JARS Journal of Applied Remote Sensing

RESEARCH PAPER

Self-supervised embedding for generalized zero-shot learning in remote sensing scene classification

Rambabu Damalla[a,*] **Rajeshreddy Datla**[b] **Chalavadi Vishnu**[c] **and Chalavadi Krishna Mohan**[a]

[a]Indian Institute of Technology Hyderabad, Kandi, Telangana, India
[b]Advanced Data Processing Research Institute, Department of Space, Secunderabad, Telangana, India
[c]University of Agder, Grimstad, Norway

ABSTRACT. Generalized zero-shot learning (GZSL) is the most popular approach for developing ZSL, which involves both seen and unseen classes in the classification process. Many of the existing GZSL approaches for scene classification in remote sensing images use word embeddings that do not effectively describe unseen categories. We explore word embedding to describe the classes of remote sensing scenes to improve the classification accuracy of unseen categories. The proposed method uses a data2vec embedding based on self-supervised learning to obtain a continuous and contextualized latent representation. This representation leverages two advantages of the standard transformer architecture. First, targets are not predefined as visual tokens. Second, latent representations preserve contextual information. We conducted experiments on three benchmark scene classification datasets of remote sensing images. The proposed approach demonstrates its efficacy over the existing GZSL approaches.

Keywords: remote sensing images; scene classification; machine learning; zero-shot learning

Paper 230083SS received Mar. 1, 2023; revised Jul. 6, 2023; accepted Jul. 25, 2023; published Aug. 12, 2023.

1 Introduction

The advancements in remote sensing platforms focus on the acquisition of high-resolution imagery and provide challenges in understanding the abundant volumes semantically. Scene classification is a preliminary task that is helpful in analyzing such volumes of remote sensing images at the coarser level. It aims to assign a label to a given scene from a set of predefined categories based on its content. Most of the works[1–3] address the problem of scene classification using supervised learning by exploring convolutional neural networks (CNNs). With the large coverage of remote sensing satellite images, the tedious annotation process for all categories of scenes becomes not possible practically. With its innate capability in accommodating the new unseen/undiscovered classes, zero-shot learning (ZSL) would benefit many existing categorization applications[4–6] of remote sensing imagery.

ZSL[7] is one such task that helps in understanding the scenes only with the description of the classes without involving any sample of new class during training. Hence, in ZSL, training and testing sets are disjointed. ZSL could be accomplished by sharing the semantic information from seen to unseen class samples. Here, semantic information is a high-level description of the

*Address all correspondence to Rambabu Damalla, cs19resch11006@iith.ac.in

classes; how it can be obtained and transferred to unseen classes is discussed in Sec. 2. ZSL approaches can be divided into two categories: conventional ZSL (CZSL) and generalized ZSL (GZSL). The objective of CZSL is to predict only unseen classes, whereas GZSL predicts both seen and unseen classes of samples.[8] These techniques are illustrated in Fig. 1. GZSL is more challenging than CZSL as many unseen classes are prone to being misclassified into one of the seen classes at the testing phase.

Mainly, the GZSL-based approaches focus on realistic images and solve the issues of mapping from visual features to semantic embeddings[9,10] and seen-unseen bias.[11,12] Nevertheless, numerous CZSL techniques have been explored in classifying remote sensing (RS) images, whereas GZSL is hardly explored in remote sensing images. To the best of our knowledge, we are the first to explore GZSL for scene classification tasks in remote sensing images. GZSL aims to categorize the RS samples for both seen and unseen classes by establishing a mapping relation between the feature and semantic spaces. With overhead imaging, the semantics of an image may ignore other modalities, such as the elevation of objects from the ground, which are described by digital elevation models.[13] To incorporate information from other modalities, word embeddings become an alternative in describing other modalities rather than utilizing them explicitly in deriving the semantics. Generally, both seen and unseen classes are represented as semantic vectors in terms of word/sentence embeddings[14] and attribute vectors[14] in the embedding space. For feature extraction, we use the models (e.g., AlexNet,[15] VGG,[16] GoogLeNet,[17] and ResNet[18]) pre-trained on ImageNet,[19] which ignore the cross-dataset bias[20] between the ImageNet dataset and remote sensing benchmark datasets. Often, the cross-dataset bias results in low-standard visual features for GZSL in remote sensing scene classification (RSSC), which would affect the classification accuracy on both seen and novel scene classes.

In general, the methods used to extract visual features of remote sensing images are poor due to cross-dataset bias[20] as they rely on ImageNet pre-trained models.[19] Also, feature vectors obtained from word2vec representation are limited by fixed representation irrespective of the context. Hence, they are not effective in achieving appropriate semantics. By alleviating these issues, extracted visual and semantic features can be enhanced to improve GZSL-RSSC classification performance. Thus, we propose a method called "GZSL for RSSC using data2vec representations," termed GZSL-RSD2V. The proposed method GZSL-RSD2V uses a feature enhancement (FE)[21] module to obtain discriminative features, which can effectively enhance the visual features. Also, the proposed method uses a data2vec model based on a standard transformer architecture to obtain continuous and contextualized hidden features. The representation of the data2vec model has two benefits: (i) targets are not fixed as visual tokens, and (ii) latent representations preserve contextual information. We conducted experiments using GZSL-RSD2V for scene classification in remote-sensing images.

