Two-Stage Fine-Tuning of Large Vision-Language Models with Hierarchical Prompting for Few-Shot Object Detection in Remote Sensing Images

Yonggi Shi *, Ruopeng Yang, Changsheng Yin, Yiwei Lu, Bo Huang, Yu Tao, Yihao Zhong

Posted Date: 22 December 2025

doi: 10.20944/preprints202512.1915.v1

Keywords: few-shot object detection; large vision-language models; hierarchical prompting; LoRA; remote sensing imagery

 doi:10.20944/preprints202512.1915.v1

Article

Two-Stage Fine-Tuning of Large Vision-Language Models with Hierarchical Prompting for Few-Shot Object Detection in Remote Sensing Images

Yongqi Shi [1,*], Ruopeng Yang [2], Changsheng Yin [2], Yiwei Lu [2], Bo Huang [1], Yu Tao [1] and Yihao Zhong [1]

[1] National University of Defense Technology, Wuhan 430030, China
[2] Information Support Force Engineering University, Wuhan 430030, China
* Correspondence: shiyongqi17@nudt.edu.cn

Highlights

What are the main findings?

- A two-stage, parameter-efficient fine-tuning framework with hierarchical prompting is proposed to adapt Qwen3-VL for few-shot object detection in remote sensing imagery. It substantially improves novel-class mAP while largely preserving base-class performance on the DIOR and NWPU VHR-10.v2 datasets.
- A three-level hierarchical prompting scheme—comprising task instructions, global remote sensing context, and fine-grained category semantics—effectively reduces class confusion and enhances the localization of small and visually similar objects in complex very-high-resolution remote sensing imagery.

What are the implications of the main findings?

- The results demonstrate that combining large vision-language models with parameter-efficient adaptation is a practical and effective strategy for transferring general-purpose multimodal models to specialized remote sensing detection tasks under scarce annotations, without incurring severe catastrophic forgetting.
- The effectiveness of injecting structured semantic priors through hierarchical prompts suggests a general paradigm for improving robustness and cross-dataset transfer in other remote sensing applications, such as instance segmentation and change detection.

Abstract

Few-shot object detection (FSOD) in high-resolution remote sensing (RS) imagery remains challenging due to scarce annotations, large intra-class variability, and high visual similarity between categories, which together limit the generalization ability of convolutional neural network (CNN)-based detectors. To address this issue, we explore leveraging large vision-language models (LVLMs) for FSOD in RS. We propose a two-stage, parameter-efficient fine-tuning framework with hierarchical prompting that adapts Qwen3-VL for object detection. In the first stage, low-rank adaptation (LoRA) modules are inserted into the vision and text encoders, and a Detection Transformer (DETR)-style detection head is trained on fully annotated base classes. In the second stage, the vision LoRA parameters are frozen, the text encoder is updated using K-shot novel-class samples, and the detection head is partially frozen, with selected components refined under a three-level hierarchical prompting scheme from superclasses to fine-grained class descriptions. To preserve base-class performance and reduce class confusion, we further introduce knowledge distillation and semantic consistency losses. Experiments on the DIOR and NWPU VHR-10.v2 datasets show that the proposed method consistently improves novel-class performance while maintaining competitive base-class accuracy and surpasses existing baselines, demonstrating the effectiveness of integrating hierarchical semantic reasoning into LVLM-based FSOD for RS imagery.

 doi:10.20944/preprints202512.1915.v1

Keywords: few-shot object detection; large vision-language models; hierarchical prompting; LoRA; remote sensing imagery

1. Introduction

Remote sensing object detection (RSOD) underpins a wide range of military and civilian applications, such as target surveillance, urban management, transportation monitoring, environmental assessment, and disaster response. With the continuous improvement of spatial resolution and revisit frequency, as well as the increasing availability of multi-modal data, automatic and robust RS object detection has become a core component of large-scale Earth observation systems. In the past decade, deep learning has significantly advanced this field: CNNs [1–3], and more recently Vision Transformers (ViTs) [4,5] together with DETR-style end-to-end detectors [6–10], have become the dominant paradigms for object detection in RS imagery. These models learn powerful hierarchical representations and have achieved state-of-the-art performance on many RS benchmarks. Nevertheless, RS imagery is characterized by arbitrary orientations, extreme aspect ratios, cluttered backgrounds, and pronounced scale and appearance variations, which have led to a diverse set of oriented and rotated detection methods [11–13]. Despite these advances, current detectors still rely heavily on large-scale, high-quality annotations, whose acquisition is labor-intensive and requires expert knowledge due to complex scenes and visually similar categories. Although semi-supervised and weakly supervised learning, self-supervised pretraining, synthetic data, and active learning can partly reduce annotation costs, data-efficient detection remains a major bottleneck, especially for newly emerging or rare categories and under distribution shifts across regions, sensors, and seasons [6,14–17].

FSOD, which aims to detect instances of novel classes given only a few labeled examples per category, provides a promising direction to alleviate this problem. Visual-only FSOD methods, such as TFA [18] and DeFRCN [19] and their variants, have achieved encouraging results on natural image benchmarks by refining feature learning and classification heads. However, in the extremely low-shot regime, these approaches still suffer from insufficient feature representation and severe overfitting. This issue becomes more pronounced in RS scenarios, where complex backgrounds, large intra-class variability, and subtle inter-class differences (e.g., among different types of vehicles, ships, or infrastructures) are ubiquitous. Moreover, most existing FSOD frameworks focus almost exclusively on visual feature learning and pay limited attention to explicit modeling of semantic information, which is essential for humans when recognizing novel objects. Semantic cues—such as textual descriptions of appearance, function, typical context, and relations between categories—can provide complementary priors to distinguish visually similar classes and to transfer knowledge from base to novel categories [20–22]. The insufficient exploitation of such semantic information has thus become a key barrier to building robust and generalizable FSOD models in RS imagery.

The rapid progress of LVLMs offers new opportunities to address this limitation. Recent studies have demonstrated that LVLMs possess strong category-level semantics and open-vocabulary recognition capabilities, which can in principle be transferred to RS object detection [23–25]. However, directly applying generic LVLMs to RS data is far from trivial. On the one hand, naïve fine-tuning of LVLMs is computationally expensive and prone to overfitting or catastrophic forgetting, especially when only a few RS samples are available. On the other hand, simple or manually designed prompts often fail to fully exploit the rich semantic knowledge embedded in LVLMs or to adapt it to the specific characteristics of RS imagery.

In this work, we address these challenges by proposing a two-stage fine-tuning framework of LVLMs with hierarchical prompting for FSOD in RS images. Our method builds upon Qwen3-VL [26] and adopts a parameter-efficient LoRA [27] scheme to adapt the model to remote sensing few-shot object detection (RS-FSOD) under scarce supervision. The core idea is to inject multi-level semantic priors into the detection pipeline through a hierarchical prompting mechanism, including task-level instructions, global RS context, and fine-grained category descriptions that encode

 doi:10.20944/preprints202512.1915.v1

appearance attributes, functional roles, and typical surroundings. Furthermore, we design a two-stage fine-tuning strategy: in the first stage, LoRA modules are inserted into the vision and text encoders and optimized jointly with a DETR-style detection head on fully annotated base classes, in order to learn RS-adapted cross-modal representations and basic detection capability; in the second stage, we freeze the visual LoRA modules and selectively adapt the text encoder and parts of the detection head using K-shot novel-class samples, while introducing knowledge distillation and semantic consistency losses to preserve base-class knowledge and better align visual features with hierarchical prompts. This selective adaptation enables the model to acquire novel-class knowledge from both visual examples and semantic prompts, while alleviating catastrophic forgetting.

The main contributions of this paper can be summarized as follows.

1. We propose a novel LVLM-based framework for FSOD in RS imagery, which builds upon Qwen3-VL and employs parameter-efficient fine-tuning to adapt LVLMs to RS-FSOD under scarce annotations.
2. We design a hierarchical prompting strategy that injects multi-level semantic priors—spanning task instructions, global RS context, and fine-grained category descriptions—into the detection pipeline, effectively compensating for limited visual supervision and mitigating confusion among visually similar categories.
3. We develop a two-stage LoRA-based fine-tuning scheme that jointly optimizes vision and language branches on base classes and selectively adapts the text encoder and detection head for novel classes with distillation and semantic consistency constraints, thereby enhancing transferability to novel RS categories while mitigating catastrophic forgetting. Extensive experiments on DIOR and NWPU VHR-10.v2 demonstrate that the proposed method achieves competitive or superior performance compared with state-of-the-art FSOD approaches.

2. Related Work

Recent years have witnessed rapid progress in object detection for RS imagery, including both fully supervised detectors and few-shot models under data-scarce settings. In Section 2.1, we review representative CNN- and Transformer-based detectors for generic RSOD. In Section 2.2, we summarize advances in FSOD for RS and position our work within this line of research.

2.1. Object Detection in Remote Sensing Imagery

RSOD has evolved significantly with the introduction of deep learning, which has enabled detectors to better cope with the unique characteristics of aerial and satellite imagery. Early methods primarily focused on adapting traditional CNN-based architectures to handle the challenges presented by RS imagery, such as scale variation and dense object distributions. Models like ABNet [28], which integrates enhanced feature pyramids and attention mechanisms, and CFIM [29], which focuses on context-aware sampling, have laid the foundation for addressing the complexities of RSOD. These approaches typically combine enhanced feature pyramids, attention mechanisms, and multiscale processing to improve detection performance under varied scenes and object densities.

In recent years, the shift towards Transformer-based models has further advanced RSOD, largely due to their ability to model global contextual information effectively. For instance, PCViT [30] introduces a pyramid convolutional vision transformer that enhances feature extraction through parallel convolutional modules, showing remarkable performance in dense and multiscale RSOD scenarios. The CoF-Net framework [31] further refines this approach with a coarse-to-fine feature adaptation strategy that boosts detection accuracy under challenging conditions, such as complex backgrounds and varying object sizes. Transformer-based models have also been optimized to address specific challenges, such as rotated objects and small targets. MSFN [11] uses spatial-frequency domain features to improve the detection of rotated objects, while CPMFNet [12] focuses on arbitrary-oriented object detection through adaptive spatial perception modules. Moreover, models like EIA-PVT [13] have extended Transformer-based architectures to integrate local and

 doi:10.20944/preprints202512.1915.v1

global context in a more compact and efficient manner, demonstrating their potential for robust RSOD in highly variable environments.

Despite these advancements, detecting small, low-contrast, and densely distributed objects remains challenging in RSOD due to the high variation in size and shape, along with the presence of cluttered backgrounds. Approaches like RS-TOD [32] and DCS-YOLOv8 [33] have tackled these challenges by introducing innovative modules that enhance sensitivity to small objects and improve performance on datasets with small target objects. For instance, RS-TOD employs a remote sensing attention module (RSAM) and a lightweight detection head specifically designed for tiny objects. Similarly, methods such as GHOST [34] and Lino-YOLO [35] focus on reducing model complexity and improving efficiency, which is crucial for applications where real-time detection is needed. Additionally, advancements such as Bayes R-CNN [1] and YOLOv8-SP [36] are pushing the boundaries of object detection by incorporating uncertainty estimation and small-object focused architecture, respectively.

Furthermore, the challenges of domain adaptation and the scarcity of labeled data in RS have spurred interest in semi-supervised and few-shot learning methods. FADA [14] and FIE-Net [15] propose foreground-alignment strategies to mitigate domain shift, while UAT [16] leverages uncertainty-aware models for semi-supervised detection. Few-shot detection has seen significant progress with FSDA-DETR [6] and MemDeT [37], which enhance detection performance under limited data conditions by incorporating memory-augmented modules and cross-domain feature alignment. We provide a more detailed review of few-shot RSOD in Section 2.2. These techniques are crucial in situations where labeled data is sparse or not available for certain target domains. Self-training and pseudo-labeling techniques, such as those found in PSTDet [17], have also contributed to improving detection performance in these scenarios.

Finally, with the rise of LVLMs, the integration of semantic priors from language and multimodal pretraining has opened up new possibilities for RSOD. LLaMA-Unidetector [23] leverages the power of open-vocabulary detection by utilizing large language models to identify unseen object categories, significantly reducing the reliance on manually labeled data. This approach is further enhanced by frameworks like CDATOD-Diff [38], which combine CLIP-guided semantic priors with generative diffusion modules to improve detection in SAR imagery. Although these models show promise, challenges remain in effectively aligning vision-language representations with the unique characteristics of RS imagery, especially under few-shot supervision. This paper aims to bridge this gap by introducing a two-stage fine-tuning strategy coupled with hierarchical prompting, thereby enhancing the adaptability of LVLMs to few-shot RSOD scenarios.

2.2. Few-Shot Object Detection in Remote Sensing Imagery

FSOD aims to alleviate the data-hungry nature of deep detectors and has been actively explored in both natural and RS imagery. In RS, early works mainly followed meta-learning paradigms, learning class-agnostic meta-features and class-specific prototypes from episodic tasks [2,3,39–41]. Examples include prototype-guided frameworks such as P-CNN [2], memory-augmented MM-RCNN [3], the YOLOv3-based meta-learner [39], multiscale contrastive MSOCL [40], and discriminative prototype learning in DPL-Net [41]. Although these methods demonstrate the feasibility of FSOD in high-resolution aerial scenes, they often suffer from optimization complexity, sensitivity to episodic design, and limited scalability to large vocabularies and ultra-high-resolution images.

