scientific reports

Check for updates

OPEN

Enhancing zero-shot scene recognition through semantic autoencoders and visual relation transfer

Chen Wang[1]✉, Man Wang[1], Guohua Peng[2], Bernard De Baets[3] & Xiong Pan[1]✉

Zero-shot learning enables the recognition of images from unseen classes by leveraging auxiliary semantic information. Traditional methods typically learn either the relationship between the visual features and the semantic vectors or that between the seen and the unseen semantic vectors. However, their zero-shot recognition performances are not ideal for scene images due to large intra-class variations. To address this challenge, we propose a novel approach combining semantic autoencoders (SAEs) and visual relation transfer (VRT), termed SAEVRT. Specifically, we learn two semantic autoencoders for both the seen and the unseen scene classes, which help to alleviate the domain shift between the visual and the semantic spaces. Considering that semantic vectors (no attribute vectors available) are less effective than visual features for scene images, we propose an interpretable seen and unseen visual relation transfer method to learn more effective unseen semantic vectors. By combining SAEs and VRT in a unified learning framework, we exploit both the visual-semantic and seen-unseen relationships. Extensive experiments on four scene datasets demonstrate the superior performance of SAEVRT, achieving recognition accuracies of 63.77%, 67.75%, 58.68%, and 53.26% on Scene15, MIT67, UCM21, and NWPU45, respectively.

Keywords Semantic autoencoders, Visual relation transfer, Zero-shot learning, Scene recognition

Traditional scene recognition tasks have achieved great success due to the development of deep learning technology[1,2]. However, they need a large number of training images and fail to recognize test images from an unseen class[3]. Benefiting from auxiliary semantic information, zero-shot learning is capable of recognizing new images from an unseen class. Therefore, in this paper, we explore zero-shot learning in scene recognition to tackle zero-shot scene recognition.

The key to zero-shot learning is how to build the relationship between the visual features and the semantic vectors. Accordingly, many researchers make great efforts to learn this relation, which can be divided into the following three embedding categories. (1) Visual→Semantic Embedding learns an embedding function that projects the visual features to the semantic vectors. Representative methods are the *Semantic AutoEncoder* (SAE) method[4] and variants[5,6] thereof. (2) Semantic→Visual Embedding learns an embedding function that projects the semantic vectors to the visual features. Based on the vision transformer framework, the *Cross Attribute-Guided Transformer* (CAGT) method[7] devises an attribute-guided transformer network to refine visual features. (3) Visual→Common Space←Semantic Embedding learns two embedding functions that project the semantic vectors and the visual features to a common space. For example, to address the domain shift problem, the simple *Discriminative Dual Semantic Autoencoder* (DDSA) method[8] learns two bidirectional mappings in an aligned space.

Learning the relationship between the visual features and the semantic vectors is conducive to promote the zero-shot recognition performance. However, it neglects the relationship between the seen and the unseen scene classes, which leads to the domain shift problem between these two groups of classes. To overcome this problem, researchers have began to exploit the relationship between the seen and the unseen classes using the unseen semantic information[9]. One of the representative methods is the *Relational Knowledge Transfer* (RKT) method[10],

[1]School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China. [2]Department of Applied Mathematics, Northwestern Polytechnical University, Xi'an 710072, China. [3]Department of Data Analysis and Mathematical Modelling, Ghent University, Coupure links 653, Ghent 9000, Belgium. ✉email: wangchen@wtu.edu.cn; xiongp@wtu.edu.cn

which extracts the relational knowledge between the seen and the unseen semantic vectors using sparse coding theory. Following this line of research, the *Transferring Knowledge and preserving Data Structure* (TKDS) method[11] directly transfers knowledge from the seen domain to the unseen domain, relieving the domain shift to some extent. With the help of the unseen semantic information, the *Zero-Shot Learning based on Class Prototype and dual Latent Subspace learning with Reconstruction* (ZSL-CPLSR) method[12] builds a relationship between the seen and unseen classes, aiming to learn the unseen class prototypes and facilitate the subsequent subspace learning. Although these methods help to relieve the domain shift issue in zero-shot learning, their zero-shot recognition performances are not ideal for scene images due to large intra-class variations. To address this challenge, more performant zero-shot learning methods are still anticipated for scene recognition.

In this paper, in order to take advantage of the above two kinds of relationships, we propose to combine semantic autoencoders and visual relation transfer to tackle zero-shot scene recognition. Concretely, we first learn the relationship between the visual features and the semantic vectors by two semantic autoencoders for both the seen and the unseen scene classes. Then, considering that the semantic vectors (no attribute vectors available) are less effective than the visual features for scene images, we propose to learn the relationship between the seen and the unseen visual features and transfer this relational knowledge to generate unseen semantic vectors. Finally, we combine semantic autoencoders and visual relation transfer in a unified learning framework. Consequently, the proposed SAEVRT method is beneficial to generate more effective unseen semantic vectors and align the visual features and the semantic vectors in a better way, thereby enhancing the performance of zero-shot scene recognition. Our contributions are summarized as follows:

(1) We propose to combine semantic autoencoders and visual relation transfer in a unified framework to tackle zero-shot scene recognition. To the best of our knowledge, we are the first to take advantage of both the visual-semantic and seen-unseen relationships.
(2) Different from the seen and unseen semantic relation transfer, we propose an interpretable seen and unseen visual relation transfer method, which directly transfers the visual relation knowledge from the visual space to the semantic space, and hence learns more representative unseen semantic vectors.
(3) An iterative optimization algorithm is devised for the proposed unified learning framework. Comprehensive experiments on four public scene datasets demonstrate that the proposed method achieves a superior zero-shot recognition performance.

The remainder of this paper is structured as follows. Section “Related work” reviews the related work. Section "Combining semantic autoencoders and visualrelation transfer" introduces the proposed SAEVRT method for zero-shot scene recognition. Section “Experiments” presents the experimental results. Section “Conclusion” concludes this paper.

Related work

Zero-shot learning

Zero-shot learning has been attracting increasing attention as its powerful ability of recognizing new images from an unseen class[13,14]. In the early years, most researchers focused on learning more effective attribute vectors, since auxiliary semantic information plays the dominant role in zero-shot learning. For example, Lampert et al[15]. introduced attribute-based classification, where objects are classified based on the semantic attributes. Yazdanian et al[16]. used different kinds of descriptions to learn new discriminative attributes, which helps to improve the zero-shot recognition performance. Having obtained representative semantic vectors, researchers make great efforts to learn the relationship between the visual features and the semantic vectors. Li et al[17]. developed a *Dual Visual-Semantic Mapping Paths*(DMaP) method, which helps to train the visual-semantic mapping between the visual space and the semantic embedding space. Instead of inductive learning, Li et al[18]. proposed a transductive zero-shot learning method for zero-shot learning based on knowledge graph, where the relationship between the visual features and the semantic vectors is exploited in a graph convolutional network. Moreover, benefiting from the powerful ability of *Generative Adversarial Nets* (GANs), researchers are able to generate a sufficient number of seen images, and then transfer zero-shot classification into traditional classification tasks[19]. Xian et al[20]. integrated the Wasserstein GAN with a classification loss, which aims to produce sufficient and discriminative deep features. Tang et al[21]. proposed a *Structure-Aligned Generative Adversarial Network* (SAGAN) framework to improve the performance of zero-shot learning. However, these methods may generate some noisy images, resulting in a degenerated performance[22].

Recently, considering that the visual and semantic feature alignment plays a critical role in zero-shot learning, Cheng et al[23]. proposed a *Discriminative and Robust Attribute Alignment*(DRAA) method, which improves the discrimination power of the learned attribute embedding by contrastive learning. Given that the attribute vectors and the visual features contain complementary information, Hu et al[24]. proposed a complementary semantic information learning method, which consists of two branches: the *Attribute Refinement by Localization* branch and the *Visual-Semantic Interaction*branch. Yao et al[25]. proposed a *Deep Semantic Canonical Correlation Embedding*(DSCCE) model, which learns the visual and semantic feature interaction in an embedding reconstruction manner. From the perspective of contrastive learning, Jiang et al[26]. designed a *Dual Prototype Contrastive Network* (DPCN). However, the above methods neglect the relationship between the seen and the unseen spaces, which fails to promote the zero-shot recognition performance for scene images. For this reason, we focus on learning not only the relationship between the visual and the semantic spaces, but also the relationship between the seen and the unseen spaces.

Zero-shot scene recognition

Scene recognition has been widely used in many latent computer vision applications[27]. With the great success of deep learning and massive labelled images, the performance of traditional scene recognition has reached a saturation point. However, in real-world applications, the number of labelled images is insufficient. The extreme case is that the number of labelled images is zero, called zero-shot learning. To deal with this situation, Li et al[28]. first introduced zero-shot scene classification in the remote sensing community. Considering that there are no attribute vectors available, the authors used the natural language processing model Word2Vec to project the name of scene classes to the semantic vectors. In addition, they learned the relationship between the seen and the unseen scene classes by constructing a semantic-directed graph. Since then, more and more researchers have paid their attention to zero-shot remote sensing scene classification. For instance, Li et al[29]. introduced the *Locality-Preserving Deep Cross-Modal Embedding Network*(LPDCMEN), which helps to tackle the problem of class structure inconsistency in an end-to-end way. Quan et al[30]. proposed a semi-supervised Sammon embedding method to learn semantic prototypes, allowing to align a consistent class structure between the visual and the semantic prototypes. Different from these methods, other researchers utilize generative networks to generate unseen class samples for the zero-shot scene classification task. Ma et al[31]. proposed to augment the *Variational Autoencoder*(VAE) models with the GAN model to deal with zero-shot remote sensing image scene classification. Liu et al[32]. used two VAEs to project the visual features of image scenes and the semantic features of scene classes into a shared latent space. However, these generative networks face challenges in generating high-quality scene images[19]. Therefore, it is necessary to explore more refined zero-shot scene recognition methods.

