drones

MDPI

Article

Generalized Category Discovery in Aerial Image Classification via Slot Attention

Yifan Zhou [1], Haoran Zhu [1], Yan Zhang [1], Shuo Liang [2], Yujing Wang [2] and Wen Yang [1,*]

[1] School of Electronic Information, Wuhan University, Wuhan 430072, China; zhouyifan24@whu.edu.cn (Y.Z.); zhuhaoran@whu.edu.cn (H.Z.); zhangyan@whu.edu.cn (Y.Z.)
[2] The 54th Research Institution of CETC, Shijiazhuang 050081, China; liangs@nuaa.edu.cn (S.L.); afreeboat@whu.edu.cn (Y.W.)
* Correspondence: yangwen@whu.edu.cn

Abstract: Aerial images record the dynamic Earth terrain, reflecting changes in land cover patterns caused by natural processes and human activities. Nonetheless, prevailing aerial image classification methodologies predominantly function within a closed-set framework, thereby encountering challenges when confronted with the identification of newly emerging scenes. To address this, this paper explores an aerial image recognition scenario in which a dataset comprises both labeled and unlabeled aerial images, intending to classify all images within the unlabeled subset, termed Generalized Category Discovery (GCD). It is noteworthy that the unlabeled images may pertain to labeled classes or represent novel classes. Specifically, we first develop a contrastive learning framework drawing upon the cutting-edge algorithms in GCD. Based on the multi-object characteristics of aerial images, we then propose a **slot** attention-based **GCD** training process (**Slot-GCD**) that contrasts learning at both the object and image levels. It decouples multiple local object features from feature maps using slots and then reconstructs the overall semantic feature of the image based on slot confidence scores and the feature map. Finally, these object-level and image-level features are input into the contrastive learning module to enable the model to learn more precise image semantic features. Comprehensive evaluations across three public aerial image datasets highlight the superiority of our approach over state-of-the-art methods. Particularly, Slot-GCD achieves a recognition accuracy of 91.5% for known old classes and 81.9% for unknown novel class data on the AID dataset.

Keywords: aerial image classification; generalized category discovery; contrastive learning; slot attention

Citation: Zhou, Y.; Zhu, H.; Zhang, Y.; Liang, S.; Wang, Y.; Yang, W. Generalized Category Discovery in Aerial Image Classification via Slot Attention. *Drones* **2024**, *8*, 160. https://doi.org/10.3390/drones8040160

Academic Editor: Pablo Rodríguez-Gonzálvez

Received: 13 March 2024
Revised: 13 April 2024
Accepted: 15 April 2024
Published: 19 April 2024

1. Introduction

Deep neural networks have demonstrated excellent performance in aerial image classification [1–5]. However, existing methodologies rely on the closed-set assumption, meaning that the image categories in the test set must be a subset of the categories in the training set. This assumption often proves invalid in open-world scenarios, where the model encounters unlabeled aerial images that may include categories unseen during the training phase. This constitutes the core issue addressed in this paper: given a dataset of aerial images, where only a portion of the images are labeled, the remaining aerial images require the model to predict their categories (see Figure 1, left). Importantly, these categories for prediction may encompass some novel classes not present in the labeled dataset. This task, known as Generalized Category Discovery (GCD) [6], holds considerable potential applications in the field of Unmanned Aerial Vehicles (UAVs). For instance, it could facilitate autonomous scene comprehension by drones operating in unfamiliar environments.

In the GCD task, the model is designed to learn known old category data as well as previously unseen novel category data in the absence of labels. In response, some researchers advocate for discarding the classification head and directly training the feature extractor

using labeled and unlabeled data within a contrastive learning framework [6]. Others propose sharpening the model's category probability outputs to generate high-quality pseudo-labels for unlabeled data, thereby enhancing the classification performance of such data [7]. DCCL [8] decomposes the model's learning process into two levels: the conception level and the instance level. Then the model separately learns the feature representations of these levels through contrastive learning to obtain more precise features. PromptCAL [9] suggests storing historical features during the training process and comparing the current feature set with historical features. By adopting a KNN approach, pseudo-labels are assigned to each feature pair, thereby improving the accuracy of contrastive learning. Although the aforementioned methods have demonstrated promising outcomes in the GCD task, their application to aerial image recognition encounters significant challenges. Since models are required to learn novel category data without labels, the resulting image feature encoding inherently lacks precision. Furthermore, since most aerial images are scene images, they have multi-object characteristics. An object in a scene can be understood as a semantic abstraction, while an aerial image is composed of one or multiple semantics. Therefore, encoding aerial images solely from a global perspective would further amplify the noise in the final features. Additionally, the phenomenon of inter-class object sharing in aerial images can introduce substantial biases in the learned image features by the model.

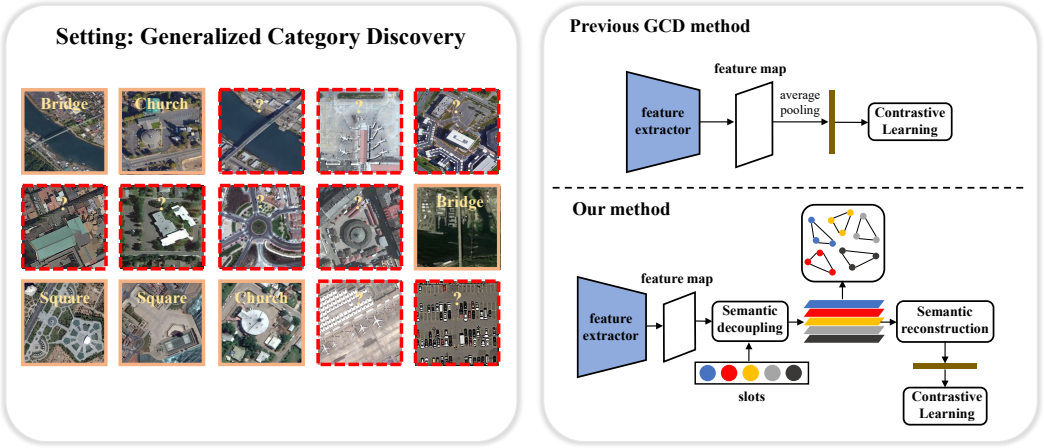


Figure 1. An overview of the task setting and our proposed pipeline. (**Left**): Illustration of GCD setting. The question mark and the red dashed box denote unlabeled samples, while the green solid line box denotes labeled samples. It is noteworthy that unlabeled data may also contain categories of labeled data. (**Right**): The previous GCD learning framework versus the proposed method in this paper.

To address the aforementioned issues, this paper introduces a learning framework tailored for the GCD task in aerial images, designed from both image-level and object-level perspectives. We design a slot-attention-based learning pipeline to learn both image-level and object-level features, termed Slot-GCD. The learning pipeline is shown in Figure 1. In contrast to the former approach, our method no longer performs average pooling on the feature maps. Instead, it decouples the feature maps into multiple semantic features using the slot attention mechanism, where each feature represents a specific object. Subsequently, contrastive learning is employed to learn these semantic features. Additionally, our method reconstructs the overall image feature by linearly combining the confidence weights of the feature map and the slot features. These reconstructed features are input into the contrastive learning module to learn the overall semantics. By these approaches, the model can learn more accurate features for aerial images.

The contributions of this paper are as follows:

(1) We develop a learning framework based on contrastive learning to address the GCD problem in aerial image classification.
(2) We introduce the slot attention mechanism within the framework of contrastive learning, allowing the model to explicitly learn local semantic features. This en-

hancement leads to an improvement in the precision of the features learned through contrastive learning.

(3) We conduct extensive experiments on three aerial image datasets to demonstrate the effectiveness of the proposed method. Slot-GCD achieves superior performance compared to the existing state-of-the-art methods.

