Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification

Article *in* ISPRS Journal of Photogrammetry and Remote Sensing · September 2021

DOI: 10.1016/j.isprsjprs.2021.08.001

CITATIONS
83

READS
1,201

5 authors, including:

Yansheng Li
Wuhan University
111 PUBLICATIONS **3,869** CITATIONS
SEE PROFILE

Yihua Tan
Huazhong University of Science and Technology
95 PUBLICATIONS **1,398** CITATIONS
SEE PROFILE

ISPRS Journal of Photogrammetry and Remote Sensing 179 (2021) 145–158

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification

Yansheng Li [a], Deyu Kong [a], Yongjun Zhang [a,*], Yihua Tan [b,c,*], Ling Chen [d,*]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[b] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[c] Shenzhen Huazhong University of Science and Technology Research Institute, Shenzhen 518000, China
[d] College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Keywords:
Robust deep alignment network (DAN)
Remote sensing knowledge graph (RSKG)
Remote sensing image scene classification
Zero-shot learning (ZSL)
Generalized zero-shot learning (GZSL)

ABSTRACT

Although deep learning has revolutionized remote sensing (RS) image scene classification, current deep learning-based approaches highly depend on the massive supervision of predetermined scene categories and have disappointingly poor performance on new categories that go beyond predetermined scene categories. In reality, the classification task often has to be extended along with the emergence of new applications that inevitably involve new categories of RS image scenes, so how to make the deep learning model own the inference ability to recognize the RS image scenes from unseen categories, which do not overlap the predetermined scene categories in the training stage, becomes incredibly important. By fully exploiting the RS domain characteristics, this paper constructs a new remote sensing knowledge graph (RSKG) from scratch to support the inference recognition of unseen RS image scenes. To improve the semantic representation ability of RS-oriented scene categories, this paper proposes to generate a Semantic Representation of scene categories by representation learning of RSKG (SR-RSKG). To pursue robust cross-modal matching between visual features and semantic representations, this paper proposes a novel deep alignment network (DAN) with a series of well-designed optimization constraints, which can simultaneously address zero-shot and generalized zero-shot RS image scene classification. Extensive experiments on one merged RS image scene dataset, which is the integration of multiple publicly open datasets, show that the recommended SR-RSKG obviously outperforms the traditional knowledge types (e.g., natural language processing models and manually annotated attribute vectors), and our proposed DAN shows better performance compared with the state-of-the-art methods under both the zero-shot and generalized zero-shot RS image scene classification settings. The constructed RSKG will be made publicly available along with this paper (https://github.com/kdy2021/SR-RSKG).

1. Introduction

Benefiting from the rapid advances in aerospace, sensor and communication technologies, human beings have entered an era of remote sensing (RS) big data (Chi et al., 2016; Li et al., 2021a; Lobry et al., 2020). Automatically accurate classification of these oversized RS images is one basic but important task for mining the value of RS big data (Cheng et al., 2016; Gu et al., 2019; Li et al., 2020; Marcos et al., 2018). Along with the spatial resolution improvement of RS imagery, pixel-level or object-level classification methods show great limitations (Blaschke 2010; Li et al., 2016; Cheng et al., 2017). As a consequence, more attention has been given to scene-level RS image classification due to its stable classification performance and its wide applications in natural disaster monitoring (Cheng et al., 2013), multimodal data fusion (Gerke et al., 2014), functional zone classification (Zhang et al., 2018), object detection (Tao et al., 2019a; Tao et al., 2019b), and image retrieval (Demir and Bruzzone, 2015; Li et al., 2018).

Until now, deep learning (LeCun et al., 2015) has greatly improved

* Corresponding authors at: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (Y. Zhang). School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China. Shenzhen Huazhong University of Science and Technology Research Institute, Shenzhen 518000, China (Y. Tan). College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China (L. Chen).
E-mail addresses: zhangyj@whu.edu.cn (Y. Zhang), yhtan@hust.edu.cn (Y. Tan), lingchen@cs.zju.edu.cn (L. Chen).

https://doi.org/10.1016/j.isprsjprs.2021.08.001
Received 10 February 2021; Received in revised form 13 June 2021; Accepted 2 August 2021
Available online 10 August 2021

RS image scene classification (Li et al., 2021c; Zhang et al., 2015). However, the current deep learning models have good classification performance only when each scene category has sufficient samples. In the era of RS big data, the number of RS scene categories presents an explosive growth trend. It is unrealistic to collect sufficient RS image samples and construct their labels for all categories at once. Hence, identifying RS image scenes that never appear in the training stage has important practical value (Li et al., 2017a) in the era of RS big data. Inspired by humans' inference ability, embedding prior knowledge into the learning process is an ideal method for addressing this issue (Li et al., 2021b).

In the literature, the development of zero-shot learning (ZSL) (Larochelle et al., 2008; Palatucci et al., 2009; Ji et al., 2020) in recent years has provided promising solutions to recognize samples from unseen categories. By leveraging the prior knowledge of categories, including seen and unseen categories, as auxiliary information, ZSL can learn samples from seen categories to identify samples from the unseen categories. Generally, the semantic information of seen and unseen classes is the common sense of human beings, which is universal and can be used in both of the training and testing stages, but the image samples of unseen classes do not exist in the training stage. Hence, how to express semantics is the key to pursue the superior performance of ZSL. For example, we can recognize the zebra image through the images of tiger, panda and horse, combined with the semantic information such as tiger stripes, panda colors and horse shapes. From this intuitive finding, we can also see the indispensable importance of semantic information in the ZSL task. As an extension of ZSL, generalized zero-shot learning (GZSL) attempts to learn samples from seen categories to simultaneously recognize seen and unseen samples in the testing stage, which is a more challenging but practical task. In the field of computer vision, large numbers of ZSL and GZSL methods have been proposed. In contrast, ZSL and GZSL are rarely discussed in the field of RS (Sumbul et al., 2017). Compared with the computer vision field, the following characteristics in the RS field limit the development of ZSL and GZSL. On the one hand, the names of RS scene categories often have domain specificity. If the semantic representations of RS scene categories are generated by directly leveraging the general natural language processing model (e.g., Word2Vec) to map the names of RS scene categories, the semantic representations cannot reflect the intrinsically semantic information of the RS category. On the other hand, RS image scenes, presenting large intraclass differences and large interclass similarities, generally have more complex appearances than natural images in the computer vision field. Generally, the ZSL and GZSL methods that have achieved excellent results in the field of computer vision cannot be directly extended to address the task in the RS domain. Overall, it deserves much more exploration to promote zero-shot and generalized zero-shot RS image scene classification.

With the aforementioned considerations, this paper mainly focuses on exploiting zero-shot and generalized zero-shot RS image scene classification. The quality of semantic representation of categories plays an important role in ZSL and GZSL (Li et al., 2017a,b,c). To generate the high-quality semantic representations of RS scene categories, this paper constructs a new remote sensing knowledge graph (RSKG) based on the domain prior knowledge from human experts, where RSKG fully considers the rich connections between RS scene elements. To the best of our knowledge, this paper, for the first time, proposes to calculate the Semantic Representations of RS scene categories by representation learning of RSKG (SR-RSKG). Based on SR-RSKG, this paper proposes a new deep alignment network with a series of well-designed constraints (DAN), which can robustly match the visual features and semantic representations in the latent space, to address zero-shot and generalized zero-shot RS image scene classification. Experimental results on one integrated RS image scene dataset show that our proposed SR-RSKG is superior to traditional knowledge types (e.g., Word2Vec (Mikolov et al., 2013), BERT (Devlin et al., 2018), and manually annotated attribute vectors). In addition, the proposed DAN performs better than the state-of-the-art methods under both the ZSL and GZSL settings. The major contributions of this paper are summarized as follows.

1) To the best of our knowledge, this paper, for the first time, proposes to generate the semantic representations of RS scene categories by representation learning of RSKG. Extensive experiments verify its superiority compared with traditional prior knowledge types. The constructed RSKG will be made publicly available along with this paper.
2) By pursuing the stable cross-modal alignment of the same category and scattered distribution of different categories, this paper proposes a novel DAN to robustly match visual features and semantic features in the latent space. Extensive experiments show that the proposed DAN outperforms the existing methods under both the ZSL and GZSL settings.

The remainder of this paper is organized as follows. Section 2 discusses the related works. Section 3 introduces the construction process of RSKG and depicts representation learning of RSKG. Section 4 introduces the DAN model in detail. Section 5 summarizes the experimental results. Finally, the conclusion is detailed in Section 6.

2. Related work

In this section, we briefly review the most relevant works in the literature that include semantic representations of RS scene categories and zero-shot RS image scene classification.

2.1. Semantic representations of remote sensing scene categories

To obtain prior knowledge for ZSL, there are two main methods: class name embedding by a natural language processing model and manual annotation attribute vectors. However, for the RS domain, both of them have insurmountable defects. Natural language processing models based on generalized corpora have a weak pertinence and often fail to finely describe RS scene categories. Although manual annotation attribute vectors are often generated by considering the specific scene categories, they also fail to consider the rich connections between different scene categories. Therefore, it is urgent to explore a priori knowledge acquisition method that has a strong pertinence for the RS domain.

Along with the rapid development of artificial intelligence (AI), the tremendous success of knowledge graph (KG) has attracted much attention. The concept of a KG originates from Tim Berners-Lee's vision of the semantic web (Shadbolt et al., 2006), where a KG mainly aims to use graphs to build relations between objects in the real world. More specifically, a KG uses nodes to represent objects in the real world and edges to represent relations between objects. To promote the application of KG, representation learning (Bordes et al., 2013) is proposed to learn the low-dimensional vectors of the entities and relations in the KG. Unlike the representation learning in self-supervised learning approaches such as SimCLR (Chen et al., 2020), MoCO (He et al., 2020), PIRL (Misra et al., 2020), which mainly aims to learn the visual feature representations of images, the representation learning based on KG aims to learn the semantic representation of entities and relations in the KG. The learned representation vector contains semantic information, so the information in the KG can be extracted and used more conveniently in wide downstream tasks. Intuitively, the semantic representations generated through KG representation learning are very suitable as prior knowledge for ZSL. Unfortunately, there is no mature KG in the RS field. Therefore, how to construct RSKG and explore appropriate representation learning methods to generate semantic representations of RS scene categories has become very urgent.

2.2. Zero-shot learning in remote sensing scene classification

In the computer vision field, early ZSL methods mainly focus on

learning mapping from visual space to semantic space or from semantic space to visual space. For example, direct attribute prediction (DAP) and indirect attribute prediction (IAP), proposed in (Lampert et al., 2009), learn the mapping from the visual space to the semantic space. To solve the hubness problem caused by the mapping of visual space to semantic space, Shigeto et al. 2015) proposed mapping semantic representations to visual feature space. With the rapid development of generative models, some ZSL methods based on visual sample generation have emerged. Long et al. 2017 proposed a method for generating unseen class samples by adding diffusion regularization to ensure that the generated invisible samples retain as much of the semantic structure information as possible. Inspired by the conditional generative adversarial network (GAN), Xian et al. (2018b) used category attributes as input to generate visual features corresponding to the category and added the classification loss of the category to the discriminator. Elhoseiny and Elfeki (2019) introduced hallucinated text in the process of training the generator to encourage the generated visual features to deviate from the seen category and pursue the diversity of the generated samples. Schonfeld et al. 2019 recommended reconstructing semantic and visual features through a variational autoencoder (VAE) so that the constructed features contain basic multimodal information related to unseen classes. Unfortunately, RS image scenes have the characteristics of complex content, unique viewing angle, high similarity of images between classes, and diversity of images within the same class. The methods in the field of computer vision cannot be directly extended to address RS classification.

