ISPRS Journal of Photogrammetry and Remote Sensing 198 (2023) 140–152

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Deep Semantic-Visual Alignment for zero-shot remote sensing image scene classification

Wenjia Xu [a,c,d], Jiuniu Wang [b,c,d,*], Zhiwei Wei [c,d], Mugen Peng [a], Yirong Wu [c,d]

[a] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
[b] City University of Hong Kong, 999077, Hong Kong, China
[c] University of Chinese Academy of Sciences, Beijing 100864, China
[d] Aerospace Information Research Institute, CAS, Beijing 100094, China

ARTICLE INFO

Keywords:
Remote sensing scene classification
Zero-shot learning
Deep semantic-visual alignment model
Automatic attribute annotation

ABSTRACT

Deep neural networks have achieved promising progress in remote sensing (RS) image classification, for which the training process requires abundant samples for each class. However, it is time-consuming and unrealistic to annotate labels for each RS category, given the fact that the RS target database is increasing dynamically. Zero-shot learning (ZSL) allows for identifying novel classes that are not seen during training, which provides a promising solution for the aforementioned problem. However, previous ZSL models mainly depend on manually-labeled attributes or word embeddings extracted from language models to transfer knowledge from seen classes to novel classes. Those class embeddings may not be visually detectable and the annotation process is time-consuming and labor-intensive. Besides, pioneer ZSL models use convolutional neural networks pre-trained on ImageNet, which focus on the main objects appearing in each image, neglecting the background context that also matters in RS scene classification. To address the above problems, we propose to collect visually detectable attributes automatically. We predict attributes for each class by depicting the semantic-visual similarity between attributes and images. In this way, the attribute annotation process is accomplished by machine instead of human as in other methods. Moreover, we propose a Deep Semantic-Visual Alignment (DSVA) that take advantage of the self-attention mechanism in the transformer to associate local image regions together, integrating the background context information for prediction. The DSVA model further utilizes the attribute attention maps to focus on the informative image regions that are essential for knowledge transfer in ZSL, and maps the visual images into attribute space to perform ZSL classification. With extensive experiments, we show that our model outperforms other state-of-the-art models by a large margin on a challenging large-scale RS scene classification benchmark. Moreover, we qualitatively verify that the attributes annotated by our network are both class discriminative and semantic related, which benefits the zero-shot knowledge transfer.

1. Introduction

With the rapid advances of sensors and remote sensing (RS) technology (Toth and Jóźków, 2016), RS image scene classification (Cheng et al., 2017; Gu et al., 2019) draws increasing attention as it is playing an essential role in urban construction (Chen et al., 2021), environmental monitoring (Li et al., 2020c), etc. While deep neural networks (DNNs) have led to impressive successes in image scene classification (Cheng et al., 2020; Wang et al., 2020), learning the hyperplane for discriminating various categories requires abundant training samples (Deng et al., 2009; Cheng et al., 2020). However, it is unrealistic to collect sufficient RS scene images for all circumstances (Xue et al., 2017; Alajaji et al., 2020; Li et al., 2020a). For instance, the earth observation system can collect a huge size of data (up to 100TB) every day, while it is notably time consuming to annotate all category labels at once. Moreover, as the RS target database is increasing dynamically, there is an urgent demand for recognizing new RS scenes that never appear in the training stage.

Zero-shot learning (ZSL), which aims to identify unseen classes without training samples (Xian et al., 2019), provides a promising solution for the aforementioned problem, and are widely explored in image classification (Xu et al., 2022b), object detection (Zhu et al., 2019a), etc. By leveraging semantic knowledge and visual information for each seen category, ZSL models can generalize the learned knowledge to unseen classes simultaneously. As shown in Fig. 1, by learning

* Corresponding author.
E-mail addresses: xuwenjia@bupt.edu.cn (W. Xu), jiuniwang2-c@my.cityu.edu.hk (J. Wang).

https://doi.org/10.1016/j.isprsjprs.2023.02.012
Received 3 September 2022; Received in revised form 2 February 2023; Accepted 24 February 2023

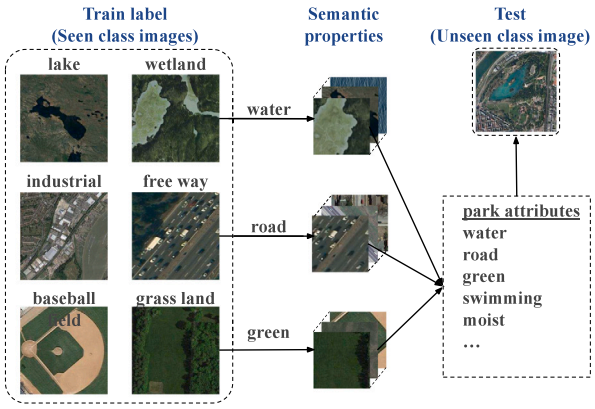

Fig. 1. Illustration of zero-shot learning. By learning the semantic properties, *e.g.*, "water", "road", from seen classes such as lake and wetland, the model would be able to recognize the unseen class "park" composing of those semantic properties.

the visual property of semantic knowledge demonstrating each seen class (*e.g.*, what is "water", "road", and "green"), the model would be able to recognize unseen category *park* composing of that semantic knowledge. The two key factors of ZSL are the class embeddings and the ZSL model (Xu et al., 2020b). The class embeddings aggregated for every class live in a semantic vector space that can be used to associate different classes even when visual examples of these classes are not available. Meanwhile, the ZSL model learns how to transfer knowledge from seen classes to unseen classes.

The prior attempts in improving class embeddings are two folds, manually-labeled attributes and semantic embeddings from pre-trained language models. Attributes are the characteristic properties of objects (Patterson et al., 2014; Wah et al., 2011; Farhadi et al., 2009), which are both human interpretable and discriminative among various categories. Therefore, attribute embeddings describing the characteristics of each class have become the most widely used and powerful class embeddings for ZSL (Xian et al., 2019; Xu et al., 2022a; Akata et al., 2015). Obtaining manually-labeled attributes is often a two-step process (Patterson et al., 2014), which is time-consuming and labor-intensive. First, domain experts carefully design an attribute vocabulary containing various attributes, *e.g.*, color, shape, etc. Second, human annotators would check each category and annotate the presence or absence of an attribute in an image or a class, which is called the labeling process. Although there are some attempts to collect attributes for remote sensing scenarios, *e.g.*, for aircraft recognition (Xu et al., 2020a) and scene classification (Li et al., 2021b), those attributes are either incomplete in visual space (with less than 60 attributes per class), or require enormous human efforts. Some pioneers tackle this problem by replacing attributes with semantic class embeddings extracted from pre-trained language models, *e.g.*, word2vec (Mikolov et al., 2013), glove (Pennington et al., 2014), and BERT (Devlin et al., 2018), or from knowledge graphs constructed for RS scenes (Li et al., 2021b). However, each dimension of these embeddings does not contain concrete semantic properties, and cannot be visually detectable, thus resulting in inferior ZSL performance. Therefore, to promote the development of ZSL in the remote sensing era, there is an urgent need to collect rich attributes depicting the visual properties of RS categories, and build a manual-free attribute labeling system that can collect visually detectable attributes.

Since RS images are usually collected from satellites or aircraft, while ordinary optical images are collected from the ground, the image distribution between the two differs from each other in the following two aspects. The generalization of ordinary ZSL models to the RS scene classification task is limited due to the following characteristics of the RS scene. First, dislike ordinary optical images that usually focus on the local objects (Deng et al., 2009; Xu et al., 2020a), both the global context and the interaction information between local objects are important. As shown in Fig. 1, the road, buildings, and the background grasses all matter in recognizing the scene as the *park* instead of *industrial* or *grass lands*. Second, RS images have high inter-class variance and intra-class similarity (Zhao et al., 2021), thus requiring the ZSL models to concentrate on the intra-class discriminative features and the inter-class shared features (Cheng et al., 2017). However, most of the remote sensing ZSL models (Li et al., 2017b; Wang et al., 2021; Li et al., 2021c,b) utilize convolutional neural network (CNN) (He et al., 2016) pre-trained on large-scale ordinary image dataset, *e.g.*, ImageNet (Deng et al., 2009), ignoring the domain gap between the ordinary and RS images. Furthermore, CNN networks inherently have smaller receptive fields by design, which focus on local object features and cannot fully correlate global and local information for large-scale RS scenes.

To tackle the key problems analyzed above, we propose two networks to improve the quality of RS attribute embeddings and the ZSL model, which integrates semantic and visual information to boost the ZSL performance. First, we propose an automatic attribute annotation process, where we use the CLIP model (Radford et al., 2021) to predict remote sensing multi-modal attributes (denoted as RSMM-Attributes) for remote sensing scenes. We first construct an attribute vocabulary containing attributes with rich semantic and visual information for RS scene classes. To ensure the attribute embeddings for each class are visually detectable, the attribute labeling process is finished by calculating the semantic-visual similarity between the attribute and the example images. The CLIP model is pre-trained with a large-scale RS dataset containing images and the corresponding text descriptions, thus being able to associate the semantic text and visual RS images in one common space. In this way, the labor-intensive labeling process is replaced with an automatic model, significantly decreasing the time consumption.

To tackle the problems of remote sensing ZSL tasks, we develop a Deep Semantic-Visual Alignment (DSVA) model that uses the RSMM-Attributes to transfer knowledge between seen and unseen classes, and perform ZSL considering both the local details and global contexts of images. We adopt a vision transformer (Dosovitskiy et al., 2020), which use long-range interaction between local image region to extract image representations. The self-attention mechanism (Vaswani et al., 2017) in transformer helps to correlate the global context and local information together, which helps to integrate global information for ZSL. This is critical for remote sensing scene recognition as the global interaction of different image areas contributes to scene prediction. To associate semantic attributes with visual features, we then learn attribute prototypes that encode the visual properties for each attribute. Meanwhile, we generate the attribute attention maps by calculating the similarity between prototypes and the visual image features and accurately map the image into attribute space by calculating the image-attribute similarity. We further propose an attention concentration module to focus on the informative attribute regions. Since attributes represent intra-class discriminative features and inter-class shared features, this module will help the network to take advantage of discriminative image regions and benefit the knowledge transfer in ZSL.

In summary, the contributions of this work are three folds.

- We propose to automatically predict visual attributes for remote sensing scenes, alleviating the manual effort needed for labeling attributes. Our RSMM-Attributes are visually detectable and can facilitate the knowledge transfer between seen and unseen classes.
- We elaborate on the difference between ordinary images and remote sensing images, and develop a Deep Semantic-Visual Alignment (DSVA) model that uses RSMM-Attributes to tackle the RS zero-shot learning problem. Our model adopts a vision transformer with self-attention mechanism to enlarge the receptive field and learn both the local details and global contexts for RS scenes. Furthermore, an attention concentration module is proposed to focus on the informative attribute regions.

