Class-Incremental Novel Category Discovery in Remote Sensing Image Scene Classification via Contrastive Learning

Yifan Zhou, Haoran Zhu, Chang Xu, Ruixiang Zhang, Guang Hua,
and Wen Yang, *Senior Member, IEEE*

***Abstract*—Remote sensing (RS) imagery captures the earth's ever-changing landscapes, reflecting evolving land cover patterns propelled by natural processes and human activities. However, existing RS scene classification methods mainly operate under a closed-set hypothesis, which stumbles when encountering novel emerging scenes. This article addresses the intricate task of RS scene classification without labels for novel scenes under incremental learning, termed class-incremental novel category discovery. We propose a contrastive learning-based novel category discovery pipeline tailored for RS image scene classification, enhancing the ability to learn unlabeled novel class data. Furthermore, within this pipeline, we introduce a positive pair filter to identify more positive sample pairs from novel classes, improving the feature representation capability on unlabeled data. Besides, our contrastive learning pipeline incorporates an old-feature replaying method to alleviate catastrophic forgetting in old classes. Extensive evaluations across three public RS datasets showcase the superiority of our method over state-of-the-art approaches.**

***Index Terms*—Contrastive learning, incremental learning, novel category discovery (NCD), remote sensing (RS), scene classification.**

I. Introduction

Deep neural networks have achieved remarkable results in remote sensing (RS) image scene classification [1], [2], [3], [4]. Existing research works are mostly performed under a closed-set hypothesis, where the test set shares the same classes with training images. However, due to the diverse earth environment and frequent human activities, aerial sensors are facing continuously emerging novel RS scenes, making this closed-set hypothesis hard to hold in practical use. Toward the automatic discovery of novel classes, the novel category discovery (NCD) task is proposed [5], [6], [7], [8], [9], which assumes that the closed-set training data are always available when learning novel class data. Nevertheless, for the continuously emerging aerial images that take up large storage and are of sensitivity risk, this task setting will be rarely feasible. Thus, we turn to the setting that combines the merits of incremental learning and NCD, termed class-incremental novel category discovery (class-iNCD) [10], to discover novel scenes for aerial images. Under this setting, there is no explicit existence of old class images at the stage of discovering unlabeled classes, and simultaneously, the model is expected to maintain the ability to classify the images of labeled classes (see Fig. 1, left).

For the class-iNCD setting, researchers leverage incremental learning techniques to avoid the catastrophic forgetting of old classes in the process of learning novel classes. More specifically, one main research line among them suggests providing additional supervision signals for novel class images by pseudo-labels obtained using the model trained on labeled classes, such as using the novel task classification head and Sinkhorn–Knopp algorithm [11]. However, due to the intraclass diversity and interclass similarity properties of RS image scene datasets [12], [13], the generated pseudo-labels can be severely noisy, leading to the difficulty in learning novel class data and overfitting old class data.

To tackle the above problems, we pursue class-iNCD for RS image scenes incorporating contrastive learning, termed the RS-ConNCD pipeline. First, we replace the pseudo-label generation with contrastive learning [14] for fine-tuning the feature extractor on unlabeled classes [15], [16], [17] (see Fig. 1, right) and replace the classification head with parameter-free clustering algorithms (i.e., K-means). This eliminates the need for generating noisy pseudo-labels on RS image scenes and classification heads, making it easier to learn novel unlabeled data and avoid the overfitting issue on old labeled data. Second, we find that the composition of contrastive learning's loss function misclassifies many positive pairs as negative pairs in the sample pair similarity measure during training. To avoid this, we propose to excavate positive pairs in contrastive learning via a newly designed dimension activation similarity based on the property of RS image scene datasets, providing more positive pairs to unlabeled novel class data and enhancing the corresponding feature representation ability. Third, we propose to use old-feature replaying to avoid catastrophic forgetting of old class data, which can also reduce

Manuscript received 16 January 2024; revised 27 March 2024; accepted 15 April 2024. Date of publication 19 April 2024; date of current version 1 May 2024. This work was supported in part by the National Natural Science Foundation of China (NSFC) Regional Innovation and Development Joint Fund under Grant U22A2010 and in part by the NSFC general program under Grant 62271355. *(Yifan Zhou and Haoran Zhu contributed equally to this work.) (Corresponding author: Wen Yang.)*

Yifan Zhou, Haoran Zhu, Chang Xu, Ruixiang Zhang, and Wen Yang are with the School of Electronic Information, Wuhan University, Wuhan 430072, China (e-mail: zhouyifan24@whu.edu.cn; zhuhaoran@whu.edu.cn; xuchangeis@whu.edu.cn; zhangruixiang@whu.edu.cn; yangwen@whu.edu.cn).

Guang Hua is with the Infocomm Technology Cluster, Singapore Institute of Technology, Singapore 138683 (e-mail: ghua@ieee.org).

Digital Object Identifier 10.1109/JSTARS.2024.3391512

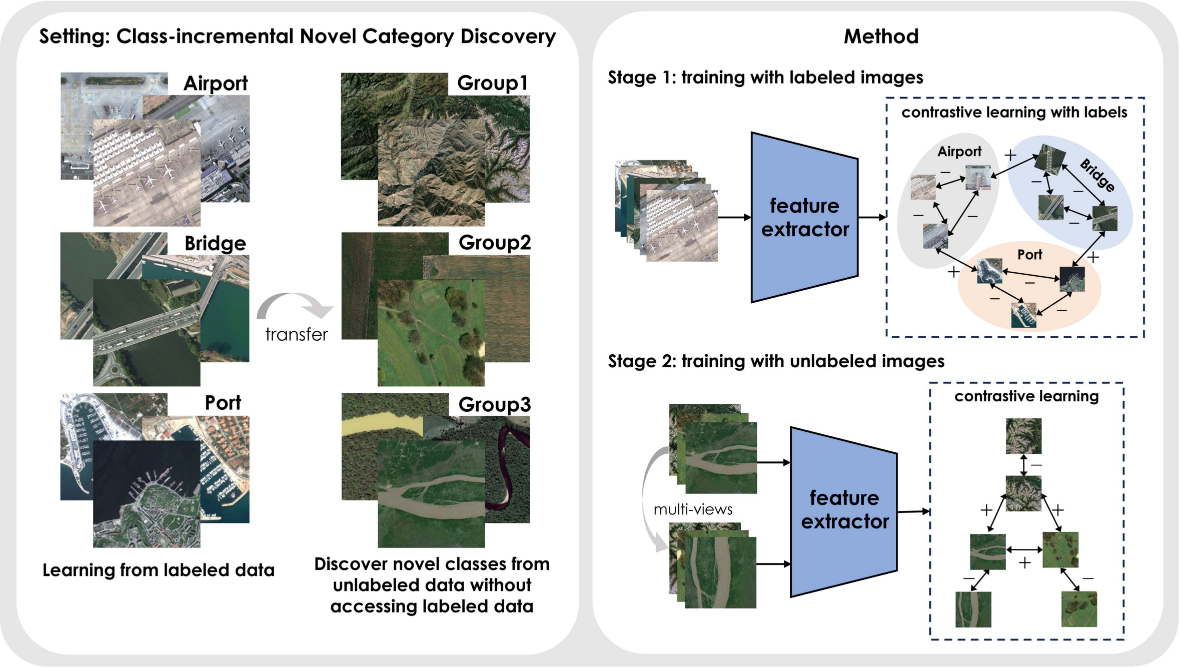

Setting: Class-incremental Novel Category Discovery
Airport
Bridge
Port
Group1
Group2
Group3
transfer
Learning from labeled data
Discover novel classes from unlabeled data without accessing labeled data
Method
Stage 1: training with labeled images
feature extractor
contrastive learning with labels
Airport
Bridge
Port
Stage 2: training with unlabeled images
multi-views
feature extractor
contrastive learning

Fig. 1. Overview of the task setting and our proposed pipeline. Left: Illustration of class-iNCD setting. Right: The proposed contrastive learning pipeline. The "+" sign indicates an increase in the Euclidean distance between two features, while the "–" sign indicates a decrease in the distance. In Stage 1, we obtain positive and negative pairs with the help of ground truth labels and then let the model learn the semantic features of these pairs by contrastive learning. In Stage 2, we first obtain positive pairs by geometric transformation and the positive pair filter method. Then, we use them to guide the model to perform contrastive learning on unlabeled data.

the old class learning's negative impact on learning new class data.