In the RS field, Li et al. 2017b proposed the first study on zero-shot RS image scene classification by leveraging the visual similarity among images from the same scene class and the label refinement approach based on sparse learning. Quan et al. 2018 employed a semi-supervised Sammon embedding algorithm to modify semantic space prototypes to have a more consistent class structure with visual space prototypes. Sumbul et al., 2017 proposed addressing fine-grained object recognition with ZSL in remotely sensed imagery and showed how the compatibility function can be estimated from the seen classes by using the maximum likelihood principle during the learning phase. Li et al. (2021b) proposed a zero-shot RS image scene classification method based on locality-preservation deep cross-modal embedding networks. Overall, the existing works ignore the diversity of RS images and learn the fixed mapping between visual space and semantic space. In addition, the existing works still use natural language models based on generalized corpora to obtain semantic representations of RS scene categories, which makes the performance of the existing approaches unsatisfactory. Furthermore, there are very few studies involving GZSL in the RS field. If the ZSL method is directly extended to GZSL, the classification accuracy of unseen classes will generally be much lower than that of seen classes.

3. Representation learning of remote sensing knowledge graph

In this section, we first introduce the construction process of RSKG and then discuss representation learning of RSKG.

3.1. Construction of remote sensing knowledge graph

In the literature, there exist many general KG, including Freebase (Bollacker et al., 2008), WikiData (Erxleben et al., 2014), DBpedia (Auer et al., 2007) and Yago (Hoffart et al., 2013). Generally, KG contain abundant explicit relational information, which is very beneficial for describing complex RS scenes. However, the current general KG are not applicable in the RS field. To support zero-shot RS image scene classification, we build a new KG (i.e., RSKG) based on RS scene elements. It is worth noting that the RS scene is not just a collection of objects but also contains rich relations between interconnected objects. An accurate description of the relation between objects can effectively promote the ability of deep learning methods to understand the semantics of RS scenes (Liang et al., 2019). By combining the characteristics of RS image content and the related research on geographic spatial relations (Clementini, 2009; Shen et al., 2017), we define the relations in RSKG into the following categories. We first divide relations into two categories: attribute relations and spatial relations. Attribute relations are used to describe the characteristics of the object or the child-parent relation with other objects. Attribute relations can be subdivided into data relations and object relations, among which data relations include shape, color, width, distribution, and height, and object relations include has, component of, part of, and member of. Spatial relations mainly describe the different location relations between different objects in space. Spatial relations can be subdivided into position relations, topological relations and vague relations, among which position relations include marked on, dock at, stop at, over, and on; topological relations include surrounded by, intersect at, pass through, meet, connect to, cover, contain, and in, and vague relations include near, next to, around and along. More specifically, the frequency of relation types is visually shown in Fig. 2.

In our implementation, 10 domain experts, who are familiar with RS image interpretation tasks, participate in constructing the RSKG. More specifically, we first determine the relation between the objects in the RS images from one specific RS category and store it as an entity-relation-entity or entity-attribute-attribute value to construct the category-specific KG. Then, we analyze the relation between different RS image scenes and aggregate the category-specific KG to form a unified RSKG, as shown in Fig. 1. More specifically, blue represents the entity, and yellow represents the attribute value. The current version of RSKG has 117 entities, 26 relations and 191 triples. As a first attempt towards the RSKG construction, the volume of the current RSKG is relatively small, but extensive visual and quantitative experiments show its superior knowledge representation ability because it has contained the intrinsic RS elements and geographic relations. Similar to other KGs, our constructed RSKG can also be easily extended after its public release.

3.2. Learning semantic representations of entities and relations

Inspired by the phenomenon that word vectors have translation invariance in the semantic space (Mikolov et al., 2013), Bordes et al. proposed the classic representation learning model TransE (Bordes et al., 2013). For each triple (h, r, t) in the KG, the TransE model assumes $\mathbf{c_h}+\mathbf{c_r} \approx \mathbf{c_t}$ where $\mathbf{c_h}$, $\mathbf{c_r}$ and $\mathbf{c_t}$ stand for the semantic representations in the unified feature space of h, r and t. The assumption $\mathbf{c_h}+\mathbf{c_r} \approx \mathbf{c_t}$ means that the head entity vector plus the relation vector is approximately equal to the tail entity vector. The TransE model has significant effects on datasets such as WordNet and Freebase and has become a classic model in the field of representation learning. However, TransE cannot deal with the complex relations such as 1-N or N-1 appearing in the KG. Unfortunately, this situation is unavoidable in the RSKG. For example, <commercial_area, have, tree>, <commercial_area, have, buildings>; if the TransE model is used for representation learning, the corresponding embedding vectors of trees and buildings cannot be effectively distinguished. Considering the limitations of the TransE model in the face of complex relations such as 1-N and N-1, we recommend using the improved representation learning model TransH (Wang et al., 2014), which flexibly models a relation as a hyperplane together with a translation operation on it and improves the performance by handling complex relations to a certain extent. As shown in Fig. 3, $c_h \in \mathbb{R}^d$, $c_r \in \mathbb{R}^d$, $c_t \in \mathbb{R}^d$ is given, where d denotes the dimension of the embedding semantic representation vector, TransH maps c_h and c_t to the hyperplane, let $c_{h_\perp} = c_h - w_r^T c_h w_r$ and $c_{t_\perp} = c_t - w_r^T c_t w_r$, where w_r is the norm vector of the hyperplane, and then the translation operation let $c_{h_\perp} + c_r \approx c_{t_\perp}$. Thus, TransH can learn complex relations well. More specifically, the objective function of TransH is formulated by Eq. (1).

$$f_r(h,t) = \|c_{h_\perp} + c_r - c_{t_\perp}\|_2^2 \quad (1)$$

By pushing the correct triple to get a lower score in Eq. (1), the

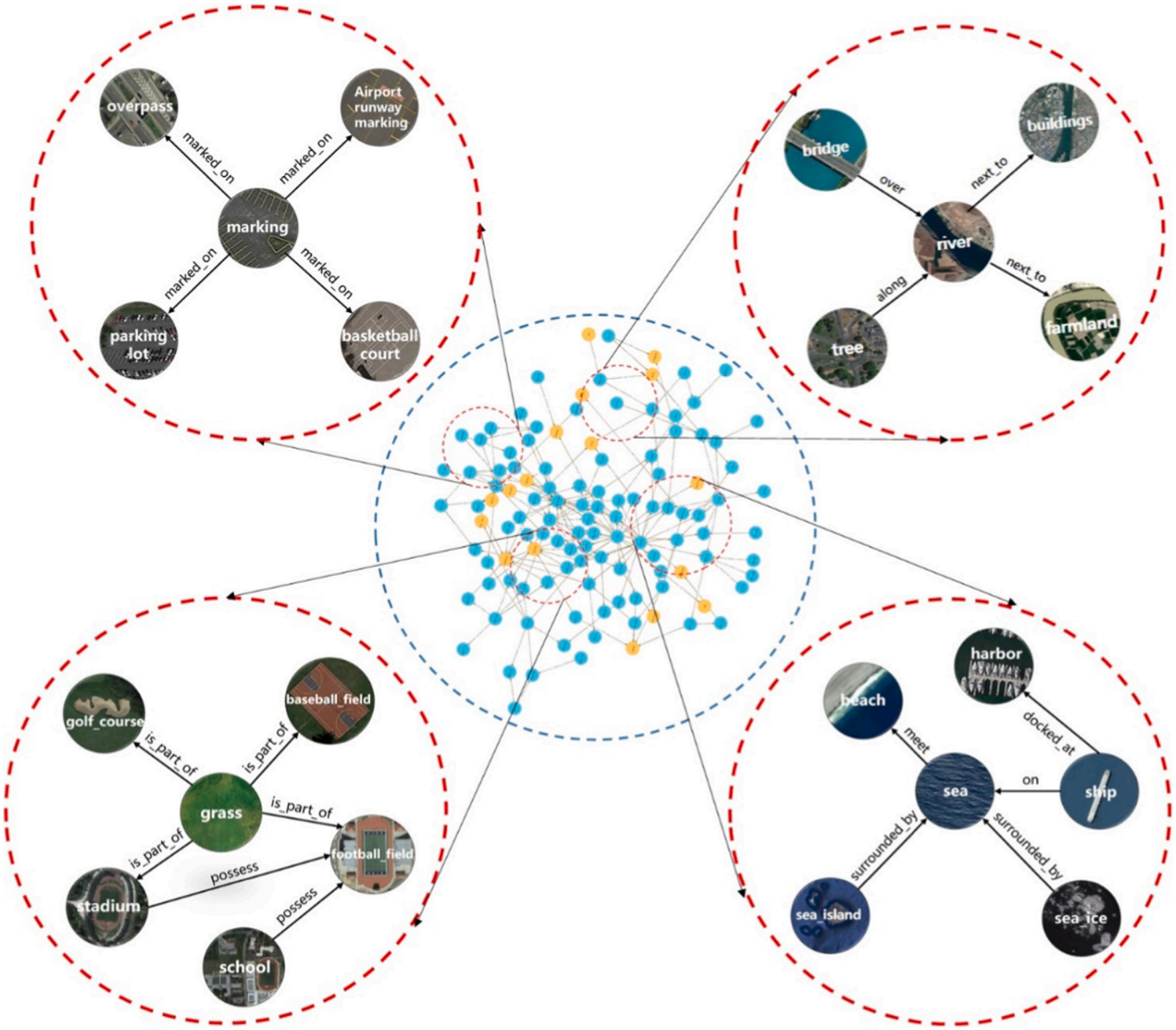


Fig. 1. Visual illustration of the constructed RSKG.