Motivated by progress on natural image benchmarks, fine-tuning-based FSOD has become the mainstream in RS. A large body of methods adopts a two-phase “base-training then few-shot fine-tuning” pipeline, while introducing RS-specific designs for scale, orientation, and class imbalance. Prototype-centric approaches refine the classification branch with more expressive class representations: HPMF [20] and InfRS [42] construct discriminative or hypergraph-guided prototypes, while G-FSDet [43] and related generalized/incremental FSOD methods [44–47] further address catastrophic forgetting through imprinting, drift rectification, multiscale knowledge

 doi:10.20944/preprints202512.1915.v1

distillation, and frequency-domain prompting. Other works explicitly tackle geometric and multiscale variations via transformation-invariant alignment [48], multiscale positive refinement [49], contrastive multiscale proposals [40], and multilevel query-support interaction [21,50,51]. Domain-specific issues such as complex backgrounds, small/dense objects, and biased proposals are handled by unbiased proposal filtration and tiny-object-aware losses [52], label-consistent classification and gradual regression [53], subspace separation with low-rank adapters [54], and augmentation-/self-training-based strategies that cope with sparse or incomplete annotations [22,55,56].

In parallel, contrastive learning, memory, transformers, and domain adaptation have further enriched RS-FSOD. Contrastive objectives are used to enhance inter-class separability under high visual similarity [25,40,56,57], while memory or prototype-guided alignment stabilizes representations and mitigates forgetting [3,37,58]. Transformer-based detectors such as CAMCFormer [21], MemDeT [37], Protoalign [58], FSDA-DETR [6], and CD-ViTO [59] leverage long-range self-/cross-attention for joint modeling of query-support relations and global context, with AsyFOD [60] and FSDA-DETR [6] extending FSOD to cross-domain scenarios via asymmetric or category-aware alignment.

A clear trend in recent work is moving beyond purely visual cues toward semantic and language-guided FSOD. CLIP-based and multimodal prototype methods [24,25,37,58,61,62] exploit textual labels or LVLMs to provide category-level priors that are agnostic to specific images, thereby improving generalization to novel classes under severe data scarcity. At the same time, generative models and frequency-domain techniques are used to enrich supervision, e.g., controllable diffusion-based data synthesis in CGK-FSOD and related work [63,64], and frequency compression and new benchmarks for infrared FSOD in FC-fsd [65]. YOLO-based RS-FSOD systems [39,66,67] also incorporate global context, segmentation assistance, or distillation heads to balance efficiency and accuracy.

Despite these advances, existing RS-FSOD methods typically employ either task-specific meta-learning or single-stage visual fine-tuning, and most semantic- or language-enhanced approaches use relatively shallow fusion strategies or fixed textual prompts [24,25,37,58,61]. In the extremely low-shot regime, they still struggle with limited visual supervision, complex and cluttered backgrounds, and subtle inter-class differences (e.g., among vehicles, ships, and infrastructure). This gap motivates our work, **Two-Stage Fine-Tuning of Large Vision-Language Models with Hierarchical Prompting for Few-Shot Object Detection in Remote Sensing Images**, which explicitly leverages large-scale vision-language pretraining and a hierarchical prompting scheme within a two-stage fine-tuning framework, aiming to more effectively transfer semantic knowledge to novel RS categories under very scarce annotations.

3. Methods

In this section, we first formalize the few-shot detection setting and introduce the Qwen3-VL base model in Section 3.1. We then describe the overall architecture of our RS-FSOD framework in Section 3.2, followed by the hierarchical prompting mechanism in Section 3.3 and the DETR-style detection head in Section 3.4. Finally, we present the two-stage LoRA-based fine-tuning strategy with knowledge distillation and semantic consistency regularization in Section 3.5.

3.1. Preliminary Knowledge

3.1.1. Problem Setting

We adhere to the standard problem definition of FSOD established in previous works such as TFA [18] and G-FSDet [43]. The objective is to train a detector capable of identifying novel objects given only a few annotated examples, leveraging knowledge transfer from abundant base classes.

Formally, the object categories are divided into two disjoint sets: Base Classes (C_{base})) and Novel Classes (C_{novel}), where $C_{base} \cap C_{novel} = \varnothing$. The corresponding dataset consists of a base dataset D_{base} and a novel dataset D_{novel}.

 doi:10.20944/preprints202512.1915.v1

- **Base Training Phase:** The model is trained on D_{base}, which contains a large number of annotated instances for categories in C_{base}.
- **Fine-tuning Phase:** The model is adapted using D_{novel}, which follows a K-shot setting. For each category in C_{novel}, only K annotated instances (e.g., $K = 3,5,10,20$) are available.

The ultimate goal is to obtain a model that achieves high detection performance on C_{novel} while minimizing catastrophic forgetting on C_{base}.

3.1.2. Base Model: Qwen3-VL

We build our method upon Qwen3-VL, an LVLM that jointly processes images and text. As illustrated in Figure 1, Qwen3-VL comprises a vision encoder, a text encoder, and a multimodal fusion module.

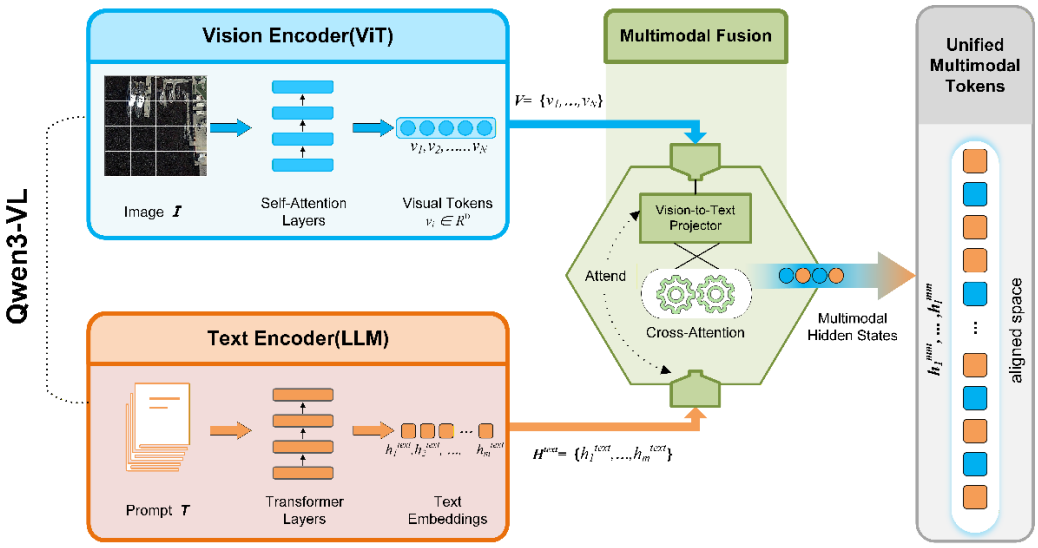

Figure 1. Architecture of the base model Qwen3-VL. Qwen3-VL is a LVLM that jointly processes images and text, and serves as the backbone of our proposed framework. It consists of three main components: (i) a ViT-based vision encoder that extracts visual features from the input RS images, (ii) a Transformer-based language model acting as the text encoder to obtain semantic representations from textual prompts, and (iii) a multimodal fusion module that integrates visual and textual features to produce cross-modal representations for subsequent FSOD.

Vision encoder. The vision encoder is a Vision Transformer (ViT)-style network that takes an input image **I** and converts it into a sequence of visual tokens. The image is first divided into patches, which are linearly projected and added with positional embeddings. The resulting patch embeddings are processed by multiple self-attention layers, producing a sequence of visual features:

$$V = \{v_1, \dots, v_N\}, v_i \in \mathbb{R}^D \quad (1)$$

where N is the number of image tokens and D is the hidden dimensionality. In our framework, these visual tokens serve as the primary representation of RS images.

Text encoder (LLM). The text branch of Qwen3-VL is built on top of a large autoregressive Transformer language model. Given a tokenized prompt sequence $T = \{t_1, \dots, t_M\}$, which in our case encodes the hierarchical instructions and category descriptions, the LLM produces contextualized text embeddings:

$$H^{text} = \{h_1^{text}, \dots, h_M^{text}\} \quad (2)$$

These embeddings capture rich semantic information about the detection task, RS context, and fine-grained categories, and can interact with visual tokens through multimodal fusion.

Multimodal fusion. Qwen3-VL integrates visual and textual information via a series of cross-attention and projection layers that map visual tokens into the language model's representation space. Concretely, the visual features V are first transformed by a vision-to-text projector and then injected

 doi:10.20944/preprints202512.1915.v1

into the LLM as special tokens or through cross-attention blocks. This produces a sequence of multimodal hidden states:

$$H^{mm} = \{h_1^{mm}, \dots, h_L^{mm}\} \quad (3)$$

where each token can attend to both visual and textual information. These multimodal tokens form a unified representation space in which visual patterns and language semantics are aligned.

In our work, we treat Qwen3-VL as a frozen multimodal backbone and adapt it to the RS-FSOD task using parameter-efficient fine-tuning. We take the multimodal hidden states H^{mm} as the input memory for a lightweight detection head, and insert LoRA modules into selected linear layers of the vision encoder and text encoder, as detailed in the following sections.

3.2. Overall Architecture

In this work, we propose a two-stage fine-tuning framework built upon the Qwen3-VL, combined with a hierarchical prompting mechanism to enhance FSOD in RS imagery. The overall architecture is illustrated in Figure 2. It consists of a Qwen3-VL backbone that performs cross-modal feature learning, a DETR-style detection head for box and class prediction, and a two-stage LoRA-based parameter-efficient fine-tuning strategy guided by hierarchical prompts, knowledge distillation, and semantic consistency constraints.

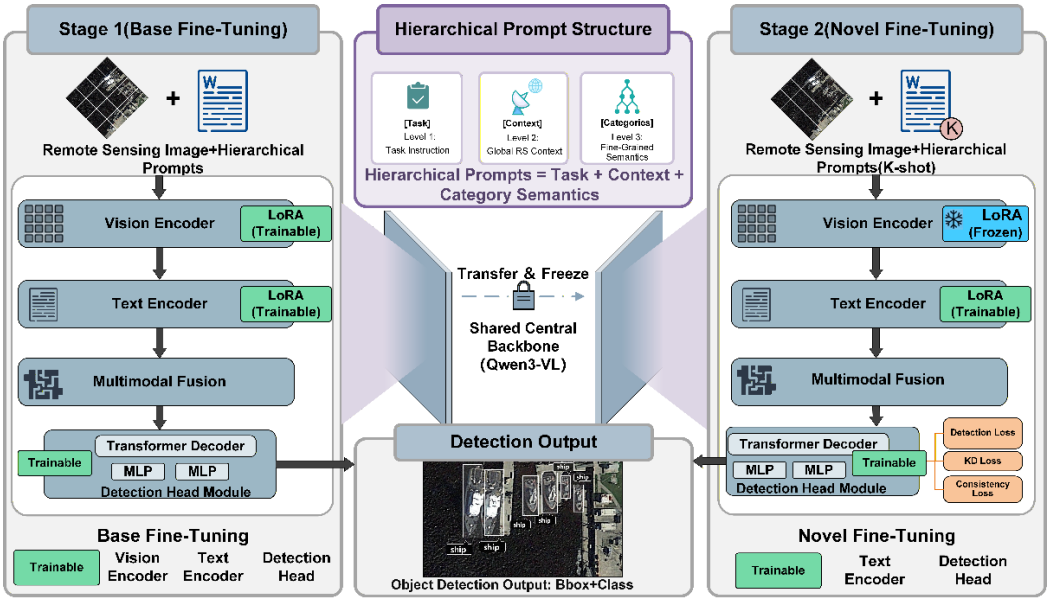

Figure 2. Overview of the proposed Qwen3-VL-based few-shot detection framework. The architecture consists of a Qwen3-VL backbone for cross-modal feature learning, a DETR-style detection head, and a two-stage LoRA-based fine-tuning strategy guided by hierarchical prompts. In the Base Fine-tuning stage (left), LoRA adapters are inserted into the vision and text encoders and trained jointly with the fully learnable detection head on fully annotated base classes, enabling the model to acquire remote-sensing-adapted cross-modal representations and basic detection capability. In the Novel Fine-tuning stage (right), visual LoRA modules are frozen, while LoRA in the text encoder and selected layers of the detection head remain trainable and are updated using K-shot novel-class samples. Hierarchical prompts encode superclass-to-fine-class semantics for both stages, and knowledge distillation together with semantic consistency loss is applied in Stage 2 to preserve base-class knowledge and reduce semantic drift during few-shot adaptation.

We first adapt Qwen3-VL to the detection setting by attaching a detection head on top of the fused multimodal features. Concretely, we follow a DETR-style design: the multimodal hidden states produced by Qwen3-VL are treated as the encoder memory, and a set of learnable object queries are fed into a multi-layer Transformer decoder. Each decoded query feature is then passed through two small multilayer perceptrons (MLPs): a classification head that outputs category logits and a regression head that predicts bounding box parameters. This design preserves the simplicity and set-

 doi:10.20944/preprints202512.1915.v1

prediction nature of DETR-style detectors, while making it straightforward to consume the sequence of multimodal tokens produced by Qwen3-VL.

