Review of Zero-Shot Remote Sensing Image Scene Classification

Xiaomeng Tan, Bobo Xi, *Member, IEEE*, Jiaojiao Li, *Senior Member, IEEE*, Tie Zheng, Yunsong Li, *Member, IEEE*, Changbin Xue, and Jocelyn Chanussot, *Fellow, IEEE*

***Abstract*—In recent years, remote sensing (RS) image scene classification methods have experienced notable development due to the powerful feature extraction ability of deep learning. However, current methods for RS image scene classification (RSSC) tasks struggle with handling unseen scene categories because of their reliance on large amounts of high-quality labeled data. To address this issue, zero-shot learning has been introduced into RSSC tasks, leading to the emergence of zero-shot remote sensing image scene classification (ZSRSSC) approaches. These methods refer to inference and recognition of unseen category images by comprehending seen category image features and combining category semantic information in a multimodal learning paradigm. The inference capability possessed by ZSRSSC methods can significantly improve the accuracy and applicability of RSSC in dealing with large amounts of data in the RS domain. However, this newly emerging field has produced numerous results without a comprehensive survey to organize research progress systematically. Therefore, this article systematically reviews the research progress made on these approaches, summarizes commonly used datasets and experimental setups, and compares the results of different methods. Specifically, the definition of zero-shot learning (ZSL) is reviewed first, and the mainstream development of ZSL classification methods is briefly introduced. Then, the definition of ZSRSSC is outlined, followed by an introduction to the prevalent ZSRSSC approaches and their categorization. Finally, the critical issues in these methods are analyzed, and their future development directions are discussed.**

***Index Terms*—Multimodal learning, remote sensing (RS) image scene classification, zero-shot learning.**

Manuscript received 30 January 2024; revised 22 April 2024; accepted 2 June 2024. Date of publication 7 June 2024; date of current version 18 June 2024. This work was supported in part by the Youth Open Project of National Space Science Data Center under Grant NSSDC2302004, in part by the General Financial Grant from the China Postdoctoral Science Foundation under Grant 2023M732742, in part by the Young Talent Fund of University Association for Science and Technology in Shaanxi, China, under Grant 20210104, in part by the National Nature Science Foundation of China under Grant 62371359, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515110504, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2024JC-YBQN-0641, in part by the Postdoctoral Science Foundation of Shaanxi Province under Grant 2023BSHYDZZ97, in part by the Fundamental Research Funds for the Central Universities under Grant XJSJ23086, and in part by the Youth Innovation Team of Shaanxi Universities. *(Corresponding authors: Bobo Xi; Changbin Xue.)*

Xiaomeng Tan is with the Key Laboratory of Electronics and Information Technology for Space Systems, National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China, also with the National Space Science Data Center, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: tanxiaomeng22@mails.ucas.ac.cn).

Bobo Xi is with the State Key Laboratory of Integrated Service Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China, and also with the National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China (e-mail: xibobo1301@foxmail.com).

Jiaojiao Li and Yunsong Li are with the State Key Laboratory of Integrated Service Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: jjli@xidian.edu.cn; ysli@mail.xidian.edu.cn).

Tie Zheng and Changbin Xue are with the Key Laboratory of Electronics and Information Technology for Space Systems, National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhengtie@nssc.ac.cn; xuechangbin@nssc.ac.cn).

Jocelyn Chanussot is with Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France (e-mail: jocelyn@hi.is).

Related code and papers are available at https://github.com/XM-Tan/Review-for-ZSRSSC.

Digital Object Identifier 10.1109/JSTARS.2024.3410995

I. Introduction

REMOTE sensing (RS) image is obtained using remotely sensed technology, which is generally captured by the RS platforms, such as airplanes and satellites, under long-distance conditions. The research of RS images has also consistently been a hot topic in academia and industry. These images contain detailed and accurate information, which may be very useful for scene classification tasks if we can identify and retrieve their inherent scene information. In the interpretation and analysis of RS images [1], scene image classification refers to the division of scene images into human-defined, meaningful, and interpretable categories, such as airports, grasslands, and parks, to name a few. This is an essential and challenging task in practical applications [2], [3], [4], [5] of the RS field. However, with the rapid blossoming of RS technology, an ever-growing volume of data brings the bulk of costs in manpower and material resources. This poses new challenges in reducing the burden of manual labor and improving efficiency. For this reason, more intelligent and automatic approaches for diverse RS applications must be developed. As a result, artificial intelligence is being increasingly applied to the RS domain [6].

In the procedure of empowering the RS field with artificial intelligence technology, there are often challenges related to fragmented and diversified demands. Typically, deep learning-based methods, such as convolutional neural networks (CNNs), have a powerful feature extraction capability and have achieved remarkable performance in RS image scene classification (RSSC) tasks. However, these methods usually require a lot of labeled data, which is relatively scarce and expensive to obtain in the RS domain. Meanwhile, new/unseen scene categories not in the training datasets may appear at any time in practical applications, making it challenging to classify them accurately. To improve the recognition accuracy of new/unseen category images, the traditional strategy is to retrain the classification model using

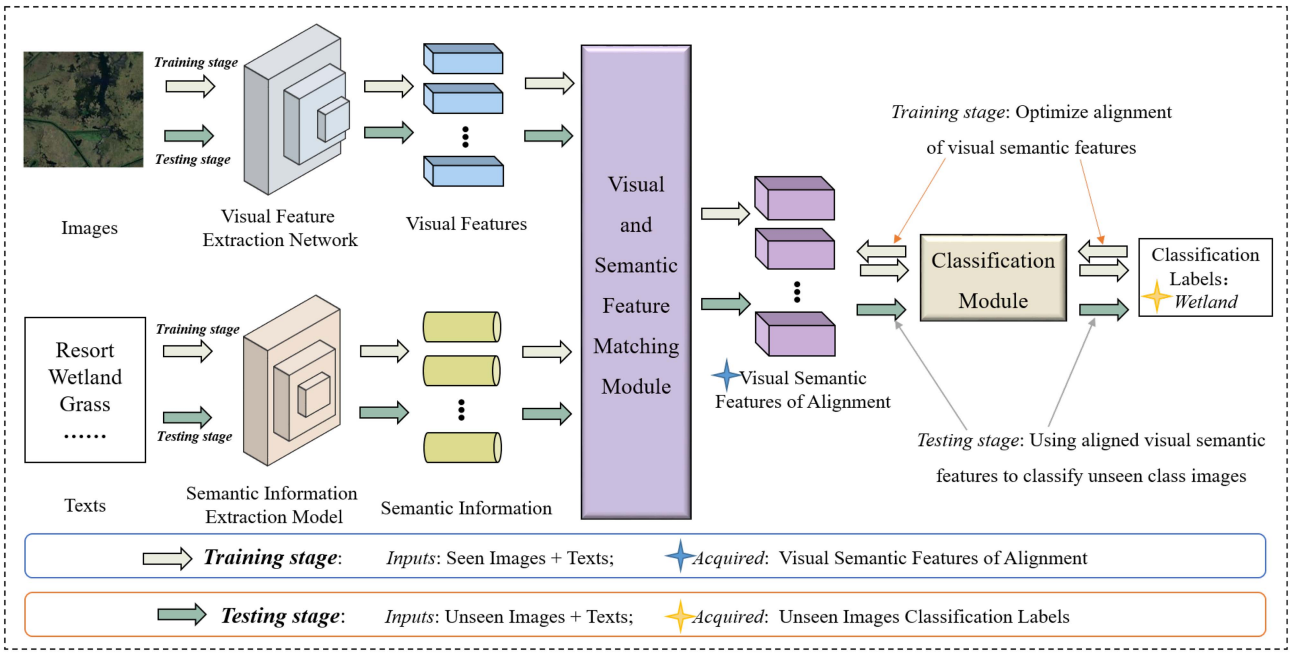

Fig. 1. Schematic diagram of ZSRSSC paradigm.

a dataset that includes the newly appeared and unseen categories, leading to trouble with computational resources, human resources, and time costs.

To break the above limitation, some scholars have explored few-shot learning [7], [8], [9], and others have introduced the idea of zero-shot learning (ZSL) to tackle the RSSC tasks involving new/unseen scene categories, defined as the zero-shot remote sensing image scene classification (ZSRSSC) paradigm, as portrayed in Fig. 1. This paradigm encompasses acquiring the aligned process between seen class images and their associated semantic information and optimizing the alignment of visual-semantic features in the training stage. Subsequently, it enables the classification of unseen categories by employing visual-semantic alignment features during the testing stage.

In practice, the powerful ability to infer unseen categories makes the ZSRSSC method a promising application. For instance, in environmental monitoring, natural or human-induced changes may lead to the emergence of new land cover types. The ZSRSSC paradigm can adapt quickly and accurately classify these new categories. Similarly, rapid response and decision-making are crucial in disaster management, and this paradigm can help identify previously unseen features or objects essential for effective disaster response. In addition, considering that the inference ability of the ZSRSSC paradigm has promising development prospects in the RS field, it has also been extended to numerous tasks such as polarized-SAR image classification [10], side-scan sonar image recognition [11], hyperspectral image denoising [12], RS image retrieval [13], [14]. Therefore, it is significant to study the ZSRSSC model and mine more valuable information from massive test data.

The rest of this article is organized as follows. In Section II, we briefly introduce the mainstream and categories of the ZSL methods. Section III introduces the key content of this article as the development and categories of the ZSRSSC methods. Section IV describes the datasets and experimental setup. Then, we analyze and compare the experimental results of most ZSRSSC methods introduced in Section III, and discuss the existing issues and future development trends of the methods. Finally, Section V concludes the article.

II. Zero-Shot Learning

A. Introduction to ZSL

ZSL aims to recognize new/unseen category images by learning seen class images with the assistance of textual description information. The problem clarification of ZSL is as follows: given a training set $\boldsymbol{D}^S = \{(\boldsymbol{x}_i^S, \boldsymbol{y}_i^S)|_{i=1}^{N^S}\}$ consisting of N^S training samples, where $\boldsymbol{y}_i^S \in \boldsymbol{Y}^{tr}$ and $\boldsymbol{Y}^{tr}$ are training classes. In the training stage, zero-shot image classification aims to learn the mapping relationship $f : \boldsymbol{X} \longrightarrow \boldsymbol{Y}$ [15] from image samples $\boldsymbol{X}$ to labels $\boldsymbol{Y}$ by minimizing the regularized empirical risk, that is

$$\frac{1}{N^S}\sum_{i=1}^{N^S} L(\boldsymbol{y}_i^S, f(\boldsymbol{x}_i^S; \boldsymbol{W})) + \Omega(\boldsymbol{W}) \tag{1}$$

where $L(\cdot)$ is the loss function and $\Omega(\cdot)$ represents the regularization term. The mapping relationship $f : \boldsymbol{X} \longrightarrow \boldsymbol{Y}$ from image sample $\boldsymbol{X}$ to label $\boldsymbol{Y}$ is defined as

$$f(\boldsymbol{x}; \boldsymbol{W}) = \arg\max_{\boldsymbol{y} \in \boldsymbol{Y}} F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{W}) \tag{2}$$

During the testing phase, the zero-shot image classification sets the test images as unseen class image samples, denoted as $\boldsymbol{y}^{ts} \subset \boldsymbol{Y}$.