The remainder of this paper is organized as follows: Section 2 is devoted to related work. Section 3 elaborates on our proposed method. Section 4 reports the experimental results compared with other methods and ablation studies. Section 5 provides a discussion of the advantages of the proposed method and possible directions for further improvement in the future. Finally, Section 6 concludes the paper.

2. Related Work

In this section, we will first summarize the progress of the aerial image classification task. Following this, we will discuss techniques related to the open-set classification method proposed in this paper, including open-set recognition, Generalized Category Discovery, contrastive learning, and slot attention.

2.1. Aerial Image Classification

Aerial image classification is a crucial research topic in the field of UAVs, opening up new possibilities in environmental monitoring. For example, techniques such as the extraction of the canopy shadow fraction using UAV images highlight the potential of leveraging bidirectional reflectance characteristics for a better understanding of vegetation structure. The primary objective of this task entails the accurate prediction of classes attributed to given aerial images. The early literature predominantly relies on manually crafted features encompassing texture, contour, color, and spatial information [10–13]. However, these conventional methods often falter in scenarios where the context of aerial images becomes intricate. Subsequently, middle-level feature approaches employ encoding methodologies to derive high-level representations from local features, exemplified by methods like Bag of Visual Words (BoVW) [14] and the Fisher kernel [15]. While these techniques have demonstrated effectiveness in aerial image classification, they are limited in their capacity to encapsulate global semantic information, owing to the reliance on handcrafted local features for representing images. With the recent advancements in convolutional neural networks (CNNs), a plethora of methodologies have emerged within the domain of aerial image classification [1–5]. CNNs exhibit commendable prowess in capturing both global and local representations of complex aerial images without necessitating additional manual intervention, thus significantly augmenting the efficacy of aerial image classification. However, these methodologies are typically trained on closed-set datasets utilizing supervised learning paradigms, rendering them ill-suited for open-world scenarios. In this study, we delve into the task of GCD for aerial image classification, aiming to empower models to autonomously acquire robust representations of novel class data in open-world settings.

2.2. Open-Set Recognition

The issue of open-set recognition (OSR), as outlined in [16], involves the classification of unlabeled instances from known semantic classes while identifying test instances from previously unseen categories. OpenMax [17] represents the pioneering deep learning approach to tackle this challenge using Extreme Value Theory. GANs are frequently leveraged to generate adversarial samples for training open-set classifiers, as observed in works such as [18,19]. Various methodologies have been proposed to train models to consider images with substantial reconstruction errors as indicative of open-set samples, as seen in [20]. Additionally, some techniques involve learning prototypes for labeled categories and utilizing their distances to identify images from unknown categories, as demonstrated in works like [21,22]. Furthermore, some researchers propose a joint training approach integrating a flow-based density estimator and a classification-based encoder for OSR [23].

2.3. Generalized Category Discovery

Similar to the OSR task, the Generalized Category Discovery (GCD) task also involves encountering unlabeled data from novel classes. However, unlike OSR, which solely requires the identification of novel class data from unlabeled samples, GCD additionally necessitates the classification of these novel class data. GCD is a branch of Novel Category Discovery (NCD). In the NCD task, the model encounters unlabeled data consisting only of novel class data [24–29]. With the deepening research into the NCD task, researchers have discovered that in practice, unlabeled data may contain both novel class and old class data, leading to the emergence of the GCD task. Initially, researchers attempted to simultaneously learn novel class and old class data through contrastive learning [6]. However, experimental results revealed that the learning effect of new class data was not satisfactory. SimGCD [7] improves upon contrastive learning by sharpening probability distributions to generate high-quality pseudo-labels, thereby enhancing the precision of the model's learned features. DCCL [8] proposes to conduct contrastive learning separately at two levels, the conception level and the instance level, to further improve the model's representational capacity. In addition, PromptCAL [9] continuously corrects pseudo-labels for feature pairs in contrastive learning based on historical features using a KNN approach, thereby mitigating errors caused by label absence. Unlike the aforementioned methods, this paper designs a slot-attention-based semantic disentanglement module tailored to the characteristics of aerial images, enabling the model to learn the semantics of multiple objects within these images.

2.4. Contrastive Learning

Contrastive learning represents a pivotal paradigm within the domain of self-supervised learning, striving to facilitate models in acquiring comprehensive and interpretable representations from unannotated data. In the absence of labels, a common approach involves the generation of diverse renditions of individual inputs through transformations that preserve semantic information [30,31], such as geometric transformations. These modified inputs, termed positive pairs, are contrasted against samples from distinct categories, termed negative pairs, to discern similarities. Through this mechanism, the features of positive pairs are drawn closer together, while those of negative pairs are pushed further apart. Other methodologies utilize triplet loss with active triplet selection to extract hard positive and hard negative pairs. These pairs can be derived either online from the current mini-batch or retrieved from a prior checkpoint, resembling the mechanisms observed in momentum networks [32,33]. Extensive research in contrastive learning suggests that feature extractors trained through this paradigm exhibit adaptability across various downstream tasks [34–36], thereby instigating exploration into learning from both historically labeled data and novel unlabeled data. In this study, we adhere to the tenets of contrastive learning, endeavoring to enhance the generalizability of our model to novel unlabeled data.

2.5. Slot Attention

Slot attention is a convolutional autoencoder module that employs a revised attention mechanism on latent vectors iteratively, aiming to derive a permutation-invariant collection of object-specific representations termed "slots". The concept of slot attention was initially introduced to address object-centric learning objectives [37], which pertains to a machine learning paradigm focused on objects and their interrelations within specific tasks. Slot attention utilizes a variant of dot-product attention where slots act as queries in a competitive process to interpret the encoder output. Initially, the efficacy of slot attention was only validated on a few simplistic handcrafted datasets. However, as research progressed, it has also been applied to numerous large-scale datasets [38], demonstrating its capability to enhance the generalization ability of pre-trained models on downstream tasks such as object detection, instance segmentation, and scene comprehension. Moreover, some researchers propose a classifier based on slot attention designed to offer transparent and accurate classification [39], providing intuitive interpretations and positive or negative

explanations for each category regulated by a custom loss function. In this paper, we leverage slot attention to enable the model to automatically learn the semantic features of multiple objects within an aerial image.

3. Method

In this section, we elaborate on the proposed method. First, we define the problem and notations in Section 3.1. Then, we give an overview of our framework in Section 3.2. We introduce contrastive learning and slot attention in Sections 3.3 and 3.4, respectively. Finally, we introduce the total objective function in Section 3.5.

3.1. Problem Definition and Notations

In the setting of GCD, the model $\mathcal{M}$ will be trained on the dataset $\mathcal{D}$, which comprises two parts, $\mathcal{D}_l = \{(\mathbf{x}_i, y_i) \sim P(\mathcal{X}|\mathcal{Y}_l)\}$ and $\mathcal{D}_u = \{(\mathbf{x}_i) \sim P(\mathcal{X}|\mathcal{Y}_u)\}$. During training, labels of $\mathcal{D}_u$ are unavailable. Since $\mathcal{Y}_l \in \mathcal{Y}_u$, we refer to the data belonging to $\mathcal{Y}_l$ as the old class data, and the data belonging to $\mathcal{Y}_u \setminus \mathcal{Y}_l$ as the novel class data.

3.2. The Overall Framework

The overall framework is shown in Figure 2. In this paper, we augment the contrastive learning framework of GCD with object-level contrastive learning based on slot attention. For a given aerial image, two different perspective samples, $\mathbf{x}$ and $\mathbf{x}'$, are obtained through geometric transformations such as cropping, rotation, and scaling. Subsequently, they are separately fed into the feature extractor to obtain corresponding feature maps, $\mathbf{f}$ and $\mathbf{f}'$. In conventional methods [6,7], a single feature vector is directly obtained by averaging pooling operations on the feature maps, followed by contrastive learning. However, in our learning framework, multiple slots are utilized to semantically disentangle the feature maps, akin to the mechanism found in cross attention in Transformer [40], which will be elaborated in Section 3.4. After semantic disentanglement, each pixel position on the feature map offers confidence scores for all corresponding slots. The slot with the highest confidence is then activated at that pixel position, thereby generating the corresponding semantic feature. All activated slots on this feature map constitute the fundamental semantics of the aerial image sample. In object-level contrastive learning, for two samples from the same image, semantic features from the same slot are pulled closer together in the feature space, while those from different slots are pushed further apart. For different images, their corresponding semantic features are all pushed apart. Additionally, based on the activation confidence and slot features, we reconstruct the image-level features of the aerial image via linear combinations and then learn these features through contrastive learning.