To adapt the large backbone to RS-FSOD in a parameter-efficient manner, we adopt LoRA and implement it through the Parameter-Efficient Fine-Tuning (PEFT) framework [68]. We insert LoRA modules into the key linear layers of the ViT-based vision encoder and the Transformer-based language model (text encoder), while the DETR-style detection head is trained with its own parameters but without LoRA adapters. During fine-tuning, we mainly update the LoRA parameters in the encoders together with the parameters of the detection head, while keeping the original Qwen3-VL weights frozen. This design significantly reduces the number of trainable parameters in the backbone and allows the pretrained multimodal knowledge to be preserved, which is particularly beneficial when transferring to RS domains with limited labeled data.

Figure 2 is organized into two panels. The left panel shows the Base Fine-tuning stage: given RS images and hierarchical prompts, the vision encoder and text encoder are equipped with trainable LoRA adapters (highlighted in green), and the DETR-style detection head is fully trainable. The model is trained on fully annotated base classes to learn remote-sensing-adapted cross-modal representations and basic detection capability. The right panel shows the Novel Fine-tuning stage: the LoRA modules in the vision encoder are frozen (rendered in blue), while only the LoRA parameters in the text encoder and selected layers of the detection head remain trainable. In this stage, the model is adapted using K-shot samples of novel classes, under the guidance of hierarchical prompts that encode superclass-to-fine-class semantics. Detection loss on the novel classes is augmented with knowledge distillation loss and semantic consistency loss to preserve base-class performance and reduce semantic confusion, without mixing additional base-class images into the few-shot training set.

Overall, our architecture exploits (i) cross-modal feature learning in Qwen3-VL, (ii) hierarchical prompts to inject structured semantic priors, (iii) a two-stage LoRA-based fine-tuning strategy that selectively updates the text encoder and parts of the detection head while keeping the visual backbone largely frozen, and (iv) knowledge distillation and consistency regularization to mitigate catastrophic forgetting and semantic drift, thereby achieving robust few-shot detection in complex RS scenes.

3.3. Hierarchical Prompting Mechanism

A key component of our framework is the hierarchical prompting mechanism, which injects structured semantic priors into Qwen3-VL in order to cope with the complexity of RS scenes and the scarcity of novel-class annotations. Simple prompts such as “Detect objects in this image.” provide only coarse task instructions and are insufficient to distinguish visually and semantically similar categories under few-shot supervision, for example separating “airplane” from “airport” or “storage tank” from “chimney” in the DIOR dataset. To address this limitation, we design a three-level hierarchical prompt that explicitly encodes task instructions, global remote sensing context, and fine-grained category semantics. The dataset-specific instantiations (covering all DIOR 20 classes and NWPU VHR-10.v2 10 classes) are provided in Appendix A and Appendix B, respectively.

Level 1: Task Instruction.

The first level specifies the detection task and the desired output format. It instructs the model to localize all objects in an overhead image and output bounding boxes with category labels. This aligns the generic instruction-following capability of Qwen3-VL with the concrete objective of remote sensing object detection. For example:

> *You are an expert in remote sensing object detection. Given an overhead image, detect all objects of interest and output their bounding boxes and category labels.*

Level 2: Global Remote Sensing Context.

The second level provides high-level priors about remote sensing imagery, including the bird’s-eye viewpoint, multi-scale and densely distributed objects, and typical object families. This guides

 doi:10.20944/preprints202512.1915.v1

the model to interpret visual patterns under RS-specific assumptions and reduces ambiguity when visual evidence is weak. For example:

> *The image is an 800×800-pixel overhead view acquired by satellite or aerial sensors under varying ground sampling distances (approximately 0.5–30 m). Objects can be extremely multi-scale, densely packed, and arbitrarily oriented, with frequent background clutter, shadows, and repetitive textures. Scenes cover airports and airfields; expressways and highway facilities (service areas, toll stations, overpasses, bridges); ports and shorelines; large industrial zones (storage tanks, chimneys, windmills, dams); and urban or suburban districts with sports venues (baseball fields, basketball/tennis courts, golf courses, stadiums). Backgrounds can be cluttered and visually simi-lar, and discriminative cues often come from fine-grained shapes, markings, and spatial context.*

Level 3: Fine-Grained Category Semantics.

The third level builds a semantic dictionary of categories, where each superclass and its fine-grained classes are described by short textual definitions, including appearance attributes, functional roles, and typical locations. For each superclass c^{sup}, we enumerate its fine-grained classes $\{c^{fine}\}$ and construct a template-based description. For example:

Superclass: *Road transportation and facilities (vehicles and highway-related structures).*

Fine-grained classes:

- *vehicle: Small man-made transport targets on roads/parking lots (cars, buses, trucks); compact rectangles with short shadows.*
- *expressway service area: Highway rest areas with large parking lots, gas pumps and service buildings; near ramps.*
- *expressway toll station: Toll plazas spanning multiple lanes with booths and canopies; strong lane markings at entries/exits.*
- *overpass: Elevated roadway segments crossing other roads/rails; ramps, pylons and cast shadows.*
- *bridge: Elevated linear structures spanning water or obstacles; approach ramps and structural shadows.*

Similarly, we define categories such as “storage tank”, “ship”, and “tennis court” based on their shapes, textures, and contextual surroundings. These detailed descriptions encourage the model to focus on discriminative cues that separate visually similar categories and to exploit the category hierarchy encoded by the prompt.

In practice, we concatenate the three levels into a single prompt sequence with explicit segment markers (“[Task]”, “[Context]”, “[Categories]”) and feed it into the Qwen3-VL text encoder. The same hierarchical prompt structure is used in both the base and novel stages; **only the fine-grained category list is expanded to include novel classes in the second stage**. By consistently conditioning on hierarchical prompts, the model learns to align visual features with structured semantic information and to leverage language priors for few-shot generalization, especially when novel-class examples are extremely limited.

3.4. DETR-Style Detection Head

To enable end-to-end object detection in the proposed framework, we integrate a DETR-style detection head that directly processes the multimodal features extracted by Qwen3-VL. Unlike conventional anchor-based detectors, our detection head formulates object detection as a set prediction problem solved via bipartite matching, which is particularly advantageous for RS imagery characterized by varying object scales and arbitrary spatial distributions. The detailed architecture of the detection head is illustrated in Figure 3.

 doi:10.20944/preprints202512.1915.v1

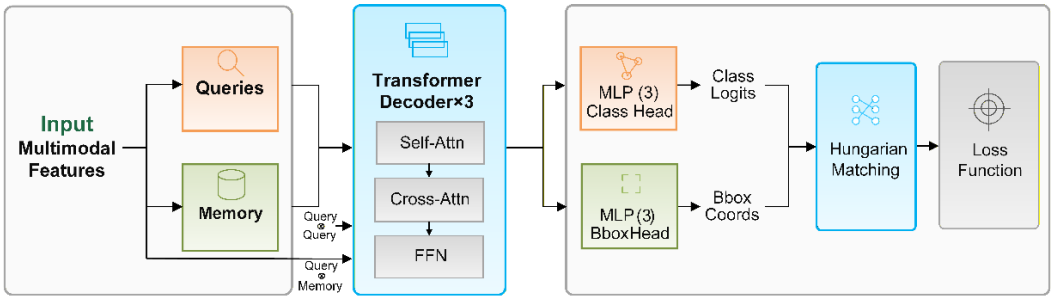

Figure 3. Architecture of the DETR-style detection head.

As illustrated in Figure 3, the detection head consists of three core components:

- A set of $N = 100$ learnable object queries $Q \in \mathbb{R}^{N\times D}$, which serve as input tokens to the decoder and are optimized jointly with the multimodal features H, where D denotes the hidden dimension.
- A Transformer decoder with $N_{\text{dec}} = 3$ layers that iteratively refines query representations through self-attention and cross-attention with the input multimodal features $H \in \mathbb{R}^{B\times L\times D}$ extracted from the vision-language encoder;
- Dual prediction heads that output class probabilities and bounding box coordinates in parallel. Specifically, for each query q_i, the decoder produces an enhanced representation $z_i \in \mathbb{R}^D$, which is subsequently fed into a three-layer multi-layer perceptron (MLP) for classification and a separate three-layer MLP for bounding box regression. The classification head predicts $(C + 1)$-dimensional logits (where $C = 20$ for the DIOR dataset, with an additional background class), while the regression head outputs normalized bounding box parameters $(c_x, c_y, w, h) \in [0,1]^4$.

Detection Loss. We employ a DETR-style detection loss L_{det} based on bipartite matching between predicted queries and ground-truth objects. For each image, the Hungarian algorithm is used to find an optimal assignment that minimizes a matching cost combining classification scores, L1 box regression, and the Generalized Intersection over Union (GIoU) loss. Given a matched pair $(b, \hat{b})$ with class c_i, the detection loss is defined as:

$$L_{\text{det}} = L_{\text{ce}} + \lambda_{\text{bbox}} L_{\text{bbox}} + \lambda_{\text{giou}} L_{\text{giou}} \quad (4)$$

where L_{ce} denotes the classification loss, L_{bbox} represents the L1 bounding box regression loss, and L_{giou} corresponds to the GIoU loss. To address the class imbalance prevalent in RS datasets, we implement L_{ce} as focal loss with $\alpha = 0.25$ and $\gamma = 2.0$, and assign a reduced weight $\varepsilon = 0.1$ to the background class. The loss weights are empirically set to $\lambda_{\text{bbox}} = 5.0$ and $\lambda_{\text{giou}} = 2.0$, prioritizing localization accuracy over classification. The matching cost used in Hungarian assignment is

$$C = -\hat{p}(c_i) + 5\left\|b - \hat{b}\right\|_1 + 2\left(1 - \text{GIoU}(b, \hat{b})\right) \quad (5)$$

where $\hat{p}(c_i)$ denotes the predicted probability for class c_i, and b and $\hat{b}$ denote the ground-truth and predicted boxes, respectively.

Adaptation for Few-Shot Learning. In the few-shot fine-tuning stage (Stage 2), we adopt a partial-freezing strategy, where the first $N_{\text{dec}} - 1$ decoder layers are frozen to preserve learned general object representations, while the final decoder layer and both prediction heads remain trainable to adapt to novel classes with limited samples. This design mitigates overfitting on scarce training data while retaining the capacity to discriminate fine-grained class-specific features. The specific detection loss and training objectives for the two stages are detailed in Section 3.5.

3.5. Two-Stage Fine-Tuning Strategy

We design a two-stage fine-tuning strategy to efficiently adapt Qwen3-VL to RS-FSOD while mitigating catastrophic forgetting and exploiting hierarchical semantic priors.

Stage 1: Base Training.

 doi:10.20944/preprints202512.1915.v1

In the first stage, we insert LoRA modules into both the vision encoder and the text encoder, and jointly optimize them together with the DETR-style detection head on fully annotated base classes. Given an input RS image and its corresponding hierarchical prompt, the model outputs query-wise class probabilities and bounding box predictions. Let $L_{\text{det}}^{\text{base}}$ denote the standard DETR-style detection loss on base classes, which combines classification loss and box regression loss under Hungarian matching. The Stage-1 objective is:

$$\mathcal{L}^{stage1} = L_{\text{det}}^{\text{base}} \tag{6}$$

Through this stage, the model acquires remote-sensing-adapted cross-modal representations and basic detection capability on the base-class set C_{base}. The resulting model is then frozen as a teacher to guide the subsequent few-shot adaptation.

Stage 2: Few-Shot Adaptation.

In the second stage, we adapt the model to novel classes using only K-shot labeled samples per novel class. To preserve the visual representations learned in Stage 1, we freeze all LoRA parameters in the vision encoder. For the DETR-style detection head, we adopt a partial-freezing strategy, where the first $N_{\text{dec}} - 1$ decoder layers are frozen to retain general object representations, while the final decoder layer and both prediction heads remain trainable. At the same time, we update the LoRA modules in the text encoder. This selective adaptation allows the model to refine its semantic understanding and detection behavior for novel categories, while keeping the visual backbone features and most of the detection head stable. Compared with full-parameter fine-tuning, this strategy reduces overfitting to scarce novel data and better preserves base-class detection ability.

The overall loss in Stage 2 combines the standard detection loss with a knowledge distillation loss and a semantic consistency loss:

$$\mathcal{L}^{stage2} = \mathcal{L}_{det} + \lambda_{KD}\mathcal{L}_{KD} + \lambda_{sem}\mathcal{L}_{sem} \tag{7}$$

Knowledge distillation loss.

To prevent catastrophic forgetting of base classes, we distill both classification and localization knowledge from the Stage-1 teacher model. For images containing base-class instances, let z_k^T and z_k^S be the classification logits of the k-th query from the teacher and student, and let b_k^T, b_k^S denote their corresponding box predictions. The knowledge distillation loss is defined as:

$$\mathcal{L}_{KD} = \sum_k \left[KL\left(\sigma\left(\frac{z_k^T}{T}\right) \middle\| \sigma\left(\frac{z_k^S}{T}\right)\right) + \beta \left\| b_k^T - b_k^S \right\|_1 \right] \tag{8}$$

where $\sigma(\cdot)$ is the softmax function, T is a temperature, and β controls the weight of the regression term. This loss encourages the student to preserve the teacher's behavior on base classes, thereby alleviating catastrophic forgetting.

Semantic consistency loss.