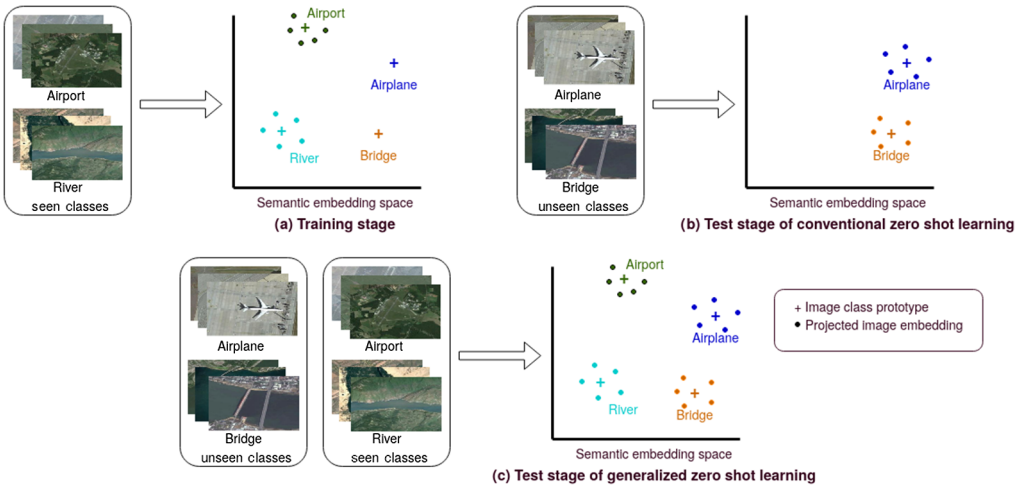


Fig. 1 Illustration of CZSL and GZSL. (a) Training stage, (b) test stage of CZSL, and (c) test stage of GZSL.

The main contributions of this paper are summarized as follows.

- We propose feature - variational autoencoder generative adversial networks (f-VAEGAN) to learn a mapping of semantics to the visual domain for visual feature generation.
- A practical embedding approach based on a standard transformer architecture is developed to represent semantic features of remote sensing images.
- We introduce an FE module to refine both seen and unseen class visual features.
- Our representation demonstrates the compactness of within-class similarity and separability of inter-class variations.

The remainder of this paper is organized as follows. Section 2 presents various methods for scene classification using ZSL and embedding approaches for encoding semantic information. In Sec. 3, we explain the proposed GZSL-RSD2V. The experimental results and the analysis of the proposed approach over the existing GZSL approaches are discussed in Sec. 4. The conclusion of this paper is provided in Sec. 5.

2 Related Work

This section presents the existing ZSL approaches and some important methods explored for encoding semantic information.

2.1 Zero-Shot Learning

ZSL-based scene classification in remote sensing images is divided into two categories, as explained in the following subsections.

2.1.1 *Embedding-based methods*

These methods aim to map seen class samples and their class semantic vectors into embedding space, and then a nearest neighbour search in the embedding space is used to classify unseen class samples with their class semantic vectors. In the domain of remote sensing images, a method based on label propagation is proposed for ZSL.[22] The label propagation mechanism helps construct a semantic-directed graph to share the semantic information from seen to unseen classes, thereby classifying the test image into one of the unseen classes.

Quan et al.[23] employed the Shannon embedding method to implement ZSL for scene classification in remote sensing images. This method alters the features in the semantic space with the respective features in the visual space for maintaining the class structure consistency between visual and semantic space. Another work[24] proposes a semantic auto-encoder-based method to impose conditions on the distance to align the visual and semantic spaces for ZSL in remote sensing images. Further, a technique[25] is used to map semantic space from visual space by training a projection network to perform ZSL tasks in remote sensing images. With the learned mapping function, semantic knowledge is perhaps transferred during the inference of unseen classes. However, embedding-based methods cannot perform well in GZSL settings considering unseen classes are essentially biased[26,27] to seen classes during the testing process. This motivates us to explore generative-based methods for ZSL in remote sensing images.

2.1.2 *Generative methods for zero-shot learning*

Initially, we train a generative model to generate unseen class image features for data augmentation. Later, we learn a classifier (CLS) to classify seen features and generate novel class features to perform the ZSL task. To implement the ZSL task, we utilize the latest works on generative models, such as variational autoencoders (VAEs),[26,28,29] GANs,[11,30,31] and generative flows.[32] Xian et al.[11] were the first to use generative adversarial networks (GANs)[33] to map semantic to visual features, giving a state-of-the-art proposal for ZSL. Li et al.[34] first implemented the ZSL task in remote sensing images using GANs by achieving within-class similarity and outside-class discrimination.

The description of the semantic information is as follows.

2.2 Semantic Information

In ZSL, only seen class images are available during training. Semantic vectors of remote sensing scene categories are a bridge between both seen and unseen class images to classify unseen classes. These semantics enable us to perform ZSL. Semantic information can be extracted from semantic attributes or word vectors.

2.2.1 *Semantic attributes*

Semantic or manually defined attributes are high-level descriptions of objects, such as objects' color or shape. Unseen classes can be recognized based on semantic attributes, but human annotation is required. As an example of natural image analysis, the "CUB dataset" was annotated with 312 semantic attributes corresponding to 200 different bird classes.[35] However, the remote sensing benchmark datasets' semantic attributes have not yet been explored.

2.2.2 *Word embeddings*

In general, natural language processing models such as word2vec,[36] glove,[37] fastText,[38] which are trained over a corpus of one trillion words often results in very high dimensional vector representation. They do not require human annotation. However, they have some limitations. In the word2vec model, each word has a fixed representation irrespective of its importance in the context, so the vector representation does not provide the promised performance. Also, they contain intense noise, which compromises the model's performance. To overcome the above limitations of the word2vec model, we explore representation from the data2vec model[39] as word embedding in ZSL. The data2vec tries to predict a contextualized latent representation based on the limited view of the input sample. The representation of data2vec has two benefits. First, targets are not fixed as visual tokens. Second, hidden representations preserve contextual information.

3 Generalized Zero-Shot Learning for Remote Sensing Scene Classification Using Data2vec Representations

The block diagram of the proposed GZSL-RSD2V for GZSL is shown in Fig. 2. It comprises f-VAEGAN,[40] an FE module[21] and a CLS. In Fig. 2(a), f-VAEGAN aims to synthesize visual features during the training process from the semantics vector of data2vec embedding "d." Here, we introduce the FE module to determine discriminative seen visual features in conjunction with f-VAEGAN. Specifically, the FE module is optimized to learn discriminative features using joint center-triplet (JCT) loss and iterative semantic consistency (ISC) loss.[21] In Fig. 2(b), we enhance the visual features for both seen and unseen class samples with the help of trained FE. Then, we train both enhanced seen and unseen class features using a CLS for classification purposes. Finally, we classify the enhanced unseen features using the trained CLS at the testing phase.