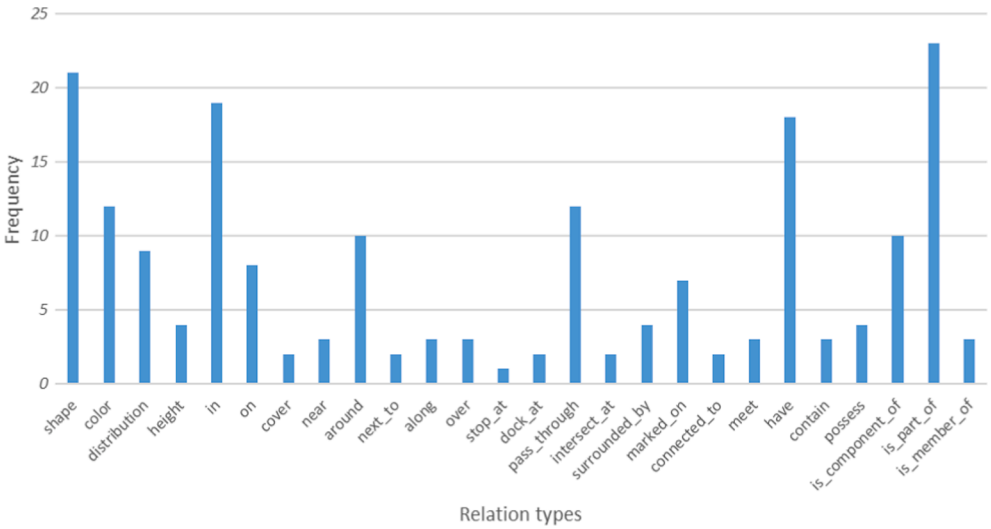
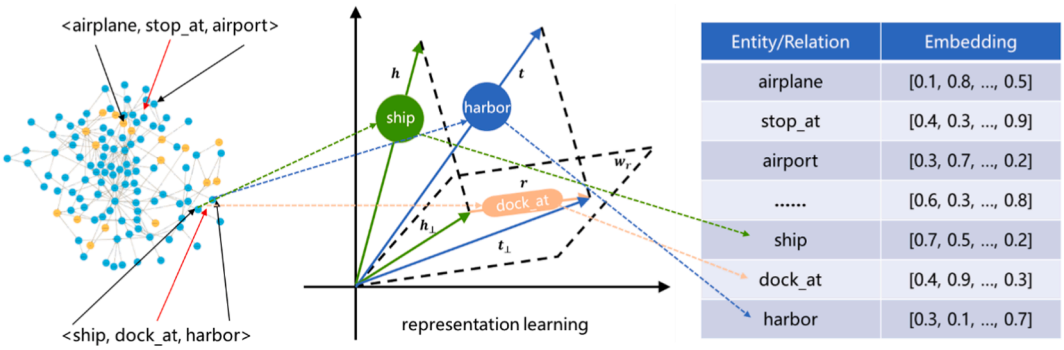


Fig. 2. Frequency of relation types in the RSKG.



Entity/Relation	Embedding
airplane	[0.1, 0.8, ..., 0.5]
stop_at	[0.4, 0.3, ..., 0.9]
airport	[0.3, 0.7, ..., 0.2]
.....	[0.6, 0.3, ..., 0.8]
ship	[0.7, 0.5, ..., 0.2]
dock_at	[0.4, 0.9, ..., 0.3]
harbor	[0.3, 0.1, ..., 0.7]

Fig. 3. Visual illustration of representation learning of RSKG.

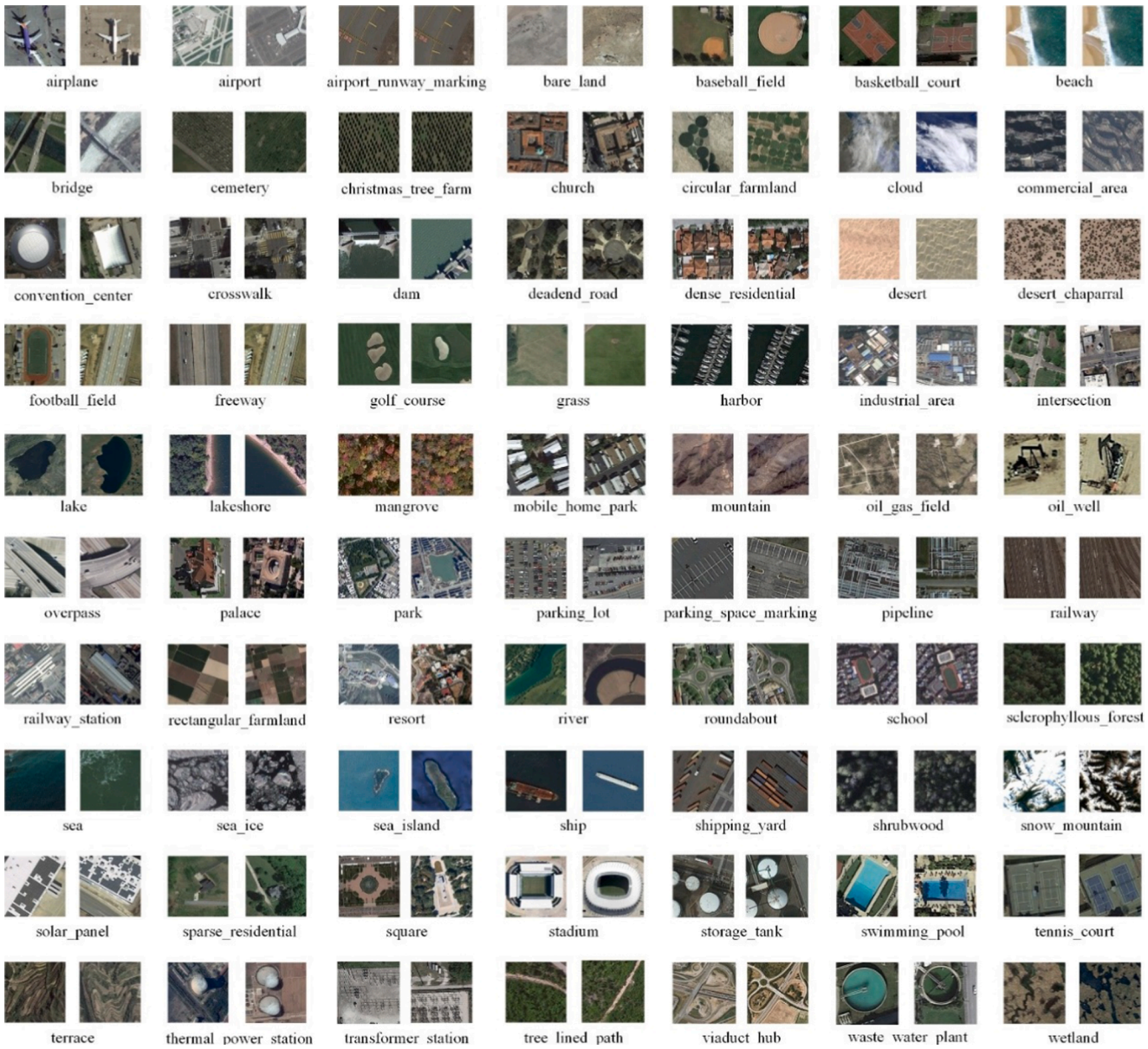


Fig. 4. The merged RS image scene dataset.

objective function in Eq. (1) can be specifically optimized by minimizing the loss function in Eq. (2).

$$L_{TransH} = \sum_{(h,r,t)\in\Delta} \sum_{(h',r',t')\in\Delta'} max(f_r(h,t) + \tau - f_{r'}(h',t'), 0) \quad (2)$$

where Δ is the set of correct triples and Δ' denotes the set of wrong triples. τ indicates the minimum interval between the scores of correct triples and wrong triples, and is usually set to 1.

More optimization details can refer to (Wang et al., 2014). By optimizing the objective function in Eq. (1), we can obtain the SR-RSKG.

3.3. Creating semantic representations of remote sensing scene categories

To fully evaluate the performance of zero-shot and generalized zero-shot RS image scene classification, we adopt a combined dataset by integrating five public datasets: UCM (Yang and Newsam, 2011), AID (Xia et al., 2017), NWPU-RESISC45 (Cheng et al., 2017), RSI-CB256 (Li et al., 2017b), and PatternNet (Zhou et al., 2018). The merged RS image scene dataset is composed of 70 scene categories, and each category contains 800 image scenes with a size of 256 × 256. The image scene dataset is visually shown in Fig. 4.

As aforementioned, the construction of RSKG considers as more details about RS objects and scene categories as possible, so the entities in the RSKG often cover the scene categories in the specific dataset. In short, the scene categories in the specific task can find their corresponding entities in RSKG. Therefore, the learned semantic representations of entities in Section 3.2 can be used to generate the semantic representations of scene categories. To facilitate understanding, we explain it from the formulation perspective in the following. Let $Y = \{y_1, y_2, \cdots, y_M\}$ denote the label set of RS scene categories, where M denotes the number of scene categories of the dataset. For each label $y_i \in Y$, there is a one correspondence between the entity in RSKG and y_i (i.e., the scene category), we take the corresponding semantic representation of entity as its semantic representation $c_i \in C$.

It is noted that the entities in the constructed RSKG include but are not limited to the scene categories of the adopted RS scene classification dataset in this paper. Hence, other RS scene classification tasks can also be flexibly implemented by the RSKG as long as the scene categories can find the entity or synonym from the RSKG. We will continue to improve and expand the RSKG in the future. For example, the follow-up work may automatically extend the RSKG by the newly proposed knowledge graph construction method (Tempelmeier et al., 2021), so that it will gradually have the coverage ability of the RS field. It can be applied to more RS tasks. This work mainly aims to verify the rationality and effectiveness of the new kind of knowledge. Generally, the performance of ZSL and GZSL can get further improved along with the continuous expansion of RSKG. Objectively, the number of entities in RSKG is empirically set without the sufficient experimental verification. In the follow-up works, the sensitivity analysis of the number of entities in the RSKG to the final performance is also very necessary.