Combining semantic autoencoders and visual relation transfer

In order to take advantage of both the visual-semantic and seen-unseen relationships, we propose to combine semantic autoencoders and visual relation transfer to tackle zero-shot scene recognition. The framework of our proposed SAEVRT method is illustrated in Fig. 1. First, we learn semantic autoencoders for both the seen and the unseen scene classes, which helps to build the visual-semantic relationship. Then, we construct the seen visual relation and transfer it into unseen classes, which helps to build the seen-unseen relationship. Third, we combine the above objectives into a unified learning framework. Subsequently, an alternating iterative strategy is devised for our unified learning framework, which generates more effective unseen semantic vectors. Finally, we can predict these unseen semantic vectors into the ground truth class prototypes, which achieves improved zero-shot learning performance.

Notation

For the N_s seen scene images, we denote the visual features as $\mathrm{X}^s = [\mathrm{x}_1^s, \mathrm{x}_2^s, \ldots, \mathrm{x}_{N_s}^s] \in \mathbb{R}^{d \times N_s}$, the semantic vectors as $\mathrm{S}^s = [\mathrm{s}_1^s, \mathrm{s}_2^s, \ldots, \mathrm{s}_{N_s}^s] \in \mathbb{R}^{r \times N_s}$, and the label values as $\mathrm{y}^s = [y_1^s, y_2^s, \ldots, y_{N_s}^s] \in \mathbb{R}^{1 \times N_s}$, where d and r represent the dimension of the visual features and the semantic vectors, respectively. For the N_u unseen

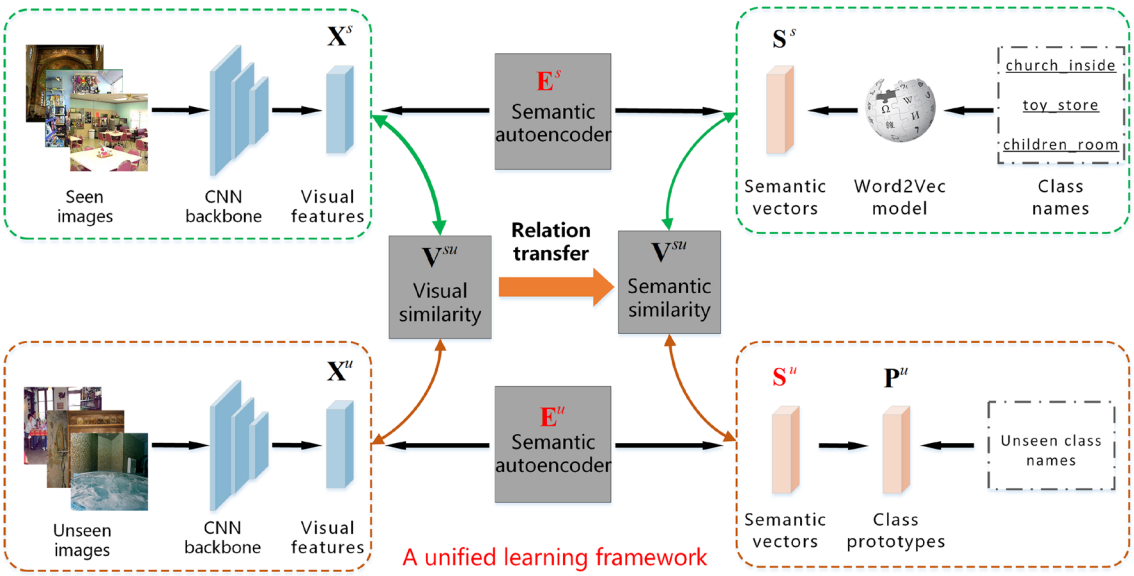


Fig. 1. The framework of the proposed SAEVRT method. The visual features and the semantic vectors are extracted based on the CNN backbone and the Word2Vec model, respectively. In the training phase, we learn two semantic autoencoders for both the seen and the unseen scene classes and transfer the seen visual relation into unseen classes, which are formulated in a unified learning framework, generating effective unseen semantic vectors. In the testing phase, we compare these unseen semantic vectors with ground truth class prototypes, yielding the final zero-shot learning results. Three unknown matrices that are learned based on the seen classes are marked with red colors. Note: three seen images and three unseen images in this figure are from the MIT67 dataset (https://web.mit.edu/torralba/www/indoor.html).

scene images, we denote the visual features as $\mathrm{X}^u = [\mathrm{x}_1^u, \mathrm{x}_2^u, \ldots, \mathrm{x}_{N_u}^u] \in \mathbb{R}^{d \times N_u}$, the semantic vectors as $\mathrm{S}^u = [\mathrm{s}_1^u, \mathrm{s}_2^u, \ldots, \mathrm{s}_{N_u}^u] \in \mathbb{R}^{r \times N_u}$, and the class labels as $\mathrm{y}^u = [y_1^u, y_2^u, \ldots, y_{N_u}^u] \in \mathbb{R}^{1 \times N_u}$. In zero-shot learning, the class labels in y^u are totally different from those in y^s, i.e., $\mathrm{y}^s \cap \mathrm{y}^u = \emptyset$. Note that in the training phase, the seen visual features X^s, the seen semantic vectors S^s, the seen label values y^s and the unseen visual features X^u are known, which are jointly used to learn the unseen semantic vectors S^u. In the testing phase, we can use the learned S^u to classify an unseen test image.

Formulation

1) Semantic autoencoders for both the seen and the unseen scene classes: Semantic autoencoders learn the relationship between the visual features and the semantic vectors, which helps to alleviate the domain shift between the visual and the semantic spaces. For the seen scene classes, in order to take the advantage of the semantic autoencoder, we use it to align the visual features and the semantic vectors, i.e.,

$$\min_{\mathrm{E}^s} \|\mathrm{X}^s - (\mathrm{E}^s)^{\mathrm{T}}\mathrm{S}^s\|_F^2 + \alpha\|\mathrm{E}^s\mathrm{X}^s - \mathrm{S}^s\|_F^2 , \tag{1}$$

here, $\mathrm{E}^s \in \mathbb{R}^{r \times d}$ represents the seen encoder matrix, and α is a hyperparameter. Our previous work[33] has demonstrated the necessity of training another semantic autoencoder for the unseen classes, since it helps to address the domain shift between the unseen visual features and the unseen semantic vectors. Consequently, we formulate the unseen semantic autoencoder as follows:

$$\min_{\mathrm{E}^u, \mathrm{S}^u} \|\mathrm{X}^u - (\mathrm{E}^u)^{\mathrm{T}}\mathrm{S}^u\|_F^2 + \beta\|\mathrm{E}^u\mathrm{X}^u - \mathrm{S}^u\|_F^2 , \tag{2}$$

where $\mathrm{E}^u \in \mathbb{R}^{r \times d}$ represents the unseen encoder matrix, and β is a hyperparameter.

According to Eqs. (1) and (2), E^s learns the relational knowledge of the seen domain, while E^u learns the relational knowledge of the unseen domain. Since these two encoder matrices come from two different domains, E^s is different from E^u to some extent. To address this issue, we minimize the discrepancy between E^s and E^u, i.e.,

$$\min_{\mathrm{E}^s, \mathrm{E}^u} \|\mathrm{E}^s - \mathrm{E}^u\|_F^2 . \tag{3}$$

This minimization problem aims to align the seen encoder matrix E^s and the unseen encoder matrix E^u.

2) Seen and unseen visual relation transfer: Relational knowledge transfer learns the relationship between the seen semantic vectors and the unseen semantic vectors, which helps to alleviate the domain shift issue between the seen and the unseen spaces, thus generating more effective unseen visual features. However, when we apply this method to zero-shot scene recognition, the performance is not optimal. The main reason lies in that the semantic vectors of scene images extracted by the word2vec model (no attribute vectors available) are less effective than the visual features trained by the deep CNN models. To this end, we propose an interpretable seen and unseen visual relation transfer, which learns the relationship between the seen visual features and the unseen visual features, and transfers this knowledge to obtain more effective unseen semantic vectors.

The motivation of the proposed visual relation transfer method is shown in Fig. 2. Basically, the greater the distances between the seen and the unseen visual features, the greater the distances between the seen and the unseen semantic vectors. Motivated by this, in order to transfer seen and unseen visual relation knowledge to the semantic space, we expect that the seen images with large distances to the unseen images (in the visual space) should have the smallest influence to represent the unseen images (in the semantic space). Therefore, the transfer weight coefficient is defined as:

$$\mathrm{v}_{ij} = \exp\left(-\frac{\|\mathrm{x}_i^s - \mathrm{x}_j^u\|_2^2}{2\sigma^2}\right) , \tag{4}$$

where σ is a predefined non-negative value. One can see that this coefficient is inversely proportional to the distance between the seen and the unseen visual features, which is consistent with our expectation. Thus, we can use the corresponding transfer weight matrix $\mathrm{V}^{su} \in \mathbb{R}^{N_s \times N_u}$ to represent the unseen semantic vectors, i.e.,

$$\mathrm{S}^u = \mathrm{S}^s\mathrm{V}^{su} . \tag{5}$$

This representation helps to generate unseen semantic vectors by transferring the visual relationship between the seen and the unseen classes. It is reasonable for a seen semantic vector to have a relatively large weight for a neighbor unseen semantic vector and a correspondingly small weight for other unseen semantic vectors.

