016.

[18] P. Kankuekul, A. Kawewong, S. Tangruamsub, and O. Hasegawa, "Online incremental attribute-based zero-shot learning," in *IEEE Conference Computer Vision and Pattern Recognition*, 2012, pp. 3657–3664.

[19] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Transactions on Pattern Analaysis and Machine Intelligence*, vol. 37, no. 11, pp. 2332–2345, 2014.

[20] Z. Akata, M. Malinowski, M. Fritz, and B. Schiele, "Multi-cue zero-shot learning with strong supervision," in *IEEE Conference Computer Vision and Pattern Recognition*, 2016, pp. 59–67.

[21] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE Transactions on Pattern Analaysis and Machine Intelligence*, vol. 38, no. 7, pp. 1425–1438, 2016.

[22] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," Tech. Rep., 2011.

[23] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Annual Conference on Neural Information Processing Systems*, 2014, pp. 1188–1196.

[24] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*, 2015, pp. 1180–1189.

[25] D. Zhou, J. Huang, and B. Scholkopf, "Learning from labelled and unlabelled data on a directed graph," in *International Conference on Machine Learning*, 2005, pp. 1036–1043.

[26] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, 2008.

[27] J. Wright, A. Yang, A. Ganesh, and S. Sastry, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analaysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[28] X. Chen, Q. Lin, S. Kim, J. Carbonell, and E. Xing, "Smoothing proximal gradient method for general structured sparse learning," in *Conference on Uncertainty in Artificial Intelligence*, 2011, pp. 105–114.

[29] G. Z. B.-K. Bao, J. Shen, and S. Yan, "Robust image analysis with sparse representation on quantized visual features," *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 860–871, 2013.

[30] Z. Fu, T. Xiang, E. Kodirov, and S. Gong, "Zero-shot object recognition by semantic manifold distance," in *IEEE Conference Computer Vision and Pattern Recognition*, 2015, pp. 2635–264.

[31] B. R. Paredes and P. H. S. Torr, "An embarrassingly simple approach to zero-shot learning," in *International Conference on Machine Learning*, 2015, pp. 2152–2161.

[32] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *IEEE Conference Computer Vision and Pattern Recognition*, 2015, pp. 4166–4174.

[33] S. Antol, C. L. Zitnick, and D. Parikh, "Zero-shot learning via visual abstraction," in *European Conference on Computer Vision*, 2014, pp. 401–416.

[34] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analaysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014.

[35] T. Mensink, E. Gavves, and C. G. M. Snoek, "Costa: Co-occurrence statistics for zero-shot classification," in *IEEE Conference Computer Vision and Pattern Recognition*, 2014, pp. 2441–2448.

[36] X. Li, Y. Guo, and D. Schuurmans, "Semi-supervised zero-shot classification with label representation learning," in *IEEE International Conference on Computer Vision*, 2015, pp. 4211–4219.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[39] Z. Lu and L. Wang, "Noise-robust semi-supervised learning via fast sparse coding," *Pattern Recognition*, vol. 48, no. 2, pp. 605–612, 2015.

[40] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *IEEE Conference Computer Vision and Pattern Recognition*, 2011, pp. 1465–1472.

Aoxue Li received the B.S. degree in electronic science and technology from Beijing Normal University, Beijing, China, in 2015. She is currently working toward the Ph.D. degree in computer science and technology at Peking University, Beijing. Her research interests include computer vision and machine learning.

Zhiwu Lu received the M.S. degree in applied mathematics from Peking University in 2005, and the Ph.D. degree in computer science from City University of Hong Kong in 2011. He is currently an associate professor of School of Information, Renmin University of China. He won the Best Paper Award at CGI 2014 and IBM SUR Award 2015. His research interests lie in machine learning, pattern recognition, and computer vision.

Liwei Wang received the Ph.D. degree from School of Mathematical Sciences, Peking University in 2005; the B.S. and M.S. degrees from Department of Electronic Engineering, Tsinghua University in 1999 and 2002, respectively. He is currently a full professor of School of Electronics Engineering and Computer Sciences, Peking University. He was named among "AI's 10 to Watch" in 2010. His research interest is machine learning, with application to computer vision.

Tao Xiang received the Ph.D. degree in electrical and computer engineering from the National University of Singapore in 2002. He is currently a reader (associate professor) in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include computer vision, machine learning, and data mining. He has published over 120 papers in international journals and conferences.

Ji-Rong Wen is a full professor at School of Information, Renmin University of China. He worked at Microsoft Research Asia for fourteen years and many of his research results have been integrated into important Microsoft products (e.g. Bing). He serves as an associate editor of ACM Transactions on Information Systems (TOIS). His main research interests include web data management, information retrieval, data mining and machine learning.