To visually show the superiority of SR-RSKG, we visualize various semantic representations including Word2Vec, BERT, Attribute and SR-RSKG, through t-SNE (Maaten and Hinton, 2008) in Fig. 5. As Word2vec and BERT are two widely used models in the natural language processing domain and adopted in the ZSL task in computer vision, they are selected as baselines in this paper. Considering that there does not exist any natural language corpus, specially set up for RS at present, we have to adopt the pre-trained Word2vec and BERT models with the general

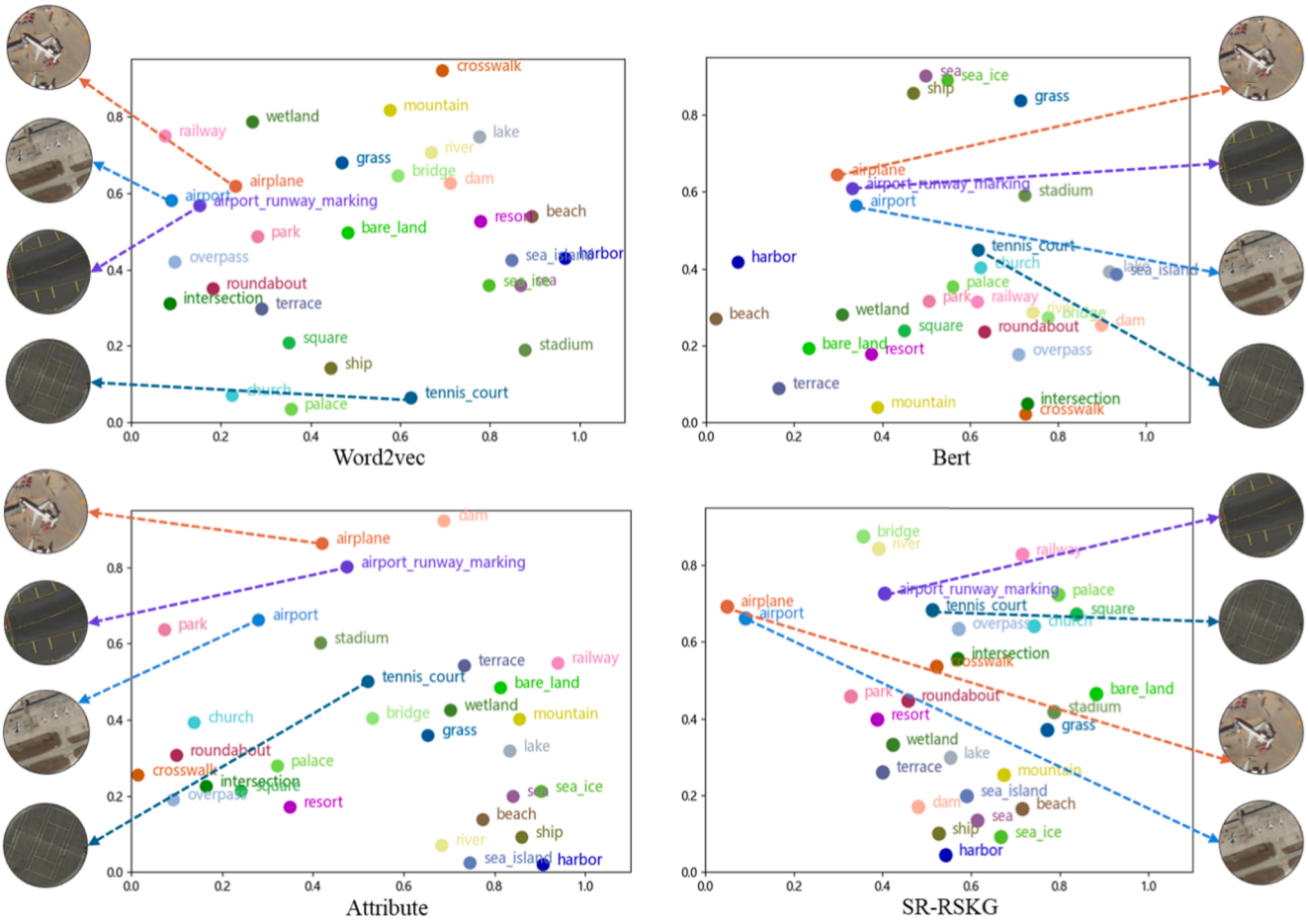


Fig. 5. Visualization of different knowledge types.

natural language corpus to generate semantic representations of scene categories. In the follow-up work, we may try to collect the natural language corpus in the RS domain to train the Word2vec and BERT models from scratch, which will be the objectively competitive baselines. Attribute is created by manual annotation, and its potential attribute elements are refined as much as possible to pursue the best performance. Hence, Attribute should be a competitive baseline of this kind of manual method.

Generally, Word2Vec, BERT and Attribute focus on depicting abstract semantics but ignore visual details. Taking the airport runway marking category as an example, in the semantic space, it is much closer to airports and airplanes in terms of Word2Vec, BERT and Attribute. In addition to semantic considerations, the SR-RSKG further considers the degree of visual relevance. As far as the illustration of SR-RSKG, the position of the airport runway marking in the semantic space is closer to the tennis court that has a certain visual similarity to it (both of them have marking elements). Benefiting from this merit, SR-RSKG is more suitable for zero-shot and generalized zero-shot RS image scene classification than Word2Vec, BERT and Attribute.

4. Robust deep alignment network for zero-shot and generalized zero-shot remote sensing image scene classification

Section 4.1 introduces the definition of ZSL and GZSL. In Section 4.2, we clarify the robust deep alignment network for zero-shot and generalized zero-shot RS image scene classification. In addition, we introduce the process of classifying RS image scenes from unseen categories.

4.1. Formulated definition of ZSL and GZSL

The task of ZSL is defined as follows. Let $D^s = \{(x_i^s, y_i^s, c(y_i^s))|i = 1, 2, \cdots, N\}$ denote the set of training examples (i.e., the seen samples). More specifically, $x_i^s \in X^s$ denotes the visual image feature of the i-th RS image scene from the seen categories where the image feature is extracted by the CNN model. $y_i^s \in Y^s$ denotes the label of the i-th RS image scene from the seen categories, and $c(y_i^s) \in C^s$ denotes the semantic representation (e.g., SR-RSKG) of the corresponding category. N represents the number of training samples. In the same way, we define X^u, Y^u, C^u as the unseen visual image features, the corresponding labels and the semantic representations of unseen categories. As well known, for ZSL and GZSL, the seen classes and unseen classes are disjoint, i.e. $Y^s \cap Y^u = \varnothing$. Given the training datasets D^s and $\{Y^u, C^u\}$, in the conventional ZSL, the task is to learn a classifier $F_{ZSL}: X^u \rightarrow Y^u$. In GZSL, the task is to learn a classifier $F_{GZSL}: X^s \cup X^u \rightarrow Y^s \cup Y^u$.

4.2. Robust deep alignment network in the latent space

Instead of learning the mapping from visual space to semantic space or from semantic space to visual space, we learn the mapping of visual features and semantic representations in latent space so that we can alleviate the hubness problem (Lazaridou et al., 2015) in ZSL and enhance visual-semantic coupling. In Fig. 6, we show an overview of our model. First, we minimize the visual and semantic representation reconstruction loss. Then, we align the distribution of vision and semantics in the hidden space, which further separates the distribution of features between different categories on the basis of aligning visual features and semantic representations, which improves the performance in ZSL tasks. In addition, the method is based on latent space mapping and the method of generating training samples to train the classifier, which balances the classification performance of seen and unseen categories, so it also has excellent performance in GZSL tasks. It is noted that the mentioned deep alignment network intrinsically tries to address the existing coordinated representation problem (Guo et al., 2019) in literature. Considering the visual feature and semantic representation alignment in latent space and multicategory distribution dispersion, the overall loss of the proposed model is defined as Eq. (3). In the rest of this section, we will introduce the details of each module.

$$\mathcal{L} = \mathcal{L}_{VAE} + \alpha\mathcal{L}_{CMFR} + \beta\mathcal{L}_{VSDM} + \gamma\mathcal{L}_{MCDD} \tag{3}$$

where α, β and γ are the weighting factors of the cross-modal feature reconstruction loss, visual and semantic distribution matching loss, and multicategory distribution dispersion loss, respectively.

4.2.1. Visual feature and semantic representation reconstruction

As our proposed method learns the mapping of visual features and semantic representations in the latent space, we first need to ensure the

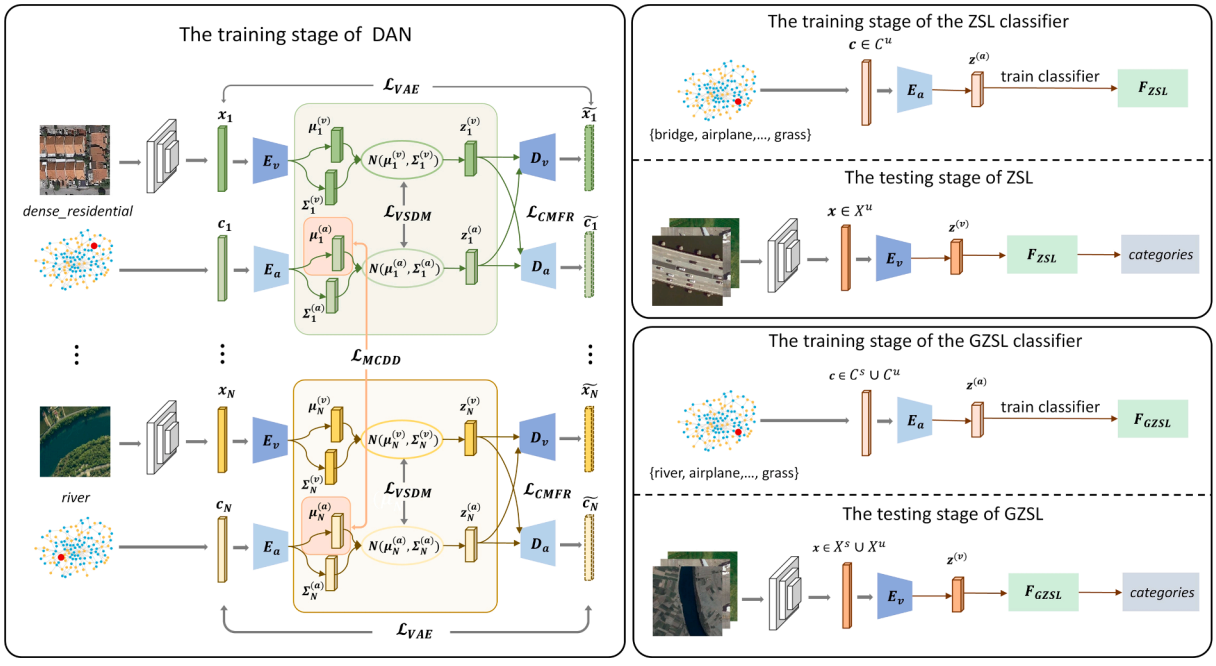


Fig. 6. Flowchart of the proposed DAN. In the training phase of ZSL and GZSL, we only use seen data to train the DAN. In the testing phase of ZSL, the testing data only comes from unseen data. In the testing phase of GZSL, the testing data comes from seen data and unseen data.

representation ability of each modality in the latent space. In addition, to minimize the loss of information, the original data should be reconstructed as much as possible using the latent vector. Therefore, we follow the architecture of the VAE net (Kingma and Welling, 2013) to learn the reconstruction model for visual features and semantic representations, which projects visual features and semantic representations into the latent space. The loss function can be defined by Eq. (4).

$$L_{VAE} = E_{q_{\phi^{(v)}}}\left(z^{(v)}|x\right)\left[logp_{\theta^{(v)}}\left(x|z^{(v)}\right)\right] - D_{KL}\left(q_{\phi^{(v)}}\left(z^{(v)}|x\right)\|p_{\theta^{(v)}}\left(z^{(v)}\right)\right) + E_{q_{\phi^{(a)}}}\left(z^{(a)}|c\right)\left[logp_{\theta^{(a)}}\left(c|z^{(a)}\right)\right] - D_{KL}\left(q_{\phi^{(a)}}\left(z^{(a)}|c\right)\|p_{\theta^{(a)}}\left(z^{(a)}\right)\right) \quad (4)$$

Where $\times$ represents the visual feature of the original image, $q_{\phi}^{(v)}$ corresponds to the encoder of visual features, $p_{\theta^{(v)}}$ corresponds to the decoder of visual features, c represents the semantic representation, $q_{\phi^{(a)}}$ corresponds to the encoder of semantic representations, and $p_{\theta^{(a)}}$ corresponds to the decoder of semantic representations.

