Learning Positive–Negative Prompts for Open-Set Remote Sensing Scene Classification

Hao Sun, *Member, IEEE*, Hanlizi Chen, Wenjing Chen, *Member, IEEE*, Chengji Wang, Wei Xie, and Xiaoqiang Lu, *Senior Member, IEEE*

***Abstract*—Most remote sensing scene classification (RSSC) methods are primarily built upon the closed-set assumption, which assumes that all test samples definitely belong to one of the categories seen during training. However, practical applications are usually open environments, where samples of other categories that have never been seen during training will appear, which is called open-set RSSC (OS-RSSC). These methods may mistakenly classify samples of unseen categories into those seen categories, resulting in a decrease in application potential. In this article, we propose a positive–negative prompt learning (PNPL) framework for OS-RSSC. PNPL aims to tune the powerful contrastive language-image pretraining (CLIP) model for OS-RSSC through learning positive and negative prompts. First, positive textual prompts and visual prompts are trained to provide the model with basic classification capabilities for known classes. Then, negative textual prompts are indirectly learned from positive prompts and images, enabling the model to capture the semantics of unknown classes. PNPL significantly increases the discriminative power between known and unknown classes, enhancing the model's ability to accurately distinguish them. Extensive experiments on three RSSI datasets have shown that PNPL outperforms compared methods. Code is available at https://github.com/ChenHanlizi/PNPL.**

***Index Terms*—Open-set classification (OSC), prompt learning, remote sensing imagery, scene classification.**

I. Introduction

RAPID advancement of remote sensing satellite technology has facilitated the acquisition of unprecedented volumes of remote sensing scene images that encapsulate diverse land covers and intricate spatial details. Remote sensing scene classification (RSSC) refers to classifying scene images into several meaningful land use and land cover classes by analyzing image contents [1]. It plays a crucial role in various practical applications such as natural disaster detection and urban morphology analysis [2].

Most RSSC methods are usually based on deep learning frameworks to design various multiscale networks and attention modules to address challenges such as large variations in target scales and complex backgrounds [3], [4]. These methods focus on extracting and fusing global and local features while suppressing irrelevant information in scene images, making great advances. However, these methods adhere to the closed-set paradigm, which assumes that the label space of training samples and the label space of test samples are strictly consistent [5]. This inherent limitation of the closed-set paradigm becomes particularly problematic when confronting open-set scenarios, where novel scene categories inevitably emerge due to dynamic environmental transformations in remote sensing scenes.

When a test sample belongs to a category not represented in the training data, closed-set RSSC methods will assign it into an existing category in the training set [6], [7]. The misidentification has a huge impact on subsequent researches [8]. Such task where new categories may emerge during testing is called the open-set RSSC (OS-RSSC) that aims to identify unknown categories and classify known categories during testing [9], [10]. Unlike few-shot learning methods that require at least one unknown class sample to construct prototypes [11], [12], OS-RSSC does not rely on any unknown class data, instead leveraging only data from known classes to construct decision boundaries for distinguishing between known and unknown categories. Recently, some works have attempted to using techniques such as graph convolutional networks, energy-based models, and adversarial learning to extract discriminative features of known classes for OS-RSSC [13], [14], [15]. There are also some works designing unique threshold selection methods by analyzing the distribution characteristics of output data to distinguish unknown classes [16], [17]. However, these methods only convert the category name texts carried by training data into discrete one-hot vectors, leaving the semantics encapsulated in texts largely unexploited [18].

The emergence and application of vision–language models (VLMs) have demonstrated the effectiveness of visual–textual interaction in improving image classification [19]. Among these, contrastive language–image pretraining (CLIP) [20] has emerged as a pioneering and widely adopted VLM that

Received 9 March 2025; revised 5 June 2025 and 14 July 2025; accepted 11 August 2025. Date of publication 15 August 2025; date of current version 26 August 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62201222 and Grant 62377026, in part by the Fundamental Research Funds for the Central Universities under Grant CCNU25ai041, and in part by Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning under Grant 2025AISL011. *(Corresponding author: Wei Xie.)*

Hao Sun, Hanlizi Chen, Chengji Wang, and Wei Xie are with Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, the School of Computer Science, and the National Language Resources Monitoring and Research Center for Network Media, Central China Normal University, Wuhan 430079, China (e-mail: xw@mail.ccnu.edu.cn).

Wenjing Chen is with the School of Computer Science, Hubei University of Technology, Wuhan 430068, China.

Xiaoqiang Lu is with the College of Physics and Information Engineering, Fuzhou University, Fuzhou 350002, China.

Digital Object Identifier 10.1109/TGRS.2025.3599188

1558-0644 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Central China Normal University. Downloaded on January 28,2026 at 12:43:44 UTC from IEEE Xplore. Restrictions apply.

establishes a joint embedding space for visual and textual modalities. By training on 400 million image–text pairs through contrastive learning, CLIP learns to associate relevant images with their textual descriptions, enabling remarkable zero-shot transfer capabilities across diverse visual tasks. Some methods employ CLIP and design templates such as “a remote sensing image of [class].” or generate specific descriptions as hard prompts for text input for OS-RSSC [21]. Zhou et al. [18] and Miyai et al. [22] utilize prompt learning to obtain suitable learnable vectors as soft prompts that enables the recognition of unknown classes. Soft prompts trained on known-class images convey positive semantics similar to “a photo of a [class]” [18]. These methods only train prompts for known classes, and naturally classify samples with low relevance to prompts into unknown classes by maximizing the cosine similarity between known class image features and prompt features [23]. However, they completely ignore the relevant information of unknown classes, which hinders the potential of VLMs.

In this article, we propose a positive–negative prompt learning (PNPL) framework to harness powerful representation capability of CLIP for OS-RSSC. Unlike constructing complex networks, PNPL incorporates trainable visual prompts and positive–negative textual prompts into CLIP to enhance discrimination between known and unknown scene categories. PNPL trains both visual prompts and positive textual prompts for known classes, while simultaneously addressing unknown classes through training negative textual prompts with negative semantics. Similar to positive textual prompts, negative textual prompts are also learnable continuous vectors. By designing appropriate loss functions, we indirectly imbue these negative textual prompts with semantics that represent unknown classes, thereby increasing their cosine similarity with the features of images belonging to unknown classes.

The training process of PNPL includes two stages. In the first stage, PNPL mainly trains visual prompts and positive textual prompts for recognizing known scene categories by maximizing the cosine similarity between the corresponding image features and text features of known classes. Additionally, a contrastive loss function is designed for images to expand interclass differences and reduce intraclass variations. In the second stage, negative textual prompts are introduced to enhance the ability of PNPL to distinguish unknown classes. To preserve the classification accuracy for known classes, the visual prompts and positive textual prompts trained in the first stage are frozen. Due to the inability to access any information from unknown classes, we design two loss functions to indirectly train negative prompts. First, to avoid learning negative prompts unrelated to remote sensing images, we introduce a prompt constraint loss to ensure a certain similarity between negative prompt features and known-class image features. Second, to ensure that negative prompts represent unknown classes, they should exhibit different or even opposite semantics compared to positive prompts. We exploit a prompt separation loss to widen the difference between positive and negative prompts. Open-set classification (OSC) scores are then calculated by measuring the cosine similarity between the features of positive and negative textual prompts and the visual input features. Finally, a threshold is obtained through k-means algorithm to achieve recognition of unknown classes.

The contributions of this article mainly include.

1) Our proposed PNPL introduces negative prompts into CLIP, providing a new perspective to recognize known classes and enhancing model ability to identify unknown classes for OS-RSSC task.
2) Unlike many OS-RSSC methods that rely solely on image information, PNPL also incorporates textual information from known class names to further improve model performance.
3) Experiments on popular UCM, AID, and NWPU RSSI datasets demonstrate that PNPL outperforms several state-of-the-art methods.

The rest is organized as follows. Section II briefly introduces recent related methods. Section III details the proposed PNPL. Section IV describes experiments. Finally, the conclusion is provided in Section V.

II. Related Work

A. Remote Sensing Scene Image Classification

RSSC is an important component of remote sensing interpretation, and there has been extensive research on it. Initially, researchers use low-level descriptors such as local color histograms, color autocorrelograms, and local binary patterns to identify different remote sensing scenes [24], [25]. However, these shallow features that only focus on texture, color, shape, or combinations of these features perform poorly in complex scene classification.

With the advent of deep learning, researchers have employed deep neural networks to extract high-level features from RSSIs, leading to significant advancements in classification performance [26]. Nogueira et al. [27] used various convolutional neural networks to extract features from scene images, demonstrating the importance of extracting deep features in RSSC. Bi et al. [3] proposed all grains one scheme (AGOS), which extends the classic multiple instance learning into multigrain formulation, extract and fuse multigrain features. Ning et al. [28] proposed to learn robust features with scale-invariance and background-independent information through knowledge distillation.

In recent years, large-scale pretrained models have emerged, achieving remarkable performance across various tasks [29]. This success has motivated several studies to apply these models to the RSSC task, resulting in significant improvements in classification accuracy. Jiao et al. [30] confirmed that pretrained large-scale foundation model can provide excellent performance in remote sensing tasks. Al Rahhal et al. [24] conducted extensive experiments on VLMs based on CLIP with multiple different parameter complexities, and used six different predefined prompts to demonstrate the superiority of large-scale CLIP models. Liu et al. [31] proposed a text-guided remote sensing image pretrained model based on CLIP architecture for RSSC. During the pretraining process, five different text prompts were used for each image. In the inference phase, the text input “a remote sensing image of

Authorized licensed use limited to: Central China Normal University. Downloaded on January 28,2026 at 12:43:44 UTC from IEEE Xplore. Restrictions apply.

[class].” was employed, enabling the classification of scene images.

B. Open-Set Classification

OSC requires models to distinguish images that do not belong to any category in the training set during testing. Its main goal is to classify known categories and identify unknown categories. Methods for OSC can be broadly categorized into two types: discriminative methods and generative methods.

