Learning Deep Cross-Modal Embedding Networks for Zero-Shot Remote Sensing Image Scene Classification

Yansheng Li, *Member, IEEE*, Zhihui Zhu, *Member, IEEE*, Jin-Gang Yu, and Yongjun Zhang, *Member, IEEE*

***Abstract*— Due to its wide applications, remote sensing (RS) image scene classification has attracted increasing research interest. When each category has a sufficient number of labeled samples, RS image scene classification can be well addressed by deep learning. However, in the RS big data era, it is extremely difficult or even impossible to annotate RS scene samples for all the categories in one time as the RS scene classification often needs to be extended along with the emergence of new applications that inevitably involve a new class of RS images. Hence, the RS big data era fairly requires a zero-shot RS scene classification (ZSRSSC) paradigm in which the classification model learned from training RS scene categories obeys the inference ability to recognize the RS image scenes from unseen categories, in common with the humans' evolutionary perception ability. Unfortunately, zero-shot classification is largely unexploited in the RS field. This article proposes a novel ZSRSSC method based on locality-preservation deep cross-modal embedding networks (LPDCMENs). The proposed LPDCMENs, which can fully assimilate the pairwise intramodal and intermodal supervision in an end-to-end manner, aim to alleviate the problem of class structure inconsistency between two hybrid spaces (i.e., the visual image space and the semantic space). To pursue a stable and generalization ability, which is highly desired for ZSRSSC, a set of explainable constraints is specially designed to optimize LPDCMENs. To fully verify the effectiveness of the proposed LPDCMENs, we collect a new large-scale RS scene data set, including the instance-level visual images and class-level semantic representations (RSSDIVCS), where the general and domain knowledge is exploited to construct the class-level semantic representations. Extensive experiments show that the proposed ZSRSSC method based on LPDCMENs can obviously outperform the state-of-the-art methods, and the domain knowledge further improves the performance of ZSRSSC compared with the general knowledge. The collected RSSDIVCS will be made publicly available along with this article.**

***Index Terms*— Latent space, locality-preservation deep cross-modal embedding networks (LPDCMENs), remote sensing (RS) imagery, transcendental knowledge, zero-shot RS scene classification (ZSRSSC).**

I. Introduction

ALONG with the rapid development of the remote sensing (RS) technology [1]–[3], massive spacecraft cameras have owned the ability to capture the high-resolution RS images of the earth surface, which provides abundant information about the ground objects [4]. As well known, accurate interpretation of the high-resolution RS imagery is the basic precondition of the wide RS-driven applications [5], [6]. For this reason, kinds of methods [7] have been exploited to automatically interpret the high-resolution RS imagery. Through years of efforts, scene-level classification [8], [9], which takes the scene (i.e., the image block) as the basic classification unit and aims at predicting the semantic category of one scene by perceiving the objects in the scene and their spatial relationships, has been widely realized to be a promising way to cope with the high-resolution RS image classification problem [10]. Compared with scene classification in the computer vision domain [11], RS image scene classification suffers from many additional challenges, such as arbitrary orientation and dense distribution of geospatial objects. As a consequence, RS image scene classification is still an open problem and deserves much more exploration.

To promote the development of RS image scene classification, multiple generous research groups in the RS community have publicly released their RS image scene data sets, such as UCM [12], AID [13], NWPU-RESISC45 [8], RSI-CB256 [14], and PatternNet [15]. Benefiting from this data sharing mechanism, experts from multiple related fields can contribute their efforts to the RS scene classification task. As a consequence, a great deal of progress has been made in RS scene classification. Among all existing RS scene classification methods, the methods based on deep networks [16] obviously outperform the others. However, the superior performance of such deep networks-based methods is highly conditioned to large amounts of labeled data [17]. In the RS big data era, new applications appear continually and bring

Manuscript received June 29, 2020; revised October 24, 2020; accepted December 22, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0505003; in part by the National Natural Science Foundation of China under Grant 41971284 and Grant 41601352; in part by the China Postdoctoral Science Foundation under Grant 2016M590716 and Grant 2017T100581; and in part by the Hubei Provincial Natural Science Foundation of China under Grant 2018CFB501. *(Corresponding authors: Yansheng Li; Zhihui Zhu; Jin-Gang Yu.)*

Yansheng Li and Yongjun Zhang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: yansheng.li@whu.edu.cn; zhangyj@whu.edu.cn).

Zhihui Zhu is with the Department of Electrical and Computer Engineering, University of Denver, Denver, CO 80208 USA (e-mail: zhihui.zhu@du.edu).

Jin-Gang Yu is with the School of Automation Science and Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: jingangyu@scut.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TGRS.2020.3047447.

Digital Object Identifier 10.1109/TGRS.2020.3047447

forward a series of new demands for the RS scene classification standard [18], which means that emerging applications often require the extension of an old RS scene category set. Therefore, it is impossible to annotate the RS scene samples on a complete RS scene category set containing all potential RS scene categories as the RS scene category set will change along with the emergence of new applications. Apparently, the RS big data era fairly requires a zero-shot RS scene classification (ZSRSSC) paradigm that the classification model learned on an existing RS scene category set should own the inference ability to recognize the RS image scenes from unseen categories (i.e., categories that do not exist in the given RS scene category set). Unfortunately, such ZSRSSC is rarely studied, i.e., the zero-shot classification technique is largely unexploited in the RS field.

On the one hand, the aforementioned ZSRSSC highly resembles the zero-shot learning (ZSL) task [19], [20] in the computer vision field. In the training stage of ZSL, only the labeled visual samples from seen categories are available, and there do not exist any visual samples from unseen categories. By leveraging the transcendental knowledge of the seen and unseen categories, ZSL is principally concerned with recognizing visual instances from unseen categories. Overall, the classic ZSL task mainly addresses the evolutionary recognition of natural images with the aid of universal knowledge. On the other hand, compared with the classic ZSL task, ZSRSSC suffers from the following two additional challenges.

1) There does not exist any kind of RS-domain-specific knowledge that can pertinently represent the RS scene categories [21]. In this case, the universal knowledge can be reluctantly used, but it lacks pertinence to use universal knowledge for describing the RS-domain-specific categories.
2) Compared with the natural images, it is much harder to mine the intrinsic content of the RS imagery due to the scale variation and arbitrary orientation of geospatial elements in the RS imagery [22].

As an early attempt of ZSRSSC, the well-known natural language process model (i.e., word2vec) is used to map the names of RS scene categories including seen and unseen categories to semantic vectors, and then, a graph propagation algorithm based on the semantic vectors is proposed to classify unseen visual samples [21]. To tackle the problem of class structure inconsistency between the visual image space and the semantic space that severely affects the performance of ZSRSSC, a semisupervised Sammon embedding algorithm [23] is proposed to address ZSRSSC. Similar to ZSRSSC, zero-shot RS image tree recognition is implanted by learning a compatibility function between the deep image features and the semantic features [24]. As a whole, deep learning only plays a role in the feature extractor in these methods [21], [23], [24], and the powerful ability of deep learning is still not fully embodied in the ZSRSSC task. Intuitively, how to utilize deep learning to thoroughly connect the visual image space and the semantic space would be a promising way to improve the performance of ZSRSSC.

With the aforementioned consideration, this article proposes a novel ZSRSSC method based on locality-preservation deep cross-modal embedding networks (LPDCMENs), which mainly aim to alleviate the problem of class structure inconsistency between the visual image space and the semantic space. By contrast to the existing ZSRSSC-related methods [21], [23], [24], the proposed LPDCMENs have two distinct characteristics: 1) LPDCMENs are composed of totally learnable modules in one unified framework to thoroughly match the visual images, instead of the frozen visual image features, and the semantic representations in the latent space and 2) LPDCMENs can assimilate the pairwise intramodal relationship supervision to preserve the aggregation of visual image samples from the same class (i.e., the locality-preservation characteristic). Under the supervision of the seen visual image samples and semantic representations, a set of explainable constraints, considering cross-modal matching via two different distance metrics (i.e., the cosine-like distance and the Euclidean distance) and latent feature regularization from two varied perspectives (i.e., the distribution balance of latent features and the variance maximum of latent features), is proposed to learn LPDCMENs in an end-to-end manner. Through a straightforward transfer, the learned LPDCMENs can be employed to recognize the unseen visual samples with the aid of unseen semantic representations. To fully verify the effectiveness of the proposed ZSRSSC approach, we construct a new RS scene data set, including the instance-level visual images and class-level semantic representations (RSSDIVCS). By integrating several existing RS image scene data sets, this article collects a unified one with more RS scene categories than any one of the publicly open data sets. The unified RS image scene data set is taken as the image part of the RSSDIVCS. To analyze the effect of knowledge types, two kinds of knowledge (i.e., the general and domain knowledge) are exploited to construct the corresponding class-level semantic representations of the RSSDIVCS. Extensive experiments show that the proposed ZSRSSC method based on LPDCMENs can obviously outperform the state-of-the-art methods and that the domain knowledge helps to further improve the performance of ZSRSSC compared with the general knowledge. As a whole, the main contributions of this article can be summarized as follows.

1) This article proposes a novel ZSRSSC approach based on LPDCMENs. The proposed LPDCMENs have two distinct characteristics.
    a) LPDCMENs are composed of totally learnable modules in one unified framework to thoroughly match the visual images, instead of the frozen visual image features, and the semantic representations in the latent space.
    b) LPDCMENs benefit from preserving the locality of visual image samples from the same class by assimilating the pairwise intramodal relationship supervision.
2) To pursue robust cross-modal matching in the latent space, a set of meaningful constraints, such as the cross-modal matching constraint via two different distance metrics (i.e., the cosine-like distance and the Euclidean distance) and the latent feature regularization constraint

using two varied subconstraints (i.e., the distribution balance subconstraint and the variance maximum subconstraint), is specifically designed to build the objective function for learning LPDCMENs. Benefiting from the well-designed objective function, LPDCMENs can be optimized in an end-to-end manner.

3) This article releases a large-scale RSSDIVCS, which can be taken as a benchmark for ZSRSSC and facilitates more knowledge-driven RS image scene understanding tasks.

The rest of this article is organized as follows. Section II reviews the related work. Section III specifically depicts the proposed LPDCMENs for ZSRSSC. Section IV reports and discusses the experimental results. Finally, Section V gives the conclusion of this article.