- Extensive quantitative experiments demonstrate that our model achieves state-of-the-art performance on a large-scale RS scene benchmark, outperforming the other pioneers by up to 30%. Qualitative analysis indicates that the learned RSMM-Attributes are both class discriminative and can reflect the visual and semantic relatedness.

The remainder of the paper is organized as follows. In Section 2, we present related works, including ZSL models and class embeddings. In Section 3, we introduce the process for collecting visually detectable and automatic-labeled attributes for RS scenes. In Section 4 we introduce our Deep Semantic-Visual Alignment (DSVA) model that performs ZSL considering the global context of RS scenes. The experimental setting and quantitative results are presented in Section 5, demonstrating the superior performance of our work. In Section 6, we present the visualization of the learnt RSMM-Attributes. Finally, we conclude the paper in Section 7.

2. Related works

In this section, we review the related literature concerning the two key factors in ZSL, *i.e.*, the zero-shot learning models and the class embeddings.

2.1. Zero-shot learning in remote sensing scene classification

In ZSL, the goal is to train a model with abundant training samples from seen classes, then recognize unseen classes that are not observed during training. This is accomplished by using class embeddings, also known as side information (Xian et al., 2019), that describe the semantic properties among seen and unseen classes. One of the earliest attempts in ZSL can be traced to Lampert et al. (2009) who perform direct attribute prediction for unseen classes. After that, approaches in this direction can be divided into two groups. The first group originates from attribute latent embeddings (ALE) (Akata et al., 2015), where the model maps the image representations to the space of class embeddings, and learns a compatibility function between them. Some following notable works aim at learning better embedding spaces (Zhang and Saligrama, 2016), improving the compatibility functions (Liu et al., 2021), or enhancing the image encoders (Xu et al., 2022a). An alternative line of work argues that the zero-shot problem can be complemented by generating fake samples for unseen classes (Xian et al., 2018). This is achieved by synthesizing image representations for unseen classes with generative models, *e.g.*, generative adversarial networks (GAN) (Goodfellow et al., 2014) and variational auto encoder (VAE) (Kingma and Welling, 2013), conditioned on the class embeddings (Zhu et al., 2019b; Schonfeld et al., 2019).

The first adaptation in the RS field can be traced to Li et al. (2017b) who utilize an unsupervised domain adaptation model to predict the unseen class labels, and develop a label propagation algorithm to leverage the visual similarity among images from the same scene class. Sumbul et al. (2017) further utilize a bilinear function that models the compatibility between the visual images and the class embeddings. Instead of using image features extracted from a fixed CNN network, LPDCMENs (Li et al., 2021c) train the network end-to-end and preserve the locality in RS scene images by emphasizing the pairwise similarity in one class. To maintain the cross-modal alignment between visual and semantic space, Li et al. (2021b) propose a novel generative model DAN to synthesize image features and match semantic features in a latent space, and Wang et al. (2021) propose an autoencoder model (DSAE) constrained by Euclidean distance between samples.

Despite the rapid development of ZSL models in the field of computer vision, the unique characteristics of remote sensing images limit the generalization of the above models in remote sensing ZSL tasks. First, since RS images usually have high inter-class variance and high intra-class similarity, the ZSL model designed for ordinary optical images cannot discriminate between similar categories. Second, while above approaches improve the ZSL performance gradually, they all utilize CNN as the backbone to extract image features, and ignore the intrinsic difference between RS images and ordinary optical images, *i.e.*, the global information. Dislike ordinary images that usually contain forehead regions that should draw much attention and background information that could be discarded, there is not an obvious distinction between foreground and background in RS images and every pixel matters when performing scene classification task. To this end, we adopt a vision transformer with self-attention mechanism to enlarge the receptive field and learn both the local details and the global contexts of RS scenes. Besides, our DSVA network is encoded with an attention concentration module focusing on the attribute-related informative image regions, which is helpful for discriminating between different classes.

2.2. Class embeddings for zero-shot learning

Class embeddings are crucial in relating different categories with shared characteristics in the semantic space, and can transfer knowledge from seen classes to unseen classes in ZSL. There are three types of commonly used class embeddings. The human annotated attributes (Patterson et al., 2014; Farhadi et al., 2009), describing the visual and semantic properties of objects, are the most popular class embeddings in zero-shot learning. Though attributes show powerful capability in discriminating and associating different classes, they are limited as the annotation process is time-consuming and labor-intensive (Xu et al., 2022b). An alternative to the manual attributes is to extract class embeddings using pre-trained language models such as BERT (Devlin et al., 2018), and word2vec (Mikolov et al., 2013). Other works utilize knowledge graphs (Nayak and Bach, 2021) and online encyclopedia (Al-Halah and Stiefelhagen, 2017) to extract class embeddings to encode more knowledge for each category.

Despite their importance, class embeddings are relatively under-explored in zero-shot learning RS scene classification. Previous works mainly use class embeddings extracted from pre-trained language models, or attributes manually-labeled by experts. Li et al. (2017b) directly leverage pre-trained word2vec model to map the name of RS scene category to the semantic space, and Li et al. (2022) further investigate more language models such as fasttext (Bojanowski et al., 2017) and BERT (Devlin et al., 2018). Although word vector can work as class embeddings, they are typically not comprehensive in capturing the visual properties of categories, and the following works improve this situation with the help of domain experts. Sumbul et al. (2017) collect 25 attributes determining visually distinctive characters of each category, such as their parts, texture, and shape. Li et al. (2021c) ask multiple domain experts who are familiar with the RS field to observe 10 images from each category and summarize them with one sentence. The class embeddings are then extracted from a pre-trained BERT model. Li et al. (2021b) collect 59 attributes for 700 remote sensing scenes considering the color, shape, and objects. They further construct an RS scene knowledge graph with the help of 10 domain experts. However, the attributes and knowledge graphs cannot fully describe the rich visual properties of each category, and the time-consuming labeling process limits the generalization to new classes in practical. To this end, we propose to alleviate the manual burden by replacing the manual labeling process with an remote sensing multi-modal network. Our network depicts a few images and annotates the attribute value for each category automatically by measuring the similarity between the attributes and the images, which ensures that the annotated attributes can be visually detectable and describes the visual properties for all categories.

color	red white yellow green blue brown tan black orange purple
object presence	plane boat car wide-road narrow-road curved-road ring-road cross-road bungalow high-rise rotunda square-building pavement railroad mountain fencing marble flowers tree grass ocean smoke shrubbery wire brick dirt-soil
materials	cement wood cloth fire paper ice still-water metal stone rock cloud running-water rubber-plastic railing glass asphalt water soil snow sand leaves
texture	symmetrical messy moist dry dirty rusty horizontal vertical soft sharp dense sparse flat monotone vivid warm cold vegetation
shape	round rectangle square triangle oval parallel-lines rhombus
functions	industrial agriculture forestry residential commercial eating cleaning shopping working transporting swimming farming buildings climbing hiking

Fig. 2. Attribute vocabulary. We split 98 attributes into 6 groups, *i.e.*, "color", "object presence", "material", "texture", "shape", "functions".

3. Automatic attribute annotation with remote sensing multi-modal similarity

In this section, we are interested in the automatic attribute annotation process considering the rich visual information in remote sensing scenes while reducing human labor in attribute labeling. We would first construct an attribute vocabulary $\mathcal{A}$ covering the semantic and visual properties of all remote sensing scenes. Then we propose to adapt the CLIP (Radford et al., 2021) model which links the semantic-visual space together to annotate the Remote Sensing Multi-Modal Attribute (RSMM-Attribute) for each remote sensing scene category.

Given a remote sensing scene category y and one attribute $a \in \mathcal{A}$, the purpose is to annotate the attribute value for this category as $r_a(y) \in \mathbb{R}$, indicating the possibility that the attribute a appearing in class y. With the CLIP model, we can measure the strength of association $r_a(y)$ between the attribute a and category y. In the end, we will get a class embedding containing all attribute values for each category as $r_{\mathcal{A}}(y) \in \mathbb{R}^{N_a}$, where N_a is the size of attribute vocabulary. In this section, we would first introduce the procedure of attribute vocabulary construction and automatic attribute annotation, then introduce how the CLIP model is trained and fine-tuned.

3.1. Attribute vocabulary construction

To discover attributes with rich semantic and visual information, we consider the following six types of attributes as shown in Fig. 2. (1) The "color" group includes the color appearing in the scene (*e.g.*, green, brown). (2) The "object presence" group lists the objects that would show up (*e.g.*, tree, soil). (3) The "materials" group describes the material constructing the scene (*e.g.*, cement, metal). (4) The "texture" group describes the texture and pattern of each category (*e.g.*, symmetrical, flat). (5) The "shape" group indicates the main shape shown up in each scene (*e.g.*, round, rectangle). (6) The "functions" group indicates the social-economic function of each category (*e.g.*, industrial, agriculture). Then we go through all the properties and objects related to the remote sensing scene to fill each group, and finally get an attribute vocabulary $\mathcal{A}$ with N_a attributes. Note that building the attribute vocabulary is relatively easy, with only one annotator working for three hours. Besides, the vocabulary is not limited to the known scene categories, as the attributes can be shared between classes, therefore can be generalized to novel RS scene categories and describe their discriminate properties.

3.2. Automatic attribute annotation with CLIP model

As shown in Fig. 3 (right), the CLIP model we adopt is composed of a semantic encoder $E_t(\cdot)$ and a visual encoder $E_v(\cdot)$ that maps the attribute name and the probe images into a shared semantic-visual space. The semantic encoder is a masked self-attention Transformer (Radford et al., 2019) and the visual encoder is a Vision Transformer (Dosovitskiy et al., 2020) ViT-B/32 with 12 layers and the input patch size is 32×32. Then the real-valued confidence $r_a(y)$ is calculated by the similarity measurement f_{sim} as:

$$r_a(y) = \sum_{i=1}^{m} f_{sim}\big(E_t(a), E_v(x_i)\big), \tag{1}$$

where the attribute value $r_a(y)$ indicates the possibility that attribute a would appear in category y. Following CLIP, we turn a single attribute name (*e.g.*, "narrow-road") into a sentence containing the attribute (*e.g.*, "This photo contains narrow-road") as the input of $E_t(\cdot)$. The similarity measurement is the dot product:

$$f_{sim}(\alpha, \beta) = <\alpha, \beta> . \tag{2}$$

In this way, we can easily replace humans in associating each attribute and the corresponding category, and predict the class semantic attributes for each class as $r_{\mathcal{A}}(y) \in \mathbb{R}^{N_a}$. In spite of decreasing manual labor significantly, the real-valued attribute value works better in discriminating and comparing between different classes than binary attributes annotated by the human. We denote the attributes annotated by CLIP model as Remote Sensing Multi-Modal Attribute (RSMM-Attribute) in the following text.