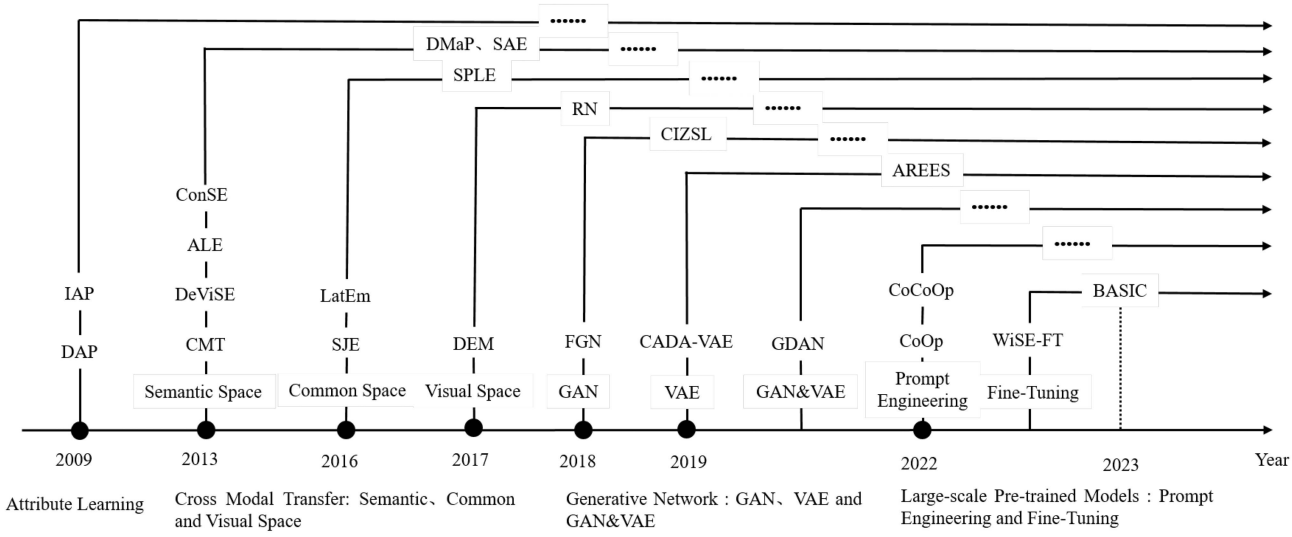

Fig. 2. Brief development diagram of the ZSL methods.

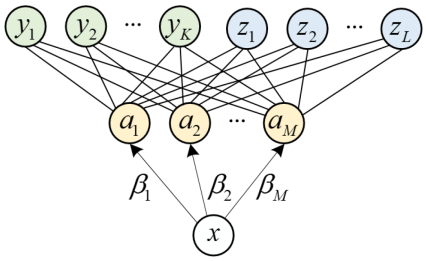

Fig. 3. Schematic diagram of the DAP model.

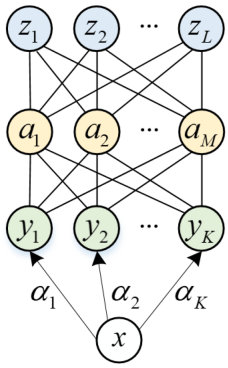

Fig. 4. Schematic diagram of the IAP model.

B. ZSL Methods

This section will briefly introduce the current research progress of ZSL methods according to the development trend of ZSL, as depicted in Fig. 2. There are four parts to provide an overview of ZSL methods, namely: early exploration (ZSL based on attribute learning), cross-modal transfer (CMT) (ZSL based on embedding models), ZSL based on generative networks, and ZSL based on pretrained models.

1) Early Exploration: ZSL Based on Attribute Learning: The pioneering work in ZSL was proposed by Lampert et al. [16] through attribute-based learning, which is a method of recognizing unseen category images by learning attribute transfer between categories. Two classical models are known as direct attribute prediction (DAP) and indirect attribute prediction (IAP). In the DAP model, the attribute layer is directly employed as an intermediate layer to decouple the relationship between images and class labels, as displayed in Fig. 3. The IAP model utilizes the attribute layer as an intermediate layer to learn the mapping between seen class labels and unseen category labels, as shown in Fig. 4. This demonstrates that the difference between the two methods lies in the position of the attribute layer. Namely, the DAP sets the attribute layer between images and class labels, while the IAP fits it between seen class labels and unseen category labels.

2) CMT: ZSL Based on Embedding Models: The idea of CMT was first introduced to ZSL by Socher et al. [17]. This strategy transforms the ZSL problem into a subspace embedding problem. The core idea is to simultaneously map images and class labels into a subspace and then classify them based on a distance metric. According to different embedding subspaces, current approaches based on embedding models can be divided into three sorts: semantic space, common space, and visual space, as depicted in Fig. 5.

a) Embedding methods based on semantic space: With the introduction of CMT ideas, embedding methods based on semantic space were initially developed. The first CMT model was proposed by [17], which is based on unsupervised deep learning for word and image representations. It achieves interclass knowledge transfer and uses outlier detection in the semantic space, implementing image classification with two separate recognition models. Subsequently, Frome et al. [18] designed the deep visual-semantic embedding (DeViSE) model, which leverages labeled image tags and unlabeled textual semantic information for training and image recognition in semantic space. Furthermore, some scholars explored new forms of mapping images to

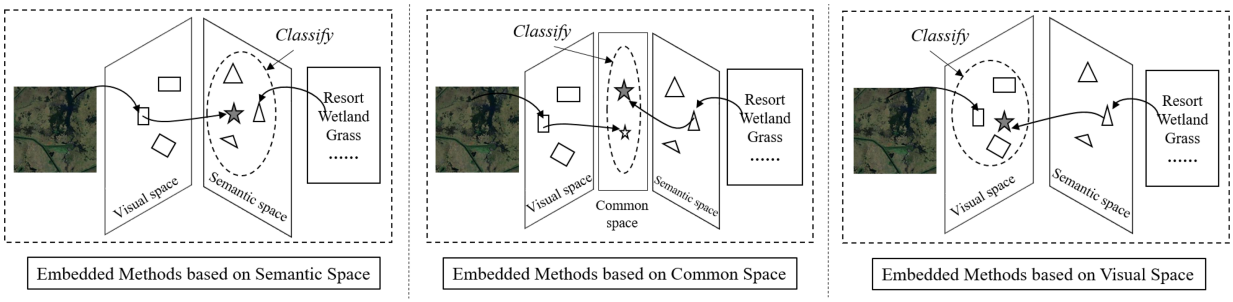

Fig. 5. Schematic diagram of the ZSL methods based on embedding models.

the semantic space. For instance, Norouzi et al. [19] proposed the convex combination of semantic embedding (ConSE) model, which maps images to the semantic space through the convex combination of embedded class labels.

On the other side, Akata et al. [20] reconsidered the importance of the attribute space and constructed the attribute label embedding (ALE) model, which treats attribute-based image classification as a label embedding problem. The attribute space is selected as a semantic space and then embeds each class within the space. The issue of domain shift caused by significant class bias between training and testing sets has become increasingly prominent with the development of cross-modal methods. To address this problem, Kodirov et al. [21] designed the semantic autoencoders (SAE) method using an encoder–decoder paradigm, which projects visual features into the semantic space by imposing additional constraints on the decoder. In addition, Li et al. [22] investigated which manifold structures in the embedding space are more advantageous for recognizing unseen categories. Then, they devise the dual visual-semantic mapping paths (DMaP) network, which treats zero-shot recognition as a joint optimization problem. It infers the underlying semantic manifold of image feature space using prior semantic knowledge and then generates an optimized semantic space to enhance the transfer ability of visual-semantic mapping to unseen categories.

b) Embedding methods based on common space: The promising deep learning network provides remarkable visual features of images and excellent class embedding of corresponding labels, and some scholars have started to focus on learning a discriminative compatibility function between the structured visual features and class embedding. These methods simultaneously map images and class labels into a latent common space, denoted as embedding methods based on common space. For instance, Reed et al. [23] designed the structured joint embedding (SJE) model, which completes the matching of image visual features and semantic information in the common space through end-to-end training and then assists in classification. To adaptively design the mapping functions of image and class label vectors in the common space, Xian et al. [24] constructed the latent embedding (LatEm) model. Latent variables are introduced to enhance the bilinear compatibility model, and then the model learns the compatibility function of image and class embedding in the latent space. Besides, to alleviate the inconsistency in visual and semantic observations of interclass similarity, Tao et al. [25] introduced semantics-preserving local embedding and devised the semantics-preserving locality embedding (SPLE) method. A derived space is chosen as a common space, embedding both images and semantic data into it for classification. Subsequently, the method performed cross-domain concept matching through subspace learning to improve the performance.

c) Embedding methods based on visual space: The embedding methods based on semantic space have greatly improved the performance of ZSL approaches. However, in the distance metric procedure, multiple semantic vectors may be closest to the mapped images, leading to the "hubness" problem. To solve this dilemma, some scholars have proposed methods that map the subspace to the visual space, acknowledged as embedding methods based on visual space. For instance, Zhang et al. [26] designed the deep embedding model (DEM), which introduces a multimodal fusion method to select the CNN output visual feature space as the embedded space and map semantic features such as attributes and word vectors to the visual space. Recurrent neural networks can be operated for end-to-end learning of semantic space representation. Furthermore, Sung et al. [27] assembled the relation network (RN) model, which consists of embedding modules and relation modules. The two modules are end-to-end metalearning, where the embedding module embeds semantic features into the visual space, and the relation module is operated for unseen category discrimination.

3) ZSL Based on Generative Networks: ZSL methods based on the embedding model have achieved remarkable results. Still, they cannot address the problem of imbalanced image numbers among diverse classes, which severely limits the improvement of ZSL performance. With the growth of generative networks, techniques such as generative adversarial networks (GANs) and variational autoencoders (VAEs) have emerged, followed by the gradual maturity of methods based on generative networks for ZSL. According to the generative models employed, this type of method can be divided into three categories: GAN, VAE, and GAN&VAE.