II. Related Work

In this section, we briefly review the most relevant works in the literature from two perspectives: RS scene classification and ZSL.

During the past decades, remarkable efforts have been made in developing kinds of methods for RS scene classification because of its wide applications including economic assessment [5], humanitarian aid [25], geospatial object detection [26], [27], and RS image retrieval [28], [29]. As RS scene classification aims to categorize image scenes to land-use and land-cover (LULC) classes based on the visual contents of image scenes, feature representation, being a straightforward way to depict the visual contents, plays a decisive role in RS scene classification. According to the adopted features, the existing RS scene classification methods can be coarsely divided into three kinds: the methods based on handcrafted-feature [12], [30]–[32], the methods based on unsupervised feature learning [33]–[35], and the methods based on supervised feature learning [9], [16], [36]. It is noted that the first two kinds of methods often further need a small set of labeled samples to train a shallow feature classifier, and the third kind generally tackles the feature representation and classification in one deep network with the aid of massive labeled samples. As a whole, the traditional RS scene classification methods can only recognize the samples from seen scene categories whose visual instances participate in the training stage and cannot handle the classification of the samples from unseen scene categories whose visual instances do not appear in the training stage. However, as explained earlier, how to recognize samples from unseen scene categories is highly required in the RS big data era.

Inspired by the phenomena that humans can easily recognize a new class even they have not seen a single instance before, pioneers in the computer vision field realize the possibility of evolutionary classification and exploit various methods for ZSL [37]. Specifically, based on the auxiliary knowledge (e.g., the semantic representations of categories), ZSL aims at learning a computational model on visual samples from the seen classes to recognize visual samples from the unseen classes. Overall, the matching strategy between the visual space and semantic space plays a key role in ZSL. According to the used matching strategies, the existing ZSL methods can be divided into three categories: matching in the semantic space [19], [38], matching in the visual space [39], and matching in the latent space [40]–[43]. Afterward, ZSL has been successfully extended to address multilabel classification [44] and large-scale retrieval [45]. In addition to these efforts in the computer vision domain, there are also many interesting attempts toward ZSL in the RS applications. Considering the association between the optical and SAR samples, the visual features of optical samples are utilized to generate the semantic representations of SAR labels [46], which further supports the recognition of SAR targets from unseen categories. From the simulation perspective, ZSL based on deep learning has effectively addressed the SAR target recognition task [47]. By learning a compatibility function between deep features of images and semantic representations of categories, zero-shot street tree classification using aerial imagery can significantly outperform the existing ones [24]. As far as the argued topic (i.e., ZSRSSC) in this article, the existing works [21], [23] mainly focus on operations on semantic representations where Li *et al.* [21] mainly aim at modeling the relationship between the seen and unseen categories in the semantic space and Quan *et al.* [23] try to align the structure between the visual image space and the semantic space. Although deep learning has been considered in the existing ZSRSSC methods [21], [23], deep learning only works as the prearranged deep feature extractor, and the existing ZSRSSC methods lack the joint treatment of deep learning between the visual image space and the semantic space, which makes the performance of the existing ZSRSSC approaches still unsatisfactory.

III. Zero-Shot Scene Classification via Deep Cross-Modal Embedding

In this section, we introduce the proposed ZSRSSC method. As visually shown in Fig. 1, the presented ZSRSSC method is implemented by bridging the visual images and semantic representations in the latent space. In the context of ZSRSSC, this article recommends LPDCMENs to thoroughly match the visual images and semantic representations in the latent space. As shown in Fig. 2, LPDCMENs are composed of the visual image mapping subnetworks via convolutional neural networks (V-CNNs) and the semantic representation mapping subnetworks via neural networks (S-NNs). After the mapping of V-CNNs and S-NNs, the RS image scenes and semantic representations are projected to the latent space. In the latent space, the RS image scenes from the same category will be close to each other (i.e., the locality-preservation property) and grouped around the semantic representation of the given category (i.e., the cross-modal matching property). In addition, it is straightforward to extend LPDCMENs to match image scenes and semantic representations from unseen categories. Hence, LPDCMENs are qualified to address evolutionary classification (i.e., the aforementioned ZSRSSC task).

To facilitate the following discussion, we give a formal definition of ZSRSSC in Section III-A. As mentioned earlier, in the context of ZSRSSC, the number of semantic representations is very limited. Under this limitation, Section III-B

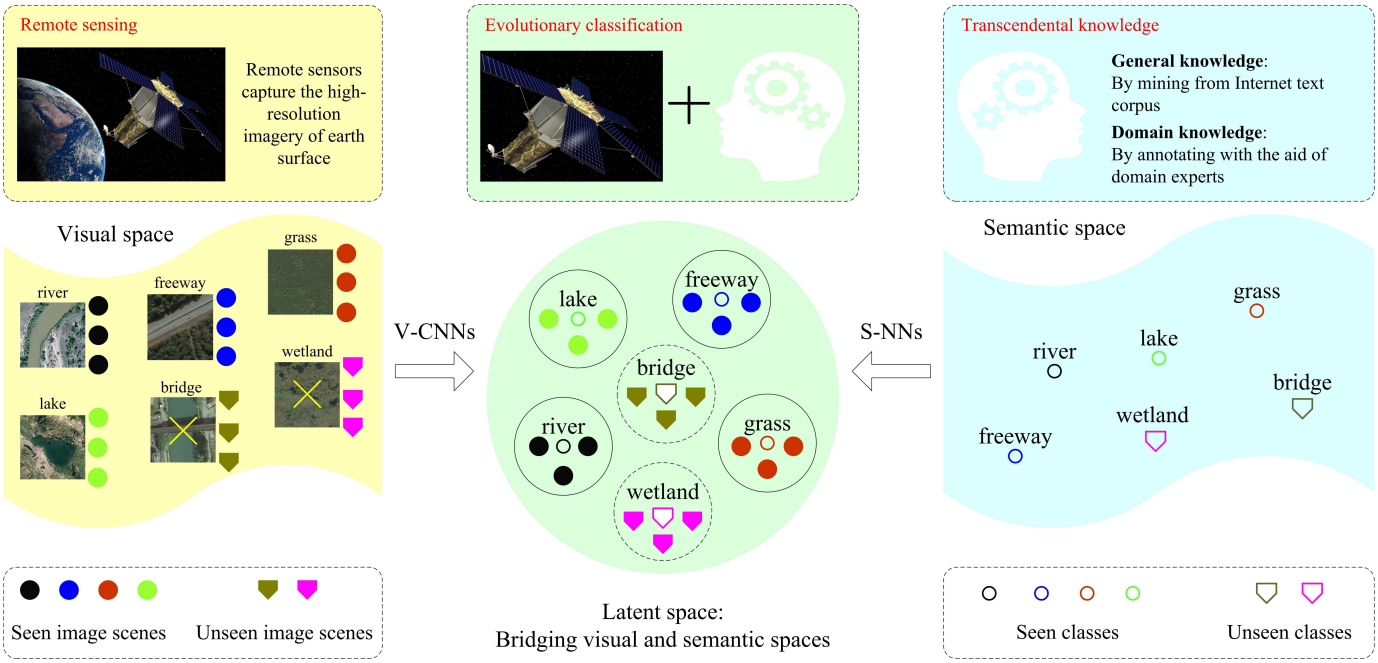


Fig. 1. Flowchart of the proposed ZSRSSC method. The instance-level visual images and class-level semantic representations from the seen categories are employed to train LPDCMENs. In the latent space, the dotted circles stand for the unseen scene categories that do not participate in the training stage. In the testing stage, LPDCMENs are competent for ZSRSSC due to their good generalization ability in bridging the visual images and semantic representations from unseen classes.

depicts how to robustly optimize LPDCMENs. In addition, Section III-C depicts how to conduct ZSRSSC based on the learned LPDCMENs.

A. Formulated Definition of Zero-Shot Scene Classification

Let $\mathbf{D}^S = \{\mathbf{h}_i^S, l_i^S\}_{i=1}^{N^S}$ denote the training data in the visual space, where $\mathbf{h}_i^S$ stands for the ith RS image scene, and l_i^S represents its corresponding label from $\ell^S = \{1^S, 2^S, \ldots, C^S\}$. In addition, we have $\mathbf{D}^U = \{\mathbf{h}_i^U, l_i^U\}_{i=1}^{N^U}$ as the testing data in the visual space, where $\mathbf{h}_i^U$ denotes the ith RS image scene, and l_i^U stands for the associated label from $\ell^U = \{1^U, 2^U, \ldots, C^U\}$. For ZSRSSC, the label sets ℓ^S and ℓ^U from the training data and testing data are disjoint (i.e., $\ell^S \cap \ell^U = \varnothing$). By simulating the humans' prior knowledge, each class is associated with a semantic vector. More specifically, let $\mathbf{F}^S = \{\mathbf{f}_c^S \in \mathbb{R}^d\}_{c=1}^{C^S}$ and $\mathbf{F}^U = \{\mathbf{f}_c^U \in \mathbb{R}^d\}_{c=1}^{C^U}$ denote the semantic representations of the training data and testing data, where d denotes the dimension of the semantic representation.

The goal of ZSRSSC is to predict the unseen labels of $\{\mathbf{h}_i^U\}_{i=1}^{N^U}$ by leveraging the visual data $\mathbf{D}^S = \{\mathbf{h}_i^S, l_i^S\}_{i=1}^{N^S}$ and the semantic representations $\mathbf{F}^S = \{\mathbf{f}_c^S \in \mathbb{R}^d\}_{c=1}^{C^S}$ and $\mathbf{F}^U = \{\mathbf{f}_c^U \in \mathbb{R}^d\}_{c=1}^{C^U}$. It should be noted that semantic representations from both the training data and testing data are accessible across the training and testing stages, which is the basic premise of ZSRSSC and easily satisfied because semantic representations come from the humans' common sense, and we can construct them in advance.

B. Deep Cross-Modal Embedding in the Latent Space

Instead of matching in the visual space or semantic space, we perform a match in the latent space for ZSRSSC as matching in the latent space alleviates the data inconsistency across hybrid spaces and enforces the within-class data locality to be preserved in the derived manifold (i.e., the latent space). Intuitively, matching in the latent space needs an embedding model. As illustrated in Fig. 2, the embedding function is implemented by LPDCMENs. More specifically, LPDCMENs are composed of V-CNNs and S-NNs where V-CNNs hierarchically map the RS image scene to the vector in the latent space and S-NNs work for projecting the semantic vector to the vector in the latent space. In the following, we introduce the proposed objective function and optimization strategy for learning LPDCMENs.