3.3. Training and fine-tuning of the CLIP model

The pre-trained CLIP (Radford et al., 2021) model is fine-tuned with remote sensing dataset (Lu et al., 2017) and the corresponding text descriptions. Here we first introduce the training procedure of the CLIP (Radford et al., 2021) model as the background for our methodology.

The CLIP network should be able to associate visual and semantic information in a common embedding space, for which the training procedure involves pre-training a semantic-visual network with a large-scale dataset containing images and the corresponding text descriptions. The training procedure is the same as the fine-tune process in Fig. 3 (left), where a batch of B images $\{x_1, x_2, \ldots, x_B\}$ are passed to the visual encoder $E_v(\cdot)$ to extract image representations $\{E_v(x_1), E_v(x_2), \ldots, E_v(x_B)\}$ and the text descriptions $\{t_1, t_2, \ldots, t_B\}$ are processed by the semantic encoder $E_t(\cdot)$ and output text representations $\{E_t(t_1), E_t(t_2), \ldots, E_t(t_B)\}$. Then a contrastive learning paradigm is adopted where the cosine similarity between positive image-text pairs (*i.e.*, x_i and t_i, where $i \in \{1, 2, \ldots, B\}$) are optimized to be 1, while the similarity of negative image-text pairs (*i.e.*, x_i and t_j, where $i \neq j$) are optimized to be close to 0. In particular, the sum of two InfoNCE loss L_{NCE} (Oord et al., 2018) is minimized to learn a joint representation of image and texts as follows:

$$\begin{aligned}\mathcal{L}_{con} = -\sum_{i=1}^{B} \Big(&\log L_{\text{NCE}}\big(E_v(x_i), E_t(t_j)\big) \\ &+ \log L_{\text{NCE}}\big(E_t(t_i), E_v(x_j)\big) \Big),\end{aligned} \tag{3}$$

where $L_{\text{NCE}}(E_v(x_i), E_t(t_j))$ denotes the visual to text similarity:

$$L_{\text{NCE}}\big(E_v(x_i), E_t(t_j)\big) = \frac{exp\big(E_v(x_i) \cdot E_t(t_i)/\tau\big)}{\sum_{j=1}^{B} exp\big(E_v(x_i) \cdot E_t(t_j)\big)}, \tag{4}$$

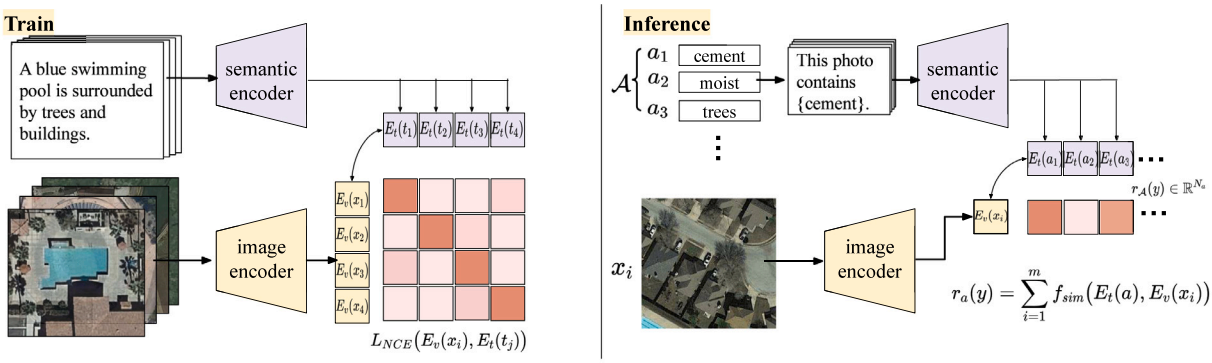

Fig. 3. The attribute annotation process. Left: We show the fine-tune process with a batch of 4 images and the corresponding text descriptions. Right: The inference process of CLIP network where we measure the similarity between the probe image x_i and attributes.

and the $L_{\text{NCE}}(E_t(t_i), E_v(x_j))$ is the text to visual similarity:

$$L_{\text{NCE}}(E_t(t_i), E_v(x_j)) = \frac{exp(E_v(x_i) \cdot E_t(t_i)/\tau)}{\sum_{j=1}^{B} exp(E_t(t_i) \cdot E_v(x_j))}, \tag{5}$$

with τ being the temperature hyper-parameter. After pre-training, the text and visual encoder in the CLIP network is able to project the texts and images in one common space.

Since the pre-training dataset of CLIP (Radford et al., 2021) mainly contains ordinary optical images collected from the internet, which lacks domain knowledge for remote sensing scenes. In this way, the pre-trained model would suffer from domain gaps. To this end, we fine-tune the CLIP model with remote sensing scene images and corresponding descriptions from RSICD dataset (Lu et al., 2017). The fine-tune procedure is illustrated in Fig. 3 (left), where a batch of images and corresponding text descriptions are encoded to a common space with the visual encoder $E_v(\cdot)$ and semantic encoder $E_t(\cdot)$. Then we optimize the sum of two InfoNCE loss L_{NCE} (Oord et al., 2018) to train the encoders as in Eq. (3). Afterwards, the CLIP model can be used to map the target image and all attributes into a common visual-semantic space, where the attribute value reflects the strength of association between the attribute and image as in Eq. (1).

4. Deep Semantic-Visual Alignment model for zero-shot learning

We start by formalizing the zero-shot learning (ZSL) and generalized zero-shot learning (GZSL) tasks. Then we introduce the architecture of our Deep Semantic-Visual Alignment (DSVA) model. Afterwards, we introduce the loss functions used to supervise the model training, and describe the inference process for (generalized) zero-shot.

4.1. Problem definition for (generalized) zero-shot learning

We are interested in the zero-shot learning problem where the training and test classes are disjoint. The training set is $S = \{x, y, r_{\mathcal{A}}(y) | x \in \mathcal{X}, y \in \mathcal{Y}^s\}$, which consists of remote sensing image x in the RGB image space $\mathcal{X}$, label y from seen classes $\mathcal{Y}^s$, and attribute embedding vector $r_{\mathcal{A}}(y) \in \mathbb{R}^{N_a}$ derived from Section 3. The unseen class label is denoted with $\mathcal{Y}^u$, and the attribute embedding vectors for unseen classes $\{r_{\mathcal{A}}(y) | y \in \mathcal{Y}^u\}$ is also known. The ZSL task aims to predict the label of images from unseen classes, *i.e.*, $\mathcal{X} \rightarrow \mathcal{Y}^u$. While in practice, it is hard to tell if the test image comes from seen or unseen classes. To this end, generalized zero-shot learning (GZSL) aims to classify images from both the seen and unseen classes, *i.e.*, $\mathcal{X} \rightarrow \mathcal{Y}^u \cup \mathcal{Y}^s$.

4.2. Deep Semantic-Visual Alignment (DSVA) model

As shown in Fig. 4, the DSVA model utilizes a transformer equipped with self-attention layers to extract the image representations, then maps the representations to the attribute space with a Visual-Attribute Mapping (VAM) module and finally predicts the class label for each image according to the attribute similarity. Besides, we propose an Attention Concentration (AC) module to concentrate on the informative image region with the help of attribute-related attention. In the following section, we will formulate the model architecture mathematically.

4.2.1. Vision transformer

Vision transformer use long-range interaction to extract image representations, which is critical for remote sensing scene recognition as the global interaction of different spatial areas in the image contributes to the scene prediction. Different from convolutional neural network that has image-specific inductive bias such as small receptive field, the self-attention mechanism in transformer results in a much larger receptive field and introduces long-range interactions between different image regions. As shown in Fig. 5, the input image $x \in \mathbb{R}^{H \times W \times C}$, with H, W and C being the height, width and channel, is reshaped to a sequence of image patches $\{\tilde{x}_n\}_{n=1}^{N}$, and $\tilde{x}_n \in \mathbb{R}^{\frac{H}{k} \times \frac{W}{k} \times C}$. k is the number of rows (or columns) for patches and $N = k \times k$.

To extract the image representation, we first learn a linear embedding layer $f_0(\cdot)$ to map the image patches to D dimensions patch embeddings $Z^0 \in \mathbb{R}^{N \times D}$:

$$Z^0 = [f_0(\tilde{x}_1), f_0(\tilde{x}_2), \ldots, f_0(\tilde{x}_N)]. \tag{6}$$

Then the patch embeddings will be forward to the transformer encoder consisting of L layers of multi-head self-attention (MHSA) subnet and multi-layer perceptron (MLP) subnet, following Dosovitskiy et al. (2020):

$$Z'^l = \text{MHSA}(\text{LN}(Z^{l-1})) + Z^{l-1} \tag{7}$$

$$Z^l = \text{MLP}(\text{LN}(Z'^l)) + Z'^l \tag{8}$$

where $l = 1, \ldots, L$ denotes the layers index, and each of the latent embeddings $Z^0, Z^1, \ldots, Z^L$ is with shape $\mathbb{R}^{N \times D}$. Where the first input vector Z^0 is from Eq. (9). Residual connection (He et al., 2016) is employed for each of the two subnets followed by layer normalization (LN) operation (Ba et al., 2016).