In a nutshell, the contributions of this article are as follows.

1) We propose a contrastive learning pipeline named RS-ConNCD to cope with the class-iNCD task for RS images, which can avoid overfitting the old labeled data.
2) We introduce a positive pair filter to discover unlabeled novel classes and an old-feature replaying method to avoid catastrophic forgetting in old labeled classes.
3) We conduct extensive experiments on three RS image scene datasets to demonstrate the effectiveness of the proposed method. RS-ConNCD achieves superior performance compared to the existing state-of-the-art methods.

The rest of this article is organized as follows. Section II is devoted to related work. Section III elaborates on our proposed method. Section IV reports the experimental results compared with other methods and ablation studies. Section V provides a discussion of the advantages of the proposed method and possible directions for further improvement in the future. Finally, Section VI concludes this article.

II. Related Work

In this section, we will first summarize the progress of the RS image scene classification task. Following this, we will discuss techniques related to the open-set scene classification method proposed in this article, including incremental learning, NCD, and contrastive learning.

A. RS Image Scene Classification

RS image scene classification is a crucial research topic in RS. The goal of this task is to predict the class of given RS image scenes. Early literature in this field mainly relies on handcrafted features, including the information of texture, contour, color, and space [18], [19], [20], [21]. However, these methods often fail when the context of RS image scenes becomes complex. Furthermore, the middle-level feature methods use encoding techniques to extract high-level representations from local features, such as bag-of-visual-words (BoVW) [22] and the Fisher kernel [23]. These techniques are effective for RS image scene classification, but the global semantic information cannot be well described since image scenes are represented by handcrafted local features. Recently, with the rapid development of convolutional neural networks (CNNs), many related methods have been developed in RS image scene classification [1], [2], [3], [4]. CNNs have excellent global and local representation ability for complex scenes without additional artificial significantly facilitating RS image scene classification. However, these methods are trained on closed-set datasets with supervised learning, which is unsuitable for open-world situations. In this article, we study the class-iNCD task, which aims to enable models to automatically learn good representations of new data in open-world scenarios.

B. Incremental Learning

Incremental learning tries to solve the problems that arise in the online learning setting. The basic setting is that the model is trained on a sequence of tasks and only data from the current task are available. In this setting, models often suffer from catastrophic forgetting [24], which means that models are prone to lose the ability to classify the data from old tasks during the training of the current data. To solve this problem, the existing methods can be divided into three categories: the regularization-based method, the exemplar-based method, and the parameter-isolation method. The first one tries to avoid catastrophic forgetting by constraining the learning range of the model parameters [25], [26], [27], [28]. The second makes the model remember the old tasks by introducing information about old tasks when training on the new tasks, such as old exemplars or old features [29], [30], [31], [32], [33]. The final one avoids catastrophic forgetting by assigning different model parameters to different tasks [34], [35]. In this article, we follow the idea of incremental learning methods and design an old-feature replaying method under the frame of contrastive learning.

C. *Novel Category Discovery*

For the NCD task, researchers attempt to transfer the knowledge learned from labeled data to unlabeled data with novel classes so that the model can classify the novel class data without labels. According to different task constraints, the current mainstream methods can be divided into two subcategories. The first is that the labeled data are available when training on the unlabeled data. Some approaches propose to jointly train both labeled and unlabeled data, employing pseudo-label generation algorithms to annotate these unlabeled instances, thereby enabling their utilization in supervised learning [5], [7], [8]. In addition, some research pointed out that it may be better to train labeled and unlabeled data in a self-supervised way [6]. However, labeled data are often not permitted when training on unlabeled data due to privacy reasons or memory constraints. Thus, the second category is training unlabeled data without the presence of labeled data, which is the class-iNCD setting. Based on AutoNovel [5] and unified objective function (UNO) [8], some literature adopted regularization- and exemplar-based methods to avoid catastrophic forgetting at the feature extractor and classification head, respectively [10], [36]. Others used the unsupervised deep clustering method from deep embedded clustering (DEC) [37] to train the unlabeled data and tried to use regularization-based methods to avoid catastrophic forgetting [38]. In this article, we study RS image scene classification in the class-iNCD setting. Nevertheless, different from the previous methods, our method discards the linear classification head and only fine-tunes the feature extractor with contrastive learning to better learn unlabeled novel class data.

D. *Contrastive Learning*

Contrastive learning is one of the learning paradigms in self-supervised learning. It aims to enable models to learn descriptive and intelligible representations from unlabeled data. Since the data are not labeled, to identify similar inputs, it is common to generate variants of individual inputs using transformations that preserve semantics, such as geometric transformations. The variants of the inputs are referred to as positive pairs, and the samples with different categories are called negative pairs. Through this method, the features of positive pairs are brought closer, while the features of negative pairs are pushed further apart. The literature on contrastive learning suggests that the feature extractor, trained through contrastive learning, is adaptable to different downstream tasks [15], [16], [17], [39], [40], [41], which inspires us how to learn from old labeled data as well as novel unlabeled data. In this article, we will follow the paradigm of contrastive learning and try to make our model generalize better to novel unlabeled data.

III. Method

In this section, we elaborate on the proposed method. First, we define the problem and notations in Section III-A. Then, we give an overview of our framework in Section III-B. We introduce contrastive learning, positive pair filter, and old-feature replaying in Sections III-C–III-E, respectively. Finally, we introduce the total objective function in Section III-F.

A. *Problem Definition and Notations.*

In the setting of class-iNCD, the model $\mathcal{M}$ will be trained on the labeled data $\mathcal{D}_l = \{(\mathbf{x}_i, y_i) \sim P(\mathcal{X}|\mathcal{Y}_l)\}$ first. After that, the labeled data $\mathcal{D}_l$ are discarded, and the model needs to learn to classify unlabeled data $\mathcal{D}_u = \{(\mathbf{x}_i) \sim P(\mathcal{X}|\mathcal{Y}_u)\}$. Since the category sets of labeled and unlabeled data are disjoint, i.e., $\mathcal{Y}_l \cap \mathcal{Y}_u = \emptyset$, we can also refer to the labeled data as the old class data and the unlabeled data as the novel class data.

B. *Overall Framework*

The overall framework is shown in Fig. 2. In Stage 1, we train the model with labeled data $\mathcal{D}_l$. We follow the paradigm of contrastive learning and calibrate the infoNCE loss [15] with the ground truth labels. After training, we reserve M old class features output by the third layer of the feature extractor and obtain C^l category center prototypes from $\mathcal{D}_l$ by calculating the feature mean for each old class.