4.2.2. Cross-modal feature reconstruction (CMFR)

Through the reconstruction of visual features and semantic representations, we learned the representations of visual features and semantic representations in the latent space. Next, we need to align their representations in the latent space. We achieve this from two aspects. The first is cross-modal feature reconstruction (CMFR). Here, visual features and semantic representations are cross-input to the encoder corresponding to another modality, and the loss function of cross-modal feature reconstruction can be defined by Eq. (5).

$$\mathscr{L}_{CMFR} = \sum_{i=1}^{N}\left|x_i - p_{\theta^{(v)}}\left(q_{\phi^{(a)}}\left(c_i\right)\right)\right| + \left|c_i - p_{\theta^{(a)}}\left(q_{\phi^{(v)}}\left(x_i\right)\right)\right| \quad (5)$$

where N represents the number of training samples, x_i and c_i represent the visual features and semantic representations of the same category.

4.2.3. Visual and semantic distribution matching (VSDM)

The second is visual and semantic distribution matching (VSDM). The distribution of visual features and semantic representations in the latent space is determined by $\mu_i^{(v)}$, $\mathrm{E}_i^{(v)}$ and $\mu_i^{(a)}$, $\mathrm{E}_i^{(a)}$. We further match the distribution of visual features and semantic representations in latent space by reducing the distance between them, and the loss function of visual and semantic distribution matching can be defined by Eq. (6).

$$\mathscr{L}_{VSDM} = \sum_{i=1}^{N}\sqrt{\|\mu_i^{(v)} - \mu_i^{(a)}\|_2^2 + \|\sqrt{\mathrm{E}_i^{(v)}} - \sqrt{\mathrm{E}_i^{(a)}}\|_F^2} \quad (6)$$

where N represents the number of training samples, $\mu_i^{(v)}$ and $\sqrt{\mathrm{E}_i^{(v)}}$ represent the mean and standard deviation of the visual feature distribution in latent space, respectively, and $\mu_i^{(a)}$ and $\sqrt{\mathrm{E}_i^{(a)}}$ represent the mean and standard deviation of the semantic representation distribution in latent space, respectively.

4.2.4. Multi-category distribution dispersion (MCDD)

As we mentioned before, RS image scenes have significant characteristics of high interclass similarity, which is very unfavorable for classification tasks. For this, we add constraints to make the distribution of different categories in the latent space more dispersed, and the loss function of multi-category distribution dispersion can be defined by Eq. (7).

$$\mathscr{L}_{MCDD} = \|\mathbf{VHV}^T - \mathbf{I}\|_F^2 \quad (7)$$

where $\mathbf{V} = [\mu_1^{(a)}, \mu_2^{(a)}, \cdots, \mu_N^{(a)}] \in \mathbb{R}^{d\times N}$, $\mathbf{H} = (N\bullet\mathbf{P} - \mathbf{W})/N$, $\mathbf{P} \in \mathbb{R}^{N\times N}$ stands for an identity matrix and $\mathbf{W} \in \mathbb{R}^{N\times N}$ stands for a matrix with all elements equal to 1. $\mathbf{I} \in \mathbb{R}^{d\times d}$ is an identity matrix.

4.3. Classification of remote sensing image scenes under the ZSL and GZSL settings

It is worth noting that the ZSL and GZSL tasks share the same process to train the encoder of visual features and the encoder of semantic representations. However, they slightly differ from each other when training the final classifier, which is specifically explained in the following.

As mentioned before, ZSL aims to classify image scenes from unseen categories. Let $c \in C^u$ denote the semantic representation of the unseen category and $x \in X^u$ represent the visual feature of the testing image from the unseen category. We use the trained semantic representations encoder $q_{\phi^{(a)}}$ to map the semantic representation c to the latent space using Eq. (8) and use the trained visual feature encoder $q_{\phi^{(v)}}$ to map the visual feature x to the latent space using Eq. (9).

$$z^{(a)} = q_{\phi^{(a)}}(c) \quad (8)$$

$$z^{(v)} = q_{\phi^{(v)}}(x) \quad (9)$$

4.3.1. ZSL for classifying the unseen remote sensing image scenes

To train the ZSL classifier, which aims to recognize the image scenes from the unseen categories, the loss function can be written as $\min_{F_{ZSL}} -\sum t_i log r_i$, where $t_i \in Y^u$, and $r_i \in R$. The classification probability is calculated by $r = \sigma(z^{(a)} * F_{ZSL})$, where F_{ZSL} denotes the classification mapping matrix, $\sigma(\cdot)$ stands for the softmax activation function and $z^{(a)}$ are generated from $c \in C^u$.

In the testing phase, given one test image from the unseen category and x denotes its visual feature by Eq. (8). The label of the image from the unseen category can be inferred by the classifier mapping matrix F_{ZSL} and $z^{(v)} = q_{\phi^{(v)}}(x)$.

4.3.2. GZSL for classifying both of the seen and unseen remote sensing image scenes

Different from ZSL, GZSL needs to classify image scenes from the seen or unseen category. Hence, let $c \in C^u \cup C^s$ denote the semantic representation of the seen or unseen category, and $x \in X^u \cup X^s$ represent the visual feature of the image scene from the seen or unseen category. We use the encoder of semantic representations to map the semantic representation c to the latent space using Eq. (8) and use the encoder of visual features to map the visual feature x to the latent space using Eq. (9).To train the GZSL classifier, which tries to simultaneously recognize the image scenes from both the seen and unseen categories, we first generate $z^{(a)}$ from $c \in C^u \cup C^s$, and the loss function is defined as $\min_{F_{GZSL}} -\sum g_i log s_i$, where $g_i \in Y^s \cup Y^u$ and $s_i \in S$. More specifically, $s = \sigma(z^{(a)} * F_{GZSL})$, where F_{GZSL} stands for the classification mapping matrix.

Given the visual feature X of one image scene from the seen or unseen category, its label can be referred to by the classifier mapping matrix F_{GZSL} and $z^{(v)} = q_{\phi^{(v)}}(x)$.

5. Experimental analysis and discussion

In this section, we design extensive experiments to evaluate our proposed approach. In Section 5.1, we introduce the experimental settings. Then, we analyze the sensitivity of critical parameters in our proposed approach in Section 5.2. Finally, we compare our method with the state-of-the-art methods in Section 5.3.

5.1. Experiment settings

Section 5.1.1 introduces our RS dataset. In Section 5.1.2, we introduce the evaluation metric of ZSL and GZSL, and we give implementation details of our proposed method in Section 5.1.3.

5.1.1. Evaluation dataset

To fully evaluate the performance of zero-shot and generalized zero-shot RS image scene classification, we adopt a combined dataset by integrating five public datasets including UCM, AID, NWPU-RESISC45, RSI-CB256, and PatternNet. The merged image scene dataset is composed of 70 scene categories, and each category contains 800 image scenes. The RS image scene dataset is visually shown in Fig. 4.

5.1.2. Evaluation metric

The overall accuracy (OA) is taken as the quantitative metric to evaluate the classification performance under the ZSL setting. Under the GZSL setting, the testing dataset is composed of seen and unseen images, so the accuracy is evaluated on seen classes, denoted as SA, and unseen classes, denoted as UA. As suggested by (Xian et al., 2018), the harmonic mean accuracy (HMA), defined as $\mathrm{HMA} = (2 \times \mathrm{SA} \times \mathrm{UA})/(\mathrm{SA} + \mathrm{UA})$, is used to evaluate the classification performance under the GZSL setting.

5.1.3. Implementation details

In the module for latent space mapping, the encoder $q_{\phi^{(v)}}$ and decoder $p_{\theta^{(a)}}$ had 512 hidden units, the encoder $q_{\phi^{(a)}}$ and decoder $p_{\theta^{(a)}}$ had 256 hidden units, and the latent embedding dimensions were 32. Regarding the dimension of representation learning d and the hyperparameters of the objective function α, β, γ, we specifically analyze their sensitivity and give the recommendation settings in Section 5.2. In the module for classifying latent features, we train a linear classifier to classify latent features.

Regarding the knowledge types, we consider four kinds of semantic representations. 1) Word2Vec: we use a Word2Vec model (Bojanowski et al., 2017) that is trained on the Wikipedia corpus to obtain the 300-D vector for each class name. 2) BERT: we depict each RS scene category by one summarized sentence after checking over 10 random RS image scenes from one given category. Then, the BERT model (Devlin et al., 2018) maps the sentence description of each RS scene category to one different semantic representation with 1,024 dimensions. 3) Attribute: Considering the color, shape and objects contained in each RS scene category, we manually design each dimension of the vector. If the RS scene category has a certain attribute, the corresponding dimension is 1; otherwise, it is marked as 0. We create the 59-dimensional vector in this way. 4) SR-RSKG: We generate the semantic representations of RS scene categories by representation learning of RSKG, as depicted in Section 3.

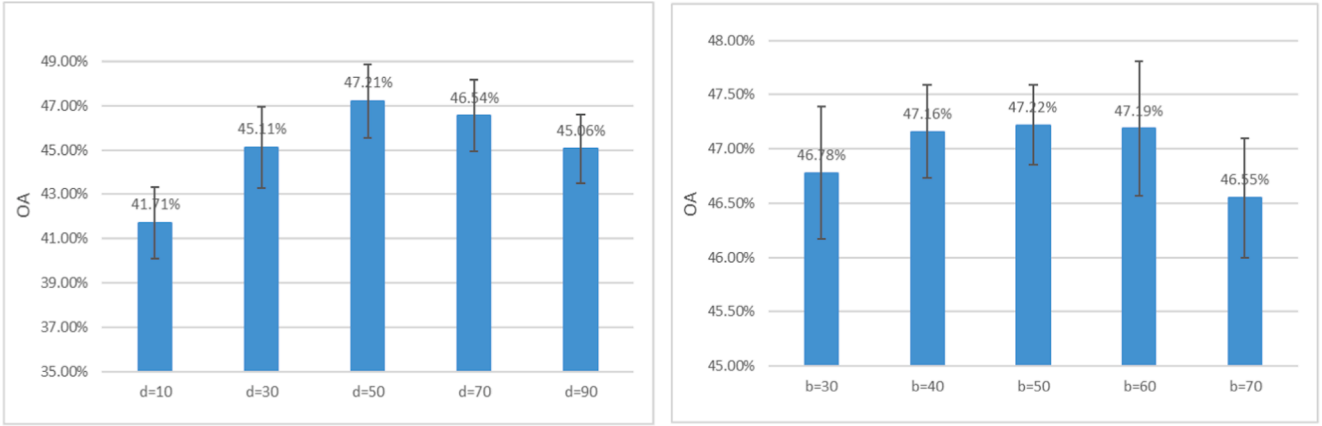
Regarding the visual features, we consider the 512-dimensional CNN feature vector of the RS image scene using ResNet-18 (He et al., 2016) and 2048-dimensional CNN feature vector of the RS image scene using EfficientNet (Tan et al., 2019). For both the ZSL and GZSL tasks, we leverage only the samples from the seen categories to fine-tune model. In the ZSL task, we use all 800 images for each seen class. In the GZSL task, considering that some samples from the seen categories are involved in the testing phase, we select 600 images of each seen class to train the ResNet-18 model, and the remaining 200 images attend the testing phase in GZSL.