In each multi-head self-attention module, with the normalized input vector $\text{LN}(Z^{l-1}) \in \mathbb{R}^{N \times D}$ (we use Z for simplicity), we calculate the weighted sum of each element in the input sequence, where the weight is based on the pairwise similarity between two elements of the sequence. The input sequence $Z \in \mathbb{R}^{N \times D}$ is mapped into three tensors, *i.e.*, the query (Q), key (K), and value (V), by a linear layer, as Dosovitskiy et al. (2020) do:

$$[Q, K, V] = ZU_{qkv}, \qquad U_{qkv} \in \mathbb{R}^{D \times D_{qkv}}. \tag{9}$$

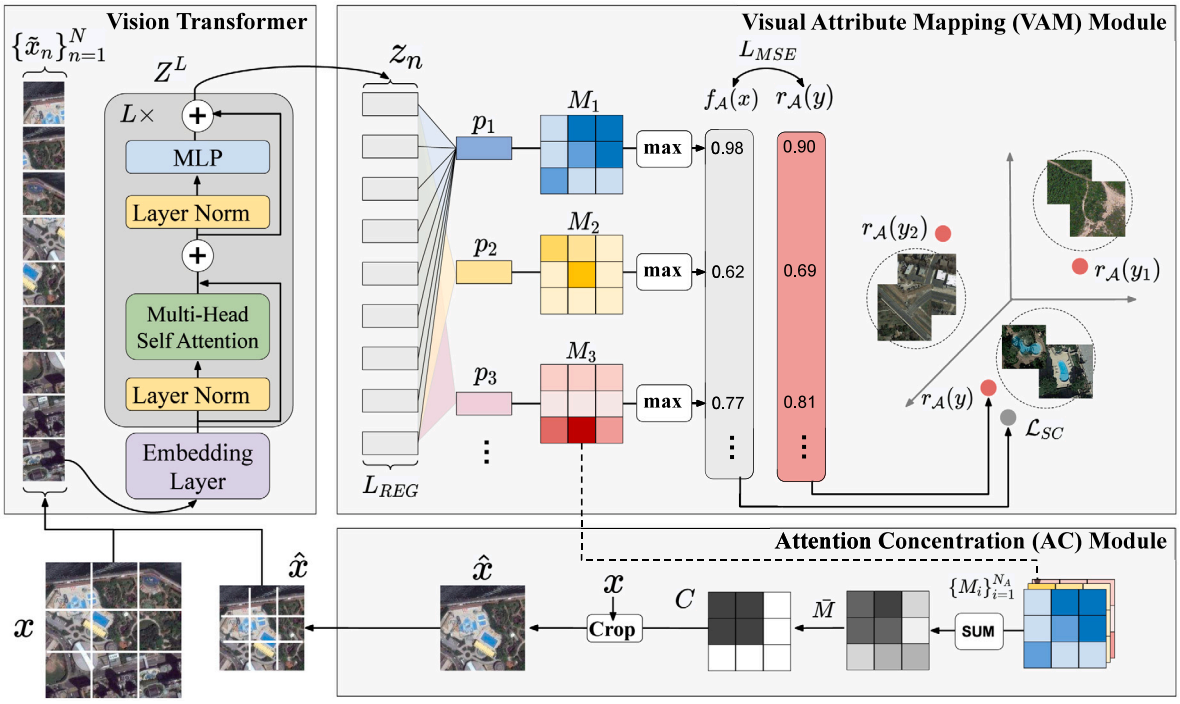

Fig. 4. The main architecture of the DSVA model. The proposed DSVA model consists of a vision transformer that extracts image features, a Visual Attribute Mapping Module that maps the image into attribute space for ZSL classification, and an Attention Concentration Module that focuses on the informative attribute regions.

Fig. 5. Three groups of similar scene categories from the RSSDIVCS (Li et al., 2021c) dataset. We displace three randomly sampled images from each category.

Afterwards, we follow Vaswani et al. (2017) to calculate the self-attention over the query and key via scaled dot product attention:

$$\text{Attention}(Q, K) = \text{softmax}(\frac{QK^T}{\sqrt{D_{qkv}}}) . \tag{10}$$

Then we multiply the attention with the V tensors:

$$\text{SA(Z)} = \text{Attention}(Q, K)V , \tag{11}$$

where SA(Z) is the output of self-attention module. In the multi-head self-attention (MHSA) subnet, the above self-attention operation is duplicated for s times, namely "heads", and the output is concatenated and projected by a linear layer as follows:

$$\text{MHSA}(Z) = [\text{SA}_1(Z), \dots, \text{SA}_s(Z)]U_{msa}, U_{msa} \in \mathbb{R}^{s \cdot D_{qkv} \times D} . \tag{12}$$

To make sure the number of parameters constant when changing s, the D_{qkv} in Eqs. (9) and (12) is typically set to D/s. The interaction between all local image patches combines visual cues from two local spots that are far away, thus helps the image representations to incorporate necessary information for image recognition. Given the output image feature $Z^L \in \mathbb{R}^{N \times D}$ with $k \times k$ local embeddings, each local embedding $z_n \in \mathbb{R}^D$ encodes the information from local image patch $\tilde{x}_n$ as well as its interaction with all other image patches.

4.2.2. Visual-attribute mapping module

In this section, we propose a Visual-attribute mapping (VAM) module to regress the attribute values from the image features, and predict the image category. We construct a visual-attribute mapping layer to learn N_a prototype vector $p_1, p_2, \dots p_{N_a}$. The ith prototype vector $p_i \in \mathbb{R}^D$ encodes the visual cues for the ith attribute a. Note that the prototype vectors are trainable vectors.

Since the prototype vector should encode the visual property for each attribute, we use the dot-product as the similarity between the attribute prototype vector p_i and each local image embedding z_n, which represents the possibility that image region $\tilde{x}_n$ contain the specific attribute:

$$\text{sim}(p_i, z_n) = p_i \cdot z_n . \tag{13}$$

As shown in Fig. 4, we reshape the similarity values between the ith attribute and $N = k \times k$ local image embeddings to form an attention map $M_i \in \mathbb{R}^{k \times k}$, indicating the possibility that attribute a_i appearing in the image x:

$$M_i = \{p_i \cdot z_n\}_{n=1}^{N}. \tag{14}$$

Then we predict the attribute value $f_a(x)$ by maximizing the similarity between each image patch and the attribute prototype:

$$f_a(x) = \max_n \{z_n \cdot p_i\}, \tag{15}$$

where $f_a(x) \in \mathbb{R}$ is the predicted attribute value. Overall, the predicted attribute embedding for image x is $f_{\mathcal{A}}(x) \in \mathbb{R}^{N_a}$. To assign the image to a specific class, we calculate the compatibility score between the predicted attribute embeddings and the ground truth attribute embedding among all training classes as follows:

$$S = f_{\mathcal{A}}(x) \cdot r_{\mathcal{A}}(y). \tag{16}$$

4.3. Attention concentrating (AC) module

As shown in Fig. 4, the attention map M_i indicates the image regions that share similar properties with the attribute prototypes, thus taking advantage of these attention maps would help the model locate attribute information and transfer the attribute knowledge between classes. To this end, we propose an attention concentrating module to crop and highlight the attribute informative image regions and train the DSVA network with the cropped images $\hat{x}$ once more.

Given N_a attribute attention maps $\{M_i\}_{i=1}^{N_a}$ generated from the VAM module (see Eq. (14)), the goal is to concentrate on the attribute related image regions and crop the original image x. We first sum the attention maps to get the mean attention:

$$\bar{M} = \frac{1}{N_a} \sum_{i=1}^{N_a} M_i. \tag{17}$$

Then the average attention value is calculated as follows:

$$\bar{m} = \frac{1}{N} \sum_{\alpha=1}^{k} \sum_{\beta=1}^{k} \bar{M}_{\alpha,\beta}, \tag{18}$$

with α and β being the spatial coordinate of the mean attention, and $N = k \times k$. Then we generate a concentration mask C with size $k \times k$ to highlight the informative image regions,

$$C_{\alpha,\beta} = \begin{cases} 1 & \text{if} \quad \bar{M}_{\alpha,\beta} \geq \bar{m} \\ 0 & \text{if} \quad \bar{M}_{\alpha,\beta} < \bar{m}. \end{cases} \tag{19}$$

Where the regions with attribute attention higher than the average value $\bar{m}$ is marked as one, and the regions with attribute attention value lower than $\bar{m}$ is marked as zero. Then we use the smallest bounding box that covers all the non-zero values in C to crop the original image x into a cropped image $\hat{x}$, and feed the cropped image into the vision transformer and VAM module again. We run the VAM module and the AC module iteratively in each batch, where the AC module will help the network to focus on informative attribute regions that point out the discriminative details between various classes.

4.4. Training loss

In this section, we introduce the loss functions $\mathcal{L}_{\text{DSVA}}$ to train our DSVA model.

4.4.1. Semantic compatibility loss

Semantic compatibility loss is used to supervise the training of the DSVA model. Given the input image x with label y and the ground truth attribute embedding $r_{\mathcal{A}}(y)$, we propose to use cross-entropy loss, encouraging the image to have a high compatibility score with its corresponding attribute label as follows:

$$L_{SC}(x) = -\log \frac{\exp\big(f_{\mathcal{A}}(x) \cdot r_{\mathcal{A}}(y)\big)}{\sum_{y_i \in \mathcal{Y}^s} \exp\big(f_{\mathcal{A}}(x) \cdot r_{\mathcal{A}}(y_i)\big)}, \tag{20}$$

where $f_{\mathcal{A}}(x) \cdot r_{\mathcal{A}}(y_i)$ is the compatibility score between the target image x and class y_i. Similarly, the semantic compatibility loss for the cropped image $\hat{x}$ generated from the AC module is

$$L_{SC}(\hat{x}) = -\log \frac{\exp\big(f_{\mathcal{A}}(\hat{x}) \cdot r_{\mathcal{A}}(y)\big)}{\sum_{y_i \in \mathcal{Y}^s} \exp\big(f_{\mathcal{A}}(\hat{x}) \cdot r_{\mathcal{A}}(y_i)\big)}. \tag{21}$$

Here we use cross entropy loss to enforce the compatibility score between target image x and its label y to be as high as possible, and the compatibility score between unmatched image and labels to be small.

4.4.2. Semantic regression loss

To facilitate the training of the visual-semantic mapping module, we further consider the attribute prediction as a regression problem and minimize the Mean Square Error (MSE) between the predicted attribute and the ground truth attribute embedding as follows:

$$L_{MSE}(x) = \left\| f_{\mathcal{A}}(x) - r_{\mathcal{A}}(y) \right\|_2, \tag{22}$$

where y is the label for image x. By optimizing the regression loss, we enforce the image representations learned by the transformer to contain semantic information and encode visual cues for each attribute, thus improving the knowledge generalization ability for ZSL. The semantic regression loss for the cropped image $\hat{x}$ is:

$$L_{MSE}(\hat{x}) = \left\| f_{\mathcal{A}}(\hat{x}) - r_{\mathcal{A}}(y) \right\|_2. \tag{23}$$

Overall, in each batch, our network optimizes the transformer and the visual-attribute mapping module with the following two objective functions iteratively,

$$\mathcal{L}_{\text{DSVA}} = \mathcal{L}_{SC}(x) + \lambda \mathcal{L}_{MSE}(x) + \mathcal{L}_{SC}(\hat{x}) + \lambda \mathcal{L}_{MSE}(\hat{x}), \tag{24}$$

with λ being the scaling factors.

