remote sensing

MDPI

Article

Mask-Guided Teacher–Student Learning for Open-Vocabulary Object Detection in Remote Sensing Images

Shuojie Wang, Yu Song, Jiajun Xiang, Yanyan Chen, Ping Zhong and Ruigang Fu *

National Key Laboratory of Science and Technology on ATR, National University of Defense Technology, Changsha 410073, China; wangshuojie@nudt.edu.cn (S.W.); songyu22@nudt.edu.cn (Y.S.); xiangjiajun24@nudt.edu.cn (J.X.); chenyanyan23@nudt.edu.cn (Y.C.); zhongping@nudt.edu.cn (P.Z.)
* Correspondence: furuigang08@nudt.edu.cn

Highlights

What are the main findings?

- A selective masking strategy enables direct utilization of partially annotated data, eliminating strict data separation requirements and improving annotation efficiency.
- Dynamic frequency class weighting based on information theory automatically balances class distributions.

What is the implication of the main finding?

- The approach eliminates the need for strict data separation and complex preprocessing, providing a more practical and data-efficient solution compared to large-scale pretraining methods.
- The dynamic frequency class weighting mechanism can be integrated into other open-vocabulary frameworks to address class imbalance issues, offering broad applicability.

Abstract

Open-vocabulary object detection in remote sensing aims to detect novel categories not seen during training, which is crucial for practical aerial image analysis applications. While some approaches accomplish this task through large-scale data construction, such methods incur substantial annotation and computational costs. In contrast, we focus on efficient utilization of limited datasets. However, existing methods such as CastDet struggle with inefficient data utilization and class imbalance issues in pseudo-label generation for novel categories. We propose an enhanced open-vocabulary detection framework that addresses these limitations through two key innovations. First, we introduce a selective masking strategy that enables direct utilization of partially annotated images by masking base category regions in teacher model inputs. This approach eliminates the need for strict data separation and significantly improves data efficiency. Second, we develop a dynamic frequency-based class weighting that automatically adjusts category weights based on real-time pseudo-label statistics to mitigate class imbalance issues. Our approach integrates these components into a student–teacher learning framework with RemoteCLIP for novel category classification. Comprehensive experiments demonstrate significant improvements on both datasets: on VisDroneZSD, we achieve 42.7% overall mAP and 41.4% harmonic mean, substantially outperforming existing methods. On DIOR dataset, our method achieves 63.7% overall mAP with 49.5% harmonic mean. Our framework achieves more balanced performance between base and novel categories, providing a practical and data-efficient solution for open-vocabulary aerial object detection.

Academic Editor: Rongjun Qin

Received: 18 August 2025
Revised: 20 September 2025
Accepted: 7 October 2025
Published: 9 October 2025

Citation: Wang, S.; Song, Y.; Xiang, J.; Chen, Y.; Zhong, P.; Fu, R. Mask-Guided Teacher–Student Learning for Open-Vocabulary Object Detection in Remote Sensing Images. *Remote Sens.* **2025**, *17*, 3385. https://doi.org/10.3390/rs17193385

Keywords: open vocabulary; multi-object detection; remote sensing imagery; deep learning; teacher–student learning

1. Introduction

Remote sensing object detection (RSOD) aims to determine whether objects of interest exist in given remote sensing images and return the category and location of each predicted object. Here, "objects" refer to discrete, man-made structures such as aircraft, vehicles, and ships, as opposed to unstructured scene elements like terrain, sky, and vegetation [1]. As a fundamental task in remote sensing image interpretation, remote sensing object detection has undergone significant evolution from traditional methods to deep learning approaches over the past two decades. Initially inspired by successful natural image detection methods, many studies have applied these frameworks (such as Faster R-CNN [2], YOLO [3], and SSD [4]) to remote sensing object detection, achieving breakthroughs. Based on the characteristics of geospatial objects in remote sensing images, many methods specifically designed for remote sensing images have emerged in recent years [5,6]. For example, ROI-Transformer [7], R3Det [8], and RSDet++ [9] for rotated object detection; CA-CNN [10], ASPP [11], and RFB [12] for mining remote sensing-related context information; and edge-enhanced GAN [13] and YOLO-DCTI [14] for small object detection. However, these methods operate under closed-set assumptions and require extensive manual annotations for new categories.

Open-vocabulary object detection (OVD [15]) aims to locate and identify new categories not annotated in the dataset by leveraging base categories and linguistic vocabulary knowledge, i.e., large-scale language models or vocabulary data, which are of significant importance in remote sensing image analysis. With the proliferation of drone technology and advances in satellite image acquisition capabilities, the diversity of observable targets in aerial images far exceeds the limited categories covered by existing annotated datasets [16,17]. Traditional closed-set detection methods require collecting and annotating large amounts of training data for each new category, which is particularly expensive and time-consuming in the remote sensing domain. Therefore, developing open-vocabulary detection techniques that can detect additional categories using limited annotated data is of great value for advancing practical applications of aerial image interpretation. First in natural images, OVR-CNN [18] introduced the first method for OVD. In recent years, due to the rapid development of large-scale language models in the remote sensing field, models such as RemoteCLIP [19] and GeoRSCLIP [20] have obtained rich semantic understanding capabilities through pre-training on large-scale remote sensing image–text pairs, providing strong foundational support for remote sensing open-vocabulary tasks. Remote sensing open-vocabulary detection is an emerging research direction that has gained attention only recently.

Unlike natural image OVD, which has been extensively studied with numerous established methods, specialized approaches for remote sensing scenarios are still limited. In this nascent field, CastDet [21] proposed the first CLIP-activated student–teacher open-vocabulary detection framework specifically for aerial images in ECCV 2024. The core contribution of CastDet lies in integrating RemoteCLIP as an external teacher into a student–teacher learning mechanism, effectively leveraging the rich knowledge of pre-trained vision–language models for novel category discovery. This design significantly improves the detection performance of novel categories in aerial images. LAE-DINO [22] defines the task as "locating anything on Earth," constructing the first large-scale remote sensing object detection dataset LAE-1M, and proposing a foundation model with dynamic vocabulary

construction and vision-guided text prompt learning. OpenRSD [23] proposes a general open prompt detection framework supporting multi-modal prompts, integrating alignment heads and fusion heads to balance speed and accuracy, and adopting a multi-stage training pipeline to enhance generalization capabilities. This method constructs a large-scale dataset ORSD+, containing 470k images and 200 categories.

However, large-scale pre-training methods (such as LAE-DINO and OpenRSD) require constructing large-scale datasets, which results in extremely high data collection and training costs, making them difficult to implement in resource-constrained practical application scenarios. Unlike their pursuit of more data, we focus more on how to make good use of limited datasets. CastDet is precisely based on this setting, but its data filtering constraints prevent images containing mixed categories from being fully utilized. Specifically, these methods require strict separation of annotated base categories data and completely unlabeled data during training. When base categories and novel categories coexist in the same image, this constraint becomes problematic because the presence of novel targets makes the entire image unusable for supervised training of base categories, requiring the removal of all annotations from the image for unsupervised learning. This design stems from the strict assumptions of traditional semi-supervised learning but appears overly conservative in open-vocabulary scenarios, resulting in an enormous waste of annotation resources. This limitation significantly introduces considerable annotation overhead in practical application scenarios. When facing potentially increasing novel categories, strict inspection and re-partitioning of annotated datasets are required.

Furthermore, severe class imbalance problems are often encountered when generating pseudo-labels for novel categories. The inherent difficulty of detecting rare or small targets in aerial images exacerbates this problem, resulting in training bias and poor performance on certain categories. CastDet adopts a static weighting scheme that cannot adapt to the dynamic distribution of pseudo-labels during training and changes in datasets.

To address these limitations, we propose an enhanced open-vocabulary detection framework that supports one-step training with partially annotated data. The main contributions of this work are summarized as follows:

- Selective masking strategy: We propose a selective masking strategy that enables direct utilization of images containing both base and novel categories, relaxing the strict data separation constraints of existing methods and achieving more flexible dataset utilization.
- Dynamic frequency-based class weighting: We propose a dynamic weighting mechanism based on pseudo-label queue frequency, which automatically adjusts category weights by monitoring pseudo-label category distribution to alleviate class imbalance issues in pseudo-labels.
- We conduct comprehensive experiments on aerial open-vocabulary detection baselines VisDroneZSD and DIOR, demonstrating significant improvements over existing methods.

2. Methods

2.1. Problem Formulation

Given a detection dataset containing only annotations of base categories, $D_{\text{labeled}} = \{(I_i, y_i)\}_{i=1}^{N}$, where N denotes the total number of images in the dataset, $I_i \in \mathbb{R}^{H\times W\times 3}$ represents the input image, and $y_i = \{(b_{i,k}, c_{i,k})\}_{k=1}^{K_i}$ represents the corresponding annotations with K_i being the number of objects in the i-th image. Each annotation includes bounding box coordinates $b_{i,k} \in \mathbb{R}^4$ and category labels $c_{i,k} \in \mathcal{C}_{\text{base}}$, where $\mathcal{C}_{\text{base}}$ denotes the set of base categories with available annotations during training.

Existing methods require strict separation of labeled and unlabeled data. The dataset must be traversed first, and if an image contains targets of novel categories $\mathcal{C}_{\text{novel}}$, the entire image must be assigned to the unlabeled dataset, thus discarding the base categories $\mathcal{C}_{\text{base}}$ annotations in that image. In contrast, our approach aims to fully utilize partially annotated images. For images containing both base categories $\mathcal{C}_{\text{base}}$ and novel categories $\mathcal{C}_{\text{novel}}$, we no longer need to assign them completely as unlabeled data but can instead directly train a detector that can simultaneously detect both category types.