3) Combining semantic autoencoders and visual relation transfer: Semantic autoencoders learn the relationship between the visual features and the semantic vectors, while visual relation transfer learns the relationship between the seen visual vectors and the unseen visual vectors. In order to promote the zero-shot recognition performance for scene images with large intra-class variations, we take advantage of not only the relationship between the visual and the semantic spaces, but also the relationship between the seen and the unseen spaces, and propose to combine these two relationships in a unified learning framework. To that end, we formulate the optimization problem of the proposed SAEVRT method by combining Eqs. (1)–(3) and (5) as follows:

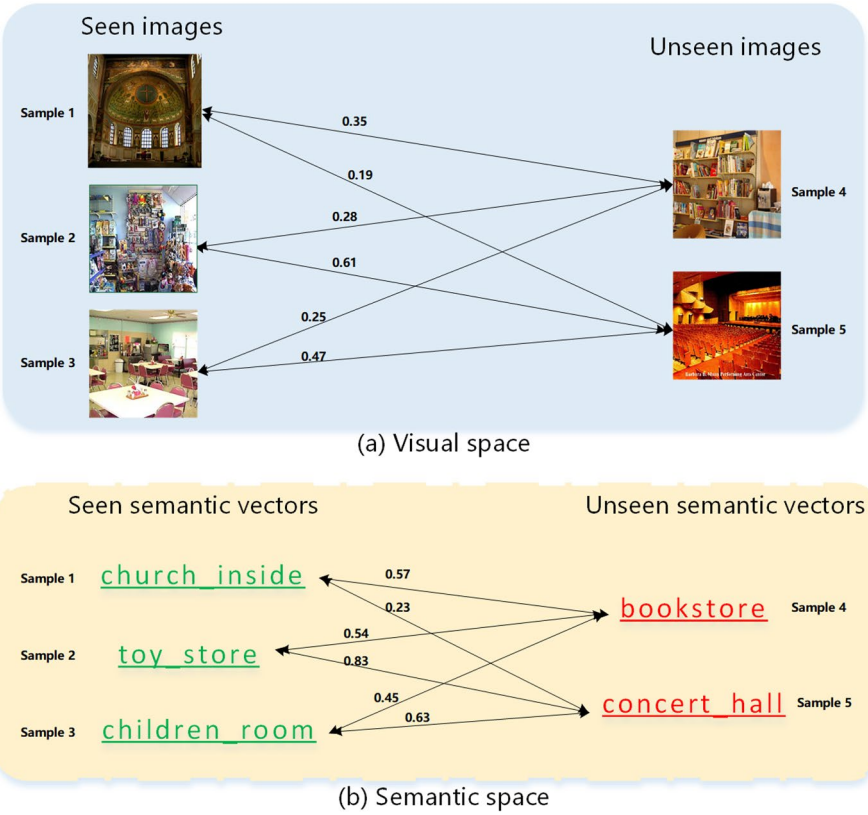


Fig. 2. The motivation of the proposed visual relation transfer method, where the values on the line represent the distances between two different samples. Generally, the greater the distances between the seen and the unseen visual features, the greater the distances between the seen and the unseen semantic vectors. Note: three seen images and two unseen images in this figure are from the MIT67 dataset (https://web.mit.edu/torralba/www/indoor.html).

$$\begin{aligned}\min_{\mathrm{E}^s,\mathrm{E}^u,\mathrm{S}^u} &\|\mathrm{X}^s-(\mathrm{E}^s)^{\mathrm{T}}\mathrm{S}^s\|_F^2+\alpha\|\mathrm{E}^s\mathrm{X}^s-\mathrm{S}^s\|_F^2\\ &+\lambda_1(\|\mathrm{X}^u-(\mathrm{E}^u)^{\mathrm{T}}\mathrm{S}^u\|_F^2+\beta\|\mathrm{E}^u\mathrm{X}^u-\mathrm{S}^u\|_F^2)\\ &+\lambda_2(\|\mathrm{E}^s-\mathrm{E}^u\|_F^2)\\ \mathrm{s.t.}\,&\mathrm{S}^u=\mathrm{S}^s\mathrm{V}^{su},\end{aligned} \tag{6}$$

where λ_1 and λ_2 are two hyperparameters. The equality constraint in Eq. (6) being difficult to satisfy, we relax it as:

$$\min_{\mathrm{S}^u}\|\mathrm{S}^s\mathrm{V}^{su}-\mathrm{S}^u\|_F^2\,. \tag{7}$$

Thus, Eq. (6) can be further recast as:

$$\begin{aligned}\min_{\mathrm{E}^s,\mathrm{E}^u,\mathrm{S}^u} &\|\mathrm{X}^s-(\mathrm{E}^s)^{\mathrm{T}}\mathrm{S}^s\|_F^2+\alpha\|\mathrm{E}^s\mathrm{X}^s-\mathrm{S}^s\|_F^2\\ &+\lambda_1(\|\mathrm{X}^u-(\mathrm{E}^u)^{\mathrm{T}}\mathrm{S}^u\|_F^2+\beta\|\mathrm{E}^u\mathrm{X}^u-\mathrm{S}^u\|_F^2)\\ &+\lambda_2(\|\mathrm{E}^s-\mathrm{E}^u\|_F^2)+\lambda_3(\|\mathrm{S}^s\mathrm{V}^{su}-\mathrm{S}^u\|_F^2)\,,\end{aligned} \tag{8}$$

where λ_3 is a hyperparameter.

It is noted that the proposed SAEVRT method has three advantages: (1) semantic autoencoders allows to address the domain shift issue between the visual and the semantic spaces; (2) visual relation transfer helps to address the domain shift issue between the seen and the unseen spaces; (3) by combining semantic autoencoders and visual relation transfer, we can transfer two kinds of relationships to generate more effective unseen semantic vectors, thereby improving the zero-shot scene recognition result.

Optimization

Considering that the proposed objective function in Eq. (8) is not jointly convex in E^s, E^u and S^u, an alternating iterative strategy is devised for our unified learning framework. The update of E^s depends on the calculation of E^u, the update of E^u depends on the calculation of E^s and S^u, and the update of S^u depends on the calculation of E^u. Therefore, we first need to initialize E^u. Since the semantic vectors of the unseen scene classes are unknown in zero-shot learning, the initialization of E^u depends on that of E^s.

Initializing E^s and E^u: In the training procedure, the visual features and semantic vectors of the seen scene classes are known, thus we can initialize E^s by learning the seen semantic autoencoder,

$$\min_{\mathrm{E}^s} \|\mathrm{X}^s - (\mathrm{E}^s)^{\mathrm{T}}\mathrm{S}^s\|_F^2 + \alpha\|\mathrm{E}^s\mathrm{X}^s - \mathrm{S}^s\|_F^2 . \tag{9}$$

Then, we take the derivative of Eq. (9) w.r.t. E^s and set it to 0, i.e.,

$$\mathrm{S}^s(\mathrm{S}^s)^{\mathrm{T}}\mathrm{E}^s + \mathrm{E}^s(\alpha\mathrm{X}^s(\mathrm{X}^s)^{\mathrm{T}}) = (1+\alpha)\mathrm{S}^s(\mathrm{X}^s)^{\mathrm{T}} . \tag{10}$$

Eq. (10) is a Sylvester equation, and it has the closed-form solution by using the Bartels–Stewart algorithm[34]. As an illustration, let $\mathrm{A} = \mathrm{S}^s(\mathrm{S}^s)^{\mathrm{T}}$, $\mathrm{B} = \alpha\mathrm{X}^s(\mathrm{X}^s)^{\mathrm{T}}$, $\mathrm{C} = (1+\alpha)\mathrm{S}^s(\mathrm{X}^s)^{\mathrm{T}}$, then the seen encoder matrix E^s is calculated as:

$$\mathrm{E}^s = \mathrm{sylvester}(\mathrm{A}, \mathrm{B}, \mathrm{C}) . \tag{11}$$

In Matlab, we can implement this step with the `sylvester` function.

By transferring the encoder knowledge from the seen space to the unseen space, we can initialize E^u with the help of E^s,

$$\mathrm{E}^u = \mathrm{E}^s . \tag{12}$$

Calculating S^u: With fixed E^s and E^u, the objective function in Eq. (8) is recast as:

$$\min_{\mathrm{S}^u} \lambda_1(\|\mathrm{X}^u - (\mathrm{E}^u)^{\mathrm{T}}\mathrm{S}^u\|_F^2 + \beta\|\mathrm{E}^u\mathrm{X}^u - \mathrm{S}^u\|_F^2) + \lambda_3\|\mathrm{S}^s\mathrm{V}^{su} - \mathrm{S}^u\|_F^2 . \tag{13}$$

We take the derivative of Eq. (13) w.r.t. S^u and set it to 0, i.e.,

$$\mathrm{S}^u = \left(\lambda_1(\mathrm{E}^u(\mathrm{E}^u)^{\mathrm{T}} + \beta\mathrm{I}) + \lambda_3\mathrm{I}\right)^{-1}((1+\beta)\lambda_1\mathrm{E}^u\mathrm{X}^u + \lambda_3\mathrm{S}^s\mathrm{V}^{su}) , \tag{14}$$

here, I means the identity matrix.

Updating E^u: With fixed E^s and S^u, the objective function in Eq. (8) is recast as:

$$\min_{\mathrm{E}^u} \lambda_1(\|\mathrm{X}^u - (\mathrm{E}^u)^{\mathrm{T}}\mathrm{S}^u\|_F^2 + \beta\|\mathrm{E}^u\mathrm{X}^u - \mathrm{S}^u\|_F^2) + \lambda_2(\|\mathrm{E}^s - \mathrm{E}^u\|_F^2) . \tag{15}$$

We take the derivative of Eq. (15) w.r.t. E^u and set it to 0, i.e.,

$$(\lambda_1\mathrm{S}^u(\mathrm{S}^u)^{\mathrm{T}} + \lambda_2\mathrm{I})\mathrm{E}^u + \mathrm{E}^u(\lambda_1\beta\mathrm{X}^u(\mathrm{X}^u)^{\mathrm{T}}) = \lambda_1(1+\beta)\mathrm{S}^u(\mathrm{X}^u)^{\mathrm{T}} + \lambda_2\mathrm{E}^s .$$

Let $\mathrm{A}^* = \lambda_1\mathrm{S}^u(\mathrm{S}^u)^{\mathrm{T}} + \lambda_2\mathrm{I}$, $\mathrm{B}^* = \lambda_1\beta\mathrm{X}^u(\mathrm{X}^u)^{\mathrm{T}}$, $\mathrm{C}^* = \lambda_1(1+\beta)\mathrm{S}^u(\mathrm{X}^u)^{\mathrm{T}} + \lambda_2\mathrm{E}^s$. The unseen encoder matrix E^u can be updated as:

$$\mathrm{E}^u = \mathrm{sylvester}(\mathrm{A}^*, \mathrm{B}^*, \mathrm{C}^*) . \tag{16}$$

Updating E^s: With fixed E^u and S^u, the objective function in Eq. (8) is recast as:

$$\min_{\mathrm{E}^s} \|\mathrm{X}^s - (\mathrm{E}^s)^{\mathrm{T}}\mathrm{S}^s\|_F^2 + \alpha\|\mathrm{E}^s\mathrm{X}^s - \mathrm{S}^s\|_F^2 + \lambda_2(\|\mathrm{E}^s - \mathrm{E}^u\|_F^2) . \tag{17}$$

We take the derivative of Eq. (17) w.r.t. E^s and set it to 0, i.e.,

$$(\mathrm{S}^s(\mathrm{S}^s)^{\mathrm{T}} + \lambda_2\mathrm{I})\mathrm{E}^s + \mathrm{E}^s(\alpha\mathrm{X}^s(\mathrm{X}^s)^{\mathrm{T}}) = (1+\alpha)\mathrm{S}^s(\mathrm{X}^s)^{\mathrm{T}} + \lambda_2\mathrm{E}^u .$$

Let $\mathrm{A}^{**} = \mathrm{S}^s(\mathrm{S}^s)^{\mathrm{T}} + \lambda_2\mathrm{I}$, $\mathrm{B}^* = \alpha\mathrm{X}^s(\mathrm{X}^s)^{\mathrm{T}}$, $\mathrm{C}^{**} = (1+\alpha)\mathrm{S}^s(\mathrm{X}^s)^{\mathrm{T}} + \lambda_2\mathrm{E}^u$. The seen encoder matrix E^s can be updated as:

$$\mathrm{E}^s = \mathrm{sylvester}(\mathrm{A}^{**}, \mathrm{B}^{**}, \mathrm{C}^{**}) . \tag{18}$$

Finally, we detail the pseudo-code for solving the proposed optimization problem in Algorithm 1.

