remote sensing

MDPI

Article

Addressing Dense Small-Object Detection in Remote Sensing: An Open-Vocabulary Object Detection Framework

Menghan Ju [1,2,3,4], Yingchao Feng [1,4,*], Wenhui Diao [1,4] and Chunbo Liu [1,4]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; jvmenghan23@mails.ucas.ac.cn (M.J.); diaowh@aircas.ac.cn (W.D.); liucb@aircas.ac.cn (C.L.)
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
[4] Key Laboratory of Target Cognition and Application Technology (TCAT), Beijing 100094, China
* Correspondence: fengyc@aircas.ac.cn

Highlights

What are the main findings?

- An open-vocabulary object detection method tailored for complex remote sensing scenarios has been proposed. By incorporating text information as supplementary data, it effectively mitigates the occurrence of missed detections and false positives for densely clustered small targets.
- Compared to the most advanced methods in recent years, this model demonstrates superior detection results in terms of detection accuracy, scalability, and classification reliability for remote sensing images.

What are the implications of the main findings?

- This study proposes a novel open-vocabulary object detection method tailored for remote sensing imagery. By incorporating techniques such as Multi-scale Large Kernel Attention and convolutional gated feedforward networks, the approach effectively identifies high-density objects within complex backgrounds. This holds significant implications for practical applications including urban planning and environmental monitoring.
- The model maintains high detection accuracy even when the dataset size is reduced to a quarter of its original size, demonstrating exceptional data scalability.

Abstract

Remote sensing open-vocabulary object detection focuses on identifying and localizing unseen categories within remote sensing imagery. However, constrained by characteristics such as dense target distribution, complex background interference, and drastic scale variations inherent to remote sensing scenarios, existing methods are prone to background noise interference when extracting features from dense, small target regions. This leads to weakened semantic representation and reduced localization accuracy. Therefore, we propose RS-DINO to address these challenges. Specifically: Firstly, to address the issue of small features being obscured by the background, the feature extraction module incorporates a multi-scale large-kernel attention mechanism. This expands the receptive field while enhancing local detail modelling, significantly improving the feature representation of minute targets. Secondly, a cross-modal feature fusion module employing bidirectional cross-attention achieves deep alignment between image and textual features. Subsequently, a language-guided query selection mechanism enhances detection accuracy through hybrid query strategies. Finally, to enhance the spatial sensitivity and channel adaptability of fusion features, the multimodal decoder integrates a convolutional gated

Check for updates

Academic Editor: Wen Yang

Received: 2 February 2026
Revised: 8 March 2026
Accepted: 9 March 2026
Published: 10 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

feedforward network, significantly boosting the model's robustness in dense, multi-scale scenes. Experiments on DIOR, DOTA v2.0, and NWPU-VHR10 demonstrate substantial gains, with fine-tuned RS-DINO surpassing existing methods by 3.5%, 3.7%, and 4.0% in accuracy, respectively.

Keywords: open-vocabulary object detection; convolutional gated feedforward neural networks; remote sensing

1. Introduction

As an indispensable component of geospatial image processing workflows, the object detection task focuses on the robust detection and pixel-level localization of region-specific targets across multispectral or panchromatic imagery. Owing to its pivotal role, this capability serves as a critical enabler for diverse real-world applications, spanning sustainable urban planning [1–3], dynamic target tracking [4,5], intelligent traffic management [6,7], and large-scale environmental monitoring [8–10].

Contemporary object detection research has been largely confined to closed-set paradigms, wherein models are trained exclusively on datasets containing predefined object categories and deployed solely to recognize instances within these fixed classes. This approach, however, encounters profound limitations in real-world remote sensing contexts: the extraordinary diversity and near-infinite variability of geospatial objects frequently transcend rigid category boundaries. Compounding this challenge, integrating novel categories or refining existing ones demands exhaustive re-annotation of training data and full model retraining—procedures that are prohibitively time-consuming and computationally expensive. These critical constraints have catalyzed the emergence of open-vocabulary object detection (OVOD) for remote sensing: a flexible, scalable framework capable of recognizing both seen and unseen object categories with significantly enhanced adaptability and practical utility.

Recent scholarly efforts have yielded notable refinements in OVOD paradigms, enhancing their applicability to complex, real-world visual domains beyond controlled datasets. Pioneering this direction, Li et al. introduced BLIP-2 [11], a versatile pre-training framework that leverages frozen pretrained vision encoders and large language models (LLMs) for vision-language alignment. Its core innovation lies in an efficient query-based Transformer that effectively bridges the modality gap with minimal computational overhead. Subsequently, Li et al. developed GLIP [12], which unifies object detection and phrase grounding through joint vision-language training. By utilizing textual prompts to guide object localization, GLIP expands the visual semantic space, enhances cross-category generalization, and enables zero-shot detection of novel categories without requiring additional manual annotation. Building upon these foundations, Liu et al. proposed GroundingDINO [13], an efficient Transformer-based framework that integrates pretrained vision-language models. Through fine-grained fusion of multimodal features, it significantly improves cross-domain image-text alignment, delivering robust localization capabilities for both seen and unseen object categories.

Most existing methods are designed for objects in natural scenes, yielding limited effectiveness in more open and complex remote sensing scenarios. Recent scholarly endeavors have increasingly prioritized the development of OVOD methodologies tailored for remote sensing applications. Notably, Pan et al. [14] introduced LAE-DINO, a framework synergizing two novel components: the Dynamic Vocabulary Construction (DVC) mechanism and the Visually Guided Text Prompt Learning (VisGT) module. The DVC

operates at the batch level to generate context-specific lexicons during training, while VisGT implements a cross-modal projection strategy that translates visual embeddings into a semantically enriched text representation space. Complementarily, Wang et al. [15] proposed LLaMA-Unidetector, an architecture decoupling detection and recognition through a two-stage pipeline. The initial phase employs a class-agnostic detector to isolate foreground instances from background clutter; the second phase performs category inference via TerraOV-LLM—a multimodal large language model (MLLM) harnessing pre-trained linguistic knowledge to resolve unseen object classes in complex remote sensing scenes. Recently, with the rapid advancement of vision-language pretraining, open-vocabulary object detection has emerged as a prominent paradigm. These methods aim to detect and recognize categories beyond the labeled training vocabulary.

Despite marked advancements in contemporary open-vocabulary object detection (OVOD) methodologies for remote sensing, persistent constraints continue to challenge robust deployment across diverse operational scenarios. Current OVOD methods encounter significant difficulties when applied to remote sensing imagery—primarily due to the dense spatial distribution of targets, pervasive background clutter, and extreme scale variations. These factors induce visual features with pronounced multi-scale inconsistency and elevated noise levels, hindering effective alignment with textual embeddings and thereby degrading semantic representation fidelity and detection accuracy.

In response to the identified methodological constraints, we present RS-DINO—a specialized architecture meticulously formulated to advance open-vocabulary object recognition within remote sensing visual data. The proposed system encompasses four pivotal components: Feature Extraction, Cross-Modal Features Fusion, Language-Guided Query Selection, and Multimodal Decoder. Our optimization initiates at the visual encoding phase: although prevailing methodologies frequently deploy Swin-Transformer as the foundational backbone for multi-level feature extraction, its efficacy in modeling intricate spatial configurations and hierarchical contextual relations proves inadequate when confronted with high-resolution remote sensing inputs featuring densely clustered instances, pronounced scale heterogeneity, and visually complex backgrounds. To address this limitation, the Multi-Scale Large-Kernel Attention (MSLKA) mechanism is integrated to enhance the expressiveness of visual features and accurately model interactions between fine-grained local structures and holistic scene semantics. Building upon this, the Cross-Modal Representation Alignment module employs a bidirectional attention mechanism to forge coherent semantic mappings between visual embeddings and linguistic descriptors. Concurrently, the Language-Informed Query Formulation strategy adaptively pinpoints regions of interest by evaluating cross-modal semantic affinities, thereby preserving multimodal integrity while emphasizing regions with joint visual prominence and textual significance. The pipeline culminates in the Unified Multimodal Decoding Unit, which synthesizes heterogeneous feature streams into a consolidated representation. Embedded within this unit, a Convolutional Gated Feedforward Network (CGFFN) refines feature integration dynamics, ultimately yielding precise outputs for object localization and semantic classification.

The contributions of this study can be outlined as follows:

- This research presents a multi-scale open-vocabulary detection architecture specifically designed for remote sensing imagery. The framework specifically addresses the cross-domain adaptation challenge arising from intrinsic semantic feature mismatches between remote sensing and natural image domains, thereby achieving substantial improvements in detection precision and system resilience.
- To simultaneously enhance the backbone's ability to interpret holistic scenes and discriminate fine-grained structures, we integrate the Multi-scale Large-Kernel Attention (MSLKA) module. This mechanism significantly improves the network's capacity to

align hierarchical semantics with spatial patterns across diverse remote sensing inputs, thereby substantially enriching visual feature representation.

- A convolutional gated feedforward neural network is incorporated into the multimodal decoder to optimize the integration of this information, thereby generating more accurate predictions of target location and category.
- We rigorously assess our method across three widely recognized remote sensing datasets for object detection, achieving significant improvements in performance that clearly underscore the efficacy of our proposed technique.

2. Related Works

2.1. Closed-Set Object Detection

Despite significant advancements in computer vision, object detection continues to pose substantial technical challenges, driven by the transformative impact of deep learning methodologies.