The feature-generating network (FGN) [28] is a typical ZSL method based on GAN, which pairs the Wasserstein generative adversarial network (WGAN) [29] with a classification loss function to generate CNN features. These features are discriminative when training softmax classifiers and multimodal embedding methods. On this basis, Sariyildiz et al. [30] combined WGAN with gradient matching loss to obtain more discriminative synthetic features. Besides, Elhoseiny et al. [31] connected

ZSL with creativity and presented the creativity-inspired ZSL (CIZSL) method. The hallucinated class descriptions are utilized to explore unseen classes and generate visual features that transfer knowledge between seen and unseen categories.

Nevertheless, the drawbacks of GAN, such as instability of generated features, and pattern collapse, limit the development of ZSL methods based on GAN. To tackle this problem, similar to the development trend of generative networks, Schonfeld et al. [32] introduced the VAE into the embedding model framework and proposed the cross and distribution aligned VAE (CADA-VAE) method. Modality-specific aligned VAEs are employed to learn shared latent space features of images and class embedding. This approach utilizes the feature distributions learned from image and class embedding auxiliary information to construct multimodal information latent features related to unseen categories to help their classification. Subsequently, Liu et al. [33] designed the attentive region embedding with enhanced semantics (AREES) framework based on VAE. It acquires a shared LatEm space for visual features and semantics through three network branches. The first branch is an attentive region embedding, the second is a decomposition structure and a semantic pivot regularization, and the last is a multimodal VAE with cross-reconstruction loss and distribution alignment loss. Notably, the AREES approach can effectively alleviate the domain shift problem to a large extent due to the simultaneous optimization of feature extraction and feature generation.

To fully excavate the benefits of generative networks in ZSL, many scholars have combined GANs and VAEs to complement the strengths and weaknesses of the two generative networks and improve ZSL performance. For instance, Huang et al. [34] designed the generative dual adversarial network (GDAN), which exploits conditional variational autoencoders as feature generators. The visual-to-semantic mapping, semantic-to-visual mapping, and metric learning are combined in a unified framework. Meanwhile, this approach designs a novel dual adversarial loss to achieve better matching learning of features between modalities via utilizing bidirectional mapping of vision and semantics.

4) ZSL Based on Pretrained Models: Large-scale pretrained models (LPM) demonstrate powerful ZSL capabilities and achieve state-of-the-art performance in zero-shot image classification tasks, such as contrastive language image pretraining (CLIP) [35] and a large-scale image and noisy-text embedding (ALIGN) [36] methods.

Many scholars have proposed methods with prompt engineering and parameter fine-tuning to enhance the performance of the LPM models in downstream tasks, such as image classification. Prompt engineering refers to the approach of adjusting prompts by adding meaningful context for the task, such as the context optimization (CoOp) [37], the unsupervised prompt learning [38], and the conditional context optimization (CoCoOp) methods [39]. They utilize learnable vectors to model the context words of prompts while keeping all the pretrained model parameters fixed, effectively improving the performance of LPM in zero-shot image classification tasks.

Parameter fine-tuning involves a small amount of labeled downstream data, such as the BASIC method constructed by Pham et al. [40]. The BASIC method can achieve combined scaling of batch, data, and model size, effectively reducing the performance gap between general LPM and specific downstream supervised/semisupervised models and significantly improving model performance. In addition, to enhance the robustness of fine-tuning, a method called ensembling the weights of the zero-shot and fine-tuned models (WiSE-FT) was designed in [41]. A more reliable zero-shot model is built into this method by targeting the distribution of the objective and utilizing the weight space integration of fine-tuning, effectively reducing the model's relative error rate.

The work above focuses on adjusting the effectiveness of input information while neglecting the improvement of its richness. To address this issue, Novack et al. [42] proposed the classification method with hierarchical label sets (CHiLS). The main idea is to generate a set of subclasses for each class before inputting it into the CLIP model. The final category prediction is achieved by mapping the predicted subclasses back to their parent classes. Thereby, it improves the zero-shot image classification ability of the CLIP model without additional training costs.

III. Zero-Shot RS Image Scene Classification

A. Introduction to ZSRSSC

ZSRSSC refers to recognizing unseen scene categories with the assistance of unseen semantic information by learning the correspondence between the seen scene category images and their semantic information. The description of the ZSRSSC problem is as follows.

Given $\boldsymbol{D}^S = \{(\boldsymbol{x}_i^S, \boldsymbol{y}_i^S)|_{i=1}^{N^S}\}$ and $\boldsymbol{D}^U = \{(\boldsymbol{x}_i^U, \boldsymbol{y}_i^U)|_{i=1}^{N^U}\}$ as training and testing data in the visual space, where $\boldsymbol{x}_i^S$ and $\boldsymbol{x}_i^U$ represent the visual features of the ith RS image of seen scene categories and unseen scene categories, respectively, and $\boldsymbol{y}_i^S$ and $\boldsymbol{y}_i^U$ denote the corresponding labels of the ith RS image of seen scene categories $\boldsymbol{l}^S = \{1^S, 2^S, \ldots, C^S\}$ and unseen scene categories $\boldsymbol{l}^U = \{1^U, 2^U, \ldots, C^U\}$, respectively. The label sets $\boldsymbol{l}^S$ and $\boldsymbol{l}^U$ of the training data and testing data are disjoint, that is, $S \cap U = \emptyset$. In addition, each category corresponds to a semantic vector, let $\boldsymbol{R}^S = \{(\boldsymbol{r}_C^S \in \mathbb{R}^{d_r})|_{c=1}^{C^S}\}$ and $\boldsymbol{R}^U = \{(\boldsymbol{r}_C^U \in \mathbb{R}^{d_r})|_{c=1}^{C^U}\}$ be the training and testing data in the semantic space, where U represents the dimension of the semantic vector. ZSRSSC aims to learn the visual representation $\boldsymbol{D}^S = \{(\boldsymbol{x}_i^S, \boldsymbol{y}_i^S)|_{i=1,2}^{N^S}\}$ of seen category data and its corresponding label $\boldsymbol{l}^S = \{1^S, 2^S, \ldots, C^S\}$, semantic representation $\boldsymbol{R}^S = \{(\boldsymbol{r}_C^S \in \mathbb{R}^{d_r})|_{c=1}^{C^S}\}$, and predict the label $\boldsymbol{l}^U$ of unseen category scene images by inputting the semantic representation $\boldsymbol{R}^U = \{(\boldsymbol{r}_C^U \in \mathbb{R}^{d_r})|_{c=1}^{C^U}\}$ of unseen category data during the testing stage.

B. ZSRSSC Methods

The early ZSRSSC approaches first explored semantic space embedding. Subsequently, ZSRSSC approaches based on common space embedding emerged. Moreover, some scholars have attempted to extend ZSL methods based on generative networks to ZSRSSC tasks. With the emergence of LPM, there has also been some research on methods based on pretrained models in RS. As a consequence, similar to the development of ZSL

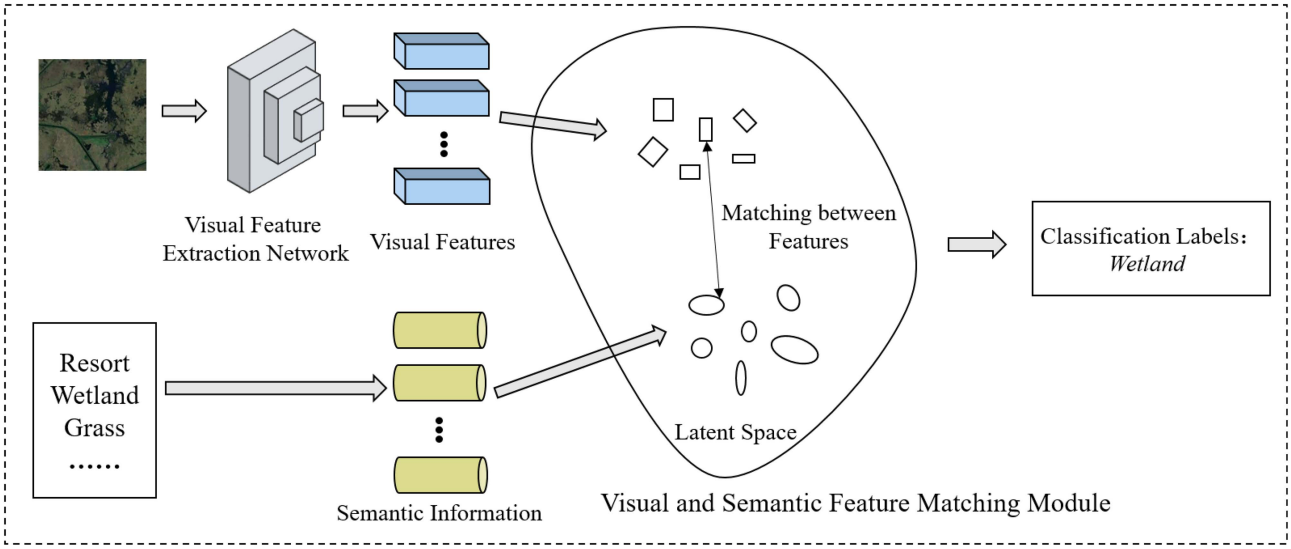

Fig. 6. Schematic diagram of the locality-preservation deep cross-modal embedding network.

approaches, the current ZSRSSC methods have also developed with the development of neural network technology, resulting in three branches of methods: methods based on embedding models, methods based on generative networks, and methods based on pretrained models. In the following sections, we will provide a detailed introduction to the ZSRSSC methods for these three branches.

1) ZSRSSC Based on Embedding Models: The ZSL method based on embedding models desires to use embedding space to classify unseen categories. It trains an embedding model and utilizes existing data and corresponding labels in the embedding space to extract the correspondence between label semantics and visual images. Subsequently, this method can classify new category images with the help of unseen label semantics. Similarly, the ZSRSSC method based on embedding models is the extension of the ZSL method based on embedding models in RS. Specifically, the ZSRSSC method based on embedding models utilizes an embedding space to transform images into low-dimensional vectors, enabling differentiation between different categories of RS images. The method embeds new category images into the existing embedding space by learning the similarity relationships between different categories.

ZSL methods based on embedding models focus on learning the mapping relationship between semantic and visual vectors. Li et al. [43] first introduced ZSL methods into RSSC tasks. They propose a label propagation (ZSRSSC-LP) method for the ZSRSSC tasks. An unsupervised domain adaptation model is utilized to predict the labels of test images. A semantic-directed graph and the initial prediction are operated to design label propagation. Two strategies are employed to suppress noise in zero-shot classification results. They are the visual similarity between images from the same scene class and a label refinement method based on sparse learning.

a) Performance improvements in different directions: Subsequently, scholars have conducted numerous studies in diverse directions to improve the performance of the ZSL approaches on ZSRSSC tasks.