To better exploit the hierarchical prompts, we explicitly enforce consistency between fine-grained predictions and superclass-level semantics. Let M be a mapping from each superclass c^{sup} to its fine-grained classes $\{c^{fine}\}$. Given the fine-grained class probabilities $p_{fine}(c^{fine}|q)$ of a query q, we aggregate them to obtain an induced superclass distribution:

$$\hat{p}_{sup}(c^{sup}|q) = \sum_{c^{fine} \in M(c^{sup})} p_{fine}(c^{fine}|q) \tag{9}$$

Meanwhile, guided by superclass-level textual descriptions in the hierarchical prompt, the model produces an explicit superclass-level distribution $p_{sup}^{model}(\cdot\,|q)$. The semantic consistency loss is then defined as:

$$\mathcal{L}_{sem} = \sum_q KL\left(\hat{p}_{sup}(\cdot\,|q) \middle\| p_{sup}^{model}(\cdot\,|q)\right) \tag{10}$$

which enforces alignment between the aggregated fine-grained predictions and the superclass semantics encoded by the hierarchical prompts.

By freezing the visual LoRA modules, selectively adapting the text encoder and parts of the detection head, and incorporating knowledge distillation and semantic consistency regularization, the proposed two-stage strategy enables the model to acquire novel-class knowledge from both visual

 doi:10.20944/preprints202512.1915.v1

examples and hierarchical prompts, while effectively alleviating catastrophic forgetting of base classes.

4. Experiments and Results

In this section, we present experimental results to evaluate the proposed Qwen3-VL-based RS-FSOD framework. In Section 4.1, we introduce the datasets and few-shot evaluation protocols. In Section 4.2, we describe the evaluation metrics and implementation details. In Section 4.3, we report main results on the DIOR and NWPU VHR-10.v2 benchmarks, and in Section 4.4, we conduct ablation studies on hierarchical prompting, the two-stage fine-tuning strategy, and the choice of LoRA rank.

4.1. Datasets and Evaluation Metrics

4.1.1. Datasets

To evaluate the effectiveness of the proposed method, we conducted extensive experiments on two widely used optical RS datasets: DIOR and NWPU VHR-10.v2.

DIOR Dataset [69]**:** The DIOR dataset is a large-scale and challenging benchmark for optical RSOD. It consists of 23,463 images with a fixed size of 800×800 pixels and covers 20 categories. The images are collected from Google Earth, with spatial resolutions ranging from 0.5m to 30m. Characterized by high inter-class similarity and substantial variations in object scale, this dataset serves as a rigorous testbed for evaluating few-shot learning in complex scenes. The 20 object categories include: *airplane, airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, golf course, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill.*

NWPU VHR-10.v2 Dataset [70]**:** The NWPU VHR-10.v2 dataset is a high-resolution benchmark for geospatial object detection. It comprises 1,173 optical images with a standardized size of 400×400 pixels and covers 10 categories. The images are cropped from Google Earth and the Vaihingen dataset, featuring very high spatial resolution details. Distinguished by its fine-grained object textures and diverse orientations, this dataset is utilized to verify the generalization ability of models across different domains. The 10 object categories include: *airplane, baseball diamond, basketball court, bridge, ground track field, harbor, ship, storage tank, tennis court, and vehicle.*

We adopt standard class splitting and K-shot sampling protocols [43,61]. For DIOR, 15 categories are used as base classes and the remaining 5 are treated as novel classes; for NWPU VHR-10.v2, 7 categories are designated as base and 3 as novel. The detailed compositions of base and novel sets for both datasets are provided in Table 1 and Table 2. During training, the model is first trained on the base classes with abundant annotations, and then fine-tuned in a second stage on a balanced few-shot dataset containing exactly K annotated instances per class. We evaluate performance under four low-data regimes with $K \in \{3,5,10,20\}$.

Table 1. Base/Novel Class Splits on DIOR.

Dataset	Split	Novel					Base
DIOR	1	Baseball field	Basketball court	Bridge	Chimney	Ship	others
	2	Airplane	Airport	Expressway toll station	Harbor	Ground track field	others
	3	Dam	Golf course	Storage tank	Tennis court	Vehicle	others

 doi:10.20944/preprints202512.1915.v1

4	Express service area	Overpass	Stadium	Train station	Windmill	others

Table 2. Base/Novel Class Splits on NWPU VHR-10.v2.

Dataset	Split	Novel			Base
NWPU VHR-10.v2	1	Airplane	Baseball diamond	Tennis court	others
	2	Basketball	Ground track field	Vehicle	others

4.1.2. Evaluation Metrics

We adopt the mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5 as the primary evaluation metric for our few-shot object detector. To comprehensively analyze the model's performance and the problem of catastrophic forgetting, we report three specific indicators:

Base mAP (mAP_{base}): The mAP on the base classes, used to monitor the extent of knowledge retention from the base training stage.

Novel mAP (mAP_{novel}): The mAP calculated only on the unseen novel classes. This is the most critical metric for evaluating few-shot learning efficacy.

Overall mAP (mAP_{all}): The mAP averaged over all categories.

4.2. Implementation Details

4.2.1. Environment and Pretraining

All experiments are implemented in PyTorch using the HuggingFace Transformers and PEFT libraries. Training is conducted on two NVIDIA RTX 6000 Ada GPUs with mixed-precision (FP16) to accelerate both stages, and we employ DeepSpeed (stage 2) for memory-efficient data-parallel training. We adopt Qwen3-VL-8B as the backbone model. The backbone is initialized from the publicly released Qwen3-VL-8B checkpoint, and all original backbone parameters are kept frozen during fine-tuning; only the LoRA adapters in the vision and text encoders, together with the lightweight DETR-style detection head, are updated.

4.2.2. Hyperparameters

We use the same set of hyperparameters for all experiments with Qwen3-VL-8B. AdamW is used for optimization with a weight decay of 1×10^{-4} in both stages.

Stage 1 (base training): initial learning rate 1×10^{-4}, batch size 16 images per GPU (total 32), maximum 30 epochs.

Stage 2 (few-shot adaptation): initial learning rate 5×10^{-5}, batch size 8 images per GPU (total 16), maximum 15 epochs.

In both stages, we adopt a cosine learning-rate schedule with a linear warm-up of 1 epoch. The LoRA rank r is set to 128 by default, with $lora\_alpha = 256$ and $lora\_dropout = 0.1$.

4.2.3. Training Details

In Stage 1, the model is trained on the full base dataset with hierarchical prompts covering all base classes. In Stage 2, we construct the few-shot novel set by randomly sampling K instances per novel class, and fine-tune the model using only these K-shot novel samples under the detection loss, while employing knowledge distillation and semantic consistency losses to preserve base-class knowledge and reduce semantic drift. For both stages, we use early stopping based on validation mAP: if the validation performance does not improve for 5 consecutive epochs, training is stopped and the checkpoint with the best validation mAP is retained.

To reduce the variance caused by few-shot sampling, each K-shot configuration is repeated 5 times with independent sampling trials for the Qwen3-VL-8B model. For each trial, we re-sample K

 doi:10.20944/preprints202512.1915.v1

instances per novel class and re-train Stage 2 starting from the same Stage-1 checkpoint. We report the mean mAP over these 5 runs as the final performance.

4.3. Main Results

To rigorously evaluate our method, we conduct comprehensive experiments on the DIOR and NWPU VHR-10.v2 benchmarks, comparing it against six recent state-of-the-art FSOD approaches that collectively span the two dominant paradigms in RS-FSOD. Specifically, we consider fine-tuning-based methods—such as G-FSDet [43] and RIFR [50]—which emphasize efficient transfer learning and careful balancing of base and novel class performance to alleviate catastrophic forgetting; and prototype-/meta-learning-based methods, including Meta-RCNN [71], P-CNN [2], TPG-FSOD [61], and GIDR [44], which construct class prototypes from limited support samples and align query features via metric learning, attention mechanisms, or hypergraph-guided refinement. Under a unified evaluation protocol—using standard base/novel splits and mAP@0.5 across 3/5/10/20-shot settings—our approach attains consistently competitive performance and, in many settings, yields clear improvements over existing methods.

4.3.1. Results on DIOR

Table 3 presents a detailed comparison on DIOR under 3/5/10/20-shot settings and four splits. Several trends align well with the design of our Qwen3-VL-based two-stage fine-tuning framework. First, our method achieves strong or competitive base mAP compared with fine-tuning-based baselines such as G-FSDet [43] and RIFR [50]. For example, in Split 1, our base mAP remains comparable to RIFR [50] across all shot settings, and is slightly higher at medium and high shots (e.g., 71.03% vs. 69.72% at 10-shot and 71.14% vs. 70.96% at 20-shot), while in Split 2 we obtain the highest base mAP at 3-shot and 20-shot (71.63% and 71.89%). Similar trends can be observed in Splits 3 and 4, where our base performance closely tracks or slightly improves over G-FSDet [43] and RIFR [50]. These results indicate that freezing the vision-side LoRA adapters in the second stage, together with knowledge distillation and semantic consistency losses, effectively preserves the generic visual representations learned from base classes and mitigates catastrophic forgetting.

Second, our approach delivers consistent and often substantial gains on novel mAP across almost all splits and shot settings. On Split 1, our method achieves the best novel mAP under every K-shot condition, e.g., 34.94% vs. 28.51% (RIFR [50]) and 34.20% (TPG-FSOD [61]) at 3-shot, and 44.21% vs. 41.32% (RIFR [50]) and 43.80% (GIDR [44]) at 20-shot. Similar improvements are observed in Split 2, where our novel mAP clearly surpasses fine-tuning-based methods and prototype/meta-learning approaches as K increases (e.g., 27.24% vs. 24.10% for RIFR [50] and 24.40% for GIDR [44] at 20-shot). On Splits 3 and 4, our method also achieves consistently higher or highly competitive novel mAP across all K-shot settings. These gains suggest that the cross-modal Qwen3-VL backbone, when combined with hierarchical prompting, LoRA-based adaptation in the text encoder, and partial fine-tuning of the detection head in the second stage, can effectively encode structured semantic priors and better exploit limited novel-class examples, thereby improving category discrimination in few-shot regimes.

Finally, the overall mAP (mAP_{all}) of our method is consistently competitive and, in most cases, superior to all baselines, especially in the medium- and high-shot settings. For instance, in Split 1, our method achieves the highest overall mAP for 10- and 20-shot (63.85% and 64.41%), and in Split 2 it attains the best overall mAP across all shot numbers (e.g., 60.73% at 20-shot vs. 59.80% for RIFR[50]). In Splits 3 and 4, our overall mAP is comparable to RIFR [50] while consistently outperforming prototype/meta-learning methods, and in Split 4 it further surpasses RIFR [50] across all K-shot configurations. This reflects a favorable trade-off between base and novel performance: the model not only improves recognition of novel categories, but also maintains strong base-class accuracy. The robustness of these gains across different class splits further indicates that the proposed two-stage LoRA fine-tuning strategy, guided by hierarchical prompts and distillation/consistency constraints, provides a stable and effective solution for FSOD in complex RS scenes. To further provide an

 doi:10.20944/preprints202512.1915.v1

intuitive understanding of the detection quality under the few-shot setting, Figure 4 visualizes representative 10-shot detection results on DIOR Split 1, where the first row shows base-class predictions and the second row shows novel-class predictions.

Table 3. Comparison of FSOD performance (mAP@0.5) on the DIOR dataset under different K-shot settings. We report Base mAP, Novel mAP, and Overall mAP for four splits. The best and second-best results in each column are highlighted in red and blue, respectively.

Split	Method	3-shot			5-shot			10-shot			20-shot		
		Base	Novel	All	Base	Novel	All	Base	Novel	All	Base	Novel	All
1	Meta-RCNN [71]	60.62	12.02	48.47	62.01	13.09	49.78	61.55	14.07	49.68	63.21	14.45	51.02
	P-CNN [2]	47.00	18.00	39.80	48.40	22.80	42.00	50.90	27.60	45.10	52.20	29.60	46.80
	G-FSDet [43]	68.94	27.57	58.61	69.52	30.42	59.72	69.03	37.46	61.16	69.80	39.83	62.31
	TPG-FSOD [61]	62.70	34.20	55.58	62.70	35.60	55.93	64.40	40.20	58.35	65.50	43.00	59.88
	RIFR [50]	70.22	28.51	59.79	70.81	32.34	61.19	69.72	37.74	61.73	70.96	41.32	63.55
	GIDR [44]	-	33.80	-	-	35.90	-	-	40.10	-	-	43.80	-
	Ours	69.73	34.94	61.03	69.31	36.76	61.17	71.03	42.31	63.85	71.14	44.21	64.41
2	Meta-RCNN [71]	62.55	8.84	49.12	63.14	10.88	50.07	63.28	14.90	51.18	63.86	16.71	52.07
	P-CNN [2]	48.90	14.50	40.30	49.10	14.90	40.60	52.50	18.90	44.10	51.60	22.80	44.4
	G-FSDet [43]	69.20	14.13	55.43	69.25	15.84	55.87	68.71	20.70	56.70	68.18	22.69	56.86
	TPG-FSOD [61]	61.30	14.30	49.55	61.80	19.00	51.1	63.00	24.80	53.45	63.30	25.60	53.88
	RIFR [50]	70.83	15.11	56.90	71.11	18.75	58.02	70.77	21.93	58.56	71.70	24.10	59.80
	GIDR [44]	-	16.20	-	-	18.70	-	-	22.10	-	-	24.40	-
	Ours	71.63	19.31	58.55	70.05	21.55	57.93	70.14	25.68	59.03	71.89	27.24	60.73
3	Meta-RCNN [71]	61.93	9.10	48.72	63.44	12.29	50.66	62.57	11.96	49.92	65.53	16.14	53.18
	P-CNN [2]	49.50	16.50	41.30	49.90	18.80	42.10	52.10	23.30	44.90	53.10	28.80	47.00
	G-FSDet [43]	71.10	16.03	57.34	70.18	23.25	58.43	71.08	26.24	59.87	71.26	32.05	61.46
	TPG-FSOD [61]	65.60	20.10	54.23	65.10	23.10	54.60	66.40	28.90	57.03	65.90	32.60	57.58
	RIFR [50]	72.16	17.78	58.57	73.91	24.15	61.47	70.15	26.46	59.23	72.13	29.76	61.54
	GIDR [44]	-	19.80	-	-	25.40	-	-	28.70	-	-	32.40	-
	Ours	70.06	21.31	57.87	70.26	25.68	59.12	69.28	28.15	59.00	69.76	34.69	60.99
4	Meta-RCNN [71]	61.73	13.94	49.78	62.60	15.84	50.91	62.23	15.07	50.44	63.24	18.17	51.98
	P-CNN [2]	49.80	15.20	41.20	49.90	17.50	41.80	51.70	18.90	43.50	52.30	25.70	45.70
	G-FSDet [43]	69.01	16.74	55.95	67.96	21.03	56.30	68.55	25.84	57.87	67.73	31.78	58.75
	TPG-FSOD [61]	60.10	13.10	48.35	62.30	22.00	52.23	60.70	26.90	52.25	60.60	31.00	53.20
	RIFR [50]	69.10	18.22	56.38	69.05	22.87	57.51	69.22	28.49	59.03	68.78	32.12	59.62
	GIDR [44]	-	21.90	-	-	25.10	-	-	32.20	-	-	37.70	-
	Ours	69.95	21.68	57.88	70.79	23.77	59.04	70.58	30.62	60.59	70.24	35.21	61.48