The advent of convolutional neural networks (CNNs) has catalyzed paradigm shifts in object detection research, revolutionizing technological capabilities within this domain. Ren et al. [16] pioneered the Faster R-CNN architecture, which synergistically combines a region proposal network (RPN) with shared parameter optimization in the detection framework—effectively minimizing computational complexity in candidate region generation while expediting the overall detection process. The YOLO framework, proposed by Redmon [17,18] and further extended by Redmon and Farhadi [19], operates through a grid-based spatial decomposition mechanism that concurrently predicts bounding box coordinates, confidence scores, and class probability distributions for each discrete grid cell. Sun et al. [20] developed Sparse R-CNN, a fully sparse approach that eliminates the need to enumerate object candidate locations across dense image grids, thereby avoiding interaction with global dense image features. Transformer-based object detection models have also seen rapid development. Carion et al. [8] pioneered DETR, an end-to-end detection paradigm that eliminates reliance on handcrafted components. However, the original Transformer-based architecture exhibited slow convergence and limited feature resolution due to its global self-attention mechanism. To mitigate these limitations, Zhu et al. [21] proposed Deformable DETR, which strategically restricts attention to adaptively sampled locations around reference points—thereby enhancing both computational efficiency and localization accuracy. Liu et al. [22] introduced a novel query formulation in DETR using dynamic anchor boxes, offering deeper insights into the functional role of queries in detection transformers. Building upon the DETR framework, Zhang's research group introduced DINO [23], a state-of-the-art end-to-end detection system incorporating three pivotal innovations: (1) denoising training objectives to stabilize optimization, (2) hybrid query initialization combining anchor-based and content-aware strategies, and (3) a dual-query iterative refinement mechanism that progressively refines both localization and classification outputs. This architectural evolution marks a milestone in transformer-based detection methodologies. In remote sensing applications, extensive research efforts (Tao et al. [24]; Cong et al. [25]; Reed et al. [26]; Bastani et al. [27]; Sun et al. [28]; Guo et al. [29]) have focused on leveraging self-supervised and unsupervised paradigms to extract latent representations from unlabeled datasets—advancing feature learning capabilities in data-scarce scenarios.

Some approaches—e.g., CALNet (He et al. [30])—improve detection performance by fusing visible (RGB) and infrared (IR) imagery. Zhang et al. [31] proposed a Unified Object Contour Detector (UniconDet), which unifies the multi-granularity object detection requirements of horizontal bounding boxes (HBB), oriented bounding boxes (OBB), and instance segmentation (InSeg) into a single task of arbitrary-shaped object contour

detection—thereby addressing multi-granularity detection challenges in complex remote sensing scenarios. Although these methods exhibit broad applicability, their capacity to substantially enhance detection robustness remains limited.

2.2. Open-Vocabulary Object Detection

Lately, the swift progress in vision-language pre-training has spurred the development of open vocabulary object detection. These new methods aim at detecting and recognizing categories beyond the labeled label space.

Radford et al. [32] proposed CLIP, an efficient and scalable natural language–supervised approach for vision–language pretraining, which aligns textual and visual representations through contrastive learning. Li et al. [11] introduced BLIP-2 in their 2023 work—a flexible and streamlined pretraining framework that leverages frozen pretrained image encoders and large language models (LLMs) for vision–language alignment. Its core innovation is a lightweight query-based Transformer that effectively bridges the modality gap with minimal computational overhead. GLIP, proposed by Li et al. [12], unifies object detection and phrase grounding via joint vision–language training; by conditioning localization on textual prompts, it expands the visual semantic space, enhances cross-category generalization, and enables zero-shot detection of novel categories without manual annotation. Du et al. [33] introduced Detection of Prompts (DetPro), an innovative OVOD method that leverages pretrained multimodal architectures to generate context-adaptive prompt embeddings—thereby extending the capabilities of open-vocabulary detection. This approach establishes a paradigm shift in cross-modal representation learning, offering a more sophisticated framework for aligning visual observations with linguistic semantics to achieve superior detection accuracy. Gao's research group [34] developed a novel pipeline that automates pseudo-bounding-box annotation using the spatial reasoning capabilities of vision–language models. Their system exploits large-scale image–caption datasets to iteratively refine localization hypotheses, transforming unstructured textual descriptions into geometrically precise detection targets without manual intervention. These generated annotations can then be used to train object detection models in a fully self-supervised manner. Meanwhile, Kim et al. [35] proposed RO-ViT, a region-aware contrastive vision–language pretraining framework explicitly designed to bridge image-level pretraining with open-vocabulary detection tasks.

Earlier research in this field had mainly focused on improving zero-shot recognition performance through visual-language pretraining approaches. Grounding-DINO was proposed by Liu et al. [13], which gained stronger target localization capabilities by introducing more powerful detector structures and fine-grained multimodal feature fusion. Bi et al. [36] proposed RingMoE, a unified multimodal remote sensing foundation model employing a modular hierarchical MoE (Mixture-of-Experts) design. This architecture integrates three expert types: modality-specific, collaborative, and shared experts, enabling simultaneous intra-modal knowledge modeling and cross-modal dependency capture while resolving representation conflicts between heterogeneous modalities. Wang et al. [15] developed LLaMA-Unidetector, which implements a disentangled two-stage paradigm decoupling detection and recognition tasks. The first stage utilizes an agnostic detection head to distinguish foreground objects from background clutter, while the second stage employs TerraOV-LLM—a multimodal large language model (MLLM)—to perform category inference by leveraging pre-trained linguistic knowledge, thereby exploiting the robust generalization capacity of large language models for open-vocabulary classification. Recent advancements in vision-language pretraining have catalyzed the emergence of OVOD, enabling systems to recognize unseen categories through cross-modal semantic alignment rather than relying solely on fixed label sets. This paradigm shift leverages

large-scale vision-language datasets to establish bidirectional knowledge transfer between visual and textual modalities, significantly enhancing detection flexibility in data-scarce remote sensing scenarios. These new methods aim at detecting and recognizing categories beyond the labeled label space.

2.3. Feed-Forward Neural Network

The feed-forward neural network (FFN), as a core component within the Transformer architecture, is essentially a position-independent two-layer perceptron. Typically positioned after the attention mechanism module, the FFN employs an "expand-compress" dimensionality path. It enhances representational capacity through nonlinear transformations in high-dimensional space, capturing richer information in high-dimensionality. Subsequently, it applies the ReLU activation function to extract more representations before finally mapping back to the original dimension, as illustrated in Figure 1a.

Studies [37,38] demonstrate that integrating 3 × 3 depthwise convolution into convolutional feedforward networks (ConvFFN) [39] effectively implements conditional positional encoding (CPE): the channel-wise 3 × 3 local convolution explicitly models spatial dependencies (e.g., contextual relationships between adjacent pixels), while zero-padding preserves boundary information, enabling content-adaptive encoding. This eliminates the need for additional positional embedding layers, achieving computational efficiency superior to conventional self-attention mechanisms, with the corresponding design illustrated in Figure 1b.

Previous work, exemplified by the Squeeze-and-Excitation (SE) mechanism [40], pioneered the introduction of channel attention into computer vision. This mechanism employs branches with activation functions to gate the outputs of feedforward neural networks. Within the gated channel attention mechanism, the gating branch holds greater decision-making authority than the ratio branch, ultimately determining whether corresponding output elements are set to zero. From this perspective, the SE mechanism ingeniously employs averaged features globally as input to the gating branch, achieving the largest receptive field for improved decision-making while addressing the limited receptive field issue in CNN architectures. However, in the ViT era, global receptive fields are no longer scarce. Various global token mixers, exemplified by self-attention, achieve higher-quality global information aggregation than global average pooling. This highlights certain limitations of the global pooling approach adopted by the SE mechanism. For instance, it causes all tokens on a feature map to share identical gating signals, resulting in inflexible and overly coarse channel attention. Nevertheless, recent studies have demonstrated that incorporating the SE mechanism into feedforward neural networks effectively enhances model robustness, as illustrated in Figure 1c.

The Gated Linear Unit (GLU) [41] is a neural network activation mechanism—empirically validated to outperform multilayer perceptrons (MLPs) across diverse natural language processing benchmarks. It comprises two parallel linear transformation branches whose outputs are combined via element-wise multiplication: one branch undergoes linear projection only, while the other is modulated by a sigmoid-activated gating function, as depicted in Figure 1d. In contrast to the Squeeze-and-Excitation (SE) mechanism, GLU generates token-level gating signals solely from the input token itself; its contextual awareness is thus inherently limited to the receptive field of the value branch and does not involve global aggregation.