Discriminative models focus on learning representations of known classes. Al Rahhal et al. [15] proposed energy-based learning on vision transformers (EViT), which improves ViT via energy-based learning to jointly model class labels and data distributions in remote sensing images. It classifies low log-likelihood samples as unknown, thereby achieving OS-RSSC. Yang et al. [32] proposed convolutional prototype network (CPN), which utilizes a clustering algorithm to create a prototype set. The model then determines the category of samples by calculating the similarity between the sample and the prototypes. Xie et al. [33] proposed feature consistency-based prototype network (FCPN), which applied the prototype network method to the OSC task and achieved good performance. Chen and Wang [13] proposed a multiorder graph convolutional network (MGCN) framework for OS-RSSC, incorporating feature dispersion weighting and cross-domain adaptation mechanisms. Liu et al. [16] introduced an incremental learning with open-set recognition (ILOSR) framework, combining convex hull-based sample selection with a novel loss function integrating prototype learning and uncertainty measurement, significantly enhancing the accuracy of OS-RSSC. Zhang et al. [21] proposed a frequency distribution-based multimodal fine-tuning strategy (FreqDiMFT) that enhances few-shot open-set recognition in remote sensing through local–global frequency encoding and adaptive feature refinement. Chen et al. [34] proposed adversarial reciprocal point learning (ARPL), which selects the most distinct feature within each category as the reciprocal point. This method increases interclass differences by maximizing the distance between reciprocal points and samples, while constraining the points to reduce open space risks. Inliers and outliers match (IOMatch) [35] and adaptive negative evidential deep learning (ANEDL) [36] adopted the idea of one-vs-set method, and achieved the recognition of unknown classes in semi-supervised open-set tasks through the combination of multiple binary classifiers.

Generative model approaches focus on generating representations of known or unknown classes. Liu et al. [37] proposed an image-level reconstruction method called multitask deep learning for open world (MDL4OW). This method determines the category of samples by calculating the reconstruction loss between input images and their reconstructed counterparts. Sun et al. [38] proposed deep feature reconstruction learning (DFRL) to perform feature level reconstruction on input images and achieved OS-RSSC by determining the category of the image through reconstruction loss. Moon et al. [39] presented difficulty-aware simulator (DIAS) to simulate the open world by introducing copycat and GAN to produce easy-, moderate-, and hard-difficulty samples. Wu and Deng [40] proposed two-stream information bottleneck (TIB) to obtain simulated unknown features for training and realize the identification of unknown classes.

C. Prompt Learning

Prompt learning is a technique for fine-tuning VLMs with a small number of parameters for various downstream tasks [41], [42]. Based on the type of modality information, prompts can be classified into two categories: textual prompts and visual prompts [43].

Textual prompt represents specific semantic information and can be divided into two types: hard prompt and soft prompt. “A photo of a [class].” used in CLIP belongs to hard prompt. Although this type of prompt is intuitive, it demands considerable language expertise and manual design. Moreover, subtle changes in prompt syntax can significantly impact model performance, so it cannot be guaranteed that the design of hard prompt is the best choice. Context optimization (CoOp) [18] applied soft prompt to VLM, converting the contexts in prompt into learnable vectors. Compared with hard prompt, CoOp achieved excellent performance in downstream image classification tasks. Conditional CoOp (CoCoOp) [44] added a lightweight neural network to CoOp to generate for each image an input-conditional token, resulting in stronger domain generalization performance. NegPrompt [23] learned soft negative prompt during the training process, achieving recognition of unknown classes. Zhao et al. [45] introduced diverse text-prompt generation learning (DPL) for RSSC, where multiple independent text prompts are learned and optimized via a cosine-based diversity loss to maximize pairwise orthogonality, enhancing CLIP’s adaptability to RS datasets. Lin et al. [46] proposed FedRSCLIP, a federated learning framework that employs shared prompts and private prompts with alignment constraints for CLIP-based remote sensing image classification, effectively addressing data heterogeneity and large-scale model transmission challenges.

Similar to text prompts, visual prompts add learnable perturbations on images to help the model adapt to new tasks. Visual prompts can bridge the gap between pretrained targets and specific task requirements by adjusting the representation of visual features. Jia et al. [47] proposed visual prompt tuning (VPT), which adapts to classification tasks by adding learnable prompts to the inputs of either the first layer or all layers of frozen transformer. Zhu et al. [48] introduced meta visual prompt tuning (MVP) that adapts VPT through meta-learning while employing random patch recombination for data augmentation, ultimately optimizing performance for RSSC. Fang et al. [49] proposed instance-aware visual prompting (IVP), a parameter-efficient tuning method that adaptively generates instance-specific prompts for RSSC. Zhang et al. [50] proposed EarthMarker, which leverages box and point prompts with cross-domain learning to enable multigranularity image interpretation. Bahng et al. [51] added learnable prompts to the edges of images, directing the model’s attention to objects in the central region of the image. Tsai et al. [52] proposed convolutional visual prompt (CVP), which reduces the self-supervised loss encountered by the model when processing

Authorized licensed use limited to: Central China Normal University. Downloaded on January 28,2026 at 12:43:44 UTC from IEEE Xplore. Restrictions apply.

corrupted images. Xu et al. [53] proposed progressive visual prompt (ProVP), which enhanced the interaction of visual prompts between layers, addressing the frequent performance fluctuations observed in VPT variant models.

III. Proposed Method

A. Problem Definition

In OS-RSSC, the training set is represented as $S_{\text{train}} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) | \ y_i \in \mathcal{C}_{\text{known}}\}$, where x_i is the ith scene image, y_i is its category, and $\mathcal{C}_{\text{known}} = \{c_1, c_2, \ldots, c_K\}$ represents a set of category names in training set. The c_i from S_{train} is called the known category, and K is the number of known categories. For OS-RSSC, the test set $S_{\text{test}} = \{(x'_1, y'_1), (x'_2, y'_2), \ldots, (x'_m, y'_m)\}$ will contain new categories that do not belong to the training set S_{train}. We refer to the new categories appearing in the test set as an unknown category $\mathcal{C}_{\text{unknown}} = \{c_{K+1}\}$, and $y'_i \in \mathcal{C}_{\text{known}} \cup \mathcal{C}_{\text{unknown}}$. In practical applications, we may not know the specific names of the unknown category c_{K+1}. The goal of OS-RSSC is to build a suitable model to fully utilize the data in S_{train} for training, so as to identify whether test image x'_i belongs to one of the known categories $\mathcal{C}_{\text{known}}$ or the unknown category $\mathcal{C}_{\text{unknown}}$.

B. Overview of PNPL

PNPL employs the pretrained CLIP as the main framework and adopts a prompt tuning scheme to learn visual prompts and positive–negative textual prompts for OS-RSSC, avoiding the need to retrain the weight parameters of text and visual encoder in CLIP. Since complex remote sensing scenes may include various objects and disparate backgrounds, and even the scenes of unknown categories may contain the same objects or backgrounds as the scenes of known categories [54], it is challenging to directly learn positive–negative prompts suitable for OS-RSSC. Therefore, PNPL adopts a two-stage training scheme. In the first stage, following [18], [47], visual prompts and positive textual prompts are optimized for recognizing known categories, ensuring the classification performance of known categories. In the second stage, the visual and positive textual prompts from the first stage are frozen, and the focus shifts to optimizing negative textual prompts. The connection between visual and textual branches is established through cosine similarity computation between image features and text features. Finally, PNPL determines whether an input belongs to a known or unknown class based on the output OSC score. The detailed process of obtaining positive visual and textual prompts is shown in Fig. 1.

C. Positive Prompt Learning

In visual branch, we introduce visual prompt $\xi \in \mathbb{R}^{3\times w\times h}$, a learnable visual representation with the same dimensions as the input image $x \in \mathbb{R}^{3\times w\times h}$. This prompt is combined with the input image and fed into the frozen visual encoder $\text{Encoder}_I(\cdot)$ of CLIP [20] to obtain the image features $F \in \mathbb{R}^d$, where d is the feature dimension

$$F = \text{Encoder}_I\left(x \oplus \xi\right) \tag{1}$$

where $\oplus$ is element-wise addition.

A contrastive loss $\mathcal{L}_{\text{con}}$ is introduced to enhance interclass separability and reduce intraclass variability by pulling image features F_I of the same category closer and pushing features of different categories further apart [55], [56]. Given a batch of samples $\{x_1, x_2, \ldots, x_B\}$, F_b represents the features of x_b

$$\mathcal{L}_{\text{con}} = -\sum_{b=1}^{B} \frac{1}{|\mathcal{P}_b|} \sum_{\rho\in\mathcal{P}_b} \log\left(\frac{\exp\left(F_b F_\rho^\top/\tau_1\right)}{\sum_{j=1}^{B} \exp\left(F_b F_j^\top/\tau_1\right)}\right) \tag{2}$$

where $\mathcal{P}_b$ represents the set of samples with the same category as the sample x_b in a batch during training, and $\tau_1 \in (0, 1)$ is a trainable temperature parameter.

In textual branch, for the kth category, the word embedding of its category name c_k is α_k and its learnable positive textual prompt is represented as $\mathcal{W}_k = \{\omega_1, \omega_2, \ldots, \omega_l, \alpha_k\} \in \mathbb{R}^{(l+1)\times r}$, where each $\omega_i \in \mathbb{R}^r$ is a learnable vector with the same dimension as the α_k and l is the number of learnable vectors. The initialization of positive textual prompts $\mathcal{W} = [\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_K]$ is carried out using the embeddings of “a photo of a [class].” Then, $\mathcal{W}$, which consists of shared learnable prompt tokens $\{\omega_1, \omega_2, \ldots, \omega_l\}$ and a class-specific token α_k, is fed into the frozen text encoder $\text{Encoder}_T(\cdot)$ of CLIP to obtain the text feature $F_T^\omega \in \mathbb{R}^{d\times K}$

$$F_T^\omega = \text{Encoder}_T\left(\mathcal{W}\right). \tag{3}$$

The cosine similarity between the text feature F_T^ω and the image feature F is computed to obtain the class confidence p_k that the input image x belongs to class c_k

$$p_k = \frac{\exp\left(\text{F}\text{F}_{T_k}^{\omega\top}/\tau_2\right)}{\sum_{j=1}^{K} \exp\left(\text{F}\text{F}_{T_j}^{\omega\top}/\tau_2\right)} \tag{4}$$

where $F_{T_k}^\omega$ denotes the features of $\mathcal{W}_k$ and τ_2 is a frozen temperature parameter of CLIP. During the training phase, the cross-entropy function is utilized to compute classification loss for each sample

$$\mathcal{L}_{\text{cls}} = -\sum_{k=1}^{K} \mathbb{1}\left(y_i = c_k\right) \log p_k \tag{5}$$

where $\mathbb{1}(\cdot)$ is an indicator function, which takes 1 when $y_i = c_k$ and 0 otherwise.