4.5. (Generalized) zero-shot inference

Zero-shot inference classifies the images into unseen classes $\mathcal{Y}^u$. Given input image x, the DSVA model first extracts image representations Z_L, then maps the visual feature into the attribute space by the VAM module and gets the predicted attribute value $f_{\mathcal{A}}(x)$. In the end, the network searches for the predicted category $\hat{y}$ that has the highest compatibility score with the predicted attribute

$$\hat{y} = \arg\max_{y \in \mathcal{Y}^u} f_{\mathcal{A}}(x) \cdot r_{\mathcal{A}}(y). \tag{25}$$

For generalized zero-shot inference where the images are classified into both seen and unseen classes, the network searches the predicted category $\hat{y}$ that has the highest compatibility score as follows:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}^u \cup \mathcal{Y}^s} \big(f_{\mathcal{A}}(x) \cdot r_{\mathcal{A}}(y) - \gamma \mathbb{I}(y \in \mathcal{Y}^s)\big). \tag{26}$$

Since the network trained with only seen classes would have bias over training classes, we adopt Calibrated Stacking (Chao et al., 2016) to decrease the compatibility score of seen classes by a constant indicator $\mathbb{I}(y \in \mathcal{Y}^s)$, where the indicator is one when y belongs to the seen classes and is zero otherwise, and γ is a scaling factor.

5. Experiment

In this section, we first introduce the dataset and the experiment settings. Then we showcase the ablation study of each component. Afterwards, we compare our DSVA model with other state-of-the-art models on both ZSL and GZSL experiments. Finally, we demonstrate the effectiveness of the RSMM-Attributes with qualitative results.

Table 1
Ablation study over the proposed DSVA model where we report the Top-1 (T1) accuracy on unseen classes for ZSL, as well as the Top-1 accuracy on unseen (U), seen (S) and the harmonic mean (H) for GZSL. We train a single VAM module with only the semantic compatibility loss $\mathcal{L}_{SC}$ as the baseline. Note that the last row denotes the full DSVA model, which combines the VAM module and the AC module (trained with $\mathcal{L}_{SC}$, $\mathcal{L}_{MSE}$, and $\mathcal{L}_{AC}$).

Model	Zero-shot learning			Generalized zero-shot learning								
	60/10	50/20	40/30	60/10			50/20			40/30		
	T1	T1	T1	U	S	H	U	S	H	U	S	H
VAM + $\mathcal{L}_{SC}$	74.1	53.8	48.8	40.9	**79.9**	54.1	36.1	**75.3**	48.8	31.5	**71.2**	43.7
+ $\mathcal{L}_{MSE}$	80.3	63.3	56.0	62.3	72.9	67.2	50.0	62.2	55.4	42.4	58.9	49.3
+ AC (DSVA)	**84.0**	**64.2**	**60.2**	**68.4**	67.1	**67.7**	**53.5**	59.8	**56.5**	**43.7**	58.1	**49.9**

5.1. Experiment settings

5.1.1. Dataset

We conduct the zero-shot learning experiment on a widely used large-scale benchmark dataset RSSDIVCS (Li et al., 2021c), which integrates the image scenes from four datasets, *i.e.*, UCM (Yang and Newsam, 2011), AID (Xia et al., 2017), NWPU-RESISC45 (Cheng et al., 2017), RSI-CB256 (Li et al., 2017a). The dataset consists of 56,000 images from 70 categories ranging from natural scenes, *e.g.*, lake, mountain, and sea ice, to scenes containing human activity, *e.g.*, school, stadium, and thermal power station. Due to the fine-grained nature of remote sensing scenes where the land covers will appear in different scenes, many categories in the RSSDIVCS dataset are hard to discriminate from each other. We display representative images from three groups of fine-grained scene categories in Fig. 5. We adopt the same dataset split as pioneer works (Li et al., 2021b,a), where the 70 categories are randomly split into three different seen/unseen ratios, *i.e.*, 60/10, 50/20, 40/30. Here 60/10 denotes adopting 60 categories as seen classes and the rest 10 categories as unseen classes.

We adopt the Remote Sensing Image Captioning Dataset (RSICD) (Lu et al., 2017) to fine-tune the CLIP model. The dataset contains 10,921 remote sensing images collected from Google Earth, Baidu Map, MapABC, and Tianditu. The image size is fixed to 224 × 224 pixels with various resolutions and there are five text descriptions for each image. All the images in RSICD dataset are used for the fine-tune procedure.

5.1.2. Metrics

For ZSL task, we adopt the overall accuracy of all unseen classes as the evaluation metric. For the GZSL scenario where the network need to recognize images from both the seen and unseen classes, we follow Xian et al. (2019) to use the harmonic mean accuracy by considering both the accuracy of seen (S) and unseen (U) classes as follows:

$$H = \frac{(2 \times S \times U)}{S + U}. \tag{27}$$

5.1.3. Training details

To fine-tune the CLIP model, we adopt the officially pre-trained CLIP (Radford et al., 2021) model as the backbone. The Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ is adopted to optimize the network. The learning rate is linearly increased from 2×10^{-6} to 5×10^{-5} for the first 20 epochs to warm up, then decreased by 0.5% every epoch for 200 epochs. To avoid tuning much hyper-parameters, the temperature hyper-parameter τ in Eq. (4) and Eq. (5), which controls the range of the logits in the softmax function, is directly optimized as a log-parameterized multiplicative scalar following Radford et al. (2021). During inference, the number of probe images for each class m in Eq. (1) is set as 10.

To train the DSVA network, we adopt the pre-trained Vision Transformer (Dosovitskiy et al., 2020) ViT-B/32 with 12 layers (Radford et al., 2021) as the backbone. The Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ is adopted to optimize the network in an end-to-end manner. We first fix the transformer backbone, and warm up the Visual-attribute mapping (VAM) module with 4 epochs by setting the learning rate as 1×10^{-4}, and then the entire network is trained with a learning rate 10^{-6} for 26 epochs. We set $\lambda = 0.08$ for all experiments. The factor for calibrated stacking is set as 1×10^{-4}.

5.2. Ablation study

In this section, we present ablation studies of your proposed DSVA model and the RSMM-Attributes.

5.2.1. Ablation study on the Deep Semantic-Visual Alignment model (DSVA)

To measure the influence of each component of the proposed DSVA model, we design an ablation study. We train the vision transformer and the visual-attribute mapping (VAM) module with only the semantic compatibility loss $\mathcal{L}_{SC}$ as the baseline, and train two variants of DSVA by adding the semantic regression loss and the attention concentrate (AC) module gradually. The model is trained with the RSMM-Attributes automatically annotated by our work.

The ZSL results in Table 1 (left) demonstrate that the full DSVA model improves over the baseline model consistently under different ZSL splits by a large margin. For instance, the accuracy of the baseline model VAM + $\mathcal{L}_{SC}$ is improved from 74.1% to 84.0% (60/10), and from 53.8% to 64.2% (50/20), and from 48.8% to 60.2% (40/30). In specific, the semantic regression loss $\mathcal{L}_{MSE}$ enforces the image representation learned by the transformer backbone to contain semantic information, which boosts the performance by a large margin, *i.e.*, 6.2% (60/10), 9.5% (50/20), and 7.2% (40/30). This indicates that the semantic regression loss encodes the attribute information in image representations and thus improves the knowledge transfer ability of the ZSL model. The attention concentrate module, which helps the network to focus on the informative attribute regions, also provides significant accuracy gain, *i.e.*, 3.7% (60/10), 0.9% (50/20), and 4.2% (40/30). The results indicate that highlighting the attribute related region can help the model to discriminate different classes.

The results under the generalized zero-shot learning (GZSL) setting are shown in Table 1 (right), which witness a similar trend as the ZSL results. First, introducing the semantic regression loss and the attention concentration module helps the model to recognize unseen classes in the GZSL setting correctly. For instance, the accuracy of unseen classes (U) is improved from 40.9% to 68.4% (60/10), and from 36.1% to 53.5% (50/20), and from 31.5% to 43.7% (40/30). Notably, the semantic regression loss improves the performance by a large margin. The reason is that with the semantic regression loss, we enforce the image representations learned by the transformer to contain attribute information and encode visual cues for each attribute prototype, thus facilitating the knowledge transfer between seen and unseen classes. The results indicate that improving the model's ability to focus on important attributes and informative image regions will significantly decrease the bias on seen classes. Consequently, the seen class accuracy (S) decreased a bit with our full DSVA model, for which the better model would not have a strong bias on seen classes and result in more balanced accuracy on all classes. Overall, the harmonic mean (H) is significantly improved by 13.6% (60/10), 7.7% (50/20), and 6.2% (40/30).

5.2.2. Ablation study on the class embeddings

To evaluate the effectiveness of our automatically collected RSMM-Attribute, we compare it with the following three class embeddings widely used by state-of-the-art models. (1) SR-RSKG (Li et al., 2021b) is a semantic representation of RS scenes extracted from knowledge

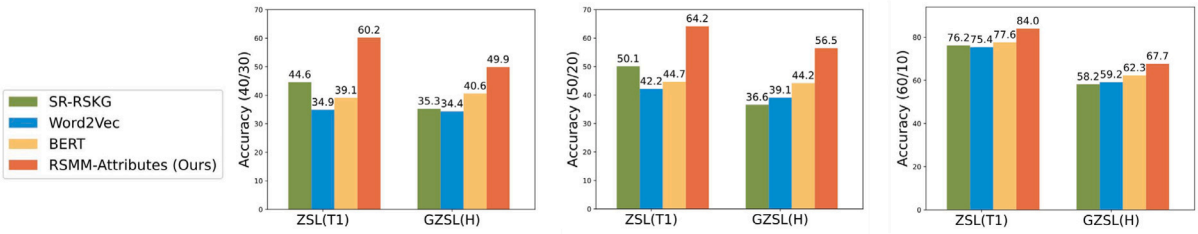

Fig. 6. Ablation study over four kinds of class embeddings under three different dataset splits. We report the Top-1 (T1) accuracy of ZSL and the harmonic mean (H) of GZSL when applying different class embeddings to our proposed DSVA model.

Table 2
Comparison between our attention concentration module with other attention modules, where we report the Top-1 (T1) accuracy on unseen classes for ZSL. We train the DSVA model with the semantic compatibility loss $\mathcal{L}_{SC}$ as the basemodel. Then we compare four different attention modules.