First, researchers have developed different approaches to address the issue of inconsistent space in ZSL methods based on embedding models on ZSRSSC tasks. For instance, a semisupervised Sammon embedding method was studied in [44], which is employed to modify the semantic space prototypes. This allows them to transfer unseen knowledge from the semantic space to the visual space, and effectively synthesize unseen category prototypes in the visual space. Besides, Li et al. [45] proposed a LPDCMEN, which fully absorbs pairwise intramodal and multimodal supervision in an end-to-end manner, as illustrated in Fig. 6.

Second, to alleviate the problem of mapping domain shift in ZSL methods based on embedding models on ZSRSSC tasks, a distance-constrained SAE was designed in [46], which aligns the visual space and semantic space. Discriminative distance metric constraints are used to enhance the discriminative ability of the SAE for categories in this method. This constraint minimizes the Euclidean distance between coding vectors of same-class samples and maximizes the distance between coding vectors of different-class samples.

Third, Li et al. [47] developed a method to enhance the robustness of the mapping, which utilizes robust cross-domain mapping and progressive semantic benchmark correction. The method employs category semantic vectors from seen classes and scene image samples to achieve deep feature extractor learning and robust mapping from the visual space to the semantic space. It progressively corrects the semantic vectors of unseen categories through corepresentation learning and the k-nearest neighbor method.

To enhance the discriminative ability of visual feature extraction, Wang et al. [48] designed a local-global feature fusion module, which can adaptively label and model the internal local landscape. They effectively fuse the local features and global information and present RS scenes with more differentiation.

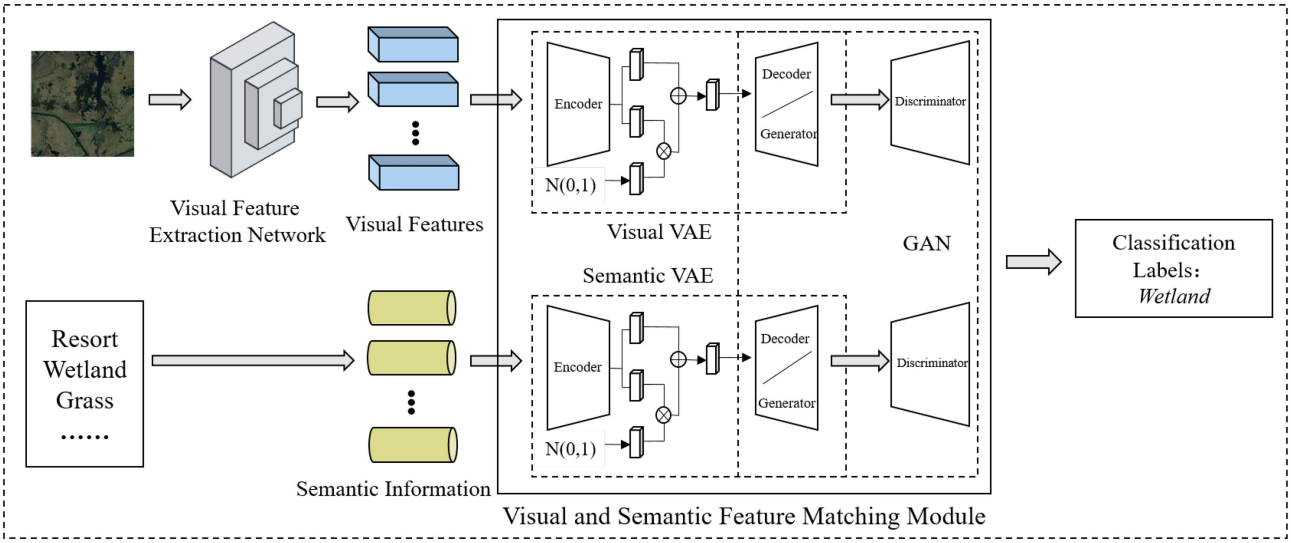

Fig. 7. Schematic diagram of the ZSRSSC method based on GAN-VAE.

Besides, a graph convolutional network was introduced to fully utilize the correlation between class information in [49]. The main idea of this approach is to explore the correlation structure between categories in a unified framework via a semantic graph convolutional network. In summary, the network achieves zero-shot image classification in an end-to-end manner through graph-based semantic embedding.

Moreover, in [50], a new road tree dataset was introduced to research the performance of ZSRSSC in fine-grained object recognition. In the design of the fine-grained ZSRSSC method, a compatibility function is employed to establish relationships between unseen and seen categories. Note that the compatibility function measures the similarity between image features extracted by CNN and auxiliary information describing interested category semantics. On the other hand, Li et al. [51] integrated embedding models and generative networks. They designed a deep alignment network (DAN) that trains the classifier based on latent space mapping and generated training samples, balancing the classification performance of seen and unseen categories.

b) Research on the semantic information of the RS: In addition to the above performance improvements in different directions, some scholars have also studied the richness of semantic information. They fully explored the characteristics of the RS field and constructed new semantic information. For instance, a new RS knowledge graph (RSKG) was constructed from scratch in [52]. Thus, the generation of scene category semantic representations can utilize the representation learning of RSKG. To reduce the cost of manually annotating RS images, Xu et al. [53] first introduced attribute semantic representation into the RS field. They then propose automatically collecting visually detectable attributes by fine-tuning the CLIP model.

2) ZSRSSC Based on Generative Networks: The ZSRSSC method based on generative models aims to utilize generative networks to generate unseen class samples and achieve recognition of unseen class samples through supervised classification. Among them, the generation of unseen class samples is mainly achieved through generative models to learn from seen class samples. The so-called zero-shot refers to a method in which the ZSRSSC method can accurately classify unseen categories during the training phase without any samples of the target class. This is achieved by learning the feature distribution of RS images and the relationships between categories.

It can be found that ZSRSSC approaches based on generative networks transform the zero-shot classification problem into a data-missing issue by generating unseen class samples. Then, they utilize the supervised learning classification paradigm to achieve zero-shot recognition. For instance, Ma et al. [54] employed GAN to enhance the VAE model and proposed the GAN-VAE method, as displayed in Fig. 7. The discriminator of GAN is used to learn a reconstruction quality metric suitable for VAE, and a cross-modal feature matching loss is designed to ensure the alignment between the visual and semantic features of a category. On this basis, Liu et al. [55] utilized two VAEs to project visual features and semantic features of scene classes into a shared latent space, and further mined the contrast relationship between cross-modal features. They promoted the alignment of instance images with their corresponding semantic features through a multilevel feature alignment method, while ensuring that the instance image remained distinct from the semantic features of other categories. Furthermore, to enhance the generation quality of unseen category images, Li et al. [56] utilized conditional Wasserstein generative adversarial network (WGAN) to generate image features. The method introduces classification loss, semantic regression module, and class prototype loss to constrain the generator to ensure that the generated unseen class RS images still exhibit interclass similarity and intraclass diversity.

3) ZSRSSC Based on Pretrained Models: The powerful ZSL inference capability of LPM has prompted more scholars in RS to research ZSRSSC methods using LPM. In summary, there are two main research directions: One is to build a general LPM of the RS domain similar to the CLIP model from scratch, and the

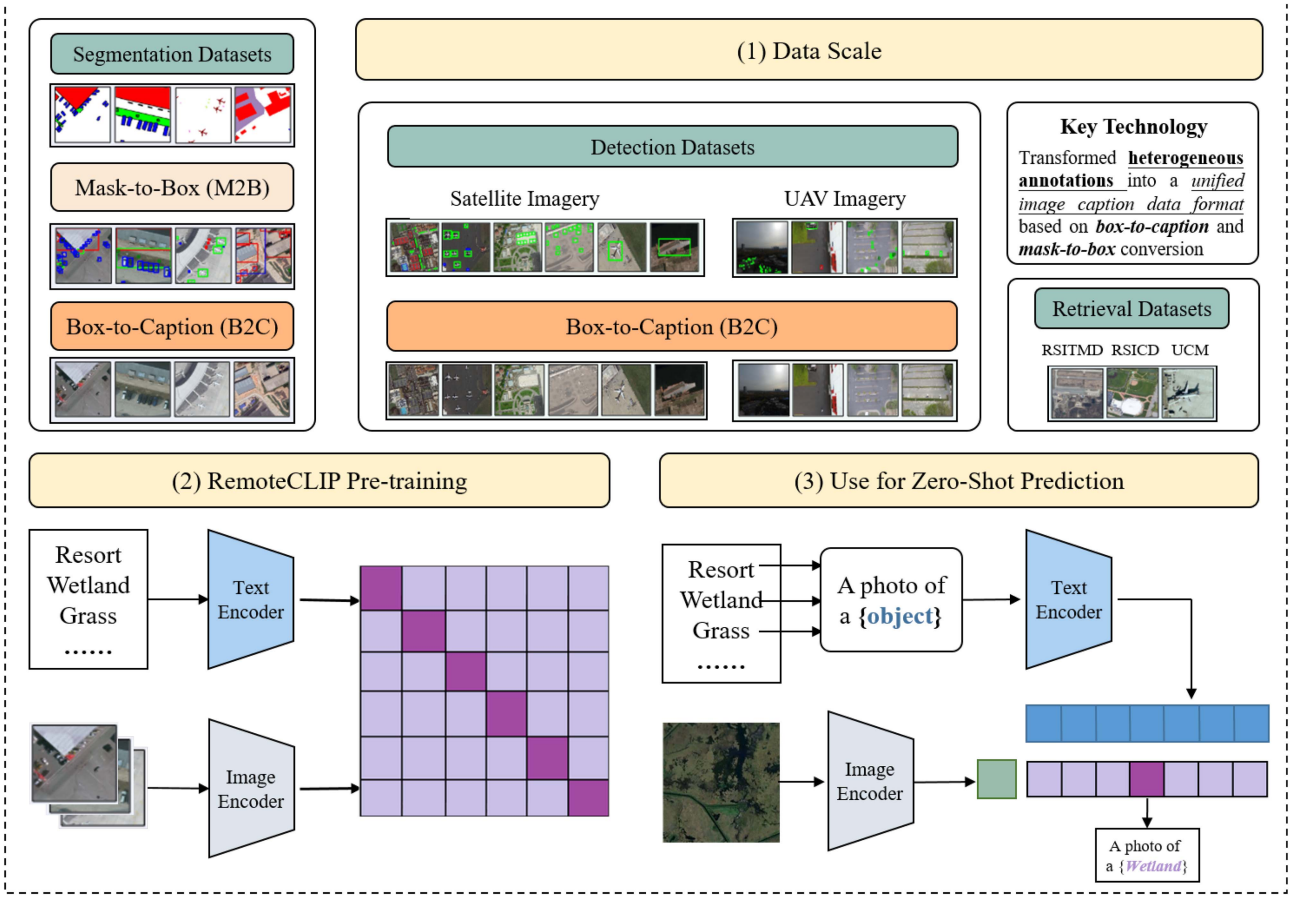

Fig. 8. Schematic diagram of establishment process for the RemoteCLIP model.

other is to fine-tune the CLIP model to make it applicable to the RS domain and achieve good performance.