3.1 Formulation

Let L^s and L^u indicate the sets of seen and unseen class samples, respectively. We indicate seen class samples as $S = \{f_i, l_i\}_{i=1}^{N}$, where f_i represents the visual feature; l_i is a respective class label $\in L^s$; and N is the total number of seen samples. The relationship between seen and unseen sets is defined as $L^s \cap L^u = \phi$, and $L = L^s \cup L^u$. We denote a set of semantic vectors for every seen and unseen class as $dj \in D$, $\forall\ j \in L$, which helps to share semantic information from the seen to unseen class samples.

3.2 Data2vec Embedding

The data2vec model is "a general framework for self-supervised learning"[39] that works over several methods, such as vision, speech, and language. But, here, we use only data2vec for the vision model. The model data2vec obtains continuous and contextualized hidden features of input data. The main idea of data2vec is to regress contextualized hidden representations based on a masked view of the input. It has a teacher and student network trained on a standard

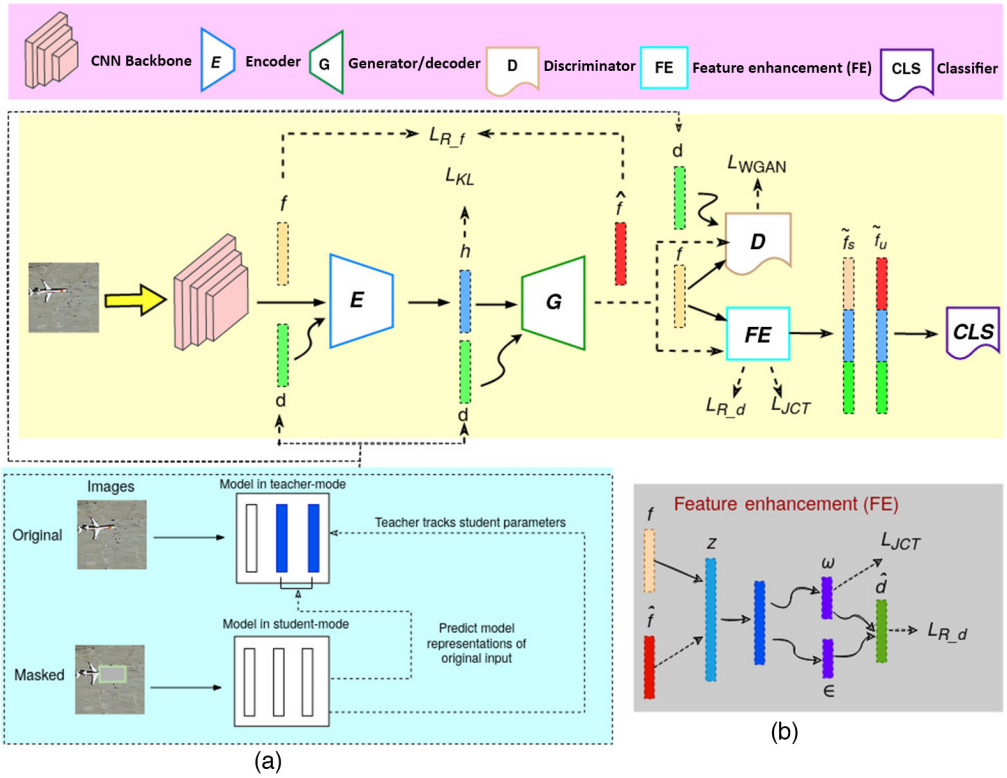


Fig. 2 Block diagram for the proposed method GZSL-RSD2V. (a) The GZSL-RSD2V comprises three modules: f-VAEGAN, an FE module, and a CLS. The f-VAEGAN module aims to synthesize visual features during the training process from the semantics vector of data2vec embedding "*d*," whereas the FE module determines discriminative seen visual features in conjunction with f-VAEGAN. We enhance the visual features for both seen and unseen class samples with the help of trained FE. Then, we train to classify both enhanced seen and unseen class features using a CLS. (b) FE module architecture. FE is optimized to learn discriminative features using JCT loss and ISC loss; its discriminative features from different layers are then concatenated to obtain enhanced features.

transformer architecture.[41] The teacher network generates contextualized representations of the full input data. The student network tries to predict full data representations based on the "block-wise masking view"[42] of the input sample. Despite that, data2vec uses a mask for 60% of the patches instead of 40%. The weights of the teacher network are updated based on the exponentially decaying average[43,44] of the student. Then, the target is made using the transformer's top K blocks, which are continual and contextualized. Prior methods predict targets lacking contextualized information. On the other hand, the data2vec model predicts contextualized latent target representations by embodying related features from the total image contrary to targets that accommodate information solitary to the present patch, such as visual tokens or pixels.