Require:
The seen visual features X^s, the seen semantic vectors S^s, and the seen label values y^s; the unseen visual features X^u;
The hyperparameters α, β, λ_1, λ_2 and λ_3.
Ensure:
Initialize E^s and E^u by Eq. (11);
while not converge **do**
  Calculate S^u by Eq. (14);
  Update E^u by Eq. (16);
  Update E^s by Eq. (18);
end while
return The seen encoder matrix E^s; the unseen encoder matrix E^u and the unseen semantic vectors S^u.

Algorithm 1. SAEVRT

Classification

Having obtained the unseen encoder matrix E^u, we are able to use it to classify an unseen test image. Let x_i^u represent the visual features of the i-th unseen test image, and $\mathrm{P}^u = [\mathrm{p}_1^u, \mathrm{p}_2^u, \ldots, \mathrm{p}_{C_u}^u] \in \mathbb{R}^{r \times C_u}$ (C_u is the number of unseen scene classes) represent the unseen class prototypes extracted from the unseen scene classes. After that, the predicted class label of the unseen image x_i^u can be conducted in the semantic space ($\mathcal{S}$) or in the visual space ($\mathcal{V}$).

(1) $\mathcal{S}$: The unseen visual features x_i^u can be encoded onto the unseen semantic space as:

$$\mathrm{s}_i^u = \mathrm{E}^u \mathrm{x}_i^u \,. \tag{19}$$

Then, the predicted class label is obtained by calculating the cosine distance ($\mathrm{d}_{\cos}(\mathrm{x}, \mathrm{y}) = 1 - \frac{\mathrm{x} \cdot \mathrm{y}}{\|\mathrm{x}\| \|\mathrm{y}\|}$),

$$y_i^u = \underset{j \in \{1, \ldots, C_u\}}{\arg\min} \; \mathrm{d}_{\cos}(\mathrm{s}_i^u, \mathrm{p}_j^u) \,. \tag{20}$$

(2) $\mathcal{V}$: The unseen class prototypes p_j^u can be decoded back onto the unseen visual space as:

$$\tilde{\mathrm{x}}_j^u = (\mathrm{E}^u)^{\mathrm{T}} \mathrm{p}_j^u \,, \qquad j = 1, \ldots, C_u \,. \tag{21}$$

Then, the predicted class label is obtained by calculating the cosine distance,

$$y_i^u = \underset{j \in \{1, \ldots, C_u\}}{\arg\min} \; \mathrm{d}_{\cos}(\mathrm{x}_i^u, \tilde{\mathrm{x}}_j^u) \,. \tag{22}$$

Convergence and computational complexity analysis

In Algorithm 1, with initialized encoder matrices E^s and E^u, the calculation of S^u guarantees the closed-form solution. In additional, the updates of E^s and E^u have closed-form solutions by solving the corresponding Sylvester equations. To this end, with the iterative update rule, the SAEVRT method is able to converge to a local optimum.

For the computational cost of the proposed SAEVRT method, we only take the three iteration steps into account (the other steps could be pre-calculated). First, the main computational cost of calculating S^u depends on computing the inverse matrix, *i.e.*, $O(r^3)$. Second, the main computational cost of updating E^u and E^s depends on calculating the Sylvester equation[4], *i.e.*, $O(d^3 + r^3)$. To sum up, the total computational cost of Algorithm 1 is $O(T(2d^3 + 3r^3))$ (T indicates the total number of iteration steps).

Experiments

Datasets and set-up

Scene15[35]: This dataset contains 4,485 images from 15 indoor and outdoor scene classes, and each class contains 200 to 400 images. We show some example images in Fig. 3. Some scene classes present a high between-class similarity, which leads to greater difficulties in classifying unseen scene images. For the splits of seen/unseen classes, we randomly select 10 seen classes and 5 other classes as unseen classes.

MIT67[36]: This dataset contains 15,620 images from 67 indoor scene classes, and the number of images varies with at least 100 images per class. We show some example images in Fig. 4. Most indoor scene images consist of multiple objects and present large visual variations, which results in a degenerated zero-shot recognition performance. For the splits of seen/unseen classes, we randomly select 57 seen classes and 10 other classes as unseen classes.

UCM21[37]: This dataset contains 2,100 images from 21 remote sensing scene classes, and each class contains 100 images. We show some example images in Fig. 5. This is a widely-used scene dataset in the remote sensing

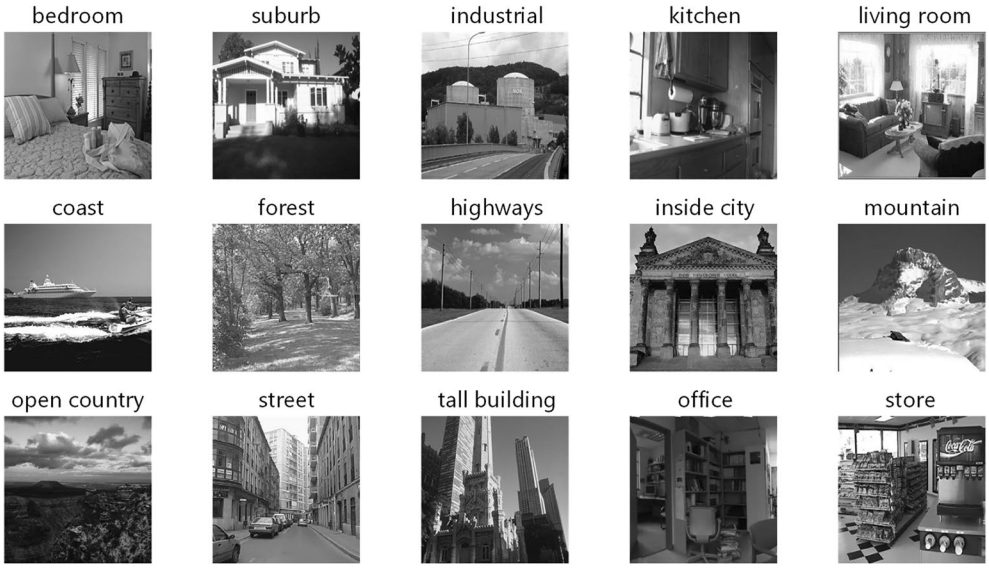


Fig. 3. Example indoor and outdoor scene images from the Scene15 dataset (https://github.com/TrungTVo/spatial-pyramid-matching-scene-recognition).

Fig. 4. Example indoor scene images from the MIT67 dataset (https://web.mit.edu/torralba/www/indoor.html).

community. For the splits of seen/unseen classes, we randomly select 16 seen classes and 5 other classes as unseen classes.

NWPU45[38]: This dataset contains 31,500 images from 45 remote sensing scene classes, and each class contains 700 images. We show some example images in Fig. 6. This is the most challenging large-scale remote sensing scene dataset. For the splits of seen/unseen classes, we randomly select 35 seen classes and 10 other classes as unseen classes.

Visual features: In order to obtain a better zero-shot recognition performance, we extract the visual features of all scene images by using the ResNet method[39] (pre-trained on the large-scale Places dataset[40]), which results in a 2048-dimensional vector.

Semantic vectors: Since there are no attribute vectors available for scene images, we extract semantic vectors of all scene classes by using the Word2Vec method (pre-trained on the large-scale Google News Corpus), which results in a 300-dimensional vector. When a class contains multiple words, we calculate the mean of the individual semantic vectors to obtain the final semantic vector[41].

We repeat all experiments 25 times under a random seen/unseen split and report the average recognition results. In our experiments, we discover that the zero-shot recognition performance in the visual space is superior to that in the semantic space, therefore, we only report the recognition accuracies in the visual space. For the five parameters, we first fix α and β at appropriate values, since they are relatively independent of the three other parameters. Then, we tune λ_1, λ_2 and λ_3 (with one of them fixed) by a grid search. As a result, on the

Fig. 5. Example remote sensing scene images from the UCM21 dataset (http://weegee.vision.ucmerced.edu/datasets/landuse.html).

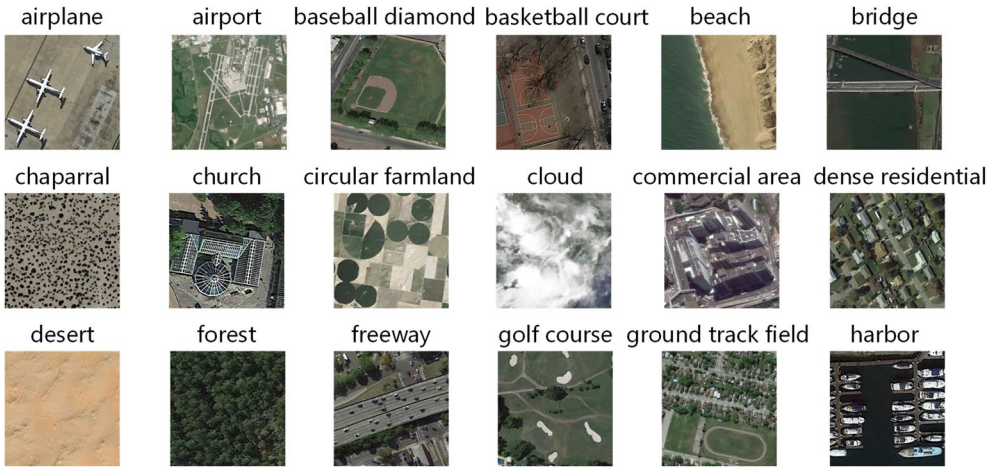


Fig. 6. Example remote sensing scene images from the NWPU45 dataset (https://gcheng-nwpu.github.io/).