Our goal is to train a detector capable of detecting objects from the complete category set $\mathcal{C}_{\text{test}} = \mathcal{C}_{\text{base}} \cup \mathcal{C}_{\text{novel}}$ during inference, where $\mathcal{C}_{\text{test}}$ represents the union of base and novel categories that the model should recognize at test time. Therefore, the key challenge lies in how to effectively discover and learn novel categories while maintaining supervised learning performance on base categories.

2.2. Overall Framework

We propose a mask-guided teacher–student framework for open-vocabulary object detection. The core idea is to achieve functional decoupling through differentiated data input strategies, enabling an image containing both base and novel categories to simultaneously serve supervised learning and pseudo-label generation. As shown in Figure 1, our framework contains four components: a data processing module, a teacher model, a dynamic label queue, and a student model.

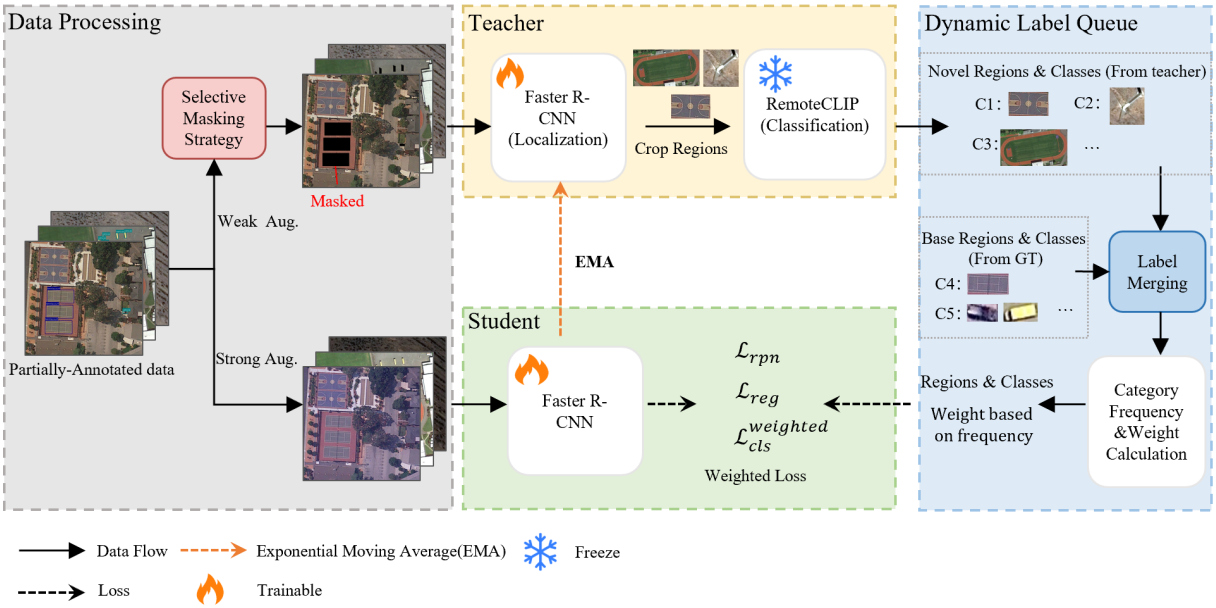

Figure 1. Overview of the proposed method. It contains four components: a data processing module, a teacher model, a dynamic label queue, and a student model.

The data processing module handles input preparation with two parallel pathways. For the teacher branch, we first apply weak augmentation to the original images. Subsequently, our selective masking strategy masks base category regions, forcing the teacher model to focus on potential novel category objects. For the student branch, we apply strong augmentation directly to the complete images to improve model robustness and generalization capability during training.

The teacher model operates on selectively masked images where base category regions are removed through our selective masking strategy, forcing the model to focus on discovering novel category objects. The teacher branch employs Faster R-CNN for object localization, generating high-quality region proposals that are subsequently classified us-

ing RemoteCLIP's powerful vision–language capabilities. This design leverages the rich semantic knowledge from large-scale remote sensing image–text pre-training to achieve robust novel category identification.

The dynamic label queue serves as the central coordination module that manages both novel category pseudo-labels generated by the teacher model and base category ground-truth annotations from the original dataset. This queue continuously tracks category frequency statistics and implements our information-theoretic dynamic weighting mechanism to address class imbalance issues. The queue performs label merging operations, combining reliable pseudo-labels with ground-truth annotations to create comprehensive supervision signals for the student model.

The student model receives complete unmasked images along with the merged supervision signals from the dynamic queue. Using the same Faster R-CNN architecture as the teacher, the student model trains on both base and novel categories simultaneously through our weighted supervision learning approach. The teacher model parameters are updated through an exponential moving average (EMA) from the student model, ensuring consistent learning progress and stability.

This design eliminates the need for prior data separation and directly utilizes original annotated data containing multiple categories for end-to-end training, significantly improving data utilization efficiency and simplifying actual deployment workflows. The entire training process is completed through a single-stage weighted supervision approach, avoiding complex multi-stage training and loss balancing issues between different learning paradigms.

2.3. Selective Masking Strategy

As illustrated in Figure 2, to achieve one-step training, we design a selective masking strategy to process data. For image I_i, we extract all the base categories' bounding boxes from their annotations $y_i = \{(b_{i,k}, c_{i,k})\}_{k=1}^{K_i}$:

$$\mathcal{B}_{\text{base}}^{(i)} = \{b_{i,k} \mid c_{i,k} \in \mathcal{C}_{\text{base}}, k = 1, \ldots, K_i\} \tag{1}$$

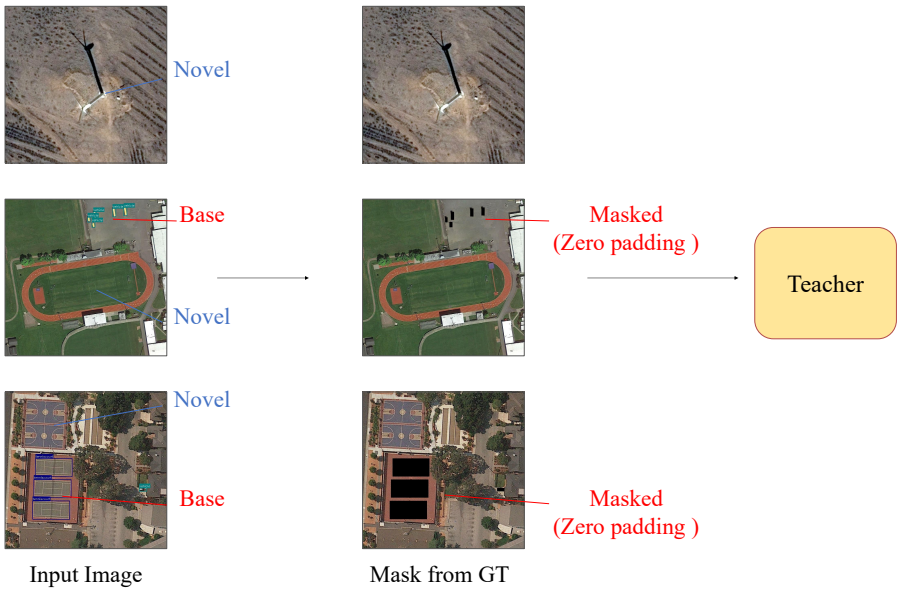

Figure 2. Illustration of the selective masking strategy. The figure demonstrates how our method processes images containing both base categories (marked in red) and novel categories (marked in blue). Base category regions are masked with zero padding (black areas), while novel category regions remain unchanged, forcing the teacher model to focus on discovering novel categories.

Based on the extracted bounding box set, we construct a binary mask matrix, written as follows:

$$M_{\text{base}}^{(i)}(u,v) = \begin{cases} 0 & \text{if } \exists b \in \mathcal{B}_{\text{base}}^{(i)} \text{ such that } (u,v) \in b \\ 1 & \text{otherwise} \end{cases}, \tag{2}$$

where (u, v) represents pixel coordinates in the image.

For images input to the teacher model, we mask the base category regions, calculated as follows:

$$I_{\text{teacher}}^{(i)} = I_i \odot M_{\text{base}}^{(i)} + f_{\text{mask}} \odot (1 - M_{\text{base}}^{(i)}), \tag{3}$$

where $I_i \odot M_{\text{base}}^{(i)}$ represents preserving original pixel values of non-base category regions, and $f_{\text{mask}} \odot (1 - M_{\text{base}}^{(i)})$ represents applying mask function f_{mask} to base category regions.

The mask function can adopt strategies such as Gaussian blur GaussianBlur(I) or zero padding 0. Here, we adopt zero padding. Zero padding provides complete information suppression, which is theoretically superior to other masking strategies for our objective. Theoretically, the advantage of zero padding lies in reducing the information entropy of masked regions to zero, completely blocking the visual information propagation of base categories. In contrast, while Gaussian blur reduces information clarity, it still preserves low-frequency information (such as shape contours), and this residual information may be exploited by the strong fitting capability of deep networks, thereby weakening the masking effect. Meanwhile, zero padding provides deterministic background suppression, avoiding the random interference that noise injection may bring, ensuring that the teacher model learns genuine novel category features rather than possible noise patterns.