The loss function of training the visual prompt ξ and the positive textual prompt $\mathcal{W}$ is as follows:

$$\mathcal{L}_{\text{pos}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{con}}. \tag{6}$$

This process aims to learn the prompts ξ and $\mathcal{W}$ tailored for known classes.

D. Negative Prompt Learning

The fundamental challenge of OS-RSSC resides in enabling discriminative capability for unknown classes that exhibit high feature similarity with known classes, under complete absence of unknown class exemplars during training. When faced with unknown class samples, if their features are very different from those of known class samples, it is easy for the model to distinguish the unknown class samples. On the contrary, when

Authorized licensed use limited to: Central China Normal University. Downloaded on January 28,2026 at 12:43:44 UTC from IEEE Xplore. Restrictions apply.

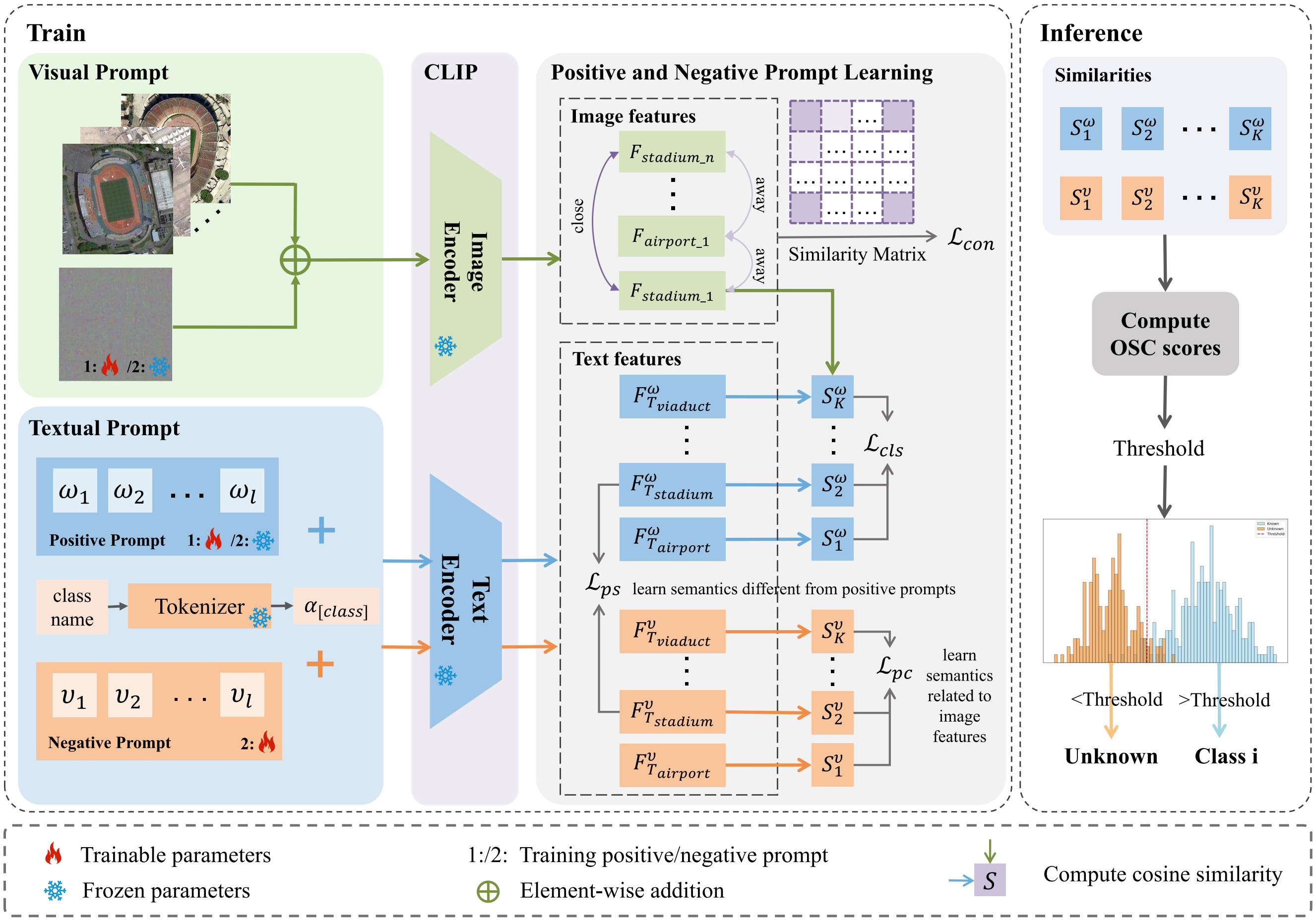


Fig. 1. Detailed structure of training positive and negative prompts. In the stage of training positive prompts, the visual prompt branch combines input images with learnable visual prompts, while the textual branch generates text features using learnable textual positive prompts. The contrastive loss $\mathcal{L}_{\text{con}}$ is employed to reduce intraclass variations among input image features, while the classification loss $\mathcal{L}_{\text{cls}}$ ensures that input image features align with the textual features of their corresponding classes. In the stage of training negative prompts, the visual prompts and positive textual prompts are frozen. Through prompt constraint loss $\mathcal{L}_{\text{pc}}$ and prompt separation loss $\mathcal{L}_{\text{ps}}$, the model is guided to learn negative prompts that effectively capture the semantics of unknown classes.

the features of unknown class samples are close to those of known class samples, it is difficult to distinguish the unknown class samples.

Remote sensing scene images typically exhibit large intraclass variations and small interclass differences, making OS-RSSC more challenging. The distribution of known and unknown classes in the aerial image dataset (AID) is visualized in Fig. 2, revealing that some known and unknown classes are adjacent or even overlap in feature space. This suggests that certain known and unknown classes share similar features in high-dimensional space. Therefore, negative prompts are intended to guide the model to focus on features that resemble those of the known classes without belonging to them.

However, only images of known classes are available during training process, making it difficult for the negative prompts to directly encounter the features of unknown classes. Consequently, we propose two loss functions: the prompt constraint loss prevents the model from learning features that are irrelevant to all categories, while the prompt separation loss distinguishes negative prompts from positive prompts.

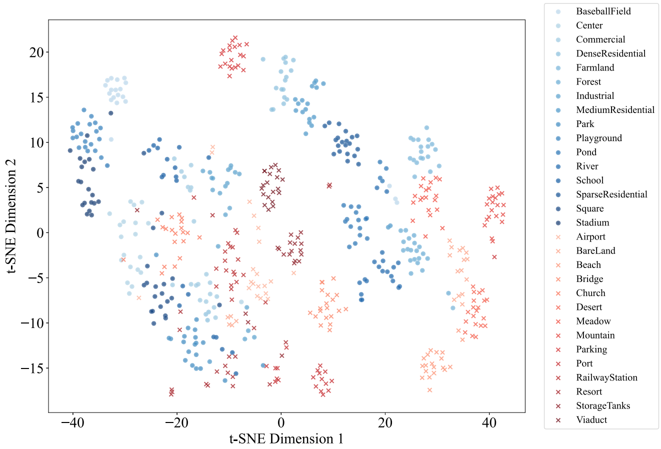


Fig. 2. Visualization results of extracting features of the AID image through the CLIP model and reducing the high-dimensional features to two dimensions using the t-SNE algorithm. Blue circle represents known categories, and red cross represents unknown categories.

When training negative prompts, the learned visual prompts and positive prompts from the first stage are frozen. And the learned positive textual prompts are used to initialize the

Authorized licensed use limited to: Central China Normal University. Downloaded on January 28,2026 at 12:43:44 UTC from IEEE Xplore. Restrictions apply.

negative textual prompts $\mathcal{V}_{m,i} = \{v_{m,1}, v_{m,2}, \ldots, v_{m,l}, \alpha_k\} \in \mathbb{R}^{(l+1)\times r}$, where m denotes the mth negative prompt, with a total of M negative textual prompts. Then, we can obtain the negative text feature $F^{v}_{T_m} \in \mathbb{R}^{d\times K}$.

1) Prompt Constraint Loss: We aim to ensure that the features of the negative prompts are closer to the features of unknown class images while avoiding the learning of irrelevant features. To achieve this, we employ a prompt constraint loss $\mathcal{L}_{\mathrm{pc}}$.

Similar to the positive textual prompts, the negative textual prompts are generic templates of negative semantics and do not depend on class labels. As explicit semantic descriptions cannot be defined for unknown classes in the same way as for known classes, we design M negative prompts to model the unknown semantic space with diversity. Tokens of different class names are added to the shared negative prompts to obtain $\mathcal{V}_{m,k} = \{v_{m,1}, v_{m,2}, \ldots, v_{m,l}, \alpha_k\}$. Then, the cosine similarity $S = \{S^{v}_{1,1}, S^{v}_{1,2}, \ldots, S^{v}_{M,K-1}, S^{v}_{M,K}\} \in \mathbb{R}^{(M\times K)}$ of the negative text features of all categories and the input image's feature can be obtained. S_t represents the tth cosine similarity, where $t \in [1, M \times K]$. For each sample in a batch, we define $\mathcal{L}_{\mathrm{pc}}$ as

$$\mathcal{L}_{\mathrm{pc}} = -\frac{1}{M \times K} \sum_{t=1}^{M\times K} \left(L_t \log \sigma(S_t) + (1 - L_t) \log(1 - \sigma(S_t))\right) \tag{7}$$

where $L = \{l_1, l_2, \ldots, l_{M\times K}\} \in \mathbb{R}^{(M\times K)}$ and $\sigma(\cdot)$ is sigmoid function. L_t represents the binary classification label corresponding to each cosine similarity S_t, which is used to indicate whether the current input image belongs to the category related to the tth cosine similarity S_t. If the category name of the textual input corresponding to S_t is the category of the input image, $L_t = 1$, otherwise $L_t = 0$.

