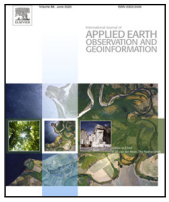
RS-CLIP: zero shot remote sensing scene classification via contrastive vision-language supervision

International Journal of Applied Earth Observation and Geoinformation 124 (2023) 103497

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



RS-CLIP: Zero shot remote sensing scene classification via contrastive vision-language supervision

Xiang Li [a], Congcong Wen [b], Yuan Hu [c], Nan Zhou [d,*]

[a] King Abdullah University of Science and Technology, Jeddah, 23955, Saudi Arabia
[b] New York University Abu Dhabi, Abu Dhabi, 129188, United Arab Emirates
[c] Institute of Remote Sensing and Geographic Information Systems, School of Earth and Space Sciences, Peking University, Beijing, 100871, China
[d] Jiangsu University of Science and Technology, Zhenjiang, 212100, China

ARTICLE INFO

Keywords:
Remote sensing scene classification
Vision-language model
CLIP
Pseudo labeling
Curriculum learning

ABSTRACT

Zero-shot remote sensing scene classification aims to solve the scene classification problem on unseen categories and has attracted numerous research attention in the remote sensing field. Existing methods mostly use shallow networks for visual and semantic feature learning, and the semantic encoder networks are usually fixed during the zero-shot learning process, thus failing to capture powerful feature representations for classification. In this work, we introduced a vision-language model for remote sensing scene classification based on contrastive vision-language supervision. Our method is capable of learning semantic-aware visual representations using a contrastive vision-language loss in the embedding space. By pretraining on large-scale image–text datasets, our baseline method shows good transferring ability on remote sensing scenes. To enable model training in zero-shot settings, we introduced a pseudo-labeling technique that can automatically generate pseudo labels from unlabeled data. A curriculum learning strategy is developed to boost the performance of zero-shot remote sensing scene classification with multiple stages of model finetuning. We conducted experiments on four benchmark datasets and showed considerable performance improvement on both zero-shot and few-shot remote sensing scene classification. The proposed RS-CLIP method achieved a zero-shot classification accuracy of 95.94%, 95.97%, 85.76%, and 87.52% on the novel classes of UCM-21, WHU-RS19, NWPU-RESISC45, and AID-30 datasets respectively. Our code will be released at https://github.com/lx709/RS-CLIP.

1. Introduction

Remote sensing scene classification (RSSC) takes a whole scene image as input and tries to predict a semantic category that characterizes ground objects and structures in the image. It is a crucial task in remote sensing image analysis and has numerous real-world applications, such as land use mapping, object detection, and image retrieval (Chen and Tsou, 2021; Jin et al., 2022; Zhu et al., 2023). In recent years, deep learning methods, especially Convolutional Neural Networks (CNNs), have shown impressive performance in this task. However, existing methods typically require a large amount of annotated data for training and cannot generalize to new categories without additional data.

Unlike machine vision methods, humans can recognize objects from new categories by comparing object descriptions to previously learned notions (Romera-Paredes and Torr, 2015). For example, a child who knows what a horse looks like can readily spot a zebra by recognizing it as resembling a horse with black-and-white stripes. Inspired by human vision, zero-shot learning (ZSL) has been developed to tackle the problem of identifying objects from unseen classes by transferring knowledge from seen classes. ZSL has received extensive study in the field of machine learning (Wang et al., 2019).

In the remote sensing field, Li et al. (2017) introduced the first attempt at zero-shot learning for the RSSC task. They leveraged the word2vec (Mikolov et al., 2013) technique to generate semantic vectors for both seen and unseen classes and built a semantic directed graph to characterize the relationships between different semantic classes. Classification of unseen classes can be accomplished by label-propagation on the semantic graph. To address the inconsistency between the visual space and semantic space, Quan et al. (2018) employed a semi-supervised Sammon embedding algorithm (Sammon, 1969) to align semantic and visual prototypes, improving the synthesis capabilities of unseen class prototypes in the visual space. Li et al. (2021) introduced a deep cross-modal embedding network for zero-shot RSSC and developed several locality-preservation constraints on both visual and semantic embeddings to address class structure inconsistency. Wang et al. (2021) proposed a distance-constrained semantic autoencoder

* Corresponding author.
E-mail addresses: xiangli92@ieee.org (X. Li), cw3437@nyu.edu (C. Wen), huyuan@pku.edu.cn (Y. Hu), zhounan@just.edu.cn (N. Zhou).

https://doi.org/10.1016/j.jag.2023.103497
Received 13 March 2023; Received in revised form 9 September 2023; Accepted 18 September 2023
Available online 6 October 2023

to align visual features and semantic representations for the zero-shot RSSC task.

Nevertheless, existing methods mostly use a word2vec model pretrained on the Wikipedia corpus to extract semantic embeddings from category names or descriptions. The semantic embeddings are preprocessed and fixed during the zero-shot learning process without adapting to visual features to be aligned. This can lead to insufficient representation capability of the extracted semantic embeddings and considerable discrepancies between visual and semantic features. Another challenge posed by previous methods is that they typically utilize shallow networks for learning visual and semantic features. This is because remote sensing scene datasets used in these methods, such as UCM (Yang and Newsam, 2010) and WHU-RS19 (Dai and Yang, 2011) datasets, are at small scales, with less than 100k scene images. Using too deep networks will cause over-fitting issues of these methods. However, shallow networks are incapable of learning high-level representative features.

To solve these issues, we introduced a vision-language model for remote sensing scene understanding in this paper. In recent years, vision-language models have been widely explored in computer vision, and numerous foundation models are built for various visual recognition tasks (Radford et al., 2021; Huang et al., 2022; Jia et al., 2021; Yuan et al., 2021), especially for zero-shot and few-shot learning. Unlike self-supervised visual feature learning methods, vision-language models can learn powerful visual feature representations and directly connect visual representations with natural languages in a holistic framework, thus enabling better zero-shot transfer under the guidance of semantic knowledge (Radford et al., 2021).

In this work, we introduced a vision language model for remote sensing scene classification based on the pretrained CLIP (Radford et al., 2021) model, denoted as RS-CLIP. Note that other vision language models (e.g., ALIGN Jia et al., 2021) can be used to replace the CLIP model in our method. In contrast to previous zero-shot RSSC methods, which primarily depend on distinct visual and semantic feature extraction, our approach excels in acquiring semantic-aware visual representations through unified visual–semantic feature learning. We experimentally found that the CLIP model, pretrained on extensive image–text datasets, showed strong transferability when applied to remote sensing scenes. To adapt the model to the remote sensing domain, we introduced a pseudo labeling technique that can automatically generate pseudo labels from unlabeled datasets, thus enabling model finetuning on the remote sensing domain. Furthermore, a curriculum learning strategy is developed to boost the performance of zero-shot remote sensing scene classification with multiple stages of model finetuning. We conducted experiments on four public benchmark datasets, and the experimental results demonstrated that our model yields considerable performance improvement on both zero-shot and few-shot remote sensing scene classification.

Our contributions are summarized as follows:

- In this paper, we introduced a CLIP-based vision-language model for zero-/few-shot remote sensing scene classification.
- We introduced a pseudo-labeling technique that can automatically generate pseudo-labels from unlabeled data. Moreover, a curriculum learning strategy is developed to boost the performance of zero-/few-shot remote sensing scene classification.
- We conducted experiments on four benchmark datasets. Our model performed significantly better than previous state-of-the-art methods on both zero-shot and few-shot remote sensing scene classification.

2. Related works

2.1. Zero-shot classification

Zero-shot learning (ZSL) aims to learn a model that can identify objects of unseen classes by transferring knowledge learned from seen classes, where semantic information of both seen and unseen classes is provided. Semantic information can be obtained from pre-defined attribute vectors (Lampert et al., 2009), word or context-based embedding (Socher et al., 2013; Fu et al., 2017), or their combinations (Song et al., 2020). They will be used to build connections between seen and unseen classes. ZSL methods usually work by transforming images and semantic descriptors into a shared embedding space, where samples from the same class are supposed to cluster around the corresponding class-level semantic descriptor. In the test stage, the model can predict labels for images of unseen classes by searching the nearest semantic descriptor in the embedding space.