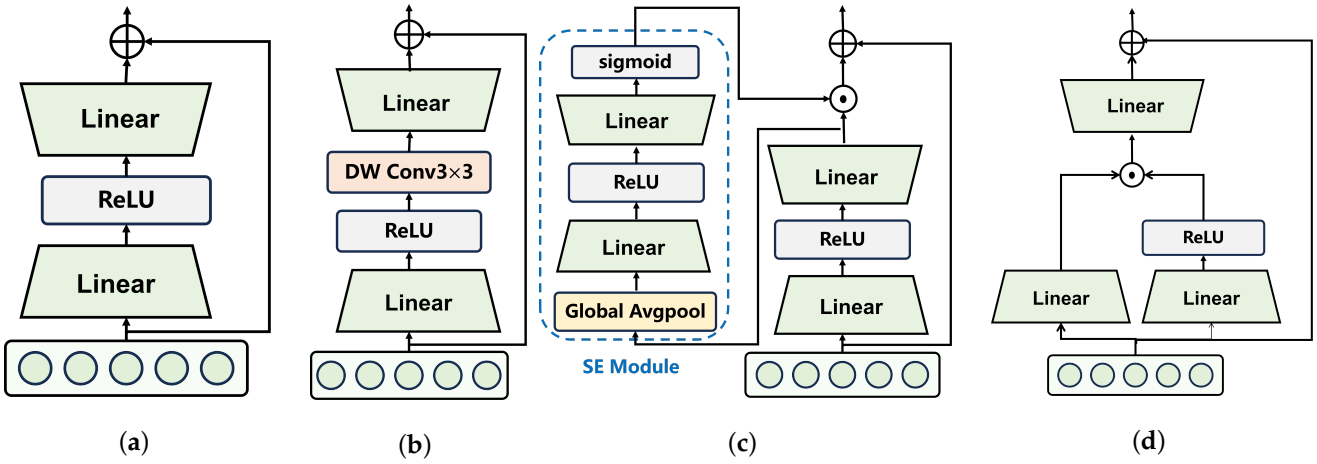


Figure 1. Research on Feedforward Neural Networks. (**a**) Original feed-forward neural network. (**b**) Convolutional feed-forward neural network. (**c**) FFN with SE module. (**d**) Gated Linear Unit.

3. Materials and Methods

3.1. Overview

The paper proposes RS-DINO, an open-vocabulary object detection (OVOD) framework specifically designed for aerial remote sensing imagery. This methodology addresses the persistent challenges of false negatives and false positives in open-vocabulary scenarios while improving detection efficiency and localization precision. The framework comprises four core components: Feature Extraction, Cross-Modal Feature Fusion, Language-Guided Query Selection, and Multimodal Decoder. It achieves robust cross-modal feature fusion and efficient detection through: (1) a Feature Extraction module incorporating a Multi-Scale Large-Kernel Attention (MSLKA) mechanism—enhancing joint modeling of global context and fine-grained local details; (2) a Cross-Modal Feature Fusion module employing bidirectional cross-attention to establish deep semantic alignment between visual and textual embeddings; (3) a Language-Guided Query Selection mechanism that improves detection accuracy via dynamic anchor initialization and hybrid query strategies; and (4) a Multimodal Decoder integrating a Convolutional Gated Feedforward Network (CGFFN) to optimize multimodal feature integration, strengthen spatial positional encoding, and refine channel-wise attention modulation. The technical flowchart in Figure 2 fully illustrates our proposed remote sensing OVOD framework, which unifies vision–language reasoning with precise target localization.

The model outputs multiple sets of object bounding boxes and noun phrases for a given image-text pair. For instance, the model locates an airplane and a vehicle within the input image, while extracting the words ‘airplane’ and ‘vehicle’ from the input text as corresponding labels. We employ all category names as text prompts for the object detection task, with one bounding box corresponding to each text input. The highest-scoring output object is selected as the task’s output.

3.2. Feature Extraction

3.2.1. Linguistic Branch

The linguistic branch will select the pretrained case-insensitive BERT-base [42] as the text encoder to extract text features, which consists of 12 Transformer layers [43]. For the input linguistic expressions, the text encoder generates text features $F_l \in R^{l \times d}$, where l denotes the number of linguistic feature tokens and d represents the feature dimensionality. Each token embedding in F_l captures contextualized semantic information for corresponding linguistic units. Owing to their semantically dense nature and capacity to

encode high-dimensional object attributes, textual features demonstrate superior efficacy in establishing high-resolution cross-modal alignment with multiscale visual representations. This property enables precise harmonization of semantic representations across modalities, particularly for targets exhibiting complex intra-class variations.

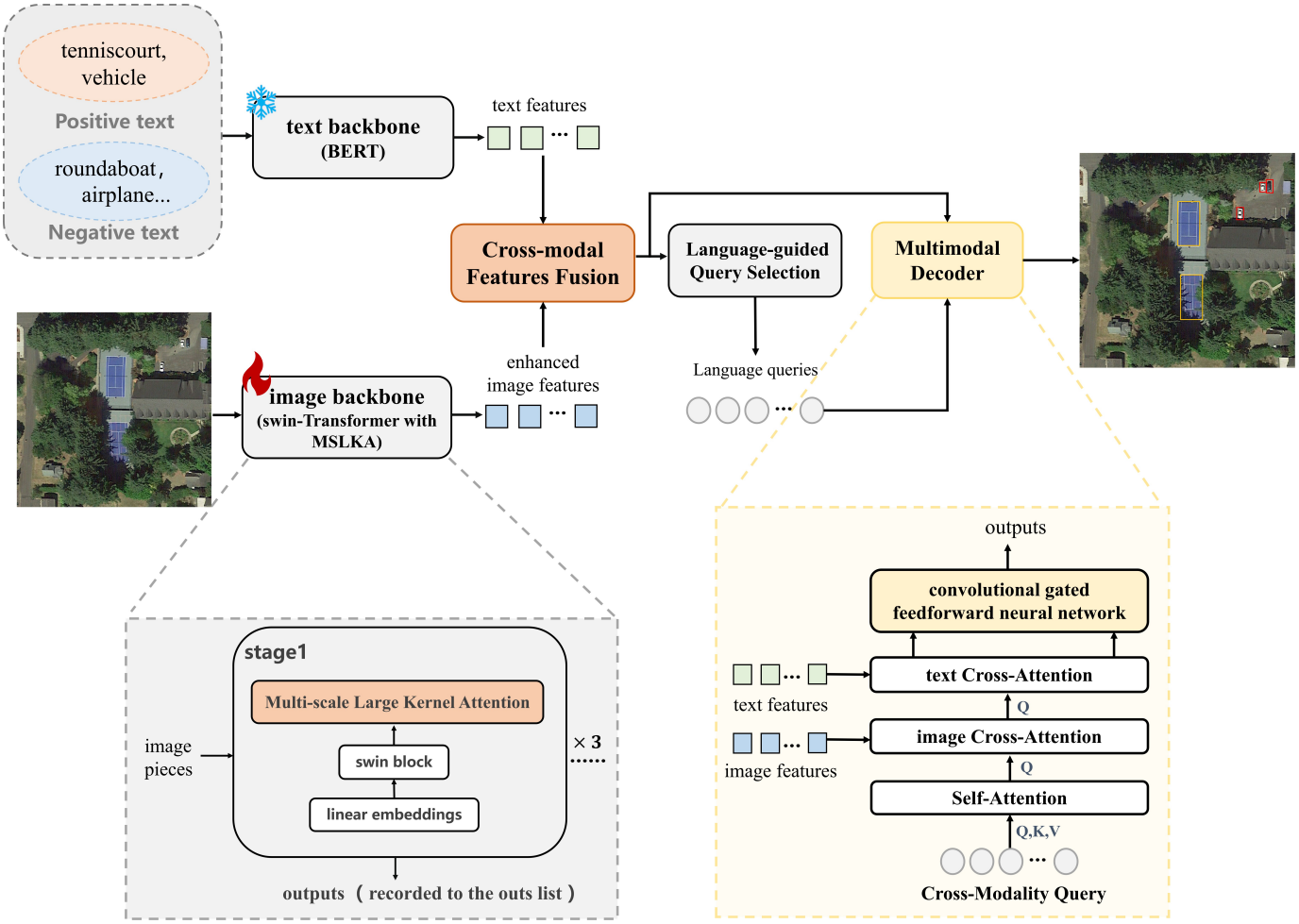


Figure 2. The overall architecture of RS-DINO.

3.2.2. Visual Branch

The image encoder integrates the Swin-Transformer backbone with our proposed Multi-scale Large Kernel Attention (MSLKA) modules, which are specifically designed to address the challenges of multi-scale object representation and dense target discrimination in remote sensing imagery. The encoder progressively extracts multi-scale visual features $F_v = \{F_v^i\}_{i=1}^4$ from four stages, and an MSLKA module is strategically inserted at the end of each stage to enhance feature refinement. Unlike conventional attention mechanisms that operate at a single scale or with fixed receptive fields, MSLKA innovatively combines three core functions: large kernel attention (LKA) to capture long-range contextual dependencies, multi-scale convolutions to aggregate heterogeneous-scale correlations, and gated aggregation to dynamically recalibrate feature responses. This tripartite design enables the encoder to simultaneously model global context and preserve fine-grained spatial details, which is critical for detecting densely packed small objects. By progressively incorporating MSLKA modules, the encoder achieves hierarchical global–local feature interactions that systematically strengthen its capacity to handle scale variations and complex backgrounds, thereby setting a new foundation for robust remote sensing interpretation.

Given input feature maps $x \in R^{C\times H\times W}$, the Large Kernel Attention (LKA) module establishes adaptive long-range spatial dependencies through a decomposed operation sequence. Specifically, it factorizes a $K \times K$ convolution into three components: a point-wise convolution $f_{PW}(\bullet)$, $a\lceil\frac{K}{d} \times \frac{K}{d}\rceil$ depth-wise d-dilation convolution $f_{DWD}(\bullet)$, and $a(2d-1) \times (2d-1)$ depth-wise convolution $f_{DW}(\bullet)$. The overall transformation is mathematically formulated as:

$$LKA(X) = f_{PW}(f_{DWD}(f_{DW}(X))) \tag{1}$$

To acquire the attention maps containing omni-scale information, we improve the fixed Large Kernel Attention (LKA) by implementing the group-wise multi-scale mechanism. Given the input feature maps $X \in R^{C\times H\times W}$, the module initially divides it into n pieces $\{X_1, X_2, \ldots, X_n\}$ of $\left\lfloor\frac{C}{n}\right\rfloor \times H \times W$. For i-th group of features X_i, an LKA decomposed by $\{K_i, d_i\}$ is employed to generate a uniform-scale attention map LKA_i. In detail, we utilize three groups of LKA: 7, 2 implemented via the 3-5-1 configuration, 21, 3 via the 5-7-1 configuration, and 35, 4 via the 7-9-1 configuration. Here, the a-b-1 notation represents the cascading of a × a depth-wise, b × b depth-wise-dilated, and point-wise convolutions.