Hyperparameter	Scene15	MIT67	UCM21	NWPU45
α	10^3	10^1	10^3	10^4
β	10^2	10^2	10^3	10^4
λ_1	10^2	10^2	10^2	10^3
λ_2	10^3	10^3	10^2	10^3
λ_3	1	10^{-3}	1	10^1

Table 1. Five hyperparameters on the different scene datasets.

four scene datasets, we set the values of these five hyperparameters as in Table 1. In Sect. "Parameter sensitivity analysis", we will discuss the parameter sensitivity in detail. The source code is available online at https://github.com/chenchenwang93/SAEVRT.

Comparison experiments

(1) Comparison with SAE-based methods: To investigate the effectiveness of the proposed SAEVRT method, we first compare it with four widely-used SAE-based methods. SAE[4] trains an encoder to project the visual features into the semantic space and sets a decoder (based on the encoder) to reconstruct the visual space; DDSA[8] learns two semantic autoencoders to project the visual features and the semantic vectors to a shared and discriminative space; CAE[42] learns two semantic autoencoders for both the seen and the unseen classes. Different from the CAE method, DIPL[5] uses the nearest unseen class to learn the unseen semantic autoencoder. To make a fair comparison, we conduct the experiments in the same settings. The experimental results are

Method	Scene15	MIT67	UCM21	NWPU45
SAE	56.87 ± 8.77	63.28 ± 9.65	49.50 ± 8.42	44.81 ± 6.73
DDSA	56.89 ± 8.69	64.18 ± 9.92	51.71 ± 7.65	47.50 ± 6.05
CAE	59.20 ± 8.95	65.18 ± 10.19	54.89 ± 9.28	49.13 ± 6.68
DIPL	61.43 ± 9.14	65.97 ± 10.23	56.54 ± 9.95	50.42 ± 5.49
SAEVRT	**63.77 ± 9.71**	**67.75 ± 10.35**	**58.68 ± 10.04**	**53.26 ± 6.11**

Table 2. Zero-shot recognition results (± standard deviation) of the SAEVRT method compared with those of the SAE-based methods (%).

Method	Scene15	MIT67	UCM21	NWPU45
RKT	55.68 ± 9.62	60.73 ± 9.53	45.26 ± 7.83	43.72 ± 6.30
TKDS	56.50 ± 8.98	61.05 ± 10.11	48.86 ± 8.54	46.25 ± 6.75
ZSL-CPLSR	59.71 ± 10.40	65.82 ± 10.25	54.26 ± 9.48	50.87 ± 5.66
SAEVRT	**63.77 ± 9.71**	**67.75 ± 10.35**	**58.68 ± 10.04**	**53.26 ± 6.11**

Table 3. Zero-shot recognition results (± standard deviation) of the SAEVRT method compared with those of the RKT-based methods (%).

reported in Table 2. Inspecting these results, we have the following conclusions. (1) On the four scene datasets, CAE and DIPL achieve a superior performance compared with SAE and DDSA, since the first two methods take the unseen classes into consideration. (2) The proposed method provides a better performance than the four other SAE-based methods. Since SAEs solely build the relationship between the visual features and the semantic vectors, they fail to build the relationship between the seen and the unseen classes. VRT significantly compensates this inherent limitation in SAEs, which helps to learn more effective unseen semantic vectors, thus improving the zero-shot recognition performance.

(2) Comparison with RKT-based methods: Considering that SAEVRT combines semantic autoencoders and visual relation transfer, we also compare it with three RKT-based methods. RKT[10] learns the relational knowledge between the seen semantic vectors and the unseen semantic vectors based on sparse coding theory; TKDS[11] directly transfers knowledge from the seen domain to the unseen domain; ZSL-CPLSR[12] builds the relationship between the seen classes and the unseen classes to learn the unseen class prototypes. We conduct the these experiments in the same settings. The experimental results are presented in Table 3. Inspecting these results, we draw two conclusions. (1) ZSL-CPLSR performs better than RKT and TKDS. This is because that ZSL-CPLSR contains two stages, i.e., class prototype learning by using RKT and dual latent subspace learning with the help of reconstruction. (2) On the four scene datasets, the proposed method performs better than ZSL-CPLSR, since the visual relation transfer uses the more effective visual features rather than the semantic vectors.

(3) Comparison with state-of-the-art methods: We also compare the proposed SAEVRT method with some state-of-the-art (SOTA) zero-shot learning methods. Two baselines are the SAE and RKT methods; SSE[43] learns similarity functions to project the unseen domain samples into the seen semantic space; DMaP[17] learns the intrinsic connection between the manifold structure and the transferability of the mapping between the visual features and the semantic vectors; WDVSc[44] learns three categories of visual structure constraint to address the domain shift for transductive zero-shot learning; moreover, VSOP[45] exploits the shared subspace using both the visual and the semantic features. Based on GAN, LsrGAN[46] transfers knowledge from the seen classes to the unseen classes by leveraging semantic information; SAGAN[21] introduces a structure-aligned module to explore the structural consistency between the visual and the semantic spaces. Moreover, three recently proposed zero-shot learning methods, i.e., DSCCE[25], DRAA[23] and ZGLR[47], are used for comparison for the Scene15 and MIT67 scene datasets; two recently proposed zero-shot scene classification methods, i.e., Ma et al[31]. and Liu et al[32]., are used for comparison for the UCM21 and NWPU45 remote sensing scene datasets.

To fully evaluate the effectiveness of the SAEVRT method, we conduct comparison experiments under different splits. For the four scene datasets, the experimental results are reported in Tables 4–7. Observing these results, we obtain the following three conclusions. (1) SAE achieves a better performance than RKT. For example, the performance of SAE increases 4.37%, 1.19% and 6.66% points compared to RKT under three different splits for the Scene15 dataset. (2) LsrGAN and SAGAN perform better than two baseline zero-shot learning methods (SAE and RKT). The main reason is that the GAN-based methods help to generate more appropriate unseen samples with the relationship between the visual features and the semantic vectors, or the relationship between the seen classes and the unseen classes. (3) The recently proposed methods achieve a better performance than the generative network methods. For example, the performance of DRAA is 2.54% points higher than that of SAGAN under the split of 57/10 for the MIT67 dataset; the performance of the method of Liu et al. is 2.21% points higher than that of LsrGAN under the split of 35/10 for the NWPU45 dataset. (4) SAEVRT obtains the highest recognition accuracy among the other zero-shot learning methods. By combining SAEs and VRT in a unified learning framework, we exploit both the visual-semantic and seen-unseen relationships, which generates more effective unseen semantic vectors, thereby improving the zero-shot scene recognition accuracy.

Method	8/7	10/5	12/3
SSE	28.63 ± 9.78	49.83 ± 9.09	66.49 ± 10.34
RKT	35.46 ± 9.85	55.68 ± 9.62	71.37 ± 10.12
DMaP	37.44 ± 9.69	56.24 ± 7.87	76.65 ± 11.34
SAE	39.83 ± 10.25	56.87 ± 8.77	78.03 ± 11.89
WDVSc	40.62 ± 12.17	61.05 ± 8.58	80.16 ± 13.76
VSOP	40.56 ± 10.90	60.31 ± 11.14	79.55 ± 12.01
LsrGAN	41.40 ± 10.24	62.54 ± 9.58	80.47 ± 12.14
SAGAN	41.12 ± 11.65	63.00 ± 9.16	81.42 ± 12.90
DSCCE	41.71 ± 11.41	63.31 ± 10.27	80.98 ± 12.21
DRAA	**42.32 ± 12.43**	63.25 ± 9.89	82.25 ± 12.19
ZGLR	41.96 ± 12.15	63.40 ± 10.01	81.67 ± 12.57
SAEVRT	42.05 ± 11.08	**63.77 ± 9.71**	**82.81 ± 12.73**

Table 4. Zero-shot recognition results (± standard deviation) of the SAEVRT method compared with those of SOTA methods (under 8/7, 10/5 and 12/3 splits) for the Scene15 dataset (%).

Method	52/15	57/10	62/5
SSE	33.51 ± 6.80	52.23 ± 8.08	65.84 ± 9.95
RKT	42.36 ± 7.53	60.73 ± 9.53	71.51 ± 9.67
DMaP	45.27 ± 8.79	62.64 ± 9.81	73.14 ± 10.62
SAE	49.08 ± 7.16	62.28 ± 9.65	74.65 ± 11.61
WDVSc	54.45 ± 9.05	63.96 ± 11.90	76.62 ± 10.93
VSOP	53.98 ± 8.71	64.09 ± 10.36	76.43 ± 11.22
LsrGAN	56.74 ± 8.20	64.85 ± 10.19	77.26 ± 10.05
SAGAN	58.13 ± 9.57	64.02 ± 9.78	76.88 ± 10.27
DSCCE	57.87 ± 8.49	65.63 ± 9.40	78.01 ± 12.38
DRAA	58.54 ± 8.83	66.56 ± 10.19	77.90 ± 12.55
ZGLR	58.10 ± 8.07	66.87 ± 10.23	**78.73 ± 12.40**
SAEVRT	**59.96 ± 8.29**	**67.75 ± 10.35**	78.58 ± 12.32

Table 5. Zero-shot recognition results (± standard deviation) of the SAEVRT method compared with those of SOTA methods (under 52/15, 57/10 and 62/5 splits) for the MIT67 dataset (%).