The typical representative of the former direction is the bridging-vision-and-language (BriVL) model proposed by Fei et al. [57]. They follow the weak semantic correlation assumption and construct a huge dataset crawled from the Internet to make it closer to real-world data distribution. By pretraining on this dataset, the generalization performance of the BriVL model can be improved. Besides, a momentum mechanism is also employed to maintain a negative sample queue between different training batches. This helps to reduce the demand for GPU resources and achieve better zero-shot classification performance in ZSRSSC tasks.

The latter direction improves the zero-shot classification performance of the RS domain pretrained model by collecting large-scale RS datasets and fine-tuning the CLIP model using these datasets. Many scholars have explored various methods to collect large-scale RS datasets. For instance, Liu et al. [58] transformed heterogeneous annotations into a unified image caption data format based on box-to-caption and mask-to-box conversion, as displayed in Fig. 8. They further merge UAV imagery to generate a pretraining dataset 12 times larger than the combination of RSITMD, RSICD, and UCM datasets, which are the common RS Domain retrieval datasets. Through large-scale pretraining, they obtain the RemoteCLIP model specifically designed for the RS domain. The model has strong generalization ability and outperforms the CLIP model in zero-shot classification performance on multiple datasets.

On the other hand, some researchers introduce the pseudolabeling strategy to overcome the limitations of image-text pairs available for training in the RS domain. For instance, Mo et al. [59] employed a semisupervised learning method for training the CLIP model (S-CLIP) using additional unpaired images. In this approach, the pseudolabeling technique is used to create pseudolabels at the caption and keyword levels, which helps the model better comprehend both the global structure and local language phrases. As a result, the zero-shot scene classification accuracy on the WHU-RS19 dataset increased significantly. Besides, Li et al. [60] built the RS-CLIP model, which introduced the pseudolabeling technique to automatically generate pseudolabels from unlabeled datasets for model fine-tuning in the RS domain. They also develop a curriculum learning strategy to improve the performance of ZSRSSC tasks through multistage model fine-tuning.

In addition, to handle cross-category and multiscale variations in high-resolution RS images, and maintain consistency between multimodal instruction retrieval and user queries, Kuckreja

TABLE I
BASIC INFORMATION OF EACH DATASET IN RSSC TASKS

Dataset	Origin	Released	Categories	Images	Image Size
RSC-11	Google Earth imagery	–	11 scenes	about 100	512 × 512 pixels
SIRI-WHU	Google Earth imagery	Wuhan University	12 urban areas	200	200 × 200 pixels
RSD46-WHU	Google Earth and Tianditu imagery	Wuhan University	46 scenes	500-3000	–
WHU-RS19	Google Earth imagery	Wuhan University	19 aerial scenes	50	600 × 600 pixels
RSSCN7	Google Earth imagery	Wuhan University	7 aerial scenes	400	400 × 400 pixels
UCM21	–	University of California, Merced (UCM)	21 land-use types	100	–
AID30	–	–	30 aerial scenes	ranging from 220 to 420	600 × 600 pixels
NWPU45 (NWPU-RESISC45)	–	Northwestern Polytechnical University (NWPU)	45 scenes	700	–
PatternNet	–	–	38 scenes	800	256 × 256 pixels
PSI-CB256	–	–	6 major classes and 35 subcategories	an average of around 690	256 × 256 pixels
RSSDIVCS	UCM21, AID30, NWPU45, PatternNet, and PSI-CB256	Wuhan University	70 scenes	800	256 × 256 pixels

et al. [61] proposed the first multipurpose RS vision-language model named GeoChat. It provides multitask dialog capabilities with high-resolution RS images, which can answer image-level queries and accept region inputs to complete region-based dialogs.

Besides, Rahhal et al. [62] explored the usefulness of visual-linguistic models (VLMs) for ZSRSSC. Specifically, they employed 13 VLMs derived from the CLIP/Open-CLIP technique to experimentally identify the most effective RS image cues for querying the linguistic competence of the VLMs. Superior results for zero-shot classification accuracy were demonstrated when using a large-scale CLIP model on three widely recognized RSSC datasets in this work.

IV. DATASETS, EXPERIMENTAL RESULTS, AND ANALYSIS

A. Datasets

We will review, to the extent possible, the whole range of datasets commonly used for RSSC tasks in this section. There are ten datasets in total and they are SIRI-WHU [63], RSD46-WHU [64], [65], WHU-RS19 [66], RSC11 [67], RSSCN7 [68], UCM21 [69], AID30 [70], NWPU45 [71], PatternNet [72], and PSI-CB256 [73]. The latter five datasets form the RSSDIVCS [51] dataset, which is the only publicly released dataset specifically for the ZSRSSC task. The basic information of each dataset is detailed in Table I.

The RSC-11 dataset is extracted from Google Earth. Three spectral bands were used including red, green, and blue. There are 11 complicated scene categories including *dense forest, grassland, harbor, high buildings, low buildings, overpass, railway, residential area, roads, sparse forest, and storage tanks*. The dataset contains 1232 images, with each category about 100 images. Each image has a size of 512 × 512 pixels.

The SIRI-WHU dataset is dragged from Google Earth. There are 12 urban area categories, including *agriculture, commercial, harbor, idle land, industrial, meadow, overpass, park, pond, residential, river, and water*. The dataset comprises 2400 images, each having dimensions of 200 × 200 pixels.

The RSD46-WHU dataset is collected from Google Earth and Tianditu, with 117 000 images and 46 classes. The categories include *airplane, airport, artificial dense forest land, artificial sparse forest land, bare land, basketball court, blue structured factory building, building, construction site, cross river bridge, crossroads, dense tall building, dock, fish pond, footbridge, graff, grassland, low scattered building, irregular farmland, medium density scattered building, medium density structured building, natural dense forest land, natural sparse forest land, oil tank, overpass, parking lot, plastic greenhouse, playground, railway, red structured factory building, refinery, regular farmland, scattered blue roof factory building, scattered red roof factory building, sewage plant-type-one, sewage plant-type-two, ship, solar power station, sparse residential area, square, steelworks, storage land, tennis court, thermal power plant, vegetable plot, and water*.

The WHU-RS19 dataset is derived from Google Earth and released by Wuhan University. The dataset composes 19 categories of aerial scenes that include *airport, bridge, river, forest, grassland, pond, parking lot, port, overpass, residential area, industrial area, commercial area, beach, desert, farmland,*

TABLE II
CATEGORY DETAILS OF PSI-CB256 DATASETS

Major class	Subcategory
Agricultural land	*green farmland, dry farm, bare land*
Woodland	*artificial grassland, sparse forest, forest, mangrove, river protection forest, shrubwood, sapling*
Transportation and facility	*airport runway, avenue, highway, marina, parking lot, crossroads, bridge, airplane*
Water area and facility	*coastline, dam, hirst, lakeshore, river, sea, stream*
Construction land and facility	*city building, container, residents, storage room, pipeline, town*
Other land	*desert, snow mountain, mountain, sandbeach*

soccer field, mountain, park, and train station. Each category has 50 samples, and the image size is 600 × 600 pixels.

The RSSCN7 dataset also comes from Google Earth imagery and is released by Wuhan University. It includes 2800 aerial scene images, categorized into *grassland, forest, farmland, parking lot, residential area, industrial area, and river/lake*. Each scene category has 400 images, and the image size is 400 × 400 pixels.

The UCM21 dataset is a ground truth dataset created by the University of California, Merced (UCM) in 2010, which comprises 21 land-use types. They are *agriculture, airplane, baseball diamond, beach, building, church, dense residential, forest, highway, golf course, port, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, tank, and tennis court*. There are 100 images for each land-use type, resulting in a total of 2100 images. The AID30 dataset consists of 30 aerial scene types such as *airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking lot, playground, pond, port, train station, resort, river, school, sparse residential, square, stadium, oil tank, and overpass*. The dataset has varying numbers of sample images for different aerial scene types, ranging from 220 to 420 images, and the image size is 600 × 600 pixels.

The NWPU45 dataset, also known as NWPU-RESISC45, is a public benchmark for RSSC created by Northwestern Polytechnical University (NWPU) in 2017. It contains 31 500 images covering 45 scene classes, 700 images per class.

The PatternNet dataset is composed of 38 classes, including *airplane, baseball field, basketball court, beach, bridge, cemetery, jungle, Christmas tree farm, closed road, coastal mansion, crosswalk, dense residential, ferry terminal, football field, forest, highway, golf course, port, intersection, mobile home park, sanatorium, oil field, oil well, overpass, parking lot, parking spot, railway, river, runway, runway marking, shipyard, solar panel, sparse residential, tank, swimming pool, tennis court, substation, and sewage treatment plant*. Each class has 800 images with a size of 256 × 256 pixels.

The PSI-CB256 dataset has six major classes and 35 subcategories, totaling approximately 24 000 images, with an average of around 690 images per category. The image size is 256 × 256 pixels. The category details are shown in Table II.

Moreover, Li et al. [51] integrated the UCM21, AID30, NWPU45, PatternNet, and PSI-CB256 to construct a dataset named RSSDIVCS specifically for the ZSRSSC task. The dataset contains 70 scene categories, with 800 images per category. The image size is 256 × 256 pixels. We select one image from each category set to form an example, as displayed in Fig. 9.

B. Experimental Results and Analysis

1) Experimental Setup: This article performs comparison experiments to evaluate the ZSRSSC methods using the RSSDIVCS dataset. Since the RSSDIVCS consists of UCM21, AID30, NWPU45, PatternNet, and PSICB256 datasets, the methods are considered to be validated on five commonly used RSSC benchmark datasets. All chosen methods utilize two types of vectors, general and domain knowledge, for semantic information. It is worth mentioning that semantic vectors are typically obtained through natural language processing models. The Word2Vec model and the bidirectional encoder representations from transformers (BERT) model are the two commonly employed models for extracting general and domain knowledge, respectively.

To effectively validate the performance of various ZSRSSC methods, we selected four representative models from the computer vision (CV) domain and extended them to the RS domain for baseline comparison. These models contain SAE, DMaP, SPLE, and CIZSL approaches. Similarly, the experiment selected the CADA-VAE and GDAN approaches in CV as baseline comparisons for ZSRSSC approaches based on generative networks. To cover most of the current approaches in the comparison range, this article includes a wide range of ZSRSSC methods. The overall accuracy (OA) of each method is calculated, and the mean and standard deviation are depicted in Table III.