In recent years, numerous methods have been developed for zero-shot learning. Early efforts use well-defined hand-engineered semantic descriptions for different classes. In general, there are three commonly used hand-engineered semantic descriptors, including visual attribute (Lampert et al., 2009; Palatucci et al., 2009), lexical (Ma et al., 2016; Palatucci et al., 2009), and text-keyword (Lei Ba et al., 2015; Elhoseiny et al., 2013). Lampert et al. (2009) introduced an attribute-based zero-shot learning approach for animal classification. This type of method requires manually annotated visual attributes from human experts, which is time-consuming and less practical for large-scale datasets. Recent studies primarily focus on learning-based semantic descriptions. These methods usually use a pretrained language model from a large-scale text database to extract semantic embeddings for each class, eliminating the need for human annotations of visual attributes. Existing studies have exploited different embedding techniques, such as Word2Vec (Wang et al., 2016) and GloVe (Xian et al., 2016). For example, Wang et al. (2016) adopted Word2Vec to learn category representations from Wikipedia for zero-shot classification tasks. Hybrid semantic embedding methods have also been explored to improve the representation abilities for diverse class descriptions.

Based on semantic embeddings, existing zero-shot learning methods can be divided into three categories. The first category of methods transforms visual features to semantic space (Norouzi et al., 2013) and predicts labels for images from unseen classes by measuring the similarities of semantic embeddings. Bucher et al. (2016) proposed a ZSL method that projects visual features to semantic space and applied a metric learning technique to control the structure of the embedding space. Guo and Guo (2020) leveraged an autoencoder model to generate auxiliary semantic features from visual features to better align the manifold structures between visual and semantic features. The second category of methods projects semantic features into a visual space (Zhang et al., 2017; Pan et al., 2020; Shigeto et al., 2015). Zhang et al. (2017) argued that visual space is more discriminative than semantic space and developed a deep neural network to map semantic features to visual space. Similarly, Pan et al. (2020) proposed a zero-shot classification method that maps semantic features to visual space through a cosine distance-based objective function. The last category of methods transforms visual and semantic features into a shared subspace. For example, Demirel et al. (2017) proposed to learn discriminative word representations such that semantic class similarities are aligned with visual similarities. Ding et al. (2017) introduced a new technique called low-rank embedded semantic dictionary learning to link visual and semantic representations. These zero-shot learning methods have achieved promising results for natural image classification. A comprehensive review of zero-shot learning can be found at Wang et al. (2019).

2.2. Zero-shot remote sensing scene classification

Unlike natural images, objects in remote sensing images tend to have significant structural and contextual variations, which makes it harder for a model to learn robust visual features for scene understanding. Li et al. (2017) introduced the first zero-shot learning-based remote sensing scene classification method. In Li et al. (2017), the

authors leveraged a word2vec model pretrained on the Wikipedia corpus to extract semantic embeddings from category names. A semantic graph was then built to characterize the relationships between semantic classes. Quan et al. (2018) further improved the method by introducing a semi-supervised Sammon embedding algorithm (Sammon, 1969) to align semantic and visual prototypes. Sumbul et al. (2017) introduced a zero-shot learning method for fine-grained remote sensing image classification. A compatibility function was learned to build the connection between image features and semantic embeddings that enables knowledge transferring from seen to unseen classes. Inspired by Kodirov et al. (2017), Wang et al. (2021) developed a distance-constrained semantic autoencoder to align the visual features and semantic representations for zero-shot RSSC task. In Li et al. (2021), the author adopted a transformer-based language model (e.g., BERT (Kenton and Toutanova, 2019)) to extract semantic embeddings from expert-defined text descriptions of all classes. Li et al. (2022) introduced a generative adversarial network (GAN)-based method for zero-shot RSSC, where a generator network was trained to generate image features from class semantics, converting the zero-shot classification problem into a traditional image classification problem. Moreover, the authors investigated different language processing models, i.e., Word2vec (Wang et al., 2016), Fasttest (Joulin et al., 2017; Bojanowski et al., 2017), Glove (Xian et al., 2016), and BERT (Kenton and Toutanova, 2019), for semantic embedding extraction from class names or descriptions.

2.3. Contrastive vision-language model

Thanks to the powerful and flexible feature representation capabilities of deep learning models, the fields of natural language processing and computer vision have merged onto a shared trajectory, resulting in a flourishing research landscape in the realm of vision-language understanding. Vision-language models are developed to learn visual representations from language supervision. These models typically consist of two parts: a vision encoder network, such as ResNet (He et al., 2016a), ViT (Dosovitskiy et al., 2021), or Swin Transformer (Liu et al., 2021), and a language encoder network using standard Transformers (Vaswani et al., 2017). To learn useful features, contrastive learning objectives are often applied to align image and language features in the embedding space. In recent years, vision-language models have demonstrated excellent performance in visual representation learning and transfer learning (Radford et al., 2021; Huang et al., 2022). For example, CLIP (Radford et al., 2021) introduced a simple but effective vision-language model pre-trained on large-scale image–text pairs and achieved remarkable results on over 30 diverse computer vision datasets. ALIGN (Jia et al., 2021) built a vision-language model using 1.8 billion noisy image–text pairs. Florence (Yuan et al., 2021) developed a new foundation model that enables fine-grained, dynamic, and multi-modality vision tasks, trained on a 900 million image–text pair dataset using universal visual-language representations. Unlike self-supervised pre-trained foundation models for visual representation learning, vision-language models have inherent transfer abilities based on semantic supervision and have been successfully applied to various vision tasks, such as object detection (Du et al., 2022; Gu et al., 2021), semantic segmentation (Xu et al., 2022), and 3D recognition (Zhang et al., 2022b).

In remote sensing, several works explored the CLIP model for remote sensing image analysis. The most similar work comes from (Qiu et al., 2022). In this work, Qiu et al. introduced a similar idea that uses the pretrained CLIP model for remote sensing image feature extraction and achieved promising scene classification results using only a few labels. In Djoufack Basso (2022), the author developed a cross-modal remote sensing image retrieval platform based on a pretrained CLIP and a text-based image retrieval model. In Bazi et al. (2022), Bazi et al. leveraged the pretrained CLIP model to extract image and text features and developed a visual question answering method based on fused feature representations. Other vision-language foundation models have also been explored for remote sensing image understanding tasks. For example, Sun et al. (2022) built a vision foundation model based on masked image modeling and achieved remarkable performance on eight remote sensing image datasets across four downstream tasks. Chen et al. (2023) introduced a visual foundation model based on SAM (Kirillov et al., 2023) for instance segmentation in remote sensing images. Hu et al. (2023) introduced a Generative Pre-trained Transformers (GPT)-based vision-language foundation model named RSGPT that enables remote sensing image captioning and visual question answering.

2.4. Pseudo labeling and curriculum learning

Pseudo Labeling (Lee et al., 2013) is a semi-supervised learning (SSL) technique that generates artificial labels for unlabeled data based on predictions of the model trained firstly on the labeled data. In Rosenberg et al. (2005), a confidence-based strategy was applied in combination with pseudo-labeling, ensuring that unlabeled data is only utilized when model predictions are sufficiently confident. Similarly, UDA (Xie et al., 2020), ReMixMatch (Berthelot et al., 2019), and FixMatch (Sohn et al., 2020) utilized the confidence-based thresholding approach, but they heavily depend on the implementation of robust data augmentations to enhance consistency regularization.

The combination of curriculum learning and pseudo labeling has gained popularity recently (Gong et al., 2016; Yu et al., 2020; Han et al., 2020; Zheng and Yang, 2021; Cascante-Bonilla et al., 2021) due to its competitive results for various image datasets. Unlike previous methods that use a fixed threshold, it utilizes adaptive scores to determine which samples to select as pseudo labels. Initially, samples with top $r\%$ confidence scores are selected as pseudo-labels, and the percentile is gradually increased during the self-training cycle until all unlabeled data is utilized. For example, Cascante-Bonilla et al. (2021) proposed curriculum labeling that selected unlabeled samples using a threshold that takes into account the distribution skew of the prediction scores on unlabeled samples. More recently, Kim et al. (2023) introduced a novel pseudo-labeling approach that was aimed to obtain more reliable pseudo-labels that are located in high-density regions by regularizing the confidence scores based on the likelihoods of the pseudo-labels.

3. Methods

In this section, we introduce our proposed zero-shot RSSC method based on vision-language models. We first briefly review the classical CLIP model in Section 3.1. Then, the proposed pseudo labeling and curriculum learning techniques are illustrated in Section 3.2 and Section 3.3, respectively. It should be noted that alternative vision language models, such as ALIGN (Jia et al., 2021) and Florence (Yuan et al., 2021), can also be employed in place of CLIP within our method.