Method	14/7	16/5	18/3
SSE	20.87 ± 6.85	35.59 ± 5.90	49.38 ± 9.91
RKT	32.15 ± 7.07	45.26 ± 7.83	58.45 ± 10.67
DMaP	36.41 ± 7.96	48.92 ± 8.71	62.33 ± 9.86
SAE	37.94 ± 8.75	49.50 ± 8.42	64.02 ± 10.90
WDVSc	40.28 ± 10.12	55.91 ± 11.77	66.69 ± 11.14
VSOP	40.21 ± 9.71	46.48 ± 7.83	66.72 ± 12.37
LsrGAN	41.40 ± 10.54	53.71 ± 9.03	68.44 ± 12.45
SAGAN	42.56 ± 10.68	55.65 ± 10.67	67.31 ± 13.13
Ma et al.	43.38 ± 10.05	56.14 ± 10.06	69.00 ± 13.55
Liu et al.	43.63 ± 9.76	57.77 ± 10.50	68.83 ± 13.89
SAEVRT	**44.69 ± 10.65**	**58.68 ± 10.04**	**70.55 ± 13.32**

Table 6. Zero-shot recognition results (± standard deviation) of the SAEVRT method compared with those of SOTA methods (under 14/7, 16/5 and 18/3 splits) for the UCM21 dataset (%).

To further discuss the zero-shot recognition results for each unseen scene class in detail, we present the confusion matrices of the proposed SAEVRT method. Here, we conduct experiments under the seen/unseen splits illustrated in Section 4.1. For the Scene15 dataset, when the five unseen classes are 'Kitchen', 'Tall building', 'Inside city', 'Mountain', and 'Street', the experimental results are reported in Fig. 7. Most of the images in the class 'Inside city' are misclassified onto the class 'Street'. For the MIT67 dataset, when the ten unseen classes are 'Computer room', 'Elevator', 'Laboratory wet', 'Church inside', 'Bowling', 'Warehouse', 'Corridor', 'Shoe shop',

Method	30/15	35/10	40/5
SSE	23.30 ± 2.48	33.36 ± 3.58	44.56 ± 6.13
RKT	34.18 ± 3.41	43.72 ± 4.30	53.36 ± 6.64
DMaP	38.07 ± 4.83	49.53 ± 6.31	57.17 ± 6.85
SAE	35.07 ± 3.91	44.81 ± 4.73	55.42 ± 7.49
WDVSc	40.92 ± 4.59	50.68 ± 6.60	60.73 ± 7.18
VSOP	36.09 ± 4.63	45.32 ± 5.71	59.49 ± 7.38
LsrGAN	37.69 ± 4.54	50.26 ± 5.75	62.31 ± 8.40
SAGAN	38.50 ± 4.40	51.51 ± 6.10	62.66 ± 8.53
Ma et al.	40.33 ± 3.71	52.23 ± 6.48	64.15 ± 8.81
Liu et al.	41.06 ± 4.38	52.47 ± 6.50	63.90 ± 9.45
SAEVRT	**42.56 ± 3.45**	**53.26 ± 6.11**	**65.70 ± 8.37**

Table 7. Zero-shot recognition results (± standard deviation) of the SAEVRT method compared with those of SOTA methods (under 30/15, 35/10 and 40/5 splits) for the NWPU45 dataset (%).

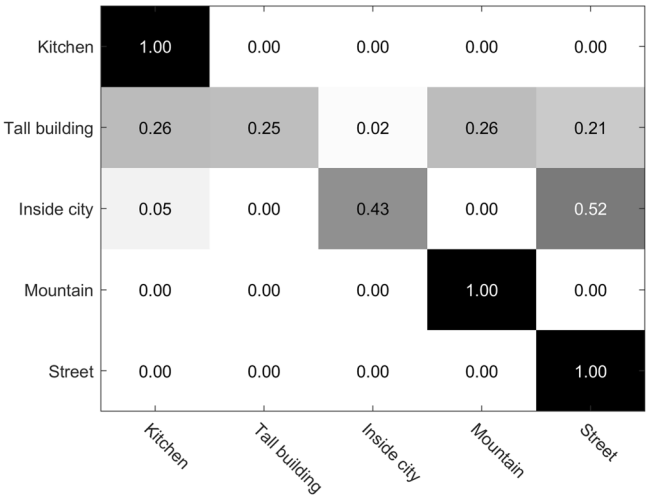


Fig. 7. Confusion matrices of the zero-shot learning results for the Scene15 dataset.

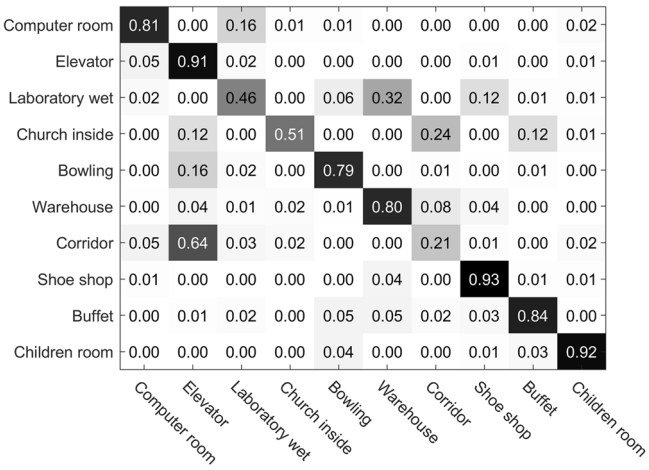


Fig. 8. Confusion matrices of the zero-shot learning results for the MIT67 dataset.

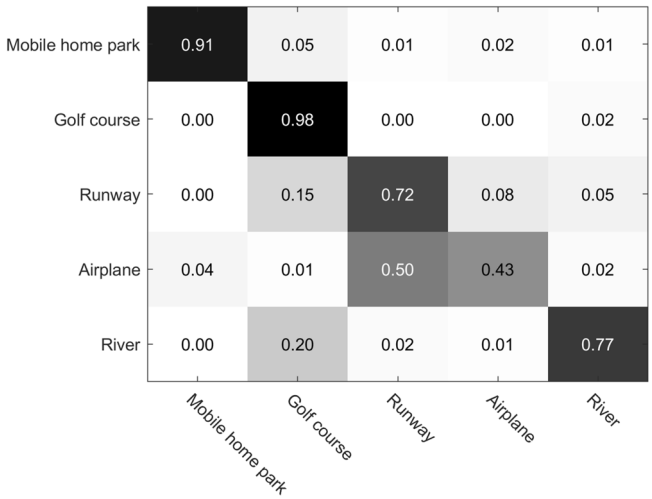


Fig. 9. Confusion matrices of the zero-shot learning results for the UCM21 dataset.

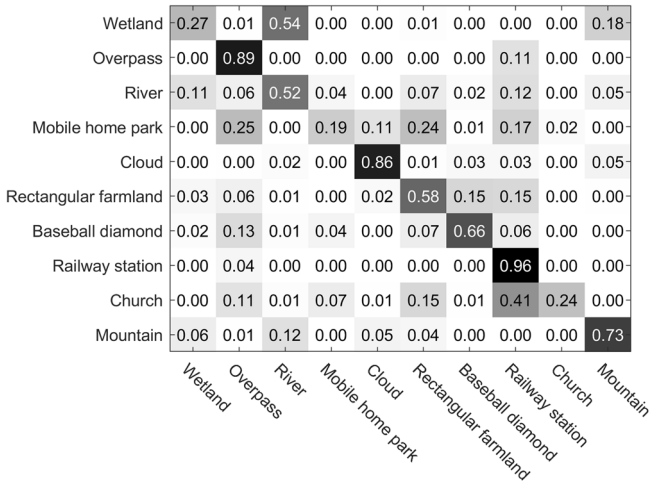


Fig. 10. Confusion matrices of the zero-shot learning results for the NWPU45 dataset.

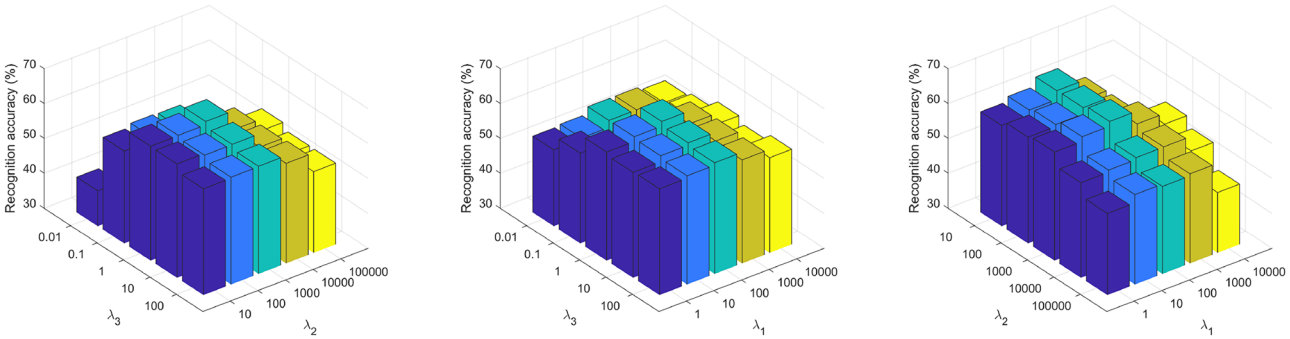
'Buffet', and 'Children room', the experimental results are reported in Fig. 8. Most of the images in the class 'Corridor' are misclassified onto the class 'Elevator', since these two classes share some semantic information. For the UCM21 dataset, when the five unseen classes are 'Mobile home park', 'Golf course', 'Runway', 'Airplane', and 'River', the experimental results are reported in Fig. 9. The recognition accuracies for the classes 'Mobile home park' and 'Golf course' almost reach the maximum. For the NWPU45 dataset, when the ten unseen classes are 'Wetland', 'Overpass', 'River', 'Mobile home park' 'Cloud', 'Rectangular farmland', 'Baseball diamond', 'Railway station', 'Church', and 'Mountain', the experimental results are reported in Fig. 10. The recognition accuracy for the class 'Railway station' almost reaches the maximum, whereas the recognition accuracies for the classes 'Mobile home park' and 'Church' are rather poor. The main reason is that the remote sensing scene images in these two classes have complex backgrounds.