In Stage 2, the model is trained on unlabeled data $\mathcal{D}_u$. Following the training paradigm in the previous NCD literature [5], [10], the first three layers of the feature extractor Φ^l_{1-3} are frozen. Then, we screen the positive pairs with the positive pair filter in Section III-D and conduct contrastive learning with infoNCE loss [15]. Meanwhile, the old class features cached in Stage 1 are fed into the fourth layer of Φ_l and the new feature extractor Φ_u, respectively, pulling their outputs closer through the method in Section III-E. Notice that before training the features with contrastive learning, we will first map the features into a higher dimensional space through a projector. The projector can accelerate convergence, improving the model's ability to transfer between different data [42], and is then discarded after training. After training, C^u category center prototypes are extracted from $\mathcal{D}_u$ by the K-means algorithm. C^u is the number of unlabeled data categories. Since C^u is unknown, we determine the value of C^u by the elbow method [43]. Thus, we can obtain $C^l + C^u$ category center prototypes from $\mathcal{D}_l \cup \mathcal{D}_u$. During inference, the input image is assigned to the class whose prototype is closest to it.

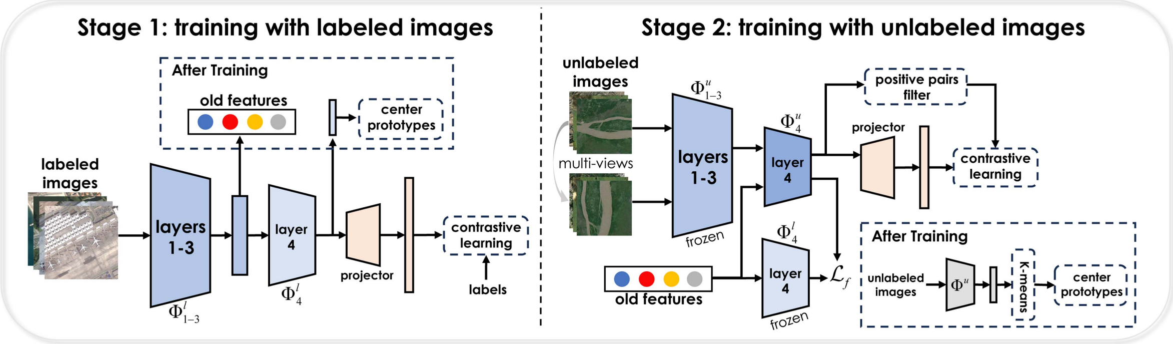


Fig. 2. Overall workflow of RS-ConNCD. Left: The feature extractor is trained on the labeled old class data through contrastive learning with labels. Old class features and center prototypes are stored after training. Right: The unlabeled novel class data are learned with contrastive learning. Forgetting old classes is prevented by the old-feature replaying method. After training, the novel class center prototypes are generated through K-means.

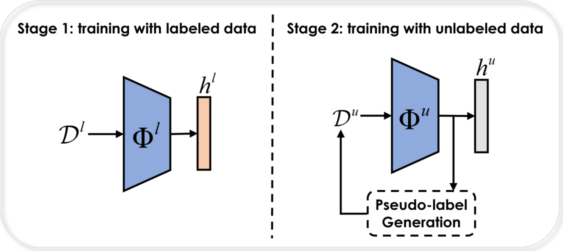


Fig. 3. Previous learning paradigm of NCD. In the first stage, the model trains with labeled classes. In the second, the model first generates pseudo-labels on unlabeled images and then supervises the learning of novel classes.

C. Contrastive Learning

According to the NCD literature [5], [24], [36], [44], the model is usually considered as the concatenation of a feature extractor Φ and a classification head h, as shown in Fig. 3. Based on this figure, in Stage 2, the feature extractor Φ^u, which is initialized with the parameters of Φ^l from Stage 1, is connected with a brand new classification head h^u, which is trained with the novel class data and the corresponding pseudo-labels, and then, the feature extractor is fine-tuned by gradient backpropagation.

However, as shown in Fig. 4, RS image scene datasets exhibit intraclass diversity and interclass similarity, resulting in severe noise in generated pseudo-labels. This can make it difficult to learn unlabeled data, as the classification head is susceptible to noisy labels [45]. Meanwhile, in Stage 1, the model is trained on the old class data with ground truth, which makes the model perform much better on the old class data than on the new class data. We note this as the "overfitting on the old class data."

In this article, we address this issue through contrastive learning mainly considering two properties: 1) contrastive learning has been widely used as pretraining to achieve robust representations for different data [14], [46], [47] and 2) it directly trains the feature extractor without labels while still allowing the model to learn features with good class discrimination [39], [48], [49], [50]. Therefore, we discard the classification head and no longer generate pseudo-labels, but directly fine-tune the feature extractor with contrastive learning.

In detail, we perform a geometric transformation of $\mathcal{D}_u$ to the input image first, so the unlabeled data will be $\mathcal{D}_u = \{(\mathbf{x}, \mathbf{x}')\}$. For a batch of unlabeled training data, we use infoNCE loss to enable Φ^u to learn good semantic representation [15]. This loss is given by

$$\mathcal{L}_u^{\text{infoNCE}} = -\sum_{i \in B} \log \frac{\exp\left(\Phi^u(\mathbf{x}_i) \cdot \Phi^u(\mathbf{x}'_i)/\tau\right)}{\sum_n \mathbf{1}_{[n \neq i]} \exp\left(\Phi^u(\mathbf{x}_i) \cdot \Phi^u(\mathbf{x}_n)/\tau\right)} \tag{1}$$

where B is the batch size, $\mathbf{1}_{[n \neq i]}$ is an indicator determining whether these two features come from the same image, and τ is the temperature coefficient. The key insight is to create positive pairs through geometric transformation and default to different images under a mini-batch as negative pairs. Then, the model is forced to make the features of positive pairs close to each other and the features of negative pairs far away from each other in the feature space. After training, we can use clustering algorithms like K-means to cluster the output features to obtain the final classification results.

As for labeled data $\mathcal{D}_l$, we can also train it with infoNCE loss. However, since the labels are available, they can be used to generate positive and negative pairs. Therefore, we opt to leverage the well-established contrastive loss with labels, as described in [51], which is given by

$$\mathcal{L}_l^{\text{infoNCE}} = -\sum_{i \in B} \sum_{j \in \mathcal{S}(i)} \log \frac{\exp\left(\Phi^l(\mathbf{x}_i) \cdot \Phi^l(\mathbf{x}_j)/\tau\right)}{\sum_n \mathbf{1}_{[n \neq i]} \exp\left(\Phi^l(\mathbf{x}_i) \cdot \Phi^l(\mathbf{x}_n)/\tau\right)} \tag{2}$$

where $\mathcal{S}(i)$ is the set of images under the same mini-batch that belong to the same category as $\mathbf{x}_i$.