Although the larger LKA captures broader pixel responses, blocking artifacts emerge in the generated attention maps of the larger LKA. For the i-th group input X_i, in order to avoid the block effect and learn more local information, we utilize a spatial gate to dynamically adapt $LKA_i(\bullet)$ into $MSLKA_i(\bullet)$ through the following method:

$$MSLKA_i(X_i) = G_i(X_i) \bigotimes LKA_i(X_i) \tag{2}$$

where $G_i(\bullet)$ represents the i-th gate generated through an $a_i \times a_i$ depth-wise convolution, and $LKA_i(\bullet)$ denotes the LKA decomposed from $a_i - b_i - 1$. Thus, $MSLKA_i$ with larger receptive fields exhibit stronger responses to long-range dependencies, while smaller $MSLKA_i$ tend to preserve local textural features.

3.3. Cross-Modal Features Fusion

To achieve effective cross-modal fusion between multi-scale visual features and textual embeddings, we introduce a Collaborative Feature Enhancement (CFE) layer framework inspired by GroundingDINO [13]. Specifically, a dynamic differential attention mechanism enhances textual representations by adaptively modeling fine-grained semantic associations and improving contextual discriminability. Concurrently, a deformable self-attention mechanism refines multi-scale visual features extracted from the visual backbone via learnable spatial sampling offsets—precisely capturing critical geometric deformations and structural details in remote sensing imagery. Building upon this, a bidirectional cross-modal attention mechanism establishes symmetric interaction pathways (image-to-text and text-to-image), enabling precise semantic alignment and cross-modal coordination across heterogeneous data sources. This unified framework mitigates modality-specific semantic biases in heterogeneous remote sensing environments while enhancing cross-modal reasoning capabilities for multi-scale, multi-class objects. The architecture systematically constructs a discriminative multimodal feature space, laying a robust foundation for downstream detection and recognition pipelines. Its implementation is detailed in Figure 3. Within the Cross-Modal Feature Fusion module, we employ six Collaborative Feature Enhancement layers to realize this objective.

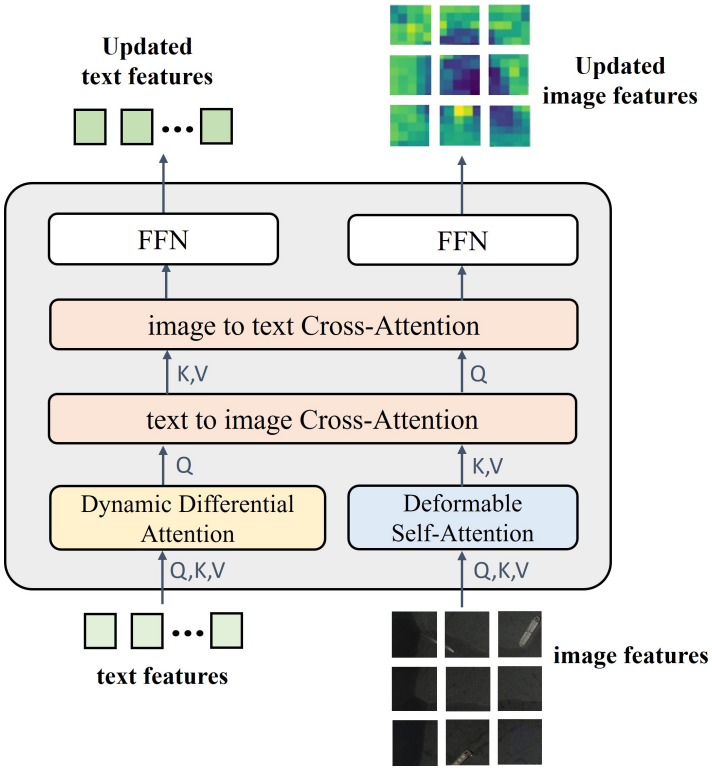


Figure 3. Architecture of collaborative feature enhancement layer.

3.4. Language-Guided Query Selection

We denote image features as $F_v \in R^{N_v d}$ and text features as $F_l \in R^{N_l d}$. Here, N_v represents the count of image tokens, N_l stands for the number of text tokens, and d indicates the feature dimension. Our objective is to derive N_q queries from the image features produced by the encoder to feed into the decoder. Following the approach in DINO [23], we fix N_q at 900. The highest-ranked N_q queries, indexed by the image features and labeled I_{N_q}, are chosen using the following formulation:

$$I_{N_q} = Top_{N_q}(Max^{(-1)})\, F_v F_l^{\top} \tag{3}$$

In this expression, Top_{N_q} denotes the operation of selecting the top N_q indexes. The function $Max^{(-1)}$ performs the maximum operation along the -1 dimension, and the symbol $\top$ denotes the matrix transpose. The Language-guided Query Selection component generates N_q candidate indices, from which corresponding feature vectors are retrieved to construct the initial query set. Adhering to the methodological principles established in DINO, a composite query initialization strategy is implemented to produce the decoder's foundational query embeddings, ensuring semantic coherence between linguistic cues and visual representations. The position part is defined as a dynamic anchor frame initialized from encoder outputs, which encodes spatial priors through learnable bounding box coordinates. The other part is the content query that can be learned during training. This language-guided query selection mechanism dynamically identifies salient regions by measuring cross-modal semantic correlations, ensuring the initialization features retain both visual prominence and linguistic relevance.

3.5. Multimodal Decoder

Every multimodal query passes through several key components within each decoder layer: self-attention, visual cross-attention, textual cross-attention, and the Convolutional Gated Feedforward Network (CGFFN). This architecture enables the system

to effectively aggregate and synthesize relevant target representations across modalities. The self-attention layer first processes the query internally; the cross-attention layers then align it with corresponding visual and textual embeddings. Finally, the CGFFN refines these integrated representations to produce coherent outputs. Together, these components operate in sequence to extract and consolidate semantically meaningful information from heterogeneous data sources. The multimodal decoder outputs object embeddings for localization prediction and generates bounding box coordinates and category logits via a detection head.

To enhance feature discrimination in dense remote sensing scenarios, we introduce the Convolutional Gated Feedforward Network (CGFFN)—a lightweight yet powerful redesign of the conventional Gated Linear Unit (GLU). The key innovation is the insertion of a minimal 3 × 3 depthwise convolution before the activation function within the GLU, transforming the module into a spatially aware, channel-wise gating mechanism. Unlike the Squeeze-and-Excitation (SE) block—which relies on coarse global average pooling to generate channel-wise weights—CGFFN derives each token's gating signal from its high-resolution local neighborhood, enabling fine-grained, content-adaptive recalibration of feature representations. This design effectively integrates local spatial detail with channel-wise selectivity.

As illustrated in Figure 4, the value branch maintains architectural parity with standard MLPs and GLUs, preserving gradient flow integrity throughout backpropagation. Despite having identical parameter counts, stride configurations, and kernel sizes as conventional feedforward networks, CGFFN achieves significantly lower computational complexity by virtue of its depthwise separable design and localized gating. Consequently, CGFFN acts as a more interpretable and robust feature mixer, particularly suited for OVOD in remote sensing imagery where both global semantics and local details are critical.

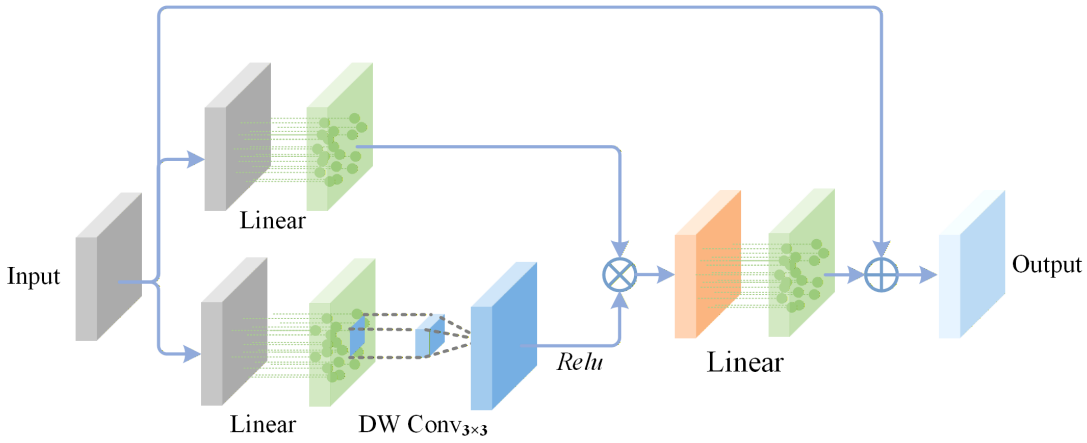


Figure 4. Architecture of convolutional gated feedforward neural network.

4. Results

4.1. Implementation Details

To comprehensively evaluate the efficacy of the proposed RS-DINO framework, we conducted systematic experiments across three open-access remote sensing object detection benchmarks: the DIOR dataset, the DOTAv2.0 dataset, and the NWPU-VHR10 dataset. Due to varying pixel dimensions across different datasets, we uniformly resize input images to 800 × 800 pixels for processing. The image cropping overlap ratio is set to 50%. During testing, we did not employ a segmentation strategy but instead directly scaled the entire image. Predictions were merged using the NMS method with a preset IoU threshold of 0.5.