3.1. Review of CLIP

The CLIP model learns visual representations using language supervision, as depicted in Fig. 1. Given a batch of N image–text pairs, the CLIP model attempts to predict the correct correspondences between the image and text inputs. To achieve this, the CLIP model employs a vision encoder network E_i to learn visual representations and a language encoder network E_t to learn text representations. During training, the CLIP model predicts a similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$, where each row indicates the probabilities of matching one image to all N texts. The CLIP model is optimized by maximizing the similarity scores of the N positive pairs and minimizing the similarity scores of the $N^2 - N$ negative pairs. This is achieved by optimizing a symmetric cross-entropy loss over the similarity matrix.

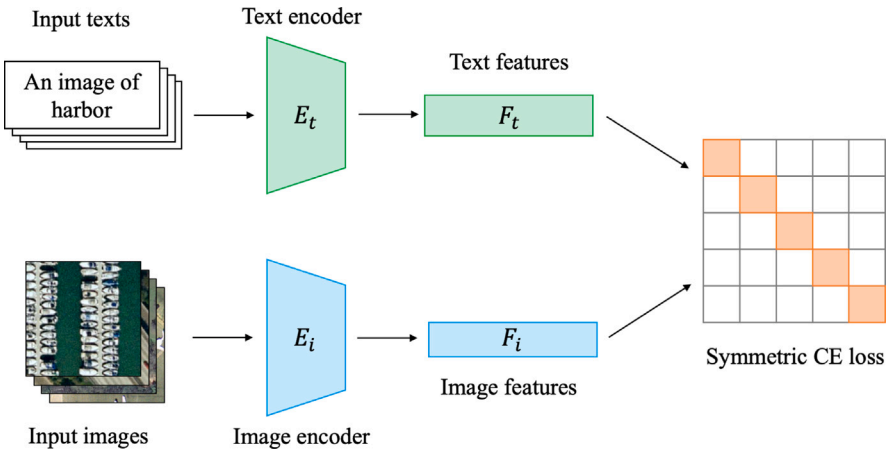


Fig. 1. Overview of CLIP model (Radford et al., 2021). An image encoder and a text encoder are utilized for feature extraction from visual and language modalities, respectively, with the inclusion of an asymmetric contrastive loss for model training.

For a downstream classification task with C categories, $\{1, 2, \ldots, C\}$, CLIP uses a pre-defined prompt, e.g., "an image of a [CLASS]", to formulate text inputs T, where the [CLASS] token denotes class names/descriptions of each category. Then, the semantic features of all classes can be generated using the text encoder network, i.e., $\mathbf{F}_t = E_t(T) \in \mathbb{R}^{C\times d}$, where d denotes the feature dimension. Given a batch of input images denoted as $\mathbf{I} \in \mathbb{R}^{B\times H\times W\times 3}$, where B, H, and W denote batch size, image height and image width, we can generate their visual features by passing them through the image encoder network, i.e., $\mathbf{F}_i = E_i(\mathbf{I}) \in \mathbb{R}^{B\times d}$. After that, the classification probability matrix can be obtained by,

$$\mathbf{P} = \text{Softmax}(\mathbf{F}_i \mathbf{F}_t^T / \tau) \tag{1}$$

where $\mathbf{F}_i$ and $\mathbf{F}_t$ are L2-normalized, and their matrix multiplication is equivalent to computing their cosine similarity. τ represents a learnable temperature parameter. A Softmax layer is applied to the class dimension, resulting in a probability matrix denoted as $\mathbf{P} \in \mathbb{R}^{B\times C}$. Each row of $\mathbf{P}$ denotes the probability of assigning one image to all possible classes. The final classification prediction can be obtained by selecting the class with the maximum probability,

$$\hat{Y} = \arg\max(\mathbf{P}) \tag{2}$$

3.2. Pseudo labeling

We do not have labeled samples of the target classes in the zero-shot setting. We resort to the pseudo labeling technique to enable the model training on target domain datasets, which is commonly used in semi-supervised learning for automatically generating pseudo samples from unlabeled data. The intuition behind pseudo labeling is that if a model gives high confidence scores on some samples, we can use the predicted labels as pseudo labels to re-train the model and improve the performance.

The CLIP model provides prior knowledge for diverse vision tasks, including related knowledge for remote sensing image understanding, as it was trained on a large-scale vision-language dataset. Thus, we use the CLIP as a prior model to generate pseudo labels for the remote sensing images in our zero-shot classification task. Previous semi-supervised learning methods usually select pseudo samples with a confidence score higher than a pre-defined threshold or dynamically adjust the threshold when more unlabeled samples are selected for training (Cascante-Bonilla et al., 2021). However, Huang et al. (2022) found that using pre-defined thresholds for the CLIP baseline model can lead to an imbalanced distribution of pseudo labels and, therefore, hurt the performance on downstream tasks. We follow (Huang et al., 2022) to select an identical number of samples for each class as pseudo labels, which prevents class overwhelming issues when selecting pseudo samples. Fig. 2 illustrates our pseudo labeling process. Specifically, for each class c, we select top-K samples with the highest confidence scores from the probability matrix in Eq. (1), which can be expressed as:

$$\pi_c = \text{top-K}(\mathbf{P}_c), \tag{3}$$

where P_c denotes the cth column of probability matrix $\mathbf{P}$. The overall pseudo-labeled samples can be obtained by the union of pseudo samples from all possible classes, calculated as,

$$D_L = U(\{\pi_c\}_{c=1}^{C}) \tag{4}$$

where U denotes set union.

3.3. Curriculum learning

The aforementioned pseudo labeling strategy only selects a small number of pseudo samples for model training. Directly selecting a large number of pseudo samples will inevitably include incorrect samples with low confidence, thus hurting the classification performance. To include more reliable unlabeled samples and improve the classification performance, we resort to a curriculum learning strategy that gradually selects more samples for model training in multiple rounds. In the early rounds, the model is less tuned on the target datasets, and thus, only a few confident samples are selected as pseudo data for training. In the later rounds, the model becomes more confident in classifying the target datasets, enabling the selection of more unlabeled samples as pseudo labels. More specifically, at iteration r, we select K_r samples as pseudo labels, where K_r is determined according to pseudo accuracy. Generally, we set $K_r \geq K_{r-1}$, which means more pseudo samples will be selected in the latter stage. The proposed curriculum learning process is illustrated in Fig. 3. Algorithm 1 shows the proposed curriculum learning process for zero-shot RSSC.

Algorithm 1 Curriculum learning for zero-shot RSSC.

1: Initialize the CLIP model using weights trained on WIT for the WebImageText dataset.
2: **for** iteration r={1,2,...R} **do**
3: (1) Predict classification probabilities for all unlabeled samples using Eq. (1) and Eq. (2).
4: (2) Select top-K_r samples with the highest probabilities for each class as pseudo labels according to Eq. (3) and Eq. (4).
5: (3) Retrain the CLIP model using pseudo labels.
6: (4) Update the CLIP model.
7: **end for**

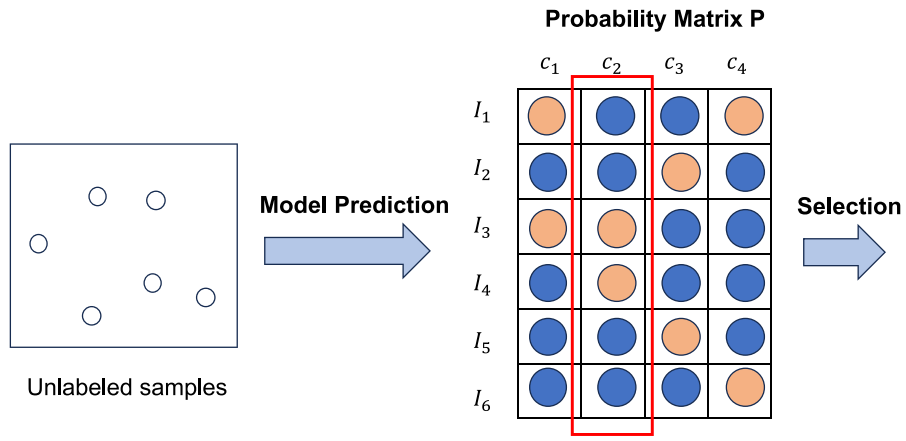


Fig. 2. Illustration of pseudo labeling. The example shows images from 4 classes and we select images with top-2 classification probabilities as pseudo labels. Note some samples can be selected by multiple classes.