D. Positive Pair Filter

When we analyze the contrastive learning mechanism based on infoNCE loss, it is found that the positive pairs in a mini-batch are not well utilized during training. Take Fig. 5 as an example; the positive pairs toward feature $\mathbf{z}_1$ in this mini-batch should be $\{(\mathbf{z}_1, \mathbf{z}'_1), (\mathbf{z}_1, \mathbf{z}_4)\}$. However, it can be seen that the pair $(\mathbf{z}_1, \mathbf{z}_4)$ is treated as a negative pair, resulting from the drawbacks in

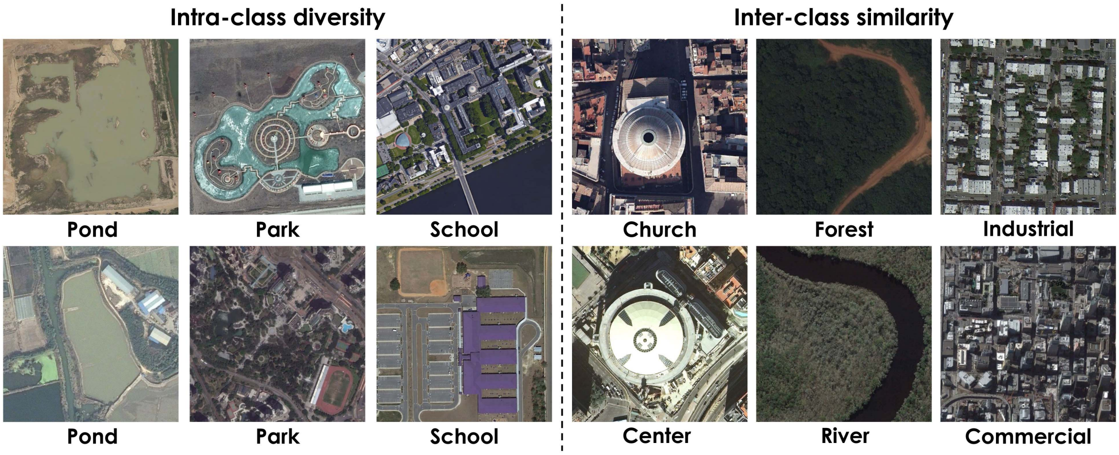


Fig. 4. Properties of RS image scene datasets. The figures on the left of the dotted line show images of intraclass diversity, and the figures on the right show interclass similarity.

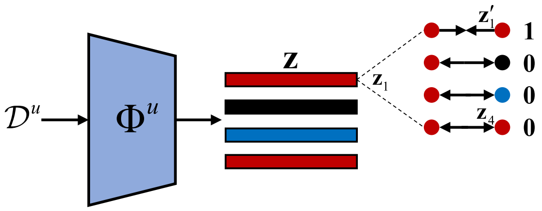


Fig. 5. Feature embeddings in a mini-batch under the infoNCE loss. Different colors denote different classes.

the contrastive learning setting, where different images even of the same class in a mini-batch are considered as negative samples. Therefore, the performance on unlabeled data can be further improved if the model can identify more correct positive pairs.

In this article, we propose to measure whether two features are positive pairs by their semantic region coincidence. Since the feature extractor Φ^u is initialized with the parameters of Φ^l from Stage 1, we hypothesize that it obtains feature generalization ability to some extent from the beginning of training. Furthermore, because of the existence of interclass similarity in RS image scenes, knowledge acquired from old class data can be effectively applied to new class data. This implies that utilizing the encoded information within the model's output features enables the identification of some genuine positive pairs, even in the absence of prior exposure to the novel class data by the model. We consider that each dimension of the output feature represents a high-level semantic contained in the original image. The semantic set represented by the dimension with the highest activation value should be the main semantic composition of this image. Therefore, we can determine whether two features belong to the same category according to the degree of overlap of their semantic sets. The determination mechanism is given by

$$r_{i,j} = \begin{cases} 1, & \frac{|\mathrm{top}_k(\mathbf{z}_i) \cap \mathrm{top}_k(\mathbf{z}_j)|}{k} > \eta \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $\mathbf{z}$ is the output feature from Φ^u, and $\mathrm{top}_k(\mathbf{z})$ is the set of dimension indices in $\mathbf{z}$ with the top k activation values. The function $|\cdot|$ computes the size of the given set. η is the threshold to determine whether two features are from the same class. Based on (3), we can modify the infoNCE loss from (1), which is given by

$$\mathcal{L}_u^{\text{infoNCE\_}r} = -\sum_{i\in B}\sum_{j\in\mathcal{R}(i)} \log \frac{\exp\left(\Phi^u(\mathbf{x}_i)\cdot\Phi^u(\mathbf{x}_j)/\tau\right)}{\sum_n \mathbf{1}_{[n\neq i]}\exp\left(\Phi^u(\mathbf{x}_i)\cdot\Phi^u(\mathbf{x}_n)/\tau\right)} \tag{4}$$

where $\mathcal{R}(i) = \{j | r_{i,j} = 1\}$, and $\mathbf{x} \in \mathcal{D}_u$. The model can learn more semantic representations from the same category through this method, which is similar to weak supervision.

E. Old-Feature Replaying

When dealing with class-iNCD tasks, catastrophic forgetting often occurs because the old class data are not available when training the novel class data. One main approach in the NCD literature is constraining the learning range of the feature extractor Φ^u for novel class data, which is fulfilled through the loss function given by

$$\mathcal{L}_f = \sum_i \left\|\Phi^l\left(\mathbf{x}_i\right) - \Phi^u\left(\mathbf{x}_i\right)\right\|_2, \quad \mathbf{x}_i \in \mathcal{D}_u \tag{5}$$

where Φ^l is the old feature extractor and it remains frozen during the training of the novel class data. However, this conflicts with fully learning the novel class data. To address this issue, we propose to decouple and avoid forgetting from learning the novel class data. We note that, first, when training the novel class data in Stage 2, only the parameters of the last layer of Φ^u need to be updated, and second, although the old class data cannot appear when training the new class data, it is feasible to preserve some of their features [10], [29]. Accordingly, we propose to use old-feature replaying to avoid catastrophic forgetting. Take ResNet-50 [52] as an example; we can save the M old features $\mathbf{f}$ for each class output by the third layer of the feature extractor Φ_l after the first stage of training. Then, input the saved old class features into the fourth layer of Φ_l and the new feature extractor Φ_u, respectively, when training the new class data in Stage 2, and encourage their outputs to be similar. The details are shown

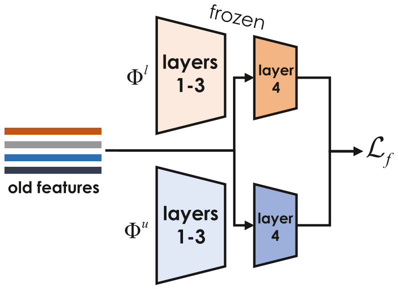


Fig. 6. Old-feature replaying method. This is a method proposed to avoid forgetting.

in Fig. 6. Therefore, (5) can be modified to

$$\mathcal{L}_f = \sum_i \left\|\Phi_4^l\left(\mathbf{f}_i\right) - \Phi_4^u\left(\mathbf{f}_i\right)\right\|_2 \tag{6}$$

where Φ_4 is the fourth layer of Φ. In this way, we decouple and avoid forgetting from the novel class data learning task, preventing them from interacting with each other.

F. Learning Objectives

In this section, we introduce the objective loss function of the proposed method. In Stage 1, the feature extractor Φ^l is trained on label data $\mathcal{D}_l$ by infoNCE loss in (2), which is

$$\mathcal{L}_{\text{Stage1}} = \mathcal{L}_l^{\text{infoNCE}}. \tag{7}$$

In Stage 2, the model is trained on unlabeled data $\mathcal{D}_u$. The loss functions in (4) and (6) are adopted to learn novel class data and prevent forgetting old class data, respectively. Moreover, since the category center prototypes of $\mathcal{D}_l$ are preserved, the model can learn to pull the features of the novel class data away from these center prototypes in the feature space to better distinguish the novel class data from the old class data. This is fulfilled through the loss given by

$$\mathcal{L}_c = -\frac{1}{C^l}\sum_i^B\sum_j^{C^l}\left\|\mathbf{z}_i - \mathbf{c}_j^l\right\|_2 \tag{8}$$

where $\mathbf{z}_i = \Phi^u(\mathbf{x}_i), \mathbf{x}_i \in \mathcal{D}_u$, and $\mathbf{c}_j^l$ is the category center prototype of class j in $\mathcal{D}_l$. Therefore, the total loss function in Stage 2 is

$$\mathcal{L}_{\text{Stage2}} = \mathcal{L}_u^{\text{infoNCE\_}r} + \alpha\mathcal{L}_f + \lambda\mathcal{L}_c. \tag{9}$$

IV. Experiments

In this section, we experiment on three datasets to verify the effectiveness of the proposed method and then verify the effectiveness of each component of the method by ablation studies.