In this study, the adoption of horizontal bounding boxes (HBB) rather than oriented bounding boxes (OBB) is a reasonable and targeted choice. This is because the core challenge addressed in this paper centers on the small scale and dense distribution of objects, rather than precise orientation. Furthermore, considering the dataset characteristics, the official annotations for both the NWPU VHR-10 and DIOR datasets are HBB, while the DOTA-v2 dataset also provides high-quality HBB annotations. Adopting HBB enables the unification of evaluation protocols across these three datasets, ensuring consistency and comparability in cross-dataset experiments. Furthermore, from both the model design perspective and the standpoint of mainstream benchmark methods in the existing remote sensing object detection field, HBB can directly leverage strategies designed for small objects within mature, general-purpose detection frameworks—such as feature pyramids and multi-scale training. This avoids the angular regression branches and complex computational overhead introduced by OBB, allowing research to focus more intently on addressing scale and density issues while maintaining efficient training and inference. Therefore, adopting HBB uniformly throughout this study represents a reasonable approach that focuses on core challenges while balancing efficiency and comparability.

We employ the Swin-Transformer model as the image encoder, dividing it into four stages with an MSLKA module introduced at the end of each stage. This encodes the input remote sensing image into multi-scale visual features $F_v = \{F_v^i\}_{i=1}^4$, where i denotes the features from the i-th stage. On the linguistic side, we use a widely used BERT-base model as a text encoder which initializes pretrained weights to initially extract text features.

We use the RS-DINO model with the Pytorch framework and train the model in a distributed fashion on 4 NVIDIA 4090 GPUs using a batch size of 2 on each GPU across 50 cycles, with the learning rate decreasing by a factor of 0.1 at cycles 15 and 35, respectively. During training, the model employs the Adam [44] optimizer with an initial learning rate of 1×10^{-5} for the text and image encoders, and 1×10^{-4} for the other parameters. The parameters in the network that are not pretrained are randomly initialized using Xavier [45].

To systematically evaluate model performance, the mean average precision (mAP) is adopted as the core evaluation metric. Specifically, we report mAP at an intersection-over-union (IoU) threshold of 0.50 (denoted as mAP50). For each object category, a predicted bounding box is classified as a true positive if its IoU with the ground-truth bounding box exceeds 0.50, based on which the average precision (AP) is computed per class. The mAP50 value is then derived by averaging the AP scores across all object categories at this IoU threshold. Mathematically, the AP for a class is calculated as follows:

$$AP = \int_0^1 Precision(r)dr \tag{4}$$

where Precision(r) is the precision at a given recall level r :

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

where TP refers to true positives and FP refers to false positives.

The mAP is computed by averaging the AP values across all object categories:

$$mAP = \frac{1}{C} \sum_{i=1}^{C} AP_i \tag{6}$$

where C indicates the complete count of object categories.

In this study, two metrics, mAP@0.5 and mAP@0.5:0.95, are employed as the primary evaluation criteria, while APs, APm, and APl are used to assist in assessing the model's

performance. mAP@0.5 determines detection correctness using an Intersection over Union (IoU) threshold of 0.5. This criterion is relatively lenient and emphasizes overall detection capability. In contrast, mAP@0.5:0.95 evaluates mAP across IoU thresholds ranging from 0.5 to 0.95 in 0.05 increments, providing a stricter and more comprehensive assessment of the model's ability to match predicted objects with ground truth. APs, APm, and APl represent the detection accuracy for small, medium, and large targets, respectively, reflecting the model's performance in detecting objects of varying scales.

Frames Per Second (FPS) denotes the number of images a model can process per second, directly reflecting its real-time inference capability. Model parameters (Params) represent the total number of trainable parameters within the model, indicating its storage requirements and fundamental complexity. FLOPs measure the total number of floating-point operations required for a single forward pass, providing insight into computational complexity and thus enabling assessment of operational efficiency in real-world remote sensing environments.

4.2. Comparisons with SOTA Methods

The model is pretrained across several extensive datasets and its performance is evaluated on an untrained dataset in a zero-shot scenario. For a more detailed comparison, we show several fine-tuning results to benchmark our model against previous work. The experimental results show that our algorithm is effective in zero-shot scenarios and highlight its strong generalization capability for predicting new classes.

4.2.1. DIOR Benchmark

We pretrained our RS-DINO using the NWPU-VHR10 and DOTAv2.0 datasets and fine-tuned our RS-DINO model using the DIOR datasets. The model was rigorously assessed on DIOR benchmark, with the results shown in Table 1 showcasing its superior performance. Under the zero-shot transfer setting, our RS-DINO achieved an mAP@0.5 of 26.5%, significantly outperforming LLAMA-Unidetector and LAE-DINO by 1.9% and 3.7%, respectively. We also fine-tuned the model on the DIOR dataset and trained four models based on CNN and Transformer architectures, including the closed-set classifiers DINO [23], SatLas [27], RingMo [28], and SkySense [29]. Pre-trained object detectors perform well on baseline categories but fail to recognise novel categories. After fine-tuning, our model achieves 3.5% and 2.7% improvements over existing state-of-the-art methods on RS-DINO in mAP@0.5 and mAP@0.95, demonstrating superior performance across both strict and lenient IoU thresholds. Furthermore, our model achieves optimal performance over existing approaches in terms of APs (28.6%), APm (49.7%), and APl (75.6%), demonstrating that RS-DINO effectively leverages multi-scale features to detect desired objects with greater accuracy.

4.2.2. DOTAv2.0 Benchmark

We pretrained our model using the NWPU-VHR10 and DIOR datasets and fine-tuned our RS-DINO using the DOTAv2.0 datasets. The model was rigorously evaluated on the prestigious DOTAv2.0 benchmark, and the results, as detailed in Table 2. This model achieves a mAP@0.5 of 27.60% in zero-shot detection and reaches 70.10% in fine-tuned detection, surpassing the state-of-the-art method LLAMA-Unidetector by 1.6% and 3.7%, respectively. Additionally, the proposed method demonstrates superior performance in closed-set detection tasks, achieving the highest performance across all key evaluation metrics, with its superiority being both statistically significant and practically meaningful.

Table 1. Results on DIOR for open-vocabulary object detection.

Methods	Zero-Shot	Fine-Tuning				
	mAP@0.5	mAP@0.5	mAP@0.5:0.95	APs	APm	APl
DINO [23]	-	73.81%	46.33%	18.02%	36.74%	56.38%
SatLas [27]	-	74.10%	48.31%	20.44%	38.16%	57.92%
RingMo [28]	-	75.90%	49.56%	21.72%	39.05%	59.64%
SkySense [29]	-	78.73%	54.33%	24.85%	45.66%	70.50%
GLIP [12]	19.20%	76.50%	49.30%	20.60%	39.60%	64.90%
Grounding DINO [13]	21.90%	78.30%	53.20%	24.60%	43.10%	68.40%
LAE-DINO [14]	22.80%	79.90%	55.90%	25.30%	45.60%	71.00%
LLAMA-Unidetector [15]	24.60%	82.80%	58.90%	26.10%	47.50%	73.80%
RS-DINO	26.50%	86.30%	61.60%	28.60%	49.70%	75.60%

Table 2. Results on DOTAv2.0 for open-vocabulary object detection.

Methods	Zero-Shot	Fine-Tuning				
	mAP@0.5	mAP@0.5	mAP@0.5:0.95	APs	APm	APl
DINO [23]	-	63.77%	50.27%	18.36%	36.54 %	59.12%
SatLas [27]	-	65.32%	53.55%	20.84%	38.63%	61.34%
RingMo [28]	-	69.94%	56.36%	22.30%	40.77 %	64.38%
SkySense [29]	-	70.03%	57.84%	22.97%	41.92%	66.65%
GLIP [12]	20.20%	58.34%	49.10%	16.50%	33.60 %	55.20%
Grounding DINO [13]	23.50%	62.50%	52.30%	19.40%	37.60%	59.60 %
LAE-DINO [14]	25.80%	64.90%	54.10%	20.40%	39.80%	61.00%
LLAMA-Unidetector [15]	26.00%	66.40%	55.20%	21.60%	41.50%	63.90 %
RS-DINO	27.60%	70.10%	58.00%	23.30%	43.30%	66.90 %

4.2.3. NWPU-VHR10 Benchmark

The NWPU-VHR10 dataset contains fewer images and targets compared with the other two benchmarks, making it relatively easier. Since the categories in the DIOR and DOTAv2.0 datasets encompass all the categories present in NWPU-VHR10, the challenge lies in the cross-domain setting, akin to zero-shot domain transfer. We evaluated the performance of the proposed RS-DINO method against LAE-DINO and LLAMA-Unidetector on the NWPU-VHR10 dataset, as shown in Table 3. We pretrained our model using the DIOR and DOTAv2.0 datasets, and fine-tuned our RS-DINO using the NWPU-VHR10 datasets. The model was directly tested on the NWPU-VHR10 benchmark. Under the zero-shot transfer setting, our model achieved an mAP@0.5 of 45.50%, outperforming LAE-DINO and LLAMA-Unidetector by 3.2% and 5.7%, respectively. Following fine-tuning, the RS-DINO model demonstrated a significant performance improvement, particularly achieving a 2.62% increase in APs compared to the current state-of-the-art, highlighting our model's superiority in small object detection.