1) Objective Function for Learning Deep Cross-Modal Embedding Networks: To pursue the within-class aggregation and between-class separation, we adopt the pairwise relationship measure. Before giving the detailed measure, we first define the pairwise matrices based on the training data $\mathbf{D}^S = \{\mathbf{h}_i^S, l_i^S\}_{i=1}^{N^S}$, and $\mathbf{F}^S = \{\mathbf{f}_c^S \in \mathbb{R}^d\}_{c=1}^{C^S}$. If $\mathbf{h}_i^S$ and $\mathbf{h}_j^S$ have the same label, $\mathrm{V}_{i,j}^1 = 1$; otherwise, $\mathrm{V}_{i,j}^1 = 0$.$\mathrm{V}_{i,j}^2 = 1 - \mathrm{V}_{i,j}^1$. Similarly, we let $\mathrm{E}_{i,j}^1 = 1$ if $i = j$ and $\mathrm{E}_{i,j}^1 = 0$ otherwise. $\mathrm{E}_{i,j}^2 = 1 - \mathrm{E}_{i,j}^1$. If $\mathbf{h}_i^S$ and $\mathbf{f}_j^S$ come from the same class, $\mathrm{M}_{i,j}^1 = 1$; otherwise, $\mathrm{M}_{i,j}^1 = 0$. $\mathrm{M}_{i,j}^2 = 1 - \mathrm{M}_{i,j}^1$.

Let Θ^X denote the hyperparameters of V-CNNs, and Θ^Y stands for the hyperparameters of S-NNs. Furthermore, $\mathbf{X}_i = \varphi(\mathbf{h}_i^S; \Theta^X) \in \mathbb{R}^m$ denotes the mapped vector in the latent space where $\varphi(\cdot; \Theta^X)$ embeds the visual space to the latent space and m denotes the vector dimension in the latent space, and $\mathbf{Y}_i = \varphi(\mathbf{f}_i^S; \Theta^Y) \in \mathbb{R}^m$ also denotes the mapped vector in the latent space where $\varphi(\cdot; \Theta^Y)$ conducts embedding from the semantic space to the latent space.

Based on the aforementioned notations, the pairwise intramodal similarity between the instance-level visual image

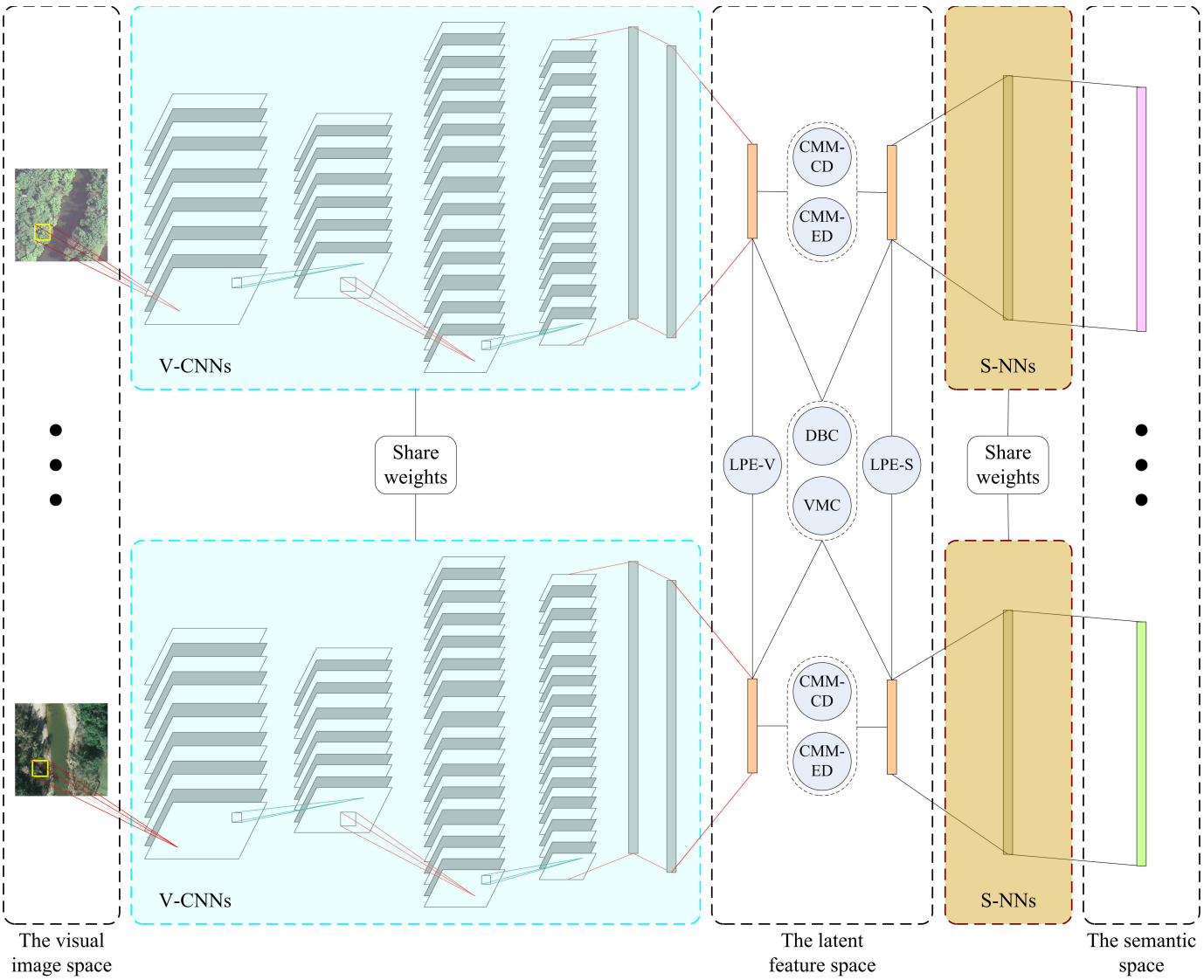


Fig. 2. Architecture of LPDCMENs. The LPDCMENs are composed of two hybrid subnetworks (i.e., V-CNNs and S-NNs). It is worth noting that our proposed LPDCMENs fully consider the joint optimization of multiple images and semantic features, which benefits preserving the locality of images in the latent space.

scenes can be formulated by the following equation:

$$\begin{cases} p\left(\mathrm{V}_{i,j}^{1}=1|\mathbf{X}\right)=\sigma\left(\Omega_{i,j}\right) \\ p\left(\mathrm{V}_{i,j}^{2}=1|\mathbf{X}\right)=1-\sigma\left(\Omega_{i,j}\right) \end{cases} \tag{1}$$

where $\mathbf{X}=[\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_{N^S}]$, $\sigma(\Omega_{i,j})=1/(1+e^{-\Omega_{i,j}})$ denotes the sigmoid function, $\Omega_{i,j}=\mathbf{X}_i^T\cdot\mathbf{X}_j/\delta$ stands for the weighted inner product, and δ is a weighted constant.

The pairwise intramodal similarity between the class-level semantic representations can be defined by the following equation:

$$\begin{cases} p\left(\mathrm{E}_{i,j}^{1}=1|\mathbf{Y}\right)=\sigma\left(\Pi_{i,j}\right) \\ p\left(\mathrm{E}_{i,j}^{2}=1|\mathbf{Y}\right)=1-\sigma\left(\Pi_{i,j}\right) \end{cases} \tag{2}$$

where $\mathbf{Y}=[\mathbf{Y}_1,\mathbf{Y}_2,\ldots,\mathbf{Y}_{C^S}]$, $\sigma(\Pi_{i,j})=1/(1+e^{-\Pi_{i,j}})$, and $\Pi_{i,j}=\mathbf{Y}_i^T\cdot\mathbf{Y}_j/\delta$.

The pairwise intramodal similarity between the visual images and semantic representations can be formulated by the following equation:

$$\begin{cases} p\left(\mathrm{M}_{i,j}^{1}=1|\mathbf{X},\mathbf{Y}\right)=\sigma\left(\Psi_{i,j}\right) \\ p\left(\mathrm{M}_{i,j}^{2}=1|\mathbf{X},\mathbf{Y}\right)=1-\sigma\left(\Psi_{i,j}\right) \end{cases} \tag{3}$$

where $\sigma(\Psi_{i,j})=1/(1+e^{-\Psi_{i,j}})$ and $\Psi_{i,j}=\mathbf{X}_i^T\cdot\mathbf{Y}_j/\delta$.

As explained earlier, LPDCMENs should own the locality preservation, cross-modal matching, and generalization abilities. To robustly learn LPDCMENs, we design a new objective function based on these constraints, which is formulated in (4). Note that, in (4), the pairwise intramodal relationship supervision (i.e., the pairwise matrices V and E) is adopted to form the locality-preservation constraint, which is composed of the locality-preservation embedding constraint for visual samples (LPE-V) and the locality-preservation embedding constraint for semantic representations (LPE-S). To pursue robust cross-modal matching, the cross-modal matching constraint is implemented by fusing two different distance metrics, including the cosine-like distance (CMM-CD) and the Euclidean distance