This loss function encourages the features of negative prompts to align with the features of known class images, ensuring that they remain consistent with the true data distribution. The prompt constraint loss effectively prevents the negative prompts from capturing completely unrelated or misleading information.

2) Prompt Separation Loss: By applying the prompt constraint loss to negative prompts, it prevents the negative prompts from becoming completely irrelevant to all classes. However, this constraint alone does not establish a direct connection between the negative prompts and the unknown classes. Positive prompts are closer to the features of known class images and have semantics similar to "a photo of a [class]." In order to learn features that are closer to unknown classes, the semantics of negative prompts should be significantly different from those of positive prompts. Therefore, we designed a prompt separation loss $\mathcal{L}_{\mathrm{ps}}$

$$\mathcal{L}_{\mathrm{ps}} = \frac{1}{M \cdot K} \sum_{m=1}^{M} \sum_{j=1}^{K} \cos\left(F^{\omega}_{T_j}, F^{v}_{T_{m,j}}\right) \tag{8}$$

where $F^{v}_{T_{m,j}}$ represents the text feature obtained by the text encoder for the mth negative prompt of the jth category.

This loss function encourages the model to learn semantics different from positive prompts by enlarging the feature distance between positive and negative textual prompts, thereby improving the correlation between negative prompts features and unknown class image features.

The loss function for training negative prompt is

$$\mathcal{L}_{\mathrm{neg}} = \mathcal{L}_{\mathrm{pc}} + \mathcal{L}_{\mathrm{ps}}. \tag{9}$$

The combination of prompt separation loss and prompt constraint loss helps improve the model's performance in distinguishing known and unknown classes, making classification tasks more accurate and effective.

E. Inference of PNPL

During the inference process, we improve the original probability calculation method in CLIP by incorporating the influence of negative prompts. Specifically, we compute the similarity between the input image and both the positive and negative prompts to get the confidence p_k for each image across different classes

$$p_k = \frac{\exp\left(\mathrm{FF}^{\omega\top}_{T_k}/\tau_2\right)}{\sum_{j=1}^{K} \exp\left(\mathrm{FF}^{\omega\top}_{T_j}/\tau_2\right) + \sum_{m=1}^{M}\sum_{j=1}^{K} \exp\left(\mathrm{FF}^{v\top}_{T_{m,j}}/\tau_2\right)} \tag{10}$$

where $F^{v}_{T_{m,j}}$ represents the text feature obtained by the text encoder for the mth negative prompt of the jth category. For images belonging to known classes, they tend to match closely with the positive prompts, resulting in a higher $\mathrm{FF}^{\omega\top}_{T}$ and a lower $\mathrm{FF}^{v\top}_{T}$. Consequently, the confidence p_k becomes higher. Conversely, for unknown classes, the similarity with the negative prompts is higher, leading to a higher $\mathrm{FF}^{v\top}_{T}$ and a lower $\mathrm{FF}^{\omega\top}_{T}$, which in turn leads to a lower confidence p_k. We then take the maximum value $S_{\mathrm{OSC}} = \max(p_k)$ as the OSC score. Thus, the model can ultimately increase the gap between known and unknown classes, achieving a clear distinction between them.

Nevertheless, due to the similarity between different classes, the distribution of OSC scores may overlap, making the boundary between known and unknown classes ambiguous. We refer to this overlapping region as the transition zone, where the model struggles to clearly assign a label of either a known or unknown class. To address this issue, we have designed a prediction method based on OSC scores with a threshold mechanism to effectively identify unknown classes.

Specifically, to accurately determine the threshold, we employ the k-means clustering algorithm to group the OSC scores of all test samples into three clusters. These three clusters are interpreted as: unknown class, transition class, and known class. The inclusion of the transition class allows the model to better capture the inherent uncertainty in this region.

Based on the clustering results, we define the mean of the cluster centers for the unknown and transition classes as the threshold, which enables an effective differentiation between unknown and known classes

$$\text{class} = \begin{cases} \mathrm{index}\left(S_{\mathrm{OSC}}\right), & \text{if } S_{\mathrm{OSC}} > \text{threshold} \\ \text{unknown}, & \text{otherwise.} \end{cases} \tag{11}$$

If an OSC score falls below the threshold, the image is classified as an unknown class, and if it exceeds the threshold, it is assigned to the known class with the highest confidence.

Authorized licensed use limited to: Central China Normal University. Downloaded on January 28,2026 at 12:43:44 UTC from IEEE Xplore. Restrictions apply.

TABLE I

UNKNOWN CLASS SETTINGS OF EACH DATASET UNDER DIFFERENT OPENNESS

Datasets	O^*	Unknown Classes
UCM	3.92%	agricultural, beach, storagetanks
	7.00%	agricultural, beach, chaparral, forest, river
	14.72%	baseball diamond, buildings, dense residential, golfcourse, medium residential, mobile homepark, parkinglot, sparse residential, tenniscourt
AID	3.64%	bareland, beach, desert, mountain
	9.25%	airport, bridge, church, parking, port, railway station, resort, storagetanks, viaduct
	16.59%	airport, bareland, beach, bridge, church, desert, meadow, mountain, parking, port, railway station, resort, storage tanks, viaduct
NWPU	2.99%	beach, harbor, island, lake, river, sea ice
	7.22%	airplane, bridge, church, freeway, intersection, overpass, palace, railway, railway station, roundabout, runway
	10.56%	airplane, bridge, church, freeway, harbor, intersection, island, lake, overpass, railway, railway station, river, roundabout, runway, sea ice

IV. EXPERIMENTS

A. Datasets

We conducted experiments using the following three popular remote sensing scene datasets.

1) UC Merced Land Use Dataset (UCM) [57]: UCM consists of 21 classes and 100 images per class. The image size is 256 × 256 and the spatial resolution is 0.3 m. 80% of the data are randomly selected as the training set.

2) Aerial Image Dataset [58]: AID is composed of 30 categories and a total of 10 000 images ranging from 220 to 420 images per category. The image size is 600 × 600, and the spatial resolution ranges from 0.5 to 8 m/pixel. 50% of the data is randomly selected as the training set.

3) NWPU-RESISC45 Dataset (NWPU) [59]: NWPU contains 45 categories and 700 images per category. The image size is 256 × 256, and the spatial resolution ranges from 0.2 to 30 m/pixel. 20% of the data are randomly selected as the training set.

B. Evaluation Measures

1) Openness: One of the important factors affecting OSC is openness O^* [5], which quantifies the degree of openness of the dataset. Its definition is as follows:

$$O^* = 1 - \sqrt{\frac{2 \times N_{\text{train}}}{N_{\text{train}} + N_{\text{test}}}}. \tag{12}$$

The method of dividing the dataset is consistent with [38], and the openness is calculated according to (12). The detailed settings of the openness are shown in Table I, where the classes listed in the table represent the unknown classes that are completely excluded from training process.

2) Area Under the Receiver Operating Characteristic (AUROC): To evaluate the ability of PNPL to distinguish between known and unknown classes, we use the evaluation metric AUROC which is suitable for testing the performance of binary classifiers. The closer the AUROC value is to 1, the stronger the ability of the method used to distinguish between known and unknown classes.

3) Open OA: OSC not only requires distinguishing unknown classes but also aims to maintain high classification accuracy for known classes. Therefore, the evaluation metric of open OA is used, which considers both the classification of known classes and the recognition of unknown classes. It extends the evaluation index of closed OA from class K to class $K+1$, and adds this class as an unknown class. The formula is as follows:

$$\text{Open OA} = \frac{\sum_{i=1}^{K+1} (\text{TP}_i + \text{TN}_i)}{\sum_{i=1}^{K+1} (\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i)}. \tag{13}$$

TP_i, TN_i, FP_i, and FN_i represent true positive, true negative, false positive, and false negative of class i, respectively.

C. Experiment Settings

We use CLIP-B/16 as backbone, which is pretrained from OpenCLIP. Images are resized to 224 × 224 to fit the input of ViT. Random cropping and random horizontal flipping are utilized for data augmentation. The number of epochs for training positive and negative prompts is set to 100. The batch size is set to 64. The number M of negative prompts is set to 2. The number of learnable vectors l is set to 4 because we initialize the prompt with the phrase “a photo of a,” which consists of four tokens after tokenization [18]. Trainable temperature parameter τ_1 is initialized to 0.5. The parameters are optimized using SGD optimizer with learning rate of 0.1, and the cosine annexing learning rate scheduler is used to dynamically adjust the learning rate. The experimental results are the average values obtained by setting the random seed to 1, 2, and 3. All our experiments are conducted on Nvidia RTX 4090 and 12th Gen Intel[1] Core[2] i7-12700.