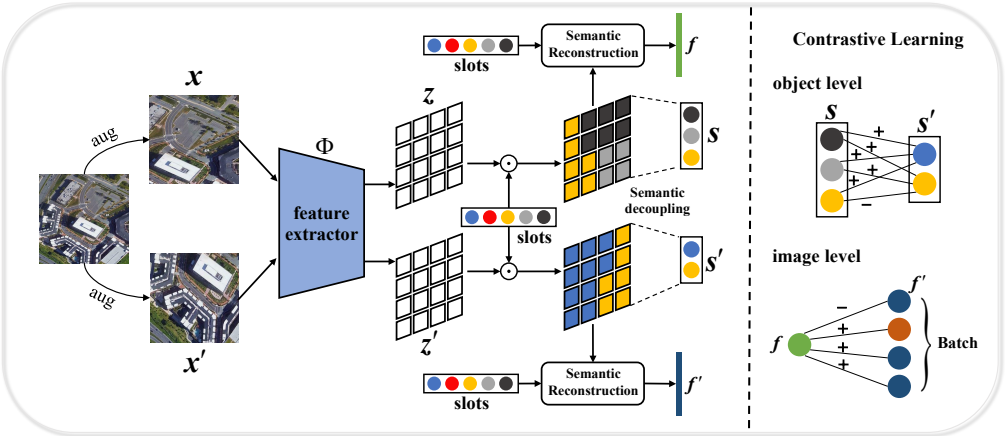


Figure 2. The overall workflow of Slot-GCD. (**Left**): Illustration of semantic decoupling and reconstruction based on slot attention. $\mathbf{x}$ and $\mathbf{x}'$are two different views from a single aerial image after geometric transformations. On thefeature map, different colors represent distinct semantic regions.

(**Right**): Image-level and object-level contrastive learning. **f** represents image-level features, while $\mathcal{S}$ represents a set of object-level features for an aerial image. The "+" sign indicates an increase in the Euclidean distance between two features, while the "−" sign indicates a decrease in the distance.

3.3. Contrastive Learning

When learning from unlabeled data, there are generally two learning methods. One involves assigning pseudo-labels to unlabeled data, followed by supervised learning. This can be achieved by either adding a classification head for novel classes [27] or employing an assignment algorithm like the Sinkhorn–Knopp algorithm [24,41]. Another method leverages the powerful representation capability of contrastive learning to directly learn from these unlabeled data in a self-supervised learning manner. Nevertheless, as depicted in Figure 3, aerial image datasets demonstrate intra-class diversity and inter-class similarity [42,43], leading to significant noise in the generated pseudo-labels. This noise can pose challenges in effectively learning from unlabeled data, particularly due to the susceptibility of the classification head to erroneous labels [44]. In contrast, contrastive learning possesses two pivotal characteristics that can effectively address the challenges above: (1) the extensive application of contrastive learning as a pre-training technique to yield resilient representations across diverse datasets [45–47]; and (2) its capacity to train the feature extractor directly without relying on labels, yet enabling the model to acquire features characterized by strong class discrimination [48–51]. Consequently, we opt to fine-tune the feature extractor directly using contrastive learning.

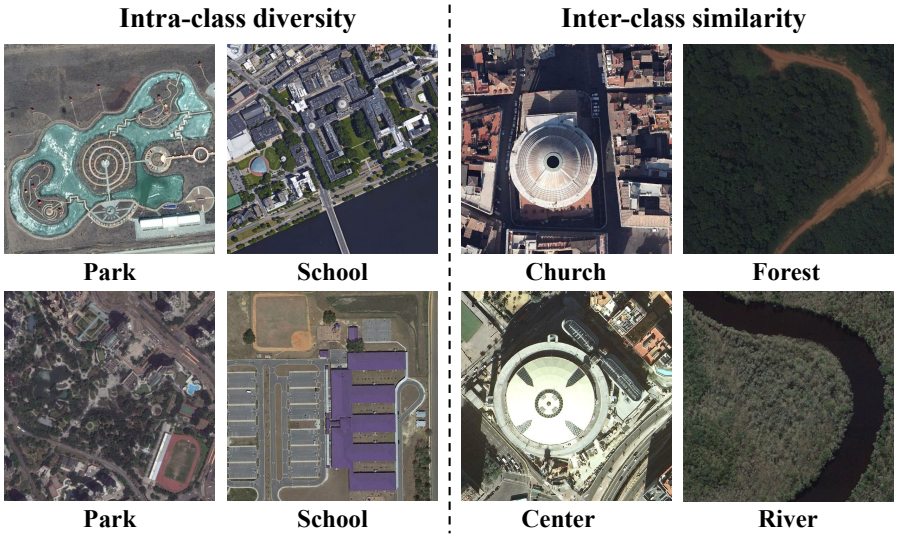


Figure 3. The properties of aerial image datasets. The figures on the left of the dotted line show images of intra-class diversity, and the figures on the right show inter-class similarity.

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 the feature extractor Φ to learn good semantic representation [30]. This loss is given by

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

where B is the mini-batch during training and M is the batch size for unlabeled data. $\mathbf{1}_{[n\neq i]}$ is an indicator determining whether these two features come from the same image, and τ is the temperature coefficient. It is noteworthy that the present study employs a conventional approach prevalent in contrastive learning methodologies. Before engaging in contrastive learning, features undergo projection into a higher-dimensional space with a

projector $\mathbf{p}$. This procedure serves to expedite convergence and bolster the representational efficacy of features [47]. The fundamental idea revolves around generating positive pairs via geometric transformations and contrasting them with different images within a mini-batch to serve as negative pairs. Subsequently, the model is compelled to minimize the distance between features belonging to positive pairs while maximizing the distance between features belonging to negative pairs within the feature space. Upon completion of training, clustering algorithms such as K-means can be employed to partition the resultant features and derive the ultimate classification outcomes.

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. Thus, the loss for learning labeled data is given by

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

where N is the batch size for labeled data and $\mathcal{I}(i)$ is the set of images under the same mini-batch that belong to the same category as $\mathbf{x}_i$.

3.4. Slot Attention

To enhance the precision of feature encoding for aerial images, this study proposes employing slot attention for contrastive learning at the object level. In aerial image datasets, many images comprise multiple types of objects, with certain objects appearing across various class categories, as depicted in Figure 4. Leveraging the characteristics of multi-object scenarios and inter-class object sharing, we advocate utilizing slot attention to disentangle semantic information in aerial images, enabling the model to learn semantic feature encodings for each object through contrastive learning. This approach facilitates a finer-grained understanding of aerial images. After obtaining semantic feature encodings for the objects constituting an aerial image, the entire image's features are reconstructed based on the confidence scores of these features on the feature map, thereby mitigating model bias towards individual objects.