 doi:10.20944/preprints202512.1915.v1

Figure 4. Visualization of detection results in the 10-shot experiments on the DIOR dataset under Class Split 1. The first row presents examples from the base classes, while the second row presents examples from the novel classes.

4.3.2. Results on NWPU VHR-10.v2

Table 4 reports the few-shot detection results on NWPU VHR-10.v2 under 3/5/10/20-shot settings and two splits. Overall, the trends further validate the effectiveness and generalization ability of our Qwen3-VL-based two-stage fine-tuning framework.

On Split 1, our method consistently achieves the best performance across all shot settings. For base classes, we obtain the highest base mAP in every K -shot configuration (e.g., 96.67%/96.54%/96.61% at 3/10/20-shot), clearly surpassing fine-tuning-based methods such as G-FSDet [43] and RIFR [50]. At the same time, we substantially improve novel mAP over these baselines (e.g., 67.51% vs. 53.17% for RIFR [50] at 3-shot, and 86.74% vs. 79.46% at 20-shot), while also outperforming prototype/meta-learning approaches such as GIDR [44] under all shot settings. As a result, our method achieves the best overall mAP across all K-shot settings, with particularly large margins in the high-shot regime (e.g., 93.65% vs. 90.61% for RIFR [50] and 85.00% for TPG-FSOD [61] at 20-shot). These results indicate that the proposed two-stage LoRA fine-tuning, combined with knowledge distillation and semantic consistency regularization, can effectively preserve base-class knowledge while significantly enhancing novel-class detection.

On Split 2, our approach shows similarly strong performance and stable behavior across different K-shot settings. At low-shot (3-shot), we already achieve the best novel mAP (51.48%) and overall mAP (79.58%), while maintaining a higher base mAP than P-CNN [2] and comparable or better base mAP than G-FSDet [43] and TPG-FSOD [61]. As K increases, the advantages of our framework become more pronounced: at 5/10/20-shot, we obtain the highest base mAP (91.51%, 91.27%, and 91.02%), the highest novel mAP (59.68%, 70.19%, and 78.47%), and the highest overall mAP (81.96%, 84.95%, and 87.26%) among all compared methods. This demonstrates that, as more support examples are available, our cross-modal Qwen3-VL backbone—adapted via hierarchical prompts, LoRA-based updates in the text encoder, and partial fine-tuning of the detection head—can better exploit both visual and semantic cues to boost novel-class discrimination without sacrificing base-class performance.

Taken together, the NWPU VHR-10.v2 results complement our findings on DIOR and show that the proposed Qwen3-VL-based two-stage fine-tuning framework generalizes well across RS datasets. By jointly leveraging cross-modal feature learning, hierarchical prompting, parameter-efficient LoRA adaptation, and distillation/consistency constraints, our method achieves a favorable balance between base-class retention and novel-class generalization, leading to consistently superior overall mAP in diverse few-shot settings. To further provide an intuitive understanding of the detection quality under the few-shot setting, Figure 5 visualizes representative 10-shot detection results on the NWPU VHR-10.v2 dataset under Class Split 1, where the first row shows base-class predictions and the second row shows novel-class predictions.

 doi:10.20944/preprints202512.1915.v1

Table 4. Comparison of FSOD performance (mAP@0.5) on the NWPU VHR-10.v2 dataset under different K-shot settings. We report Base mAP, Novel mAP, and Overall mAP for two splits. The best and second-best results in each column are highlighted in red and blue, respectively. Note that RIFR [50] reports results only on Class Split 1.

Split	Method	3-shot			5-shot			10-shot			20-shot		
		Base	Novel	All	Base	Novel	All	Base	Novel	All	Base	Novel	All
1	Meta-RCNN [71]	87.00	20.51	67.05	85.74	21.77	66.55	87.01	26.98	69.00	87.29	28.24	69.57
	P-CNN [2]	82.84	41.80	70.53	82.89	49.17	72.79	83.05	63.29	78.11	83.59	66.83	78.55
	G-FSDet [43]	89.11	49.05	77.01	88.37	56.10	78.64	88.40	71.82	83.43	89.73	75.41	85.44
	TPG-FSOD [61]	89.40	56.10	79.41	89.80	64.70	82.27	89.20	75.60	85.12	88.90	75.90	85.00
	RIFR [50]	94.64	53.17	82.20	95.26	62.39	85.40	96.23	72.29	89.04	95.38	79.46	90.61
	GIDR [44]	-	67.30	-	-	73.00	-	-	78.10	-	-	86.00	-
	Ours	96.67	67.51	87.92	96.84	69.49	88.64	96.54	75.63	90.27	96.61	86.74	93.65
2	Meta-RCNN [71]	86.86	21.41	67.23	87.38	35.34	71.77	87.56	37.14	72.43	87.26	39.47	72.92
	P-CNN [2]	81.03	39.32	68.52	81.18	46.10	70.70	80.93	55.90	73.41	81.21	58.37	75.50
	G-FSDet [43]	89.99	50.09	78.02	90.52	58.75	80.99	89.23	67.00	82.56	90.61	75.86	86.13
	TPG-FSOD [61]	90.10	48.70	77.68	89.80	63.50	81.91	89.30	69.40	83.33	90.20	75.90	85.91
	RIFR [50]	-	-	-	-	-	-	-	-	-	-	-	-
	GIDR [44]	-	50.90	-	-	58.50	-	-	68.30	-	-	76.40	-
	Ours	91.62	51.48	79.58	91.51	59.68	81.96	91.27	70.19	84.95	91.02	78.47	87.26

Figure 5. Visualization of detection results in the 10-shot experiments on the NWPU VHR-10.v2 dataset under Class Split 1. The first row presents examples from the base classes, while the second row presents examples from the novel classes.

4.4. Ablation Studies

We conduct ablation studies to analyze the contributions of the key components in our framework, including hierarchical prompting, the two-stage fine-tuning strategy, and the choice of LoRA rank. All ablations are performed on the **DIOR dataset** using **Class Split 1** under the standard few-shot setting.

4.4.1. Effect of Hierarchical Prompting

To verify the effectiveness of semantic injection via prompts, we compare three variants:

No Prompt: the text encoder is not conditioned on any instruction; only class labels are used as identifiers.

 doi:10.20944/preprints202512.1915.v1

Simple Prompt: we use a single generic instruction, e.g., "Detect objects in this image and output bounding boxes and category labels," without global context or fine-grained descriptions.

Hierarchical Prompt (ours): the full three-level design described in Section 3.3, including task instruction, global remote sensing context, and fine-grained category semantics.

Table 5 reports the ablation results of hierarchical prompting under Class Split 1 on DIOR. As shown in Figure 6, three clear trends can be observed. First, introducing any form of prompt brings substantial gains over the No Prompt setting. Compared with using only class labels, the Simple Prompt already improves both base and novel performance across all K-shot settings. For instance, at 3-shot, Simple Prompt boosts Base mAP from 58.71% to 65.35% (+6.64) and Novel mAP from 11.62% to 23.51% (+11.89), leading to an Overall mAP increase from 46.94% to 54.89% (+7.95). Similar improvements are observed at 5/10/20-shot, indicating that conditioning the text encoder on even a generic detection instruction can effectively encode task-aware semantics and enhance vision-language alignment.

Second, the proposed Hierarchical Prompt further provides consistent and significant improvements over the Simple Prompt in all metrics. At 3-shot, our hierarchical design raises Base mAP from 65.35% to 69.73% (+4.38), Novel mAP from 23.51% to 34.94% (+11.43), and Overall mAP from 54.89% to 61.03% (+6.14). The gains become even more pronounced as K increases: at 20-shot, Hierarchical Prompt achieves 71.14% / 44.21% / 64.41% on base / novel / overall mAP, outperforming Simple Prompt (65.15% / 31.78% / 56.81%) by +5.99 / +12.43 / +7.60, respectively. Notably, the improvements on novel classes are consistently larger than those on base classes (e.g., +23–26 points over No Prompt in novel mAP at 10/20-shot), showing that structured semantic priors are particularly important when only a few examples are available.

Finally, comparing the two extremes, the Hierarchical Prompt clearly delivers the best results under all K-shot settings, while the No Prompt variant performs worst. From No Prompt to Hierarchical Prompt, Overall mAP improves by about 14–16 points across different K-shot settings (e.g., from 48.40% to 63.85% at 10-shot, and from 48.82% to 64.41% at 20-shot), alongside more than doubling the novel mAP (e.g., 16.39% → 42.31% at 10-shot, 18.47% → 44.21% at 20-shot). These observations confirm that removing prompts severely limits the text encoder's ability to provide meaningful semantic guidance, and that full three-level hierarchical prompting—combining task instruction, global remote-sensing context, and fine-grained category semantics—most effectively injects semantic knowledge into the cross-modal Qwen3-VL backbone, thereby enhancing detection performance on both base and novel classes in the few-shot regime.

Table 5. Effect of hierarchical prompting on FSOD performance (mAP@0.5) under Class Split 1 on DIOR across different K-shot settings. The best results are highlighted in bold.

Split	Prompt	3-shot			5-shot			10-shot			20-shot		
		Base	Novel	All	Base	Novel	All	Base	Novel	All	Base	Novel	All
1	No	58.71	11.62	46.94	58.53	13.25	47.21	59.07	16.39	48.40	58.93	18.47	48.82
	Simple	65.35	23.51	54.89	65.59	25.19	55.49	65.74	29.27	56.62	65.15	31.78	56.81
	Hierarchical	**69.73**	**34.94**	**61.03**	**69.31**	**36.76**	**61.17**	**71.03**	**42.31**	**63.85**	**71.14**	**44.21**	**64.41**

 doi:10.20944/preprints202512.1915.v1

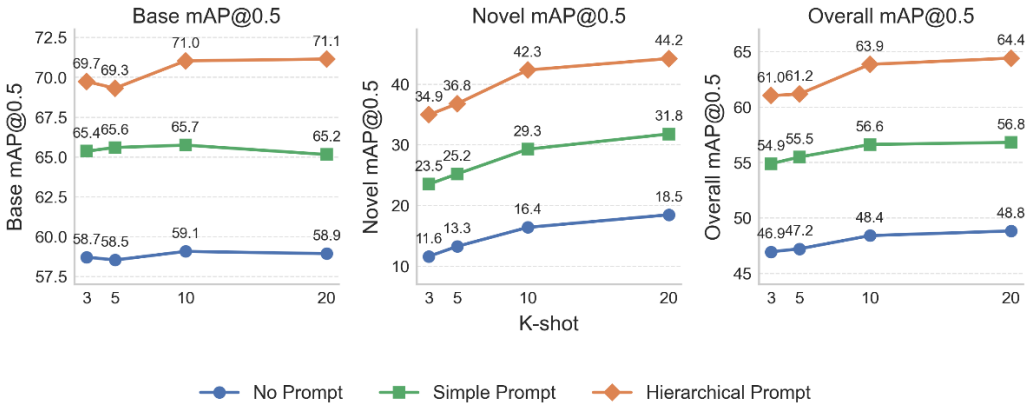

Figure 6. Effect of different prompting strategies on FSOD performance (mAP@0.5) under Class Split 1 on the DIOR dataset across varying K-shot settings. From left to right, we report Base, Novel, and Overall mAP, respectively. Each curve corresponds to a different prompting strategy (No Prompt, Simple Prompt, and the proposed Hierarchical Prompt). Hierarchical Prompt consistently achieves the best performance, especially on novel classes, and maintains clear gains as the number of shots increases.