Theoretically, selective masking addresses the fundamental challenge in open-vocabulary detection: how to prevent the model from being dominated by well-represented categories while encouraging exploration of under-represented ones. The selective masking strategy directly impacts the behavior of the Region Proposal Network (RPN) in Faster R-CNN. Without masking, the RPN generates proposals with a strong bias toward base categories, as these regions produce higher objectness scores. By masking base category regions, we eliminate these high-confidence proposals and force the RPN to generate proposals from regions that might contain novel categories. This redistribution of proposal generation increases the likelihood of discovering novel category objects that would otherwise be overshadowed by dominant base category responses.Through this masking strategy, base category regions are "hidden" in images received by the teacher model, forcing it to focus on discovering and localizing potential novel category targets, while the student model still receives the complete image I_i for training.

2.4. Pseudo-Label Generation and Quality Control

The quality of pseudo-labels directly affects the learning effectiveness of the student model. Therefore, after the teacher model receives masked images, we adopt a complete pseudo-label generation and quality control mechanism to ensure that only reliable predictions are used as supervision signals.

Teacher model pseudo-label generation. The teacher model receives masked images where base category regions have been selectively masked out, forcing it to focus on potential novel category objects. The process begins with feature extraction using the backbone network, followed by the Region Proposal Network (RPN), which generates an initial set of object proposals. These proposals represent potential object locations but lack category-specific information.

To refine the localization accuracy, we employ an iterative optimization strategy where the predicted bounding boxes undergo multiple rounds of regression refinement. In each iteration t, the current bounding boxes $\hat{b}_i^{(t)}$ (where $i = 1, 2, \ldots, N$ indexes the proposals)

are first converted to ROI format, then fed through the ROI pooling layer to extract fixed-size feature representations $f_i^{(t)}$. The regression head processes these features to predict coordinate refinement offsets $\Delta_i^{(t)}$:

$$\Delta_i^{(t)} = \text{BBoxHead}(f_i^{(t)}) \tag{4}$$

The refined bounding boxes are obtained by decoding the current coordinates with the predicted offsets:

$$\hat{b}_i^{(t+1)} = \text{BBoxDecode}(\hat{b}_i^{(t)}, \Delta_i^{(t)}), \tag{5}$$

where $\text{BBoxDecode}(\cdot, \cdot)$ transforms the regression deltas into absolute coordinates based on the current bounding box positions. This iterative refinement process continues for a predetermined number of steps, progressively improving the spatial accuracy of the bounding box predictions.

Once high-quality bounding box proposals, $\{\hat{b}_i\}_{i=1}^K$, are obtained, the critical step of category classification begins. We crop the corresponding image regions based on the refined bounding boxes and resize them to the input resolution required by RemoteCLIP (typically 224 × 224 pixels). These cropped regions, $\{x_i\}_{i=1}^K$, are then fed into the visual encoder of RemoteCLIP, which extracts rich semantic features:

$$v_i = \text{VisualEncoder}(x_i) \in \mathbb{R}^{d_v} \tag{6}$$

For the textual component, we construct category descriptions using a predefined template "a photo of [category]" for each target category, including both base and novel categories. These textual descriptions are processed through RemoteCLIP's text encoder:

$$t_j = \text{TextEncoder}(\text{template}_j) \in \mathbb{R}^{d_t}, \tag{7}$$

where template_j represents the text template for category j.

The classification process involves computing similarity scores between the L2-normalized visual and text embeddings, scaled by a learnable temperature parameter:

$$s_{ij} = \frac{v_i^T \cdot t_j}{\tau \|v_i\|_2 \cdot \|t_j\|_2}, \tag{8}$$

where $\frac{1}{\tau}$ corresponds to the clip logit scale parameter in the implementation. The final classification probabilities are obtained by applying a softmax function over these similarity scores:

$$p_{ij} = \frac{\exp(s_{ij})}{\sum_{k=1}^{|\mathcal{C}|} \exp(s_{ik})}, \tag{9}$$

where p_{ij} represents the probability that the i-th cropped region belongs to category j. The predicted category label for each region is determined by the following:

$$\hat{c}_i = \arg\max_j p_{ij} \tag{10}$$

This integration of RemoteCLIP enables our framework to leverage the rich semantic knowledge acquired from large-scale remote sensing image–text pairs, providing robust classification capabilities for both base and novel categories without requiring additional training on the target dataset.

Multi-Layer Quality Control Mechanism. We adopt a three-layer progressive filtering strategy to ensure pseudo-label quality:

First step: Non-maximum suppression. NMS processing is applied to candidate boxes of the same category to eliminate duplicate detections while removing redundancy without mistakenly deleting adjacent targets.

Second step: Geometric constraint filtering based on physical characteristics and statistical patterns of targets in aerial images. To avoid mistaking large background areas as targets, we apply a geometric constraint to filter obviously oversized detection boxes. Based on empirical analysis of target size distributions in aerial imagery, we set a maximum bounding box area threshold of 400,000 pixels (approximately equivalent to a 632 × 632 pixel square).

Third step: Overlap detection. Calculate IoU between pseudo-labels and ground-truth annotations, removing pseudo-labels with overlap exceeding 0.5 to avoid conflicts with already annotated base categories.

2.5. Dynamic Frequency-Based Class Weighting

To address imbalance issues among novel categories, commonly used methods establish higher static weights for few-shot samples, but this approach cannot adapt to dynamic changes in pseudo-label distribution during training and changes across different datasets. In the classic Class-Balanced Loss paper, weights can be defined by counting effective sample numbers:

$$w_i = \frac{1-\beta}{1-\beta^{n_{c_i}}}, \tag{11}$$

where β is a hyperparameter controlling the re-weighting strength and n_{c_i} represents the number of samples for category c_i.

However, CB Loss is based on fixed training sample statistics, assuming that the number of samples for each category is determined at the beginning of training. In our method, samples of novel categories are actually dynamically generated pseudo-labels, whose quantity and quality continuously change as teacher model capabilities improve. Static effective sample number formulas cannot capture this dynamic evolution process. Therefore, we propose a weight adjustment mechanism based on dynamic label queue frequency, which can automatically adjust loss weights according to real-time category frequency statistics in the pseudo-label queue, ensuring all rare novel categories receive adequate learning opportunities.

Long-Term and Short-Term Frequency Statistics. We maintain a pseudo-label queue to track category distribution and combine long-term and short-term statistical information. The design philosophy of combining long-term and short-term statistics is that long-term statistics reflect overall data distribution, providing a stable baseline, while short-term statistics reflect recent detection trends and can respond promptly to distribution changes. The specific frequency calculation formula is written as follows:

$$f_{c_i} = 0.5 \cdot f_{\text{queue},c_i} + 0.5 \cdot f_{\text{recent},c_i}, \tag{12}$$

where $f_{\text{queue},c_i} = \frac{N_{\text{queue},c_i}}{N_{\text{queue,total}}}$ represents the long-term frequency of category c_i in the entire queue, and $f_{\text{recent},c_i} = \frac{N_{\text{recent},c_i}}{N_{\text{recent,total}}}$ represents the short-term frequency of category c_i in the most recent 500 updates.

This design effectively prevents overfitting: pure short-term statistics might lead to excessive weight adjustment on local batches, while adding long-term statistics provides a global perspective, avoiding dramatic weight changes due to short-term fluctuations while still being able to respond promptly to real distribution shifts.

Adaptive Weight Calculation. Inspired by Shannon's information theory, we believe that rare categories carry a higher information value in imbalanced pseudo-label distribu-

tions. According to information theory principles, the self-information of category c_i is defined as follows:

$$I(c_i) = -\log p(c_i) = -\log f_{c_i} \tag{13}$$

Self-information measures the "information content" obtained when observing this event. Rare events contain more information and should therefore receive more attention in training. Based on this principle, we use self-information as the theoretical foundation for weight adjustment, ensuring the model allocates more learning resources to rare categories with high information content. We design the training weight for each novel category as follows:

$$w_{c_i}^{\text{raw}}(t) = 1 + \mu \cdot I(c_i) = 1 + \mu \cdot (-\log(f_{c_i} + \epsilon)), \tag{14}$$

where μ is the information intensity coefficient controlling the influence of self-information on weights. $\epsilon = \frac{1}{N_{\text{queue,total}}}$ is a smoothing term preventing numerical issues caused by zero frequency. The base weight 1 ensures all novel categories have weights no lower than base categories.

Smooth Weight Updates. To ensure stability in weight updates and avoid training shocks caused by sudden weight changes, we adopt exponential moving average for weight updates:

$$W_{c_i}(t) = \alpha \cdot W_{c_i}(t-1) + (1-\alpha) \cdot w_{c_i}^{\text{raw}}(t), \tag{15}$$

where α is the smoothing coefficient and t represents the update step.

The exponential moving average update mechanism ensures training stability by smoothing weight fluctuations. The smoothing parameter provides a balance between responsiveness to distribution changes and stability against noise in pseudo-label generation.

The advantage of an exponential moving average over directly using the latest predicted values is that it assigns higher weights to recent data while retaining historical trend information, making gradient descent smoother. We update weights every fixed number of iterations (e.g., 100), which can respond to distribution changes promptly while avoiding computational overhead from overly frequent updates.