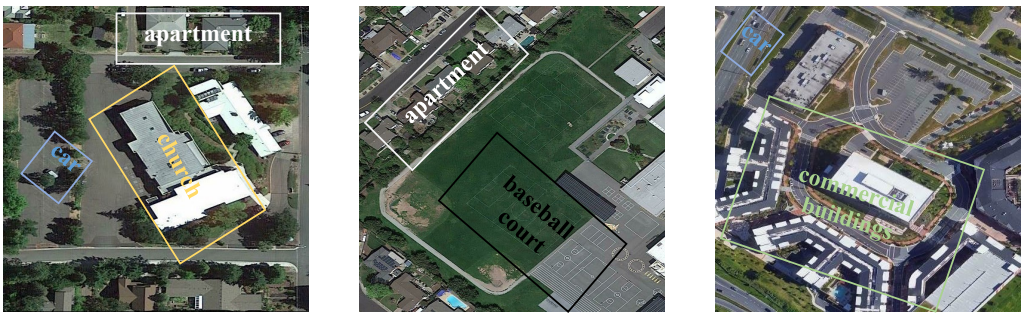


Figure 4. The multi-object nature and inter-class object sharing characteristics of aerial images. From these three different category images, it can be observed that each image contains multiple objects, and there are cases where the same object appears in aerial images of different categories, such as the presence of house objects in both the church and baseball field.

3.4.1. Semantic Decoupling

For a feature map of an aerial image, there may exist multiple objects. Drawing inspiration from the implementation mechanism of cross attention in Transformer [40], we introduce K slot vectors, representing potential objects (semantics) in the image. Subsequently, we calculate the attention of each pixel in the feature map on the slot vectors and compute confidence using softmax. The specific formula is given by

$$\mathcal{A} = \underset{K}{\text{softmax}}\left(\mathbf{z} \cdot \overline{\mathcal{S}}^{\top} / \tau\right) \in \mathbb{R}^{H \times W \times K}, \tag{3}$$

where $\mathbf{z} \in \mathbb{R}^{H \times W \times D}$ is the feature map, and $\overline{\mathcal{S}} \in \mathbb{R}^{K \times D}$ are the slot vectors. Since each slot represents a specific semantic feature or object, the slot vectors should be kept orthogonal to ensure non-overlapping semantics among slots. In the confidence matrix $\mathcal{A}$, the confidence scores in the pixel dimension represent the weights of the slots on each pixel. Combining these weights with the corresponding pixel features linearly yields the slot features based on this feature map. The calculation formula is given by

$$\mathcal{S} = \frac{1}{\sum_{i,j} \mathcal{A}[i,j]} \sum_{i,j} \mathcal{A}[i,j] \odot \boldsymbol{z}[i,j] \in \mathbb{R}^{K \times D} = \left[\mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^K\right], \tag{4}$$

where $\odot$ denotes the Hadamard product and $[i,j]$ represents the position index of the feature map. Since not all slots represent semantics present in the current aerial image, it is necessary to filter out slot features irrelevant to this aerial image. This can be achieved by retaining only the slot features with the highest confidence in the slot dimensions of the confidence matrix. The calculation method is given by

$$\mathbf{1}^k = \exists_{i,j} \quad \text{such that } \underset{K}{\text{argmax}}(\mathcal{A}[i,j]) = k, \quad \mathbf{1} = \left[\mathbf{1}^1, \mathbf{1}^2, \ldots, \mathbf{1}^K\right], \tag{5}$$

where $\mathbf{1}^k$ is a binary indicator representing whether the slot with index k is activated on the feature map. In this way, the feature map $\mathbf{z}$ can be decomposed into multiple local semantic features $\mathcal{S}$.

This paper enhances the model's representation capability of local semantic features through contrastive learning. The basic idea is that, within a batch, for different image samples from the same image, the activated slot features with the same index are pulled closer together in space, while those with different indices are pushed farther apart. The slot features from different images are also pushed farther apart in space. The specific calculation method is given by

$$\mathcal{L}_{slot}^{InfoNCE} = -\frac{1}{M} \sum_{i}^{M} \sum_{k} \log \frac{\exp\left(\mathbf{p}_s(\mathbf{s}_i^k) \cdot \mathbf{p}_s(\mathbf{s}_{i'}^k) / \tau\right)}{\sum_j \sum_{k'} \exp\left(\mathbf{p}_s(\mathbf{s}_i^k) \cdot \mathbf{p}_s(\mathbf{s}_j^{k'}) / \tau\right)}, \quad \mathbf{1}_i^k = \mathbf{1}_{i'}^k = \mathbf{1}_j^{k'} = 1, \tag{6}$$

where $\mathbf{p}_s$ refers to the projector mapping the slot features into a high-dimensional space.

3.4.2. Semantic Reconstruction

In aerial image classification tasks, the model ultimately needs to output a vector representing the entire aerial image. Directly average pooling the feature map would result in significant information loss. In this paper, the feature map is decomposed into multiple slot features. We can then utilize these features along with their confidence scores on the feature map to reconstruct the feature vector of the entire aerial image. The specific calculation is given by

$$\mathbf{f} = \frac{1}{\sum_{i,j} \mathcal{A}[i,j]} \sum_{i,j} \mathcal{A}[i,j] \mathbf{s}^k, \quad k = \underset{K}{\text{argmax}} \mathcal{A}[i,j]. \tag{7}$$

Since we use $\mathbf{f}$ to represent the semantics of the entire image, the image-level contrastive learning loss function (1) can be rewritten as

$$\mathcal{L}_u^{infoNCE} = -\frac{1}{M} \sum_{i \in B} \log \frac{\exp\left(\mathbf{p}(\mathbf{f_i}) \cdot \mathbf{p}(\mathbf{f}_i') / \tau\right)}{\sum_n \mathbf{1}_{[n \neq i]} \exp(\mathbf{p}(\mathbf{f}_i) \cdot \mathbf{p}(\mathbf{f}_n) / \tau)}, \tag{8}$$

where $\mathbf{f}_i'$ represents the global features of the image sample from another perspective.

3.5. Learning Objectives

In this section, we present the objective loss function employed in the proposed methodology. The overall loss function is given by

$$\mathcal{L} = \mathcal{L}_l^{infoNCE} + (1-\lambda)\mathcal{L}_u^{infoNCE} + \lambda\mathcal{L}_{slot}^{infoNCE}, \tag{9}$$

where λ represents the loss balancing weight controlling the learning at the object and image levels. The specific training procedure is illustrated in Algorithm 1.

Algorithm 1 The process of the proposed Slot-GCD

Input: The labeled dataset $\mathcal{D}_l$, the unlabeled dataset $\mathcal{D}_u$, the feature extractor Φ, the slot vectors $\overline{\mathcal{S}} \in \mathbb{R}^{K\times D}$, the feature projector $\mathbf{p}$, the slot projector $\mathbf{p}_s$, and the optimizer $\mathcal{O}$.
Output: optimal Φ and $\overline{\mathcal{S}}$.

```
Initialize Φ, S̄, p and p_s; Perform geometric transformations on D_l and D_u.
repeat
    Input D_l and D_u into Φ to obtain corresponding feature maps.
    Calculate the slot confidence matrix A using Equation (3).
    Use Equation (4) to obtain slot features S.
    Filter out irrelevant slot features by Equation (5).
    Compute the object-level contrastive learning loss by Equation (6).
    Reconstruct the overall image semantic features by Equation (7).
    Compute the image-level contrastive learning loss by Equation (8).
    Calculate the overall loss by Equation (9).
    Train the corresponding parameters with O.
until a preset number of training iterations is satisfied.
```

4. Experiment

In this section, we conduct experiments on three public aerial image datasets to validate the efficacy of the proposed method. Subsequently, we perform ablation studies to assess the effectiveness of each component of the method.

4.1. Experimental Setup

4.1.1. Datasets

This paper conducts experiments on three publicly available aerial image datasets: AID [52], Million-AID [53], and NWPU-RESISC45 [13]. The AID dataset, established by Wuhan University in 2017, comprises 30 distinct aerial image categories such as bridges, rivers, forests, grasslands, schools, factories, and other types. Each category contains 200–400 aerial images, totaling 10,000 images, with a resolution of 600 × 600. The Million-AID dataset, created by Wuhan University in 2021, consists of 51 categories, with images per category ranging from approximately 2000 to 45,000, totaling 1,000,848 images. The NWPU-RESISC45 dataset, provided by Northwestern Polytechnical University (NWPU), is widely employed in aerial image recognition. It comprises 45 distinct categories, each containing 700 images, resulting in a total of 31,500 images. Each image has a resolution of 256 × 256.