5.2. Sensitivity analysis of critical parameters

To evaluate the effect of the hyperparameters, we summarize the classification performance of our proposed method under the ZSL setting. More specifically, we first analyze the sensitivity of the dimension of semantic representation d in the representation learning module and the sensitivity of the batch size b , where d specifically denotes the dimension of the semantic vector h, r and t in Eq. (1).

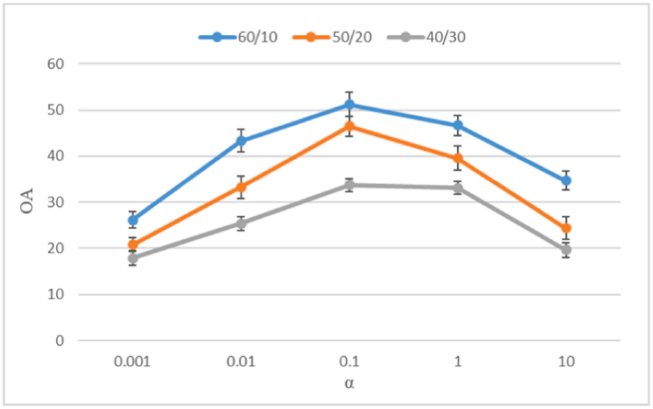
Considering the time consumption, the seen/unseen ratio is set to 60/10. Given this seen/unseen ratio, we calculate the average and standard deviation of the classification results over 5 random seen/unseen splits to determine the optimal value of the parameter. As shown in Fig. 7(a), when α, β and γ are set to 1, our proposed method can achieve the best performance when d equals 50. As shown in Fig. 7(b), when d is fixed to 50, our proposed method is not significantly sensitive to the batch size between the interval from 40 to 60, so b is recommended to be set to 50.

Furthermore, we analyze the sensitivity of the hyperparameters α, β, and γ in our proposed DAN. In the evaluation experiments, we vary one parameter at each time while fixing the others to their optimal value. It is worth noting that in the ZSL task and the GZSL task, the training processes of the latent space mapping model are equal. The difference lies in the number of samples used and the test samples in the test phase. Therefore, considering the time consumption, we only conduct experiments under the ZSL task when analyzing the sensitivity of critical

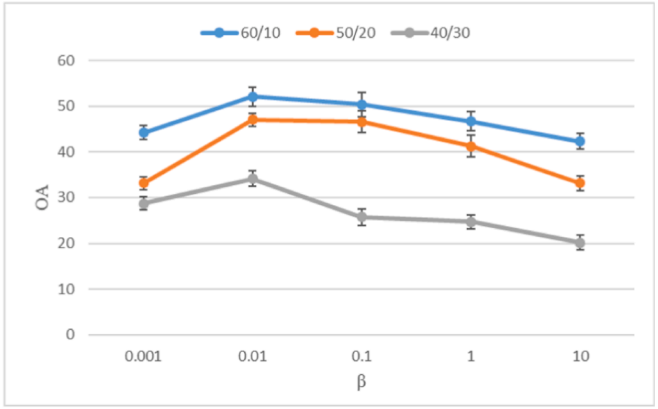


(a) The performance variation under different d. (b) The performance variation under different b.

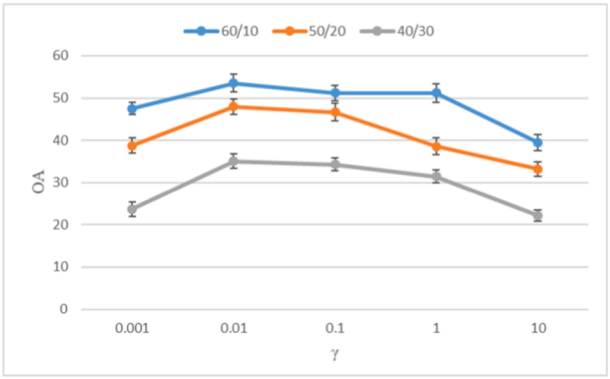
Fig. 7. The performance variation along with different representation learning vector dimensions and batch sizes. (a) The performance variation under different d. (b) The performance variation under different b.



(a) The performance variation under different α.



(b) The performance variation under different β.



(c) The performance variation under different γ.

Fig. 8. The classification performance variation under different α, β and γ. (a) The performance variation under different α. (b) The performance variation under different β. (c) The performance variation under different γ.

parameters.

Fig. 8 reports the classification performance of our proposed method under different α, β and γ. To determine the optimal value of each parameter more accurately, we conduct experiments under three ratios of seen/unseen: 60/10, 50/20, and 40/30. As shown in Fig. 8(a), we set β and γ to 1, and our proposed method can achieve the best performance when α equals 0.1. As shown in Fig. 8(b), we set α as 0.1 and γ as 1. Our proposed method can achieve the best performance when β equals 0.01. As shown in Fig. 8(c), we set α as 0.1 and β as 0.01. Our proposed method can achieve the best performance when γ equals 0.01.

To pursue generalization, the hyperparameter setting of our method is fixed under both the ZSL and GZSL settings.

5.3. Comparison with the state-of-the-art methods

To fully analyze the ZSL and GZSL methods, we report the classification results under different seen/unseen ratios (e.g., 60/10, 50/20, and 40/30). More specifically, in each given seen/unseen ratio, we evaluate each method over 5 random seen/unseen splits. As mentioned above, four kinds of semantic representations, including Word2Vec, BERT, Attribute and SR-RSKG, are evaluated.

5.3.1. Comparison with the existing ZSL methods

To show the superiority of our proposed method, we consider the following baselines in ZSL: SAE (Kodirov et al., 2017), dual visual-semantic mapping (DMaP) (Li et al., 2017c), semantics-preserving locality embedding (SPLE) (Tao et al., 2017), creativity inspired zero-shot

Table 1
Comparison of different methods with ResNet-18 under the ZSL setting using OA.

Knowledge type	Seen/Unseen ratio	SAE	DMaP	SPLE	CIZSL	CADA-VAE	ZSC-SA	Our DAN
Word2Vec	60/10	23.5 ± 4.2	26.0 ± 3.6	20.1 ± 3.7	20.6 ± 0.4	41.4 ± 2.3	26.7 ± 5.3	**44.3 ± 2.6**
	50/20	13.7 ± 1.7	16.7 ± 2.2	13.2 ± 1.9	10.6 ± 3.7	30.3 ± 2.7	15.2 ± 1.0	**34.7 ± 1.7**
	40/30	9.6 ± 1.4	10.4 ± 0.9	9.8 ± 1.4	6.0 ± 1.2	21.2 ± 2.9	12.1 ± 0.8	**24.3 ± 3.7**
BERT	60/10	22.0 ± 1.7	16.4 ± 1.9	19.0 ± 3.8	20.4 ± 4.1	48.1 ± 2.9	29.3 ± 3.8	**50.2 ± 2.5**
	50/20	12.4 ± 1.9	15.6 ± 1.9	13.2 ± 2.6	10.3 ± 1.9	37.1 ± 3.5	18.3 ± 1.3	**43.4 ± 2.7**
	40/30	8.8 ± 1.3	10.0 ± 0.8	8.3 ± 2.0	6.2 ± 2.1	26.3 ± 2.2	13.1 ± 3.0	**31.5 ± 2.0**
Attribute	60/10	23.6. ± 2.8	31.2 ± 4.1	26.8 ± 2.1	16.4 ± 3.1	47.1 ± 2.9	28.5 ± 3.2	**50.1 ± 3.3**
	50/20	12.1 ± 1.7	18.7 ± 2.4	16.6 ± 2.1	7.5 ± 3.2	35.2 ± 2.1	19.4 ± 2.8	**43.1 ± 1.5**
	40/30	8.6 ± 1.0	12.6 ± 1.1	10.7 ± 1.2	6.2 ± 2.9	26.1 ± 2.6	12.7 ± 2.1	**30.1 ± 1.9**
Our SR-RSKG	60/10	22.1. ± 2.3	33.1 ± 2.9	28.5 ± 2.6	18.2 ± 2.6	50.5 ± 2.6	31.3 ± 2.5	**53.3 ± 3.8**
	50/20	12.8 ± 2.3	20.3 ± 1.8	17.2 ± 2.1	8.9 ± 2.5	39.6 ± 3.1	19.1 ± 1.7	**45.2 ± 1.3**
	40/30	9.2 ± 1.5	12.9 ± 2.4	10.2 ± 1.6	7.1 ± 1.5	28.2 ± 2.6	13.6 ± 2.5	**33.4 ± 3.0**

Table 2
Comparison of different methods with EfficientNet under the ZSL setting using OA.

Knowledge type	Seen/Unseen ratio	SAE	DMaP	SPLE	CIZSL	CADA-VAE	ZSC-SA	Our DAN
Word2Vec	60/10	24.1 ± 2.4	27.1 ± 2.3	20.8 ± 2.3	20.4 ± 2.1	40.1 ± 2.2	29.3 ± 3.6	**43.2 ± 2.2**
	50/20	15.1 ± 2.5	17.9 ± 3.2	14.7 ± 1.6	11.1 ± 1.8	**33.8 ± 2.1**	20.2 ± 1.8	32.7 ± 1.9
	40/30	9.8 ± 1.1	11.1 ± 1.7	9.9 ± 1.2	7.8 ± 1.1	22.8 ± 2.1	11.9 ± 0.9	**23.5 ± 1.7**
BERT	60/10	22.9 ± 2.1	27.7 ± 1.3	22.3 ± 3.1	19.1 ± 2.1	**49.2 ± 2.2**	32.1 ± 2.6	49.0 ± 1.9
	50/20	16.2 ± 1.2	15.9 ± 1.9	14.3 ± 2.8	10.4 ± 1.5	39.2 ± 3.1	21.3 ± 1.6	**40.2 ± 2.1**
	40/30	11.5 ± 1.3	10.2 ± 0.5	9.7 ± 2.4	6.6 ± 2.2	25.2 ± 1.7	13.3 ± 2.1	**29.3 ± 1.4**
Attribute	60/10	24.1. ± 1.3	30.4 ± 2.1	27.1 ± 2.4	21.3 ± 4.1	51.9 ± 2.9	33.2 ± 2.6	**53.1 ± 2.3**
	50/20	16.1 ± 1.8	16.6 ± 2.0	18.0 ± 2.7	12.3 ± 2.2	40.9 ± 2.0	21.6 ± 2.1	**43.6 ± 1.9**
	40/30	10.3 ± 1.2	12.3 ± 1.1	12.1 ± 1.7	8.2 ± 2.4	29.2 ± 2.6	13.0 ± 2.3	**31.6 ± 1.3**
Our SR-RSKG	60/10	23.9. ± 1.2	30.3 ± 2.2	28.8 ± 2.4	20.2 ± 2.3	52.5 ± 2.6	35.4 ± 2.2	**55.2 ± 2.8**
	50/20	16.6 ± 1.9	19.1 ± 1.3	18.3 ± 1.8	10.9 ± 2.3	41.9 ± 2.3	23.2 ± 2.4	**43.1 ± 2.3**
	40/30	11.2 ± 1.6	11.7 ± 2.5	10.9 ± 1.4	7.7 ± 1.4	30.1 ± 1.8	17.3 ± 2.1	**31.5 ± 1.8**

Table 3
Comparison of different methods with ResNet-18 under the GZSL setting using HMA.