2.6. Training with Merged Labels and Weighted Loss

The final training stage integrates high-quality pseudo-labels with ground-truth annotations to train the student model. Novel category pseudo-labels that pass quality control are merged with base category ground truth to create comprehensive supervision signals. For image I_i, the merged annotation is calculated as follows:

$$y_i^{\text{merged}} = \{(b_{i,k}^{\text{base}}, c_{i,k}^{\text{base}})\}_{k=1}^{K_i^{\text{base}}} \cup \{(b_{i,j}^{\text{novel}}, c_{i,j}^{\text{novel}})\}_{j=1}^{K_i^{\text{novel}}}, \tag{16}$$

where the former represents original base category ground-truth annotations (confidence set to 1.0), and the latter represents filtered novel category pseudo-labels (maintaining original confidence).

After merging ground-truth annotations with high-quality pseudo-labels, the student model is trained using supervised learning with dynamic class weighting. Our loss function integrates multiple components to handle both learning on base categories and novel categories simultaneously:

$$\mathcal{L} = \mathcal{L}_{\text{rpn}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{cls}}^{\text{weighted}} \tag{17}$$

RPN Loss. The Region Proposal Network (RPN) loss adopts the standard formulation for object proposal generation. In our framework, RPN performs binary classification

(foreground vs. background) rather than multi-class classification, with both classification and regression components:

$$\mathcal{L}_{\text{rpn}} = \frac{1}{N_{\text{rpn}}} \sum_i \mathcal{L}_{\text{cls}}^{\text{rpn}}(p_i, p_i^*) + \lambda \frac{1}{N_{\text{reg}}} \sum_i p_i^* \mathcal{L}_{\text{reg}}^{\text{rpn}}(t_i, t_i^*), \tag{18}$$

where p_i is the predicted probability of anchor i being an object (foreground), p_i^* is the binary ground-truth label (1 for positive, 0 for negative), t_i represents the predicted bounding box regression parameters, t_i^* denotes the ground-truth regression targets, N_{rpn} and N_{reg} are normalization terms, and λ is the balance weight between classification and regression loss. Importantly, all ground-truth labels are set to binary (foreground/background) for RPN training, regardless of their original categories.

Regression Loss. The regression loss operates on the ROI head level, applying Smooth L1 loss to the sampled proposals after RPN processing:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N_{\text{roi}}} \sum_{i=1}^{N_{\text{roi}}} \mathcal{L}_{\text{smooth-L1}}(\hat{b}_i, b_i^*), \tag{19}$$

where N_{roi} is the number of ROI samples (both positive and negative), $\hat{b}_i$ represents the predicted bounding box refinement parameters from the ROI head, and b_i^* denotes the corresponding regression targets. This loss refines the initial proposals from RPN to achieve more precise object localization. The Smooth L1 loss is defined as follows:

$$\mathcal{L}_{\text{smooth-L1}}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{20}$$

This loss function is less sensitive to outliers compared to L2 loss while maintaining differentiability, making it particularly suitable for bounding box regression tasks.

Weighted Classification Loss. To address the inherent class imbalance in pseudo-label generation, we introduce a dynamically weighted classification loss:

$$\mathcal{L}_{\text{cls}}^{\text{weighted}} = \frac{1}{N_b} \sum_{i=1}^{N_b} W_{c_i} \cdot \mathcal{L}_{\text{cls}}(\hat{s}_i, c_i), \tag{21}$$

where N_b is the total number of samples in a batch, $\hat{s}_i$ represents the predicted class scores, c_i is the ground-truth class label, and W_{c_i} is the dynamic weight corresponding to category c_i. The base classification loss $\mathcal{L}_{\text{cls}}$ employs standard cross-entropy:

$$\mathcal{L}_{\text{cls}}(\hat{s}_i, c_i) = -\log\left(\frac{\exp(\hat{s}_{i,c_i})}{\sum_{k=1}^{K} \exp(\hat{s}_{i,k})}\right), \tag{22}$$

where K is the total number of classes including background.

The category-specific weights are computed based on our proposed information-theoretic approach, updated periodically according to pseudo-label queue frequency statistics as described in Section 2.5:

$$W_{c_i} = \begin{cases} 1.0 & \text{if } c_i \in \mathcal{C}_{\text{base}} \\ W_{c_i}(t) & \text{if } c_i \in \mathcal{C}_{\text{novel}} \end{cases}, \tag{23}$$

where $\mathcal{C}_{\text{base}}$ and $\mathcal{C}_{\text{novel}}$ represent base and novel category sets, respectively, and $W_i(t)$ is the time-dependent dynamic weight that increases for rare categories to ensure balanced learning.

Our weighted supervision design offers the following key advantages: (1) *Simplicity*: Unlike traditional semi-supervised methods that require careful balancing between multiple loss branches (e.g., supervised loss, consistency loss, and pseudo-label loss), our approach integrates all supervision signals into a single coherent framework, eliminating the need for manual loss weight tuning between different branches. (2) *Adaptivity*: The dynamic weighting mechanism automatically adjusts to the evolving class distribution during training, ensuring that rare novel categories receive adequate attention without requiring prior knowledge of class frequencies. (3) *Robustness*: By maintaining standard loss formulations with principled information-theoretic weighting, our approach preserves training stability while effectively addressing class imbalance issues that commonly affect pseudo-label-based learning.

3. Materials and Experimental Settings

3.1. Dataset Introduction

We conduct main experimental evaluations on the DIOR and VisDroneZSD datasets. DIOR is a large-scale benchmark dataset for optical remote sensing image object detection, with each image having a resolution of 800 × 800 pixels. The dataset contains 23,463 images and 192,472 instances, covering 20 object categories. Similar to VisDroneZSD, we have set the same 16 basic categories and 4 new categories. We use the test set of 11,738 images for evaluation.

VisDroneZSD is a subset of the DIOR dataset, following the settings of the VisDrone2023 zero-shot object detection challenge. It includes 16 base categories and 4 novel categories, where novel categories are airport, basketball court, ground track field, and windmill. It contains 8730 images for training and 3337 images for testing. The training set contains only base category annotations without any novel category labels, following the standard open-vocabulary detection setting. The test set includes 10,554 base category instances and 7693 novel category instances, totaling 18,247 instances.

3.2. Comparative Methods

To verify the effectiveness of the our proposed method, the following state-of-the-art methods are selected:

- VILD [24] distills knowledge from a pre-trained open-vocabulary image classification model (teacher) into a two-stage detector (student). VILD uses the teacher model to encode category texts and image regions of object proposals, then trains a student detector whose region embeddings are aligned with text and image embeddings inferred by the teacher.
- GLIP [25] unifies object detection and phrase grounding for pre-training, enabling learning from both detection and grounding data. The method leverages 27M grounding data, including 3M human-annotated and 24M web-crawled image–text pairs.
- OV-DETR [26] is the first end-to-end Transformer-based open-vocabulary detector based on DETR architecture. OV-DETR formulates the learning objective as binary matching between input queries (class name or exemplar image) and corresponding objects, enabling detection of any object given its class name or an exemplar image.
- Detic [27] expands the vocabulary of detectors to tens of thousands of concepts by simply training the classifiers of a detector on image classification data. Unlike prior work, Detic does not need complex assignment schemes and is compatible with various detection architectures, achieving state-of-the-art results on the open-vocabulary LVIS benchmark.

- GroundingDINO [28] combines the Transformer-based detector DINO with grounded pre-training to detect arbitrary objects with human inputs such as category names or referring expressions.
- YOLO-World [29] enhances YOLO with open-vocabulary detection capabilities through vision–language modeling and pre-training on large-scale datasets. The method proposes Re-Parameterizable Vision–Language Path Aggregation Network (RepVL-PAN) and demonstrates superior efficiency for real-time applications.
- CastDet is the first open-vocabulary object detection framework designed for aerial images. The method employs a CLIP-activated student–teacher learning mechanism to detect objects without annotations.

3.3. Evaluation Metrics

To comprehensively evaluate the performance of our proposed open-vocabulary detection framework, we adopt standard object detection metrics with specific adaptations for the open-vocabulary setting. Our evaluation focuses on Precision (P), Recall (R), Mean Average Precision (mAP), and class-specific performance metrics for base and novel categories.

(1) Precision measures the accuracy of positive predictions, defined as the fraction of correct detections among all predictions:

$$P = \frac{TP}{TP + FP}, \tag{24}$$

where TP represents true positive detections and FP represents false positive detections.

(2) Recall quantifies the model's ability to identify all relevant objects, calculated as the fraction of ground-truth objects that are successfully detected:

$$R = \frac{TP}{TP + FN}, \tag{25}$$

where FN denotes false negative detections (ground-truth objects that were missed by the model).

(3) Average Precision (AP) summarizes the precision–recall trade-off by computing the area under the precision–recall curve:

$$\text{AP} = \int_0^1 P \cdot \text{RdR}, \tag{26}$$

where P is the precision as a function of recall R.

(4) Mean Average Precision (mAP) provides an overall performance measure by averaging AP across all categories:

$$mAP = \frac{1}{|C|} \sum_{c \in C} AP, \tag{27}$$

where C represents the set of all object categories and $|C|$ is the total number of categories. In our experiments, mAP is computed at an IoU threshold of 0.5.

(5) Base categories performance (mAP_{base}) specifically evaluates the detection performance on categories with available training annotations:

$$mAP_{base} = \frac{1}{|\mathcal{C}_{base}|} \sum_{c \in \mathcal{C}_{base}} AP, \tag{28}$$

where $\mathcal{C}_{base}$ denotes the set of base categories with ground-truth annotations during training.