4.1.2. Evaluation Metrics

Upon the completion of model training, we obtain features corresponding to each image, which are then clustered using the K-means algorithm to derive the center prototype for each category. As the number of novel classes in the unlabeled data is unknown, we employ the classical elbow method [54] in unsupervised clustering to estimate the number of novel classes. The test set is inputted into the model during the testing phase to extract respective features. Following this, the Euclidean distances between these features and the category center prototypes in the training set are calculated, and subsequently, the category

of the nearest centroid is allocated to the test set sample. The classification accuracy is measured by

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

where T 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].

In this paper, ACC is computed from three perspectives, denoted as "All", "Old", and "Novel", respectively. "All" represents all samples in the test set. "Old" represents the samples in the test set whose labels belong to $\mathcal{Y}_l$, reflecting the model's classification performance on unlabeled old class data. "Novel" represents the samples in the test set whose labels belong to $\mathcal{Y}_u \setminus \mathcal{Y}_l$, reflecting the model's classification performance on unlabeled novel class data.

4.1.3. Implementation Details

This paper employs the first four layers of ResNet-50 [56] as the feature extractor, initialized with parameters pretrained on ImageNet. In slot attention, we configure the number of trainable slots differently for each dataset: 128 for AID, 256 for NWPU-RESISC45, and 1024 for Million-AID. Throughout training, these slots maintain orthogonality to each other. The projector used in contrastive learning follows the methodology outlined in DINO [51], utilizing three-layer Multi-Layer Perceptrons (MLPs). We initialize the learning rate to 0.1 and dynamically adjust it during training using the CosineAnnealingLR strategy from PyTorch. We employ the Layer-wise Adaptive Rate Scaling (LARS) optimizer [51]. Training is conducted with a batch size of 128 over 300 epochs. The weighting factor λ for the loss functions at the image and object levels is set to 0.5.

4.2. Main Results

Given the absence of methodologies specifically designed to address the GCD task within the context of aerial image classification, this paper applies several effective methods from both the NCD and GCD domains to aerial image classification, to compare them with the proposed method. These methods include RankStats+ [27], UNO+ [24], GCD [6], SimGCD [7] and DCCL [8]. RankStats+ and UNO+ are NCD algorithms adapted for the GCD task. GCD, DCCL, and SimGCD represent the state-of-the-art algorithms for the GCD task. The experimental datasets in this study are split into training and testing sets in a 4:1 ratio. Next, this paper will analyze the effectiveness of the proposed method from both quantitative and qualitative perspectives.

4.2.1. Quantitative Analysis

The study randomly selects a subset of classes from the training set, consisting of $\frac{1}{3}$ and $\frac{2}{3}$ of the total class proportion, to serve as unlabeled novel class data, while the remaining classes are designated as old class data. This is done to assess the model's sensitivity to the proportion of novel and old classes. During the training phase, 50% of the samples from the old class data are randomly selected, their labels are removed, and they are combined with the unlabeled novel class data to form the unlabeled data.

The quantitative analysis results of this study are shown in Tables 1 and 2. We conducted experiments using six algorithms, including the proposed method, on three aerial image datasets. Table 1 presents the experimental results under the setting where the ratio of old to novel categories is 2:1. RankStats+ [27] and UNO+ [24] denote NCD algorithms adapted for the GCD task. This adaptation was necessary as NCD algorithms initially do not possess the capability to handle GCD tasks, where all unlabeled data are assumed to belong to novel categories. Experimental results reveal that RankStats+ and UNO+ can learn meaningful representations from unlabeled data, especially UNO+, which

exhibits the best classification performance on old category data among the six methods. This is attributed to its parametric classifiers. However, a corresponding drawback is its suboptimal performance on novel category data. The proposed method Slot-GCD outperforms UNO+ by 8%, 6.3%, and 8.1% in classification performance on AID, NWPU-RESISC45, and Million-AID, respectively. This performance gap is even more pronounced in novel category data. GCD [6], DCCL [8], and SimGCD [7] represent current state-of-the-art GCD algorithms with good representation capabilities for both novel and old category data. However, they still generate the image feature encodings from a holistic perspective during contrastive learning, without fully exploiting the multi-object nature of partial aerial images. By utilizing slot attention for semantic disentanglement of aerial images and conducting contrastive learning at both the object and image levels, our proposed method Slot-GCD achieves better image feature representation, as evidenced by experimental results outperforming GCD, DCCL, and SimGCD on all three aerial image datasets. Table 2 presents the experimental results under the setting where the ratio of old to novel categories is 1:2. Under this setting, our method still achieves relatively good classification results. It is worth noting that, compared to Table 1, when the number of labeled categories decreases, the model's classification performance on unlabeled novel category data decreases, and the performance on labeled category data increases. This may be because the model's ability to classify unknown novel category data to some extent relies on the knowledge learned from labeled data. When the quantity of old class categories is limited, the model becomes susceptible to overfitting, thereby attenuating its capacity for generalizing to novel class data.

Table 1. Classification accuracy results (in %) on three aerial image datasets with 2:1 class partitioning. "Old" represents the classification results on the old class data; "Novel" represents the classification results on the novel class data; "All" represents the classification results on the entire test set.

Methods	AID			NWPU-RESISC45			Million-AID		
	Old	Novel	All	Old	Novel	All	Old	Novel	All
RankStats+ [27]	61.5	32.8	45.5	58.7	30.2	43.9	48.6	21.5	32.6
UNO+ [24]	**94.6**	63.7	75.2	**93.2**	60.4	74.3	**82.1**	51.3	64.7
GCD [6]	88.3	72.2	76.9	86.9	69.5	75.1	76.3	59.9	65.8
DCCL [8]	88.7	79.4	80.5	87.5	76.2	78.5	77.3	65.4	69.2
SimGCD [7]	91.0	81.2	82.6	88.4	77.9	79.5	78.6	68.3	70.5
Slot-GCD	91.5	**81.9**	**83.2**	89.3	**79.2**	**80.6**	79.9	**71.2**	**72.8**

Table 2. Classification accuracy results (in %) on three aerial image datasets with 1:2 class partitioning. "Old" represents the classification results on the old class data; "Novel" represents the classification results on the novel class data; "All" represents the classification results on the entire test set.

Methods	AID			NWPU-RESISC45			Million-AID		
	Old	Novel	All	Old	Novel	All	Old	Novel	All
RankStats+ [27]	62.2	27.8	34.2	59.1	26.9	34.9	53.2	20.6	27.4
UNO+ [24]	**94.8**	60.5	67.3	**94.1**	56.9	64.3	**82.9**	48.4	54.2
GCD [6]	89.7	70.3	71.6	87.2	65.9	68.1	79.6	54.2	57.5
DCCL [8]	89.1	76.5	78.5	88.4	72.5	74.5	80.2	62.9	63.6
SimGCD [7]	93.4	77.8	80.4	90.3	73.2	76.1	82.3	65.1	66.4
Slot-GCD	92.9	**80.7**	**81.6**	91.4	**75.4**	**78.3**	82.6	**68.5**	**68.1**

4.2.2. Qualitative Analysis

To better validate the effectiveness of the proposed method in aerial images, we compare the classification results of GCD [6], DCCL [8], SimGCD [7], and the proposed method Slot-GCD on four aerial images labeled as "apartment", "church", "parking_lot" and "commercial_area". The specific details are shown in Figure 5. Observing the first and

second images, it can be noted that although their categories are different, they share many common local objects such as cars and apartments. This characteristic may lead to the inability of the model's image feature encoding to effectively express the overall semantics in the absence of labels, even leaning towards a specific local object. For instance, GCD and DCCL both misclassify the second aerial image, possibly due to their failure to adequately learn the key local object: the church. Alternatively, they may have leaned too much towards features similar to apartments during image feature encoding. In contrast, Slot-GCD correctly identifies the first and second images. It is noteworthy that SimGCD also correctly classifies these two aerial images, possibly because its well-designed high-quality pseudo-label generation mechanism accurately distinguishes the differences between the two images. The last two aerial images also exhibit the characteristics of multi-object and inter-class object sharing, such as a large number of cars appearing in both images. GCD, DCCL, and SimGCD fail to output entirely correct results for these two images, while Slot-GCD correctly identifies them.