2) Analysis of Experimental Results: Table III illustrates the experimental results of diverse ZSRSSC methods on the RSSDIVCS dataset.

The accuracy of the method has shown significant improvement across all types. For instance, the OA of ZSRSSC approaches using embedding models now exceeds 30% with a seen/unseen category ratio of 60/10. Moreover, hybrid models that combine embedding models and generative networks can achieve an OA of 40% at the same ratio. This progress demonstrates the effectiveness and potential of ZSRSSC methods. As accuracy improves, these approaches are shifting from merely predicting to actively distinguishing unseen categories, enhancing the ability of professionals in the RS field to uncover new insights and knowledge.

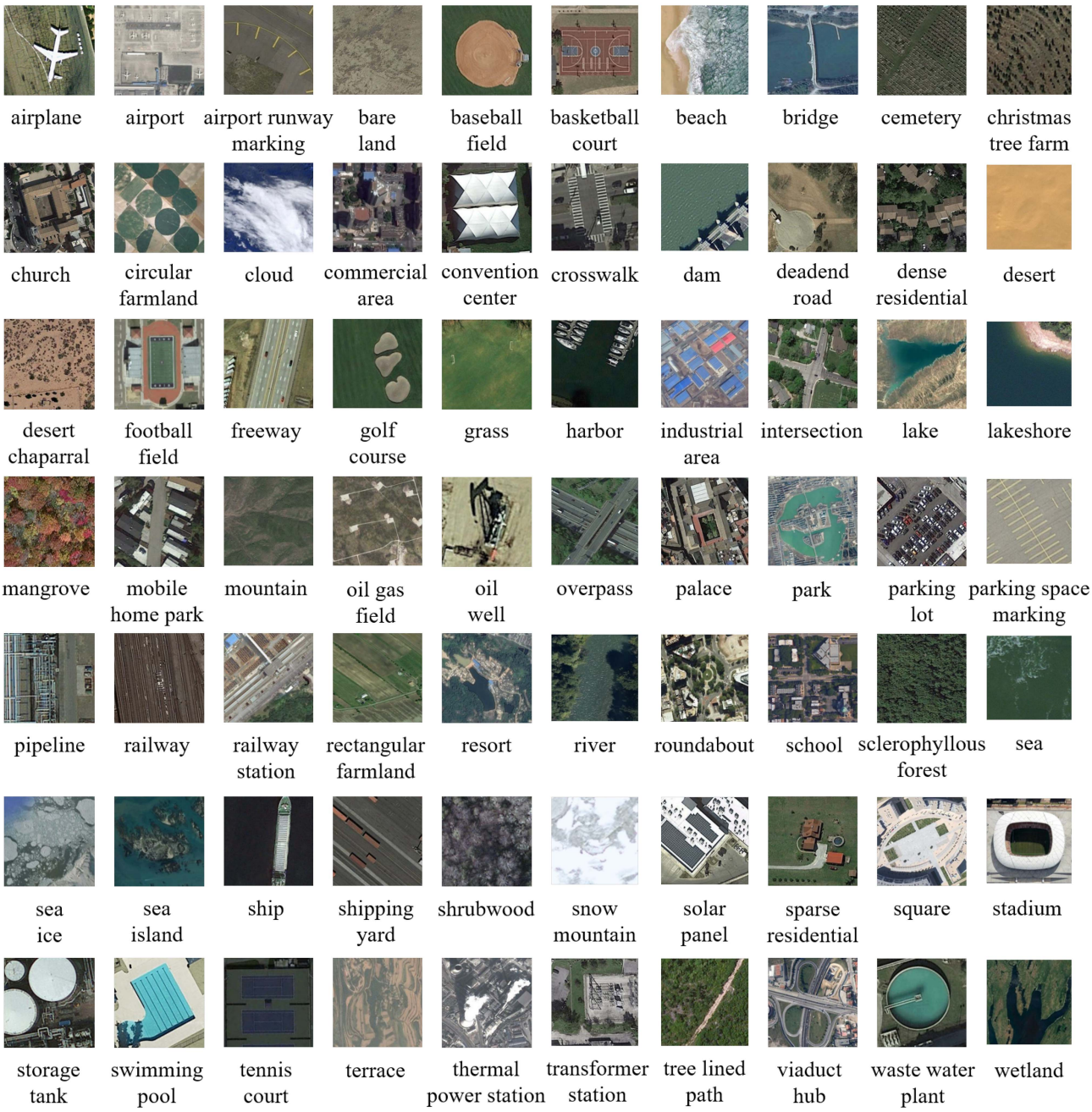

Fig. 9. Example images of the 70 scene categories from the RSSDIVCS dataset.

From the standpoint of method migration, it is feasible to directly emigrate ZSL methods in the CV domain to the RS. In general, ZSL methods based on embedding models perform well when applied to the RS domain, as seen in the comparison of SAE, DMaP, SPLE, and CIZSL baseline methods.

In the comparison of ZSRSSC methods, embedding model-based methods generally perform better at lower seen/unseen ratios (i.e., 40/30) compared to other methods, such as ZSC-SA versus GAN-VAE. However, generative network-based methods slightly outperform those based on embedding models at higher ratios (i.e., 60/10). Overall, embedding model-based methods typically provide superior classification accuracy in both general and specific knowledge contexts. This suggests that methods based on embedding models are more effective for ZSRSSC tasks due to the complexity of RS scenes and the challenges generative networks face in producing high-quality samples.

It is evident that hybrid model-based methods typically achieve higher accuracy compared to other methods in the RS domain. Traditional methods based on embedding models such as LPDCMEN, ZSRSSC-LP, and LIANHE show lower accuracy when compared to hybrid model-based approaches such as CADA-VAE and DAN, which have demonstrated superior accuracy levels. This is probably due to these methods integrating the

TABLE III
OVERALL ACCURACIES (%) OF DIFFERENT METHODS UNDER VARIOUS SEEN/UNSEEN RATIOS

Domain	Methods	Method Type	Word2Vec			BERT		
			60/10	50/20	40/30	60/10	50/20	40/30
Baseline	SAE [21]	Embedded	23.5 ± 4.2	13.7 ± 1.7	9.6 ± 1.4	*22.0 ± 1.7*	12.4 ± 1.9	8.8 ± 1.3
	DMaP [22]	Embedded	*26.0 ± 3.6*	*16.7 ± 2.2*	*10.4 ± 0.9*	16.4 ± 1.9	*15.6 ± 1.9*	*10.0 ± 0.8*
	SPLE [25]	Embedded	20.1 ± 3.7	13.2 ± 1.9	9.8 ± 1.4	19.0 ± 3.8	13.2 ± 2.6	8.3 ± 2.0
CV	CIZSL [31]	Generative	20.6 ± 1.4	10.6 ± 3.7	6.0 ± 1.2	20.4 ± 4.1	10.3 ± 1.9	6.2 ± 2.1
	CADA-VAE [32]	Embedded-Generative	*41.1 ± 2.3*	*30.3 ± 2.7*	*21.2 ± 2.9*	*48.1 ± 2.9*	*37.1 ± 3.5*	*26.3 ± 2.2*
	GDAN [34]	Generative	27.6 ± 5.9	13.6 ± 4.7	8.1 ± 1.1	32.1 ± 8.2	13.8 ± 1.1	9.8 ± 0.9
RS	ZSRSRC-LP [43]	Embedded	14.8 ± 0.4	8.3 ± 0.7	4.7 ± 0.3	14.1 ± 1.6	7.4 ± 0.3	4.6 ± 0.4
	ZSC-SA [44]	Embedded	26.7 ± 5.3	15.2 ± 1.0	12.1 ± 0.8	29.3 ± 3.8	18.3 ± 1.3	13.1 ± 1.3
	LPDCMEN [45]	Embedded	34.2 ± 9.0	19.7 ± 3.6	15.8 ± 2.9	43.8 ± 7.3	24.9 ± 3.7	21.6 ± 3.8
	LIANHE [47]	Embedded	33.1 ± 5.8	19.0 ± 1.8	12.5 ± 1.5	35.8 ± 5.1	19.6 ± 2.7	12.7 ± 2.3
	DAN [51]	Embedded-Generative	**44.3 ± 2.6**	**34.7 ± 1.7**	**24.3 ± 3.7**	*50.2 ± 2.5*	*43.4 ± 2.7*	*31.5 ± 2.0*
	GAN-VAE [54]	Generative	28.0 ± 1.4	16.6 ± 3.5	9.3 ± 1.6	35.1 ± 1.2	17.9 ± 1.1	11.4 ± 0.4
LPM	RemoteCLIP [58]	Pre-trained	–	–	–	**81.5 ± 19.3**	**64.4 ± 27.5**	**58.9 ± 27.5**

The bold value indicates that the data is the best performance, and the italic value indicates that the data is the second best performance.

advantages of embedding models and generative networks. For instance, the CADA-VAE method learns the parameters of the embedding model using VAE, which leads to higher accuracy. This also demonstrates the potential of joint embedding models and generative networks in dealing with ZSRSSC tasks and highlights the importance of embedding models in achieving higher classification accuracy.

RemoteCLIP, an LPM-based method, outperforms previous methods in numerous seen/unseen ratios. This may be attributed to the fact that the pretraining dataset of the RemoteCLIP is 12 times larger than the previous methods. Another possible reason may be that the RemoteCLIP contains a much larger number of image-text feature pairs. However, since the pretraining dataset is not publicly available, it cannot be ruled out that the unseen images in the RSSDIVCS dataset may have been trained on RemoteCLIP. Nevertheless, the advantages of methods based on LPM are fully reflected in the comparison results in Table III.

In summary, the feasibility of the ZSRSSC approach is no longer in doubt, and the overall trend indicates a more favorable development. For the ZSRSSC task, the research method direction is mainly dominated by embedding models, with hybrid models combined with generative networks also performing relatively well. Moreover, LPM-based methods also exhibit great potential.

V. DISCUSSIONS

In this section, we will summarize the issues with the current ZSRSSC methods, some of the solutions currently available, and finally point out the direction of their future development.

A. Existing Issues

Currently, ZSRSSC approaches are still in the development stage with promising prospects. However, some issues persist in the existing methods.

We summarize the issues of current ZSRSSC approaches, mainly including the following points:

1) Unique Characteristics of RS Images Often Make Extracting Useful Features for ZSRSSC Tasks Difficult: This is the fact that RS images have different spatial distributions compared to natural images. It usually fails to capture the contextual information of RS image features. The reason is that the models used to extract features from RS images are generally pretrained on large-scale natural image datasets such as ImageNet, such as GoogleNet, ResNet, etc.