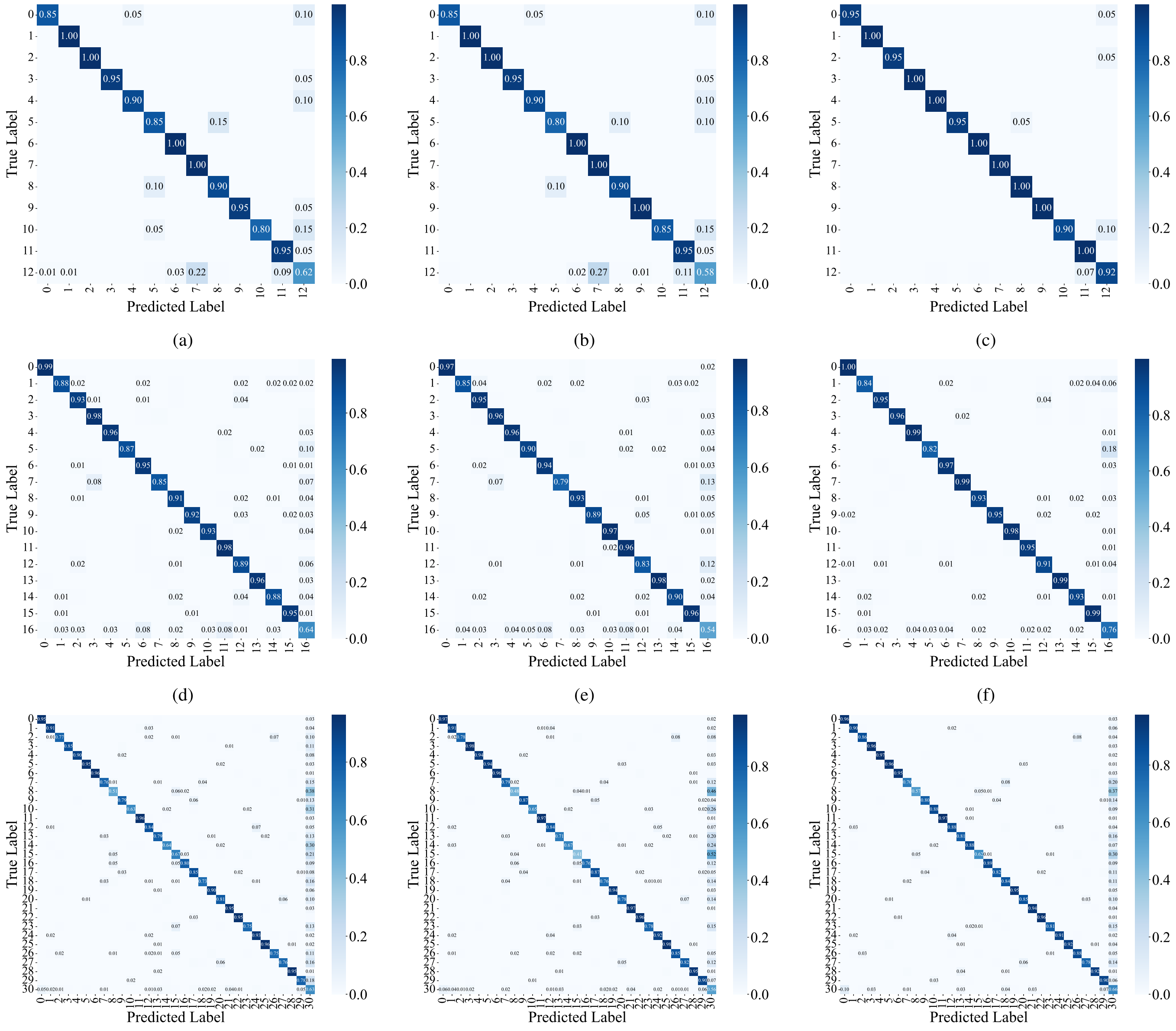
D. Compared Methods

To verify the effectiveness of PNPL in OS-RSSC, we compared it with three different categories of methods. The first category includes popular OSC methods: SoftMax, OpenMax [6], conditional Gaussian distribution learning (CGDL) [60], generative causal model (GCM-CF) [61], ARPL [34], and ARPL+ [10]. The second category consists of RSSC-based methods, including DFRL [38], remote sensing mamba (RSMamba) [62], and multiscale sparse cross-attention network (MSCN) [63], where RSMamba and MSCN employ the same setting as SoftMax, using 0.5 as the threshold to distinguish between known and unknown classes. The third category consists of CLIP-based methods, including CoOp [18], local regularized CoOp (LoCoOp) [22], consistency-guided prompt learning (CoPrompt) [64], and NegPrompt [23]. All CLIP-based methods do not specify a threshold for judging known and unknown classes. To ensure fairness, the threshold judgment method proposed in this article is uniformly used. For methods that do not use negative prompts, the GL-MCM score [22] is used as OSC score.

[1] Registered trademark.

[2] Trademarked.

Authorized licensed use limited to: Central China Normal University. Downloaded on January 28,2026 at 12:43:44 UTC from IEEE Xplore. Restrictions apply.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

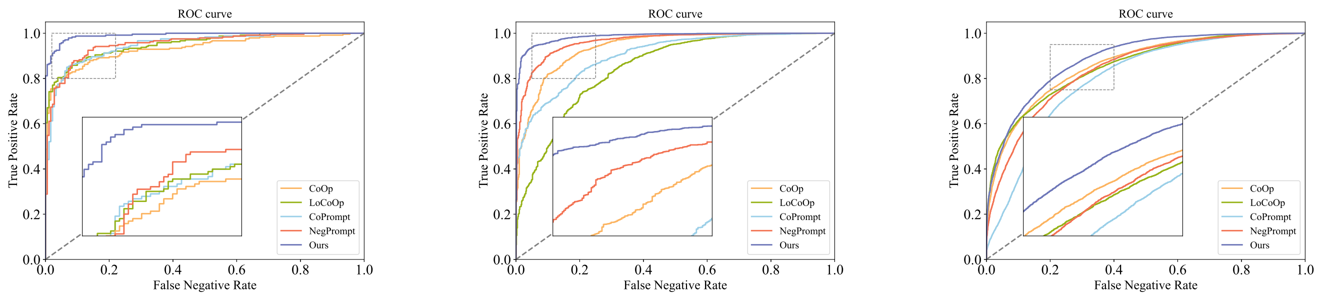
Fig. 3. Confusion matrix under different experimental settings. (a) CoOp UCM O^* = 14.72%. (b) LoCoOp UCM O^* = 14.72%. (c) PNPL UCM O^* = 14.72%. (d) CoOp AID O^* = 16.59%. (e) LoCoOp AID O^* = 16.59%. (f) PNPL AID O^* = 16.59%. (g) CoOp NWPU O^* = 10.56%. (h) LoCoOp NWPU O^* = 10.56%. (i) PNPL NWPU O^* = 10.56%. The class names in (a)–(c): 0—agricultural, 1—airplane, 2—beach, 3—chaparral, 4—forest, 5—freeway, 6—harbor, 7—intersection, 8—overpass, 9—river, 10—runway, 11—storage tanks, and 12—unknown. The class names in (d)–(f): 0—baseball field, 1—center, 2—commercial, 3—dense residential, 4—farmland, 5—forest, 6—industrial, 7—medium residential, 8—park, 9—playground, 10—pond, 11—river, 12—school, 13—sparse residential, 14—square, 15—stadium, 16—unknown. The class names in (g)–(i): 0—airport, 1—baseball diamond, 2—basketball court, 3—beach, 4—chaparral, 5—circular farmland, 6—cloud, 7—commercial area, 8—dense residential, 9—desert, 10—forest, 11—golf course, 12—ground track field, 13—industrial area, 14—meadow, 15—medium residential, 16—mobile home park, 17—mountain, 18—palace, 19—parking lot, 20—rectangular farmland, 21—ship, 22—snowberg, 23—sparse residential, 24—stadium, 25—storage tank, 26—tennis court, 27—terrace, 28—thermal power station, 29—wetland, and 30—unknown.

E. Main Results

OAs on UCM, AID, and NWPU datasets are shown in Table II, and AUROCs are shown in Table III. The confusion matrix visualization is shown in Fig. 3. The comparison of ROC curves obtained by some experimental settings is shown in Fig. 4. From Tables II and III, PNPL exhibits superior performance under different openness settings of three different datasets. Moreover, under different openness settings within the same dataset, AUROC and OA exhibit minimal fluctuation. This demonstrates that PNPL is relatively insensitive to openness, meaning it can adapt well even as the number of unknown classes increases. Taking UCM dataset as an example, when the openness is 3.92% and 14.72%, respectively, the fluctuation values of SoftMax, OpenMax, CGDL, GCM-CF, ARPL, and ARPL+ on OA are 21.21%, 10.07%, 22.14%, 23.57%, 5.28%, and 7.26%, respectively, while the fluctuation value of PNPL is only 2.22%.

RSSC-based methods generally achieve superior performance in terms of OA and AUROC compared to popular OSC methods approaches designed for natural images. This

Authorized licensed use limited to: Central China Normal University. Downloaded on January 28,2026 at 12:43:44 UTC from IEEE Xplore. Restrictions apply.



(a)

(b)

(c)

Fig. 4. Receiver operating curve under different experimental settings. (a) UCM O^* = 14.72%. (b) AID O^* = 3.64%. (c) NWPU O^* = 10.56%.

TABLE II

Comparison of OA (%) of Different Methods on UCM, AID, and NWPU Datasets With Different Openness. The Boldfaced Results Indicate the Best Performance

Methods	UCM			AID			NWPU		
	3.92%	7.00%	14.72%	3.64%	9.25%	16.59%	2.99%	7.22%	10.56%
OSC methods									
SoftMax	60.42	59.54	39.21	61.52	43.61	37.53	26.73	26.12	23.04
OpenMax	60.33	59.27	50.26	62.29	52.02	50.13	51.12	50.04	42.83
CGDL	73.81	64.05	51.67	67.58	53.58	43.84	60.67	55.10	47.83
GCM-CF	74.29	65.01	50.72	68.94	56.76	50.74	55.54	50.89	43.83
ARPL	78.61	78.13	73.33	71.18	68.50	68.43	61.33	60.57	59.39
ARPL+	86.20	84.05	78.94	81.77	73.68	69.38	64.24	62.27	61.53
RSSC methods									
DFRL	88.10	85.71	86.19	82.34	77.58	68.48	73.88	69.04	62.78
RSMamba	89.52	89.76	76.67	84.54	77.26	71.36	76.84	79.96	77.22
MSCN	85.48	85.95	79.29	88.92	84.72	76.82	75.13	75.96	77.48
CLIP-based methods									
CoOp	86.11	86.27	80.56	86.17	83.16	81.69	70.19	76.92	75.66
LoCoOp	77.14	83.49	83.10	79.80	82.90	75.46	66.31	77.13	74.03
CoPrompt	84.21	85.08	86.19	77.98	75.94	74.66	72.14	70.82	71.40
NegPrompt	89.02	90.93	85.36	86.26	87.37	82.02	77.87	79.87	76.44
PNPL(Ours)	**91.81**	**93.56**	**94.03**	**90.01**	**89.04**	**84.20**	**80.56**	**81.70**	**79.94**

improvement can be attributed to their explicit consideration of the distinct feature characteristics between remote sensing scene images and natural images. Remarkably, some experimental results even surpass those obtained using CLIP.