(CMM-ED). To lift the generalization performance, this article recommends the usage of the latent feature regularization constraint that considers two varied perspectives, including the distribution balance subconstraint of the latent features (DBC) and the variance maximum subconstraint of the latent features (VMC). To facilitate understanding, we intuitively illustrate the constructed constraints, including LPE-V, LPE-S, CMM-CD, CMM-ED, DBC, and VMC in Fig. 2

$$\begin{aligned}\min_{\Theta^X,\Theta^Y} J = &\overbrace{\left(\sum_{i=1}^{N^S}\sum_{j=1}^{N^S}\sum_{k=1}^{2}\left(-V_{i,j}^k \log p\left(V_{i,j}^k=1|\mathbf{X}\right)\right)\right)}^{\text{LPE-V}}\\ &+\frac{N^S}{C^S}\cdot\overbrace{\left(\sum_{i=1}^{C^S}\sum_{j=1}^{C^S}\sum_{k=1}^{2}\left(-E_{i,j}^k \log p\left(E_{i,j}^k=1|\mathbf{Y}\right)\right)\right)}^{\text{LPE-S}}\\ &+\alpha\cdot\overbrace{\left(\sum_{i=1}^{N^S}\sum_{j=1}^{C^S}\sum_{k=1}^{2}\left(-\mathrm{M}_{i,j}^k \log p\left(\mathrm{M}_{i,j}^k=1|\mathbf{X},\mathbf{Y}\right)\right)\right)}^{\text{CMM-CD}}\\ &+\beta\cdot\overbrace{\left(\sum_{j=1}^{C^S}\left\|\mathbf{Y}_j-\frac{1}{N_j^S}\sum_{i=1}^{N^S}\mathrm{M}_{i,j}^1\cdot\mathbf{X}_i\right\|_F^2\right)}^{\text{CMM-ED}}\\ &+\gamma\cdot\overbrace{\left(\|\mathbf{Z}\cdot\mathbf{1}\|_F^2\right)}^{\text{DBC}}+\eta\cdot\overbrace{\left(\left\|\mathbf{ZHZ}^T-\mathbf{I}\right\|_F^2\right)}^{\text{VMC}} \end{aligned}\tag{4}$$

where $\mathbf{Z}=[\mathbf{X},\mathbf{Y}]\in\mathbb{R}^{m\times(N^S+C^S)}$; $\mathbf{1}\in\mathbb{R}^{(N^S+C^S)\times 1}$ stands for a vector with all elements equal to 1; $\mathbf{I}\in\mathbb{R}^{m\times m}$ is an identity matrix; $\mathbf{H}=((\mathrm{N}^S+\mathrm{C}^S)\cdot\mathbf{G}-\mathbf{K})/(N^S+C^S)$; $\mathbf{G}\in\mathbb{R}^{(N^S+C^S)\times(N^S+C^S)}$ denotes an identity matrix; $\mathbf{K}\in\mathbb{R}^{(N^S+C^S)\times(N^S+C^S)}$ is a matrix with all elements equal to 1; $N_j^S=\sum_{i=1}^{N^S}\mathrm{M}_{i,j}^1$; and α,β,γ, and η stand for the empirical weights of multiple constraints (i.e., CMM-CD, CMM-ED, DBC, and VMC). For the proposed cross-modal matching constraint, α and β adjust the relative contribution degree of two different distance metrics (i.e., CMM-CD and CMM-ED). In the experimental section, ablation experiments about α and β are conducted to quantitatively explain the superiority of the joint consideration of different distance metrics. As far as the latent feature regularization constraint, γ and η tune the relative contribution degree of two varied latent feature regularization subconstraints (i.e., DBC and VMC). To check the rationality of the recommended latent feature regularization constraint, ablation experiments about γ and η are also conducted in the experimental section.

By plugging the pairwise similarity functions given in (1)–(3) into (4), the objective function for optimizing LPDCMENs can be rewritten as follows:

$$\begin{aligned}\min_{\Theta^X,\Theta^Y} J = &\overbrace{\left(\sum_{i=1}^{N^S}\sum_{j=1}^{N^S}\left(-V_{i,j}^1\cdot\Omega_{i,j}+\log\left(1+e^{\Omega_{i,j}}\right)\right)\right)}^{\text{LPE-V}}\\ &+\frac{N^S}{C^S}\cdot\overbrace{\left(\sum_{i=1}^{C^S}\sum_{j=1}^{C^S}\left(-\mathrm{E}_{i,j}^1\cdot\Pi_{i,j}+\log\left(1+e^{\Pi_{i,j}}\right)\right)\right)}^{\text{LPE-S}}\\ &+\alpha\cdot\overbrace{\left(\sum_{i=1}^{N^S}\sum_{j=1}^{C^S}\left(-\mathrm{M}_{i,j}^1\cdot\Psi_{i,j}+\log\left(1+e^{\Psi_{i,j}}\right)\right)\right)}^{\text{CMM-CD}}\\ &+\beta\cdot\overbrace{\left(\sum_{j=1}^{C^s}\left\|\mathbf{Y}_j-\frac{1}{N_j}\sum_{i=1}^{N^s}\mathrm{M}_{i,j}^1\cdot\mathbf{X}_i\right\|_F^2\right)}^{\text{CMM-ED}}\\ &+\gamma\cdot\overbrace{\left(\|\mathbf{Z}\cdot\mathbf{1}\|_F^2\right)}^{\text{DBC}}+\eta\cdot\overbrace{\left(\left\|\mathbf{ZHZ}^T-\mathbf{I}\right\|_F^2\right)}^{\text{VMC}}.. \end{aligned}\tag{5}$$

2) Optimization Strategy for Learning Deep Cross-Modal Embedding Networks: As shown in the objective function in (5), the hyperparameters of two subnetworks need to be optimized, and the latent vectors are also unobserved in advance. With this consideration, we optimize them via an alternative learning strategy where one variant is optimized, while the others are fixed. Like the popular optimization skill of deep learning [28], we adopt the stochastic gradient descent (SGD) to learn the hyperparameters of LPDCMENs using the following three steps over a fixed number of iterations iterMax.

1) Calculate $\mathbf{X}$ and $\mathbf{Y}$. More specifically, the RS image scenes can be mapped to $\mathbf{X}$ based on the hyperparameters of V-CNNs Θ^X, and the semantic representations can be projected to $\mathbf{Y}$ based on the hyperparameters of S-NNs Θ^Y.
2) Update Θ^X by fixing Θ^Y. With respect to the latent feature vector $\mathbf{X}_i$, we can obtain the closed-form gradient of the objective function in (5), where the gradient can be expressed by (6). Furthermore, the gradient is used to update Θ^X by SGD

$$\begin{aligned}\frac{\partial J}{\partial\mathbf{X}_i} = &\frac{2}{\delta}\cdot\sum_{j=1}^{N^S}\left(\sigma\left(\Omega_{i,j}\right)\cdot\mathbf{X}_j-V_{i,j}^1\cdot\mathbf{X}_j\right)\\ &+\frac{\alpha}{\delta}\cdot\sum_{j=1}^{C^S}\left(\sigma\left(\Psi_{i,j}\right)\cdot\mathbf{Y}_j-\mathbf{M}_{i,j}^1\cdot\mathbf{Y}_j\right)\\ &+\frac{2\beta}{N_j^s}\cdot\sum_{j=1}^{C^S}\left(\mathrm{M}_{i,j}^1\cdot\left(\frac{1}{N_j^S}\sum_{i=1}^{N^S}\mathrm{M}_{i,j}^1\cdot\mathbf{X}_i-\mathbf{Y}_j\right)\right)\\ &+2\gamma\cdot(\mathbf{Z}\cdot\mathbf{1})+2\eta\cdot\left(\mathbf{ZHZ}^T-\mathbf{I}\right)\mathbf{ZH}_i^T \end{aligned}\tag{6}$$

where $\mathbf{H}_i$ denotes the ith column of $\mathbf{H}$ and $i=1,2,\ldots,N^S$.
3) Update Θ^Y by fixing Θ^X. With respect to the latent feature vector $\mathbf{Y}_j$, we can obtain the closed-form gradient of the objective function in (5), where the gradient can be expressed by (7). Furthermore, the gradient is used

Algorithm 1 Optimization Algorithm for Learning LPDCMENs

Input: The training visual data $\mathbf{D}^S = \{\mathbf{h}_i^S, l_i^S\}_{i=1}^{N^S}$, the training semantic data $\mathbf{F}^S = \{\mathbf{f}_c^S \in \mathbb{R}^d\}_{c=1}^{C^S}$

Output: The hyperparameters of LPDCMENs including Θ^X and Θ^Y

Initialization: Θ^X and Θ^Y are initialized. The vector dimension in the latent space m and the weighted constant δ are empirically set. The number of iterations $iterMax$. The weights $\alpha, \beta, \gamma, \eta$ in Eq. (5) are empirically set.

repeat

Calculate the latent vectors $\mathbf{X}$ and $\mathbf{Y}$ based onΘ^X and Θ^Y;

for i=1,2,..., N^S **do**

- Calculate the visual gradient of the ith RS image scene according to Eq. (6);
- Update Θ^X by back propagating the visual gradient;

end for

for j=1,2,..., C^S **do**

- Calculate the semantic gradient of the j-th semantic representation according to Eq. (7);
- Update Θ^Y by back propagating the visual gradient;

end for

until a fixed number of iterations $iterMax$

to update Θ^Y by SGD

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{Y}_j} = &\frac{2}{\delta}\cdot\frac{N^S}{C^S}\cdot\sum_{j=1}^{C^S}\left(\sigma\left(\Pi_{i,j}\right)\cdot\mathbf{Y}_j - \mathrm{E}_{i,j}^1\cdot\mathbf{Y}_j\right)\\ &+\frac{\alpha}{\delta}\cdot\sum_{i=1}^{N^s}\left(\sigma\left(\Psi_{i,j}\right)\cdot\mathbf{X}_i - \mathbf{M}_{i,j}^1\cdot\mathbf{X}_i\right)\\ &+2\beta\cdot\left(\mathbf{Y}_j - \frac{1}{N_j^S}\sum_{i=1}^{N^s}\mathrm{M}_{i,j}^1\mathbf{X}_i\right)\\ &+2\gamma\cdot(\mathbf{Z}\cdot\mathbf{1}) + 2\eta\cdot\left(\mathbf{Z}\mathbf{H}\mathbf{Z}^T - \mathbf{I}\right)\mathbf{Z}\mathbf{H}_{N+j}^T \quad (7)\end{aligned}$$

where $\mathbf{H}_{N^s+j}$ denotes the $N^S + j$th column of $\mathbf{H}$ and $j = 1, 2, \ldots, C^S$.

To facilitate understanding, we summarize the iteratively alternative optimization procedure in Algorithm 1.

As depicted in Algorithm 1, the hyperparameters of LPDCMENs, including Θ^X and Θ^Y, can be learned in an end-to-end manner. In the following, Section III-C introduces how to conduct ZSRSSC based on the learned LPDCMENs.

C. Recognizing Scenes From Unseen Classes via the Learned Deep Cross-Modal Embedding Networks

Although LPDCMENs are learned on the training data, which does not have any label overlap with the testing data, the feature mapping function of LPDCMENs can be generalized to the testing data. Before recognizing the testing visual scenes, we should first construct the semantic templates $\{\mathbf{Y}_1^U, \mathbf{Y}_2^U, \ldots, \mathbf{Y}_{C^U}^U\}$ (i.e., the semantic representations in the latent space) based on the testing semantic representations $\mathbf{F}^U = \{\mathbf{f}_c^U \in \mathbb{R}^d\}_{c=1}^{C^U}$, which is formulated by the following equation:

$$\mathbf{Y}_c^U = \varphi\left(\mathbf{f}_c^U; \Theta^Y\right) \quad (8)$$

where $c = 1, 2, \ldots, C^U$.