3.3 Feature Generating Models

We use f-VAEGAN[40] as a baseline for generating synthetic CNN features to map from semantic vectors to visual features conditioned on the data2vec embedding d. The f-VAEGAN uses feature generating VAE (f-VAE)[45] and feature generating wasserstein GAN (f-WGAN)[11] to improve the feature generator. f-VAE comprises an encoder $E(f,d)$ and a decoder $\mathrm{Dec}(h,d)$. Here, the encoder converts input f to hidden features h, and a decoder $\mathrm{Dec}(h,d)$ rebuilds input f from h. The loss function for f-VAE is as follows:

$$L_{\mathrm{VAE}} = \mathrm{KL}(q(h|f,d)\|p(h|d)) - \mathbb{E}_{q(h|f,d)}[\log\ p(f|h,d)], \tag{1}$$

where the conditional distribution $q(h|f,d)$ is modeled as $\mathbb{E}(f,d)$, $p(h|d))$ is considered to be $N(0,1)$, KL is the Kullback–Leibler divergence, and $p(f|h,d)$ is equal to $\mathrm{Dec}(h,d)$. In f-WGAN,

generator $G(h,d)$ generates a synthetic CNN feature $\hat{f}$ from random input noise h_p. In contrast, the discriminator $D(f,d)$ tries to discriminate the real and synthetic features. f-WGAN returns a real value between 0 and 1, optimizing the following:

$$L_{\text{WGAN}} = \mathbb{E}[D(f,d)] + \mathbb{E}[D(\hat{f},d)] - \eta\mathbb{E}[(\|\nabla_{f'}D(f',d)\|_2 - 1)^2], \tag{2}$$

where $\hat{f} = G(h,d)$ is the synthetic feature, $f' = \rho f + (1-\rho f)$ with $\rho \sim \text{U}(0,1)$, and η is the penalty multiplier. The parameters of the decoder $\text{Dec}(h,d)$ and generator $G(h,d)$ are shared to improve the feature generator.

3.4 Feature Enhancement Module

The cross-dataset bias is alleviated by processing the visual features of remote-sensing images through an FE module. Here, the FE module is constrained to JCT loss and ISC loss.

3.4.1 *Joint center-triplet loss*

This loss is introduced to learn discriminative features. These features are obtained by encouraging the features of the same class label to stay together and features of different class labels to be away from each other; this is defined as the compactness of within-class similarity and the separability of inter-class variations, respectively. This could be achieved with the help of class label information, center loss, and triplet loss. JCT loss is formally defined as follows:

$$L_{\text{JCT}}(\hat{d},l,l') = \max(0,\Gamma + \psi\|\omega - c_l\|_2^2 - (1-\psi)\|\omega - c_{l'}\|_2^2), \tag{3}$$

where c_l is the l'th class center, $c_{l'}$ is the l''th class center, Γ denotes the margin that controls the separability of intra-class pairs from inter-class pairs, ω represents the encoded features in FE, and $\psi \in [0,1]$ denotes the balancing factor to indicate the compactness of within class similarity and separability of inter-class variations.

3.4.2 *Iterative semantic consistency loss*

This loss is introduced at the last layer of the FE module to learn semantic features. ISC loss generates the semantic features $\hat{d}$ from f or $\hat{f}$ using the "reparameterization trick."[45] To learn effective semantic features, ISC loss is applied to synthetic semantic features to ensure that synthesized semantic features are mapped from the original semantic vectors. This loss is achieved using the l_1 reconstruction loss and is formally defined as follows:

$$L_{R\_d} = \mathbb{E}[\|\hat{d}_{\text{real}} - d\|_1] + \mathbb{E}[\|\hat{d}_{\text{syn}} - d\|_1], \tag{4}$$

where $\hat{d}_{\text{real}}$ represents semantic features synthesized from f with the help of FE and $\hat{d}_{\text{syn}}$ represents semantic features synthesized from $\hat{f}$. Note that $\hat{d} = \hat{d}_{\text{real}} \cup \hat{d}_{\text{syn}}$ and d represents the semantic features for the given visual features f or $\hat{f}$.

3.4.3 *Extracting enhanced features*

In this stage, we take out enhanced features $\tilde{f}_s$ and $\tilde{f}_u$ from the trained FE. Using the residual connection,[18] we combine visual features f, respective latent vector $z_s \in Z$, and semantic embedding $\tilde{d}_s \in D$ as $\tilde{f}_s$. Similarly, we combine visual features $\hat{f}$, respective latent vector z_u, and semantic embedding $\tilde{d}_u$ as $\tilde{f}_u$. Figure 2(b) illustrates the fully enhanced features $\tilde{f}_s$ and $\tilde{f}_u$, formally defined as follows:

$$\tilde{f}_s = f \odot z_s \odot \tilde{d}_s, \tag{5}$$

$$\tilde{f}_u = \tilde{f}_u \odot z_u \odot \tilde{d}_u, \tag{6}$$

where $\odot$ denotes the concatentation operation and $\tilde{f}_s$ and $\tilde{f}_u \in \tilde{F}$. Hence, visual features $\tilde{f}_s$ and $\tilde{f}_u$ are enhanced as discriminative features that are class- and semantically appropriate to avoid ambiguities within feature samples of the distinct classes.

Finally, our model GZSL-RSD2V is trained with the following overall objective function:

$$L_{\text{total}} = L_{\text{VAE}} + L_{\text{WGAN}} + \lambda_{\text{JCT}} L_{\text{JCT}} + \lambda_{R\_d} L_{R\_d}, \tag{7}$$

where λ_{JCT} and λ_{R_d} are hyperparameters of the JCT loss and ISC loss multipliers, respectively.

4 Experimental Results and Analysis

This section provides the results and analysis of the proposed approach for GZSL for scene classification in remote sensing images. We demonstrated the efficacy of the proposed approach on three benchmark datasets of scene classification: UCMercedLandUse (UCM21),[46] WHU-RS19 (RS19),[47] and aerial image dataset (AID30).[48]

4.1 Details of Scene Classification Datasets

UCM21 is the 21-class land use RSSC benchmark dataset manually extracted from large images from the US Geological Survey. The RS19 dataset contains 19 scene classes extracted from Google Earth with various high-resolutions. AID30 is a 30-class scene classification in the RSI. Table 1 provides details of these datasets.