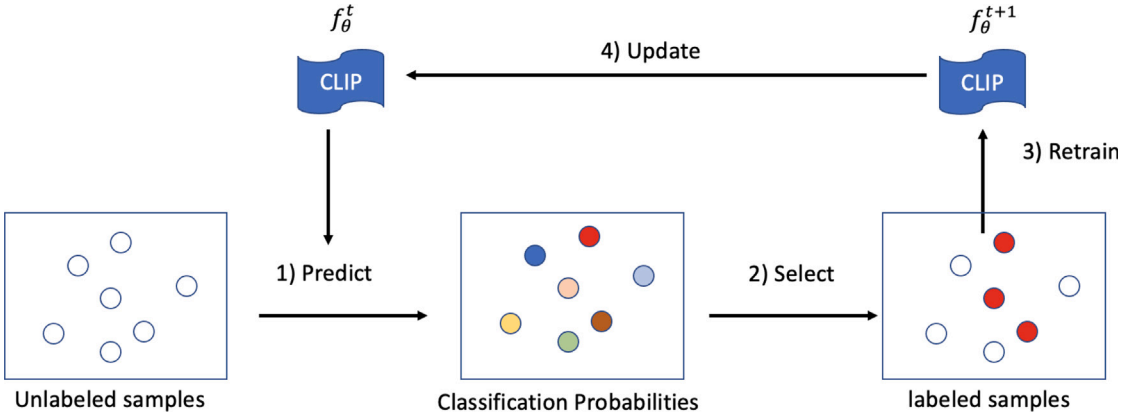


Fig. 3. Illustration of our proposed curriculum learning process. At each curriculum learning, our model predicts the classification probability for all unlabeled samples and selects pseudo samples based on their probability distribution. The CLIP model is then retrained using the selected pseudo samples and the updated model weights are used for the next curriculum learning stage.

4. Experiments and results

4.1. Datasets

We conducted experiments on four commonly used benchmark datasets for remote sensing scene classification, including UCM (Yang and Newsam, 2010), WHU-RS19 (Dai and Yang, 2011), NWPU-RESISC45 (Cheng et al., 2017), and AID (Xia et al., 2017). The UCM dataset is one of the most widely used datasets for remote sensing scene classification. It contains aerial images from 21 scene categories, and each category has 100 images of size 256 × 256. The WHU-RS19 dataset contains aerial images from 19 scene categories. There are, in total, 1,013 images with 600 × 600 pixels. The NWPU-RESISC45 dataset consists of aerial images from 45 scene categories, and each category has 700 images with a size of 256 × 256 pixels. The AID dataset contains 1000 images from 30 scene categories, and each image is of size 600 × 600 pixels. Table 1 summarizes the brief information of these datasets. Fig. 4 shows selected examples from the UCM dataset. For zero-shot RSSC, we follow (Li et al., 2017) to split the datasets.

4.2. Implementation details

In the curriculum learning stage, we train the model with selected pseudo labels, using an Adam optimizer with an initial learning rate of 1e–5. We train each model for 300 steps at each curriculum learning stage and decay the learning by 0.7 every 20 steps. The batch size is set to 24, and the momentum is set to 0.9. To improve the robustness of our model, we apply data augmentation during the training stage. Specifically, for each input image, we first resize it to 256 × 256 pixels and then crop the central area with a size of 224 × 224. A random horizontal and vertical flip is then applied to the cropped image in the training stage.

4.3. Hyper-parameter selection

We first investigate the effect of using different text prompts for language supervision. We conducted zero-shot scene classification experiments on the UCM dataset using five different prompts; the results are listed in Table 2. As shown in Table 2, different prompts show comparable performance, with a top-1 accuracy of around 70%. Fig. 6 shows the classification confusion matrix on the UCM and WHU-RS19 datasets. As can be seen, the CLIP model successfully predicts the correct class for most images, as indicated by the diagonal of the confusion matrix. The classification error mainly comes from the misclassification of semantically related classes. For example, the CLIP model incorrectly classified 49 spare residential and 76 medium residential images as dense residential, respectively.

Moreover, the CLIP model shows a top-5 accuracy of more than 94% on all benchmark datasets. The top-5 accuracy shows great potential to improve the zero-shot classification performance using pseudo labels. Fig. 5 shows examples of top-5 classification predictions from failure cases. As can be seen in the figure, although the CLIP model fails to predict the correct labels for these images in top-1 results, the error predictions mostly come from the categories that are highly related

Table 1
Dataset details.

	UCM	WHU-RS19	NWPU-RESISC45	AID
No. class	21	19	45	30
No. images	2,100	1,013	31,500	10,000
Size	256 × 256	600 × 600	256 × 256	600 × 600

Fig. 4. Remote sensing scene images from the UCM dataset. From left to right, we show scene images from categories of agriculture, airplane, baseball diamond, building, harbor, forest, and intersection.

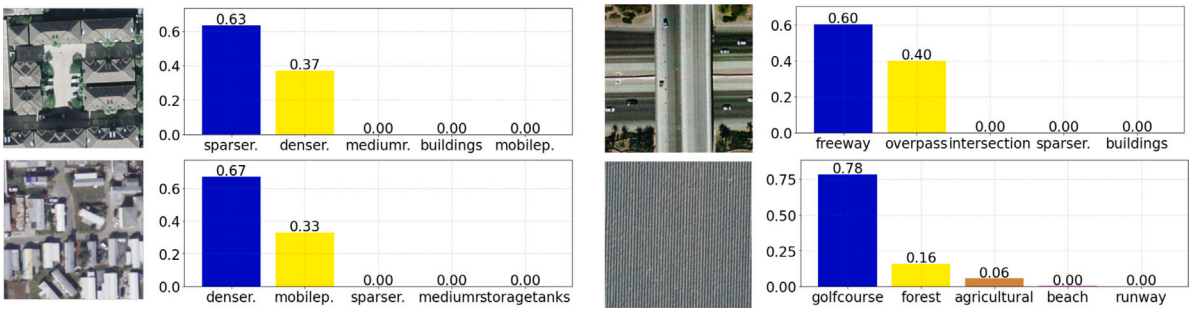


Fig. 5. Top-5 classification results on the UCM dataset. The left two figures show examples from the building category, the top right figure shows an example from the overpass category, and the bottom right figure shows an example from the agriculture category. Numbers denote the predicted probabilities. 'sparer.', 'mediumr.', 'denser.' and 'mobilep.' denote sparse residential, medium residential, dense residential, and mobile homepark respectively.

to ground truth (e.g., agriculture is misclassified as forest in the third row), and ground truth categories are included in the top-5 predictions. This is the reason why the CLIP model shows promising top-5 classification performance. Moreover, the two prompts, "This is a satellite image of a [CLASS]" and "This is an aerial image of a [CLASS]", generally perform better than other prompts. Furthermore, "This is an aerial image of a [CLASS]" shows better performance on top-5 accuracy. We will use this prompt as the default setting in the following sections.

We further explored zero-shot classification performance under different visual encoder backbones. We tried ResNet-50, ResNet-101, ViT-B/32, and ViT-L/14, keeping all other settings the same. As shown in Table 3, with the increase of model capacities of visual backbones, the performance of the CLIP model increases consistently. The CLIP model with the backbone network ViT-L/14 yields the best performance. In the following sections, we use ViT-L/14 as the vision backbone. Note that using large backbone networks will hurt the inference efficiency. Thus, we did not try larger backbones, given the increased computational burden.

4.4. Curriculum learning results

Before discussing the zero-shot remote sensing scene classification results, we analyzed our method's performance at various stages of curriculum learning. We first investigated the impact of the number of pseudo samples at each curriculum learning stage. We experimented with varying numbers of pseudo samples during different stages. For UCM, NWPU-RESISC45, and AID datasets, we tested with 10/20/30 pseudo samples in the first stage. In the second stage, we tried 30/40/50 pseudo samples; in the third stage, we explored 50/60/70 pseudo samples. Note that we use a smaller number of pseudo samples for the WHU-RS19 dataset. We test with 5/10/20 samples in the first stage, 15/20/25 in the second stage, and 30/35/40 in the third stage. Table 4 presents the zero-shot scene classification results of these experiments. The results indicate that our model benefits from curriculum learning, with an increase in pseudo labels leading to improved performance. We also included the accuracy of the selected pseudo labels. It becomes evident that selecting an appropriate number of pseudo samples is crucial for achieving high pseudo accuracy and improving zero-shot test accuracy.