(6) Novel category performance (mAP_{novel}) measures the model's open-vocabulary capability by evaluating performance on categories without training annotations:

$$mAP_{novel} = \frac{1}{|\mathcal{C}_{novel}|} \sum_{c \in \mathcal{C}_{novel}} AP, \tag{29}$$

where $\mathcal{C}_{novel}$ represents the set of novel categories that are only encountered during testing.

(7) Harmonic mean (HM) provides a balanced evaluation of both base and novel category performance by computing the harmonic mean of mAP_{base} and mAP_{novel}:

$$HM = \frac{2 \times mAP_{base} \times mAP_{novel}}{mAP_{base} + mAP_{novel}} \tag{30}$$

The HM metric is particularly valuable in open-vocabulary detection, as it penalizes approaches that achieve high performance on one category type at the expense of the other, ensuring a comprehensive assessment of the model's ability to handle both seen and unseen object categories.

mAP_{base}, mAP_{novel}, and HM are crucial for open-vocabulary evaluation, as they allow us to assess whether the proposed method maintains strong performance on known categories while effectively generalizing to unseen ones.

3.4. Implementation Details

Our method is implemented based on the MMDetection framework, using PyTorch version 1.10.0, with training and testing scripts in Python 3.8. We adopt Faster R-CNN as the detector backbone network, using ResNet-50 as the feature extractor. The model is initialized using pre-trained RemoteCLIP, and RemoteCLIP is frozen during the training process. Batch size is 8, trained on a single NVIDIA A100 GPU (40 GB VRAM, NVIDIA Corporation, Sunnyvale, CA, USA). The optimizer uses SGD with a learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.0001.

In the selective masking strategy, base category regions are masked using zero padding. The dynamic weight update interval is set to 100 iterations, and the smoothing coefficient α is set to 0.8. Long-term statistics of the pseudo-label queue are based on all accumulated pseudo-label data in the queue; the short-term statistics window is set to the most recent 500 updates.

Considering that the teacher model has weak foreground-background separation capability in the early training stages, directly using low-quality pseudo-labels would introduce noise and affect the student model's learning effectiveness. Specifically, the model is trained for a total of 40,000 iterations, with the first 30,000 iterations not using pseudo-labels generated by the teacher model. The first 30,000 iterations allow the model to focus on learning basic object detection capabilities, including foreground-background discrimination, object proposal generation, and other fundamental skills. In the early stages, the teacher model's pseudo-label quality is poor, and introducing pseudo-labels too early would generate numerous incorrect pseudo-labels due to insufficient teacher model capability, thereby interfering with the student model's learning process. Based on training experience from previous related work and loss analysis, basic object detection fundamentally converges at around 30,000 iterations. The 10,000 iterations of semi-supervised learning provide sufficient learning opportunities for novel categories, which also reference previous semi-supervised learning training methods.

4. Result and Discussion

4.1. Comparison with SOTA Methods

We compare our proposed method with several state-of-the-art approaches in the open-vocabulary detection field, including VILD, GLIP, OV-DETR, Detic, GroundingDINO, YOLO-World, and CastDet. As shown in Table 1, our method achieves significant performance improvements on the VisDroneZSD dataset. Specifically, our approach achieves 42.7% overall mAP, 43.5% base categories mAP, and 39.5% novel category mAP, with a harmonic mean of 41.4%.

Notably, our method demonstrates superior performance across all evaluation metrics compared to baseline methods. The 39.5% mAP_{novel} represents a substantial improvement over methods like VILD (14.2%) and GLIP (5.4%), showing relative improvements of 178.2% and 631.5%, respectively. This indicates that our enhanced framework with RemoteCLIP, selective masking strategy, and dynamic frequency weighting significantly improves the model's ability to detect novel categories in aerial imagery.

Compared to the original CastDet baseline (41.5% mAP_{novel}), our student model achieves competitive performance while maintaining better overall balance as reflected in the harmonic mean metric. The HM score of 41.4% demonstrates that our approach effectively balances performance between base and novel categories, avoiding the common trade-off where improvements in novel category detection come at the cost of base categories performance.

Table 1. Comparison of our model with other methods on VisDroneZSD.

Method	mAP	mAP_{base}	mAP_{novel}	HM
VILD	25.6	28.5	14.2	19.0
GLIP	33.8	41.0	5.4	9.5
OV-DETR	28.7	30.8	20.6	24.7
Detic	16.8	19.8	4.8	7.7
GroundingDINO	33.0	40.5	3.3	6.1
YOLO-World	32.9	39.1	8.5	13.9
CastDet	35.2	33.6	41.5	37.1
Ours	42.7	43.5	39.5	41.4

Extended experiments on the DIOR dataset (Table 2) further validate the generalization capability of our method. Our approach achieves 63.7% overall mAP, 70.1% base categories mAP, and 38.2% novel category mAP, with a harmonic mean of 49.5%. Our method outperforms all baseline approaches in terms of overall mAP and harmonic mean performance.

Table 2. Comparison of our model with other methods on DIOR.

Method	mAP	mAP_{base}	mAP_{novel}	HM
VILD	45.7	53.5	14.2	22.4
GLIP	56.0	69.1	3.6	6.9
OV-DETR	54.0	62.6	19.9	30.2
Detic	36.9	45.3	3.5	6.5
GroundingDINO	57.3	70.8	3.2	6.2
YOLO-World	57.7	70.2	8.0	14.4
CastDet	56.9	61.0	40.3	48.5
Ours	63.7	70.1	38.2	49.5

Specifically, our method shows substantial improvements over existing open-vocabulary detection methods on the DIOR dataset. Compared to VILD, which achieves

the most competitive overall performance among traditional methods with 45.7% mAP, our approach delivers a significant 39.4% relative improvement in overall mAP (from 45.7% to 63.7%). More impressively, for novel category detection, our method achieves 38.2% mAP_{novel} compared to VILD's 14.2%, representing a remarkable 169.0% relative improvement. The harmonic mean metric shows an even more dramatic enhancement, improving from 22.4% to 49.5% (a 120.9% relative improvement), demonstrating the balanced performance gains across both base and novel categories.

When compared to more recent methods, like GLIP and GroundingDINO, our approach demonstrates superior novel category detection capabilities. GLIP achieves 56.0% overall mAP but only 3.6% mAP_{novel}, while our method improves the overall mAP by 13.8% and achieves a massive 961.1% relative improvement in novel category detection. Similarly, GroundingDINO, despite achieving the highest base categories performance (70.8% mAP_{base}), performs poorly on novel categories (3.2% mAP_{novel}). Our method maintains a competitive base categories performance (70.1% mAP_{base}) while achieving 38.2% mAP_{novel}, resulting in a substantially higher harmonic mean (49.5% vs. 6.2%).

Compared to the original CastDet baseline, our enhanced framework shows consistent improvements across all metrics. We achieve 11.9% relative improvement in overall mAP (from 56.9% to 63.7%) and 14.9% improvement in base categories mAP (from 61.0% to 70.1%). Although CastDet demonstrates strong novel category performance (40.3% mAP_{novel}), this is achieved by selecting images containing novel categories for unsupervised training, which requires extensive dataset preprocessing and strict data separation. In contrast, our method eliminates the need for such dataset filtering and can directly utilize partially annotated data without prior separation while still maintaining competitive novel category performance (38.2% mAP_{novel}) and significantly improving the overall balance, as reflected in the harmonic mean improvement from 48.5% to 49.5%. The notable improvement in base categories performance can be attributed to our ability to simultaneously train on both base and novel categories within the same images, eliminating the data waste inherent in CastDet. Specifically, for CastDet, when an image contains novel category objects, the entire image must be used for unsupervised learning, resulting in the loss of valuable base category annotations.

4.2. Ablation Study

To validate the effectiveness of our proposed components, we conduct comprehensive ablation studies focusing on the key innovations of our framework.

4.2.1. Effectiveness of Selective Masking Strategy

Table 3 demonstrates the results of the ablation study, comparing different masking strategies. Without any masking strategy, the student model achieves an overall mAP of 37.6%, a base categories mAP of 39.4%, a novel category mAP of 30.5%, and a harmonic mean of 34.4%. The Gaussian blur masking strategy shows improvement with an overall mAP of 40.3%, a base categories mAP of 41.1%, a novel category mAP of 37.1%, and a harmonic mean of 39.0%. Our proposed zero padding masking strategy achieves the best performance with an overall mAP of 42.7%, a base categories mAP of 43.5%, a novel category mAP of 39.5%, and a harmonic mean of 41.4%.

Table 3. Ablation study of different masking strategies on VisDroneZSD.

Masking Strategy	mAP	mAP_{base}	mAP_{novel}	HM
No Masking	37.6	39.4	30.5	34.4
Gaussian Blur	40.3	41.1	37.1	39.0
Zero Padding	42.7	43.5	39.5	41.4

The progressive improvements demonstrate the effectiveness of different masking approaches. Compared to non-masking, zero padding achieves a relative improvement of 13.6% in overall mAP, 10.4% in base categories detection, and a substantial improvement of 29.5% in the detection of novel categories. More importantly, zero padding outperforms Gaussian blur with a 5.96% relative improvement in overall mAP and a 6.47% improvement in novel category detection, validating our theoretical analysis regarding complete information suppression.

These results confirm that by completely masking base category regions through zero padding, the teacher model more effectively focuses on discovering novel category targets. The generated high-quality pseudo-labels, when merged with accurate ground-truth annotations of base categories, provide more complete and consistent supervision signals for the student model. The superior performance of zero padding over Gaussian blur validates our information-theoretic approach, demonstrating that complete information suppression is more effective than partial information degradation for novel category discovery.