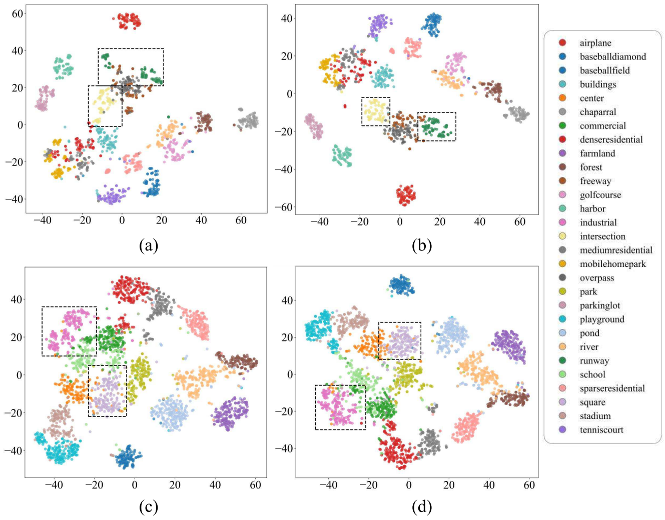
The experimental results of methods based on CLIP are generally shown better performance, proving that pretrained VLM, especially CLIP, is suitable for OS-RSSC. Compared to positive prompt training methods CoOp and LoCoOp, PNPL also shows excellent performance, validating the effectiveness of training negative prompt for OS-RSSC. To further analyze the impact of training negative prompts, we present the confusion matrices of CoOp, LoCoOp, and PNPL in Fig. 3 under the maximum openness experimental setting of the three datasets. The darker the color of the diagonal, the higher the classification accuracy. It can be seen that PNPL improves the classification performance of unknown classes compared to CoOp and LoCoOp which only train positive prompts. The last column and last row of the confusion matrix represent the case of dividing known classes into unknown classes and the case of dividing unknown classes into known classes, respectively. PNPL effectively ameliorates both types of problems, resulting in higher OA.

CoOp trains prompts using images from known classes and does not consider unknown classes. This limitation prevents it from fully harnessing the potential of CLIP in OSC. LoCoOp treats irrelevant information, such as backgrounds in known class images, as nuisances and learns to push them away from the prompt embeddings. However, this assumption does not hold for RSSIs with densely distributed objects, leading to unstable performance and even a decline compared to CoOp, which does not consider local features. Although CoPrompt incorporates visual prompts, textual prompts, and adapters simultaneously, its performance is not stable, as it does not specifically consider the characteristics of OSC. NegPrompt shows improvement over methods that only train positive prompts, but its performance is still slightly lower than PNPL.

Authorized licensed use limited to: Central China Normal University. Downloaded on January 28,2026 at 12:43:44 UTC from IEEE Xplore. Restrictions apply.

TABLE III
COMPARISON OF AUROC OF DIFFERENT METHODS ON UCM, AID, AND NWPU DATASETS WITH DIFFERENT OPENNESS. THE BOLDFACED RESULTS INDICATE THE BEST PERFORMANCE

Methods	UCM			AID			NWPU		
	3.92%	7.00%	14.72%	3.64%	9.25%	16.59%	2.99%	7.22%	10.56%
OSC methods									
SoftMax	0.554	0.595	0.652	0.609	0.654	0.608	0.400	0.643	0.505
OpenMax	0.613	0.644	0.687	0.627	0.688	0.690	0.596	0.595	0.501
CGDL	0.724	0.634	0.690	0.605	0.548	0.639	0.579	0.626	0.514
GCM-CF	0.639	0.675	0.724	0.656	0.562	0.669	0.550	0.627	0.598
ARPL	0.629	0.609	0.785	0.707	0.660	0.519	0.465	0.626	0.643
ARPL+	0.860	0.859	0.822	0.849	0.767	0.729	0.679	0.734	0.696
RSSC methods									
DFRL	0.887	0.924	0.978	0.929	0.894	0.918	0.855	**0.900**	0.877
RSMamba	0.928	0.962	0.920	0.887	0.853	0.839	0.812	0.870	0.852
MSCN	0.825	0.927	0.951	0.942	0.915	0.891	0.857	0.898	0.884
CLIP-based methods									
CoOp	0.968	0.960	0.927	0.946	0.912	0.930	0.842	0.887	0.865
LoCoOp	0.947	0.954	0.943	0.862	0.907	0.877	0.755	0.887	0.853
CoPrompt	0.960	0.971	0.938	0.924	0.868	0.884	0.895	0.772	0.802
NegPrompt	0.955	0.963	0.957	0.951	0.946	0.943	0.890	0.860	0.848
PNPL(Ours)	**0.971**	**0.978**	**0.986**	**0.962**	**0.957**	**0.963**	**0.905**	0.891	**0.888**



(a) (b) (c) (d)

Fig. 5. Image feature distributions of known classes on UCM (3.92% openness) and AID (16.59% openness) with and without visual prompts. (a) UCM without visual prompts, (b) UCM with visual prompts, (c) AID without visual prompts, and (d) AID with visual prompts.

This is because it sets a constraint to control the distance between positive prompts and negative prompts.

F. Ablation Study

1) Visualization on Image Features: Fig. 5 illustrates the features extracted from the original images and after adding visual prompts. The figure shows the impact of visual prompts when the openness of UCM is 3.92% and that of AID is 16.59%. Categories with significant improvements are marked with dashed boxes. Obviously, adding visual prompts can reduce the intraclass differences of image features and deal with the problem of large intraclass differences in RSSIs.

2) Impact of Different Prompts: Tables IV and V report the results of different prompt settings, highlighting that all three prompts set in this article are crucial for OS-RSSC. PTP, VP, and NTP stand for positive textual prompts, visual prompts, and negative textual prompts, respectively. Similar to the methods of training textual prompts such as CoOp, our method has greatly improved the model performance after training positive textual prompts compared with CLIP. Training negative prompts designed specifically for OSC improves AUROC and OA. Furthermore, incorporating image prompts on top of this improves model performance even further.

To further demonstrate the effect of negative prompts in OSC task, we present the results from both the first and second stages of the model in Fig. 6. As can be observed, training with negative prompts consistently improves the model's performance in OS-RSSC. Coupled with the results presented in Table IV, it can be concluded that, regardless of whether visual prompts are trained in the first stage, the negative prompts in the second stage lead to a noticeable enhancement in the experimental outcomes.

3) Impact of $\mathcal{L}_{con}$ and $\mathcal{L}_{ps}$: Since $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{pc}}$ are the most basic loss functions for training positive and negative prompts, we focus on verifying the effectiveness of $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{ps}}$ through ablation experiments. The impact on OA is shown in Table VI, while the impact on AUROC is shown in Table VII. The results show that in most experimental settings, adding $\mathcal{L}_{\text{con}}$ helps expand the interclass differences of images, thereby improving AUROC to a certain extent. Without considering $\mathcal{L}_{\text{ps}}$, OA and AUROC values decrease, which proves that adding $\mathcal{L}_{\text{ps}}$ to

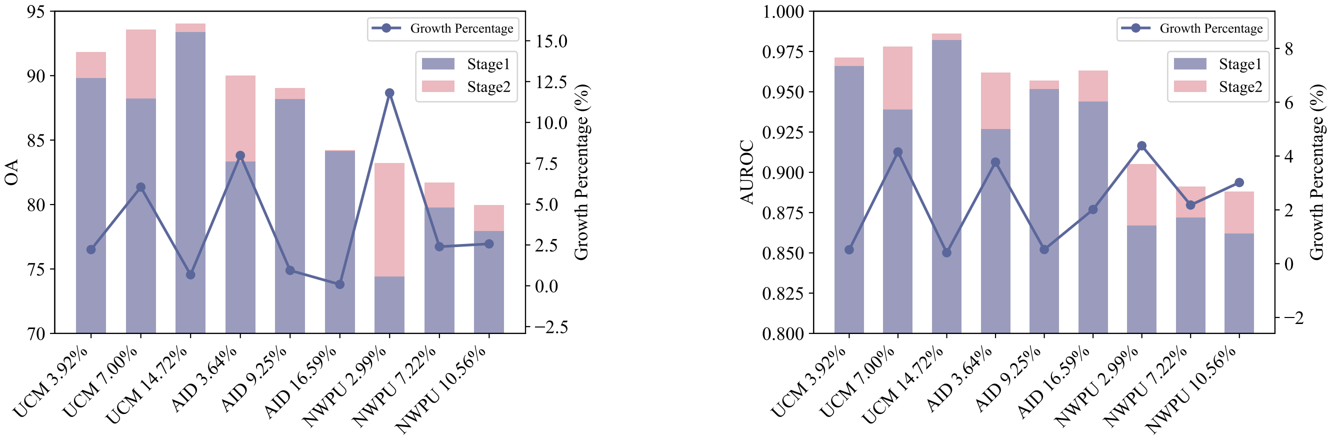
Authorized licensed use limited to: Central China Normal University. Downloaded on January 28,2026 at 12:43:44 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
ABLATION STUDY ON THE IMPACT OF DIFFERENT PROMPT COMBINATIONS ON OA

Prompts			UCM			AID			NWPU			Average
PTP	VP	NTP	3.92%	7.00%	14.72%	3.64%	9.25%	16.59%	2.99%	7.22%	10.56%	
			62.53	69.94	54.42	54.03	48.13	51.01	53.22	56.20	58.02	56.39
✓			86.64	88.39	90.14	85.42	87.18	81.70	72.87	79.14	76.53	83.11
✓		✓	88.86	92.84	89.37	89.65	88.20	83.22	77.91	80.73	78.66	85.49
✓	✓	✓	**91.81**	**93.56**	**94.03**	**90.01**	**89.04**	**84.20**	**80.56**	**81.70**	**79.94**	**87.21**

TABLE V
ABLATION STUDY ON THE IMPACT OF DIFFERENT PROMPT COMBINATIONS ON AUROC

Prompts			UCM			AID			NWPU			Average
PTP	VP	NTP	3.92%	7.00%	14.72%	3.64%	9.25%	16.59%	2.99%	7.22%	10.56%	
			0.904	0.961	0.698	0.909	0.676	0.740	0.869	0.744	0.768	0.808
✓			0.942	0.964	0.970	0.932	0.945	0.936	0.852	0.859	0.846	0.916
✓		✓	0.949	0.976	0.968	0.958	**0.961**	0.957	0.902	0.873	0.870	0.935
✓	✓	✓	**0.971**	**0.978**	**0.986**	**0.962**	0.957	**0.963**	**0.905**	**0.891**	**0.888**	**0.945**



(a) (b)

Fig. 6. Comparison of training positive prompts and training negative prompts. (a) and (b) Effects of training positive and negative prompts on OA and AUROC, respectively.