Model	Zero-shot learning		
	60/10 T1	50/20 T1	40/30 T1
basemodel	74.1	53.8	48.8
basemodel + BAM (Park et al., 2018)	75.5	54.6	49.0
basemodel + CBAM (Woo et al., 2018)	75.9	54.9	49.5
basemodel + GradCAM (Li et al., 2020b)	76.8	55.2	50.1
basemodel + Ours	**84.0**	**64.2**	**60.2**

graphs, where 10 domain experts participated in constructing the knowledge graph. (2) Word2vec embeddings are 300 dimension word embeddings for each RS category, which are extracted from a Word2Vec model pre-trained on Wikipedia corpus. (3) BERT embeddings are built by domain knowledge, where multiple RS experts are invited to depict each RS scene category and summarize them with one sentence. Then the BERT model is utilized to map the sentence to 1024 dimension embeddings.

We train our DSVA network under both ZSL and GZSL settings with the above three class embeddings separately. As shown in Fig. 6, the BERT embeddings are better than the Word2Vec embeddings in all circumstances. The reason is that the sentences describing each RS category used in BERT embeddings contain more semantic information than only the category name used in Word2Vec embeddings, and results in better generalization ability in the zero-shot learning scenario. The SR-RSKG embeddings (Li et al., 2021b) perform similarly to BERT and Word2Vec when trained with our DSVA models. Moreover, our RSMM-Attributes work much better than other alternative class embeddings under both ZSL and GZSL settings with different dataset splits. For instance, with a split 40/30, the model trained with RSMM-Attributes provides significant performance gain compared with the model trained with BERT by 21.1% (ZSL) and 9.3% (GZSL). With a split 50/20, RSMM-Attributes achieve 64.2% (ZSL) and 56.5% (GZSL), while BERT only gets 44.7% (ZSL) and 44.2% (GZSL). The results demonstrate two advantages of the RSMM-Attributes. Firstly, the visual properties encoded in the RSMM-Attributes help the ZSL network to model the visual space of unseen classes and recognize unseen images accurately. Secondly, different from other class embeddings that cannot tell the concrete semantics they encode, each dimension in our RSMM-Attributes denotes a specific semantic attribute, which is intuitive for the ZSL network to link different categories with those attributes and benefit the intra-class knowledge transfer. The results verifies the importance of using language-visual multi-modal network for annotating RSMM-Attributes.

5.2.3. Ablation study on the attention concentration module

We compare our attention concentration module with three other attention modules proposed recently. Bottleneck Attention Module (BAM) (Park et al., 2018) infers an attention map along two separate pathways, *i.e.*the channel and spatial attention, to concentrate on the important image regions and channels. Convolutional Block Attention Module (CBAM) (Woo et al., 2018) further utilize maxpooling to generate more salient features from the feature map, to concentrate on the global regions. Li et al. (2020b) propose to use Gradient Attention Map (GradCAM) as the attention module to pay attention to the salient image regions. We train the DSVA model with the semantic compatibility loss $\mathcal{L}_{SC}$ as the basemodel, then add the three attention modules separately to verify their effectiveness. As shown in Table 2, our attention concentration module with the semantic regression loss $\mathcal{L}_{MSE}$ outperforms other competitors by a large margin. For instance, compared to the GradCAM attention module, our attribute attention concentration module gains 7.2% (60/10), 9.0% (60/10), and 10.1% (40/30). The reason is that our model encodes the attribute information in the image representations and thus improves the knowledge transfer ability of the ZSL model. Besides, the attention concentrate module helps the network to focus on all the informative attribute regions guided by the attribute value for each class, which provides significant accuracy gain compared to other attention modules that only concentrate on one image region inferred by a learnable parameter.

5.3. Main results

In this section, we compare our Deep Semantic-Visual Alignment (DSVA) model with two groups of state-of-the-art models. Non-generative models learns projection function between the image features and class embeddings, *i.e.*, SPLE (Tao et al., 2017), DMaP (Li et al., 2017c), and LPDCMENs (Li et al., 2021c). Generative models learns auto-encoder or generative adversarial network to synthesize image features for unseen classes, *i.e.*, SAE (Kodirov et al., 2017), ZSC-SA (Quan et al., 2018), CIZSL (Elhoseiny and Elfeki, 2019), CADA-VAE (Schonfeld et al., 2019), TF-VAEGAN (Narayan et al., 2020), CE-GZSL (Han et al., 2021), and DAN (Li et al., 2021b). We build our DSVA model with two different backbones, *i.e.*, ResNet18 (He et al., 2016) and ViT (Dosovitskiy et al., 2020), and all other ZSL models use image representations extracted from ResNet18. We firstly compare those models trained with three different class embeddings, *i.e.*, SR-RSKG (Li et al., 2021b), Word2Vec (Mikolov et al., 2013), and BERT (Devlin et al., 2018), then we compare the performance of our DSVA model with the best performance of other SOTA models.

Table 3 displays the generalized zero-shot learning performance of different models trained with three class embeddings. As can be observed, our DSVA model yields a better harmonic mean than all other SOTA models. Specifically, when trained with BERT embeddings, our DSVA model using ResNet 18 as backbone achieves 52.0% (60/10), 39.8% (50/20), and 33.8% (40/30), which is much higher than the second best model DAN (Li et al., 2021b), which obtains 38.0% (60/10), 31.5% (50/20), and 28.2% (40/30). When trained with SR-RSKG embeddings learned from knowledge graphs, our DSVA (RN18) model still leads to significant performance gain compared to DAN, by 10.4%

Table 3
We compare the GZSL performance of our DSVA model with five SOTA models, where the models are trained with three different kinds of class embeddings (SideInfo), *i.e.*, SR-RSKG (Li et al., 2021b), Word2Vec (Mikolov et al., 2013), and BERT (Devlin et al., 2018). The results are harmonic mean (H). We build DSVA model with two different backbones, *i.e.*, ResNet18 (He et al., 2016) (denoted as RN18) and Vision Transformer (denoted as ViT) (Dosovitskiy et al., 2020).

SideInfo	Seen/Unseen ratio	SAE	DMaP	CIZSL	CADA-VAE	DAN	DSVA (RN18) (Ours)	DSVA (ViT) (Ours)
SR-RSKG	60/10	28.9	30.1	23.7	38.1	40.3	50.7	**58.2**
	50/20	23.7	23.4	13.9	32.9	34.1	36.0	**36.6**
	40/30	16.9	16.2	8.1	28.1	29.6	33.9	**35.3**
Word2Vec	60/10	28.0	28.9	25.2	32.9	34.1	48.2	**59.2**
	50/20	21.0	20.3	15.7	30.3	31.4	36.4	**39.1**
	40/30	17.2	16.8	9.1	26.1	25.6	30.1	**34.4**
BERT	60/10	28.6	26.6	25.0	36.3	38.0	52.0	**62.3**
	50/20	21.5	19.5	15.0	31.5	31.5	39.8	**44.2**
	40/30	16.7	16.3	8.6	27.1	28.2	33.8	**40.6**

Table 4
We compare the ZSL performance of our DSVA model with seven SOTA models, where the models are trained with three kinds of class embeddings (SideInfo), *i.e.*, SR-RSKG (Li et al., 2021b), Word2Vec (Mikolov et al., 2013), and BERT (Devlin et al., 2018). We report the Top-1 accuracy of all unseen classes. We build DSVA model with two different backbones, *i.e.*, ResNet18 (He et al., 2016) (denoted as RN18) and ViT (Dosovitskiy et al., 2020).

SideInfo	Seen/Unseen ratio	SAE	DMaP	SPLE	CIZSL	CADA-VAE	ZSC-SA	DAN	DSVA (RN18) (Ours)	DSVA (ViT) (Ours)
SR-RSKG	60/10	22.1	33.1	28.5	18.2	50.5	31.3	53.3	58.7	**76.2**
	50/20	12.8	20.3	17.2	8.9	39.6	19.1	45.2	46.6	**50.1**
	40/30	9.2	12.9	10.2	7.1	28.2	13.6	33.4	35.9	**44.6**
Word2Vec	60/10	23.5	26.0	20.1	20.6	41.4	26.7	44.3	55.7	**75.4**
	50/20	13.7	16.7	13.2	10.6	30.3	15.2	34.7	39.9	**42.2**
	40/30	9.6	10.4	9.8	6.0	21.2	12.1	24.3	31.0	**34.9**
BERT	60/10	22.0	16.4	19.0	20.4	48.1	29.3	50.2	59.4	**77.6**
	50/20	12.4	15.6	13.2	10.3	37.1	18.3	43.4	44.0	**44.7**
	40/30	8.8	10.0	8.3	6.2	26.3	13.1	31.5	35.3	**39.1**

Table 5
Comparing our DSVA model with SOTA models. We build DSVA model with two different backbones, *i.e.*, ResNet18 (He et al., 2016) (denoted as RN18) and ViT (Dosovitskiy et al., 2020). We report the Top-1 (T1) accuracy of unseen classes under the ZSL setting and the harmonic mean (H) of both seen and unseen classes under the GZSL setting. For fair comparison, all the models are trained with SR-RSKG (Li et al., 2021b) as the class embedding (SideInfo). To verify the effectiveness of our RSMM-Attributes, we further report the results of our DSVA model trained with RSMM-Attributes. Some results are"–" since we cannot access to their official code.

SideInfo	Model	ZSL accuracy			GZSL accuracy			Model size
		60/10 T1	50/20 T1	40/30 T1	60/10 H	50/20 H	40/30 H	(MB)
SR-RSKG	SAE (Kodirov et al., 2017)	22.1	12.8	9.2	28.9	23.7	16.9	44.59
	CIZSL (Elhoseiny and Elfeki, 2019)	18.2	8.9	7.1	23.7	13.9	8.1	50.59
	DMaP (Li et al., 2017c)	33.1	20.3	12.9	30.1	23.4	16.2	44.59
	CADA-VAE (Schonfeld et al., 2019)	50.5	39.6	28.2	38.1	32.9	28.1	26.34
	TF-VAEGAN (Li et al., 2021b)	51.5	41.9	30.0	40.1	35.0	29.2	291.73
	CE-GZSL (Li et al., 2021b)	53.6	44.7	32.1	42.9	35.9	32.1	156.08
	ZSC-SA (Quan et al., 2018)	31.3	19.1	13.6	–	–	–	–
	LPDCMENs (Li et al., 2021c)	43.8	24.9	21.6	–	–	–	–
	DAN (Li et al., 2021b)	53.3	45.2	33.4	40.3	34.1	29.6	–
	DSVA (RN18) (Ours)	58.7	46.6	35.9	50.7	36.0	33.9	43.01
	DSVA (ViT) (Ours)	76.2	50.1	44.6	58.2	36.6	35.3	334.20
RSMM-Attributes	DSVA (RN18) (Ours)	69.8	48.7	36.4	52.1	37.6	37.3	43.01
	DSVA (ViT) (Ours)	**84.0**	**64.2**	**60.2**	**67.7**	**56.5**	**49.9**	334.20

(60/10), 1.9% (50/20), and 4.3% (40/10). When using a vision transformer (ViT) as the backbone, our performance is further boosted. This indicates that our DSVA model that enforces the alignment between visual features and class embeddings is able to balance the performance of unseen and seen classes, and decrease the bias towards seen classes.