Given the ith testing image scene $\mathbf{h}_i^U$, it can be mapped to the latent space via $\mathbf{X}_i^U = \varphi(\mathbf{h}_i^U; \Theta^X)$. As LPDCMENs own a good ability to match visual and semantic data in the latent space, the class label of $\mathbf{X}_i^U = \varphi(\mathbf{h}_i^U; \Theta^X)$ can be inferred by the following equation:

$$\arg\max_c \frac{\left\langle \mathbf{X}_i^U, \mathbf{Y}_c^U \right\rangle}{\left\|\mathbf{X}_i^U\right\| \cdot \left\|\mathbf{Y}_c^U\right\|} \quad (9)$$

where $< \cdot, \cdot >$ denotes the inner product of two latent feature vectors and $c = 1, 2, \ldots, C^U$

We summarize the proposed ZSRSSC method based on LPDCMENs in Algorithm 2. In Section IV, we evaluate the proposed ZSRSSC method.

Algorithm 2 Proposed LPDCMENs-Driven ZSRSSC Method

Input: The testing visual data $\{\mathbf{h}_i^U\}_{i=1}^{N^U}$, the testing semantic data $\mathbf{F}^U = \{\mathbf{f}_c^U \in \mathbb{R}^d\}_{c=1}^{C^U}$. The hyper-parameters of LPDCMENs including Θ^X and Θ^Y

Output: The labels of the testing visual data $\{l_i^U\}_{i=1}^{N^U}$

Calculate the the semantic templates $\{\mathbf{Y}_1^U, \mathbf{Y}_2^U, \ldots, \mathbf{Y}_{C^U}^U\}$ according to Eq. (8);

for i=1,2,..., N^U **do**

- Calculate the latent representation of the i-th visual image scene via $\mathbf{X}_i^U = \varphi(\mathbf{h}_i^U; \Theta^X)$;
- Calculate the label of the i-th visual image scene according to Eq. (9);

end for

IV. Experimental Results and Discussion

Section IV-A first introduces the collected RS data set and adopted evaluation metric, and the implementation details of the presented LPDCMEN method. On the collected data set, Section IV-B analyzes the sensitivity of critical parameters in our proposed LPDCMEN method. In addition, Section IV-C gives the ablation analysis of the recommended constraints, including the cross-modal matching constraint and the latent feature regularization constraint. Finally, Section IV-D compares our proposed LPDCMENs with the state-of-the-art methods.

A. Experimental Setup

In this section, we first introduce the evaluation data sets in Section IV-A1. We then give the implementation details of our proposed method in Section IV-A2.

1) Evaluation Data Set and Metric: As aforementioned, the RSSDIVCS includes two modalities (i.e., the instance-level visual images and class-level semantic representations). In the following, these two modalities are specifically introduced, respectively.

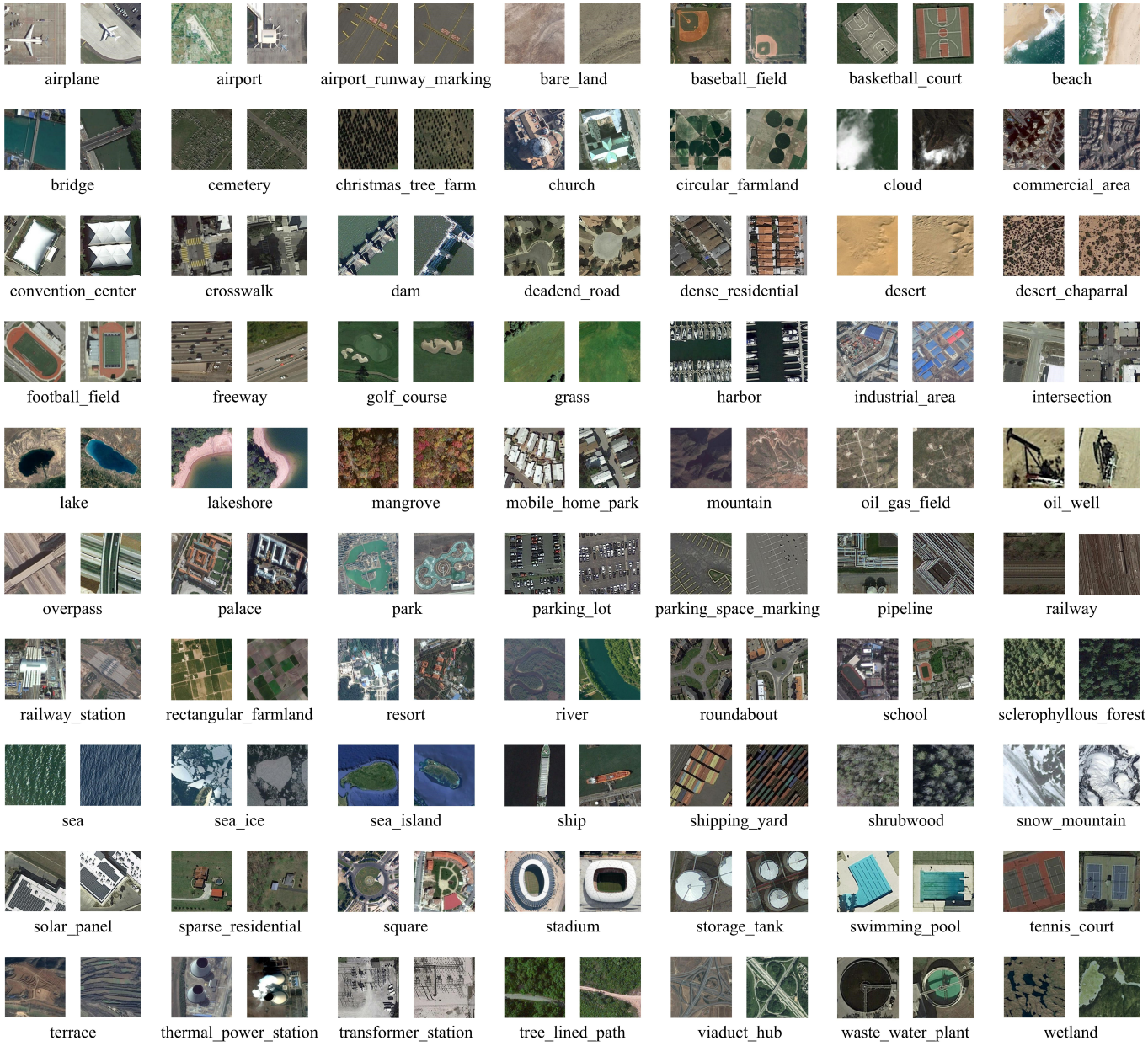


Fig. 3. Visual description of RS image scenes in the RSSDIVCS. More specifically, two RS image scenes, which are randomly selected from each category, are visually shown.

In the literature, lots of RS image scene data sets, such as UCM [12], AID [13], NWPU-RESISC45 [8], RSI-CB256 [14], and PatternNet [15], have been publicly released. However, each of these data sets has a relatively small number of scene categories, so any one of the existing data sets cannot fully verify the performance of ZSRSSC. Considering that UCM, AID, NWPU-RESISC45, RSI-CB256, and PatternNet own a complementary characteristic from the scene category perspective, and the image scenes from these five data sets have a similar image size, we collect a new one by integrating these four data sets. In the integration process, only one scene category is kept if multiple scene categories from these five data sets associate with a similar LULC type, and the rotation augmentation is adopted if one scene category has a relatively small number of image scenes. The new image scene data set is composed of 70 scene categories, and each category contains 800 image scenes with a size of 256×256. More specifically, the new image scene data set is visually shown in Fig. 3. The RS image scene category nomenclature and corresponding language-level description can refer to Table I.

Based on the integrated RS image scene data set, we construct two kinds of class-level semantic representations by two kinds of knowledge.

1) The first kind of semantic representations is built by general knowledge. More concretely, the word2vec model [48], which is trained on the Wikipedia corpus, is leveraged to map each category name to one different semantic vector with 300 dimensions. The semantic vector is taken as the first kind of semantic representation.
2) The second kind of semantic representation is built by the domain knowledge. As shown in Table I, multiple domain experts from the RS field depict each