4.4.2. Effect of Two-Stage Fine-Tuning Strategies

We next study the impact of the two-stage fine-tuning scheme. We compare:

Joint Training: base and novel classes are mixed and trained in a single stage using all available data, with all LoRA modules updated jointly.

Two-Stage (ours): the proposed Stage-1 base training followed by Stage-2 few-shot adaptation, with selective freezing of visual LoRA modules, partial fine-tuning of the detection head, and additional knowledge distillation and semantic consistency losses.

Table 6 reports the ablation results of different fine-tuning strategies under Class Split 1 on DIOR. Overall, the proposed two-stage scheme brings consistent gains over Joint Training across all K-shot settings.

First, the two-stage strategy yields substantial improvements on novel classes while maintaining or slightly enhancing base performance. In low-shot regimes (3/5-shot), the two-stage training improves Novel mAP from 25.17% to 34.94% (+9.77) and from 26.58% to 36.76% (+10.18), respectively, while Base mAP remains very similar or slightly better (69.17% → 69.73% and 68.82% → 69.31%). As K increases, the novel-class gains remain clear: at 10/20-shot, Novel mAP is boosted from 34.89% / 39.17% (Joint) to 42.31% / 44.21% (two-stage), with improvements of +7.42 and +5.04. Meanwhile, Base mAP also benefits more in the higher-shot settings (69.57% → 71.03% at 10-shot and 69.35% → 71.14% at 20-shot). These results indicate that Stage-1 base training followed by Stage-2 few-shot adaptation, together with selectively freezing visual LoRA modules and partially fine-tuning the detection head, effectively protects and even refines the base-class representations while allowing the model to better absorb the sparse novel-class data.

Second, the overall mAP (mAP_{all}) of two-stage consistently surpasses that of Joint Training for all K-shot configurations. The gains are relatively stable, around +2.6 to +3.0 points (e.g., 58.17% → 61.03% at 3-shot, 60.90% → 63.85% at 10-shot, and 61.81% → 64.41% at 20-shot). This reflects a more favorable balance between base and novel performance: Joint Training, which updates all modules on mixed base and novel data in a single stage, tends to underfit novel classes (especially at low shot) and risks mild forgetting on base classes as training proceeds. In contrast, the two-stage scheme leverages knowledge distillation and semantic consistency losses in Stage 2 to regularize the adaptation process, enabling the cross-modal Qwen3-VL backbone to integrate novel-class semantics without sacrificing, and in some cases improving, base-class detection. Overall, these findings validate the effectiveness of the proposed two-stage fine-tuning strategy for FSOD in RS scenes.

 doi:10.20944/preprints202512.1915.v1

Table 6. Effect of two-stage fine-tuning strategies on FSOD performance (mAP@0.5) under Class Split 1 on DIOR for different K-shot settings. The best results are highlighted in bold.

Split	Fine-Tuning Strategies	3-shot			5-shot			10-shot			20-shot		
		Base	Novel	All	Base	Novel	All	Base	Novel	All	Base	Novel	All
1	Joint	69.17	25.17	58.17	68.82	26.58	58.26	69.57	34.89	60.90	69.35	39.17	61.81
	Two-Stage	**69.73**	**34.94**	**61.03**	**69.31**	**36.76**	**61.17**	**71.03**	**42.31**	**63.85**	**71.14**	**44.21**	**64.41**

4.4.3. Effect of LoRA Rank

Finally, we investigate the influence of the LoRA rank r on performance and parameter efficiency. We evaluate $r \in \{32,64,128,256\}$ while keeping all other hyperparameters fixed. For each rank, we measure the total number of trainable parameters (LoRA + detection head) and the corresponding base/novel mAP.

Table 7 summarizes the impact of the LoRA rank on few-shot detection performance under Class Split 1 on DIOR, and Figure 7 provides a visual illustration of these trends. Overall, increasing the rank from 32 to 128 consistently improves both base and novel mAP across all K-shot settings, whereas further increasing the rank to 256 brings marginal or no gains and sometimes slightly degrades performance.

At the low-rank setting ($r = 32$), the model shows clear capacity limitations: both base and novel mAP are noticeably lower than those of higher-rank configurations. For example, at 10-shot, Base / Novel / Overall mAP are 61.05% / 33.12% / 54.07%, which are significantly below the corresponding values at $r = 64$ and $r = 128$. Increasing the rank to $r = 64$ yields substantial improvements, especially on novel classes. At 3-shot, Novel mAP improves from 25.32% to 31.56% (+6.24), and Overall mAP rises from 52.49% to 56.02% (+3.53). Similar trends are observed at 5/10/20-shot, indicating that a moderate increase in LoRA capacity allows the model to better leverage hierarchical prompts and scarce novel-class examples.

Further increasing the rank to $r = 128$ brings consistent and often the best performance across all K-shot settings. For instance, at 20-shot, $r = 128$ achieves 71.14% Base mAP and 44.21% Novel mAP, leading to the highest Overall mAP of 64.41%, compared with 58.92% at $r = 64$ and 55.42% at $r = 32$. The gains from $r = 64$ to $r = 128$ are smaller than those from $r = 32$ to $r = 64$, but still clear (e.g., +4.17 Novel mAP and +5.61 Overall mAP at 10-shot), suggesting that $r = 128$ provides a good balance between representational capacity and effective adaptation.

When the rank is further increased to $r = 256$, performance tends to saturate and even slightly decline in terms of Overall mAP, despite the significant growth in trainable parameters. For example, at 3 / 5 / 20-shot, Overall mAP at $r = 256$ (60.62%, 60.82%, 62.87%) is slightly lower than at $r = 128$ (61.03%, 61.17%, 64.41%). Novel mAP also shows small fluctuations (e.g., 34.94% → 33.71% at 3-shot and 44.21% → 43.95% at 20-shot). These patterns indicate diminishing returns and a potential risk of overfitting when the LoRA subspace is overly large under the current few-shot data scale.

Taken together, these results show that the LoRA rank has a non-trivial effect on few-shot performance: too small a rank ($r = 32$) underfits, while an excessively large rank ($r = 256$) offers limited benefits relative to its parameter cost. Medium ranks, especially $r = 128$ (and to a lesser extent $r = 64$), strike a favorable trade-off between parameter efficiency and detection accuracy, and are thus preferable choices for the proposed Qwen3-VL-based few-shot detection framework.

Table 7. Effect of LoRA rank on FSOD performance (mAP@0.5) under Class Split 1 on DIOR across different K-shot settings. The best results are highlighted in bold.

Split	LoRA rank	3-shot			5-shot			10-shot			20-shot		
		Base	Novel	All	Base	Novel	All	Base	Novel	All	Base	Novel	All
1	32	61.54	25.32	52.49	61.29	28.25	53.03	61.05	33.12	54.07	61.74	36.46	55.42
	64	64.17	31.56	56.02	64.59	33.67	56.86	64.94	38.14	58.24	64.72	41.52	58.92
	128	**69.73**	**34.94**	**61.03**	**69.31**	**36.76**	**61.17**	**71.03**	42.31	**63.85**	**71.14**	**44.21**	**64.41**

 doi:10.20944/preprints202512.1915.v1

	256	69.59	33.71	60.62	69.27	35.46	60.82	69.86	**42.53**	63.03	69.17	43.95	62.87

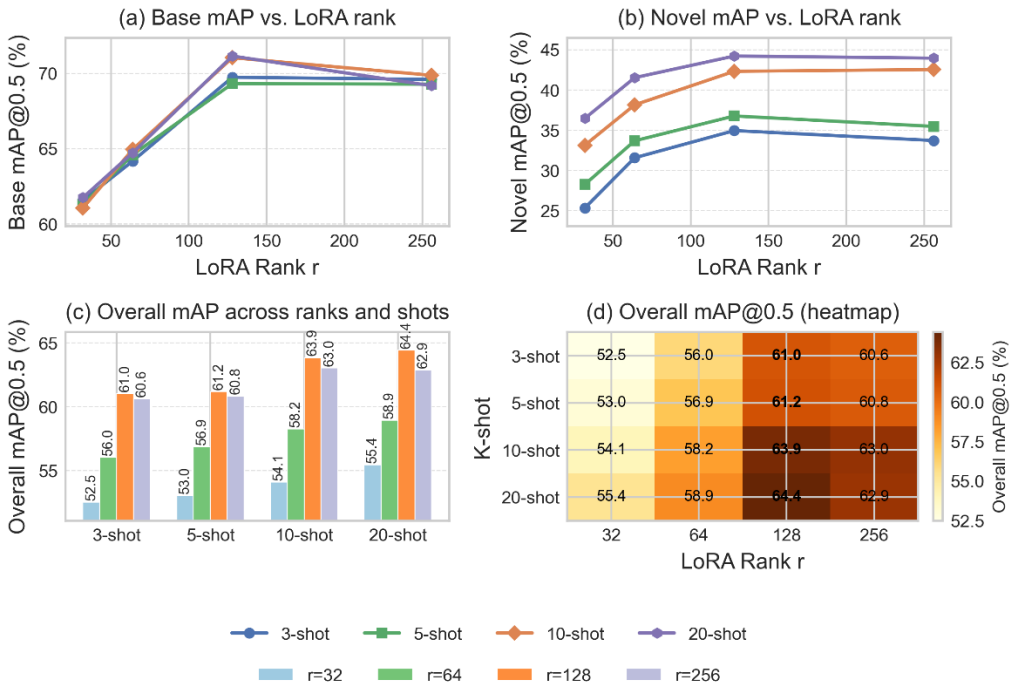

Figure 7. Effect of the LoRA rank on FSOD performance (mAP@0.5) under Class Split 1 on DIOR across different K-shot settings. (a) Base mAP@0.5 vs. LoRA rank. (b) Novel mAP@0.5 vs. LoRA rank. (c) Overall mAP@0.5 ("All") across different combinations of LoRA rank and K-shot. (d) Heatmap of overall mAP@0.5, where rows and columns denote the K-shot setting and LoRA rank r, respectively.

5. Discussion

5.1. Mechanism Analysis

From a cross-modal knowledge transfer perspective, LVLMs such as Qwen3-VL provide a strong semantic prior for FSOD. Pretrained on massive image-text pairs, they encode rich knowledge about object appearance, context, and attributes that cannot be learned from a handful of annotated RS images. In our framework, this prior is injected through the text encoder and prompts, so the detector starts from a semantically informed initialization rather than learning categories purely from sparse visual examples.

The proposed two-stage LoRA-based fine-tuning further structures this transfer. Stage 1 adapts both vision- and text-side LoRA modules to the RS domain and base-class taxonomy, anchoring cross-modal features to aerial viewpoints and cluttered backgrounds. Stage 2 then freezes the visual LoRA modules and updates the text-side LoRA together with partially fine-tuning the detection head under knowledge distillation and semantic consistency losses. The ablation on two-stage strategies shows that this design consistently improves novel mAP while maintaining or slightly boosting base mAP compared with joint training, indicating that selective updates and distillation are effective in preserving base knowledge and focusing adaptation on new categories.

Hierarchical prompting provides an additional, empirically validated mechanism for shaping decision boundaries. Compared with no prompt or a single generic instruction, the three-level prompts—task-level instruction, global remote sensing context, and fine-grained category descriptions—yield large gains in novel and overall mAP. This multi-granularity guidance structures the label space and encourages the model to separate visually similar categories (e.g., storage tank vs. stadium, basketball court vs. tennis court, baseball field vs. ground track field) by exploiting semantic differences, which is particularly beneficial under few-shot supervision.

 doi:10.20944/preprints202512.1915.v1

The LoRA rank ablation further clarifies the capacity-efficiency trade-off. Increasing the rank from 32 to 64 and 128 steadily improves both base and novel mAP, while a too-small rank underfits and a too-large rank (256) brings diminishing returns or slight degradation despite more parameters. This suggests that a medium rank (128) is sufficient to capture the task-relevant variations when combined with hierarchical prompts and the two-stage scheme.

5.2. Performance on Specific Classes

On both DIOR (20 classes) and NWPU VHR-10.v2 (10 classes), class-wise results are consistent with our mechanism analysis. Categories with distinctive shapes and stable contexts—such as storage tank, tennis court, baseball/basketball court, ground track field, airport, and harbor—benefit the most from cross-modal priors and hierarchical prompts. For example, storage tanks and harbors often appear in industrial or port areas, while tennis and basketball courts, baseball diamonds, and stadiums have clear geometric layouts. The LVLM's pretrained knowledge about these man-made facilities, combined with prompts that highlight their shapes, functions, and typical surroundings, allows the model to quickly align sparse RS evidence with rich semantic prototypes, leading to notable AP gains over purely visual FSOD baselines.

By contrast, categories with extreme scale variation, dense clutter, or ambiguous appearance—such as small vehicles in urban traffic, ships in complex harbor backgrounds, and building-like structures (e.g., overpass, bridge, chimney, windmill) with varied viewpoints—show more modest improvements. In these cases, high-level semantics (e.g., "ship on water", "vehicle on road") are harder to exploit when the targets are very small or heavily occluded, and label noise or imprecise bounding boxes in existing datasets further limit the benefit of semantic guidance. This suggests that while cross-modal priors and hierarchical prompts substantially help many typical RS categories in DIOR and NWPU VHR-10.v2, the overall gains remain bounded by data quality, target visibility, and extreme appearance variations.