A. Experimental Setup

1) Datasets and Implementation Details: In our experiments, we adopt three public RS image scene classification datasets, i.e., AID [53], Million-AID [54], and NWPU-RESISC45 [19]. The AID dataset contains 10 000 images in total with 30 scene classes. The NWPU-RESISC45 dataset consists of 45 scene classes and contains 31 500 images. Million-AID is a large dataset with 51 classes and 1 000 848 images in total. We adopt ResNet-50 [52] without the fully connected layer as the feature extractor and use a three-layer multilayer perceptron (MLP) as the projector. The parameters η and k in (3) are set to 0.5 and 100, respectively. The number of old features M for each old class in old feature replaying is set to 40. The parameters α and λ in (9) are set to 1 for all the experiments. The batch size during training is 128. All the models are trained for 150 epochs on the old class data and for 200 epochs on the novel class data.

2) Evaluation Metrics: In Stages 1 and 2, we obtain category center prototypes for labeled and unlabeled data, respectively. During inference, we assign classes to input samples by calculating the distances of their features from these prototypes, and the classification accuracy is measured by

$$ACC = \max_{p\in\mathcal{P}(\mathcal{Y})}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{y_i = p\left(\hat{y}_i\right)\right\} \tag{10}$$

where N is the size of the test set and $\mathcal{P}(\mathcal{Y})$ is the set of all permutations of the class labels in the test set. y and $\hat{y}$ are the ground truth labels and the model's predictions, respectively. We use the Hungarian optimal assignment algorithm to compute the maximum over the set of permutations [55].

B. Main Results

Since there is no existing method dealing with the class-iNCD task for RS image scene classification, we implement state-of-the-art methods of NCD on RS image scene datasets for comparison, including UNO [8], AutoNovel [5], GCD [6], ResTune [38], and FRoST [10]. Experiments are performed with two different class ratios (the number of old classes: the number of novel classes) settings, i.e., 2:1 and 1:2. Under each class split ratio setting, we randomly select the corresponding number of categories as unlabeled novel classes and subsequently conduct experiments to obtain results. This process is repeated three times, and the corresponding experimental outcomes are averaged to yield the final experimental results for the respective method.

Results under these two settings are shown in Tables I and II, respectively. The column "IL" indicates whether the method employs incremental learning. The columns "Old" and "Novel" refer to the classification accuracy of the model on old class (labeled) and novel class (unlabeled) data, respectively, and the column "All" gives the average classification accuracy over the entire dataset. According to these tables, the methods without incremental learning suffer from catastrophic forgetting of old class data, such as UNO, AutoNovel, and GCD. ResTune and FRoST utilize incremental learning mechanisms to prevent catastrophic forgetting, achieving good results on old class data. However, these two methods adopt the supervised paradigm in Fig. 3 to learn unlabeled novel class data, suffering from overfitting old class data. The classification accuracy gap between old and novel class data exceeds 11% for all three datasets. In contrast, the proposed RS-ConNCD, which utilizes contrastive learning to acquire knowledge of novel class data, produces a more balanced learning outcome for old and novel class data

TABLE I
CLASSIFICATION ACCURACY RESULTS ON THREE RS IMAGE SCENE DATASETS WITH 2:1 CLASS PARTITIONING

Methods	IL	AID			NWPU-RESISC45			Million-AID		
		Old	Novel	All	Old	Novel	All	Old	Novel	All
UNO [8]	✗	0.134	0.734	0.364	0.144	0.698	0.329	0.122	0.618	0.253
AutoNovel [5]	✗	0.288	0.663	0.436	0.258	0.657	0.391	0.190	0.585	0.309
GCD [6]	✗	0.345	0.814	0.523	0.350	0.770	0.490	0.257	0.710	0.386
ResTune [38]	✓	**0.859**	0.611	0.761	**0.853**	0.561	0.756	0.698	0.464	0.642
FRoST [10]	✓	0.815	0.701	0.769	0.813	0.673	0.766	0.715	0.609	0.687
RS-ConNCD	✓	0.844	**0.841**	**0.842**	0.821	**0.805**	**0.816**	**0.730**	**0.721**	**0.728**

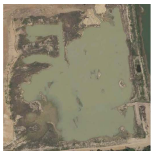
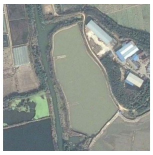
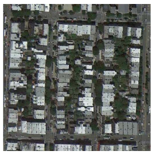
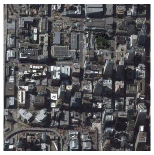
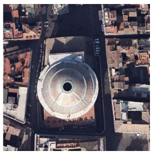
The entity in bold, such as RS-ConNCD, indicates that this is our proposed method.

TABLE II
CLASSIFICATION ACCURACY RESULTS ON THREE RS IMAGE SCENE DATASETS WITH 1:2 CLASS PARTITIONING

Methods	IL	AID			NWPU-RESISC45			Million-AID		
		Old	Novel	All	Old	Novel	All	Old	Novel	All
UNO [8]	✗	0.206	0.664	0.536	0.178	0.610	0.466	0.119	0.514	0.391
AutoNovel [5]	✗	0.353	0.549	0.491	0.227	0.565	0.468	0.188	0.489	0.396
GCD [6]	✗	0.402	0.798	0.673	0.349	0.727	0.601	0.288	0.638	0.542
ResTune [38]	✓	**0.886**	0.546	0.642	**0.856**	0.495	0.615	0.723	0.432	0.514
FRoST [10]	✓	0.859	0.639	0.695	0.851	0.616	0.694	**0.738**	0.494	0.561
RS-ConNCD	✓	0.863	**0.811**	**0.823**	0.833	**0.770**	**0.791**	0.734	**0.632**	**0.657**

The entity in bold, such as RS-ConNCD, indicates that this is our proposed method.

TABLE III
VISUALIZATION OF CLASSIFICATION RESULTS FROM THE AID DATASET

Images										
Methods	True label	Prediction	True label	Prediction	True label	Prediction	True label	Prediction	True label	Prediction
ResTune [38]	Pond	Farmland	Pond	Port	Dense residential	Industrial	Commercial	School	Church	Center
FRoST [10]	Pond	Farmland	Pond	Pond	Dense residential	Commercial	Commercial	Commercial	Church	Center
RS-ConNCD	Pond	Pond	Pond	Pond	Dense residential	Dense residential	Commercial	Commercial	Church	Center

The entity in bold, such as RS-ConNCD, indicates that this is our proposed method.

when compared to FRoST and ResTune. The gap in classification accuracy between old and novel class data is less than 9% for all three datasets. In addition, comparing the results in Tables I and II, it can be observed that the more old classes there are, the better the model performs on novel class data. This might result from the model's better generalization to novel class data when it has seen more categories.

Table III presents the visualization classification results of several RS scene images. In this table, the first and second images on the left are from the pond class, reflecting intraclass diversity. Both ResTune and FRoST misclassify the first image into another class, while our method provides the correct classification result. The third and fourth images belong to the dense residential and commercial classes, respectively, and embody interclass similarity. Neither FRoST nor ResTune could classify them correctly, while the proposed RS-ConNCD can differentiate them accurately. The last image is quite challenging as it is very similar to other classes. As a result, all the methods misclassified it.