4.2.2. Effectiveness of Queue-Based Dynamic Frequency Weighting

Table 4 presents a comprehensive comparison of different weighting strategies applied to address class imbalance in novel category detection. The results clearly demonstrate the progressive improvement achieved by more sophisticated weighting mechanisms.

Table 4. Ablation study of dynamic frequency weighting on VisDroneZSD.

Method	mAP	$\mathbf{mAP_{base}}$	$\mathbf{mAP_{novel}}$	HM
Static Weight	37.1	41.8	18.3	25.4
CB Loss	39.2	41.7	29.1	34.3
Ours	42.7	43.5	39.5	41.4

When employing static weights where all categories are assigned uniform importance (w = 1.0), the model achieves modest performance with 37.1% overall mAP and only 18.3% mAP_{novel}. The integration of Class-Balanced Loss, which adjusts weights based on effective sample numbers computed from the dynamic label queue, provides notable improvement, achieving 39.2% overall mAP and 29.1% mAP_{novel}. However, our proposed dynamic frequency weighting mechanism yields the most significant performance gains, reaching 42.7% overall mAP and 39.5% mAP_{novel}.

The quantitative analysis reveals several key insights. Compared to static weighting, our dynamic approach delivers a substantial 21.2 percentage point improvement in novel category detection (from 18.3% to 39.5%), representing a remarkable 115.8% relative enhancement. Even when benchmarked against the established CB Loss method, our information-theoretic weighting strategy maintains a significant edge with 10.4 percentage points improvement in mAP_{novel} (from 29.1% to 39.5%) and 3.5 percentage points improvement in overall mAP. The harmonic mean shows consistent improvement from 25.4% (static) to 34.3% (CB Loss) to 41.4% (ours), demonstrating better balanced performance across base and novel categories.

Table 5 provides detailed per-category analysis, revealing both the effectiveness and limitations of dynamic weighting for addressing class imbalance issues. The results are particularly striking for the most challenging rare categories. For airport detection, our method achieves 38.5% AP with 69.1% recall, compared to complete failure (0.0% AP and recall) with static weighting. The basketball court and ground track field also show consistent improvements across all methods, with our approach achieving the best performance (65.1% AP and 41.1% AP, respectively).

Table 5. Novel category performance on VisDroneZSD.

Method	Airport		Basketball Court		Ground Track Field		Windmill	
	AP	Recall	AP	Recall	AP	Recall	AP	Recall
Static Weight	0.0	0.0	39.0	79.4	34.2	55.2	0.0	0.0
CB Loss	24.8	45.5	45.1	79.1	37.5	55.6	9.1	3.3
Ours	38.5	69.1	65.1	79.7	41.1	56.3	13.2	18.1

However, the analysis also reveals persistent challenges in extremely rare categories. Windmill detection, which represents the most severe class imbalance challenge, shows limited improvement with our method achieving only 13.2% AP and 18.1% recall. While this represents a significant improvement over static weighting (complete failure) and CB Loss, the recall rate remains substantially lower compared to other novel categories.

To understand the root causes of this performance gap, we conducted detailed quality analysis of pseudo-labels across novel categories. We sampled 300 instances from each novel category during VisDroneZSD training to analyze confidence scores. The results reveal that the windmill has an average confidence of only 42.2%, while the basketball court achieves 78.1%, corresponding with the confidence patterns shown in Figure 3. Additionally, statistical analysis of the VisDroneZSD test set shows that windmill targets have an average size of only 2807 pixels, compared to the basketball court at 18,327 pixels and the ground track field at 54,760 pixels, with the windmill being nearly 20 times smaller than the ground track field.

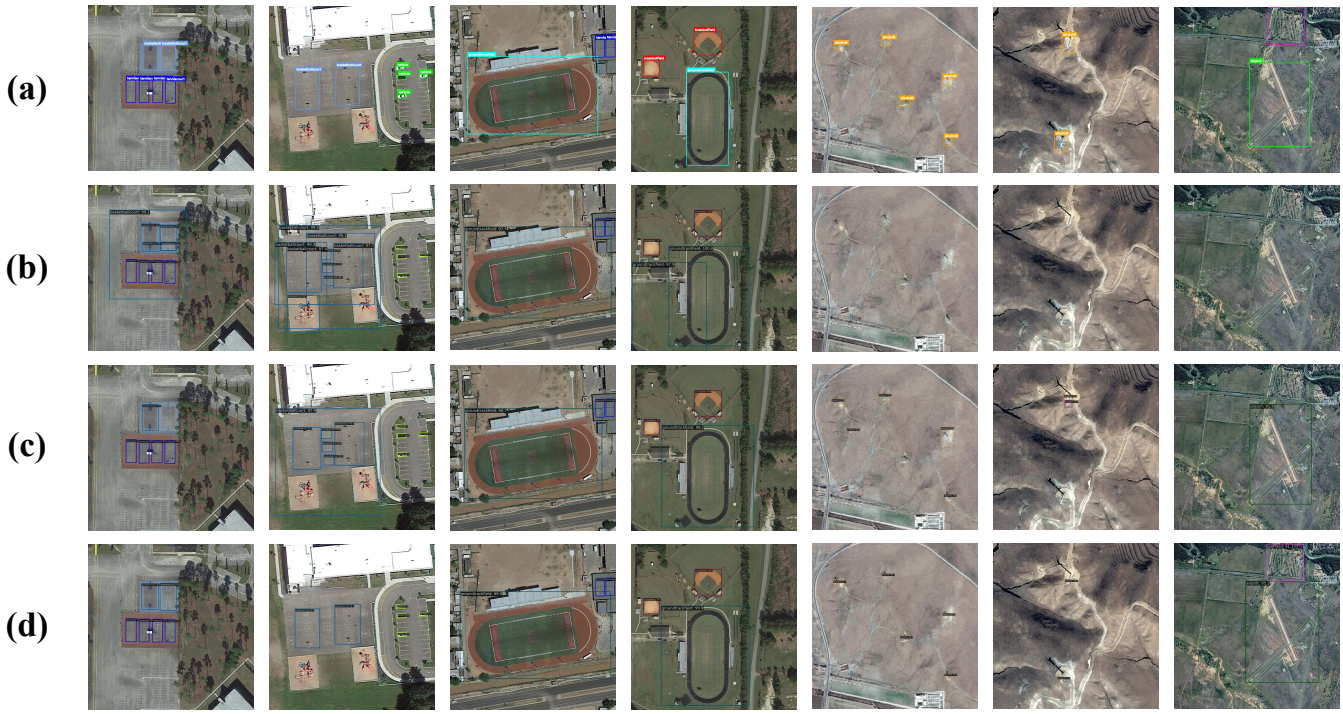

Figure 3. Qualitative comparison of different weighting strategies on novel category detection. Row (**a**) shows ground-truth annotations, row (**b**) presents results with static weighting (w = 1.0 for all categories), row (**c**) shows CB Loss weighting results, and row (**d**) demonstrates our dynamic frequency weighting results. The visualization spans seven representative images containing novel categories: airport, basketball court, ground track field, and windmill.

We believe the fundamental causes are attributed to the following: Small object problem—windmill targets are significantly smaller than other categories, leading to insufficient feature extraction and limited model recognition capability on small-scale targets. Since standard geometric bounding boxes are used, the actual pixels occupied by windmills are far smaller than the statistical results, as they are mostly composed of linear structures, further exacerbating the difficulty for RPN extraction.

Based on these findings, we propose specific improvement directions for future research: integrating small object detection enhancement modules, such as multi-scale feature fusion and specialized feature extractors for small targets, and introducing context-aware mechanisms that leverage geographic and environmental context information to assist small target identification.

Figure 3 provides a qualitative visualization of the effectiveness of different weighting strategies on novel category detection. The results corroborate the quantitative findings presented in Tables 4 and 5, demonstrating clear visual improvements across all categories.

As shown in the figure, the static weighting approach (row a) fails completely to detect the most challenging categories, airport and windmill, producing no detection results for these rare object types. This complete failure corresponds to the 0 AP values reported in Table 5 for these categories. Furthermore, the static weighting method suffers from numerous false positives in basketball court and ground track field detection, indicating poor discrimination capability. The CB Loss method (row b) shows improved performance compared to static weighting, successfully detecting some instances of airport and windmill categories, but still exhibits suboptimal localization accuracy and confidence scores.

In contrast, our proposed dynamic frequency weighting method (row c) achieves the most accurate detection results with precise bounding box localization and minimal false negatives. Notably, the windmill category shows significant confidence score improvements compared to CB Loss, demonstrating the effectiveness of our information-theoretic weighting approach in addressing severe class imbalance. The visual results confirm the quantitative analysis, showing that our method not only improves detection recall for rare categories but also enhances overall detection precision across all novel object types.

Figure 4 illustrates the frequency evolution of each novel category in the pseudo-label queue during the training process, clearly revealing the dynamic characteristics of the class imbalance problem. The horizontal axis represents the training epochs during the pseudo-label introduction phase. The following key observations can be made from the figure: The ground track field dominates in the early training stage with a frequency as high as 0.75, but gradually decreases to 0.35 as training progresses. The basketball court exhibits the opposite trend, rising from an initial frequency of 0.25 to 0.65, ultimately becoming the most frequently detected category. Meanwhile, the airport and windmill categories maintain frequencies close to 0 throughout the entire training process, demonstrating severe scarcity.