5.3. Limitations

Despite promising results, our approach has several limitations. First, the inference cost is non-trivial, especially for Qwen3-VL-8B on high-resolution RS images: large inputs and long prompts increase latency and memory usage, which may hinder deployment in real-time or resource-constrained scenarios. Second, efficient training currently assumes relatively strong hardware; the benefits of LVLM-based FSOD on much smaller models are less clear without careful tuning. Third, the design of hierarchical prompts still relies on human expertise, and the LVLM's semantic space is not perfectly aligned with RS taxonomies, so some categories remain challenging despite careful prompt engineering.

These limitations suggest several directions for future work. One promising direction is automated prompt optimization (e.g., prompt tuning, prefix tuning, or instruction tuning on RS corpora) to reduce manual effort and improve robustness across datasets. Another is to develop more lightweight LVLMs or hybrid architectures tailored to overhead imagery, combining strong semantics with compact visual backbones. Finally, knowledge distillation from LVLM-based detectors to purely visual, task-specific models is a natural next step to retain most of the semantic benefits while substantially reducing inference cost and improving deployability.

6. Conclusions

We presented an LVLM-based framework for FSOD in RS, built on the Qwen3-VL model. The framework combines a two-stage LoRA fine-tuning strategy with hierarchical prompting to better exploit cross-modal knowledge learned during large-scale vision-language pretraining. Stage 1 adapts the model to the RS domain and base classes; Stage 2 performs parameter-efficient few-shot adaptation on novel classes with selective LoRA updates, knowledge distillation, and semantic consistency constraints. Hierarchical prompts provide multi-granularity semantic guidance—task

 doi:10.20944/preprints202512.1915.v1

instruction, global remote sensing context, and fine-grained category descriptions—while an appropriate LoRA rank ensures a good balance between capacity and efficiency.

Extensive experiments on DIOR and NWPU VHR-10.v2 show that our method consistently improves novel-class mAP over state-of-the-art few-shot detectors while maintaining strong base-class performance and transferring well across datasets. Ablation studies on prompting, two-stage strategies, and LoRA rank support our design choices and clarify how each component contributes to the final performance. Future work will focus on more efficient distillation to lightweight visual detectors and on automatic prompt optimization to further reduce manual effort and enhance robustness in diverse RS scenarios.

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

Funding: This research was funded by the National Natural Science Foundation of China (NSFC) (No. 62206302).

Data Availability Statement: The original datasets generated and/or analyzed during the current study are publicly available at the following link: https://github.com/Shi-YQ/RS-FSOD-datasets.

Acknowledgments: We would like to express our sincere gratitude to the editors and the anonymous reviewers for their valuable time, careful evaluation, and constructive comments, which have greatly improved the quality of this manuscript.

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

Abbreviations

The following abbreviations are used in this manuscript:

RS	Remote sensing
FSOD	Few-shot object detection
RSOD	Remote sensing object detection
RS-FSOD	Remote sensing few-shot object detection
CNN	Convolutional neural network
LVLM	Large vision-language model
LoRA	Low-rank adaptation
DETR	Detection Transformer
ViTs	Vision Transformers
RSAM	Remote sensing attention module
MLPs	Multilayer perceptrons
GIoU	Generalized Intersection over Union
IoU	Intersection over Union
mAP	Mean Average Precision
PEFT	Parameter-Efficient Fine-Tuning

Appendix A. Hierarchical Prompting for the DIOR Dataset

[Task]

You are an expert in remote sensing object detection. Given an overhead image, detect all objects of interest and output their bounding boxes and category labels.

[Context]

The image is an 800×800-pixel overhead view acquired by satellite or aerial sensors under varying ground sampling distances (approximately 0.5–30 m). Objects can be extremely multi-scale, densely packed, and arbitrarily oriented, with frequent background clutter, shadows, and repetitive textures. Scenes cover airports and airfields; expressways and highway facilities (service areas, toll

 doi:10.20944/preprints202512.1915.v1

stations, overpasses, bridges); ports and shorelines; large industrial zones (storage tanks, chimneys, windmills, dams); and urban or suburban districts with sports venues (baseball fields, basketball/tennis courts, golf courses, stadiums). Backgrounds can be cluttered and visually similar, and discriminative cues often come from fine-grained shapes, markings, and spatial context.

[Categories]

Superclass: Aerial transportation (aircraft and associated facilities observed from overhead around airfields).

Fine-grained classes:

airplane: Fixed-wing aircraft with clear wings and fuselage; parked on aprons or taxiing near runways.

airport: Large facilities with runways, taxiways, aprons and terminals; long linear strips and rectilinear layouts.

Superclass: Maritime transportation (vessels and port infrastructure along coastlines and waterways).

Fine-grained classes:

ship: Elongated hulls on water; visible wakes at sea or moored alongside docks.

harbor: Quays, piers, berths and cranes at the water-land interface; dense maritime infrastructure.

Superclass: Road transportation and facilities (vehicles and highway-related structures).

Fine-grained classes:

vehicle: Small man-made transport targets on roads/parking lots (cars, buses, trucks); compact rectangles with short shadows.

expressway service area: Highway rest areas with large parking lots, gas pumps and service buildings; near ramps.

expressway toll station: Toll plazas spanning multiple lanes with booths and canopies; strong lane markings at entries/exits.

overpass: Elevated roadway segments crossing other roads/rails; ramps, pylons and cast shadows.

bridge: Elevated linear structures spanning water or obstacles; approach ramps and structural shadows.

Superclass: Rail transportation (stations and rail hubs with linear track patterns).

Fine-grained classes:

train station: Parallel platforms and tracks with long platform roofs; track convergence or rail yards nearby.

Superclass: Sports and recreation facilities (man-made grounds with distinctive geometric markings).

Fine-grained classes:

baseball field: Diamond-shaped infield with bases and pitcher's mound; green outfield, dirt infield contrast.

basketball court: Small rectangle with painted key and arc lines; vivid surfaces in schools/residential areas.

tennis court: Rectangular court with marked baselines and service boxes; often appears in paired clusters with fences.

golf course: Irregular fairways and greens with sand bunkers; manicured textures and curved cart paths.

ground track field: Oval running track (often reddish) enclosing a field; located in schools or sports parks.

stadium: Large oval/circular stands enclosing a pitch/field; symmetric seating rings in urban settings.

Superclass: Industrial and utility infrastructure (energy, storage and emission structures).

Fine-grained classes:

 doi:10.20944/preprints202512.1915.v1

storage tank: Circular/cylindrical tanks, often clustered; bright roofs in refineries and industrial zones.

chimney: Tall slender exhaust stacks with long linear shadows; adjacent to plants or power stations.

windmill: Wind turbines with tall tower and three blades; radial blade shadows; in fields/wind farms.

dam: Linear/curved barrier across rivers/reservoirs; water retained on one side; spillways visible.

Appendix B. Hierarchical Prompting for the NWPU VHR-10.v2 Dataset

[Task]

You are an expert in remote sensing object detection. Given an overhead image, detect all objects of interest and output their bounding boxes and category labels.

[Context]

The image is a 400×400-pixel overhead crop acquired by satellite or aerial sensors at very high spatial resolution. Objects are small, densely distributed, and appear with diverse orientations and fine textures. Scenes frequently include roads and parking areas, bridges over water or roads, port shorelines with docks and cranes, industrial storage areas, and campus or recreational complexes. Typical object families include airplanes; vehicles; ships and harbors; storage tanks; bridges; and athletic venues such as baseball diamonds, basketball/tennis courts, and ground track fields. Backgrounds can be cluttered and visually similar, so recognition relies on subtle geometric layouts, line markings, and local context.

[Categories]

Superclass: Aerial transportation (aircraft observed around runways and aprons).

Fine-grained classes:

airplane: Fixed-wing aircraft with clear wings and fuselage; parked on aprons or taxiing near runways.

Superclass: Maritime transportation (vessels and port infrastructure along coastlines and waterways).

Fine-grained classes:

ship: Elongated hulls on water; visible wakes at sea or moored alongside docks.

harbor: Quays, piers, berths and cranes at the water-land interface; dense maritime infra-structure.

Superclass: Road transportation (vehicles on roads and in parking areas).

Fine-grained classes:

vehicle: Small man-made transport targets on roads/parking lots (cars, buses, trucks); com-pact rectangles with short shadows.

Superclass: Civil infrastructure (elevated crossings over water or roads).

Fine-grained classes:

bridge: Linear elevated spans with approach ramps and shadows; connects transport corridors.

Superclass: Sports and recreation facilities (courts and tracks with distinctive geometric markings).

Fine-grained classes:

baseball diamond: Diamond-shaped infield with bases and pitcher's mound; green outfield, dirt infield contrast.

basketball court: Small rectangle with painted key and arc lines; vivid surfaces in schools/residential areas.

tennis court: Rectangular court with marked baselines and service boxes; often appears in paired clusters with fences.

ground track field: Oval running track (often reddish) enclosing a field; located in schools or sports parks.

 doi:10.20944/preprints202512.1915.v1

Superclass: Industrial and storage (tank structures in plants/refineries).

Fine-grained classes:

storage tank: Circular/cylindrical tanks, often clustered; bright roofs in refineries and industrial zones.

References

1. Sharifuzzaman, S.A.S.M.; Tanveer, J.; Chen, Y.; Chan, J.H.; Kim, H.S.; Kallu, K.D.; Ahmed, S. Bayes R-CNN: An Uncertainty-Aware Bayesian Approach to Object Detection in Remote Sensing Imagery for Enhanced Scene Interpretation. *Remote Sens.* **2024**, *16*, 2405, doi:10.3390/rs16132405.
2. Cheng, G.; Yan, B.; Shi, P.; Li, K.; Yao, X.; Guo, L.; Han, J. Prototype-CNN for Few-Shot Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–10, doi:10.1109/TGRS.2021.3078507.
3. Li, J.; Tian, Y.; Xu, Y.; Hu, X.; Zhang, Z.; Wang, H.; Xiao, Y. MM-RCNN: Toward Few-Shot Object Detection in Remote Sensing Images With Meta Memory. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14, doi:10.1109/TGRS.2022.3228612.
4. Song, H.; Xie, J.; Wang, Y.; Fu, L.; Zhou, Y.; Zhou, X. Optimized Data Distribution Learning for Enhancing Vision Transformer-Based Object Detection in Remote Sensing Images. *Photogramm. Rec.* **2025**, *40*, e70004, doi:10.1111/phor.70004.
5. Li, J.; Tian, P.; Song, R.; Xu, H.; Li, Y.; Du, Q. PCViT: A Pyramid Convolutional Vision Transformer Detector for Object Detection in Remote-Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15, doi:10.1109/TGRS.2024.3360456.
6. Yang, B.; Han, J.; Hou, X.; Zhou, D.; Liu, W.; Bi, F. FSDA-DETR: Few-Shot Domain-Adaptive Object Detection Transformer in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–16, doi:10.1109/TGRS.2025.3574245.
7. Chen, Y.; Liu, B.; Yuan, L. PR-Deformable DETR: DETR for Remote Sensing Object Detection. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5, doi:10.1109/LGRS.2024.3483217.
8. He, X.; Liang, K.; Zhang, W.; Li, F.; Jiang, Z.; Zuo, Z.; Tan, X. DETR-ORD: An Improved DETR Detector for Oriented Remote Sensing Object Detection with Feature Reconstruction and Dynamic Query. *Remote Sens.* **2024**, *16*, 3516, doi:10.3390/rs16183516.
9. Kong, Y.; Shang, X.; Jia, S. Drone-DETR: Efficient Small Object Detection for Remote Sensing Image Using Enhanced RT-DETR Model. *Sensors* **2024**, *24*, 5496, doi:10.3390/s24175496.
10. Wang, A.; Xu, Y.; Wang, H.; Wu, Z.; Wei, Z. Cde-Detr: A Real-Time End-to-End High-Resolution Remote Sensing Object Detection Method Based on Rt-Detr. In Proceedings of the 2024 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS 2024); IEEE: Athens, GREECE, 2024; pp. 8090–8094.
11. Xu, Y.; Pan, Y.; Wu, Z.; Wei, Z.; Zhan, T. Channel Self-Attention Based Multiscale Spatial-Frequency Domain Network for Oriented Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15, doi:10.1109/TGRS.2024.3500013.
12. Bai, P.; Xia, Y.; Feng, J. Composite Perception and Multiscale Fusion Network for Arbitrary-Oriented Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16, doi:10.1109/TGRS.2024.3486559.
13. Zhang, C.; Su, J.; Ju, Y.; Lam, K.-M.; Wang, Q. Efficient Inductive Vision Transformer for Oriented Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–20, doi:10.1109/TGRS.2023.3292418.
14. Xu, T.; Sun, X.; Diao, W.; Zhao, L.; Fu, K.; Wang, H. FADA: Feature Aligned Domain Adaptive Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16, doi:10.1109/TGRS.2022.3147224.
15. Zhang, J.; Zhang, X.; Liu, S.; Pan, B.; Shi, Z. FIE-Net: Foreground Instance Enhancement Network for Domain Adaptation Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–14, doi:10.1109/TGRS.2025.3557171.