C. Ablation Study

We examine the contributions of certain settings and modules of the proposed approach. Specifically, we determine the importance of the following elements in our experiments: the backbone selection, the proportion of novel to old class category splits, contrastive learning, positive pair filter, and old-feature replaying. Note that in Sections IV-C1 and IV-C3–IV-C5, our old-to-new class ratio is set at 2:1, and the experimental results are obtained from a single random sampling of the categories.

1) Effectiveness of the Backbone: To elucidate the impact of the backbone on the model's performance, we incorporated experiments with ResNet-18, ResNet-50, and ResNet-101 as backbones, as shown in Table IV. According to Table IV, it is evident that on the AID dataset, utilizing ResNet-50 as the backbone yields the best performance. Compared to ResNet-18, ResNet-50 possesses more trainable parameters and stronger feature extraction capabilities. In addition, in comparison to ResNet-101, there is a slight performance improvement, which

TABLE IV
ABLATION STUDY OF THE CHOICE OF BACKBONES

Methods	AID			Million-AID		
	Old	Novel	All	Old	Novel	All
ResNet-18	0.842	0.803	0.818	0.713	0.671	0.705
ResNet-50	**0.869**	**0.834**	**0.846**	0.742	0.704	0.736
ResNet-101	0.867	0.831	0.844	**0.748**	**0.706**	**0.741**

The data in bold indicates that this data is the largest one in the corresponding column of the table.

TABLE V
ABLATION STUDY OF THE CLASS SPLIT RATIO

Ratios	AID		
	Old	Novel	All
2:1	0.844	**0.841**	**0.842**
1:1	0.852	0.827	0.831
1:2	**0.863**	0.811	0.823

The data in bold indicates that this data is the largest one in the corresponding column of the table.

TABLE VI
ABLATION STUDY OF THE PROPOSED OLD-FEATURE REPLAYING (REPLAY), POSITIVE PAIR FILTER (FILTER), AND PROJECTOR

Methods	AID			Million-AID		
	Old	Novel	All	Old	Novel	All
Ours w/o projector	0.826	0.778	0.802	0.719	0.661	0.697
Ours w/o filter	0.867	0.815	0.844	0.743	0.681	0.732
Ours w/o replay	**0.871**	0.776	0.829	**0.763**	0.654	0.710
Ours	0.869	**0.834**	**0.846**	0.742	**0.704**	**0.736**

The data in bold indicates that this data is the largest one in the corresponding column of the table.

may be attributed to the redundant parameters of ResNet-101 for medium-sized datasets like AID. Conversely, on larger datasets, such as Million-AID, ResNet-101 exhibits the best performance.

2) Effectiveness of the Proportion of Category Splits: To further investigate the impact of varying old-to-novel class proportions on the model, we conducted experiments using the RS-ConNCD with AID dataset splits at ratios of 2:1, 1:1, and 1:2. The results are summarized in Table V. As shown in the table, a decrease in the number of old classes corresponds to a decrease in the model's classification accuracy on unlabeled novel class data. This observation can be attributed to the model's propensity to overfit the old classes, consequently leading to a weakened generalization performance on novel class data as the number of old classes declines.

3) Effectiveness of Contrastive Learning: Based on the experience of previous literature [42], before performing contrastive learning, we use an MLP as the projector to map features into a higher dimensional space for better learning of old and novel class data. We conduct experiments on contrastive learning without using the projector, and the results are shown in the row "Ours w/o projector" of Table VI. The results demonstrate that the absence of the projector results in varying degrees of decreased classification accuracy for both old and novel class data, indicating the projector's effectiveness.

To further validate the effectiveness of contrastive learning in expressing features for unlabeled novel class data, we conduct visualization experiments for the feature aggregation of ResTune, FRoST, and the proposed RS-ConNCD. RS-ConNCD uses contrastive learning to train on unlabeled novel class data, while ResTune and FRoST rely on supervised learning with pseudo-labels. We map the high-dimensional features into a 2-D space by t-SNE [56], where one point in the 2-D space corresponds to the high-dimensional feature output of the model for one sample. The experimental results of the visualization are shown in Fig. 7. We observe that the features produced by ResTune and FRoST only form a broad cluster in the 2-D space. The features belonging to different categories have a significant degree of overlap in the space, and the features of the same category are scattered in the space. In contrast, RS-ConNCD produces features that form distinct clusters in the 2-D space. The features belonging to the same class are clustered together in the space, and the features belonging to different classes are well separated.

4) Effectiveness of Positive Pair Filter: To verify the effectiveness of the proposed positive pair filter, we replaced the modified infoNCE loss in (4), which incorporated the positive pair filter, with the original infoNCE loss in (1). In Table VI, the row "Ours w/o filter" shows the model's classification accuracy results on the novel class data after removing the positive pair filter. The model's classification accuracy decreased by approximately 2% without it. These results verify the effectiveness of the positive pair filter in enhancing the performance of contrastive learning.

In addition, there are two hyperparameters η and k in the positive pair filter, where η is the threshold to determine whether two features are from the same class, and k means that the locations of the top k highest activation values of the features are selected for similarity measurement. To reveal the influence of these two hyperparameters on the similarity decision, we conducted the following experiments. First, the novel class data are input into the model trained in Stage 1, and the corresponding feature pair outputs are obtained. Then, we use the proposed method to determine the similarity of these pairs and compare the determined results with the real results to obtain the similarity determination accuracy of the proposed method.

The results on the AID dataset are shown in Fig. 8. The green line "ppf" is the curve of the positive pair filter. In the left figure, when η decreases from 1 to 0, the accuracy of similarity determination first increases and then decreases. This is because in this process, the model gradually identifies some positive feature pairs, and the corresponding accuracy of similarity determination increases. However, when the threshold is too low, many negative pairs will be identified as positive pairs, resulting in a decrease in the accuracy of similarity determination. In the right figure, when k is relatively small, the threshold for similarity determination can be easily met, which leads to a large number of negative pairs being wrongly considered positive pairs. As k increases, more and more negative pairs are correctly identified and the accuracy of similarity determination improves. However, when k exceeds 100, many positive pairs are wrongly considered negative pairs due to the difficulty in meeting the

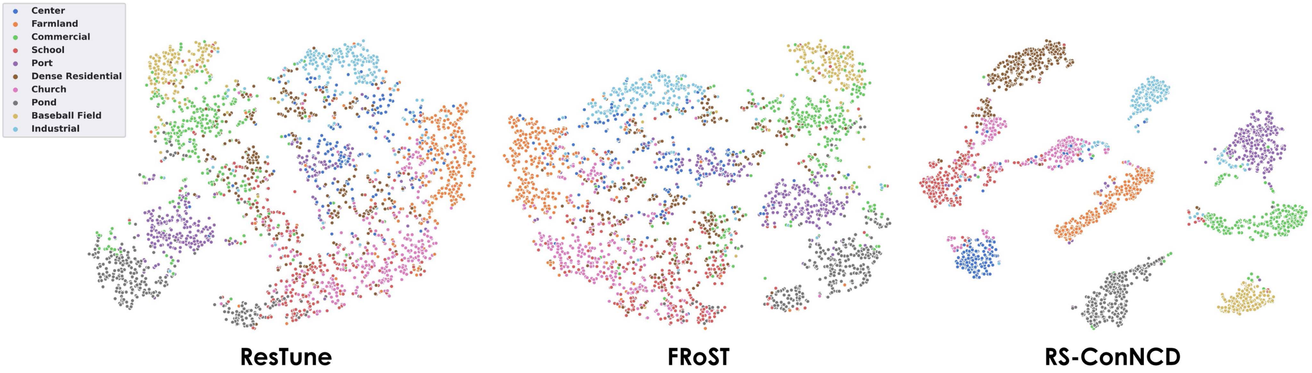
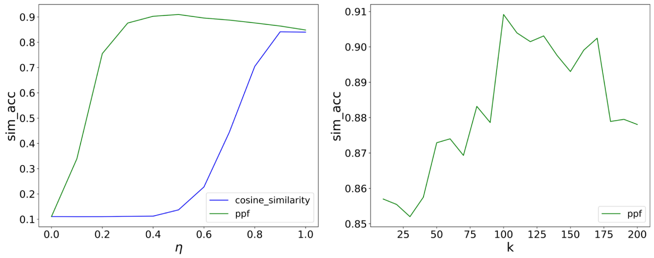


Fig. 7. Feature visualization of different methods. t-SNE visualization of unlabeled novel class instances in the AID dataset for features generated by ResTune [38], FRoST [10], and RS-ConNCD (our approach).