2) Auxiliary Semantic Information Commonly Utilized in CV May Not Work Well in RS: The category labels in the field of RS are usually unable to provide semantic solid representations like those in natural images. Methods that convert category labels into semantic vectors using natural language processing models are limited. This is due to these models being trained on common sense texts, and the semantic vectors obtained from them often lack RS domain knowledge. This constraint limits the performance of ZSRSSC approaches.

3) Inherent Problem of ZSL Methods Exists in ZSRSSC Methods: This is because ZSRSSC approaches essentially follow the paradigm of ZSL, and the problems faced by ZSL in the method transfer process have not been fundamentally eliminated. Therefore, the challenges of ZSL in CV also persist in ZSRSSC approaches, such as the semantic gap caused by inconsistent structures between visual and semantic modalities and the domain shift due to the significant class bias of training and testing sets.

4) Scarcity of Datasets in ZSRSSC Tasks is Unfavorable for Practical Applications: Currently, the only dataset available is RSSDIVCS, which contains 70 scene categories with 800 images each. However, the equal number of scene images per category masks the potential challenge of imbalanced category image quantities that ZSRSSC approaches may face when applied in practice.

B. Methodologies

In response to the abovementioned third existing issue, which is the same inherent problem in the ZSRSSC task, many scholars have proposed new theories and designed new loss functions. Commonly used distance metric functions in design are cosine similarity, the Gaussian kernel function, and Euclidean distance.

To alleviate the semantic gap problem, Quan et al. [44] designed a structural difference measurement DM using the Euclidean interclass instance to close the class structure of semantic space prototypes and visual space prototypes, which is defined below

$$DM = \sum_{i=1}^{N} \sum_{j=i+1}^{N} |\boldsymbol{\nu}_{ij} - \boldsymbol{\mu}_{ij}| \tag{3}$$

where $\boldsymbol{\nu}_{ij}, \boldsymbol{\mu}_{ij}$ denotes the normalized Euclidean interclass distances between class i and class j in visual and semantic space, respectively. N is the number of categories, including seen and unseen classes.

Li et al. [45] implemented the cosine-like distance and the Euclidean distance to pursue robust cross-modal matching constraint $\mathcal{L}_{CMM}$, which is defined as follows:

$$\mathcal{L}_{CMM} = \alpha \cdot \left(\sum_{i=1}^{N^s} \sum_{j=1}^{C^s} \sum_{k=1}^{2} \left(-M_{i,j}^{k} \log p \left(M_{i,j}^{k} = 1 | \boldsymbol{X}, \boldsymbol{Y} \right) \right) \right) + \beta \cdot \left(\sum_{j=1}^{C^s} \| \boldsymbol{Y}_j - \frac{1}{N_j^s} \sum_{i=1}^{N^s} M_{i,j}^{1} \cdot \boldsymbol{X}_i \|_F^2 \right) \tag{4}$$

where N^s, C^s denotes the average semantic relevance between seen and unseen classes, respectively. $\boldsymbol{X}, \boldsymbol{Y}$ represents the mapped vector in the latent space of visual and semantic features, respectively. $p(\cdot)$ means the pairwise intramodal similarity between the visual and semantic features. α, β are the weights of multiple constraints. The formulation of $M_{i,j}^{k}|_{k=1}^{2}$ and $p(\cdot)$ are

$$M_{i,j}^{1} = \begin{cases} 1, & \text{for} \quad i = j \\ 0, & \text{for} \quad i \neq j \end{cases} \tag{5}$$

$$M_{i,j}^{2} = 1 - M_{i,j}^{1} \tag{6}$$

$$\begin{cases} p\left(M_{i,j}^{1} | \boldsymbol{X}, \boldsymbol{Y}\right) = 1/(1 + e^{\boldsymbol{X}_i^T \cdot \boldsymbol{Y}_j}) \\ p\left(M_{i,j}^{2} | \boldsymbol{X}, \boldsymbol{Y}\right) = 1 - 1/(1 + e^{\boldsymbol{X}_i^T \cdot \boldsymbol{Y}_j}). \end{cases} \tag{7}$$

Furthermore, Wang et al. [48] designed the weight mapping loss function $\mathcal{L}_{WM}$ using a Gaussian kernel function and the popular LSE Loss (least square embedding loss), which is defined as follows.

$$\mathcal{L}_{WM} = \mathcal{L}_{s-u} \cdot \sum \| \boldsymbol{F}_i - \boldsymbol{F}_{ss} \|^2 + \lambda \| \boldsymbol{W} \|^2 \tag{8}$$

$$\mathcal{L}_{s-u} = \frac{1}{V} \sum_{v=1}^{V} e^{-\|\boldsymbol{F}_{ss} - \boldsymbol{F}_{su}\|^2} \tag{9}$$

where $\mathcal{L}_{s-u}$ denotes the average semantic relevance between seen and unseen classes. $\boldsymbol{F}_i$, $\boldsymbol{F}_{ss}$, and $\boldsymbol{F}_{su}$ represent image vectors, seen class semantic vectors, and unseen class semantic vectors, respectively. $\boldsymbol{W}$ is a randomly initialized encoding matrix for the semantic embedding, aiming to align the visual space and semantic space, and λ is a regularization parameter.

In addition, to address domain shift, Wang et al. [46] used Euclidean distance constrains the same categories to be close and different categories to be far away. The domain shift regularization constraint they designed is shown below

$$\mathcal{L}_{DS} = \| \boldsymbol{W}_s - \boldsymbol{W}_u \|_F^2 \tag{10}$$

where $\boldsymbol{W}_s$, $\boldsymbol{W}_u$ represent the encoder matrix for seen and unseen classes, respectively.

C. Future Development

The following analysis discusses the future development of ZSRSSC approaches. As a newly emerging direction in the RSSC tasks, ZSRSSC approaches have made significant progress in recent years. These approaches are derived from ZSL methods used in CV, which are closely related to deep learning. Given that RS images have larger intraclass variance and smaller interclass variance, and there is no manually annotated attribute source, the effect of extending ZSL methods directly to the RS domain is usually poor. Therefore, based on the characteristics of the RS field, the research of ZSRSSC approaches still has considerable potential for development. Presently, there are several potential directions for future research in ZSRSSC.

1) High-Quality Semantic Information Annotation is Necessary for Accurate Classification in ZSRSSC, Particularly for Recognizing Images of Unseen Categories: Word2Vec or BERT models are commonly used to extract semantic vectors in ZSRSSC. The semantic space trained by these models is generally composed of a Wikipedia corpus. However, RS images have significant differences from natural images, and the corresponding semantic information of RS images is more specific to the domain. Therefore, constructing high-quality semantic information with domain knowledge is still an area for future development.

2) ZSRSSC Involves the Input of Two Modal Features, and the Modal Feature Mapping Functions are Still the Core Determinants of Method Performance: Most current ZSRSSC approaches adopt the idea of CMT in ZSL and achieve ZSRSSC based on embedding models. Therefore, to further enhance the recognition accuracy of the method, it is necessary to refine the mapping functions of these two modalities to obtain superior experimental results.

3) The Practical Application of Methods Must Fully Consider the Distribution of Image Category Samples to Ensure Performance Guarantee: Most current ZSRSSC methods are validated on the RSSDIVC dataset, which has a balanced number of images for each category. Therefore, these methods did not consider the impact of imbalanced image quantity of categories. However, this problem often exists and significantly impacts performance in practical application scenarios. Hence, advancing the integration and enhancement of the implementation mechanisms for both few-shot RSSC and ZSRSSC could represent an additional avenue for the development.

4) With the Advancement of Large-Scale Vision-Language Models, Their Powerful Potential in ZSL Tasks Has Been

Demonstrated: Consequently, it has become a trend to focus on researching and optimizing zero-shot models specifically for the RS domain. There are already many large-scale pretraining models based on the CLIP model in RS, which significantly improve the performance of ZSRSSC by fine-tuning the CLIP model with a large amount of RS pretraining data. However, balancing method performance and fine-tuning costs remains a large-scale RS future optimization.

VI. Conclusion

The ZSRSSC methods have emerged as a new direction in recent years. With the exploration of predecessors, a series of achievements have been made. Unlike traditional RSSC tasks, the difference with this approach lies in its ability to recognize unseen categories. This is significant in the era of RS data explosion and has considerable research prospects. In this article, through a comparative analysis of the current ZSRSSC approaches, it has been found that embedding model-based research is more employed than those based on generative networks and is more suitable for ZSRSSC tasks. Besides, hybrid model methods combining embedding models and generative networks have demonstrated superior performance. Noticeably, large-scale RS models have also shown great potential in ZSRSSC tasks. More specifically, this article analyzes the problems existing in ZSRSSC approaches and the prospects of their development direction.

References

[1] L. Gómez-Chova, G. Camps-Valls, J. Munoz-Mari, and J. Calpe, "Semisupervised image classification with Laplacian support vector machines," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 3, pp. 336–340, Jul. 2008.

[2] E. Othman, Y. Bazi, F. Melgani, H. Alhichri, N. Alajlan, and M. Zuair, "Domain adaptation network for cross-scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4441–4456, Aug. 2017.

[3] H. Alhichri, "Multitask classification of remote sensing scenes using deep neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 1195–1198.

[4] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.

[5] W. Wang, Y. Chen, and P. Ghamisi, "Transferring CNN with adaptive learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5533918.

[6] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 270–294, Jun. 2022.

[7] G. Cheng et al., "SPNet: Siamese-prototype network for few-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5608011.

[8] Q. Zeng and J. Geng, "Task-specific contrastive learning for few-shot remote sensing image scene classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 191, pp. 143–154, 2022.

[9] C. Qiu, A. Yu, X. Yi, N. Guan, D. Shi, and X. Tong, "Open self-supervised features for remote-sensing image scene classification using very few samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2022, Art. no. 2500505.

[10] R. Gui, X. Xu, R. Yang, K. Deng, and J. Hu, "Generalized zero-shot domain adaptation for unsupervised cross-domain polSAR image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 270–283, 2021.

[11] H. Xu, Z. Bai, X. Zhang, and Q. Ding, "MFSANet: Zero-shot side-scan sonar image recognition based on style transfer," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 1503105.

[12] T. Itasaka and M. Okuda, "Zero-shot hyperspectral image denoising using self-completion with 3D random patterned masks," *IEEE Access*, vol. 11, pp. 79305–79314, 2023.

[13] U. Chaudhuri, R. Bose, B. Banerjee, A. Bhattacharya, and M. Datcu, "Zero-shot cross-modal retrieval for remote sensing images with minimal supervision," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4708615.

[14] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "A zero-shot sketch-based intermodal object retrieval scheme for remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 8007705.

[15] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning-the good, the bad and the ugly," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4582–4591.

[16] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 951–958.

[17] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," *Adv. Neural Inf. Process. Syst.*, vol. 26, pp. 1–10, 2013.