4.2 Implementation

Our proposed method employs an encoder, generator, and discriminator, which are basically multilayer perceptrons. Each perceptron accommodates a 4096-node hidden layer with LeakyReLU activation. The FE module is also a multilayer perceptron. It holds two hidden layers with 4096 nodes and $2 \times |\hat{d}|$ nodes with LeakyReLU, followed by an encoding layer that uses two feature vectors of size $|\hat{d}|$ to constitute the second hidden layer. Its final layer $|\hat{d}|$ corresponds to the semantic vector of the word embedding method (e.g., $|\hat{d}| = 768$ for the data2vec). We used the Adam optimizer[49] with $\beta 1 = 0.5$ and $\beta 2 = 0.999$. The penalty multiplier η is set to 10. In this study, hyper parameters of the JCT loss multiplier (λ_{JCT}), ISC loss multiplier (λ_{R_d}), and gamma (ψ) are set to 0.999.

4.3 Extraction of Visual and Semantic Features

Deep learning-based, fine-tuned features of 2048 in size from remote sensing image scenes are extracted from the ResNet-101[18] model pre-trained on ImageNet.[19] Semantic prototypes are extracted from the data2vec model. Here, the data2vec model predicts contextualized hidden features of the entire input image based on a masked version of the input sample in a self-refined setup using a standard transformer architecture.[41] We used word2vec word embeddings pre-trained on the Google News Corpus[36] for a fair comparison. The details of word embedding dimensions are shown in Table 2.

4.4 Quantitative Analysis

We used the unified evaluation protocol[11] for fair comparison to evaluate our proposed approach. We assess the top-1 accuracy for seen and unseen class samples (indicated by S and U, respectively). The harmonic mean (indicated by H) of S and U is also estimated using $H = (2 \times S \times U)/(S + U)$.

Table 1 Details of three benchmark datasets for scene classification in remote sensing images.

Characteristic	UCM21	RS19	AID30
Number of classes	21	19	30
Images per class	100	50	220 to 420
Number of images	2100	950	10,000
Dimension of each image	256 × 256	600 × 600	600 × 600
Seen/unseen split ratio	16/5	15/4	25/5

Table 2 Details of semantic vectors extracted from different embeddings over three datasets.

Method-size $\lvert\hat{d}\rvert$	UCM21	RS19	AID30
word2vec - dim	300	300	300
data2vec - dim	768	768	768

All of the zero-shot RSSC experiments are reiterated 25 times, accompanied by a random seen/unseen split, and average classification accuracies are noted. Tables 3–5 show the top-1 classification accuracies of the word2vec and data2vec methods over the UCM21, RS19, and AID30 datasets, respectively. It can be observed from the results that our approach with data2vec embedding performs better in comparison with word2vec embedding on three benchmark datasets. To the best of our knowledge, we are the first to implement GZSL for scene classification tasks in remote sensing images.

4.4.1 *Analysis on UCM21 dataset*

We evaluated our proposed GZSL-RSD2V method by considering the word2vec and data2vec embedding approaches over the four standard splits[22] of the UCM21 dataset with seen/unseen classes of 16/5, 13/8, 10/11, and 7/14. It is observed from Table 3 that data2vec shows an improvement of 4.5%, 7.4%, 0.3%, and 2.8% in seen class accuracy and 5.3%, 2.7%, 4.2%,

Table 3 Seen, unseen, and harmonic mean scene classification accuracies with standard seen/unseen splits on the UCM21 dataset.

	16/5	13/8	10/11	7/14
Method	*S U H*	*S U H*	*S U H*	*S U H*
word2vec	94.1 52.6 67.0	91.7 37.5 52.6	98.2 26.5 41.3	96.0 21.8 35.4
data2vec	98.6 57.9 72.7	99.1 40.2 57.0	98.5 30.7 46.8	98.8 24.9 39.8

Table 4 Seen, unseen, and harmonic mean scene classification accuracies with different seen/unseen splits on the RS19 dataset.

	15/4	12/7	9/10	6/13
Method	*S U H*	*S U H*	*S U H*	*S U H*
word2vec	95.4 58.2 71.9	95.2 35.6 51.4	97.8 26.8 41.9	95.7 23.3 37.1
data2vec	97.5 70.7 81.8	97.6 46.2 62.6	98.5 33.2 49.5	97.5 27.4 42.7

Table 5 Seen, unseen, and harmonic mean scene classification accuracies with standard seen/unseen splits on the AID30 dataset.

	25/5	20/10	15/15	10/20
Method	*S U H*	*S U H*	*S U H*	*S U H*
word2vec	98.4 34.3 50.5	98.9 29.2 44.5	98.4 12.9 22.7	98.3 10.4 18.8
data2vec	97.2 57.3 71.9	97.4 32.4 48.5	97.9 22.5 36.6	97.6 17.1 29.0

and 3.1% in unseen class accuracy on the standard splits, respectively. Also, the efficacy of our proposed method with data2vec embedding is demonstrated in terms of the harmonic mean with an improvement of 5.7%, 4.4%, 5.5%, and 4.4% under the same seen/unseen splits (e.g., 16/5, 13/8, 10/11, and 7/14, respectively). Our proposed method with the data2vec embedding approach exhibits better classification in comparison with word2vec. This may be due to data2vec having self-supervised word embedding, making it capable of learning semantic features from unseen classes.

4.4.2 *Analysis on RS19 dataset*

We considered the word2vec and data2vec embedding approaches to evaluate our proposed method using the four standard splits[22] of the RS19 dataset with seen/unseen classes of 15/4, 12/7, 9/10, and 6/13. It is observed from Table 4 that data2vec achieved an improvement of 2.1%, 2.4%, 0.7%, and 1.8% in seen class accuracy and 12.5%, 10.6%, 6.4% and 4.1% in unseen class accuracy on the same standard splits. Also, the efficacy of our proposed method with data2vec embedding is demonstrated in terms of the harmonic mean with an improvement of 9.9%, 11.2%, 7.6%, and 5.6% under the same seen/unseen splits. Our proposed method with the data2vec embedding approach exhibits better classification in comparison with word2vec. This may be due to data2vec having self-supervised word embedding, making it able to learn semantic features of unseen classes.