Ablation study

In order to observe the influence of each term in the proposed optimization problem on the zero-shot recognition results, we evaluate three variants of SAEVRT. First, we remove the third term in the objective function ($\lambda_1 = 0$ in Eq. (8)) and record it as SAEVRT-λ_1. Second, we remove the fourth term in the objective function ($\lambda_2 = 0$ in Eq. (8)) and record as SAEVRT-λ_2. Third, we remove the fifth term in the objective function ($\lambda_3 = 0$ in Eq. (8)) and record as SAEVRT-λ_3. The experimental results are reported in Table 8. Overall, we can observe that the zero-shot recognition results of SAEVRT are consistently higher than those of the other variants. More specifically, for the Scene15 dataset, when removing the third term, the recognition result drops 7.22% points; when removing the fourth term, the recognition result drops 2.04% points; when removing the fifth term, the recognition result drops 4.31% points. For the UCM21 dataset, when removing the third term, the recognition result drops 6.59% points; when removing the fourth term, the recognition result drops 1.06% points; when

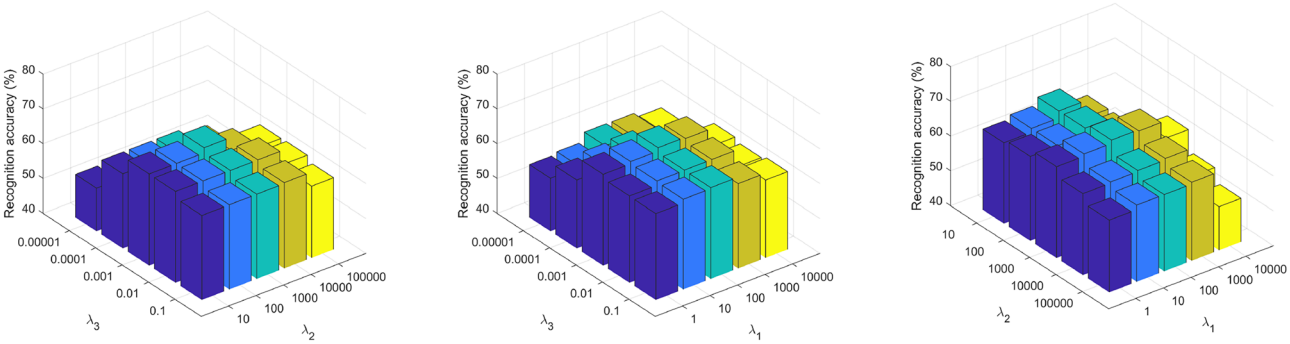
Method	Scene15	MIT67	UCM21	NWPU45
SAEVRT-λ_1	56.55 ± 8.24	63.91 ± 7.55	52.09 ± 10.68	47.08 ± 5.96
SAEVRT-λ_2	61.73 ± 10.53	66.45 ± 10.39	57.62 ± 11.81	51.75 ± 6.78
SAEVRT-λ_3	59.46 ± 9.76	65.76 ± 10.41	56.89 ± 13.24	50.48 ± 6.39
SAEVRT	**63.77 ± 9.71**	**67.75 ± 10.35**	**58.68 ± 10.04**	**53.26 ± 6.11**

Table 8. Zero-shot recognition results (± standard deviation) of the SAEVRT method compared with those of three variants (%).



(a) $\lambda_1 = 10^2$ (b) $\lambda_2 = 10^3$ (c) $\lambda_3 = 1$

Fig. 11. Recognition accuracies with different values of λ_1, λ_2 and λ_3 for the Scene15 dataset.



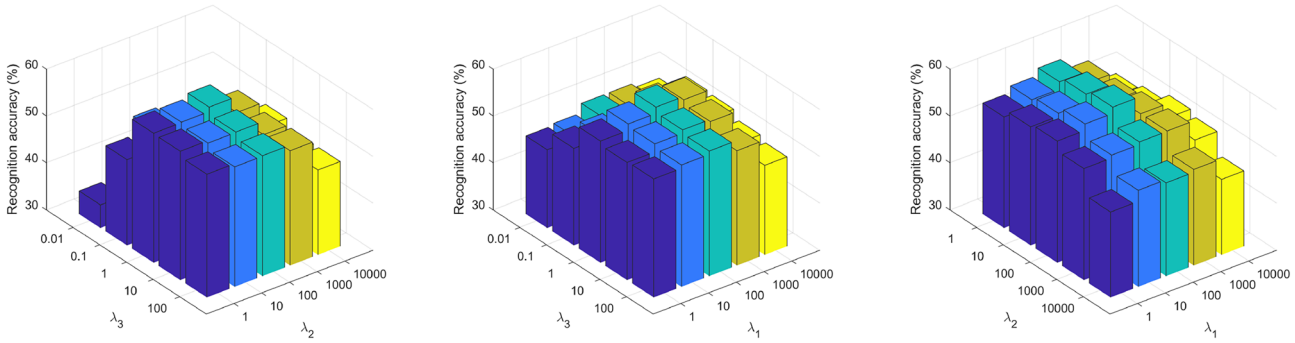
(a) $\lambda_1 = 10^2$ (b) $\lambda_2 = 10^3$ (c) $\lambda_3 = 10^{-3}$

Fig. 12. Recognition accuracies with different values of λ_1, λ_2 and λ_3 for the MIT67 dataset.

removing the fifth term, the recognition result drops 1.79% points. These experimental results demonstrate that combining semantic autoencoders and visual relation transfer is helpful to exploit not only the relationship between the visual and the semantic spaces, but also the relationship between the seen and the unseen spaces. For the two other scene datasets, we obtain the similar results. Therefore, we can confirm that the three terms play a positive role in improving the performance of zero-shot scene recognition.

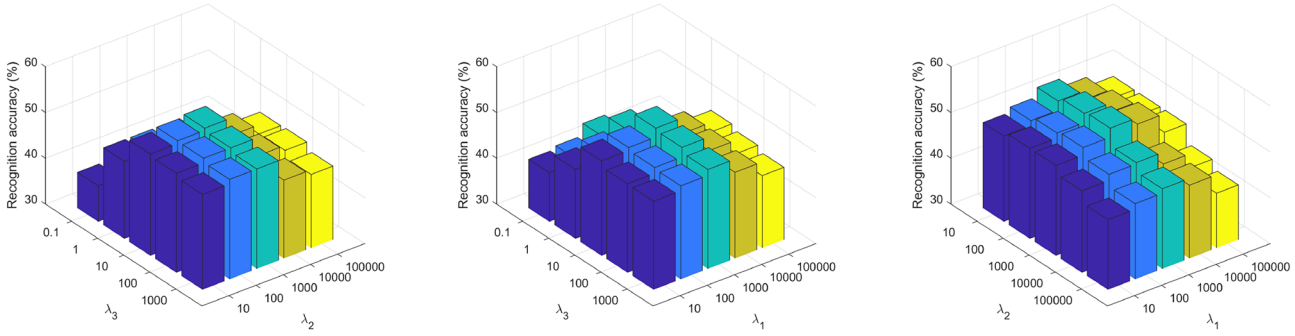
Parameter sensitivity analysis

The proposed SAEVRT method contains three important hyperparameters λ_1, λ_2 and λ_3. Parameter λ_1 adjusts the importance of the seen and the unseen semantic autoencoders; λ_2 controls the consistency of the seen and the unseen encoder matrices; and λ_3 controls the importance of the visual relation transfer. In order to fully explore how the hyperparameters impact the recognition performance, we use the four scene datasets to conduct parameter sensitivity analysis. For the Scene15 dataset, we first fix parameter λ_1 at 10^2 and observe the recognition accuracies when varying the two other parameters λ_2 and λ_3. Then, we fix parameter λ_2 at 10^3 and observe the recognition accuracies when varying the two other parameters λ_1 and λ_3. Finally, we fix parameter λ_3 at 1 and observe the recognition accuracies when varying the two other parameters λ_1 and λ_2. The experimental results are presented in Fig. 11 (a)–(c). We can notice that when the value of λ_1 is gradually



(a) $\lambda_1 = 10^2$ (b) $\lambda_2 = 10^2$ (c) $\lambda_3 = 1$

Fig. 13. Recognition accuracies with different values of λ_1, λ_2 and λ_3 for the UCM21 dataset.



(a) $\lambda_1 = 10^3$ (b) $\lambda_2 = 10^3$ (c) $\lambda_3 = 10^1$

Fig. 14. Recognition accuracies with different values of λ_1, λ_2 and λ_3 for the NWPU45 dataset.

increased, the recognition results fist improve and then decline. Moreover, when the value of λ_2 is greater than or equal to 10^3, this has an important influence on the recognition accuracies; when the value of λ_3 is less than or equal to 1, this also has an important influence on the recognition accuracies. For the three other datasets, the experimental results are presented in Figs. 12,13,14, where similar trends can be observed. Therefore, λ_1, λ_2 and λ_3 play a crucial role to compromise the importance of these objective functions in the unified learning framework.

Discussion

Examining Tables 4–7, on the four scene datasets, the proposed SAEVRT method obtains the best zero-shot recognition result among all competing methods under standard splits, i.e., 63.77%, 67.75%, 58.68%, and 53.26% on Scene15, MIT67, UCM21, and NWPU45, respectively. However, SAEVRT may be inferior to the other methods under different splits. For example, on the Scene15 dataset, SAEVRT achieves a recognition accuracy of 42.05% under the 8/7 split, while DRAA achieves a recognition accuracy of 42.32% under the same split. On the MIT67 dataset, SAEVRT achieves a recognition accuracy of 78.58% under the 62/5 split, while ZGLR achieves a recognition accuracy of 78.73% under the same split. The reason may be that the proposed visual relation transfer method cannot effectively exploit the seen-unseen relationship under a relatively small or large seen/unseen split, which fails to transfer the visual relation knowledge from the visual space to the semantic space, thus degenerating the zero-shot learning recognition result. Therefore, it is not convincing to evaluate zero-shot recognition performance based on a fixed split.

In addition, we can observe that almost all standard deviations are quite large. Actually, when conducting the randomized experiments, we find that if the randomly chosen unseen scene classes are significantly different from the seen scene classes, the zero-shot recognition performance declines sharply, since it is difficult for knowledge transfer. Accordingly, we should pay more attention to some unseen scene classes that are significantly different from seen scene classes.