Table 4 displays the Top-1 ZSL accuracy of different models trained with two class embeddings, where our performance is comparable to or better than other SOTA methods. Notably, when trained with Word2Vec and BERT embeddings, our DSVA model outperforms other models by a large margin, *e.g.*, up to 30%. For instance, the DSVA (RN18) model trained with Word2Vec improves the accuracy of DAN (Li et al., 2021b) from 44.3% (60/10) to 55.7% (60/10). The DSVA (RN18) model trained with BERT embeddings improves the accuracy of CADA-VAE (Schonfeld et al., 2019) from 48.1% (60/10) to 59.4% (60/10). The impressive improvement demonstrates the ability of our model to recognize unseen classes with the help of different class embeddings.

In Table 5, we compare our DSVA model trained with RSMM-Attribute and the best performance of other SOTA models. For fair comparison, we train our model with two different class embeddings, *i.e.*, SR-RSKG (Li et al., 2021b) and our RSMM-Attributes. Besides, We build our DSVA model with two different backbones, *i.e.*, ResNet18 (He et al., 2016) (denoted as RN18) and Vision Transformer (denoted as ViT) (Dosovitskiy et al., 2020). Under the ZSL setting, compared with all other state-of-the-art models, our model yields consistent improvement on three dataset splits. Compared to the recent proposed non-generative model LPDCMENs designed for ZSL remote sensing scene classification, our DSVA (RN18) trained with SR-RSKG gained 14.9% (60/10), 21.7% (50/20), and 14.3% (40/30), respectively. When trained with our RSMM-Attributes, the ZSL performance of our DSVA model is further improved. Our DSVA (RN18) improves generative model DAN that synthesizes images for unseen classes from 53.3% to 69.8 (60/10), from 45.2% to 48.7% (50/20), and from 33.4 to

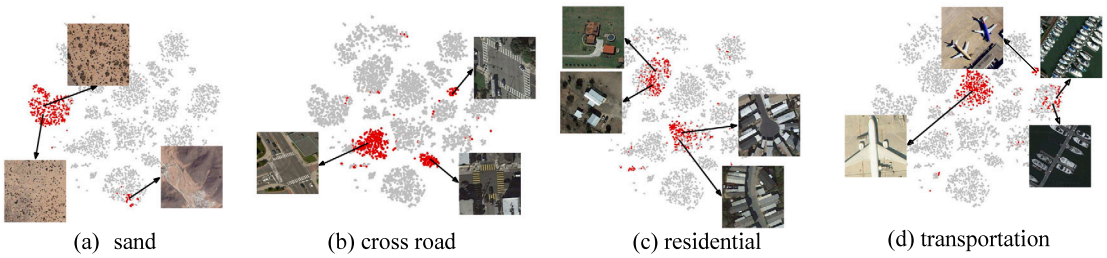

Fig. 7. The t-SNE visualization for four attributes. The red dots indicate images that activate the attribute, and we show several image examples in the cluster center. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

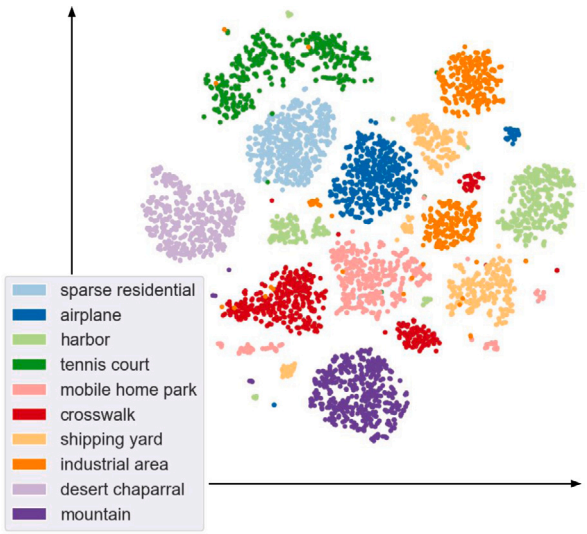

Fig. 8. Exploring scene images in 2-D attribute space, where dots with different colors represent images from various categories, and the visualization is finished by t-SNE (Van der Maaten and Hinton, 2008). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Class	Image	Top-5 attribute	Value	Bottom-5 attribute	Value
sparse residential		buildings	0.1226	flowers	0.0785
		bungalow	0.1221	ocean	0.0807
		symmetrical	0.1207	shopping	0.0843
		brick	0.1153	red	0.0858
		grass	0.1143	marble	0.0870
basketball court		symmetrical	0.1189	snow	0.0753
		flat	0.1152	sand	0.0788
		asphalt	0.1145	mountain	0.0855
		rectangle	0.1138	cloud	0.0872
		square	0.1136	hiking	0.0873
park		still water	0.1196	bungalow	0.0793
		curved road	0.1179	car	0.0799
		trees	0.1150	tan	0.0810
		forestry	0.1143	orange	0.0819
		running water	0.1140	smoke	0.0825

Fig. 9. Examples of the ground truth RSSM-Attributes embeddings. We display the example image, attribute name, and attribute value for three classes. Top-5 attributes denote the attributes with the highest attribute value for a certain class, and Bottom-5 attributes denote those with the lowest attribute value.

40.5 (40/30)%. Compared with the SOTA ZSL model CE-GZSL (Han et al., 2021) designed for ordinary optical images, our model DSVA trained with a light backbone ResNet18 already significantly improves the accuracy with only 43.01 MB parameters, *i.e.*, our model achieves 69.8% (60/10), 48.7% (60/10), and 36.4% (40/30) for ZSL accuracy, while the CE-GZSL model only achieves 53.6% (60/10), 44.7% (60/10), and 32.1% (40/30) with much more parameters (156.08 MB). When applying a larger model ViT with 334.20 MB parameters as the backbone, the ZSL accuracy of our model is significantly boosted to 84.0% (60/10), 64.2% (60/10), and 60.2% (40/30). The results indicate that our model can already outperform all other ZSL models with very few parameters. When using a vision transformer (ViT) as the backbone, associating the global information of remote scene images is quite useful for zero-shot model to recognize unseen classes. The same trend is observed under the GZSL setting in Table 5 (right), where the DSVA (ViT) model yields the best performance over all other SOTA models. It indicates that even under the realistic setting where the model needs to recognize seen and unseen classes simultaneously, our non-generative model DSVA still yields the best generalization ability and outperforms other generative models.

6. Attribute visualization

We start by visualizing the class distribution in the attribute space in Fig. 8, where 10 unseen classes are selected with 1,000 images per class. For each image x, we extract the RSMM-Attribute value $r_{\mathcal{A}}(x) \in \mathbb{R}^{N_{\mathcal{A}}}$, where each dimension indicates the similarity $f_{sim}\left(E_t(a), E_v(x)\right)$ between the image x and the corresponding attribute a. Then t-SNE (Van der Maaten and Hinton, 2008) is used to embed the attribute for all 10,000 images into a 2-D space. Fig. 8 demonstrates that the attribute space is class discriminative, and various categories can be well separated apart. Besides, attributes can reflect both visual and semantic relatedness. For instance, the visually similar classes, *e.g.*, *mobile home park* and *crosswalk*, locate near each other in the attribute space. This is because the RSSM-Attributes can encode not only the visual information in each image, but also the semantic relatedness according to the text description simultaneously. This property can benefit the zero-shot generalization, which transfers visual knowledge with semantic class embeddings.

Then we explore where images with different attributes live in the attribute space. In Fig. 7, we use red dots to represent images that activate a certain attribute, *i.e.*, the top 10% images that have the highest attribute value, and mark other images with gray color dots. The visualization demonstrated that images containing the same attributes tend to live together in the 2-D attribute space. Meanwhile, images sharing some same attributes may differ from each other according to other attributes. As shown in the figure, the image clusters for each attribute are usually split apart according to their overall appearance similarity. For instance, in Fig. 7(a), both dessert and mountain contain attribute "sand", and in Fig. 7(d), both airport and harbor has the attribute "transportation". This indicates that the RSMM-Attribute can not only link images sharing the same attribute together, but also discriminate images from various categories apart.

We display some examples of the ground truth RSSM-Attributes embedding for several classes in Fig. 9. Since there are 98 attributes for each class, to save space, we display the top 5 attributes with the highest attribute value and the bottom 5 attributes with the lowest attribute value. As shown in Fig. 9, the top 5 attributes indicate the

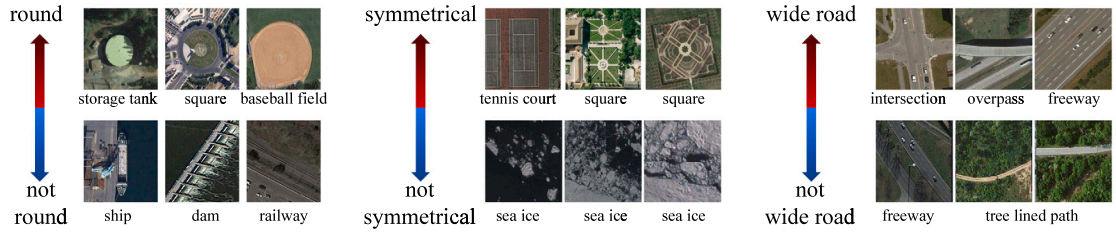

Fig. 10. Example (de-)activated images for three RSMM-Attributes, *i.e.*, "round", "symmetrical", and "wide road". For each attribute, the images near the red arrow indicate the positive samples that have the highest attribute value, while images near the blue arrow are negative samples having the lowest attribute value. The text under each image denotes the category label for them. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

most representative visual and semantic properties for each class. For example, for the class basketball court, attributes "symmetrical" and "flat" represents the texture, "asphalt" represents the materials, "rectangle" and "square" represents the shape. Conversely, attributes that are not commonly associated with basketball courts, such as "snow", "sand", "mountain", "cloud", and "hiking" have low attribute values.