4.4.3 *Analysis on AID30 dataset*

Our proposed GZSL-RSD2V is evaluated by considering the word2vec and data2vec embedding approaches over the four standard splits[22] of the AID30 dataset with seen/unseen classes of 25/5, 20/10, 15/15, and 10/20. Table 5 shows a rise in the classification accuracy of 23.0%, 3.2%, 9.6%, and 6.7% in unseen classes under these seen/unseen splits with the data2vec approach, though a marginal drop in the performance of seen class accuracy around 1% is observed with data2vec in comparison with word2vec. Our proposed method also exhibits the effectiveness of data2vec embedding in terms of the harmonic mean with an improvement of 21.4%, 4.0%, 13.9%, and 10.2% under the same seen/unseen splits. It is observed from the experiments that the data2vec provides better semantic features on unseen classes compared with seen classes.

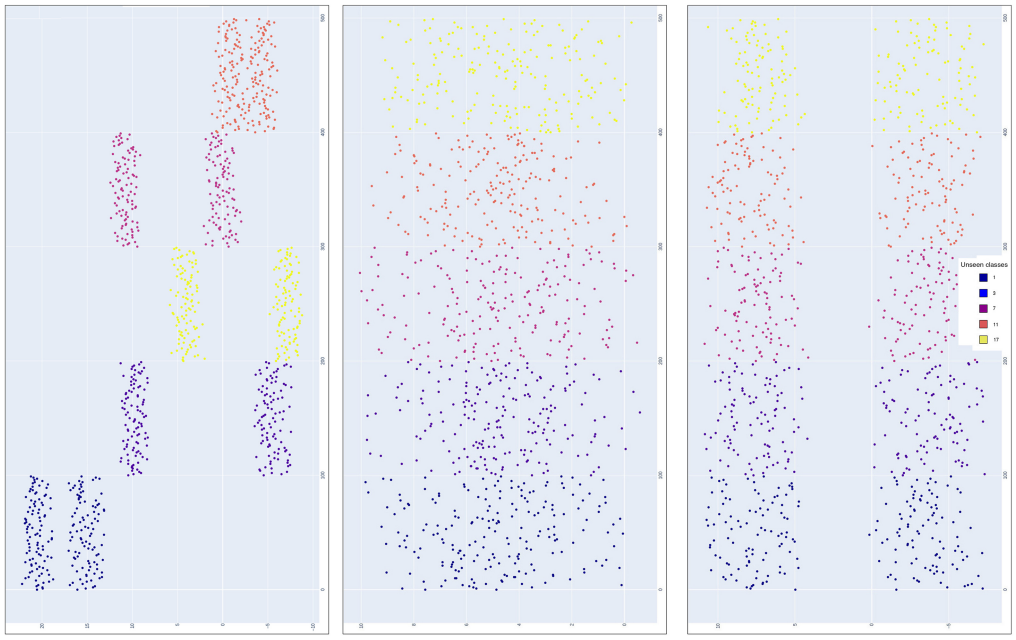
Upon evaluation of our proposed GZSL-RSD2V method over three challenging scene classification benchmark datasets, we noticed that data2vec embedding shows consistent improvement in classifying the scenes of unseen classes. However, data2vec embedding does not show improvement in classifying the scenes of seen classes from the AID30 dataset in comparison with the UCM21 and RS19 datasets, though the classification accuracy drop is relatively very small.

4.5 Qualitative Analysis

This section provides the qualitative results and their analysis of our proposed method. We used the uniform manifold approximation and projection (UMAP)[50] to visualize real unseen class visual features and the synthesized visual features through our proposed method with word2vec and data2vec embeddings. Figures 3–5 denote the UMAP visualization of the UCM21, RS19, and AID30 datasets, respectively. It is observed from Figs. 3–5 that the synthesized visual features through our proposed method with data2vec exhibit better separability in comparison with synthesized visual features with the word2vec method. It is evident from the visualization that our proposed method with data2vec embedding is able to capture meaningful semantics relevant to unseen classes.

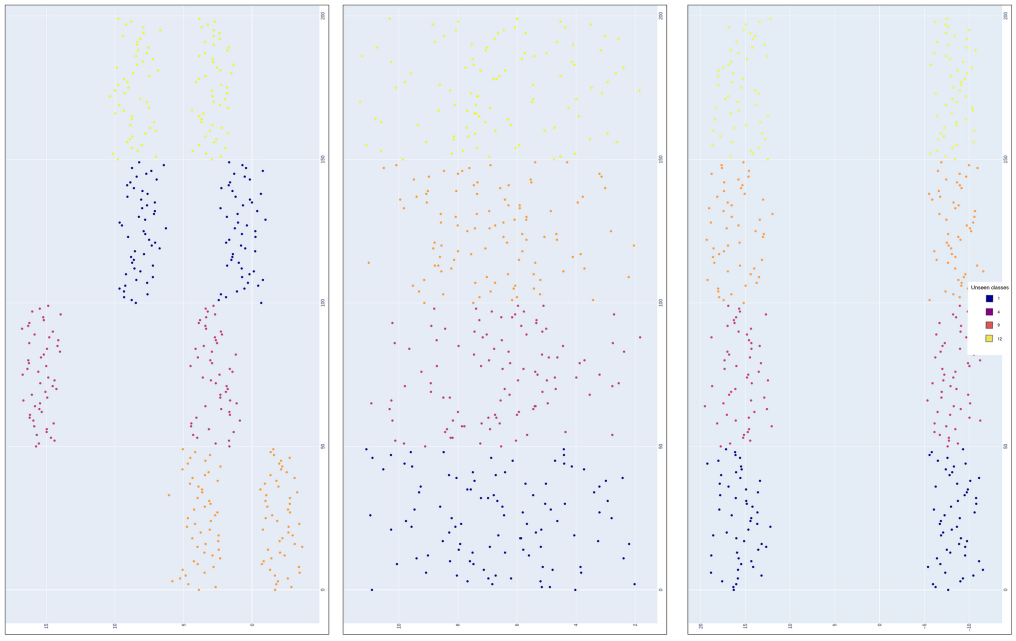
5 Conclusion

This paper proposed a self-supervised embedding to represent semantics useful for GZSL for scene classification in remote sensing images. A learning mechanism was devised to map the semantics to the corresponding visual domain during the visual feature generation. A feature refinement module was employed to improve the visual features of both seen and unseen classes of remote-sensing images. To the best of our knowledge, we were the first to explore a GZSL approach in the remote