This dynamic change in frequency distribution fully validates the necessity of our proposed dynamic weight adjustment mechanism. Traditional static weighting schemes cannot adapt to such time-varying category distributions, nor can they accommodate the variations of rare categories across different datasets. Our method, by real-time monitoring of frequency changes and correspondingly adjusting weights, can provide higher learning weights for rare categories (such as airport and windmill), ensuring that the model is not dominated by frequent categories. This effectively improves the detection performance of these difficult categories, which also explains why we observe performance improvements from the dynamic weighting mechanism in the ablation experiments shown in Table 4.

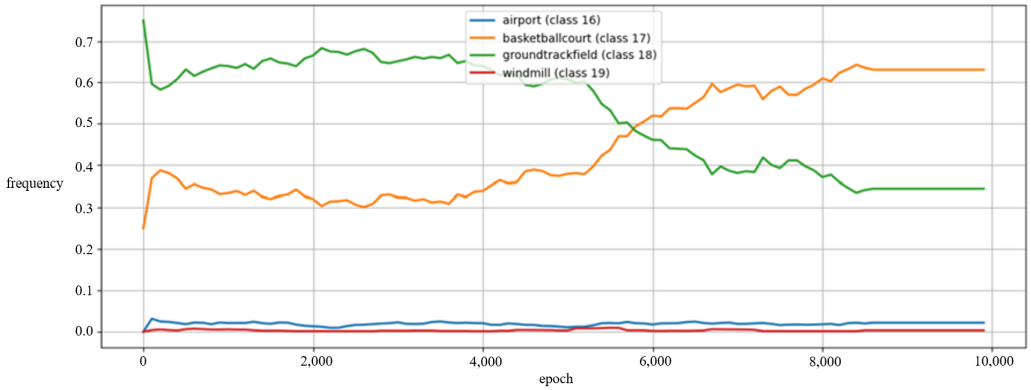

Figure 4. Class frequency evolution in the pseudo-label queue during training. The figure shows the dynamic changes in detection frequency for four novel categories (airport, basketball court, ground track field, and windmill) across training epochs, highlighting the temporal class imbalance patterns that motivate our dynamic weighting approach.

4.2.3. Parameter Sensitivity Analysis

Impact of the information intensity coefficient: To investigate the impact of the information intensity coefficient μ on our dynamic weighting mechanism, we conduct a comprehensive sensitivity analysis, as shown in Table 6. The coefficient μ controls the influence of self-information on the weight calculation, where $\mu = 0$ indicates that novel category weights are uniformly set to 1.0 (equivalent to no dynamic adjustment).

Table 6. Parameter sensitivity analysis of the information intensity coefficient μ on VisDroneZSD.

μ	mAP	mAP_{base}	$\text{mAP}_{\text{novel}}$	HM
0	37.1	41.8	18.3	25.4
2	41.1	43.2	32.6	37.2
4	42.7	43.5	39.5	41.4
8	41.9	43.3	36.3	39.5
16	37.2	39.9	26.5	31.8

The experimental results reveal a clear pattern in performance variation across different μ values. When $\mu = 0$, representing the absence of dynamic weight adjustment, the model achieves baseline performance with 37.1% overall mAP and 18.3% mAP_{novel}. As μ increases to 2, substantial improvements are observed across all metrics, with mAP_{novel} rising to 32.6% (78.1% relative improvement) and the harmonic mean increasing from 25.4% to 37.2%.

The optimal performance is achieved at $\mu = 4$, where the model reaches peak performance with 42.7% overall mAP, 39.5% mAP_{novel}, and 41.4% harmonic mean. This optimal value demonstrates that moderate weight adjustment based on self-information provides the best balance between emphasizing rare categories and maintaining training stability. The substantial improvement from $\mu = 0$ to $\mu = 4$ (115.8% relative improvement in mAP_{novel}) validates the effectiveness of our information-theoretic approach.

However, further increasing μ to 8 and 16 leads to performance degradation, with mAP_{novel} dropping to 36.3% and 26.5%, respectively. When μ is too small, the weight adjustment is insufficient to address the severe class imbalance, resulting in continued dominance by frequent categories. Conversely, when μ is too large, excessive weight adjustment leads to training instability and reduced overall performance, as the loss becomes dominated by potentially noisy pseudo-labels from rare categories. The performance decline at $\mu = 16$ (26.5% mAP_{novel}) demonstrates that overly aggressive weighting can be counterproductive.

These findings confirm that $\mu = 4$ provides the optimal trade-off between addressing class imbalance and maintaining robust training dynamics.

The analysis provides valuable insights for practitioners implementing our dynamic frequency weighting mechanism. The results demonstrate that selecting an appropriately moderate value of μ (around 4) is sufficient to achieve substantial performance improvements without requiring extensive hyperparameter tuning. Importantly, the findings suggest that pursuing excessively large μ values is not only unnecessary but also potentially detrimental to model performance. This robustness to parameter selection makes our method practically applicable across different datasets and scenarios, as users can expect reliable improvements by choosing μ values in the moderate range (2–4) rather than engaging in aggressive parameter optimization.

Impact of the Weight Update Interval: Table 7 presents the impact of different weight update intervals on both model performance and computational efficiency. The update interval determines how frequently the dynamic weights are recalculated based on the current dynamic label queue statistics, balancing between responsiveness to distribution changes and computational overhead.

Table 7. Parameter sensitivity analysis of the weight update interval on VisDroneZSD.

Weight Update Interval	Training Time/iter	mAP	mAP_{base}	$\text{mAP}_{\text{novel}}$	HM
25	0.878s	42.4	43.2	39.4	41.2
50	0.872s	42.5	43.1	40.1	41.5
100	0.865s	42.7	43.5	39.5	41.4
500	0.860s	41.9	43.3	36.3	39.5

The experimental results demonstrate that moderate update frequencies yield optimal performance. With an update interval of 50–100 iterations, the model achieves peak performance, with the 50-iteration interval reaching 40.1% mAP_{novel} and 41.5% harmonic mean, while the 100-iteration interval achieves 42.7% overall mAP. These frequencies provide an effective balance between timely response to class distribution changes and training stability.

When the update interval is too frequent (25 iterations), although the model responds quickly to distribution changes, the performance shows slight degradation (39.4% mAP_{novel}), potentially due to weight fluctuations that introduce training instability. Conversely, when the update interval is too sparse (500 iterations), significant performance deterioration occurs, with mAP_{novel} dropping to 36.3% and the harmonic mean declining to 39.5%. This demonstrates that infrequent updates fail to adequately respond to class distribution changes and cannot fully exploit the advantages of dynamic adjustment, resulting in limited improvement for novel categories.

An important practical consideration is the computational overhead introduced by the dynamic weighting mechanism. The training time analysis reveals that the weight update frequency has minimal impact on computational cost, with per-iteration training times ranging from 0.860 s to 0.878 s across all tested intervals. The difference between the most frequent (25 iterations) and least frequent (500 iterations) update schedules is merely 0.018 s per iteration. For a typical training scenario of 10,000 iterations, this translates to approximately 3 min additional training time, which is negligible considering the substantial performance improvements achieved. This analysis confirms that our dynamic weighting mechanism introduces minimal computational burden while providing significant performance gains, making it highly practical for real-world applications.

Impact of the Smoothing Coefficient: Table 8 presents the impact of different smoothing coefficient values on the stability and responsiveness of our dynamic weighting mech-

anism. The parameter sensitivity analysis for the smoothing coefficient α reveals the importance of our dynamic loss weighting mechanism. The results demonstrate a clear inverted-U-shaped performance curve across different α values.

Table 8. Parameter sensitivity analysis of the smoothing coefficient α on VisDroneZSD.

α	mAP	mAP_{base}	$\text{mAP}_{\text{novel}}$	HM
0	40.1	40.8	37.3	39.0
0.5	41.6	42.2	39.2	40.6
0.8	42.7	43.5	39.5	41.4
0.95	39.1	40.5	33.5	36.7

When $\alpha = 0$ (no smoothing), the model achieves moderate performance with 40.1% overall mAP and 37.3% $\text{mAP}_{\text{novel}}$. This setting allows weights to change drastically based on immediate results, leading to training instability and suboptimal convergence. The relatively poor performance indicates that excessive responsiveness to short-term fluctuations harms the learning process.

As α increases to 0.5, performance improves significantly to 41.6% overall mAP and 39.2% $\text{mAP}_{\text{novel}}$, demonstrating that moderate smoothing helps stabilize the training process while maintaining reasonable adaptability to distribution changes.

The optimal performance is achieved at $\alpha = 0.8$, reaching 42.7% overall mAP and 39.5% $\text{mAP}_{\text{novel}}$. This value provides an effective balance between maintaining historical weight information and adapting to recent distribution changes, ensuring stable convergence while preserving sufficient responsiveness to evolving pseudo-label statistics.

When α becomes too large (0.95), performance degrades significantly to 39.1% overall mAP and 33.5% $\text{mAP}_{\text{novel}}$. This excessive smoothing makes the weighting mechanism overly conservative, failing to adapt promptly to changing class distributions and essentially approaching static weighting behavior. The substantial performance drop confirms that insufficient responsiveness limits the effectiveness of our dynamic adjustment strategy.

Impact of the Short-Term Frequency Window Size: Table 9 presents the impact of different short-term frequency window sizes on model performance. The results demonstrate a clear performance pattern that validates our choice of window size for the dynamic loss weighting mechanism.