Knowledge type	Seen/Unseen ratio	SAE	DMaP	CIZSL	CADA-VAE	Our DAN
Word2Vec	60/10	27.97 ± 1.13	28.88 ± 1.26	25.18 ± 0.86	32.88 ± 2.54	**34.09 ± 1.34**
	50/20	20.99 ± 1.90	20.33 ± 1.13	15.70 ± 0.86	30.25 ± 3.07	**31.44 ± 1.66**
	40/30	17.15 ± 0.55	16.78 ± 1.10	9.10 ± 1.32	**26.06 ± 0.79**	25.63 ± 0.26
BERT	60/10	28.57 ± 0.94	26.57 ± 0.65	25.00 ± 1.25	36.34 ± 2.03	**37.96 ± 1.65**
	50/20	21.52 ± 1.38	19.52 ± 1.42	14.95 ± 1.51	**31.51 ± 2.27**	31.45 ± 1.85
	40/30	16.65 ± 0.40	16.31 ± 1.24	8.57 ± 0.57	27.05 ± 0.79	**28.15 ± 1.16**
Attribute	60/10	28.58 ± 0.93	30.71 ± 0.78	23.88 ± 0.87	36.00 ± 2.19	**37.60 ± 1.24**
	50/20	20.52 ± 1.75	23.55 ± 0.87	14.27 ± 1.05	32.17 ± 2.41	**32.66 ± 0.80**
	40/30	16.73 ± 1.06	16.12 ± 0.82	8.11 ± 0.98	26.13 ± 0.79	**28.79 ± 0.92**
Our SR-RSKG	60/10	28.86 ± 0.60	30.11 ± 1.39	23.65 ± 0.61	38.10 ± 1.89	**40.25 ± 0.84**
	50/20	23.66 ± 1.06	23.41 ± 1.21	13.93 ± 1.01	32.94 ± 1.42	**34.11 ± 0.45**
	40/30	16.94 ± 1.03	16.20 ± 1.62	8.14 ± 0.87	28.11 ± 0.79	**29.61 ± 0.82**

Table 4
Comparison of different methods with EfficientNet under the GZSL setting using HMA.

Knowledge type	Seen/Unseen ratio	SAE	DMaP	CIZSL	CADA-VAE	Our DAN
Word2Vec	60/10	29.11 ± 1.22	30.13 ± 1.65	24.72 ± 1.41	32.13 ± 1.81	**33.56 ± 1.15**
	50/20	22.05 ± 1.95	21.6 ± 1.41	17.05 ± 1.43	29.34 ± 2.25	**30.23 ± 1.66**
	40/30	17.88 ± 1.21	16.87 ± 1.06	8.99 ± 1.01	23.15 ± 1.95	**23.63 ± 0.71**
BERT	60/10	30.47 ± 1.45	32.17 ± 2.15	24.08 ± 1.24	35.07 ± 2.05	**37.39 ± 1.47**
	50/20	23.34 ± 1.16	22.92 ± 1.15	15.65 ± 1.12	30.15 ± 2.18	**32.85 ± 1.56**
	40/30	17.92 ± 1.40	17.13 ± 1.06	8.01 ± 0.76	24.32 ± 1.68	**26.60 ± 1.34**
Attribute	60/10	30.18 ± 0.91	31.08 ± 1.51	23.12 ± 0.56	**36.24 ± 2.28**	36.15 ± 1.33
	50/20	24.11 ± 1.34	22.55 ± 1.17	14.11 ± 1.27	30.22 ± 2.41	**31.86 ± 0.92**
	40/30	16.98 ± 1.64	16.24 ± 1.10	8.56 ± 0.58	25.41 ± 0.59	**26.07 ± 1.24**
Our SR-RSKG	60/10	32.11 ± 0.61	34.15 ± 1.21	24.56 ± 0.65	37.15 ± 1.54	**39.61 ± 1.66**
	50/20	25.06 ± 1.36	23.01 ± 1.30	15.93 ± 1.35	31.04 ± 1.26	**32.94 ± 0.81**
	40/30	18.91 ± 1.14	19.13 ± 1.44	9.12 ± 0.76	25.95 ± 1.19	**27.43 ± 1.19**

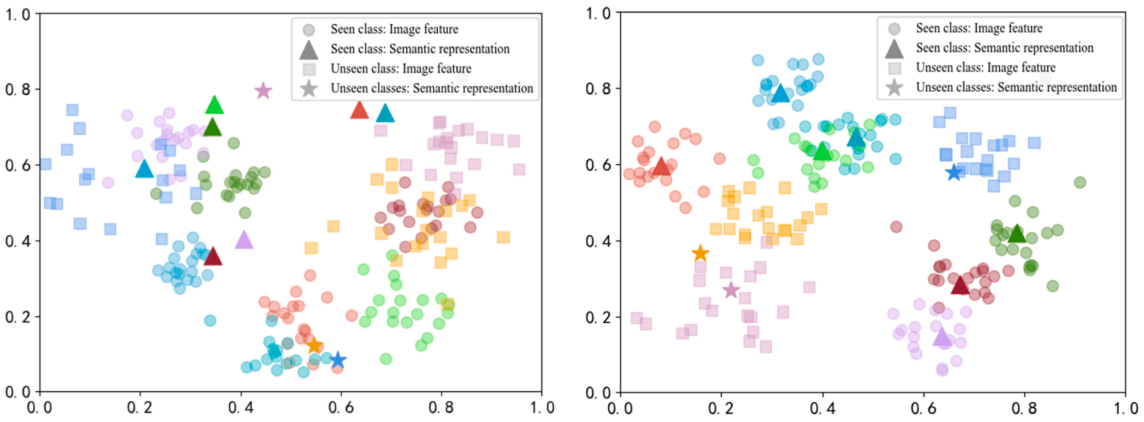
learning (CIZSL) (Elhoseiny and Elfeki, 2019), cross and distribution aligned VAE (CADA-VAE) (Schonfeld et al., 2019), and zero-shot scene classification algorithm based on structural alignment (ZSC-SA) (Quan et al., 2018). ZSC-SA is a ZSL method that is specifically proposed to address zero-shot RS image scene classification, and the other methods are recently representative ZSL methods in the computer vision field. To pursue the fair comparison, the parameters of the above baseline methods are set based on the recommended parameters in their papers.

As shown in Table 1, our proposed DAN can obviously outperform the state-of-the-art methods in terms of different seen/unseen ratios and with different knowledge types. In addition, it is worth noting that comparing the different knowledge types used in the same method, the SR-RSKG achieves the best performance in most cases. This proves that vectors obtained based on representation learning of RSKG are superior to the embedding vectors extracted by natural language processing models and manual annotation attribute vectors in describing RS scenes. As shown in Table 2, we also conduct experiments by leveraging the image features extracted by EfficientNet. To pursue the generalization, we re-use the critical parameters based on Resnet-18. As shown in Table 2, our DAN also achieves the optimal performance in most cases.

5.3.2. *Comparison with the existing GZSL methods*

To pursue generalization, we directly reuse the hyperparameter setting of ZSL to evaluate GZSL. We consider the following baselines: SAE, DMaP, CIZSL, and CADA-VAE in GZSL. It is worth noting that the SAE and DMaP are not proposed for the GZSL task, as they are specifically designed for ZSL.

In GZSL, we evaluate the performance of different methods under the GZSL setting using HMA. As shown in Table 3, our method aligns the visual latent features and semantic latent representations while separating different class distributions in the latent space, which enhances the visual-semantic coupling so that our method improves the classification accuracy and maintains a good balance between seen class accuracy and unseen class accuracy. Furthermore, we evaluate the performance of different GZSL method with the image features based on EfficientNet and summarize the evaluation results in Table 4. As shown



(a) Original features without DAN. (b) Aligned features with DAN.

Fig. 9. The t-SNE visualization of original features without DAN and aligned latent features with DAN. (a) Original features without DAN. (b) Aligned features with DAN.

SAE	River	Park	Baseball_field	River	Sea_ice
DMaP	River	Church	Baseball_field	Freeway	Desert
CIZSL	Desert_chaparral	Dense_residential	Desert	Dam	Beach
CADA-VAE	Grass	Overpass	Baseball_field	Bridge	Wetland
Our DAN	River	Dense_residential	Basketball_court	Bridge	Wetland

Fig. 10. Visual prediction results of different GZSL methods with the presented SR-RSKG. Red indicates wrong prediction, and green stands for correct prediction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in Table 4, our DAN can also outperform the most majority of methods under different kinds of semantic representations.

In order to intuitively show the effectiveness of our DAN, we visualize the original features without DAN and aligned features with DAN under the GZSL setting with the 60/10 partition in Fig. 9. Here, we also take the popular t-SNE dimension reduction tool to display features. To pursue the clear illustration, we only randomly display 20 image features from one category. Because the semantic features of the scene category is unique, the unique semantic features of the same category is displayed. The image and semantic features of the same category are represented by the same color, but the image and semantic features are represented by different shapes. Specifically, the circle represents the image features of seen class, the triangle represents the semantic features of seen class, the square represents the image features of the unseen class, and the five pointed star represents the semantic representation of unseen class. Through Fig. 9, in terms of seen and unseen classes, we can intuitively see that the distribution of visual and semantic features in the latent space has been obviously aligned after DAN. In addition, our DAN can generate the better alignment performance on the seen classes compared with the unseen classes. This result can be understood as DAN is trained on the seen classes and DAN does not see any unseen data in the training stage. It is very gratifying to see the promising alignment on unseen classes as shown in Fig. 9(b).

With the seen/unseen ratio set to 60/10, we report the visual classification results of different GZSL methods in Fig. 10. As shown, our DAN can outperform the existing competitors in terms of the seen classes and unseen classes.

6. Conclusion

Driven by the increasing practical demands of ZSL and GZSL in the RS field, this paper mainly focuses on zero-shot and generalized zero-shot RS image scene classification. Considering that natural language processing models based on generalized corpora have poor performance in describing RS-oriented scene categories appropriately, this paper, for the first time, proposes to generate semantic representations of RS scene categories through representation learning of RSKG and applies them to zero-shot and generalized zero-shot RS image scene classification. By comparison with other traditional knowledge types, we verify the superiority of SR-RSKG from intuitive illustration and quantitative analysis perspectives. In addition, we propose a robust DAN to complete zero-shot and generalized zero-shot scene classification. To evaluate our method, we conduct extensive comparative experiments under three seen/unseen ratios using a large RS image scene dataset. Our proposed DAN outperforms the state-of-the-art methods under both the ZSL and GZSL settings.