(a) (b) (c)

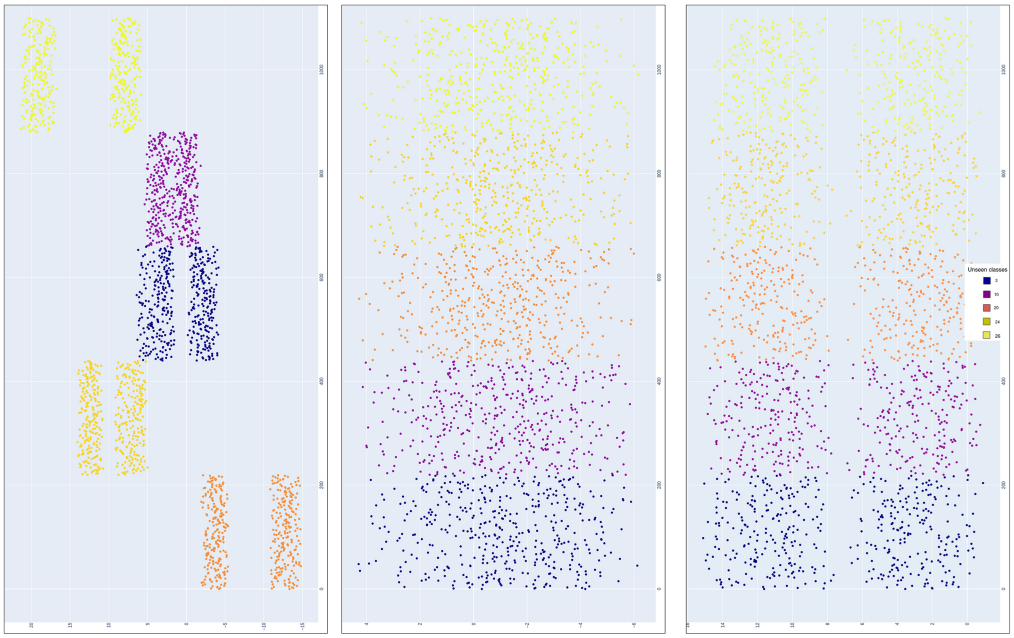
Fig. 3 UMAP visualizations for the features of five unseen class samples from the UCM21 dataset. (a) The real unseen class features. (b) The separability of the synthesized features of our proposed method with word2vec embedding. (c) The separability of the synthesized features of our proposed method with data2vec embedding.



(a) (b) (c)

Fig. 4 UMAP visualizations for the features of four unseen class samples from the RS19 dataset. (a) The real unseen class features, (b) the synthesized features of our proposed method with word2vec embedding, and (c) the synthesized features of our proposed method with data2vec embedding.

sensing domain. Our proposed approach was evaluated using both data2vec and word2vec embeddings. It is observed from the experiments that our proposed method with data2vec embedding was able to capture meaningful semantics relevant to unseen classes. In the future, we will explore weighted embeddings for representing the semantics of remote-sensing images.



(a) (b) (c)

Fig. 5 UMAP visualizations for the features of five unseen class samples from the AID30 dataset. (a) The real unseen class features, (b) the synthesized features of our proposed method with word2vec embedding, and (c) the synthesized features of our proposed method with data2vec embedding.

Code, Data, and Materials Availability

The code cannot be made publicly available due to its proprietary nature. The data presented in this article are publicly available at Refs. 46–48.

Acknowledgment

The authors have no conflicts of interest to disclose.

References

1. G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: benchmark and state of the art," *Proc. IEEE* **105**(10), 1865–1883 (2017).
2. G. Cheng et al., "Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **13**, 3735–3756 (2020).
3. R. Datla et al., "Scene classification in remote sensing images using dynamic kernels," in *Int. Joint Conf. on Neural Netw. (IJCNN)*, IEEE, pp. 1–8 (2021).
4. R. Datla, V. Chalavadi, and C. K. Mohan, "A framework to derive geospatial attributes for aircraft type recognition in large-scale remote sensing images," *Proc. SPIE* **12084**, 120840N (2022).
5. J. Sharma et al., "Aircraft type recognition in remote sensing images using mean interval kernel," in *IMPROVE*, pp. 166–173 (2022).
6. R. Datla, V. Chalavadi, and C. K. Mohan, "A multimodal semantic segmentation for airport runway delineation in panchromatic remote sensing images," *Proc. SPIE* **12084**, 1208407 (2022).
7. H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks," in *AAAI Conf. on Artif. Intell.* (2008).
8. Y. Xian et al., "Zero-shot learning–a comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 2251–2265 (2017).
9. C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 453–465 (2014).
10. Z. Akata et al., "Label-embedding for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.* **38**, 1425–1438 (2015).
11. Y. Xian et al., "Feature generating networks for zero-shot learning," in *IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, pp. 5542–5551 (2017).