Conclusion

In this paper, we have proposed a novel zero-shot scene recognition method by combining semantic autoencoders and visual relation transfer. By learning two semantic autoencoders for both the seen and the unseen classes, our proposed method has alleviated the domain shift problem between the visual features and the semantic vectors. By transferring the visual relationship from the seen classes to the unseen classes, our proposed method

is able to generate more effective unseen semantic vectors, and align the visual features and the semantic vectors in a better way. Comprehensive experiments on four public scene datasets have confirmed the effectiveness of the proposed SAEVRT method, achieving recognition accuracies of 63.77%, 67.75%, 58.68%, and 53.26% on Scene15, MIT67, UCM21, and NWPU45, respectively.

Zero-shot scene recognition assumes that the test images solely come from the unseen scene classes. However, this assumption in not really practical, since the real scene images may come from both the seen scene classes and the unseen scene classes. To overcome this shortcoming, generalized zero-shot learning (GZSL)[48,49] has been introduced. In our future work, we will try to extend the proposed method to tackle this generalized zero-shot scene recognition task.

Data availability

This study used four scene image datasets Scene15, MIT67, UCM21, and NWPU45, which ares available at https://github.com/TrungTVo/spatial-pyramid-matching-scene-recognition, https://web.mit.edu/torralba/www/indoor.html, http://weegee.vision.ucmerced.edu/datasets/landuse.html, and https://gcheng-nwpu.github.io/, respectively. The source code of this study is available online at https://github.com/chenchenwang93/SAEVRT.

Received: 31 August 2025; Accepted: 7 November 2025
Published online: 19 December 2025

References

1. Xie, L., Lee, F., Liu, L., Kotani, K. & Chen, Q. Scene recognition: A comprehensive survey. *Pattern Recognition.* **102**, 107205 (2020).
2. Wang, C., Peng, G. & De Baets, B. Deep feature fusion through adaptive discriminative metric learning for scene recognition. *Information Fusion.* **63**, 1–12 (2020).
3. Pourpanah, F. et al. A Review of Generalized Zero-Shot Learning Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **45**(4), 4051–4070 (2023).
4. Kodirov, E., Xiang, T. & Gong, S. Semantic Autoencoder for Zero-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4447–4456 (2017).
5. Zhao, A., et al. Domain-invariant projection learning for zero-shot recognition. In *Advances in Neural Information Processing Systems.* 1019–1030 (2018).
6. Han, X. et al. Design of a turbo-based deep semantic autoencoder for marine Internet of Things. *Internet of Things.* **28**, 101393 (2024).
7. Chen, S. et al. TransZero++: Cross Attribute-Guided Transformer for Zero-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **45**(11), 12844–12861 (2023).
8. Liu, Y., Li, J. & Gao, X. A Simple Discriminative Dual Semantic Auto-encoder for Zero-shot Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 940–941 (2020).
9. Yao, H., Zhang, C., Wei, Y., Jiang, M. & Li, Z. Graph Few-Shot Learning via Knowledge Transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 6656–6663 (2020).
10. Wang, D., Li, Y., Lin, Y. & Zhuang, Y. Relational knowledge transfer for zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 2145–2151 (2016).
11. Li, X., Fang, M. & Wu, J. Zero-shot classification by transferring knowledge and preserving data structure. *Neurocomputing.* **238**, 76–83 (2017).
12. Zhao, P., Zhang, S., Liu, J. & Liu, H. Zero-shot Learning via the fusion of generation and embedding for image recognition. *Information Sciences.* **578**, 831–847 (2021).
13. Fang, J. et al. Zero-shot learning via categorization-relevant disentanglement and discriminative samples synthesis. *The Visual Computer.* **40**, 3889–3901 (2024).
14. Chen, S., Hong, Z., You, X. & Shao, L. Semantics-Conditioned Generative Zero-Shot Learning via Feature Refinement. *International Journal of Computer Vision.* **133**, 4504–4521 (2025).
15. Lampert, C. H., Nickisch, H. & Harmeling, S. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **36**(3), 453–465 (2014).
16. Yazdanian, R., Shojaee, S.M. & Baghshah, M.S. An attribute learning method for zero-shot recognition. In *Proceedings of the Iranian Conference on Electrical Engineering.* 2235–2240 (2017).
17. Li, Y., Wang, D., Hu, H., Lin, Y., & Zhuang, Y. Zero-shot recognition using dual visual-semantic mapping paths. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3279–3287 (2017).
18. Li, Q., Sun, X. & Dong, J. Transductive zero-shot learning via knowledge graph and graph convolutional networks. *Scientific Reports.* **15**, 28708 (2025).
19. Xu, B., Zeng, Z., Lian, C. & Ding, Z. Generative Mixup Networks for Zero-Shot Learning. *IEEE Transactions on Neural Networks and Learning Systems.* **36**(3), 4054–4065 (2025).
20. Xian, Y., Lorenz, T., Schiele, B. & Akata, Z. Feature Generating Networks for Zero-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 5542–5551 (2018).
21. Tang, C., He, Z., Li, Y. & Lv, J. Zero-Shot Learning via Structure-Aligned Generative Adversarial Network. *IEEE Transactions on Neural Networks and Learning Systems.* **33**(11), 6749–6762 (2022).
22. Wang, Y., Hong, M., Huangfu, L., & Huang, S. Data Distribution Distilled Generative Model for Generalized Zero-Shot Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 5695–5703 (2024).
23. Cheng, D. et al. Discriminative and Robust Attribute Alignment for Zero-Shot Learning. *IEEE Transactions on Circuits and Systems for Video Technology.* **33**(8), 4244–4256 (2023).
24. Hu, X., Wang, Z. & Li, J. Learning complementary semantic information for zero-shot recognition. *Signal Processing: Image Communication.* **115**, 116965 (2023).
25. Yao, Z., Jiang, Q., Zhong, W. & Gu, X. Deep Semantic Canonical Correlation Embedding for Zero-Shot Industrial Process Fault Diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems.* **55**(7), 4458–4471 (2025).
26. Jiang, H. et al. Dual Prototype Contrastive Network for Generalized Zero-Shot Learning. *IEEE Transactions on Circuits and Systems for Video Technology.* **35**(2), 1111–1122 (2025).
27. Wang, C., Peng, G. & De Baets, B. Joint global metric learning and local manifold preservation for scene recognition. *Information Sciences.* **610**, 938–956 (2022).
28. Li, A., Lu, Z., Wang, L., Xiang, T. & Wen, J. R. Zero-shot scene classification for high spatial resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing.* **55**(7), 4157–4167 (2017).
29. Li, Y., Zhu, Z., Yu, J. G. & Zhang, Y. Learning Deep Cross-Modal Embedding Networks for Zero-Shot Remote Sensing Image Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing.* **59**(12), 10590–10603 (2021).

30. Quan, J., Wu, C., Wang, H., & Wang, Z. Structural alignment based zero-shot classification for remote sensing scenes. In *Proceedings of the IEEE International Conference on Electronics and Communication Engineering*. 17–21 (2018).
31. Ma, S., Liu, C., Li, Z. & Yang, W. Integrating Adversarial Generative Network with Variational Autoencoders towards Cross-Modal Alignment for Zero-Shot Remote Sensing Image Scene Classification. *Remote Sensing.* **14**(18), 4533 (2022).
32. Liu, C., Ma, S., Li, Z., Yang, W. & Han, Z. Mining Contrastive Relations Between Cross-Modal Features for Zero-Shot Remote Sensing Image Scene Classification. *IEEE Geoscience and Remote Sensing Letters.* **21**, 1–5 (2024).
33. Wang, C., Peng, G. & De Baets, B. A distance-constrained semantic autoencoder for zero-shot remote sensing scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.* **14**, 12545–12556 (2021).
34. Bartels, R. H. & Stewart, G. W. Solution of the matrix equation AX+ XB= C [F4]. *Communications of the ACM.* **15**(9), 820–826 (1972).
35. Lazebnik, S., Schmid, C., & Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*. 2169–2178 (2006).
36. Quattoni, A. & Torralba, A. Recognizing indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition;* 413–420 (2009).
37. Yang, Y. & Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 270–279 (2010).
38. Cheng, G., Han, J. & Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE.* **105**(10), 1865–1883 (2017).
39. He, K., Zhang, X., Ren, S., & Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 770–778 (2016).
40. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A. & Torralba, A. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **40**(6), 1452–1464 (2018).
41. Sumbul, G., Cinbis, R. G. & Aksoy, S. Fine-grained object recognition and zero-shot learning in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing.* **56**(2), 770–779 (2018).
42. Sun, G., Wu, S., Gao, G., Wu, F. & Jing, X.Y. Coupled autoencoders learning for zero-shot classification with domain shift. In *Proceedings of the International Conference on Progress in Informatics and Computing*. 66–70 (2017).
43. Zhang, Z. & Saligrama, V. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE International Conference on Computer Vision*. 4166–4174 (2015).
44. Wan, Z., et al. Transductive zero-shot learning with visual structure constraint. In *Advances in Neural Information Processing Systems*. 9972–9982 (2019).
45. Wu, H. et al. Joint Visual and Semantic Optimization for zero-shot learning. *Knowledge-Based Systems.* **215**, 106773 (2021).
46. Vyas, M.R., Venkateswara, H. & Panchanathan, S. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *Proceedings of the European Conference on Computer Vision*. 70–86 (2020).
47. Wang, Q., Mou, H., Wang, J., Wei, C. & Zhou, Y. Zero-shot learning based on the fusion of global and local representations. *Measurement Science and Technology.* **36**(3), 035905 (2025).
48. Verma, V.K., Arora, G., Mishra, A. & Rai P. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4281–4289 (2018).
49. Chen, Y. & Zhou, Y. Incorporating attribute-level aligned comparative network for generalized zero-shot learning. *Neurocomputing.* **573**, 127188 (2024).

Acknowledgements

The authors gratefully acknowledge the editors and reviewers for their insightful comments and constructive suggestions, which have significantly enhanced the quality of this work.

Author contributions

Chen Wang contributes to Writing–original draft, Validation, Methodology, Data curation. Man Wang contributes to Writing–review and editing, Validation, Methodology. Guohua Peng contributes to Writing–review and editing, Methodology. Bernard De Baets contributes to Writing–review and editing, Validation, Supervision. Xiong Pan contributes to Writing–review and editing, Supervision. All authors reviewed the manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 62301373, and in part by the Natural Science Foundation of Hubei Province under Grant 2023AFB279. BDB received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.W. or X.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.