Table 3. Results on NWPU-VHR10 for open-vocabulary object detection.

Methods	Zero-Shot	Fine-Tuning				
	mAP@0.5	mAP@0.5	mAP@0.5:0.95	APs	APm	APl
DINO [23]	-	80.74%	61.60%	29.19%	50.62%	62.31%
SatLas [27]	-	83.63%	63.22%	29.37%	52.81%	66.74%
RingMo [28]	-	84.40%	64.65%	30.23%	53.64%	66.89%
SkySense [29]	-	86.92%	67.92%	30.68%	57.66%	70.25%

Table 3. *Cont.*

Methods	Zero-Shot	Fine-Tuning				
	mAP@0.5	**mAP@0.5**	**mAP@0.5:0.95**	**APs**	**APm**	**APl**
GLIP [12]	29.20%	71.20%	52.70%	23.60%	45.50%	56.30%
Grounding DINO [13]	36.50%	80.80%	60.90%	28.30%	51.80%	62.10%
LAE-DINO [14]	39.80%	81.40%	62.80%	29.70%	53.20%	64.80%
LLAMA-Unidetector [15]	42.30%	83.50%	64.10%	30.60%	54.90%	68.90%
RS-DINO	45.50%	87.50%	68.30%	33.30%	58.80%	72.10%

4.3. Ablation Study

4.3.1. Regarding MSLKA Ablation Experiments

To rigorously evaluate the effectiveness of the proposed MSLKA module, we independently integrated CBAM [46], EMA [47], and MSLKA—each under identical architectural and training configurations—into the Swin-Transformer visual feature extractor. As shown in Figure 5, we visualize and compare the corresponding feature maps of the baseline model, the CBAM-enhanced model, the EMA-enhanced model, and the MSLKA-enhanced model across three benchmark remote sensing datasets. The results confirm that all three attention mechanisms improve object detection performance over the baseline; critically, MSLKA consistently achieves the highest gains across all datasets—particularly in detecting and localizing densely packed small objects. This superiority stems from MSLKA's novel design: multi-scale large-kernel convolutions explicitly capture long-range contextual dependencies at varying receptive fields, enabling hierarchical global–local feature interaction. Consequently, MSLKA exhibits superior robustness to scale variation and cluttered backgrounds—key challenges in remote sensing imagery—while preserving discriminative fine-grained details of crowded targets. These findings substantiate that MSLKA not only enhances feature representativeness but also strengthens the model's capacity for reliable object detection in complex, real-world remote sensing scenarios.

To systematically investigate the MSLKA module's ability to extract multi-scale features, we conducted quantitative ablation experiments on the number of visual feature scales. As shown in Table 4, when using four-level multi-scale features (i.e., retaining all four levels of the feature pyramid), the model achieved the best performance on the three benchmark datasets: mAP@0.5 reached 83.40% on the DIOR dataset, 67.10% on the DOTAv2.0 dataset, and 85.30% on the NWPU-VHR10 dataset. Notably, when only a single-level MSLKA enhancement was used, the model's performance significantly declined: the mAP@0.5 on the three datasets dropped to 78.80%, 63.80%, and 81.20% respectively, representing a decrease of 4.60%, 3.30%, and 4.10% compared to the four-level configuration. This result fully validates the critical role of multi-scale feature fusion in remote sensing small target detection, indicating that MSLKA effectively addresses the feature sparsity issue of dense small targets in remote sensing images by collaboratively capturing context information at different scales (such as large-scale scene backgrounds and small-scale target details), thereby enhancing the model's localization accuracy.

Table 4. Model analysis on the number of visual feature scales in MSLKA module.

Feature Scales	DIOR (mAP@0.5)	DOTAv2.0 (mAP@0.5)	NWPU-VHR10 (mAP@0.5)
{64}	78.80%	63.80%	81.20%
{64,32}	80.60%	64.90%	82.50%
{64,32,16}	81.30%	65.20%	83.70%
{64,32,16,8}	83.40%	67.10%	85.30%

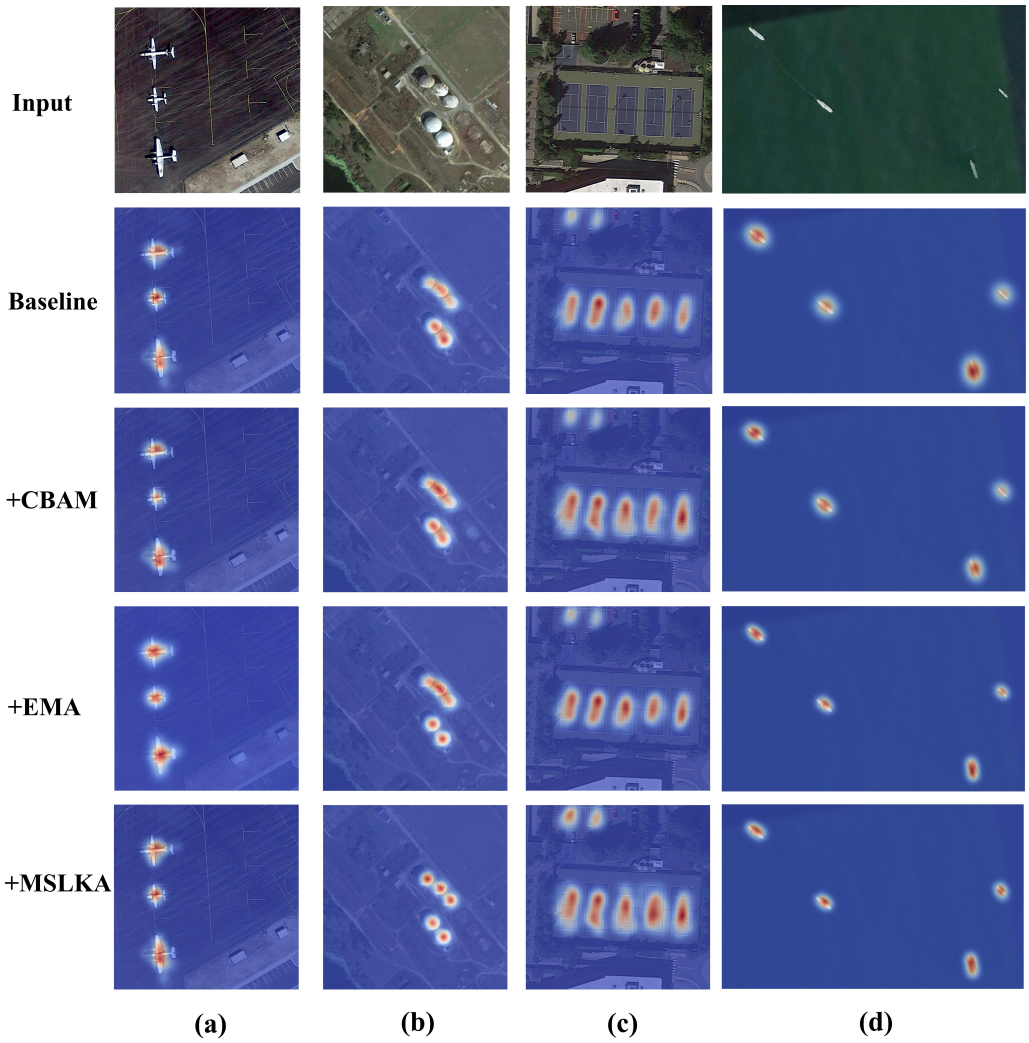


Figure 5. Visualization of the performance of baseline, CBAM-optimised model, EMA-optimised model, and MSLKA-optimised model. (**a**,**b**) from DIOR, (**c**) from DOTAv2.0 , (**d**) from NWPU-VHR10.

4.3.2. Regarding CGFFN Ablation Experiments

To evaluate the effectiveness of our proposed CGFFN, we integrated ConvFFN, GLU, and CGFFN into the cross-modal decoder of the baseline model using identical default settings. As shown in Table 5, we present the model parameter sizes and detection results on three datasets after integrating these three feedforward neural networks. The results show that integrating all three feedforward neural networks into the baseline model can improve the performance of object detection. However, compared with ConvFFN and GLU, CGFFN adopts a unique gating signal based on its most recent fine-grained features, further addressing the overly coarse-grained defect of global average pooling in existing feedforward neural networks. It effectively combines local spatial details with channel selectivity, achieving the best performance with mAP@0.5 reaching 84.9%, 69.4%, and 86.6% on the three datasets respectively. Regarding computational complexity—a critical issue for deployment on resource-constrained edge devices—the model incorporating CGFFN achieved optimal performance while maintaining relatively moderate parameter growth compared to the baseline model. Its detection accuracy surpassed the baseline model by 6.6%, 6.9%, and 5.8%, respectively. Its FLOPs (36.5 GFLOPs) and parameters (199 M) increased by 30.1% and 15.6%, respectively, compared to the baseline model. These increases are acceptable as they deliver meaningful accuracy gains and meet the requirements for model deployment in real-world remote sensing scenarios.

Table 5. Ablation experiments for CGFFN.