Table 9. Parameter sensitivity analysis of the short-term frequency window size on VisDroneZSD.

Window Size	mAP	mAP_{base}	$\text{mAP}_{\text{novel}}$	HM
50	41.7	42.7	37.8	40.1
300	42.2	42.9	39.6	41.2
500	42.7	43.5	39.5	41.4
5000	40.5	41.8	35.2	38.2

When the window size is too small (about 50 updates), the model achieves suboptimal performance with 41.7% overall mAP and 37.8% $\text{mAP}_{\text{novel}}$. This reduced performance indicates that overly small windows make the frequency statistics too sensitive to short-term fluctuations, leading to unstable weight adjustments that can negatively impact training.

The window size of 300 updates shows improved performance compared to 50, achieving 42.2% overall mAP and 39.6% $\text{mAP}_{\text{novel}}$. It is already similar to the best outcome.

Our chosen window size of 500 updates achieves the best overall performance with 42.7% overall mAP. This configuration effectively balances short-term adaptability with sufficient stability to avoid noise interference, confirming the appropriateness of our parameter selection.

When the window becomes excessively large (5000 updates), performance degrades significantly to 40.5% overall mAP and 35.2% mAP_{novel}. This substantial decline indicates that overly large windows reduce the mechanism's responsiveness to recent distribution changes, losing the benefits of dynamic adaptation.

5. Conclusions

We present a one-step training open-vocabulary detection framework that supports partially annotated data, effectively addressing key challenges faced by existing methods in aerial image open-vocabulary detection, including complex data filtering processes, low data utilization efficiency, and novel category imbalance issues. Our main contributions encompass the following three aspects:

First, we propose a selective masking strategy and label merging mechanism that eliminates the constraint of strict data separation required by existing methods. This approach no longer necessitates pre-partitioning images containing novel categories, significantly improving data utilization efficiency. Second, we design a frequency-based dynamic weighting mechanism built upon a dynamic pseudo-label queue. Through long-term and short-term frequency statistics combined with exponential moving average updates, this mechanism adaptively addresses imbalance issues among novel categories, ensuring that rare categories receive adequate learning attention. Finally, we develop a single-stage weighted supervision approach that avoids complex multi-stage training and loss-balancing problems between different learning paradigms, achieving end-to-end optimization.

Compared to existing frameworks, our approach offers a distinct technical route for open-vocabulary remote sensing detection. While LAE-DINO focuses on dynamic vocabulary construction through massive pre-training, our method addresses vocabulary expansion through intelligent utilization of existing limited annotations. OpenRSD's multi-modal prompt framework requires careful prompt engineering and supports various input modalities. Our method offers a more automated approach that adapts to data distribution without manual prompt tuning, making it more suitable for scenarios with limited domain expertise. Importantly, our dynamic weighted loss could be integrated into other frameworks when facing class imbalance issues.

Despite achieving significant progress, our method still has some limitations. For instance, generating high-quality pseudo-labels for certain novel categories remains challenging, which exacerbates class imbalance issues. Future research directions include improving pseudo-label generation quality, incorporating small object detection enhancement modules, and using context-aware detection mechanisms that leverage geographic and environmental information to address the unique challenges of aerial image interpretation and enhance the overall performance of open-vocabulary detection in remote sensing imagery.

Author Contributions: Conceptualization, S.W.; methodology, S.W. and R.F.; validation, S.W.; formal analysis, S.W. and J.X.; investigation, S.W., Y.C., and R.F.; writing—original draft, S.W. and Y.S.; writing—review and editing, S.W., Y.S., and R.F.; visualization, S.W.; supervision, R.F. and P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Fundamental Research Foundation of National Key Laboratory of Automatic Target Recognition under Grant JKWATR-240301.

Data Availability Statement: The data used in this study are public datasets. DIOR dataset can be obtained from https://opendatalab.org.cn/OpenDataLab/DIOR (accessed on 8 October 2025). VisDroneZSD can be obtained from http://aiskyeye.com/challenge-2023/zero-shot-object-detection/ (accessed on 8 October 2025).

Acknowledgments: The authors would like to express their gratitude for the valuable feedback and suggestions provided by all the anonymous reviewers and the editorial team.

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

Abbreviations

The following abbreviations are used in this manuscript:

RSOD	Remote sensing object detection
OVD	Open-vocabulary object detection

References

1. Zhang, X.; Zhang, T.; Wang, G.; Zhu, P.; Tang, X.; Jia, X.; Jiao, L. Remote Sensing Object Detection Meets Deep Learning: A Metareview of Challenges and Advances. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 8–44. [CrossRef]
2. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
3. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
4. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
5. Gui, S.; Song, S.; Qin, R.; Tang, Y. Remote Sensing Object Detection in the Deep Learning Era—A Review. *Remote Sens.* **2024**, *16*, 327. [CrossRef]
6. Li, Z.; Wang, Y.; Zhang, N.; Zhang, Y.; Zhao, Z.; Xu, D.; Ben, G.; Gao, Y. Deep Learning-Based Object Detection Techniques for Remote Sensing Images: A Survey. *Remote Sens.* **2022**, *14*, 2385. [CrossRef]
7. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858.
8. Yang, X.; Yan, J.; Feng, Z.; He, T. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; pp. 3163–3171.
9. Qian, W.; Yang, X.; Peng, S.; Zhang, X.; Yan, J. RSDet++: Point-Based Modulated Loss for More Accurate Rotated Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 7869–7879. [CrossRef]
10. Liang, Y.; Ren, Y.; Yu, J.; Zha, W. Current Trajectory Image-Based Protection Algorithm for Transmission Lines Connected to MMC-HVDC Stations Using CA-CNN. *Prot. Control Mod. Power Syst.* **2023**, *8*, 6. [CrossRef]
11. Sullivan, A.; Lu, X. ASPP: A New Family of Oncogenes and Tumour Suppressor Genes. *Br. J. Cancer* **2007**, *96*, 196–200. [CrossRef] [PubMed]
12. Sahu, R.K.; Gorripotu, T.S.; Panda, S. A Hybrid DE–PS Algorithm for Load Frequency Control under Deregulated Power System with UPFC and RFB. *Ain Shams Eng. J.* **2015**, *6*, 893–911. [CrossRef]
13. Rabbi, J.; Ray, N.; Schubert, M.; Chowdhury, S.; Chao, D. Small-Object Detection in Remote Sensing Images with End-to-End Edge-Enhanced GAN and Object Detector Network. *Remote Sens.* **2020**, *12*, 1432. [CrossRef]
14. Min, L.; Fan, Z.; Lv, Q.; Reda, M.; Shen, L.; Wang, B. YOLO-DCTI: Small Object Detection in Remote Sensing Base on Contextual Transformer Enhancement. *Remote Sens.* **2023**, *15*, 3970. [CrossRef]
15. Wu, J.; Li, X.; Xu, S.; Yuan, H.; Ding, H.; Yang, Y.; Li, X.; Zhang, J.; Tong, Y.; Jiang, X.; et al. Towards Open Vocabulary Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5092–5113. [CrossRef] [PubMed]
16. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and a New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
17. AISKYEYE Team at Lab of Machine Learning and Data Mining, Tianjin University, China. Zero-Shot Object Detection Challenge. 2023. Available online: http://aiskyeye.com/challenge-2023/zero-shot-object-detection/ (accessed on 8 October 2025).
18. Zareian, A.; Rosa, K.D.; Hu, D.H.; Chang, S.F. Open-Vocabulary Object Detection Using Captions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Event, 20–25 June 2021; pp. 14393–14402.
19. Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; Zhou, J. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5622216. [CrossRef]
20. Zhang, Z.; Zhao, T.; Guo, Y.; Yin, J. RS5M and GeoRSCLIP: A Large-Scale Vision- Language Dataset and a Large Vision-Language Model for Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5642123. [CrossRef]

21. Li, Y.; Guo, W.; Yang, X.; Liao, N.; He, D.; Zhou, J.; Yu, W. Toward Open Vocabulary Aerial Object Detection with CLIP-Activated Student-Teacher Learning. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 431–448.
22. Pan, J.; Liu, Y.; Fu, Y.; Ma, M.; Li, J.; Paudel, D.P.; Van Gool, L.; Huang, X. Locate Anything on Earth: Advancing Open-Vocabulary Object Detection for Remote Sensing Community. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Philadelphia, PA, USA, 25 February–4 March 2025; pp. 6281–6289.
23. Huang, Z.; Feng, Y.; Yang, S.; Liu, Z.; Liu, Q.; Wang, Y. OpenRSD: Towards Open-Prompts for Object Detection in Remote Sensing Images. *arXiv* **2025**, arXiv:2503.06146.
24. Gu, X.; Lin, T.-Y.; Kuo, W.; Cui, Y. Open-Vocabulary Object Detection via Vision and Language Knowledge Distillation. *arXiv* **2021**, arXiv:2104.13921.
25. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. Grounded Language-Image Pre-Training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 10965–10975.
26. Zang, Y.; Li, W.; Zhou, K.; Huang, C.; Loy, C.C. Open-Vocabulary DETR with Conditional Matching. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 106–122.
27. Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; Misra, I. Detecting Twenty-Thousand Classes Using Image-Level Supervision. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 350–368.
28. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 38–55.
29. Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; Shan, Y. YOLO-World: Real-Time Open-Vocabulary Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 16901–16911.