[18] A. Frome et al., "Devise: A deep visual-semantic embedding model," *Adv. Neural Inf. Process. Syst.*, vol. 26, pp. 1–9, 2013.

[19] M. Norouzi et al., "Zero-shot learning by convex combination of semantic embeddings," 2013, *arXiv:1312.5650*.

[20] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 819–826.

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

[22] 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.*, 2017, pp. 3279–3287.

[23] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 49–58.

[24] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 69–77.

[25] 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. 1–12.

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

[27] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.

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

[29] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[30] M. B. Sariyildiz and R. G. Cinbis, "Gradient matching generative networks for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2168–2178.

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

[32] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero-and few-shot learning via aligned variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8247–8255.

[33] Y. Liu, Y. Dang, X. Gao, J. Han, and L. Shao, "Zero-shot learning with attentive region embedding and enhanced semantics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 3, pp. 4220–4231, Mar. 2024.

[34] H. Huang, C. Wang, P. S. Yu, and C.-D. Wang, "Generative dual adversarial network for generalized zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 801–810.

[35] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[36] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[37] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.

[38] H. Huang, J. Chu, and F. Wei, "Unsupervised prompt learning for vision-language models," *ArXiv*, vol. abs/2204.03649, 2022, Art. no. 248006070. [Online]. Available: https://api.semanticscholar.org/CorpusID
[39] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16816–16825.
[40] H. Pham et al., "Combined scaling for zero-shot transfer learning," *Neurocomputing*, vol. 555, 2023, Art. no. 126658.
[41] M. Wortsman et al., "Robust fine-tuning of zero-shot models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7959–7971.
[42] Z. Novack, J. McAuley, Z. C. Lipton, and S. Garg, "Chils: Zero-shot image classification with hierarchical label sets," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 26342–26362.
[43] 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.
[44] 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.*, 2018, pp. 17–21.
[45] Y. Li, Z. Zhu, J.-G. Yu, and Y. Zhang, "Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10590–10603, Dec. 2021.
[46] C. Wang, G. Peng, and B. De Baets, "A distance-constrained semantic autoencoder for zero-shot remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12545–12556, 2021.
[47] Y. Li, D. Kong, Y. Zhang, and R. Xiao, "Zero-shot remote sensing image scene classification based on robust cross-domain mapping and gradual refinement of semantic space," *Acta Geodaetica et Cartographica Sinica*, vol. 49, no. 12, pp. 1–11, 2020.
[48] C. Wang et al., "Zero-shot remote sensing scene classification based on local-global feature fusion and weight mapping loss," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2763–2776, 2023.
[49] J. Shang, C. Niu, W. Zhou, Z. Zhou, and J. Yang, "Graph-based semantic embedding refinement for zero-shot remote sensing image scene classification," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 60, no. 1, pp. 644–657, Feb. 2024.
[50] 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, 2017.
[51] Y. Li, D. Kong, Y. Zhang, Y. Tan, and L. Chen, "Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 179, pp. 145–158, 2021.
[52] Y. Li, D. Kong, Y. Zhang, R. Chen, and J. Chen, "Representation learning of remote sensing knowledge graph for zero-shot remote sensing image scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 1351–1354.
[53] W. Xu, J. Wang, Z. Wei, M. Peng, and Y. Wu, "Deep semantic-visual alignment for zero-shot remote sensing image scene classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 198, pp. 140–152, 2023.
[54] S. Ma, C. Liu, Z. Li, and W. Yang, "Integrating adversarial generative network with variational autoencoders towards cross-modal alignment for zero-shot remote sensing image scene classification," *Remote Sens.*, vol. 14, no. 18, pp. 1–18, 2022.
[55] C. Liu, S. Ma, Z. Li, W. Yang, and Z. Han, "Mining contrastive relations between cross-modal features for zero-shot remote sensing image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 5503905.
[56] Z. Li, D. Zhang, Y. Wang, D. Lin, and J. Zhang, "Generative adversarial networks for zero-shot remote sensing scene classification," *Appl. Sci.*, vol. 12, no. 8, pp. 1–16, 2022.
[57] N. Fei et al., "Towards artificial general intelligence via a multimodal foundation model," *Nature Commun.*, vol. 13, no. 1, pp. 1–13, 2022.
[58] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, and J. Zhou, "RemoteCLIP: A vision language foundation model for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024, doi: 10.1109/TGRS.2024.3390838.
[59] S. Mo, M. Kim, K. Lee, and J. Shin, "S-CLIP: Semi-supervised vision-language learning using few specialist captions," in *Proc. 37th Conf. Neural Inf. Process. Syst.*, 2023, pp. 1–26.
[60] X. Li, C. Wen, Y. Hu, and N. Zhou, "RS-CLIP: Zero shot remote sensing scene classification via contrastive vision-language supervision," *Int. J. Appl. Earth Observation Geoinformation*, vol. 124, 2023, Art. no. 103497.
[61] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "Geochat: Grounded large vision-language model for remote sensing," 2023, *arXiv:2311.15826*.
[62] M. M. Al Rahhal, Y. Bazi, H. Elgibreen, and M. Zuair, "Vision-language models for zero-shot classification of remote sensing images," *Appl. Sci.*, vol. 13, no. 22, 2023, Art. no. 12462.
[63] B. Zhao, Y. Zhong, G.-S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.
[64] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.
[65] Z. Xiao, Y. Long, D. Li, C. Wei, G. Tang, and J. Liu, "High-resolution remote sensing image retrieval based on CNNs from a dimensional perspective," *Remote Sens.*, vol. 9, no. 7, pp. 1–24, 2017.
[66] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 173–176, Jan. 2011.
[67] L. Zhao, P. Tang, and L. Huo, "Feature significance-based multi-bag-of-visual-words model for remote sensing image scene classification," *J. Appl. Remote Sens.*, vol. 10, no. 3, pp. 035004–035004, 2016.
[68] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *Proc. IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.
[69] 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.*, 2010, pp. 270–279.
[70] 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.
[71] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE Proc. IRE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[72] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 197–209, 2018.
[73] H. Li et al., "RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data," *Sensors*, vol. 20, no. 6, pp. 1–26, 2020.

Xiaomeng Tan received the B.E. degree in process equipment and control engineering from the Beijing University of Chemical Technology, Beijing, China, in 2019. She is currently working toward the Ph.D. degree in computer application technology with the University of Chinese Academy of Sciences, Beijing, China.

Her research interests include multimodal deep learning and zero-shot learning.

Bobo Xi (Member, IEEE) received the B.E. degree in information engineering and the Ph.D. degree in information and communication engineering from Xidian University, Xi'an, China, in 2017 and 2022, respectively.

He is currently a Lecturer with the State Key Laboratory of Integrated Services Networks, School of Telecommunications, Xidian University. He has published over 20 papers in refereed journals, including IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. His research interests include hyperspectral image processing, machine learning, and deep learning.

Jiaojiao Li (Senior Member, IEEE) received the B.E. degree in computer science and technology, the M.S. degree in software engineering, and the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in 2009, 2012, and 2016, respectively.

She was an exchange Ph.D. student of Mississippi State University, supervised by Dr. Qian Du. She is currently an Associate Professor and the doctoral supervisor with the State Key Laboratory of Integrated Service Networks, School of Telecommunications, Xidian University. Her research interests include hyperspectral remote sensing image analysis and processing, pattern recognition, and data compression.

Tie Zheng received the B.E. degree in electrical engineering and automation from Northeast Agricultural University, Harbin, China, and the Ph.D. degree in computer application technology from the University of the Chinese Academy of Sciences, Beijing, China, in 2017 and 2023, respectively.

He is currently with the National Space Science Center, Chinese Academy of Sciences, with major research on space signal processing and remote sensing techniques.

Yunsong Li (Member, IEEE) received the M.S. degree in telecommunication and information systems and the Ph.D. degree in signal and information processing from Xidian University, Xi'an, China, in 1999 and 2002, respectively.

He joined the School of Telecommunications Engineering, Xidian University, in 1999, where he is currently a Professor. He is also the Director of the State Key Laboratory of Integrated Service Networks, Image Coding and Processing Center. His research interests include image and video processing, hyperspectral image processing, and high-performance computing.

Changbin Xue received the M.S. degree in communications and electronic systems from Harbin Institute of Technology, Harbin, China, and the Ph.D. degree in condensed matter physics from Beijing Institute of Technology, Beijing, China, in 1997 and 2017, respectively.

He has been engaged in the design and development of complex space electronic information systems and space science exploration payload systems for more than 20 years. He was the Chief Designer of the Chang'e-4 payload system and the Chief Designer of the space science pilot project "Shijian-10 Return Science Experiment Satellite." He has published more than 30 academic papers and more than 10 authorized patents. His research interests include full process design and simulation technology of deep space exploration scientific payload, machine learning, and deep learning.

Jocelyn Chanussot (Fellow, IEEE) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology, Grenoble, France, in 1995, and the Ph.D. degree in electrical engineering from the Université de Savoie, Annecy, France, in 1998.

Since 1999, he has been with Grenoble INP, where he is currently a Professor of signal and image processing. He has been a Visiting Scholar with Stanford University, Stanford, CA, USA; KTH, Stockholm, Sweden; and NUS, Singapore. Since 2013, he has also been an Adjunct Professor with the University of Iceland, Reykjavik, Iceland. From 2015 to 2017, he was a Visiting Professor with the University of California at Los Angeles (UCLA), Los Angeles, CA, USA. He holds the AXA Chair in remote sensing and is an Adjunct Professor with the Chinese Academy of Sciences, Aerospace Information research Institute, Beijing, China. He has coauthored more than 250 articles in international journals, gathering 31 500+ citations, with H-index = 78. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence.

Dr. Chanussot is an ELLIS Fellow and a fellow of the Asia-Pacific Artificial Intelligence Association, was a Member of the Institut Universitaire de France from 2012 to 2017, and has been a Highly Cited Researcher by Clarivate Analytics/Thomson Reuters since 2018. He is the Founding President of the IEEE Geoscience and Remote Sensing French Chapter from 2007 to 2010, which received the 2010 IEEE GRS-S Chapter Excellence Award. He has received multiple outstanding paper awards. He was the Vice President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia from 2017 to 2019. He was the General Chair for the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing. He was the Chair from 2009 to 2011 and the Co-Chair for the GRS Data Fusion Technical Committee from 2005 to 2008. He was a Member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society from 2006 to 2008 and the Program Chair for the IEEE International Workshop on Machine Learning for Signal Processing in 2009. He is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, and the Proceedings of the IEEE. He was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2011 to 2015. In 2014, he was a Guest Editor for the IEEE SIGNAL PROCESSING MAGAZINE.