12. A. Mishra et al., "A generative model for zero shot learning using conditional variational autoencoders," in *IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. Workshops (CVPRW)*, pp. 2269–22698 (2017).
13. R. Datla and C. K. Mohan, "A novel framework for seamless mosaic of Cartosat-1 DEM scenes," *Comput. Geosci.* **146**, 104619 (2021).
14. S. E. Reed et al., "Learning deep representations of fine-grained visual descriptions," in *IEEE Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 49–58 (2016).
15. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM* **60**, 84–90 (2012).
16. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR abs/1409.1556 (2014).
17. C. Szegedy et al., "Going deeper with convolutions," in *IEEE Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 1–9 (2014).
18. K. He et al., "Deep residual learning for image recognition," in *IEEE Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 770–778 (2015).
19. J. Deng et al., "ImageNet: a large-scale hierarchical image database," in *IEEE Conf. on Comput. Vis. and Pattern Recognit.*, pp. 248–255 (2009).
20. A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR*, pp. 1521–1528 (2011).
21. S. Chen et al., "Free: feature refinement for generalized zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 122–131 (2021).
22. A. Li et al., "Zero-shot scene classification for high spatial resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.* **55**, 4157–4167 (2017).
23. J. C. Quan et al., "Structural alignment based zero-shot classification for remote sensing scenes," in *IEEE Int. Conf. on Electron. and Commun. Eng. (ICECE)*, pp. 17–21 (2018).
24. C. Wang, G. Peng, and B. D. Baets, "A distance-constrained semantic autoencoder for zero-shot remote sensing scene classification," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **14**, 12545–12556 (2021).
25. G. Sumbul, R. G. Cinbis, and S. Aksoy, "Fine-grained object recognition and zero-shot learning in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.* **56**, 770–779 (2017).
26. E. Schönfeld et al., "Generalized zero- and few-shot learning via aligned variational autoencoders," in *IEEE/CVF Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 8239–8247 (2018).
27. S. Liu et al., "Generalized zero-shot learning with deep calibration network," in *Neural Inf. Process. Syst.* (2018).
28. G. Arora et al., "Generalized zero-shot learning via synthesized examples," in *IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 4281–4289 (2017).
29. Y. Liu et al., "Graph and autoencoder based feature extraction for zero-shot learning," in *Int. Joint Conf. on Artif. Intell.* (2019).
30. K. Li, M. R. Min, and Y. R. Fu, "Rethinking zero-shot learning: a conditional visual classification perspective," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 3582–3591 (2019).
31. J. Li et al., "Leveraging the invariant side of generative zero-shot learning," in *IEEE/CVF Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 7394–7403 (2019).
32. Y. Shen, J. Qin, and L. Huang, "Invertible zero-shot recognition flows," ArXiv abs/2007.04873 (2020).
33. I. J. Goodfellow et al., "Generative adversarial nets," in *NIPS* (2014).
34. Z. Li et al., "Generative adversarial networks for zero-shot remote sensing scene classification," *Appl. Sci.* **12**(8), 3760 (2022).
35. C. Wah et al., "The caltech-UCSD birds-200-2011 dataset," (2011).
36. T. Mikolov et al., "Efficient estimation of word representations in vector space," in *Int. Conf. Learn. Represent.* (2013).
37. J. Pennington, R. Socher, and C. D. Manning, "Glove: global vectors for word representation," in *Conf. Empirical Methods in Nat. Lang. Process.* (2014).
38. P. Bojanowski et al., "Enriching word vectors with subword information," arXiv:1607.04606 (2016).
39. A. Baevski et al., "Data2vec: a general framework for self-supervised learning in speech, vision and language," in *Int. Conf. on Mach. Learn.* (2022).
40. Y. Xian et al., "F-VAEGAN-D2: a feature generating framework for any-shot learning," in *IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 10267–10276 (2019).
41. A. Vaswani et al., "Attention is all you need," ArXiv abs/1706.03762 (2017).
42. H. Bao, L. Dong, and F. Wei, "BEiT: BERT pre-training of image transformers," ArXiv abs/2106.08254 (2021).
43. J.-B. Grill et al., "Bootstrap your own latent: a new approach to self-supervised learning," ArXiv abs/2006.07733 (2020).
44. M. Caron et al., "Emerging properties in self-supervised vision transformers," in *IEEE/CVF Int. Conf. on Comput. Vis. (ICCV)*, pp. 9630–9640 (2021).

45. D. P. Kingma and M. Welling, "Auto-encoding variational bayes," CoRR abs/1312.6114 (2013).
46. Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *ACM SIGSPATIAL Int. Workshop on Adv. in Geographic Inf. Syst.* (2010).
47. G.-S. Xia et al., "Structural high-resolution satellite image indexing," (2010).
48. G.-S. Xia et al., "Aid: a benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.* **55**, 3965–3981 (2016).
49. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," CoRR abs/1412.6980 (2014).
50. L. McInnes and J. Healy, "Umap: uniform manifold approximation and projection for dimension reduction," ArXiv abs/1802.03426 (2018).

Rambabu Damalla received his BTech degree from JNTU, Pulivendula, Andhra Pradesh, India, in 2012, and his MTech degree from NIT Warangal, India, in 2015. He is pursuing a PhD in CSE at IIT Hyderabad, India. His research interests include remote sensing imagery analysis, computer vision, and machine learning.

Rajeshreddy Datla received his BTech and MTech degrees in CSE from JNTU Hyderabad, India, in 2002 and 2012, respectively, and his PhD in CSE from IIT Hyderabad, India, in 2021. Currently, he is a scientist/engineer-SF at ADRIN, Department of Space, Secunderabad, India. His research interests include computer vision, pattern recognition, machine learning, and remote sensing imagery analysis. He is a life member of the Astronautical Society of India and a senior member of IEEE.

Chalavadi Vishnu received his BTech (CSE) degree from JNTU Hyderabad in 2016, and his MTech (CSE) and PhD degrees from IIT Hyderabad in 2018 and 2022, respectively. Currently, he is doing a postdoc at the University of Agder, Grimstad, Norway. His research interests include learning representations of video activities and drone data and autonomous vehicles.

Chalavadi Krishna Mohan received his MTech (SACA) degree from NIT Surathkal, India, in 2000 and his PhD in CSE from IIT Madras, India, in 2007. He is currently a professor in the Department of CSE, IIT Hyderabad, India. His research interests include video content analysis and machine learning. He is a senior member of IEEE.