In Fig. 10, we display some image examples of our annotated RSMM-Attributes. For each attribute, we select images that are annotated with the highest attribute value, and images with the lowest attribute value. We can observe the following. First, the activated images that have the highest attribute value all convey the attribute properties correctly, *e.g.*, the "wide road" existing in *intersection*, *overpass*, and *freeway*, and the "symmetrical" shapes in the *tennis court* and *square*. Moreover, instead of activating images with the same visual appearance, the positive images contain attributes with various visual cues. For instance, though indicating different objects, the round roof in the *storage tank*, the circle road in the *square*, and the round playing ground in the *baseball field* all activate the "round" attributes. This observation demonstrates that our model can discover the same attribute on different objects and facilitate attribute sharing and knowledge transfer across classes. In addition, the negative examples for each attribute are not conveying arbitrary properties but show semantically opposite properties. For example, the negative examples for "round" are images containing straight stripes, and the negative examples for "wide road" are scenes holding the narrow road. This interesting observation indicates the advantage of using the semantic-visual pre-training network, which encodes fluent semantic information visual attribute space and is beneficial for zero-shot learning where each category is described by semantic attributes.

7. Conclusion

Driven by the increasing demand for recognizing previous unseen classes in RS scenarios, we aim to improve the performance of ZSL models for RS scene classification in this work. To alleviate the manual effort needed in attribute annotation, we propose a semantic-visual multi-modal network to annotate visually detectable attributes for each category automatically. Moreover, a Deep Semantic-Visual Alignment model is proposed to map visual images into the attribute space and classify images from both seen and unseen classes simultaneously. We explicitly encourage the model to associate local image regions together for better representation learning with the help of self-attention. Moreover, an attention concentration module is proposed to focus on the informative attribute regions. With extensive experiments, we demonstrate that our model improves over the state-of-the-art model by a large margin. Moreover, we qualitatively verify that the RSMM-Attributes annotated by our network are both class discriminative and semantic related, which benefits the zero-shot knowledge transfer between seen and unseen classes.

CRediT authorship contribution statement

Wenjia Xu: Conceptualization of this study, Methodology, Software. **Yirong Wu:** Data curation, Writing - Original draft preparation.

Declaration of competing interest

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

Acknowledgments

This paper was supported in part by the National Key Research and Development Program of China under Grant 2021YFB2900200, National Natural Science Foundation of China under No. 61925101, and 61831002, 61921003, the Beijing Municipal Science and Technology, China Project NO. Z211100004421017.

References

Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C., 2015. Label-embedding for image classification. IEEE Trans. Pattern Anal. Mach. Intell. 38 (7), 1425–1438.

Al-Halah, Z., Stiefelhagen, R., 2017. Automatic discovery, association estimation and learning of semantic attributes for a thousand categories. In: IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 614–623.

Alajaji, D., Alhichri, H.S., Ammour, N., Alajlan, N., 2020. Few-shot learning for remote sensing scene classification. In: 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium. M2GARSS, IEEE, pp. 81–84.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. 5, 135–146.

Chao, W.-L., Changpinyo, S., Gong, B., Sha, F., 2016. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: European Conference on Computer Vision. Springer, pp. 52–68.

Chen, Q., Cheng, Q., Wang, J., Du, M., Zhou, L., Liu, Y., 2021. Identification and evaluation of urban construction waste with VHR remote sensing using multi-feature analysis and a hierarchical segmentation method. Remote Sens. 13 (1), 158.

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

Cheng, G., Xie, X., Han, J., Guo, L., Xia, G.-S., 2020. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 3735–3756.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: IEEE/CVF Computer Vision and Pattern Recognition Conference.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Elhoseiny, M., Elfeki, M., 2019. Creativity inspired zero-shot learning. In: IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 5784–5793.

Farhadi, A., Endres, I., Hoiem, D., Forsyth, D., 2009. Describing objects by their attributes. In: IEEE/CVF Computer Vision and Pattern Recognition Conference. IEEE, pp. 1778–1785.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: NeurIPS, Vol. 27.

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

Han, Z., Fu, Z., Chen, S., Yang, J., 2021. Contrastive embedding for generalized zero-shot learning. In: IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 2371–2381.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE/CVF Computer Vision and Pattern Recognition Conference.

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

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

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

Li, H., Cui, Z., Zhu, Z., Chen, L., Zhu, J., Huang, H., Tao, C., 2020a. RS-MetaNet: Deep meta metric learning for few-shot remote sensing scene classification. arXiv preprint arXiv:2009.13364.

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

Li, Y., Kong, D., Zhang, Y., Chen, R., Chen, J., 2021a. Representation learning of remote sensing knowledge graph for zero-shot remote sensing image scene classification. In: International Geoscience and Remote Sensing Symposium. IEEE, pp. 1351–1354.

Li, Y., Kong, D., Zhang, Y., Tan, Y., Chen, L., 2021b. Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification. ISPRS J. Photogramm. Remote Sens. 179, 145–158.

Li, J., Lin, D., Wang, Y., Xu, G., Zhang, Y., Ding, C., Zhou, Y., 2020b. Deep discriminative representation learning with attention map for scene classification. Remote Sens. 12 (9).

Li, A., Lu, Z., Wang, L., Xiang, T., Wen, J.-R., 2017b. Zero-shot scene classification for high spatial resolution remote sensing images. IEEE Trans. Geosci. Remote Sens. 55 (7), 4157–4167.

Li, J., Pei, Y., Zhao, S., Xiao, R., Sang, X., Zhang, C., 2020c. A review of remote sensing for environmental monitoring in China. Remote Sens. 12 (7), 1130.

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

Li, Z., Zhang, D., Wang, Y., Lin, D., Zhang, J., 2022. Generative adversarial networks for zero-shot remote sensing scene classification. Appl. Sci. 12 (8), 3760.

Li, Y., Zhu, Z., Yu, J.-G., Zhang, Y., 2021c. Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification. IEEE Trans. Geosci. Remote Sens. 59 (12), 10590–10603.

Liu, Y., Zhou, L., Bai, X., Huang, Y., Gu, L., Zhou, J., Harada, T., 2021. Goal-oriented gaze estimation for zero-shot learning. In: IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 3794–3803.

Lu, X., Wang, B., Zheng, X., Li, X., 2017. Exploring models and data for remote sensing image caption generation. IEEE Trans. Geosci. Remote Sens. 56 (4), 2183–2195. http://dx.doi.org/10.1109/TGRS.2017.2776321.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: NeurIPS.

Narayan, S., Gupta, A., Khan, F.S., Snoek, C.G., Shao, L., 2020. Latent embedding feedback and discriminative features for zero-shot classification. In: European Conference on Computer Vision. Springer, pp. 479–495.

Nayak, N.V., Bach, S.H., 2021. Zero-shot learning with common sense knowledge graphs. In: International Conference on Learning Representations.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Park, J., Woo, S., Lee, J.-Y., Kweon, I.S., 2018. Bam: Bottleneck attention module. In: British Machine Vision Conference.

Patterson, G., Xu, C., Su, H., Hays, J., 2014. The sun attribute database: Beyond categories for deeper scene understanding. Int. J. Comput. Vis. 108.

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing.

Quan, J., Wu, C., Wang, H., Wang, Z., 2018. Structural alignment based zero-shot classification for remote sensing scenes. In: 2018 IEEE International Conference on Electronics and Communication Engineering. ICECE, IEEE, pp. 17–21.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: ICML. PMLR, pp. 8748–8763.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI Blog 1 (8), 9.

Schonfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., Akata, Z., 2019. Generalized zero- and few-shot learning via aligned variational autoencoders. In: IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 8247–8255.

Sumbul, G., Cinbis, R.G., Aksoy, S., 2017. Fine-grained object recognition and zero-shot learning in remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 56 (2), 770–779.

Tao, S.-Y., Yeh, Y.-R., Wang, Y.-C.F., 2017. Semantics-preserving locality embedding for zero-shot learning. In: British Machine Vision Conference.

Toth, C., Jóźków, G., 2016. Remote sensing platforms and sensors: A survey. ISPRS J. Photogramm. Remote Sens. 115, 22–36.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: NeurIPS, Vol. 30.

Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., 2011. The Caltech-UCSD Birds-200–2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wang, Q., Huang, W., Xiong, Z., Li, X., 2020. Looking closer at the scene: Multiscale representation learning for remote sensing image scene classification. IEEE Trans. Neural Netw. Learn. Syst..

Wang, C., Peng, G., De Baets, B., 2021. A distance-constrained semantic autoencoder for zero-shot remote sensing scene classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 12545–12556.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: European Conference on Computer Vision.

Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Remote Sens. 55 (7), 3965–3981.

Xian, Y., Lampert, C.H., Schiele, B., Akata, Z., 2019. Zero-shot learning-A comprehensive evaluation of the good, the bad and the ugly. IEEE Trans. Pattern Anal. Mach. Intell..

Xian, Y., Lorenz, T., Schiele, B., Akata, Z., 2018. Feature generating networks for zero-shot learning. In: IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 5542–5551.

Xu, W., Wang, J., Wang, Y., Xu, G., Lin, D., Dai, W., Wu, Y., 2020a. Where is the model looking at?–Concentrate and explain the network attention. IEEE J. Sel. Top. Sign. Proces. 14 (3), 506–516.

Xu, W., Xian, Y., Wang, J., Schiele, B., Akata, Z., 2020b. Attribute prototype network for zero-shot learning. In: NeurIPS, Vol. 33. pp. 21969–21980.

Xu, W., Xian, Y., Wang, J., Schiele, B., Akata, Z., 2022a. Attribute prototype network for any-shot learning. Int. J. Comput. Vis. 1–19.

Xu, W., Xian, Y., Wang, J., Schiele, B., Akata, Z., 2022b. VGSE: Visually-grounded semantic embeddings for zero-shot learning. In: IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 9316–9325.

Xue, Z., Du, P., Li, J., Su, H., 2017. Sparse graph regularization for robust crop mapping using hyperspectral remotely sensed imagery with very few in situ data. ISPRS J. Photogramm. Remote Sens. 124, 1–15.

Yang, Y., Newsam, S., 2011. Spatial pyramid co-occurrence for image classification. In: International Conference on Computer Vision. IEEE, pp. 1465–1472.

Zhang, Z., Saligrama, V., 2016. Zero-shot learning via joint latent similarity embedding. In: IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 6034–6042.

Zhao, Q., Jia, S., Li, Y., 2021. Hyperspectral remote sensing image classification based on tighter random projection with minimal intra-class variance algorithm. Pattern Recognit. 111, 107635.

Zhu, P., Wang, H., Saligrama, V., 2019a. Zero shot detection. IEEE Trans. Circuits Syst. Video Technol. 30 (4), 998–1010.

Zhu, Y., Xie, J., Liu, B., Elgammal, A., 2019b. Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning. In: International Conference on Computer Vision.