TABLE VI
ABLATION STUDY ON THE IMPACT OF CONTRASTIVE LOSS $\mathcal{L}_{CON}$ AND PROMPT SEPARATION LOSS $\mathcal{L}_{PS}$ ON OA

$\mathcal{L}_{con}$	$\mathcal{L}_{ps}$	UCM			AID			NWPU			Average
		3.92%	7.00%	14.72%	3.64%	9.25%	16.59%	2.99%	7.22%	10.56%	
✓		90.85	**94.11**	89.98	84.04	86.88	83.94	78.80	80.35	78.22	85.24
	✓	91.57	92.60	93.48	**90.14**	**89.09**	82.17	78.29	81.12	79.09	86.39
✓	✓	**91.81**	93.56	**94.03**	90.01	89.04	**84.20**	**80.56**	**81.70**	**79.94**	**87.21**

the overall loss to reduce the similarity between negative and positive prompts is effective.

4) Impact of Number of Negative Prompts: M is the number of negative prompts. We conduct experiments with different numbers of negative prompts and present the results in Fig. 7. In most experimental settings, increasing the number of negative prompts slightly increases model performance, but overall, the performance gains are not significant. In addition, increasing the number of negative prompts leads to an increase in the number of trainable parameters, which raises computational costs. In our approach, GFLOPs is 293.12 and the average training time for one epoch on UCM, AID, and

Authorized licensed use limited to: Central China Normal University. Downloaded on January 28,2026 at 12:43:44 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
ABLATION STUDY ON THE IMPACT OF CONTRASTIVE LOSS $\mathcal{L}_{\text{CON}}$ AND PROMPT SEPARATION LOSS $\mathcal{L}_{\text{PS}}$ ON AUROC

$\mathcal{L}_{con}$	$\mathcal{L}_{ps}$	UCM			AID			NWPU			Average
		3.92%	7.00%	14.72%	3.64%	9.25%	16.59%	2.99%	7.22%	10.56%	
✓		0.962	**0.981**	0.979	0.918	0.940	0.956	0.898	0.880	0.873	0.932
	✓	0.957	0.970	0.984	0.955	**0.965**	0.959	0.895	0.889	0.879	0.939
✓	✓	**0.971**	0.978	**0.986**	**0.962**	0.957	**0.963**	**0.905**	**0.891**	**0.888**	**0.945**

TABLE VIII
ABLATION STUDY ON THE IMPACT OF THRESHOLD ON OA

Threshold	UCM			AID			NWPU			Average
	3.92%	7.00%	14.72%	3.64%	9.25%	16.59%	2.99%	7.22%	10.56%	
EVT	85.04	78.20	83.46	86.31	70.10	51.92	80.54	71.20	64.47	74.58
Otsu	87.43	89.26	93.08	78.08	86.77	**88.73**	66.23	74.07	76.18	82.20
K-means	**91.81**	**93.56**	**94.03**	**90.01**	**89.04**	84.20	**80.56**	**81.70**	**79.94**	**87.21**

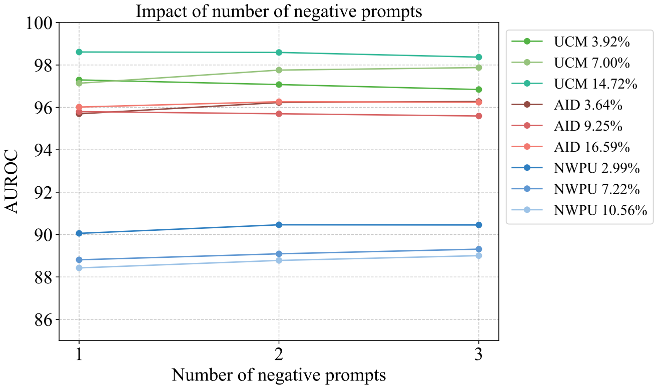


Fig. 7. Impact curve of negative prompt's number on AUROC under different experimental settings.

NWPU is 8.04, 16.50, and 31.23 s, respectively. The model contains 234.8 M parameters in total, with only 15 6672 being trainable. These learnable parameters originate from the visual prompts along with the positive and negative text prompts. The overwhelming majority of parameters remain frozen, including the complete CLIP backbone components such as both the vision encoder and text encoder. In order to minimize the number of parameters as much as possible, we ultimately chose 2 as the number of negative prompts M.

5) Impact of Threshold: To demonstrate the effectiveness of the k-means clustering algorithm in threshold determination, we conducted an ablation study comparing it with two popular threshold selection methods for OSC: extreme value theory (EVT) [5] and Otsu [38]. The results are presented in Table VIII. In the vast majority of cases, k-means outperformed the alternatives. Notably, both EVT and Otsu can produce significantly worse results when the openness level varies. For instance, EVT achieved only a 51.92% overall accuracy (OA) on the AID dataset at an openness of 16.59%.

G. Discussion

To investigate the impact of prompt learning on fine-tuning VLMs for OSC, we conducted experiments on public datasets with varying degrees of openness. The results demonstrate that PNPL can significantly enhance VLM performance in OSC. Unlike previous OS-RSSC methods, our approach introduces learnable negative prompts that aim to represent "unknown" features, which strengthens the distinction between known and unknown classes. However, PNPL's reliance on the features of positive prompts during the training of negative prompts implies that the effectiveness of the second stage is dependent on the quality of the first stage's training. If the training of positive prompts is suboptimal, the overall performance will be adversely affected. Moreover, the introduction of negative prompts and visual prompts inevitably increases the number of trainable parameters in PNPL. Therefore, developing methods to effectively train prompts that better align with the characteristics of unknown classes remains a promising area for further exploration.

V. CONCLUSION

In this article, we primarily focused on OSC of remote sensing scene images. We proposed PNPL, a CLIP-based method for training positive textual prompts, visual prompts, and negative textual prompts in OS-RSSC, which expands the distinction between known and unknown classes. PNPL leverages the power of contrastive learning and prompt tuning to enhance the model's ability to discriminate between known and unknown classes. By introducing negative prompts, PNPL effectively guides the model to learn semantic features of unknown classes, which are similar to but distinct from known classes, thereby enhancing the model's discriminative capability. The experimental results on three RSSI datasets demonstrate the effectiveness of PNPL. In future, we plan to study and explore OS-RSSC under semi-supervised conditions, further enhancing the robustness of the model.

Authorized licensed use limited to: Central China Normal University. Downloaded on January 28,2026 at 12:43:44 UTC from IEEE Xplore. Restrictions apply.

References

[1] Z. Xue et al., “Multimodal self-supervised learning for remote sensing data land cover classification,” *Pattern Recognit.*, vol. 157, Jan. 2025, Art. no. 110959.

[2] Y. Zhang, S. Yan, L. Zhang, and B. Du, “Fast projected fuzzy clustering with anchor guidance for multimodal remote sensing imagery,” *IEEE Trans. Image Process.*, vol. 33, pp. 4640–4653, 2024.

[3] Q. Bi, B. Zhou, K. Qin, Q. Ye, and G.-S. Xia, “All grains, one scheme (AGOS): Learning multigrain instance representation for aerial scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5629217.

[4] W. Wang, Y. Sun, J. Li, and X. Wang, “Frequency and spatial based multi-layer context network (FSCNet) for remote sensing scene classification,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 128, Apr. 2024, Art. no. 103781.

[5] C. Geng, S.-J. Huang, and S. Chen, “Recent advances in open set recognition: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3614–3631, Oct. 2021.

[6] A. Bendale and T. E. Boult, “Towards open set deep networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1563–1572.

[7] H. Xu, W. Chen, C. Tan, H. Ning, H. Sun, and W. Xie, “Orientational clustering learning for open-set hyperspectral image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[8] C. Robinson et al., “Global land-cover mapping with weak supervision: Outcome of the 2020 IEEE GRSS data fusion contest,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3185–3199, 2021.

[9] L. Fang et al., “Open-world recognition in remote sensing: Concepts, challenges, and opportunities,” *IEEE Geosci. Remote Sens. Mag.*, vol. 12, no. 2, pp. 8–31, Jun. 2024.

[10] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, “Open-set recognition: A good closed-set classifier is all you need,” in *Proc. Int. Conf. Learn. Represent.*, 2021.

[11] X. Chen, G. Zhu, and C. Ji, “Combining Hilbert feature sequence and lie group metric space for few-shot remote sensing scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5620415.

[12] X. Chen, G. Zhu, and J. Wei, “MMML: Multimanifold metric learning for few-shot remote-sensing image scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5618714.

[13] J. Chen and X. Wang, “Open set few-shot remote sensing scene classification based on a multiorder graph convolutional network and domain adaptation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4709517.

[14] J. Zheng et al., “Open-set domain adaptation for scene classification using multi-adversarial learning,” *ISPRS J. Photogramm. Remote Sens.*, vol. 208, pp. 245–260, Feb. 2024.

[15] M. M. Al Rahhal, Y. Bazi, R. Al-Dayil, B. M. Alwadei, N. Ammour, and N. Alajlan, “Energy-based learning for open-set classification in remote sensing imagery,” *Int. J. Remote Sens.*, vol. 43, nos. 15–16, pp. 6027–6037, Aug. 2022.