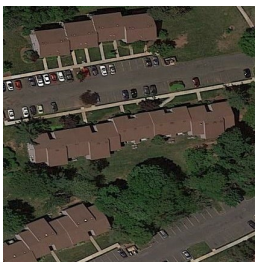
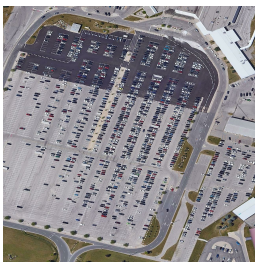
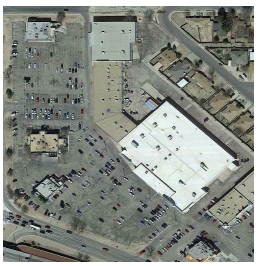
Images								
Methods	GT	Pred	GT	Pred	GT	Pred	GT	Pred
GCD	apartment	golf-course	church	detached_house	parking_lot	works	commercial_area	commercial_area
DCCL	apartment	apartment	church	apartment	parking_lot	parking_lot	commercial_area	parking_lot
SimGCD	apartment	apartment	church	church	parking_lot	parking_lot	commercial_area	parking_lot
Slot-GCD	apartment	apartment	church	church	parking_lot	parking_lot	commercial_area	commercial_area

Figure 5. Visualization of classification results for 4 aerial image samples. "GT" represents the ground truth label, while "Pred" denotes the model's predicted label.

Furthermore, because the classification performance of GCD heavily relies on the discriminative nature of the features outputted by the model's feature extractor in the spatial domain, the quality of the learned features can be assessed by observing the intra-cluster aggregation and inter-cluster discrimination of the feature clusters in space. Given the high-dimensional nature of the model's output features, visualizing them directly is impractical, necessitating their projection into a lower-dimensional space for analysis. In this study, we employ the t-SNE method [57] for dimensionality reduction to visualize the features extracted by models trained on the AID dataset using various methods, including GCD, DCCL, SimGCD, and Slot-GCD.

Figure 6 presents the visualization results of the features outputted by models trained on the AID dataset using the aforementioned methods. From this figure, it can be observed that the features in the right half of the four feature clusters exhibit good intra-cluster aggregation and inter-cluster discrimination. This is because the unlabeled data includes some old class data, and these class features have been well learned by the model during the training phase using labeled old class data. Conversely, upon observing the left half of the feature clusters, it is evident that all four feature clusters exhibit varying degrees of

intra-cluster feature dispersion and inter-cluster feature intersection. This phenomenon reflects the aggregation effect of features from unlabeled new class data. Notably, GCD demonstrates the poorest aggregation effect, with the feature points in the left half nearly coalescing, resulting in low inter-cluster discrimination. Conversely, the left half of the feature clusters from DCCL and SimGCD exhibit some distinct boundaries of class feature clusters, albeit with considerable overlap. In comparison, the feature clusters generated by Slot-GCD demonstrate higher inter-cluster discrimination, indicating that the proposed method can yield a feature extractor with notably superior category representation performance.

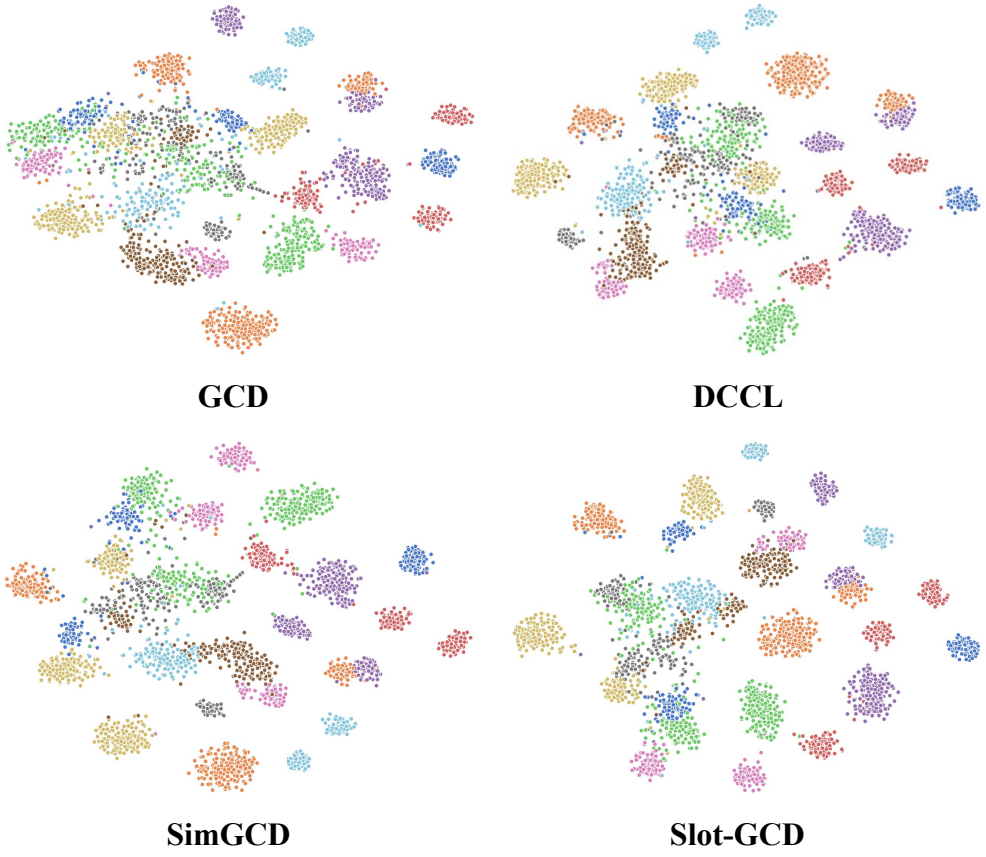


Figure 6. Feature visualization of different methods. t-SNE visualization of unlabeled instances in the AID dataset for features generated by GCD [6], DCCL [8], SimGCD [7], and Slot-GCD (our approach). Feature points of different colors represent they belong to different categories.

4.3. Ablation Study

4.3.1. Effectiveness of Each Component

To validate the effectiveness of each module in Slot-GCD, this paper conducted ablation experiments and analysis. The specific classification accuracy results (in %) are shown in Table 3. We conducted experiments on the AID dataset with a ratio of old to new categories set at 2:1. Here, "Backbone" denotes the feature extractor used by the model, "Unsup-CL" represents contrastive learning on unlabeled data, "Sup-CL" denotes contrastive learning with label correction on labeled data, and "Slot-CL" indicates contrastive learning using extracted slot features. The first row of Table 3 presents the results obtained by directly clustering the feature outputs of the pre-trained model using K-means, which showed poor performance and failed to capture meaningful features. Rows two and three reflect the impact of the two image-level contrastive learning modules on the model's performance, demonstrating a significant improvement in classification performance upon their inclusion. The comparison between these two rows reveals that the introduction of module "Unsup-CL" enhances the model's classification ability for novel class data, while module "Sup-CL" substantially improves the model's representational capacity for old

class data. The fourth row illustrates the classification performance of the model after removing the slot, essentially reverting to a conventional GCD algorithm. Comparing the results of the last row with those of the previous rows, it is evident that the introduction of slot attention significantly improves the model's classification ability for novel class data, with a minor gain in recognizing old class data.

Table 3. Ablation study on the different components of our approach.