(a) (b)

Fig. 8. Effects of hyperparameters. “ppf” indicates the proposed positive pair filter, while “cosine_similarity” indicates the cosine similarity-based method. (a) Influence of the value of η on the accuracy of similarity determination. The value of k is set to 100. (b) Influence of the value of k on the accuracy of similarity determination. The value of η is set to 0.5.

threshold for similarity determination, leading to a decrease in the accuracy of similarity determination.

Furthermore, we also experiment with the widely used cosine similarity, and the result is shown in the blue curve “cosine_similarity” in the left figure. It can be observed that when the value of η is 1, most of the feature pairs are deemed negative pairs. However, as η decreases, the accuracy of similarity determination continues to decrease, which means that this metric fails to identify the real positive pairs.

5) Effectiveness of Old-Feature Replaying: In this article, we introduce an old-feature replaying module designed to mitigate the forgetting of old class data. This module disentangles the learning of novel class data from the prevention of forgetting old class data, thereby enabling the model to better capture the features of novel class data while maintaining classification performance on old class data. To demonstrate the effectiveness of the proposed old-feature replaying mechanism, we replace it with a conventional method that confines the learning of novel class data to prevent the forgetting of old class data [10], [38]. The results are presented in the row labeled “Ours w/o replay” in Table VI. Employing the conventional method to prevent catastrophic forgetting yields a slight improvement in the classification accuracy of the model for old class data. However, it significantly impedes the model’s learning process for novel class data, resulting in a decrease of approximately 5–6% in classification accuracy.

TABLE VII
EFFECT OF THE VALUE OF SELECTED OLD CLASS FEATURES ON THE AVOIDANCE OF CATASTROPHIC FORGETTING

Number of old class features per class	5	10	20	30	40	50	60
Accuracy on old class data	0.498	0.631	0.753	0.842	0.869	0.871	0.872

Finally, Table VII illustrates how the number of selected old class features impacts the ability to avoid forgetting old class data. It can be observed that with the increase in the number of old class features per class, the classification accuracy of old class data increases. The accuracy tends to be relatively stable when the number of old class features reaches 40.

V. Discussion

Class-iNCD aims to enable unsupervised learning from unseen data in an open-world scenario while retaining knowledge of previous class data, a critical requirement for RS applications like scene interpretation in unknown environments. The ability to autonomously comprehend unfamiliar environments enhances RS performance in real-world scenarios. This study explores the application of class-iNCD to RS image scene classification, addressing specific challenges encountered during this process.

One significant challenge is the poor quality of pseudo-labels, influenced by intraclass diversity and interclass similarity in RS scene images, leading to suboptimal performance of previous methods [10], [38] on novel class data. In response, we propose a contrastive learning-based class-iNCD method that directly finetunes the feature extractor, bypassing the need for pseudo-label generation. In addition, we introduce an old-feature replaying method to mitigate catastrophic forgetting. Experimental results on three RS datasets demonstrate the effectiveness of our approach in learning unlabeled novel class data while preserving the knowledge of old class data, achieving a balanced learning paradigm.

Despite these advancements, there are potential areas for improvement. Our method still encounters challenges in accurately recognizing certain highly confusable scene images, as demonstrated in Table III. To address this issue, we propose potential enhancements from three perspectives: First, the methodology employed in this study falls within the domain of representation learning, where the quality could be further refined through advanced geometric data augmentations [50] and the utilization of multiple local crops [57]. Second, since the performance of contrastive learning relies on the quantity of positive and negative feature pairs, it is feasible to enhance the precision of contrastive learning through techniques aimed at increasing the number of feature pairs, such as utilizing a large batch size [17] or implementing the negative pair sampling mechanism [58]. In addition, the learning performance of the model is significantly influenced by the interclass similarity and intraclass diversity in RS scene images. Future research could explore the generation of hard positive and negative pairs through certain generative methods [7], followed by contrastive learning using the triplet loss [59], to mitigate the impact of interclass similarity and intraclass diversity on model performance.

Besides, the ability to learn novel class data is influenced by the transferability of knowledge between old and novel class data. However, challenges arise in scenarios with a substantial gap between old and novel class data, such as discovering novel categories in infrared image data when the model is trained on optical aerial images. Investigating novel class discovery among different modal data is valuable for future research. Moreover, despite the diverse nature of RS image scenes, many can be categorized into broad groups (e.g., a baseball field and stadium as sports land). Incorporating text information about these broad categories and leveraging language-image models for discovering new categories [60] could potentially lead to further gains.

VI. Conclusion

In this article, we address the challenging task of class-iNCD for RS image scene classification. Leveraging the unique characteristics of RS image scenes, we present a novel framework grounded in contrastive learning named RS-ConNCD. Departing from conventional methodologies, RS-ConNCD discards the classification head and pseudo-label generation, opting instead for the direct fine-tuning of the feature extractor. This strategic refinement aims to establish a more balanced learning model for both novel and existing class data. In addition, we introduce a positive pair filter to reveal more informative pairs, thereby augmenting the learning capacity for novel class data. Moreover, we integrate an old-feature replaying method to mitigate the risk of catastrophic forgetting. The experimental evaluations on three public RS image scene datasets demonstrate the advantages of our proposed framework over several state-of-the-art methods. Our work provides valuable insights into the incremental discovery of novel categories in RS scene images, contributing to the ongoing development of precise and reliable methods for uncovering novel categories in understanding RS images.

Acknowledgment

The authors would like to express their sincere thanks to the anonymous reviewers for their valuable comments and suggestions. The numerical calculations in this article have been done on the supercomputing system in the Supercomputing Center, Wuhan University.

References

[1] G. Cheng, X. Xie, J. Han, L. Guo, and G. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.

[2] E. Othman, Y. Bazi, F. Melgani, H. Al-Hichri, N. Alajlan, and M. Zuair, "Domain adaptation network for cross-scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4441–4456, Aug. 2017.

[3] C. Shi, T. Wang, and L. Wang, "Branch feature fusion convolution network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5194–5210, 2020.

[4] K. Qi, C. Yang, C. Hu, Y. Shen, and H. Wu, "Deep object-centric pooling in convolutional neural network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7857–7868, 2021.

[5] K. Han, S. Rebuffi, S. Ehrhardt, A. Vedaldi, and A. Zisserman, "AutoNovel: Automatically discovering and learning novel visual categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6767–6781, Oct. 2022.

[6] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Generalized category discovery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7482–7491.

[7] Z. Zhong, E. Fini, S. Roy, Z. Luo, E. Ricci, and N. Sebe, "Neighborhood contrastive learning for novel class discovery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10862–10870.

[8] E. Fini, E. Sangineto, S. Lathuilière, Z. Zhong, M. Nabi, and E. Ricci, "A unified objective for novel class discovery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9264–9272.