In our future work, we will attempt to enlarge the RSKG. Intuitively, it can be expected that a larger RSKG can contain richer prior knowledge and benefit generating better semantic representations of RS scene categories. In addition, we will exploit some of the advanced

representation learning methods, such as convolutional 2D knowledge graph embeddings (ConvE) (Dettmers et al., 2018) and the graph attention network (GAT) (Velickovic et al., 2017), to further improve the performance of representation learning of RSKG. In addition, how to address zero-shot multi-label scene classification (Hua et al., 2020) using RSKG is also a very worthy research direction. Finally, how to design and optimize the end-to-end network between RS image scenes and RSKG is also a very meaningful research direction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0505003; the National Natural Science Foundation of China under Grant 41971284 ; the State Key Program of the National Natural Science Foundation of China under Grants 42030102 and 92038301; the Foundation for Innovative Research Groups of the Natural Science Foundation of Hubei Province under Grant 2020CFA003; the China Postdoctoral Science Foundation under Grants 2016M590716 and 2017T100581; the Shenzhen Basic Research Program under Grant JCYJ20180305180708546.

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., 2007. Dbpedia: A nucleus for a web of open data. The Semantic Web 722–735.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 65 (1), 2–16.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5, 135–146.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J., 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In: In: Proceedings of the ACM SIGMOD international conference on Management of data, pp. 1247–1250.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O., 2013. Translating embeddings for modeling multi-relational data. Proceedings of Neural Information Processing Systems 26, 2787–2795.

Chen, T., Kornblith, S., Norouzi, M., et al., 2020. A simple framework for contrastive learning of visual representations. International conference on machine learning. PMLR 1597–1607.

Cheng, G., Guo, L., Zhao, T., Han, J., Li, H., Fang, J., 2013. Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA. International Journal of Remote Sensing 34 (1), 45–59.

Cheng, G., Han, J., 2016. A survey on object detection in optical remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing 117, 11–28.

Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE 105 (10), 1865–1883.

Chi, M., Plaza, A., Benediktsson, J.A., Sun, Z., Shen, J., Zhu, Y., 2016. Big data for remote sensing: Challenges and opportunities. Proceedings of the IEEE 104 (11), 2207–2219.

Clementini, E., 2009. A conceptual framework for modelling spatial relations. Information Technology and Control 48 (1), 5–17.

Demir, B., Bruzzone, L., 2016. Hashing-based scalable remote sensing image search and retrieval in large archives. IEEE Transactions on Geoscience and Remote Sensing 54 (2), 892–904.

Dettmers, T., Minervini, P., Stenetorp, P., Riedel, S., 2018. Convolutional 2d knowledge graph embeddings. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805v2.

Elhoseiny, M., Elfeki, M., 2019. Creativity inspired zero-shot learning. In: In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5784–5793.

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., 2014. Introducing Wikidata to the linked data web. In: In: Proceedings of International semantic web conference, pp. 50–65.

Gerke, M., Xiao J., Fusion of airborne laserscanning point clouds and images for supervised and unsupervised scene classification. ISPRS Journal of Photogrammetry and Remote Sensing, 87: 78-92.

Gu, Y., Wang, Y., Li, Y., 2019. A survey on deep learning-driven remote sensing image scene understanding: Scene classification, scene retrieval and scene-guided object detection. Applied Sciences 9 (10), 2110.

Guo, W., Wang, J., Wang, S., 2019. Deep multimodal representation learning: A survey. IEEE Access 7, 63373–63394.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

He, K., Fan, H., Wu, Y., 2020. Momentum contrast for unsupervised visual representation learning. In: In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9729–9738.

Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G., 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence 194, 28–61.

Hua, Y., Mou, L., Zhu, X.X., 2020. Relation network for multilabel aerial image classification. IEEE Transactions on Geoscience and Remote Sensing 58 (7), 4558–4572.

Ji, Z., Cui, B., Li, H., Jiang, Y.-G., Xiang, T., Hospedales, T., Fu, Y., 2020. Deep ranking for image zero-shot multi-label classification. IEEE Transactions on Image Processing 29, 6549–6560.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv: 1312.6114.

Kodirov, E., Xiang, T., Gong, S., 2017. Semantic autoencoder for zero-shot learning. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3174–3183.

Lampert, C.H., Nickisch, H., Harmeling, S., 2009. Learning to detect unseen object classes by between-class attribute transfer. In: In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 951–958.

Larochelle, H., Erhan, D., Bengio, Y., 2008. Zero-data learning of new tasks. Proceedings of AAAI 1 (2), 3.

Lazaridou, A., Dinu, G., Baroni, M., 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In: In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp. 270–280.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Li, Y., Tao, C., Tan, Y., Shang, K.e., Tian, J., 2016. Unsupervised multilayer feature learning for satellite image scene classification. IEEE Geoscience and Remote Sensing Letters 13 (2), 157–161.

Li, A., Lu, Z., Wang, L., Xiang, T., Wen, J.-R., 2017a. Zero-shot scene classification for high spatial resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 55 (7), 4157–4167.

Li, H., Dou, X., Tao, C., Hou, Z., Chen, J., Peng, J., Deng, M., Zhao, L., 2017b. RSI-CB: A large scale remote sensing image classification benchmark via crowdsource data. arXiv preprint arXiv:1705.10450.

Li, Y., Wang, D., Hu, H., Lin, Y., Zhuang, Y., 2017c. Zero-shot recognition using dual visual-semantic mapping paths. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3279–3287.

Li, Y., Zhang, Y., Huang, X., Ma, J., 2018. Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval. IEEE ransactions on Geoscience and Remote Sensing 56 (11), 6521–6536.

Liang, Y., Bai, Y., Zhang, W., Qian, X., Zhu, L., Mei, T., 2019. Vrr-vg: Refocusing visually-relevant relationships. In: In: Proceedings of the IEEE International Conference on Computer Vision, pp. 10403–10412.

Li, Y., Chen, W., Zhang, Y., Tao, C., Xiao, R., Tan, Y., 2020. Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning. Remote Sensing of Environment 250, 112045. https://doi.org/10.1016/j.rse.2020.112045.

Lobry, S., Marcos, D., Murray, J., Tuia, D., 2020. RSVQA: visual question answering for remote sensing data. IEEE Transactions on Geoscience and Remote Sensing 58 (12), 8555–8566.

Li, Y., Shi, T.e., Zhang, Y., Chen, W., Wang, Z., Li, H., 2021a. Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation. ISPRS Journal of Photogrammetry and Remote Sensing 175, 20–33.

Li, Y., Zhu, Z., Yu, J., Zhang, Y., 2021b. Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing in press.

Li, Y., Zhang, Y., Zhu, Z., 2021c. Error-Tolerant Deep Learning for Remote Sensing Image Scene Classification. IEEE Transactions on Cybernetics 51 (4), 1756–1768.

Long, Y., Liu, L., Shao, L., Shen, F., Ding, G., Han, J., 2017. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1627–1636.

Marcos, D., Volpi, M., Kellenberger, B., Tuia, D., 2018. Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models. ISPRS Journal of Photogrammetry and Remote Sensing 145, 96–107.

Maaten, L.v.d., Hinton, G.J., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research, 9: 2579-2605.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Misra, I., Maaten, L., 2020. Self-supervised learning of pretext-invariant representations. In: In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6707–6717.

Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T., 2009. Zero-shot learning with semantic output codes. Proceedings of Advances in neural information processing systems 22, 1410–1418.

Quan, J., Wu, C., Wang, H., Wang, Z., 2018. Structural alignment based zero-shot classification for remote sensing scenes. In: In: Proceedings of the IEEE International Conference on Electronics and Communication Engineering (ICECE), pp. 17–21.
Schonfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., Akata, Z., 2019. Generalized zero- and few-shot learning via aligned variational autoencoders. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8247–8255.
Shadbolt, N., Berners-Lee, T., Hall, W., 2006. The semantic web revisited. IEEE intelligent systems 21 (3), 96–101.
Shen, J., Zhou, T., Chen, M., 2017. A 27-intersection model for representing detailed topological relations between spatial objects in two-dimensional space. ISPRS International Journal of Geo-Information 6 (2), 37.
Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M., Matsumoto, Y., 2015. Ridge regression, hubness, and zero-shot learning,In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 135-151.
Sumbul, Gencer, Cinbis, Ramazan Gokberk, Aksoy, Selim, 2018. Fine-grained object recognition and zero-shot learning in remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing 56 (2), 770–779.
Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: In: International Conference on Machine Learning, pp. 6105–6114.
Tao, S.-Y., Yeh, Y.-R., Wang, Y.C.F., 2017. Semantics-Preserving Locality Embedding for Zero-Shot Learning. In: Proceedings of BMVC.
Tao, Chao, Mi, Li, Li, Yansheng, Qi, Ji, Xiao, Yuan, Zhang, Jiaxing, 2019a. Scene Context-Driven Vehicle Detection in High-Resolution Aerial Images. IEEE Transactions on Geoscience and Remote Sensing 57 (10), 7339–7351.
Tao, Chao, Qi, Ji, Li, Yansheng, Wang, Hao, Li, Haifeng, 2019b. Spatial information inference net: road extraction using road-specific contextual information. ISPRS Journal of Photogrammetry and Remote Sensing 158, 155–166.
Tempelmeier, Nicolas, Demidova, Elena, 2021. Linking openstreetmap with knowledge graphs—link discovery for schema-agnostic volunteered geographic information. Future Generation Computer Systems 116, 349–364.
Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., 2017. Graph attention networks. arXiv preprint arXiv:1710.10903.
Wang, Z., Zhang, J., Feng, J., Chen, Z., 2014. Knowledge graph embedding by translating on hyperplanes, Proceedings of the AAAI Conference on Artificial Intelligence. 28(1).
Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., Sensing, R., 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing 55 (7), 3965–3981.
Xian, Yongqin, Lampert, Christoph H., Schiele, Bernt, Akata, Zeynep, 2019. Zero-shot learning-A comprehensive evaluation of the good, the bad and the ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (9), 2251–2265.
Xian, Y., Lorenz, T., Schiele, B., Akata, Z., 2018. Feature generating networks for zero-shot learning. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5542–5551.
Yang, Y., Newsam, S., 2011. Spatial pyramid co-occurrence for image classification. In: In: Proceedings of International Conference on Computer Vision, pp. 1465–1472.
Zhang, Xiuyuan, Du, Shihong, Wang, Qiao, 2018. Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping. Remote Sensing of Environment 212, 231–248.
Zhang, Fan, Du, Bo, Zhang, Liangpei, 2016. Scene classification via a gradient boosting random convolutional network framework. IEEE Transactions on Geoscience and Remote Sensing 54 (3), 1793–1802.
Zhou, Weixun, Newsam, Shawn, Li, Congmin, Shao, Zhenfeng, 2018. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS Journal of Photogrammetry and Remote Sensing 145, 197–209.