Backbone	Unsup-CL	Sup-CL	Slot-CL	AID		
				Old	Novel	All
✓	✗	✗	✗	37.1	30.4	32.9
✓	✓	✗	✗	65.8	51.9	56.2
✓	✗	✓	✗	82.2	49.7	66.4
✓	✓	✓	✗	88.3	72.2	76.9
✓	✓	✓	✓	**91.5**	**81.9**	**83.2**

4.3.2. Effectiveness of the Number of Slots and the Loss Balancing Weight

Table 4 illustrates the impact of the number of slots, denoted as K, on the model's classification performance across the entire test set. It can be observed from the table that the effect of K varies across different datasets. The classification performance of the model on the AID dataset decreases as K increases, which could be attributed to the relatively limited number of objects in the AID dataset. On the NWPU-RESISC45 dataset, the model achieves the best classification performance when K is set to 256. This is likely because NWPU-RESISC45 is a relatively large dataset with a greater variety of objects. Meanwhile, on the Million-AID dataset, the model exhibits optimal classification performance when K is set to 1024. This is attributed to the large scale of the Million-AID dataset, which encompasses a vast array of objects.

Table 4. Effectiveness of the number of slots.

K	AID	NWPU-RESISC45	Million-AID
128	**83.2**	79.5	71.9
256	82.9	**80.6**	72.4
512	82.9	79.7	72.1
1024	82.7	79.7	**72.8**
2048	82.1	78.6	72.6

Table 5 illustrates the impact of parameter λ on model learning. λ is the loss balancing weight controlling the learning at the object and image levels. When λ is set to 0, Slot-GCD degenerates into the general GCD algorithm. As λ increases, the emphasis on object-level contrastive learning gradually intensifies, leading to improved model classification accuracy. However, when λ reaches a certain value, the model's performance begins to decline. This is because the model becomes overly focused on local semantics at this point, neglecting global semantics, which consequently leads to inaccurate image feature encoding.

Table 5. Effectiveness of the loss balancing weight.

λ	AID	NWPU-RESISC45	Million-AID
0	76.9	75.1	65.8
0.3	81.5	79.3	**73.4**
0.5	**83.2**	**80.6**	72.8
0.7	81.7	80.1	70.4
0.9	78.8	77.6	70.1
1	78.4	76.9	69.5

5. Discussion

The GCD task endeavors to equip models with the capability to autonomously learn from unlabeled data within an open-world context. In contrast to the NCD task, the unlabeled data in GCD encompasses categories previously encountered by the model. This bears considerable potential for applications in UAVs, wherein GCD algorithms can facilitate the automatic acquisition of knowledge concerning known and unknown scenes and environmental conditions. In this paper, we introduce a learning approach tailored to address the GCD task using aerial image data, taking into account the unique characteristics of aerial images.

Prior GCD algorithms predominantly focused on the holistic encoding of aerial image features, such as average pooling features to derive the overall image representations, followed by learning the semantics of input images via contrastive learning. However, these methods neglect the semantics of local objects. Particularly when dealing with unlabeled aerial images, the characteristics of multi-object and inter-class object sharing may lead to inaccuracies in the derived image features. To mitigate this issue, we propose a semantic decoupling and reconstruction method based on slot attention. This method computes the attention of each slot at various positions on the feature map to extract distinct local semantic features. Additionally, we reconstruct the overall semantic feature of the image using the slots' confidence scores and feature maps. Consequently, semantic features at both the object and image levels can be learned through contrastive learning to attain more precise feature representations. Experimental results on three public aerial image datasets demonstrate that our proposed method outperforms other advanced GCD algorithms in terms of classification performance.

Although our proposed method yields promising results, there are potential areas for improvement. Firstly, the efficacy of the GCD task heavily hinges on the model's feature extraction prowess. This suggests that employing more potent feature extractors, such as Transformer architectures [40], holds the potential to enhance the model's classification performance. Secondly, the contrastive learning framework utilized in this study falls within the realm of representation learning. Leveraging advanced representation learning techniques, such as sophisticated geometric data augmentations and employing multiple local crops [48,58], can bolster the expressive capacity of the extracted features. Moreover, during object-level and image-level contrastive learning, the positive pairs utilized are exclusively drawn from sample pairs of the same aerial image under different transformations. However, within a batch, there exist diverse images from the same category, which should also be considered positive pairs. The exploration of strategies for selecting these positive pairs constitutes a potential avenue for future research in this domain. Lastly, slot attention has demonstrated notable proficiency in extracting local object features from images [38,59,60]. Therefore, in the future, we will exploit slot attention for addressing open-world learning tasks characterized by multiple targets, such as open-set object detection and image segmentation.

6. Conclusions

This work addresses the challenging task of Generalized Category Detection (GCD) for aerial image classification. We propose Slot-GCD, a novel framework that leverages slot attention to exploit the distinctive attributes of aerial images. Unlike conventional methods, Slot-GCD incorporates contrastive learning of image features at both the image and object levels, recognizing the multi-object nature of aerial images. Slot attention facilitates the semantic decoupling of feature maps, generating object-level features that capture local semantics. Furthermore, we propose a technique to reconstruct overall image features by linearly combining feature maps with their corresponding confidence scores. Through contrastive learning, the model learns to identify similarities and differences across images, fostering a more generalized understanding of aerial images that improves its performance on unseen categories. Extensive evaluations on three benchmark aerial image datasets demonstrate the superiority of Slot-GCD compared to state-of-the-art methods. By enabling

accurate detection of novel categories, Slot-GCD has the potential to significantly extend applications of aerial image classification.

Author Contributions: Conceptualization, Y.Z. (Yifan Zhou) and Y.W.; methodology, Y.Z. (Yifan Zhou) and H.Z.; validation, S.L. and Y.W.; investigation, W.Y.; resources, W.Y.; writing—original draft preparation, Y.Z. (Yifan Zhou); writing—review and editing, H.Z., Y.Z. (Yan Zhang) and W.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by the National Natural Science Foundation of China (NSFC) Regional Innovation and Development Joint Fund (No. U22A2010) and the CETC key laboratory of aerospace information applications under Grant SKX222010021.

Data Availability Statement: The AID dataset can be acquired from https://captain-whu.github.io/AID/, and is accessed on 10 February 2024; the Million-AID dataset can be acquired from https://captain-whu.github.io/DiRS/, and is accessed on 10 February 2024; the NWPU-RESISC45 dataset can be acquired from http://www.escience.cn/people/JunweiHan/NWPU-RESISC45.html, and is accessed on 10 February 2024.

Acknowledgments: The authors would like to express their sincere thanks to the anonymous reviewers for their valuable comments and suggestions. In addition, the numerical calculations in this article were performed on the supercomputing system in the Supercomputing Center, Wuhan University.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Detka, J.; Coyle, H.; Gomez, M.; Gilbert, G.S. A Drone-Powered Deep Learning Methodology for High Precision Remote Sensing in California's Coastal Shrubs. *Drones* **2023**, *7*, 421. [CrossRef]
2. Shi, Y.; Fu, B.; Wang, N.; Cheng, Y.; Fang, J.; Liu, X.; Zhang, G. Spectral-Spatial Attention Rotation-Invariant Classification Network for Airborne Hyperspectral Images. *Drones* **2023**, *7*, 240. [CrossRef]
3. Safonova, A.; Hamad, Y.; Dmitriev, E.; Georgiev, G.; Trenkin, V.; Georgieva, M.; Dimitrov, S.; Iliev, M. Individual Tree Crown Delineation for the Species Classification and Assessment of Vital Status of Forest Stands from UAV Images. *Drones* **2021**, *5*, 77. [CrossRef]
4. Jiménez-Torres, M.; Silva, C.P.; Riquelme, C.; Estay, S.A.; Soto-Gamboa, M. Automatic Recognition of Black-Necked Swan (*Cygnus melancoryphus*) from Drone Imagery. *Drones* **2023**, *7*, 71. [CrossRef]
5. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G. Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2020**, *13*, 3735–3756. [CrossRef]
6. Vaze, S.; Han, K.; Vedaldi, A.; Zisserman, A. Generalized Category Discovery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7482–7491.
7. Wen, X.; Zhao, B.; Qi, X. Parametric classification for generalized category discovery: A baseline study. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 16590–16600.
8. Pu, N.; Zhong, Z.; Sebe, N. Dynamic Conceptional Contrastive Learning for Generalized Category Discovery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7579–7588.
9. Zhang, S.; Khan, S.; Shen, Z.; Naseer, M.; Chen, G.; Khan, F.S. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3479–3488.
10. Zhong, Y.; Zhu, Q.; Zhang, L. Scene Classification Based on the Multifeature Fusion Probabilistic Topic Model for High Spatial Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 6207–6222. [CrossRef]
11. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
12. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
13. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
14. Yang, Y.; Newsam, S. Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
15. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the Fisher Kernel for Large-Scale Image Classification. In Proceedings of the European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010; pp. 143–156.
16. Scheirer, W.J.; de Rezende Rocha, A.; Sapkota, A.; Boult, T.E. Toward Open Set Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1757–1772. [CrossRef]
17. Bendale, A.; Boult, T.E. Towards Open Set Deep Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, 26 June–1 July 2016; pp. 1563–1572.