[9] K. Han, A. Vedaldi, and A. Zisserman, "Learning to discover novel visual categories via deep transfer clustering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8400–8408.

[10] S. Roy, M. Liu, Z. Zhong, N. Sebe, and E. Ricci, "Class-incremental novel class discovery," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 317–333.

[11] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. 27nd Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2292–2300.

[12] S. Chen, Q. Wei, W. Wang, J. Tang, B. Luo, and Z. Wang, "Remote sensing scene classification via multi-branch local attention network," *IEEE Trans. Image Process.*, vol. 31, pp. 99–109, 2022.

[13] R. Du, G. Wang, N. Zhang, L. Chen, and W. Liu, "Domain adaptive remote sensing scene classification with middle-layer feature extraction and nuclear norm maximization," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2448–2460, 2024.

[14] R. Balestriero et al., "A cookbook of self-supervised learning," 2023, *arXiv:2304.12210*.

[15] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, 2018, *arXiv:1807.03748*.

[16] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 9929–9939.

[17] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[18] Y. Zhong, Q. Zhu, and L. Zhang, "Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6207–6222, Nov. 2015.

[19] 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.

[20] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[22] 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.*, 2010, pp. 270–279.

[23] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.

[24] M. D. Lange et al., "A continual learning survey: Defying forgetting in classification tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3366–3385, Jul. 2022.

[25] Z. Li and D. Hoiem, "Learning without forgetting," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 614–629.

[26] P. Dhar, R. V. Singh, K. Peng, Z. Wu, and R. Chellappa, "Learning without memorizing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5138–5146.

[27] W. Zhao, R. Peng, Q. Wang, C. Cheng, W. J. Emery, and L. Zhang, "Life-long learning with continual spectral-spatial feature distillation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5544214.

[28] L. Zhao et al., "Continual learning for remote sensing image scene classification with prompt learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6012005.

[29] 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.

[30] N. Ammour, "Continual learning using data regeneration for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8012805.

[31] X. Ma, K. Ji, S. Feng, L. Zhang, B. Xiong, and G. Kuang, "Open set recognition with incremental learning for SAR target classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5106114.

[32] S. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5533–5542.

[33] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara, "Dark experience for general continual learning: A strong, simple baseline," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 15920–15930.

[34] Y. Wu et al., "Large scale incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 374–382.

[35] X. Lu, X. Sun, W. Diao, Y. Feng, P. Wang, and K. Fu, "LIL: Lightweight incremental learning approach through feature transfer for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5611320.

[36] K. J. Joseph et al., "Novel class discovery without forgetting," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 570–586.

[37] J. Xie, R. B. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 478–487.

[38] Y. Liu and T. Tuytelaars, "Residual tuning: Toward novel category discovery without labels," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 7271–7285, Oct. 2023.

[39] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput.Vis.*, 2021, pp. 9630–9640.

[40] R. Ni, M. Shu, H. Souri, M. Goldblum, and T. Goldstein, "The close relationship between contrastive learning and meta-learning," in *Proc. Int. Conf. Learn. Represent.*, 2022.

[41] M. Bi, M. Wang, Z. Li, and D. Hong, "Vision transformer with contrastive learning for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 738–749, 2023.

[42] F. Bordes, R. Balestriero, Q. Garrido, A. Bardes, and P. Vincent, "Guillotine regularization: Improving deep networks generalization by removing their head," 2022, *arXiv:2206.13378*.

[43] R. L. Thorndike, "Who belongs in the family?," *Psychometrika*, vol. 18, pp. 267–276, 1953.

[44] C. Yeh, C. Hong, Y. Hsu, T. Liu, Y. Chen, and Y. LeCun, "Decoupled contrastive learning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 668–684.

[45] L. Feng, S. Shu, Z. Lin, F. Lv, L. Li, and B. An, "Can cross entropy loss be robust to label noise?," in *Proc. 29th Int. Jont Conf. Artif. Intell.*, 2020, pp. 2206–2212.

[46] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 15637–15648.

[47] P. Goyal et al., "Vision models are more robust and fair when pretrained on uncurated images without supervision," 2022, *arXiv:2202.08360*.

[48] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15979–15988.

[49] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9643–9653.

[50] J. Grill et al., "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[51] P. Khosla et al., "Supervised contrastive learning," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[53] 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.

[54] Y. Long et al., "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, 2021.

[55] H. W. Kuhn, "The Hungarian method for the assignment problem," in *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*. Berlin, Germany: Springer, 2010, pp. 29–47.

[56] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

[57] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 9912–9924.

[58] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.

[59] H. Wang et al., "Triplet-metric-guided multi-scale attention for remote sensing image scene classification with a convolutional neural network," *Remote. Sens.*, vol. 14, no. 12, 2022, Art. no. 2794.

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

Yifan Zhou received the B.S. degree in electronic information engineering in 2021 from Wuhan University, Wuhan, China, where he is currently working toward the M.S. degree in communication and information systems.

His research interests include computer vision, remote sensing scene classification, and novel category discovery.

Haoran Zhu received the B.S. degree in electronic information engineering in 2023 from Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree in communication and information system.

His research interests include computer vision and remote sensing image tiny object detection.

Chang Xu received the B.S. degree in electronic information engineering in 2021 from Wuhan University, Wuhan, China, where he is currently working toward the M.S. degree in communication and information systems.

His research interests include geolocalization, aerial object detection, and label noise.

Ruixiang Zhang received the B.S. degree in electronic engineering in 2019 from Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree in communication and information systems.

His research interests include remote sensing image processing, including label-efficient object detection and cross-modal object detection.

Guang Hua received the B.Eng. degree in communication engineering from Wuhan University, Wuhan, China, in 2009, and the Ph.D. degree in information engineering from Nanyang Technological University (NTU), Singapore, in 2014.

He was a Research Fellow with the School of Electrical and Electronic Engineering, NTU, from 2015 to 2016, an Associate Professor with the School of Electronic Information, Wuhan University, from 2017 to 2022, and an International Scholar Exchange Fellow with Yonsei University, Seoul, South Korea, sponsored by the CHEY Institute for Advanced Studies, Seoul, South Korea, from 2020 to 2021. From 2013 to 2015, he was a Scientist with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore, where he was also a Senior Scientist from 2022 to 2023. He is currently an Associate Professor with the Infocomm Technology Cluster, Singapore Institute of Technology, Singapore. He was the Principal Investigator for two China's National Natural Science Foundation of China projects and several industry projects. He has publications in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE/CVF Computer Vision and Pattern Recognition Conference, IEEE International Conference on Acoustics, Speech, and Signal Processing, and IEEE International Workshop on Information Forensics and Security. He holds a Singapore patent (licensed) and a Chinese patent (transferred) on media forensics. His research interests include artificial intelligence security, media security, and statistical signal processing.

Dr. Hua is an Associate Editor for IEEE SIGNAL PROCESSING LETTERS.

Wen Yang (Senior Member, IEEE) received the B.S. degree in electronic apparatus and surveying technology, the M.S. degree in computer application technology, and the Ph.D. degree in communication and information systems from Wuhan University, Wuhan, China, in 1998, 2001, and 2004, respectively.

In 2008 and 2009, he was a Visiting Scholar with the Apprentissage et Interfaces Team, Laboratoire Jean Kuntzmann, Grenoble, France. From 2010 to 2013, he was a Postdoctoral Researcher with the State Key Laboratory of Information Engineering, Surveying, Mapping, and Remote Sensing, Wuhan University, where he currently a Full Professor with the School of Electronic Information. His research interests include object detection and recognition, multisensor information fusion, and remote sensing image interpretation.