16. Li, J.; Ji, Y.; Xu, H.; Cheng, K.; Song, R.; Du, Q. UAT: Exploring Latent Uncertainty for Semi-Supervised Object Detection in Remote-Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–12, doi:10.1109/TGRS.2025.3586701.
17. Zhang, X.; Jiang, X.; Hu, Q.; Luo, H.; Zhong, S.; Tang, L.; Peng, J.; Fan, J. Enabling Near-Zero Cost Object Detection in Remote Sensing Imagery via Progressive Self-Training. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14, doi:10.1109/TGRS.2024.3415002.
18. Wang, X.; Huang, T.E.; Darrell, T.; Gonzalez, J.E.; Yu, F. Frustratingly Simple Few-Shot Object Detection. In Proceedings of the Proceedings of the 37th International Conference on Machine Learning; JMLR.org, July 13 2020; Vol. 119, pp. 9919–9928.
19. Qiao, L.; Zhao, Y.; Li, Z.; Qiu, X.; Wu, J.; Zhang, C. DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); October 2021; pp. 8661–8670.
20. Li, Y.; Hao, M.; Ma, J.; Temirbayev, A.; Li, Y.; Lu, S.; Shang, C.; Shen, Q. HPMF: Hypergraph-Guided Prototype Mining Framework for Few-Shot Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–13, doi:10.1109/TGRS.2025.3613849.
21. Wang, L.; Mei, S.; Wang, Y.; Lian, J.; Han, Z.; Feng, Y. CAMCFormer: Cross-Attention and Multicorrelation Aided Transformer for Few-Shot Object Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–16, doi:10.1109/TGRS.2025.3543583.
22. Wang, Y.; Zou, X.; Yan, L.; Zhong, S.; Zhou, J. SNIDA: Unlocking Few-Shot Object Detection with Non-Linear Semantic Decoupling Augmentation. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Seattle, WA, USA, June 16 2024; pp. 12544–12553.
23. Xie, J.; Wang, G.; 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*, 1–18, doi:10.1109/TGRS.2025.3564332.
24. Liu, Y.; Pan, Z.; Yang, J.; Zhou, P.; Zhang, B. Multi-Modal Prototypes for Few-Shot Object Detection in Remote Sensing Images. *Remote Sens.* **2024**, *16*, 4693, doi:10.3390/rs16244693.
25. Xu, Y.; Qin, J.; Zhan, T.; Wu, H.; Wei, Z.; Wu, Z. Few-Shot Object Detection in Remote Sensing Images via Dynamic Adversarial Contrastive-Driven Semantic-Visual Fusion. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–16, doi:10.1109/TGRS.2025.3602969.
26. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 Technical Report 2025. arXiv **2025**, arXiv:2505.09388, doi:10.48550/arXiv.2505.09388.
27. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models 2021. arXiv **2021**, arXiv:2106.09685, doi:10.48550/arXiv.2106.09685.
28. Liu, Y.; Li, Q.; Yuan, Y.; Du, Q.; Wang, Q. ABNet: Adaptive Balanced Network for Multiscale Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14, doi:10.1109/TGRS.2021.3133956.
29. Li, Z.; Wang, Y.; Zhang, Y.; Gao, Y.; Zhao, Z.; Feng, H.; Zhao, T. Context Feature Integration and Balanced Sampling Strategy for Small Weak Object Detection in Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5, doi:10.1109/LGRS.2024.3356507.
30. Li, J.; Tian, P.; Song, R.; Xu, H.; Li, Y.; Du, Q. PCViT: A Pyramid Convolutional Vision Transformer Detector for Object Detection in Remote-Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15, doi:10.1109/TGRS.2024.3360456.
31. Zhang, C.; Lam, K.-M.; Wang, Q. CoF-Net: A Progressive Coarse-to-Fine Framework for Object Detection in Remote-Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–17, doi:10.1109/TGRS.2022.3233881.
32. Nautiyal, R.; Deshmukh, M. RS-TOD: Tiny Object Detection Model in Remote Sensing Imagery. *Remote Sens. Appl.-Soc. Environ.* **2025**, *38*, 101582, doi:10.1016/j.rsase.2025.101582.
33. Zhao, X.; Yang, Z.; Zhao, H. DCS-YOLOv8: A Lightweight Context-Aware Network for Small Object Detection in UAV Remote Sensing Imagery. *Remote Sens.* **2025**, *17*, 2989, doi:10.3390/rs17172989.
34. Zhang, J.; Lei, J.; Xie, W.; Li, Y.; Yang, G.; Jia, X. Guided Hybrid Quantization for Object Detection in Remote Sensing Imagery via One-to-One Self-Teaching. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15, doi:10.1109/TGRS.2023.3293147.

35. Li, Y.; Fang, Y.; Zhou, S.; Long, T.; Zhang, Y.; Ribeiro, N.A.; Melgani, F. A Lightweight Normalization-Free Architecture for Object Detection in High-Spatial-Resolution Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 24491–24508, doi:10.1109/JSTARS.2025.3609658.
36. Tian, P.; Li, Z. YOLOv8-SP: Ship Object Detection in Optical Remote Sensing Imagery. *Mar. Geod.* **2025**, *48*, 688–709, doi:10.1080/01490419.2025.2493860.
37. Azeem, A.; Li, Z.; Siddique, A.; Zhang, Y.; Cao, D. Memory-Augmented Detection Transformer for Few-Shot Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–21, doi:10.1109/TGRS.2025.3551513.
38. Ma, J.; Bian, M.; Fan, F.; Kuang, H.; Liu, L.; Wang, Z.; Li, T.; Zhang, R. Vision-Language Guided Semantic Diffusion Sampling for Small Object Detection in Remote Sensing Imagery. *Remote Sens.* **2025**, *17*, 3203, doi:10.3390/rs17183203.
39. Li, X.; Deng, J.; Fang, Y. Few-Shot Object Detection on Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14, doi:10.1109/TGRS.2021.3051383.
40. Chen, J.; Qin, D.; Hou, D.; Zhang, J.; Deng, M.; Sun, G. Multiscale Object Contrastive Learning-Derived Few-Shot Object Detection in VHR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15, doi:10.1109/TGRS.2022.3229041.
41. Guo, M.; You, Y.; Liu, F. Discriminative Prototype Learning for Few-Shot Object Detection in Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–13, doi:10.1109/TGRS.2023.3326992.
42. Li, W.; Zhou, J.; Li, X.; Cao, Y.; Jin, G.; Zhang, X. InfRS: Incremental Few-Shot Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14, doi:10.1109/TGRS.2024.3475482.
43. Zhang, T.; Zhang, X.; Zhu, P.; Jia, X.; Tang, X.; Jiao, L. Generalized Few-Shot Object Detection in Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 353–364, doi:10.1016/j.isprsjprs.2022.12.004.
44. Yan, B.; Cheng, G.; Lang, C.; Huang, Z.; Han, J. Global-Integrated and Drift-Rectified Imprinting for Few-Shot Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–11, doi:10.1109/TGRS.2025.3546062.
45. Guirguis, K.; Meier, J.; Eskandar, G.; Kayser, M.; Yang, B.; Beyerer, J. NIFF: Alleviating Forgetting in Generalized Few-Shot Object Detection via Neural Instance Feature Forging. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Vancouver, BC, Canada, June 2023; pp. 24193–24202.
46. Chen, J.; Guo, Y.; Qin, D.; Zhu, J.; Gou, Z.; Sun, G. Multiscale Feature Knowledge Distillation and Implicit Object Discovery for Few-Shot Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15, doi:10.1109/TGRS.2024.3520715.
47. Lu, X.; Diao, W.; Li, J.; Zhang, Y.; Wang, P.; Sun, X.; Fu, K. Few-Shot Incremental Object Detection in Aerial Imagery via Dual-Frequency Prompt. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17, doi:10.1109/TGRS.2024.3400595.
48. Liu, N.; Xu, X.; Celik, T.; Gan, Z.; Li, H.-C. Transformation-Invariant Network for Few-Shot Object Detection in Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14, doi:10.1109/TGRS.2023.3332652.
49. Wu, J.; Liu, S.; Huang, D.; Wang, Y. Multi-Scale Positive Sample Refinement for Few-Shot Object Detection. arXiv **2020**, arXiv:2007.09384, doi: 10.48550/arXiv.2007.09384.
50. Zhu, Z.; Wang, P.; Diao, W.; Yang, J.; Kong, L.; Wang, H.; Sun, X. Balancing Attention to Base and Novel Categories for Few-Shot Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–19, doi:10.1109/TGRS.2024.3512865.
51. Wang, L.; Mei, S.; Wang, Y.; Lian, J.; Han, Z.; Chen, X. Few-Shot Object Detection With Multilevel Information Interaction for Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14, doi:10.1109/TGRS.2024.3410308.
52. Li, L.; Yao, X.; Wang, X.; Hong, D.; Cheng, G.; Han, J. Robust Few-Shot Aerial Image Object Detection via Unbiased Proposals Filtration. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–11, doi:10.1109/TGRS.2023.3300071.

53. Liu, Y.; Pan, Z.; Yang, J.; Zhang, B.; Zhou, G.; Hu, Y.; Ye, Q. Few-Shot Object Detection in Remote-Sensing Images via Label-Consistent Classifier and Gradual Regression. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14, doi:10.1109/TGRS.2024.3369666.
54. Jiang, H.; Wang, Q.; Feng, J.; Zhang, G.; Yin, J. Balanced Orthogonal Subspace Separation Detector for Few-Shot Object Detection in Aerial Imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17, doi:10.1109/TGRS.2024.3423305.
55. Zhang, F.; Shi, Y.; Xiong, Z.; Zhu, X.X. Few-Shot Object Detection in Remote Sensing: Lifting the Curse of Incompletely Annotated Novel Objects. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14, doi:10.1109/TGRS.2023.3347329.
56. Ma, S.; Hou, B.; Wu, Z.; Li, Z.; Guo, X.; Ren, B.; Jiao, L. Automatic Aug-Aware Contrastive Proposal Encoding for Few-Shot Object Detection of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–11, doi:10.1109/TGRS.2023.3294943.
57. Zhou, J.; Li, W.; Cao, Y.; Cai, H.; Huang, T.; Xia, G.-S.; Li, X. Few-Shot Oriented Object Detection in Remote Sensing Images via Memorable Contrastive Learning. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–14, doi:10.1109/TGRS.2025.3578632.
58. Azeem, A.; Li, Z.; Siddique, A.; Zhang, Y.; Li, Y. Prototype-Guided Multilayer Alignment Network for Few-Shot Object Detection in Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–23, doi:10.1109/TGRS.2025.3587389.
59. Fu, Y.; Wang, Y.; Pan, Y.; Huai, L.; Qiu, X.; Shangguan, Z.; Liu, T.; Fu, Y.; Van Gool, L.; Jiang, X. Cross-Domain Few-Shot Object Detection via Enhanced Open-Set Object Detector. In Proceedings of the Computer Vision – ECCV 2024; Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G., Eds.; Springer Nature Switzerland: Cham, 2025; pp. 247–264.
60. Gao, Y.; Lin, K.-Y.; Yan, J.; Wang, Y.; Zheng, W.-S. AsyFOD: An Asymmetric Adaptation Paradigm for Few-Shot Domain Adaptive Object Detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Vancouver, BC, Canada, June 2023; pp. 3261–3271.
61. Liu, T.; Zhou, S.; Li, W.; Zhang, Y.; Guan, J. Semantic Prototyping With CLIP for Few-Shot Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–14, doi:10.1109/TGRS.2025.3550372.
62. Sun, C.; Jia, Y.; Han, H.; Li, Q.; Wang, Q. A Semantic-Guided Framework for Few-Shot Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–13, doi:10.1109/TGRS.2025.3592484.
63. Zhang, T.; Zhuang, Y.; Wang, G.; Chen, H.; Wang, H.; Li, L.; Li, J. Controllable Generative Knowledge-Driven Few-Shot Object Detection From Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–19, doi:10.1109/TGRS.2025.3541937.
64. Zhang, T.; Zhuang, Y.; Zhang, X.; Wang, G.; Chen, H.; Bi, F. Advancing Controllable Diffusion Model for Few-Shot Object Detection in Optical Remote Sensing Imagery. In Proceedings of the 2024 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS 2024); IEEE: Athens, GREECE, 2024; pp. 7600–7603.
65. Zhang, R.; Yang, B.; Xu, L.; Huang, Y.; Xu, X.; Zhang, Q.; Jiang, Z.; Liu, Y. A Benchmark and Frequency Compression Method for Infrared Few-Shot Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–11, doi:10.1109/TGRS.2025.3540945.
66. Wang, B.; Ma, G.; Sui, H.; Zhang, Y.; Zhang, H.; Zhou, Y. Few-Shot Object Detection in Remote Sensing Imagery via Fuse Context Dependencies and Global Features. *Remote Sens.* **2023**, *15*, 3462, doi:10.3390/rs15143462.
67. Zhang, J.; Hong, Z.; Chen, X.; Li, Y. Few-Shot Object Detection for Remote Sensing Imagery Using Segmentation Assistance and Triplet Head. *Remote Sens.* **2024**, *16*, 3630, doi:10.3390/rs16193630.
68. Han, Z.; Gao, C.; Liu, J.; Zhang, J.; Zhang, S.Q. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey 2024. arXiv **2024**, arXiv:2403.14608, doi: 10.48550/arXiv.2403.14608.
69. 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, doi:10.1016/j.isprsjprs.2019.11.023.

 doi:10.20944/preprints202512.1915.v1

70. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-Class Geospatial Object Detection and Geographic Image Classification Based on Collection of Part Detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132, doi:10.1016/j.isprsjprs.2014.10.002.
71. Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; Lin, L. Meta R-CNN: Towards General Solver for Instance-Level Low-Shot Learning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); IEEE: Seoul, Korea (South), October 2019; pp. 9576–9585.