TABLE I

RS IMAGE SCENE CATEGORY LIST AND THE CORRESPONDING SENTENCE DESCRIPTION OF EACH RS IMAGE SCENE CATEGORY

Category name	Sentence description
Airplane	A few big airplanes are taxiing on the airport runway.
Airport	Many small airplanes are parked on the airport tarmac and the airport also contains runways and buildings.
Airport_runway_marking	There are some yellow marking signs on the airport runway.
Bare_land	The area is full of bare land and contains very rare vegetation.
baseball_field	A baseball diamond with a fan shape is composed of sand and grass.
Basketball_court	A few basketball courts with a rectangular shape often occur simultaneously.
Beach	The white waves with foams splash the sand beach.
Bridge	A few wide roads cross a river or sea.
Cemetery	Many gravestones are scattered on the grass.
Christmas_tree_farm	The field is planted with rows of Christmas trees.
Church	A church with a enormous dome-shaped roof is often surrounded by trees or buildings.
Circular_farmland	Many pieces of circular farmland appear together.
Cloud	The land cover information is missing as this region is covered by massive white thin cloud.
Commercial_area	Many tall and large buildings are crowded together.
Convention_center	Several small buildings and many green trees are around a giant building with a special-shape roof.
Crosswalk	An intersection with two roads vertical to each other has zebra crossing signs.
Dam	A dam is built to block a river.
Deadend_road	A circular road has paths leading to several homes.
Dense_residential	It is a dense residential area with lots of houses arranged nearly.
Desert	The desert scene is full of natural sands.
Desert_chaparral	Some plants are sparsely scattered in the loess ground.
Football_field	The playground consists of a red track and a green football field.
Freeway	Straight freeways are with some scattered vehicles on them.
Golf_course	A part of a golf course is with green turfs and some bunkers.
Grass	A meadow of green grass.
Harbor	Many boats are docked in lines at the harbor and the water is with the deep blue color.
Industrial_area	Some large factory buildings are crowded together.
Intersection	Two vehicle roads cross each other at an intersection without zebra crossing signs.
Lake	An area of water with a clear lake boundary is surrounded by land.
Lakeshore	This is a lakeshore with lake and trees on the land.
Mangrove	This is a dense forest with mangrove trees.
Mobile_home_park	Many mobile homes are arranged neatly in the mobile home park and some roads go through this area.
Mountain	A mountain is with folds and veins.
Oil_gas_field	Many bright impervious areas are scattered on the oil gas field.
Oil_well	A view of the beam-pumping unit on the bare land.
Overpass	An overpass means that a road go across another one with many vehicles on the roads.
Palace	A palace with a regular layout is surrounded by some trees and buildings.
Park	A park means a region is with many green trees and ponds, and is surrounded by some buildings and roads.
Parking_lot	Lots of cars are parked in the parking lot.
Parking_space_marking	It is a nearly empty parking lot with some white marking lines on it.
Pipeline	A series of pipelines are arranged in a crisscross pattern.
Railway	Straight railways with some green plants on both sides of the railways.
Railway_station	Many buildings are in both sides of a railway station.
Rectangular_farmland	Many pieces of rectangular farmland appear simultaneously.
Resort	Some buildings and many green trees are in a resort.
River	A curved river crosses some green trees.
Roundabout	A ring road with some cars on the roads.
School	The school area is often with many trees, one playground and many other buildings.
Sclerophyllous_forest	This is a dense forest with broad leaf trees.
Sea	There are ripples on the blue sea.
Sea_ice	Massive sea ice covers the sea surface.
Sea_island	An island is surrounded by the sea water.
Ship	A few large ships lay on the sea surface.
Shipping_yard	Many containers with a rectangular shape is put on the shipping yard.
Shrubwood	Many shrubs are closely packed together.
Snow_mountain	The mountain is covered with snow and ice.
Solar_panel	Many solar panels with a rectangular shape is put on the roof.
Sparse_residential	The houses are sparely scattered and the houses is often surrounded by trees and roads.
Square	The square is a void area with some green plants and roads in this area.
Stadium	There is a green football field in the large stadium.
Storagetanks	One or several circular storagetanks on the ground.
Swimming_pool	A swimming pool surrounded by some trees in the yard.
Tennis_court	Some tennis courts with a rectangular shape are surrounded by some green plants.
Terrace	Terrace is one kind of farmland with the terrace parcel shape.
Thermal_power_station	One or several circular thermal power stations present in the industrial area.
Transformer_station	An area is full of transformers.
Tree_lined_path	A small road runs through the woods.
Viaduct_hub	Viaduct is a hub with many criss-crossing highways knotted here.
Waste_water_plant	One or several circular waste water plant present on the ground.
Wetland	Wetland means the region is mixed with water and vegetation.

TABLE II
CONFIGURATION OF V-CNNS

Layer	Configuration
Conv1	filter: $64 \times 11 \times 11 \times 3$, stride1: 4×4, pool: 3×3, stride2: 2×2
Conv2	filter: $256 \times 5 \times 5 \times 64$, stride1: 1×1, pool: 3×3, stride2: 2×2
Conv3	filter: $256 \times 3 \times 3 \times 256$, stride1: 1×1
Conv4	filter: $256 \times 3 \times 3 \times 256$, stride1: 1×1
Conv5	filter: $256 \times 3 \times 3 \times 256$, stride1: 1×1, pool: 3×3, stride2: 2×2
Full6	4096
Full7	4096
Full8	1000
Full9	m

RS scene category by one summarized sentence after checking over ten random RS image scenes from one given category. Furthermore, the bidirectional encoder representations from transformers (BERT) model [49] maps the high-level sentence description of each RS scene category to one different semantic representation with 1024 dimensions.

Considering the expert-crafted sentence has assimilated lots of knowledge from domain experts, this kind of semantic representation based on sentence description is taken as an example of the domain knowledge in this article. In the future, we will explore more kinds of domain knowledge, such as the expert-crafted attribute vectors or semantic representations from the domain knowledge graph.

Similar to [42], the overall accuracy (OA) is taken as the quantitative metric to evaluate the classification performance of ZSL methods over all unseen classes.

2) Implementation Details: In this implementation, V-CNNs of LPDCMENs are constructed by transferring the VGG-F net [50] pretrained on ImageNet based on the fact that the image scene in our collected RSSDIVCS resembles the natural image in ImageNet in terms of spectral range and spatial resolution. The specific configuration of the transferred V-CNNs is shown in Table II, and V-CNNs can process the input image with $224 \times 224 \times 3$. In Table II, "filter" specifies the number of filters, the size of the field, and the dimension of the input data and can be formulated as $num \times size \times size \times dim$. "stride1" indicates the sliding step in the convolution operation. "pool" stands for the downsampling factor. "stride2" stands for the sliding step in the local pooling operation.

By contrast to the rich samples in the visual space, each RS scene category has only one semantic representation. To prevent overfitting of S-NNs, the dimension of the original semantic representation is first reduced, and we adopt a shallow semantic embedding network with one fully connected layer, as depicted in Table III. To reduce the dimension of semantic representation, each semantic vector $\mathbf{f}_i$ is transformed to a kernelized semantic representation

$$\psi(\mathbf{f}_i) = \left[d(\mathbf{f}_i, \mathbf{f}_1^S), \ldots, d\left(\mathbf{f}_i, \mathbf{f}_{C^S}^S\right), d\left(\mathbf{f}_i, \mathbf{f}_1^U\right), \ldots, d\left(\mathbf{f}_i, \mathbf{f}_{C^U}^U\right)\right] \quad (10)$$

where $d(\mathbf{f}_i, \mathbf{f}_j) = \exp(-\hbar\|\mathbf{f}_i - \mathbf{f}_j\|^2)$. In our implementation, $\hbar$ is empirically set to 0.01 and 0.005 for the general knowledge and domain knowledge.

TABLE III
CONFIGURATION OF S-NNS

Layer	Configuration
Full	m

More specifically, we set the learning rate to 0.01, and the weight decay is set to 0.0005. To alleviate overfitting, we freeze the convolutional layers of V-CNNs and only update the fully connected layers of V-CNNs. In addition, S-NNs are fully updated in the learning process. It is worth mentioning that the whole LPDCMENs can be optimized in an end-to-end manner. How to determine the parameters in the objective function will be discussed in Section IV-B.

B. Sensitivity Analysis of Critical Parameters

To effectively analyze the sensitivity of critical parameters of our proposed method, the seen/unseen ratio is set to 40/30. Given this seen/unseen ratio, we calculate the average and standard deviation of the quantitative classification results of unseen classes over ten random seen/unseen splits.

Overall, our proposed LPDCMENs can converge very fast. More specifically, as shown in Fig. 4(a) and (b), our proposed method with the general knowledge can achieve the best performance when the number of iterations, iterMax, equals 2. Furthermore, as illustrated in Fig. 4(c) and (d), our proposed method with the domain knowledge can achieve the best performance when the number of iterations, iterMax, equals 3.

As shown in Fig. 4(a), when the general knowledge is adopted, our proposed method can achieve the best performance when $m = 150$. Fixing $m = 150$, as shown in Fig. 4(b), $\delta = 4$ makes our proposed method with the general knowledge achieve the best performance. Similar to this analysis process, as shown in Fig. 4(c) and (d), our proposed method with the domain knowledge also can achieve the best performance when $m = 150$ and $\delta = 4$.

With α, β, and γ empirically set to 1, 100, and 0.1, we further analyze the sensitivity of the constraint coefficient η in the objective function. As illustrated in Fig. 5(a), when the general knowledge is adopted, $\eta = 10^{-4}$ can make our proposed method achieve the best performance. When the domain knowledge is adopted, Fig. 5(b) shows that our proposed method achieves the best performance when $\eta = 10^{-3}$. To pursue a generalization ability, the best setting of η under the seen/unseen ratio of 40/30 is reused in all seen/unseen ratios. Without any doubt, the performance of our proposed method can be further improved if we further tune α, β, and γ. We do not do that here because the repetitive training deep of networks under so many different parameter settings would be incredibly time-consuming. As a compromise solution, we quantitatively analyze the contribution of parameters (i.e., α, β, γ, and η) in the objective function via the following ablation analysis.

C. Ablation Analysis of Critical Parameters

To effectively conduct the ablation analysis, the seen/unseen ratio is set to 40/30 in the following experiments. To verify the superiority of the recommended cross-modal matching