[16] W. Liu, X. Nie, B. Zhang, and X. Sun, “Incremental learning with open-set recognition for remote sensing image scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622916.

[17] S. Wang, D. Hou, and H. Xing, “A self-supervised-driven open-set unsupervised domain adaptation method for optical remote sensing image scene classification and retrieval,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5605515.

[18] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, Sep. 2022.

[19] S. Pratt, I. Covert, R. Liu, and A. Farhadi, “What does a platypus look like? Generating customized prompts for zero-shot image classification,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 15645–15655.

[20] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. 38th Int. Conf. Mach. Learn.*, Jul. 2021, pp. 8748–8763.

[21] J. Zhang, Y. Rao, X. Huang, G. Li, X. Zhou, and D. Zeng, “Frequency-aware multi-modal fine-tuning for few-shot open-set remote sensing scene classification,” *IEEE Trans. Multimedia*, vol. 26, pp. 7823–7837, 2024.

[22] A. Miyai, Q. Yu, G. Irie, and K. Aizawa, “LoCoOp: Few-shot out-of-distribution detection via prompt learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 76298–76310.

[23] T. Li, G. Pang, X. Bai, W. Miao, and J. Zheng, “Learning transferable negative prompts for out-of-distribution detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17584–17594.

[24] M. Al Rahhal, Y. Bazi, H. Elgibreen, and M. Zuair, “Vision-language models for zero-shot classification of remote sensing images,” *Appl. Sci.*, vol. 13, no. 22, p. 12462, Nov. 2023.

[25] H. Yao, R. Chen, W. Chen, H. Sun, W. Xie, and X. Lu, “Pseudolabel-based unreliable sample learning for semi-supervised hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5527116.

[26] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, “Deep learning in remote sensing applications: A meta-analysis and review,” *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[27] K. Nogueira, O. A. B. Penatti, and J. A. dos Santos, “Towards better exploiting convolutional neural networks for remote sensing scene classification,” *Pattern Recognit.*, vol. 61, pp. 539–556, Jan. 2017.

[28] H. Ning, T. Lei, M. An, H. Sun, Z. Hu, and A. K. Nandi, “Scale-wise interaction fusion and knowledge distillation network for aerial scene recognition,” *CAAI Trans. Intell. Technol.*, vol. 8, no. 4, pp. 1178–1190, Mar. 2023.

[29] J. Zhang, J. Huang, S. Jin, and S. Lu, “Vision-language models for vision tasks: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5625–5644, Aug. 2024.

[30] L. Jiao et al., “Brain-inspired remote sensing foundation models and open problems: A comprehensive survey,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 10084–10120, 2023.

[31] B. Liu, X. Chen, D. Zhou, P. Wang, and R. Wang, “Text guided zero-shot scene classification of high spatial resolution remote sensing images,” *J. Appl. Remote Sens.*, vol. 18, no. 1, Mar. 2024, Art. no. 014525.

[32] H.-M. Yang, X.-Y. Zhang, F. Yin, Q. Yang, and C.-L. Liu, “Convolutional prototype network for open set recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2358–2370, May 2022.

[33] Z. Xie, P. Duan, W. Liu, X. Kang, X. Wei, and S. Li, “Feature consistency-based prototype network for open-set hyperspectral image classification,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 9286–9296, Jul. 2024.

[34] G. Chen, P. Peng, X. Wang, and Y. Tian, “Adversarial reciprocal points learning for open set recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8065–8081, Nov. 2022.

[35] Z. Li, L. Qi, Y. Shi, and Y. Gao, “IOMatch: Simplifying open-set semi-supervised learning with joint inliers and outliers utilization,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 15824–15833.

[36] Y. Yu et al., “ANEDL: Adaptive negative evidential deep learning for open-set semi-supervised learning,” in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 15, pp. 16587–16595.

[37] S. Liu, Q. Shi, and L. Zhang, “Few-shot hyperspectral image classification with unknown classes using multitask deep learning,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5085–5102, Jun. 2021.

[38] H. Sun et al., “Deep feature reconstruction learning for open-set classification of remote-sensing imagery,” *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[39] W. Moon, J. Park, H. S. Seong, C.-H. Cho, and J.-P. Heo, “Difficulty-aware simulator for open set recognition,” in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 365–381.

[40] A. Wu and C. Deng, “TIB: Detecting unknown objects via two-stream information bottleneck,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 611–625, Jan. 2024.

[41] J. Wang et al., “Review of large vision models and visual prompt engineering,” *Meta-Radiology*, vol. 1, no. 3, Nov. 2023, Art. no. 100047.

[42] Y. Wang et al., “Emotion-oriented cross-modal prompting and alignment for human-centric emotional video captioning,” *IEEE Trans. Multimedia*, vol. 27, pp. 3766–3780, 2025.

[43] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, “MaPLe: Multi-modal prompt learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19113–19122.

[44] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Conditional prompt learning for vision-language models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16816–16825.

[45] W. Zhao, X. Lv, R. He, F. Zhao, H. Wang, and Y. He, “Diverse text-prompt generation for remote sensing image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5603310.

[46] H. Lin, C. Zhang, D. Hong, K. Dong, and C. Wen, “FedRSCLIP: Federated learning for remote sensing scene classification using vision-language models,” *IEEE Geosci. Remote Sens. Mag.*, early access, May 6, 2025, doi: 10.1109/MGRS.2025.3556532.

Authorized licensed use limited to: Central China Normal University. Downloaded on January 28,2026 at 12:43:44 UTC from IEEE Xplore. Restrictions apply.

[47] M. Jia et al., “Visual prompt tuning,” in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 709–727.

[48] J. Zhu et al., “MVP: Meta visual prompt tuning for few-shot remote sensing image scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5610413.

[49] L. Fang, Y. Kuang, Q. Liu, Y. Yang, and J. Yue, “Rethinking remote sensing pretrained model: Instance-aware visual prompting for remote sensing scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5626713.

[50] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, J. Li, and X. Mao, “EarthMarker: A visual prompting multimodal large language model for remote sensing,” *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5604219.

[51] H. Bahng, A. Jahanian, S. Sankaranarayanan, and P. Isola, “Exploring visual prompts for adapting large-scale models,” 2022, *arXiv:2203.17274*.

[52] Y.-Y. Tsai, C. Mao, Y.-K. Lin, and J. Yang, “Convolutional visual prompt for robust visual perception,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 27897–27921.

[53] C. Xu et al., “Progressive visual prompt learning with contrastive feature re-formation,” *Int. J. Comput. Vis.*, vol. 133, no. 2, pp. 511–526, Feb. 2025.

[54] S. Wang et al., “Personalized multiparty few-shot learning for remote sensing scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4506115.

[55] Y. Liu, Y. Huang, S. Liu, Y. Zhan, Z. Chen, and Z. Chen, “Open-set video-based facial expression recognition with human expression-sensitive prompting,” in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 5722–5731.

[56] Y. Liu et al., “Sample-cohesive pose-aware contrastive facial representation learning,” *Int. J. Comput. Vis.*, vol. 133, no. 6, pp. 3727–3745, Jun. 2025.

[57] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2010, pp. 270–279.

[58] G.-S. Xia et al., “AID: A benchmark data set for performance evaluation of aerial scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[59] G. Cheng, J. Han, and X. Lu, “Remote sensing image scene classification: Benchmark and state of the art,” *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[60] X. Sun, Z. Yang, C. Zhang, K.-V. Ling, and G. Peng, “Conditional Gaussian distribution learning for open set recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13477–13486.

[61] Z. Yue, T. Wang, Q. Sun, X.-S. Hua, and H. Zhang, “Counterfactual zero-shot and open-set visual recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15404–15414.

[62] K. Chen, B. Chen, C. Liu, W. Li, Z. Zou, and Z. Shi, “RSMamba: Remote sensing image classification with state space model,” *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[63] J. Ma, W. Jiang, X. Tang, X. Zhang, F. Liu, and L. Jiao, “Multiscale sparse cross-attention network for remote sensing scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5605416.

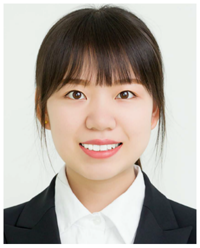
[64] S. Roy and A. Etemad, “Consistency-guided prompt learning for vision-language models,” in *Proc. 12th Int. Conf. Learn. Represent.*, 2023.

Hao Sun (Member, IEEE) received the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2021.

He is currently a Lecturer with the School of Computer Science, Central China Normal University, Wuhan, Hubei, China. His research interests include computer vision, deep learning, and hyperspectral image analysis.

Hanlizi Chen received the B.Eng. degree in computer science and technology from Central China Normal University, Wuhan, Hubei, China, in 2023, where she is currently pursuing the M.E. degree in computer science and technology with the School of Computer Science.

Her research interests include computer vision, remote sensing, and deep learning.



Wenjing Chen (Member, IEEE) received the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2021.

She is currently a Lecturer with the School of Computer Science, Hubei University of Technology, Wuhan, Hubei, China. Her research interests include deep learning and hyperspectral image processing.

Chengji Wang received the M.Sc. and Ph.D. degrees in computer science and technology from Xiamen University, Xiamen, China, in 2018 and 2022, respectively.

He is currently a Lecturer with the School of Computer Science, and Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, as well as the National Language Resources Monitoring and Research Center for Network Media, Central China Normal University, Wuhan, China. His research interests include multimodal learning and affective computing.

Wei Xie is a Professor with the School of Computer Science, Central China Normal University, Wuhan, Hubei, China. His research interests include image processing, computer vision, and deep learning.

Xiaoqiang Lu (Senior Member, IEEE) received the Ph.D. degree in signal and information processing from Dalian University of Technology, Dalian, China, in 2010.

He is currently a Full Professor with the College of Physics and Information Engineering, Fuzhou University, Fuzhou, China. His research interests include intelligent optical sensing, pattern recognition, machine learning, and hyperspectral image analysis.

Authorized licensed use limited to: Central China Normal University. Downloaded on January 28,2026 at 12:43:44 UTC from IEEE Xplore. Restrictions apply.