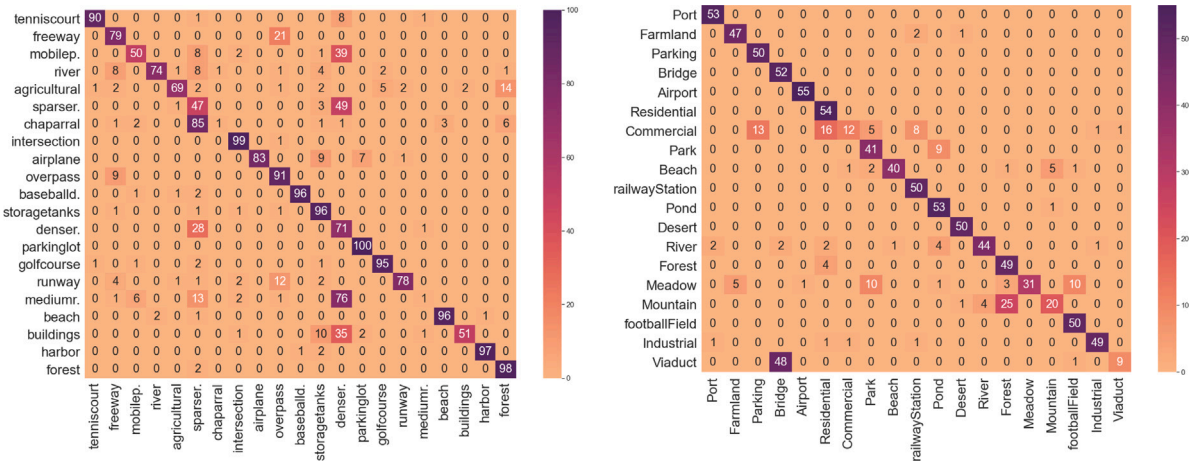


Fig. 6. Classification confusion matrix on the UCM (left) and WHU-RS19 datasets (right).

Table 2
Scene classification results using different prompts.

Prompt	UCM		WHU-RS19		NWPU-RESISC45		AID	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
This is a photo of a {CLASS}.	74.67	93.67	77.91	93.23	63.84	90.06	66.70	95.63
This is a satellite image of a {CLASS}.	74.57	**97.71**	**81.29**	98.31	66.21	93.18	**70.51**	96.92
This is a land use image of a {CLASS}.	**78.00**	96.38	75.82	92.94	**71.35**	93.47	67.51	95.28
This is a remote sensing image of a {CLASS}.	75.19	96.47	82.69	98.61	68.53	93.80	66.80	94.37
This is an aerial image of a {CLASS}.	74.28	95.91	80.30	**98.71**	68.84	**94.66**	68.87	**96.97**

Table 3
Scene classification results using different visual backbone networks.

Model	UCM		Time (s)	# Params
	top-1	top-5		
ResNet-50	52.05	87.71	231.2	**102M**
ResNet-101	54.57	87.43	353.7	119M
ViT-B/32	59.59	86.88	**167.8**	151M
ViT-L/14	**74.28**	**95.91**	2963.9	437M

Based on our observations, we determined the optimal number of pseudo samples for each curriculum learning stage for the UCM, NWPU-RESISC45, and AID datasets. Specifically, we selected 20, 40, and 50 pseudo samples for the respective stages in these datasets. Conversely, for datasets with relatively fewer images per category, i.e., the WHU-RS19 dataset, we chose a different set of pseudo samples for each curriculum learning stage, namely 10, 20, and 35. By carefully selecting the appropriate number of pseudo samples, we can ensure the effectiveness and generalizability of our curriculum learning approach for different datasets with varying characteristics.

From Table 4, our model benefits from curriculum learning with increasing pseudo labels. The original CLIP model obtains a classification accuracy of 74.38% and 80.30% on the UCM and WHU-RS19 dataset, respectively; in contrast, by using three stages of curriculum learning, our method achieves a classification accuracy of 86.71% (resp. 99.10%) on the UCM (resp. WHU-RS19) dataset respectively.

Fig. 7 shows the output features extracted by the visual encoder network from the UCM dataset at different curriculum learning stages. In Fig. 7, stage-0 means the original pretrained CLIP model, and stage-3 means our model finetuned after three rounds of curriculum learning using pseudo labels. As shown in Fig. 7, at the later stage of curriculum learning, in which more pseudo labels are used, the feature separability of our method becomes better. After three rounds of curriculum learning, only a few samples lie in the intersection of classification boundaries. This is also demonstrated by the quantitative results in Table 5.

4.5. Zero-shot results

Previous zero-shot remote sensing scene classification methods focus on a transferring setting, in which sufficient labels are provided for selected seen classes, and the models are learned to classify images of unseen categories. To enable a direct comparison with previous methods, we follow (Li et al., 2022) to divide the dataset into seen and unseen classes and provide zero-shot scene classification performance on novel classes only. We randomly divided seen/unseen classes 25 times and reported the average zero-shot classification performance. Moreover, we also report the performance in a more challenging setting where no labels are provided for all classes of target datasets.

Table 5 presents the zero-shot remote sensing scene classification performance across four benchmark datasets. Notably, our proposed RS-CLIP method significantly outperforms previous state-of-the-art approaches. For instance, the previous SOTA method CSPW-GAN (Wang et al., 2021) achieved a classification accuracy of 62.66% (resp. 55.86%) on the UCM (resp. AID) dataset, where labels were available for 16 (resp. 25) seen classes, and the model was tested on five unseen classes. In contrast, our RS-CLIP method achieves remarkable results with a classification accuracy of 95.54% and 93.34% on the UCM and AID datasets, respectively.

Additionally, we report the performance under the generalized zero-shot setting, where the model is evaluated on all classes of the target dataset. In this more challenging scenario, our RS-CLIP method achieves an impressive classification accuracy of 86.71% (resp. 79.56%) on all UCM (resp. AID) dataset classes. These results demonstrate the superior capabilities of our RS-CLIP method, showcasing its effectiveness and robustness in zero-shot remote sensing scene classification tasks.

Fig. 8 displays several examples of classification results on the UCM and WHU-RS19 datasets. As shown in this figure, our model can successfully identify the categories of a majority of images. Classification errors, as shown in the last column of each figure, mostly come from semantic-related categories. For example, our model misclassified a freeway scene as an overpass, as shown in the first row of Fig. 8(a).

Table 4
Effect of the number of pseudo samples at different curriculum learning stages. Zero-shot scene classification performance on the UCM and WHURS datasets. We report both pseudo accuracy and zero-shot test accuracy. The boldface indicates the best performance.

UCM				WHU-RS19			
# Samples	Pseudo Acc. (%)	Test Acc. (%)		# Samples	Pseudo Acc. (%)	Test Acc. (%)	
		top-1	top-5			top-1	top-5
0(baseline)	–	74.38	95.91	0(baseline)	–	80.30	98.71
10	85.71	80.48	**99.33**	5	94.74	95.42	99.60
20	85.48	**82.71**	98.57	10	96.32	**96.92**	99.60
30	83.81	81.52	98.62	20	95.00	95.32	99.90
30	85.24	84.57	98.71	15	100.00	98.10	99.88
40	86.79	**85.76**	98.14	20	99.74	**98.21**	99.90
50	87.24	85.29	97.71	25	99.37	**98.21**	99.70
50	89.05	**86.71**	97.57	30	99.47	98.81	**100.00**
60	88.81	86.29	98.24	35	99.25	**99.10**	99.60
70	87.89	86.10	97.52	40	99.21	99.00	99.60