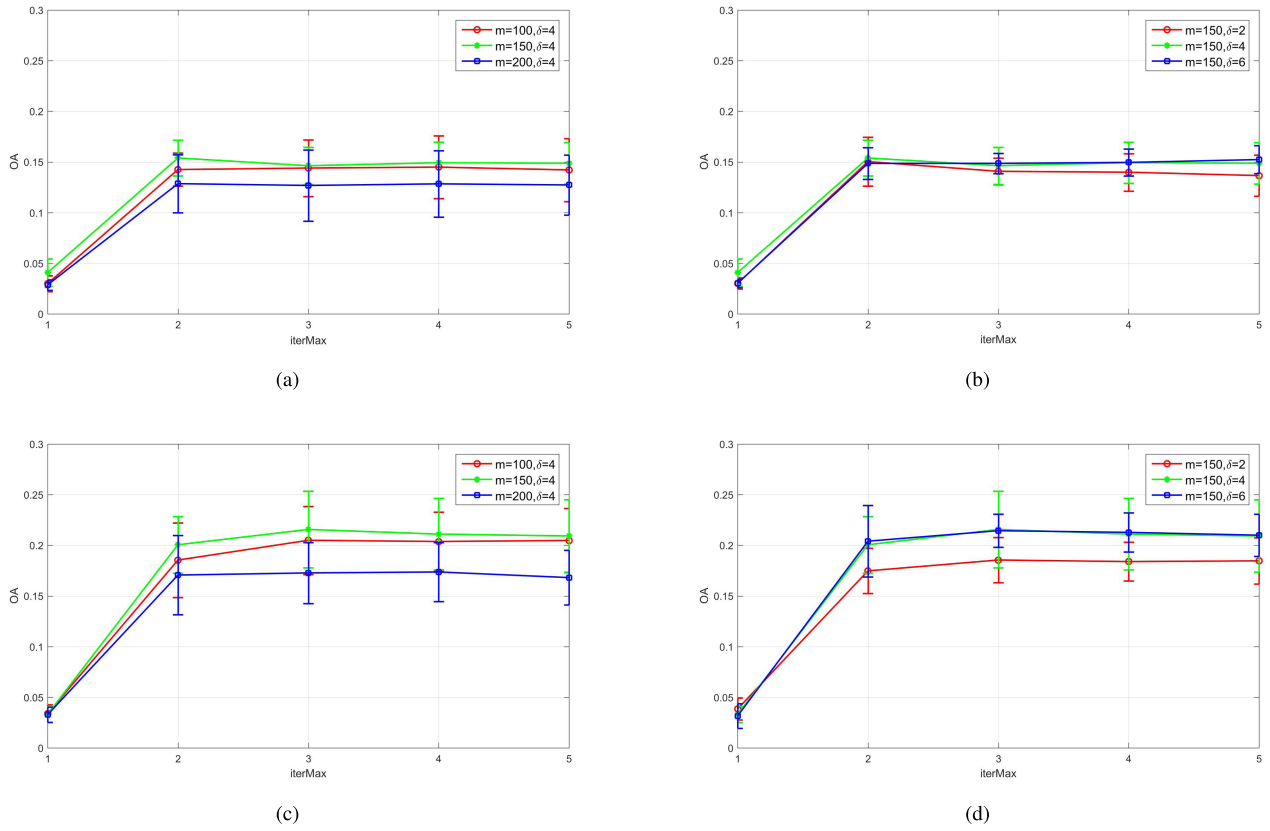


Fig. 4. Classification performance of our proposed method by varying the latent vector dimension m and weight constant δ. (a) Results of our proposed method using general knowledge under different m's. (b) Results of our proposed method using general knowledge under different δ's. (c) Results of our proposed method using the domain knowledge under different m's. (d) Results of our proposed method using the domain knowledge under different δ's.

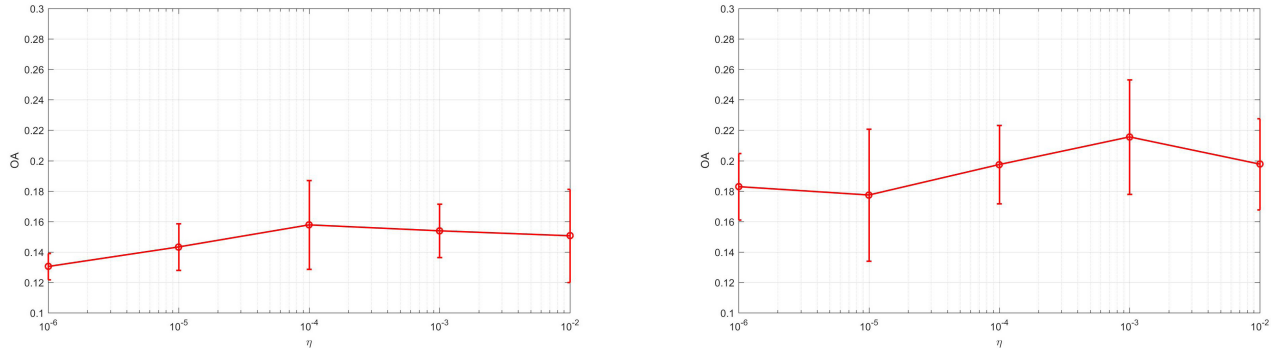


Fig. 5. Classification performance of our proposed method by varying the constraint coefficient η's. (a) Results of our proposed method using general knowledge under different η's. (b) Results of our proposed method using the domain knowledge under different η's.

constraint, we summarize the results of the ablation experiments about the cross-modal matching constraint in Table IV. As shown in Table IV, only using CMM-CD outperforms CMM-ED, and the fusion of two metrics performs better than every single one under two kinds of semantic representations (i.e., the general knowledge and domain knowledge).

Furthermore, Table V records the results of the ablation experiments about the latent feature regularization constraint. As shown in Table V, only using DBC or VMC can get a very close performance, but the fusion of them obviously outperform every single one, which reflects the complementary property between DBC and VMC.

D. Comparison With the State-of-the-Art Approaches

To give a full analysis of ZSL methods, we report the quantitative results under different seen/unseen ratios (e.g., 40/30, 50/20, and 60/10) in Table VI. More specifically, in each given seen/unseen ratio, we evaluate each method over ten random seen/unseen splits. As aforementioned, two kinds of knowledge, including general knowledge and domain knowledge, are evaluated, respectively.

To show the superiority of our proposed method, we consider the following baselines. As an extension of ridge regression [54], semantic autoencoder (SAE) [51] has been proposed

TABLE IV
ABLATION ANALYSIS OF THE CROSS-MODAL MATCHING CONSTRAINT

The cross-modal matching constraint	Parameters in the objective function	General knowledge	Domain knowledge
CMM-CD	$\alpha \neq 0; \beta = 0$	0.144 ± 0.017	0.181 ± 0.019
CMM-ED	$\alpha = 0; \beta \neq 0$	0.129 ± 0.040	0.134 ± 0.041
CMM-CD + CMM-ED	$\alpha \neq 0; \beta \neq 0$	0.158 ± 0.029	0.216 ± 0.038

TABLE V
ABLATION ANALYSIS OF THE LATENT FEATURE REGULARIZATION CONSTRAINT

The latent feature regularization constraint	Parameters in the objective function	General knowledge	Domain knowledge
DBC	$\gamma \neq 0; \eta = 0$	0.121 ± 0.025	0.175 ± 0.035
VMC	$\gamma = 0; \eta \neq 0$	0.131 ± 0.004	0.173 ± 0.039
DBC + VMC	$\gamma \neq 0; \eta \neq 0$	0.158 ± 0.029	0.216 ± 0.038

TABLE VI
QUANTITATIVE COMPARISON RESULTS (OA) OF ZSL METHODS UNDER DIFFERENT SEEN/UNSEEN RATIOS

Knowledge type	General knowledge			Domain knowledge		
Seen/unseen ratio	40/30	50/20	60/10	40/30	50/20	60/10
SAE(V → S) [51]	0.096 ± 0.014	0.137 ± 0.017	0.235 ± 0.042	0.088 ± 0.013	0.124 ± 0.019	0.220 ± 0.017
SAE(S → V) [51]	0.052 ± 0.014	0.095 ± 0.016	0.167 ± 0.041	0.050 ± 0.013	0.080 ± 0.023	0.168 ± 0.044
DMaP [52]	0.104 ± 0.009	0.167 ± 0.022	0.260 ± 0.036	0.100 ± 0.008	0.156 ± 0.019	0.164 ± 0.019
SPLE [42]	0.098 ± 0.014	0.132 ± 0.019	0.201 ± 0.037	0.083 ± 0.020	0.132 ± 0.026	0.190 ± 0.038
CIZSL [53]	0.060 ± 0.012	0.106 ± 0.037	0.206 ± 0.004	0.062 ± 0.021	0.103 ± 0.019	0.204 ± 0.041
ZSRSSC-GP [21]	0.047 ± 0.003	0.083 ± 0.007	0.148 ± 0.004	0.046 ± 0.004	0.074 ± 0.003	0.141 ± 0.016
ZSRSSC-SE [23]	0.121 ± 0.077	0.152 ± 0.010	0.267 ± 0.053	0.131 ± 0.030	0.183 ± 0.013	0.293 ± 0.038
Our proposed LPDCMENs	**0.158 ± 0.029**	**0.197 ± 0.036**	**0.342 ± 0.090**	**0.216 ± 0.038**	**0.249 ± 0.037**	**0.438 ± 0.073**

to address zero-shot classification of natural imagery. Here, SAE (V → S) is extended to address zero-shot scene classification of RS imagery in the semantic space, and SAE (S → V) denotes zero-shot scene classification of RS imagery in the visual space. In the computer vision domain, dual visual-semantic mapping path (DMaP) [52] is first proposed to address zero-shot classification. Here, we also evaluate it on our collected RS scene data set. Different from the aforementioned feature mapping methods (i.e., from one given space to another given space), semantics-preserving locality embedding (SPLE) [42] is proposed to address zero-shot classification in the latent space. From the generative generation perspective, the recently proposed creativity inspired ZSL (CIZSL) [53] is also taken as one of the baselines. In addition to these zero-shot classification methods in the computer vision field, we also consider two recently proposed ZSRSSC methods, including the ZSRSSC method via graph propagation (ZSRSSC-GP) [21] and the ZSRSSC algorithm via semisupervised Sammon embedding (ZSRSSC-SE) [23].

As shown in Table VI, our proposed ZSRSSC method based on LPDCMENs can obviously outperform the state-of-the-art methods in terms of under different seen/unseen ratios and with different knowledge types. As our proposed LPDCMENs follow the classic characteristic of deep learning that often depends on lots of labeled data, our proposed LPDCMENs can achieve a larger improvement magnitude compared with the baselines along with the increment of seen/unseen ratios. As the dimension of domain knowledge is larger than the general knowledge, the overfitting phenomena may make many existing ZSL methods with the domain knowledge perform worse than the general knowledge. However, our proposed LPDCMENs can avoid this problem and make full use of the advanced domain knowledge. As depicted in Table VI, our proposed LPDCMENs with the domain knowledge can outperform the general knowledge, remarkably.

V. CONCLUSION

Driven by more and more practical demands of ZSRSSC and the fact that ZSRSSC requires unified semantic representations (i.e., the prior knowledge about the seen and unseen categories), this article proposes a novel ZSRSSC approach based on LPDCMENs where LPDCMENs aim to address the problem of class structure inconsistency between two hybrid spaces by matching the visual and semantic information in the latent space. In addition, a set of explainable constraints is exploited to train LPDCMENs in an end-to-end manner. Extensive experiments on the new collected RSSDIVCS show that the proposed LPDCMENs can obviously outperform the state-of-the-art methods under varying cases.

There exist some pure RS image scene data sets that are publicly released. However, to the best of our knowledge, there exists only very few RS scene data sets (which have small or moderate size) that contain both visual samples and semantic representations. With this consideration, we collect a new large-scale RSSDIVCS where the general and domain knowledge are exploited to construct the class-level semantic representations. The collected RSSDIVCS will be made publicly available along with this article. Apparently, the released RSSDIVCS not only benefits the advanced progress of ZSRSSC but also promotes more knowledge-driven RS image scene understanding tasks.