18. Kong, S.; Ramanan, D. Opengan: Open-set recognition via open data generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 813–822.
19. Neal, L.; Olson, M.; Fern, X.; Wong, W.K.; Li, F. Open set learning with counterfactual images. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 613–628.
20. Sun, X.; Yang, Z.; Zhang, C.; Ling, K.V.; Peng, G. Conditional gaussian distribution learning for open set recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13480–13489.
21. Chen, G.; Peng, P.; Wang, X.; Tian, Y. Adversarial reciprocal points learning for open set recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 8065–8081. [CrossRef]
22. Liu, W.; Nie, X.; Zhang, B.; Sun, X. Incremental Learning With Open-Set Recognition for Remote Sensing Image Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
23. Zhang, H.; Li, A.; Guo, J.; Guo, Y. Hybrid models for open set recognition. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 102–117.
24. Fini, E.; Sangineto, E.; Lathuilière, S.; Zhong, Z.; Nabi, M.; Ricci, E. A Unified Objective for Novel Class Discovery. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9264–9272.
25. Zhong, Z.; Fini, E.; Roy, S.; Luo, Z.; Ricci, E.; Sebe, N. Neighborhood Contrastive Learning for Novel Class Discovery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10867–10875.
26. Han, K.; Vedaldi, A.; Zisserman, A. Learning to discover novel visual categories via deep transfer clustering. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8401–8409.
27. Han, K.; Rebuffi, S.; Ehrhardt, S.; Vedaldi, A.; Zisserman, A. AutoNovel: Automatically Discovering and Learning Novel Visual Categories. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 6767–6781. [CrossRef] [PubMed]
28. Liu, Y.; Tuytelaars, T. Residual Tuning: Toward Novel Category Discovery Without Labels. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 7271–7285. [CrossRef] [PubMed]
29. Roy, S.; Liu, M.; Zhong, Z.; Sebe, N.; Ricci, E. Class-Incremental Novel Class Discovery. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Volume 13693, pp. 317–333.
30. Oord, A.V.D.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
31. Wang, T.; Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Proceedings of the International Conference on Machine Learning, Virtual Event, 13–18 July 2020; pp. 9929–9939.
32. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
33. Sermanet, P.; Lynch, C.; Chebotar, Y.; Hsu, J.; Jang, E.; Schaal, S.; Levine, S.; Brain, G. Time-contrastive networks: Self-supervised learning from video. In Proceedings of the IEEE International Conference on Robotics and Automation, Brisbane, QLD, Australia, 21–25 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1134–1141.
34. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G.E. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the International Conference on Machine Learning, Virtual Event, 13–18 July 2020; Volume 119, pp. 1597–1607.
35. Ni, R.; Shu, M.; Souri, H.; Goldblum, M.; Goldstein, T. The close relationship between contrastive learning and meta-learning. In Proceedings of the International Conference on Learning Representations, Virtual Event, Austria, 3–7 May 2021.
36. Bi, M.; Wang, M.; Li, Z.; Hong, D. Vision Transformer With Contrastive Learning for Remote Sensing Image Scene Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 738–749. [CrossRef]
37. Locatello, F.; Weissenborn, D.; Unterthiner, T.; Mahendran, A.; Heigold, G.; Uszkoreit, J.; Dosovitskiy, A.; Kipf, T. Object-Centric Learning with Slot Attention. In Proceedings of the International Conference on Neural Information Processing Systems, Online, 6–12 December 2020.
38. Wen, X.; Zhao, B.; Zheng, A.; Zhang, X.; Qi, X. Self-Supervised Visual Representation Learning with Semantic Grouping. In Proceedings of the International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022.
39. Li, L.; Wang, B.; Verma, M.; Nakashima, Y.; Kawasaki, R.; Nagahara, H. Scouter: Slot attention-based classifier for explainable image recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1046–1055.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All you Need. In Proceedings of the International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
41. Cuturi, M. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 2292–2300.
42. Du, R.; Wang, G.; Zhang, N.; Chen, L.; Liu, W. Domain Adaptive Remote Sensing Scene Classification with Middle-Layer Feature Extraction and Nuclear Norm Maximization. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *17*, 1–12. [CrossRef]
43. Chen, S.; Wei, Q.; Wang, W.; Tang, J.; Luo, B.; Wang, Z. Remote Sensing Scene Classification via Multi-Branch Local Attention Network. *IEEE Trans. Image Process.* **2022**, *31*, 99–109. [CrossRef] [PubMed]
44. Feng, L.; Shu, S.; Lin, Z.; Lv, F.; Li, L.; An, B. Can cross entropy loss be robust to label noise? In Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021; pp. 2206–2212.

45. Hendrycks, D.; Mazeika, M.; Kadavath, S.; Song, D. Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. In Proceedings of the International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 15637–15648.
46. Goyal, P.; Duval, Q.; Seessel, I.; Caron, M.; Misra, I.; Sagun, L.; Joulin, A.; Bojanowski, P. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv* **2022**, arXiv:2202.08360.
47. Balestriero, R.; Ibrahim, M.; Sobal, V.; Morcos, A.; Shekhar, S.; Goldstein, T.; Bordes, F.; Bardes, A.; Mialon, G.; Tian, Y.; et al. A cookbook of self-supervised learning. *arXiv* **2023**, arXiv:2304.12210.
48. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap Your Own Latent—A New Approach to Self-Supervised Learning. In Proceedings of the International Conference on Neural Information Processing Systems, Online, 6–12 December 2020.
49. Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; Hu, H. SimMIM: A Simple Framework for Masked Image Modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9643–9653.
50. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R.B. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 15979–15988.
51. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9630–9640.
52. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
53. Long, Y.; Xia, G.S.; Li, S.; Yang, W.; Yang, M.Y.; Zhu, X.X.; Zhang, L.; Li, D. On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidances, and Million-AID. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4205–4230. [CrossRef]
54. Thorndike, R.L. Who belongs in the family? *Psychometrika* **1953**, *18*, 267–276. [CrossRef]
55. Kuhn, H.W. 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*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 29–47.
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 770–778.
57. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
58. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In Proceedings of the International Conference on Neural Information Processing Systems, Online, 6–12 December 2020.
59. Xu, J.; Hou, J.; Zhang, Y.; Feng, R.; Wang, Y.; Qiao, Y.; Xie, W. Learning open-vocabulary semantic segmentation models from natural language supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2935–2944.
60. Li, L.; Liniger, A.; Millhaeusler, M.; Tsiminaki, V.; Li, Y.; Dai, D. Object-centric Cross-modal Feature Distillation for Event-based Object Detection. *arXiv* **2023**, arXiv:2311.05494.