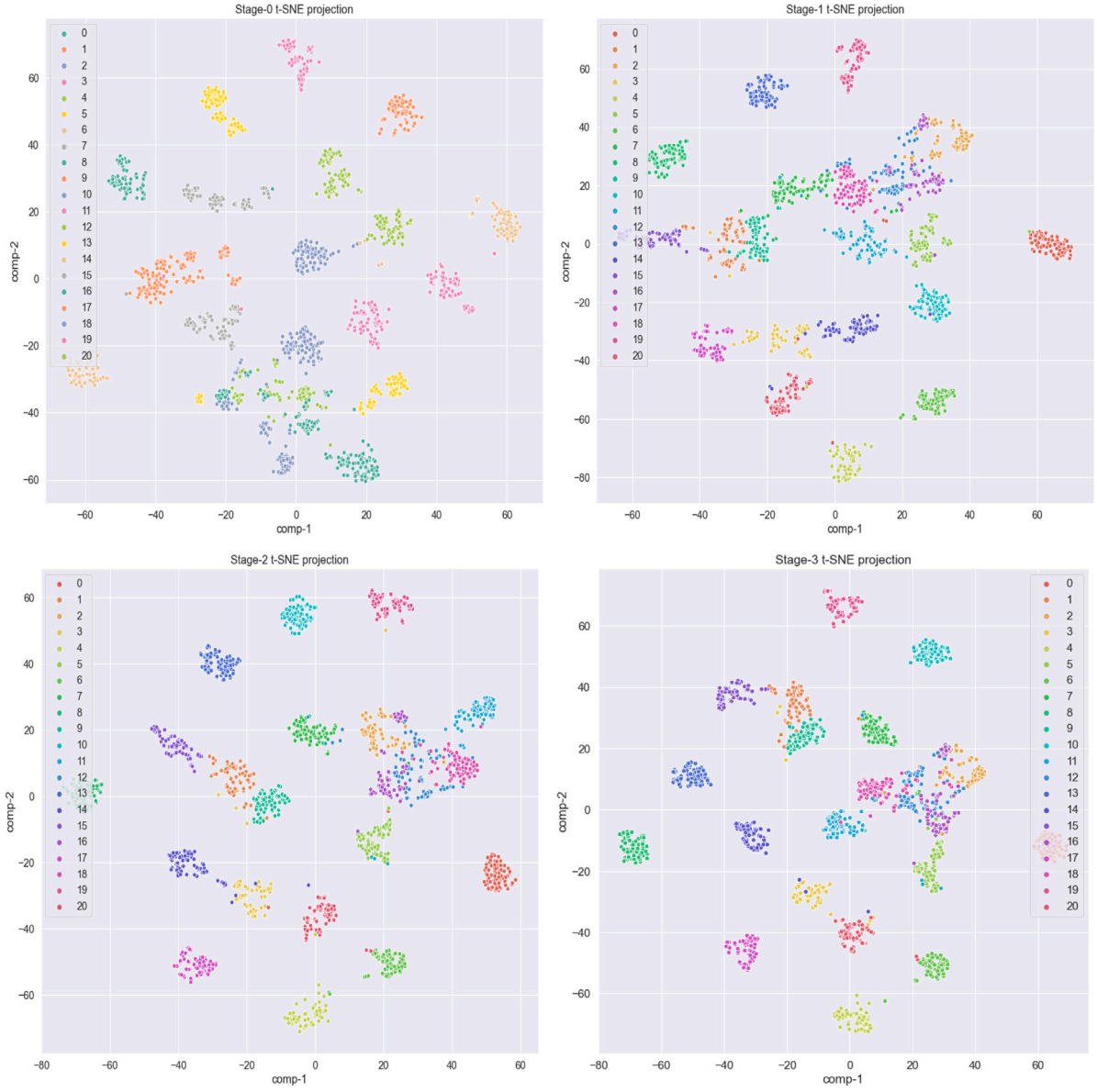


Fig. 7. t-SNE visualization of features extracted by the visual encoder network from the UCM dataset at different curriculum learning stages.

Table 5
Zero-shot scene classification results on four benchmark datasets. We report both top-1 and top-5 classification accuracy. Numbers in the bracket show the number of novel classes and all classes. In the top part, we report the classification accuracy on the novel classes, all values are borrowed from Li et al. (2022); in the middle part, we report the classification accuracy of our method on novel classes; in the bottom part, we report the classification accuracy of our method on all classes. Boldface values indicate the best performance.

Method	UCM (5/21)		WHU-RS19 (5/19)		NWPU-RESISC45 (10/45)		AID (5/30)	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
SAE (Kodirov et al., 2017)	49.50	–	–	–	44.81	–	47.34	–
ZSP-LP (Li et al., 2017)	49.01	–	–	–	47.00	–	46.77	–
ZSC-SA (Quan et al., 2018)	50.42	–	–	–	48.40	–	50.87	–
WDVSc (Wan et al., 2019)	55.91	–	–	–	50.68	–	52.61	–
RBGN (Xing et al., 2020)	57.93	–	–	–	44.60	–	51.99	–
DASE (Wang et al., 2021)	58.63	–	–	–	51.52	–	53.49	–
CSPWGAN (Li et al., 2022)	62.66	–	–	–	51.52	–	55.86	–
Ours (0-iter)	89.11	**99.92**	95.97	**100.00**	85.76	99.10	87.52	**100.00**
Ours (1-iter)	93.86	99.91	99.09	**100.00**	92.43	99.60	91.65	**100.00**
Ours (2-iter)	95.20	99.90	99.46	**100.00**	94.02	99.63	92.57	**100.00**
Ours (3-iter)	**95.54**	99.89	**99.49**	**100.00**	**96.95**	**100.00**	**93.34**	**100.00**
Ours (0-iter)	74.38	95.91	80.30	98.71	66.86	93.44	65.48	89.41
Ours (1-iter)	82.71	**98.57**	96.92	**99.60**	82.94	74.17	75.75	**93.38**
Ours (2-iter)	85.76	98.14	98.21	**99.90**	**85.44**	77.83	78.36	93.00
Ours (3-iter)	**86.71**	97.57	**99.10**	**99.60**	85.07	**79.11**	**79.56**	92.52

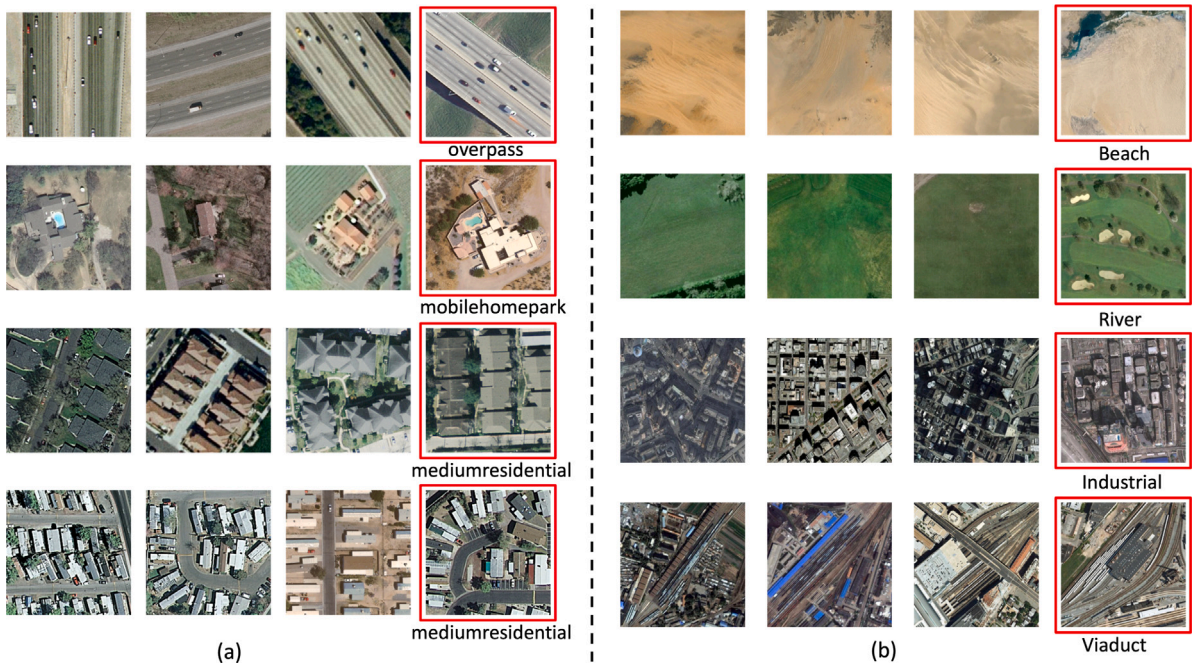


Fig. 8. Zero-shot classification results on the UCM (left) and WHU-RS19 (right) datasets. For the UCM dataset, we show results on freeway, spare residential, dense residential, and mobile homepark categories from top to bottom. For the WHS-RS19 dataset, we show results on desert, meadow, commercial, and railway station categories from top to bottom. In the last column of each figure, we show examples of false predictions, where texts below the images show predicted categories.

Other misclassification also comes from inter-class visual similarities. For example, in the second row of Fig. 8(b), our model classified the meadow scene as a river since it contains channel-like textures.

4.6. Few-shot results

We further investigate the performance of our model on few-shot remote sensing scene classification. Previous few-shot RSSC methods focus on a transferring setting where sufficient labels are provided for some seen classes, and the models are learned to perform classification on unseen categories with only a few labels. We follow (Ji et al., 2022b,a) to select unseen classes for evaluation. For our RS-CLIP model, we reloaded the model after the first round of curriculum learning and selected K_f samples from each class to finetune our model using the ground truth labels. Following previous few-shot RSSC methods (Li et al., 2020b,a; Ji et al., 2022a), we set K_f to 5 in our experiments. We name the model Ours+ft. To further show the effect of our newly designed curriculum learning technique, we finetuned the original CLIP model using the same ground truth labels. The model is denoted as CLIP+ft.