REFERENCES

[1] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

[2] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing big data: A survey," *Inf. Fusion*, vol. 67, pp. 94–115, Mar. 2021.

[3] P. Zhong and R. Wang, "Jointly learning the hybrid CRF and MLR model for simultaneous denoising and classification of hyperspectral imagery," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 7, pp. 1319–1334, Jul. 2014.

[4] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," *Proc. IEEE*, vol. 104, no. 11, pp. 2207–2219, Nov. 2016.

[5] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, Aug. 2016.

[6] Y. Li, W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan, "Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning," *Remote Sens. Environ.*, vol. 250, Dec. 2020, Art. no. 112045.

[7] X. Han, X. Huang, J. Li, Y. Li, M. Y. Yang, and J. Gong, "The edge-preservation multi-classifier relearning framework for the classification of high-resolution remotely sensed imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 57–73, Apr. 2018.

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

[9] Y. Li, Y. Zhang, and Z. Zhu, "Error-tolerant deep learning for remote sensing image scene classification," *IEEE Trans. Cybern.*, early access, May 14, 2020, doi: 10.1109/TCYB.2020.2989241.

[10] X. Zhang, S. Du, and Q. Wang, "Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping," *Remote Sens. Environ.*, vol. 212, pp. 231–248, Jun. 2018.

[11] Y. Yuan, L. Mou, and X. Lu, "Scene recognition by manifold regularized deep learning architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2222–2233, Oct. 2015.

[12] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. (GIS)*, 2010, pp. 270–279.

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

[14] H. Li *et al.*, "RSI-CB: A large scale remote sensing image classification benchmark via crowdsource data," 2017, *arXiv:1705.10450*. [Online]. Available: http://arxiv.org/abs/1705.10450

[15] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 197–209, Nov. 2018.

[16] G. Cheng, C. Ma, P. Zhou, X. Yao, and J. Han, "Scene classification of high resolution remote sensing images using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 767–770.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] M. Zhai, H. Liu, and F. Sun, "Lifelong learning for scene recognition in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1472–1476, Sep. 2019.

[19] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 935–943.

[20] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5542–5551.

[21] A. Li, Z. Lu, L. Wang, T. Xiang, and J.-R. Wen, "Zero-shot scene classification for high spatial resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 4157–4167, Jul. 2017.

[22] Y. Li, Y. Tan, J. Deng, Q. Wen, and J. Tian, "Cauchy graph embedding optimization for built-up areas detection from high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2078–2096, May 2015.

[23] J. Quan, C. Wu, H. Wang, and Z. Wang, "Structural alignment based zero-shot classification for remote sensing scenes," in *Proc. IEEE Int. Conf. Electron. Commun. Eng. (ICECE)*, Dec. 2018, pp. 17–21.

[24] G. Sumbul, R. G. Cinbis, and S. Aksoy, "Fine-grained object recognition and zero-shot learning in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 770–779, Feb. 2018.

[25] P. Aravena Pelizari, K. Spröhnle, C. Geiß, E. Schoepfer, S. Plank, and H. Taubenböck, "Multi-sensor feature fusion for very high spatial resolution built-up area extraction in temporary settlements," *Remote Sens. Environ.*, vol. 209, pp. 793–807, May 2018.

[26] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, "Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 182–196, Dec. 2018.

[27] Y. Tan, S. Xiong, and Y. Li, "Automatic extraction of built-up areas from panchromatic and multispectral remote sensing images using double-stream deep convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 3988–4004, Nov. 2018.

[28] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, "Large-scale remote sensing image retrieval by deep hashing neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 950–965, Feb. 2018.

[29] Y. Li, Y. Zhang, X. Huang, and J. Ma, "Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6521–6536, Nov. 2018.

[30] S. Newsam, L. Wang, S. Bhagavathy, and B. S. Manjunath, "Using texture to analyze and manage large collections of remote sensed image and video data," *Appl. Opt.*, vol. 43, no. 2, pp. 210–217, 2004.

[31] 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 Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4238–4249, Aug. 2015.

[32] Y. Li, Y. Tan, Y. Li, S. Qi, and J. Tian, "Built-up area detection from satellite images using multikernel learning, multifield integrating, and multihypothesis voting," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 6, pp. 1190–1194, Jun. 2015.

[33] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1349–1362, Mar. 2016.

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

[35] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017.

[36] Y. Wang *et al.*, "Learning a discriminative distance metric with label consistency for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4427–4440, Aug. 2017.

[37] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, "Learning like a child: Fast novel visual concept learning from sentence descriptions of images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2533–2541.

[38] I. Kadar and O. Ben-Shahar, "SceneNet: A perceptual ontology for scene understanding," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 385–400.

[39] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2021–2030.

[40] Z. Zhang and V. Saligrama, "Zero-shot learning via joint latent similarity embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 6034–6042.

[41] H. Jiang, R. Wang, S. Shan, Y. Yang, and X. Chen, "Learning discriminative latent attributes for zero-shot classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4223–4232.

[42] S.-Y. Tao, Y.-R. Yeh, and Y.-C.-F. Wang, "Semantics-preserving locality embedding for zero-shot learning," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 3.1–3.12, Art. no. 3.

[43] Y. Yu, Z. Ji, J. Guo, and Z. Zhang, "Zero-shot learning via latent space encoding," *IEEE Trans. Cybern.*, vol. 49, no. 10, pp. 3755–3766, Oct. 2019.

[44] Z. Ji *et al.*, "Deep ranking for image zero-shot multi-label classification," *IEEE Trans. Image Process.*, vol. 29, no. 7, pp. 6549–6560, Jul. 2020.

[45] Z. Ji, Y. Sun, Y. Yu, Y. Pang, and J. Han, "Attribute-guided network for cross-modal zero-shot hashing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 321–330, Jan. 2020.

[46] T. Toizumi, K. Sagi, and Y. Senda, "Automatic association between SAR and optical images based on zero-shot learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 17–20.

[47] Q. Song, H. Chen, F. Xu, and T. J. Cui, "EM simulation-aided zero-shot learning for SAR automatic target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 1092–1096, Jun. 2020.

[48] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[49] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[50] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: An evaluation of recent feature encoding methods," in *Proc. Brit. Mach. Vis. Conf.*, 2011, vol. 2, no. 4, p. 8.

[51] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3174–3183.

[52] Y. Li, D. Wang, H. Hu, Y. Lin, and Y. Zhuang, "Zero-shot recognition using dual visual-semantic mapping paths," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3279–3287.

[53] M. Elhoseiny and M. Elfeki, "Creativity inspired zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5784–5793.

[54] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970.

Yansheng Li (Member, IEEE) received the B.S. degree in information and computing science from Shandong University, Weihai, China, in 2010, and the Ph.D. degree in pattern recognition and intelligent system from the Huazhong University of Science and Technology, Wuhan, China, in 2015.

He is also an Associate Professor with the School of Remote Sensing and Information Engineering, Wuhan University (WHU), Wuhan. He was an Assistant Professor with WHU in 2015, where he became an Associate Research Fellow in 2017. From 2017 to 2018, he was a Visiting Assistant Professor with the Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. He has authored more than 40 peer-reviewed articles (SCI articles) in international journals from multiple domains, such as remote sensing, and computer vision. His research interests mainly lay in the field of computer vision, machine learning, and their applications in remote sensing big data analysis.

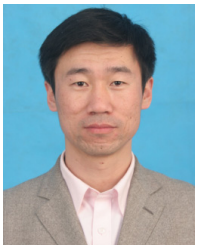
Dr. Li has been frequently serving as a Referee for over 20 international journals, such as the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, *ISPRS Journal of Photogrammetry and Remote Sensing*, and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING. He is also a Communication Evaluation Expert for the National Natural Science Foundation of China.

Zhihui Zhu (Member, IEEE) received the B.S. degree in communications engineering from the Zhejiang University of Technology, Hangzhou, China, in 2012, and the Ph.D. degree in electrical engineering from the Colorado School of Mines, Golden, CO, USA, in 2017.

He was a Post-Doctoral Fellow with the Mathematical Institute for Data Science and the Center for Imaging Science, Johns Hopkins University, Baltimore, MD, USA, from 2018 to 2019. He is also an Assistant Professor with the Department of Electrical and Computer Engineering, University of Denver, Denver, CO, USA. His research interests include exploiting inherent structures and applying optimization methods with guaranteed performance for signal processing and machine learning.

Jin-Gang Yu received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 2005, and the M.S. and Ph.D. degrees from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2007 and 2014, respectively.

He was a Post-Doctoral Research Associate with the Department of Computer Science and Technology, University of Nebraska–Lincoln, Lincoln, NE, USA, from 2014 to 2016. He spent three years as a Research and Development Engineer with ZTE Corporation, Shenzhen, China, and Nortel Networks Corporation, Guangzhou, China, before starting the Ph.D. Program at HUST. He joined the South China University of Technology, Guangzhou, China, in 2016, where he is also an Associate Professor. His research interests include computer vision, pattern recognition, and machine learning.



Yongjun Zhang (Member, IEEE) received the B.S. degree in geodesy, the M.S. degree in geodesy and surveying engineering, and the Ph.D. degree in geodesy and photography from Wuhan University, Wuhan, China, in 1997, 2000, and 2002, respectively.

From 2014 to 2015, he was a Senior Visiting Fellow with the Department of Geomatics Engineering, University of Calgary, Calgary, AB, Canada. From 2015 to 2018, he was a Senior Scientist with the Environmental Systems Research Institute, Inc. (Esri), Redlands, CA, USA. He is also the Dean of the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 150 research articles and one book. He holds 23 Chinese patents and 26 copyright registered computer software. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource data sets, artificial intelligence-driven remote sensing image interpretation, integration of LiDAR point clouds and images, and 3-D city reconstruction.

Prof. Zhang has been a Key Member of the ISPRS Workgroup II/I since 2020. He is also the PI Winner of the Second-Class National Science and Technology Progress Award in 2017 and the Outstanding-Class Science and Technology Progress Award in Surveying and Mapping from the Chinese Society of Surveying, Mapping and Geoinformation, China, in 2015. In recent years, he has also served as the session chair of above 20 international workshops or conferences. He has been frequently serving as a referee for over 20 international journals.