Methods	Params (M)	FLOPs (G)	DIOR (mAP@0.5)	DOTAv2.0 (mAP@0.5)	NWPU-VHR10 (mAP@0.5)
baseline	172	34.2	78.30%	62.50%	80.80%
+ConvFFN	188	39.9	79.60%	66.50%	82.90%
+GLU	193	40.3	82.80%	65.20%	83.50%
+CGFFN	199	44.5	84.90%	69.40%	86.60%

4.3.3. Regarding Query Ablation Experiments

To validate model performance through more language-guided queries, we conducted 1200 and 1500 additional experiments on the DIOR, DOTAv2.0, and NWPU-VHR10 datasets, respectively, with details presented in Table 6. Results indicate that across all three datasets, models with 1200 and 1500 queries slightly outperform the 900-query variant, suggesting improved coverage of target categories. However, this improvement is marginal, as the 900-query model already suffices to cover all objects in the datasets. Furthermore, introducing additional queries exacerbates data imbalance during training, as the model is trained on objects drawn from the sampled categories. This imbalance may negate the benefits gained from extra queries.

Table 6. Model analysis on the Number of language-guided queries.

Query Number	DIOR (mAP@0.5)	DOTAv2.0 (mAP@0.5)	NWPU-VHR10 (mAP@0.5)
900	86.30%	70.10%	87.50%
1200	86.60%	70.50%	87.80%
1500	87.00%	70.80%	88.10%

4.4. Scalability

Table 7 presents the object detection results achieved by our model on the DIOR, DOTA2.0, and NWPU-VHR10 test sets. These results were obtained by fine-tuning the model using different-sized datasets (full, ½, and ¼) randomly sampled from the training sets of these three datasets. We observe that using only half of the training data achieves mAP50 scores of 84.3%, 64.7%, and 83.5%, respectively. These detection results demonstrate that after open-vocabulary pretraining, only a small amount of data is required for fine-tuning to achieve satisfactory performance in practical detection tasks, highlighting the scalability of our model in terms of data utilization.

Table 7. Testing results using fine-tuned data of varying sizes randomly sampled from the training set on the test set.

Data Scale	DIOR (mAP@0.5)	DOTAv2.0 (mAP@0.5)	NWPU-VHR10 (mAP@0.5)
full	86.30%	70.10%	87.50%
½	85.20%	66.50%	84.40%
¼	84.30%	64.70%	83.50%

4.5. Qualitative Analysis

In this section, we present qualitative visualizations of test-set predictions and provide an in-depth analysis of model behavior. As illustrated in Figure 6, RS-DINO consistently outperforms traditional OVOD methods in detecting densely packed small targets—such as helicopters and ships in Figure 6—demonstrating superior completeness and localization fidelity. This improvement stems from RS-DINO's ability to selectively attend to query-annotated objects via multi-scale feature integration, while actively suppressing irrelevant

background clutter and confounding foreground objects. Critically, the enhanced MSLKA and CGFFN modules enable hierarchical cross-scale reasoning: MSLKA refines spatially localized details across pyramid levels, whereas CGFFN dynamically modulates channel-wise feature responses based on fine-grained gating signals. Together, they synergistically strengthen small-object representation under severe occlusion and scale variation, leading to more robust and accurate detection in complex, high-density remote sensing scenes.

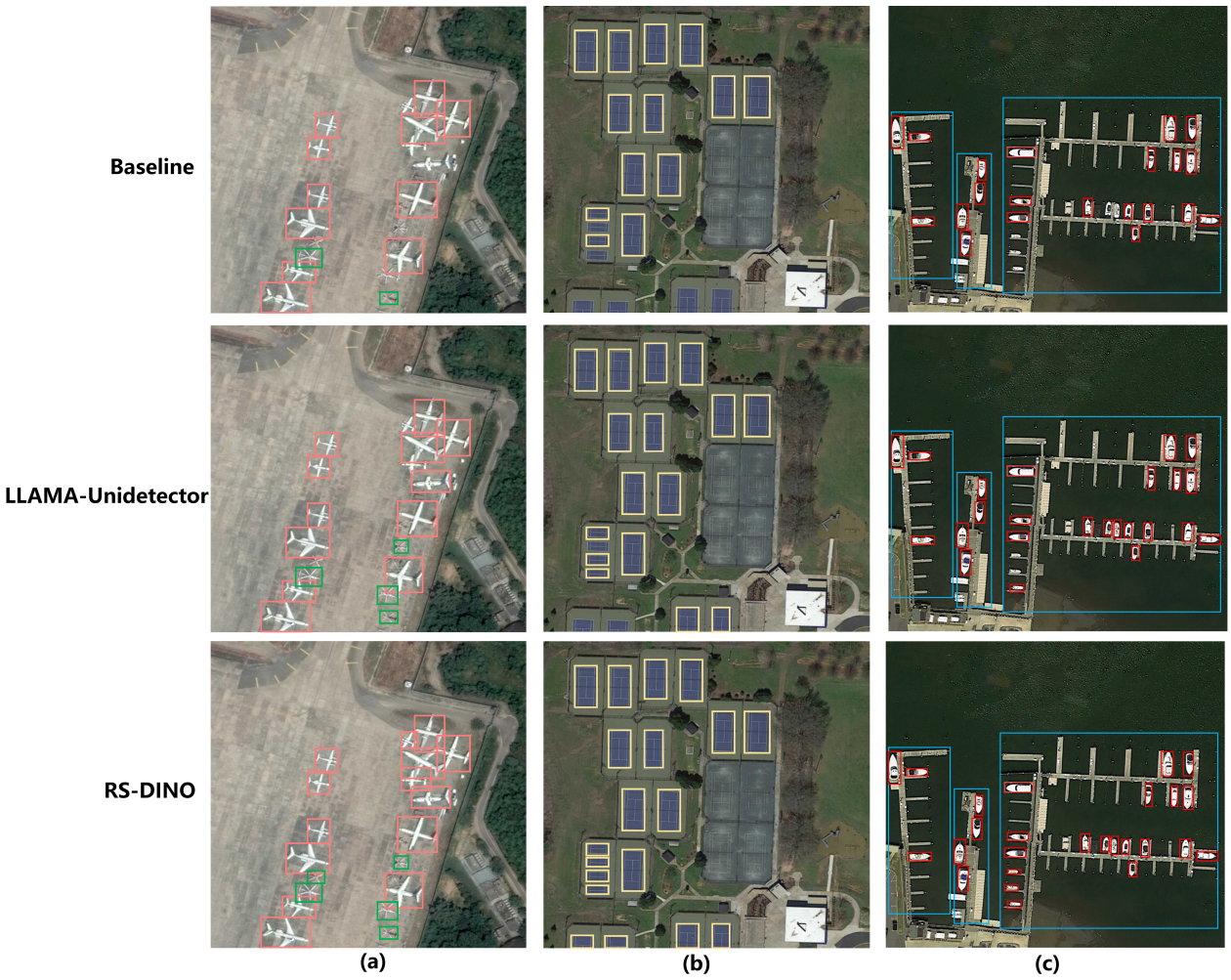


Figure 6. Visualization of the performance of baseline, LLAMA-Unidetector and RS-DINO models. (**a**) airplane, helicopter. (**b**) tennis court. (**c**) ship, harbor.

5. Discussion

To systematically assess the efficacy of our OVOD framework (RS-DINO), we conducted a comprehensive evaluation across heterogeneous fine-grained categories from three remote sensing benchmarks—DIOR, DOTAv2.0, and NWPU-VHR10—with detailed visualizations provided in Figure 7. The model exhibits strong performance on the seen categories (e.g., groundtrackfield and basketballcourt) within the DIOR dataset, achieving mean average precision (mAP) values above 0.75, with a peak of 0.852 on the tenniscourt category. On unseen categories, the model also demonstrates solid performance for chimney (mAP = 0.714), express-way-service-area (mAP = 0.709), and stadium (mAP = 0.702). For the DOTAv2.0 dataset, the model achieves an mAP of 0.746 on the seen tenniscourt category but performs poorly on unseen categories such as container crane (mAP = 0.083) and helipad (mAP = 0.056). These findings reveal a critical insight: RS-DINO excels in categories with clear visual semantics and moderate structural complexity but faces limitations when handling objects exhibiting high intra-class variability, weak visual discriminability,

or semantic abstraction beyond the scope of training prompts. Although RS-DINO outperforms most state-of-the-art methods on holistic OVOD benchmarks, this performance gap underscores the persistent challenge of fine-grained open-vocabulary detection in remote sensing. Future research should prioritize enhancing semantic grounding through prompt engineering, multi-scale feature fusion, cross-modal alignment, and the incorporation of domain-specific knowledge. Advancing models capable of precisely discerning subtle inter-class variations remains a pivotal direction for building more robust, adaptable, and intelligent remote sensing vision systems.