Table 6 shows the few-shot remote sensing scene classification performance on four benchmark datasets. We report the classification accuracy of the novel classes following previous methods. The results of our proposed method are presented in the middle and bottom parts of Table 6. As shown in Table 6, our proposed method performs significantly better when evaluated on novel classes than the previous state-of-the-art methods. Specifically, the previous SOTA method (Ji et al., 2022a) obtained a classification accuracy of 73.42% on the UCM dataset, where sufficient labels were provided in 16 seen classes, and the model was evaluated on 6 novel classes with 5-shot labels provided for each novel class. In contrast, with only 5-shot labels for each novel class, our method achieves a classification accuracy of 97.83%, significantly better than (Ji et al., 2022a).

Table 6
Few-shot scene classification results on different datasets. Numbers in the bracket show the number of novel classes and all classes. In the top and middle parts, we show classification accuracy on novel classes; while in the bottom part, we show classification accuracy on all classes. Boldface values indicate the best performance and underline values indicate the previous best performance.

Method	UCM (6/21)	WHU-RS19 (5/19)	NWPU-RESISC45 (10/45)	AID (7/30)
DLA-MatchNet (Li et al., 2020b)	63.01	79.89	81.63	73.45
RS-MetaNet (Li et al., 2020a)	67.63	87.45	79.62	73.76
SGMNet (Zhang et al., 2022a)	73.42	90.12	82.32	75.68
Ji et al. (2022b)	–	94.24	89.20	87.31
Ji et al. (2022a)	–	92.96	89.87	87.33
CLIP	89.17	92.63	80.13	76.61
CLIP+ft.	97.33	99.29	89.50	86.57
Ours+ft.	**97.83**	**100.00**	**93.96**	**94.22**
CLIP	74.38	80.30	66.85	70.67
CLIP+ft.	88.57	96.17	78.45	84.17
Ours+ft.	**92.86**	**97.41**	**85.42**	**89.22**

Additionally, we report the performance under the generalized few-shot setting, where the model is evaluated on all classes of the target dataset. Results are presented in the lower section of Table 6. It is important to note that classifying all classes is more difficult than only classifying novel classes. Our method performs even better in all classes than previous SOTA methods reported for novel classes on the UCM, WHU-RS19, and AID datasets.

4.7. Results on the SEN12MS dataset

To further demonstrate the performance of our model, we conducted experiments on the SEN12MS dataset (Schmitt et al., 2019). In the upper section of Table 7, we present the zero-shot classification performance, revealing our model's generalization capabilities when applied to Sentinel-2 images. Our model attains a top-1 accuracy exceeding 40% for the summer subset and surpassing 28% for the entire test dataset. Notably, the performance of our RS-CLIP model on the SEN12MS dataset falls behind the performance on preceding aerial image datasets. This discrepancy can be attributed to the broader domain gap between the SEN12MS and WebImageText datasets, thus rendering the pseudo-labeling process considerably more arduous. This is substantiated by considerably inferior top-1 and top-5 performance prior to curriculum learning.

To remedy this issue, we conducted experiments under few-shot settings. After the first round of curriculum learning, we reloaded the weights and finetuned our model with 5/10 samples per class, guided by ground truth labels. For comparison, we constructed a supervised baseline in which the model was trained on all training samples using a ResNet50 (He et al., 2016b) architecture. We also explored ResNet50 as the visual encoder network in our RS-CLIP model to ensure a fair comparison. The results are presented in the lower section of Table 7. As shown in the table, our RS-CLIP model achieved a substantial performance improvement through fine-tuning, even with access to only a limited quantity of labels. When using ResNet50 as the visual encoder, our model notably outperformed the fully-supervised baseline for the summer subset. With the ViT-L visual encoder, our RS-CLIP model achieved comparable performance to the fully supervised ResNet50 model on the entire test set.

Table 7
Zero-/few-shot classification on the SEN12MS dataset. We show performance on the SEN12MS dataset summer split on the left and all seasons on the right. For the ResNet50 baseline, we use the official evaluation code from https://github.com/schmitt-muc/SEN12MS.

	Summer		All	
	Top-1	Top-5	Top-1	Top-5
Ours (0-iter)	17.80	49.99	20.30	63.72
Ours (1-iter)	36.50	79.55	**28.11**	**75.95**
Ours (2-iter)	37.29	84.21	26.79	74.67
Ours (3-iter)	**41.14**	**88.18**	24.20	74.67
ResNet50	69.66	**99.141**	**58.98**	**98.86**
Ours w/ ResNet50 (5-shot)	73.84	94.35	29.57	78.62
Ours w/ ResNet50 (10-shot)	**83.31**	98.37	42.91	78.34
Ours w/ ViT-L (5-shot)	76.73	97.51	42.27	**94.11**
Ours w/ ViT-L (10-shot)	**82.07**	**99.56**	**55.57**	92.34

5. Conclusion and limitation

In this research, we present a novel vision-language model for remote sensing scene classification, leveraging the pretrained CLIP (Radford et al., 2021) baseline. Unlike conventional zero-shot RSSC methods that typically employ separate visual and semantic encoders, our approach achieves semantic-aware visual representations through joint visual–semantic feature learning. Our vision-language model is pretrained on extensive image–text datasets, providing robust general knowledge and demonstrating strong transfer capabilities to remote sensing scenes. To adapt the model to the remote sensing domain, we propose a pseudo-labeling technique that automatically generates pseudo labels for unlabeled datasets. This enables effective model fine-tuning on the target domain. Additionally, we introduce a curriculum learning strategy involving multiple stages of model fine-tuning, which significantly enhances the performance of zero-shot remote sensing scene classification. We thoroughly evaluate our approach on four benchmark datasets, and the experimental results demonstrate substantial performance improvements for both zero-shot and few-shot remote sensing scene classification scenarios.

In our research, we employed the CLIP model pretrained on the WebImageText dataset as our baseline model. However, this choice introduces two inherent limitations: Firstly, the WebImageText dataset is not publicly accessible, leading to uncertainty regarding the specific scene classes it encompasses. This lack of transparency raises concerns about fair evaluation in zero-shot scenarios. To mitigate this issue, excluding scene classes already present in the WebImageText dataset would be advisable during zero-shot evaluation. Secondly, it is essential to acknowledge that the WebImageText dataset was not tailored explicitly for the remote sensing domain. Consequently, significant domain discrepancies may arise when transferring the model to remote sensing data, posing challenges during the process of selecting initial pseudo labels. This is revealed by the performance on the SEN12MS dataset. To address these limitations and enhance the robustness of our approach, we propose future work that involves the development of a custom vision language foundation model using remote sensing data. This forthcoming endeavor entails the curation of a comprehensive benchmark dataset comprising large-scale image–text pairs. Through training this novel vision language foundation model, we aim to achieve enhanced generalization capabilities across a diverse set of remote sensing tasks. By doing so, we aspire to overcome these limitations and advance the effectiveness and generalization capabilities of our method.

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.

Data availability

Data will be made available on request.

Acknowledgment

The project is supported by the Guangdong Science and Technology Strategic Innovation Fund (the Guangdong–Hong Kong–Macau Joint Laboratory Program), China, Project No.: 2020B1212030009.

References

Bazi, Y., Al Rahhal, M.M., Mekhalfi, M.L., Al Zuair, M.A., Melgani, F., 2022. Bi-modal transformer-based approach for visual question answering in remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 60, 1–11.

Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C., 2019. ReMixMatch: Semi-supervised learning with distribution matching and augmentation anchoring. In: ICLR.

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

Bucher, M., Herbin, S., Jurie, F., 2016. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In: ECCV. Springer, pp. 730–746.

Cascante-Bonilla, P., Tan, F., Qi, Y., Ordonez, V., 2021. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In: AAAI, Vol. 35, No. 8. pp. 6912–6920.

Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., Shi, Z., 2023. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. arXiv preprint arXiv:2306.16269.

Chen, F., Tsou, J.Y., 2021. DRSNet: Novel architecture for small patch and low-resolution remote sensing image scene classification. JAG 104, 102577.

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

Dai, D., Yang, W., 2011. Satellite image classification via two-layer sparse coding with biased image representation. IEEE Trans. Geosci. Remote Sens. 8 (1), 173–176.

Demirel, B., Gokberk Cinbis, R., Ikizler-Cinbis, N., 2017. Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning. In: ICCV. pp. 1232–1241.

Ding, Z., Shao, M., Fu, Y., 2017. Low-rank embedded ensemble semantic dictionary for zero-shot learning. In: CVPR. pp. 2050–2058.

Djoufack Basso, L., 2022. CLIP-RS: A Cross-modal Remote Sensing Image Retrieval Based on CLIP (Ph.D. thesis), a Northern Virginia Case Study. Virginia Tech.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale. In: ICLR. URL https://openreview.net/forum?id=YicbFdNTTy.

Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G., 2022. Learning to prompt for open-vocabulary object detection with vision-language model. In: CVPR. pp. 14084–14093.

Elhoseiny, M., Saleh, B., Elgammal, A., 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In: ICCV. pp. 2584–2591.

Fu, Z., Xiang, T., Kodirov, E., Gong, S., 2017. Zero-shot learning on semantic class prototype graph. TPAMI 40 (8), 2009–2022.

Gong, C., Tao, D., Maybank, S.J., Liu, W., Kang, G., Yang, J., 2016. Multi-modal curriculum learning for semi-supervised image classification. IEEE Trans. Image Process. 25 (7), 3249–3260.

Gu, X., Lin, T.-Y., Kuo, W., Cui, Y., 2021. Open-vocabulary object detection via vision and language knowledge distillation. In: ICLR.

Guo, J., Guo, S., 2020. A novel perspective to zero-shot learning: Towards an alignment of manifold structures via semantic feature expansion. IEEE Trans. Multimed. 23, 524–537.

Han, Y., Liu, Y., Jin, Z., 2020. Sentiment analysis via semi-supervised learning: a model based on dynamic threshold and multi-classifiers. Neural Comput. Appl. 32, 5117–5129.

He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition. In: CVPR. pp. 770–778.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Deep residual learning for image recognition. In: CVPR. pp. 770–778.

Hu, Y., Yuan, J., Wen, C., Lu, X., Li, X., 2023. RSGPT: A remote sensing vision language model and benchmark. arXiv preprint arXiv:2307.15266.

Huang, T., Chu, J., Wei, F., 2022. Unsupervised prompt learning for vision-language models. arXiv preprint arXiv:2204.03649.

Ji, H., Gao, Z., Zhang, Y., Wan, Y., Li, C., Mei, T., 2022a. Few-shot scene classification of optical remote sensing images leveraging calibrated pretext tasks. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Ji, H., Yang, H., Gao, Z., Li, C., Wan, Y., Cui, J., 2022b. Few-shot scene classification using auxiliary objectives and transductive inference. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. PMLR, pp. 4904–4916.

Jin, J., Zhou, W., Ye, L., Lei, J., Yu, L., Qian, X., Luo, T., 2022. DASFNet: Dense-attention–similarity-fusion network for scene classification of dual-modal remote-sensing images. JAG 115, 103087.

Joulin, A., Grave, É., Bojanowski, P., Mikolov, T., 2017. Bag of tricks for efficient text classification. In: EACL. pp. 427–431.

Kenton, J.D.M.-W.C., Toutanova, L.K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186.

Kim, M., Kim, J., Bento, J., Song, G., 2023. Revisiting self-training with regularized pseudo-labeling for tabular data. arXiv preprint arXiv:2302.14013.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al., 2023. Segment anything. arXiv preprint arXiv:2304.02643.

Kodirov, E., Xiang, T., Gong, S., 2017. Semantic autoencoder for zero-shot learning. In: CVPR. pp. 3174–3183.

Lampert, C.H., Nickisch, H., Harmeling, S., 2009. Learning to detect unseen object classes by between-class attribute transfer. In: CVPR. IEEE, pp. 951–958.

Lee, D.-H., et al., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, Vol. 3, No. 2. ICML, p. 896.

Lei Ba, J., Swersky, K., Fidler, S., et al., 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. In: ICCV. pp. 4247–4255.

Li, H., Cui, Z., Zhu, Z., Chen, L., Zhu, J., Huang, H., Tao, C., 2020a. RS-MetaNet: Deep metametric learning for few-shot remote sensing scene classification. IEEE Trans. Geosci. Remote Sens. 59 (8), 6983–6994.

Li, L., Han, J., Yao, X., Cheng, G., Guo, L., 2020b. DLA-MatchNet for few-shot remote sensing image scene classification. IEEE Trans. Geosci. Remote Sens. 59 (9), 7844–7853.

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

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., 2021. Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification. IEEE Trans. Geosci. Remote Sens. 59 (12), 10590–10603.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022.

Ma, Y., Cambria, E., Gao, S., 2016. Label embedding for zero-shot fine-grained named entity typing. In: COLING. pp. 171–180.

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

Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G.S., Dean, J., 2013. Zero-shot learning by convex combination of semantic embeddings. arXiv preprint arXiv:1312.5650.

Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M., 2009. Zero-shot learning with semantic output codes. NeurIPS 22.

Pan, C., Huang, J., Hao, J., Gong, J., 2020. Towards zero-shot learning generalization via a cosine distance loss. Neurocomputing 381, 167–176.

Qiu, C., Yu, A., Yi, X., Guan, N., Shi, D., Tong, X., 2022. Open self-supervised features for remote-sensing image scene classification using very few samples. IEEE Geosci. Remote Sens. Lett. 20, 1–5.

Quan, J., Wu, C., Wang, H., Wang, Z., 2018. Structural alignment based zero-shot classification for remote sensing scenes. In: 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.

Romera-Paredes, B., Torr, P., 2015. An embarrassingly simple approach to zero-shot learning. In: ICML. PMLR, pp. 2152–2161.

Rosenberg, C., Hebert, M., Schneiderman, H., 2005. Semi-supervised self-training of object detection models. In: WACV, Vol. 1. IEEE, pp. 29–36.

Sammon, J.W., 1969. A nonlinear mapping for data structure analysis. IEEE Trans. Comput. 100 (5), 401–409.

Schmitt, M., Hughes, L.H., Qiu, C., Zhu, X.X., 2019. SEN12MS – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. In: ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci., Vol. IV-2/W7. pp. 153–160. http://dx.doi.org/10.5194/isprs-annals-IV-2-W7-153-2019.

Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M., Matsumoto, Y., 2015. Ridge regression, hubness, and zero-shot learning. In: ECML PKDD. Springer, pp. 135–151.

Socher, R., Ganjoo, M., Manning, C.D., Ng, A., 2013. Zero-shot learning through cross-modal transfer. NeurIPS 26.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.-L., 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. NeurIPS 33, 596–608.

Song, X., Zeng, H., Zhang, S., Herranz, L., Jiang, S., 2020. Generalized zero-shot learning with multi-source semantic embeddings for scene recognition. In: ACMMM. pp. 3976–3985.

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.

Sun, X., Wang, P., Lu, W., Zhu, Z., Lu, X., He, Q., Li, J., Rong, X., Yang, Z., Chang, H., et al., 2022. RingMo: A remote sensing foundation model with masked image modeling. IEEE Trans. Geosci. Remote Sens.

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

Wan, Z., Chen, D., Li, Y., Yan, X., Zhang, J., Yu, Y., Liao, J., 2019. Transductive zero-shot learning with visual structure constraint. NeurIPS 32.

Wang, D., Li, Y., Lin, Y., Zhuang, Y., 2016. Relational knowledge transfer for zero-shot learning. In: AAAI, Vol. 30, No. 1.

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.

Wang, W., Zheng, V.W., Yu, H., Miao, C., 2019. A survey of zero-shot learning: Settings, methods, and applications. TIST 10 (2), 1–37.

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., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B., 2016. Latent embeddings for zero-shot classification. In: CVPR. pp. 69–77.

Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q., 2020. Unsupervised data augmentation for consistency training. NeurIPS 33, 6256–6268.

Xing, Y., Huang, S., Huangfu, L., Chen, F., Ge, Y., 2020. Robust bidirectional generative network for generalized zero-shot learning. In: ICME. IEEE, pp. 1–6.

Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X., 2022. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: ECCV. Springer, pp. 736–753.

Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. In: SIGSPATIAL. pp. 270–279.

Yu, Q., Ikami, D., Irie, G., Aizawa, K., 2020. Multi-task curriculum framework for open-set semi-supervised learning. In: ECCV. Springer, pp. 438–454.

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al., 2021. Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432.

Zhang, B., Feng, S., Li, X., Ye, Y., Ye, R., Luo, C., Jiang, H., 2022a. SGMNet: Scene graph matching network for few-shot remote sensing scene classification. IEEE Trans. Geosci. Remote Sens. 60, 1–15.

Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H., 2022b. Pointclip: Point cloud understanding by clip. In: CVPR. pp. 8552–8562.

Zhang, L., Xiang, T., Gong, S., 2017. Learning a deep embedding model for zero-shot learning. In: CVPR. pp. 2021–2030.

Zheng, Z., Yang, Y., 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. Int. J. Comput. Vis. 129 (4), 1106–1120.

Zhu, J., Yang, K., Guan, N., Yi, X., Qiu, C., 2023. HCPNet: Learning discriminative prototypes for few-shot remote sensing image scene classification. JAG 123, 103447.