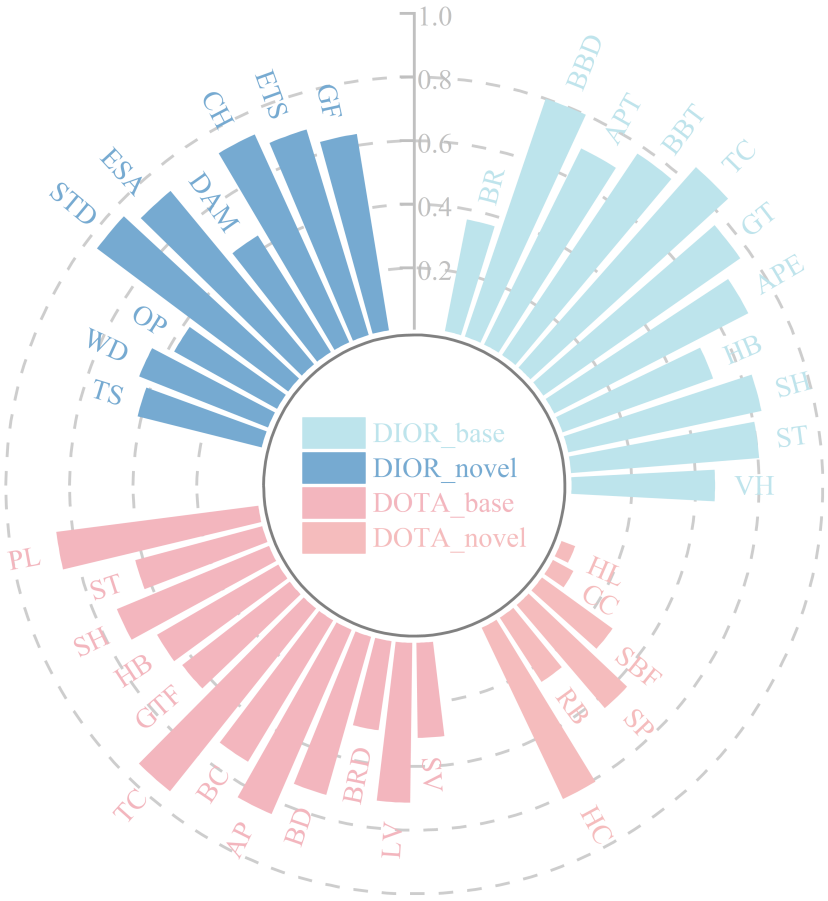


Figure 7. mAP for each category on the DIOR and DOTAv2.0 datasets, demonstrating the model's effectiveness in recognizing both base and novel categories.

6. Conclusions

This study introduces RS-DINO, an open-vocabulary object detection framework specifically designed for remote sensing scene interpretation. It systematically tackles two persistent and interrelated challenges in aerial imagery: high-frequency missed detections and false positives—both primarily induced by densely packed small targets under complex background conditions. The framework adopts a dual-branch architecture: the text branch leverages the pre-trained BERT-base model to encode category semantics, while the visual branch incorporates the Multi-scale Large-Kernel Attention (MSLKA) module into the backbone encoder. MSLKA enables simultaneous modeling of long-range global context and fine-grained local structure across pyramid levels—thereby strengthening hierarchical multi-scale representation without sacrificing spatial fidelity. To bridge the modality gap, we introduce a cross-modal attention mechanism that enforces bidirectional semantic align-

ment between vision and language features, complemented by Language-Guided Query Selection to prioritize semantically grounded queries. Finally, the multimodal decoder aggregates cross-modal representations and refines them via a Convolutional Gated Feed-forward Network (CGFFN), which dynamically modulates channel-wise activations using fine-grained spatial gating signals—ensuring precise localization and robust classification. Extensive experiments demonstrate that RS-DINO achieves state-of-the-art performance on standard remote sensing OVOD benchmarks, with quantitative results confirming systematic reductions in both omission and misclassification errors—the hallmark failure modes of prior methods. However, limitations remain, particularly in detecting fine-grained categories. The dataset's scale and diversity also impose constraints, impacting the model's generalization capability. Future research will focus on the specific challenges of remote sensing scenarios and the various limitations of existing methods.

Author Contributions: Conceptualization, M.J. and Y.F.; methodology, M.J.; software, M.J.; validation, M.J.; formal analysis, M.J.; investigation, M.J.; resources, M.J.; data curation, M.J.; writing—original draft preparation, M.J.; writing—review and editing, M.J. and Y.F.; visualization, M.J.; supervision, Y.F., W.D. and C.L.; project administration, Y.F., W.D. and C.L.; funding acquisition, Y.F., W.D. and C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Key R&D Program of China (2024YFF1401001). This research was supported by the National Natural Science Foundation of China under Grant 62301538. This research was supported by the Science and Disruptive Technology Program 2025-AIRCAS-SDTP-04.

Data Availability Statement: The DIOR dataset presented in this study is publicly accessible at https://ieee-dataport.org/documents/dior accessed on 4 April 2025. The DOTAv2.0 dataset can be accessed via https://captain-whu.github.io/DOTA/dataset.html accessed on 4 April 2025 . The NWPU-VHR10 dataset is available at https://opendatalab.org.cn/OpenDataLab/NWPU_VHR-10 accessed on 5 April 2025.

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

References

1. 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]
2. Cho, H.; Seo, Y.W.; Kumar, B.V.K.V.; Rajkumar, R.R. A multi-sensor fusion system for moving object detection and tracking in urban driving environments. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014.
3. Peeling, J.A.; Chen, C.; Judge, J.; Singh, A.; Achidago, S.; Eide, A.; Tarrio, K.; Olofsson, P. Applications of remote sensing for land use planning scenarios with suitability analysis. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 6366–6378. [CrossRef]
4. Hu, Z.; Yang, D.; Zhang, K.; Chen, Z. Object Tracking in Satellite Videos Based on Convolutional Regression Network with Appearance and Motion Features. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 783–793. [CrossRef]
5. Bi, F.; Sun, J.; Lei, M.; Wang, Y.; Sun, X. Remote Sensing Target Tracking for UAV Aerial Videos Based on Multi-Frequency Feature Enhancement. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020.
6. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
7. Yu, Y.; Gu, T.; Guan, H.; Li, D.; Jin, S. Vehicle detection from high-resolution remote sensing imagery using convolutional capsule networks. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1894–1898. [CrossRef]
8. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
9. Meng, X.; Shen, H.; Yuan, Q.; Li, H.; Zhang, L.; Sun, W. Pansharpening for cloud-contaminated very high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2840–2854. [CrossRef]
10. Yang, X.; Yu, Y. Estimating soil salinity under various moisture conditions: An experimental study. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2525–2533. [CrossRef]

11. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 19730–19742.
12. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 10965–10975.
13. 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*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 38–55.
14. 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. *arXiv* **2024**, arXiv:2408.09110. [CrossRef]
15. Wang, G.; Xie, J.; Zhang, T.; Sun, Y.; Chen, H.; Zhuang, Y.; Li, J. LLaMA-Unidetector: An LLaMA-Based Universal Framework for Open-Vocabulary Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4409318. [CrossRef]
16. 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.* **2016**, *39*, 1137–1149. [CrossRef]
17. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the Computer Vision & Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
18. Li, X.; Wang, J.; Xu, F.; Song, J. Improvement of YOLOv3 algorithm in workpiece detection. In Proceedings of the 2019 IEEE 9th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), Suzhou, China, 29 July–2 August 2019; pp. 1063–1068.
19. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
20. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14454–14463.
21. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
22. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; Zhang, L. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv* **2022**, arXiv:2201.12329. [CrossRef]
23. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv* **2022**, arXiv:2203.03605.
24. Tao, C.; Qi, J.; Zhang, G.; Zhu, Q.; Lu, W.; Li, H. TOV: The original vision model for optical remote sensing image understanding via self-supervised learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 4916–4930. [CrossRef]
25. Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D.; Ermon, S. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 197–211.
26. Reed, C.J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; Darrell, T. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 4088–4099.
27. Bastani, F.; Wolters, P.; Gupta, R.; Ferdinando, J.; Kembhavi, A. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 16772–16782.
28. Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5612822. [CrossRef]
29. Guo, X.; Lao, J.; Dang, B.; Zhang, Y.; Yu, L.; Ru, L.; Zhong, L.; Huang, Z.; Wu, K.; Hu, D.; et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 27672–27683.
30. He, X.; Tang, C.; Zou, X.; Zhang, W. Multispectral object detection via cross-modal conflict-aware learning. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 1465–1474.
31. Zhang, T.; Zhuang, Y.; Wang, G.; Chen, H.; Li, L.; Li, J. A Unified Remote Sensing Object Detector Based on Fourier Contour Parametric Learning. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5611225. [CrossRef]
32. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.

33. Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; Li, G. Learning to prompt for open-vocabulary object detection with vision-language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14084–14093.
34. Gao, M.; Xing, C.; Niebles, J.C.; Li, J.; Xu, R.; Liu, W.; Xiong, C. Open vocabulary object detection with pseudo bounding-box labels. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 266–282.
35. Kim, D.; Angelova, A.; Kuo, W. Region-aware pretraining for open-vocabulary object detection with vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 11144–11154.
36. Bi, H.; Feng, Y.; Tong, B.; Wang, M.; Yu, H.; Mao, Y.; Chang, H.; Diao, W.; Wang, P.; Yu, Y.; et al. RingMoE: Mixture-of-Modality-Experts Multi-Modal Foundation Models for Universal Remote Sensing Image Interpretation. *arXiv* **2025**, arXiv:2504.03166. [CrossRef]
37. Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; Wei, X.; Xia, H.; Shen, C. Conditional Positional Encodings for Vision Transformers. *arXiv* **2021**, arXiv:2102.10882.
38. Islam, M.A.; Jia, S.; Bruce, N.D.B. How Much Position Information Do Convolutional Neural Networks Encode? *arXiv* **2020**, arXiv:2001.08248. [CrossRef]
39. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
40. Hu, J.; Shen, L.; Sun, G.; Albanie, S. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, Salt Lake City, UT, USA , 18–23 June 2017.
41. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language Modeling with Gated Convolutional Networks. *arXiv* **2016**, arXiv:1612.08083.
42. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Stroudsburg PA USA, 2019; pp. 4171–4186.
43. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
44. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
45. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, Sardinia, Italy, 13–15 May 2010; pp. 249–256.
46. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. *CBAM: Convolutional Block Attention Module*; Springer: Cham, Switzerland, 2018.
47. Ouyang, D.; He, S.; Zhan, J.; Guo, H.; Huang, Z.; Luo, M.L.; Zhang, G.L. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. *arXiv* **2023**, arXiv:abs/2305.13563.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.