symmetry

MDPI

Article

DG-Net: Few-Shot Remote Sensing Detection with Dynamic Dual-Stream Collaboration and Generative Meta-Learning

Shanliang Liu †, Xinnan Shao *,†, Yan Dong *, Qihang He and Chunlei Li

School of Electronic and Information Engineering, Zhongyuan University of Technology, Zhengzhou 450007, China; shliucauc@163.com (S.L.)

* Correspondence: shao_work_paper@163.com (X.S.); dy@zut.edu.cn (Y.D.)

† These authors contributed equally to this work.

Abstract

Existing research has demonstrated that meta-learning methods hold considerable promise in addressing the challenges posed by few-shot object detection. However, remote sensing scenarios present two major challenges. The sparse features of small objects provide insufficient support information for query enhancement, and significant morphological variations caused by lighting and viewpoint differences hinder intra-class consistency capture via direct alignment in few-shot learning. To address these challenges, we propose a generative meta-learning detection framework. The framework first introduces a Dynamic Relation Dual-Stream Network to achieve dynamic support-query feature alignment through joint modeling of evolutionary and relational features, thereby enhancing representation in few-shot conditions. Second, an Optimal Transport-based Generative Meta-Learner is developed to mitigate feature distribution bias via generative augmentation in latent space. Additionally, an Orthogonal Frequency Decomposition Head is incorporated to adaptively separate query features into low-frequency contour and high-frequency detail components, effectively suppressing background noise interference. Experiments on multiple remote sensing datasets demonstrate that the proposed method achieves consistent performance gains over leading baseline methods in various few-shot settings. Its effectiveness is further validated across different backbone networks, highlighting strong generalization in few-shot remote sensing object detection.

Keywords: few-shot learning; remote sensing detection; meta-learning; feature enhancement

1. Introduction

In recent years, few-shot object detection (FSOD), as an important research direction in the field of computer vision, has been dedicated to addressing the challenges of object recognition and localization in scenarios with limited annotated data. Traditional object detection methods typically rely on large-scale annotated datasets for supervised training [1]. However, in specialized domains such as medical imaging and remote sensing, acquiring sufficient high-quality annotated data often involves high costs and lengthy cycles. To address this, researchers have proposed few-shot detection frameworks based on transfer learning [2] and meta-learning [3,4]. These approaches involve pre-training on base classes with abundant data, enabling the model to quickly adapt to novel classes with only a few annotated samples. In natural scene object detection tasks, few-shot learning methods have achieved significant success. Models based on meta-learning or metric learning facilitate the transfer of feature knowledge from base classes to novel classes, allowing effective

Check for updates

Academic Editors: Shuwei Zhu and Leilei Cao

Received: 11 February 2026
Revised: 4 March 2026
Accepted: 6 March 2026
Published: 7 March 2026

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

recognition of new objects with minimal annotation. Such methods typically rely on diverse and uniformly distributed training samples to enhance feature generalization capabilities. Therefore, they are particularly suited for natural image scenarios where object appearances vary but maintain structural consistency. These approaches have demonstrated strong detection performance and generalization capabilities in widely used public datasets such as COCO [5] and Pascal-VOC [6].

In the context of few-shot object detection in natural scenes, prior research has confirmed that meta-learning-based detection frameworks can partially alleviate the generalization challenges caused by scarce annotated data. However, when such methods are directly applied to satellite remote sensing image analysis, their performance is still constrained by the following two key factors. First, the performance of meta-learning mechanisms is constrained by the low resolution and sparse features of small objects in remote sensing imagery [7,8]. In satellite images, both artificial and natural targets often occupy minimal pixel areas, with their textures and structures simplified by imaging constraints. The limited and insufficiently representative support samples hinder the meta-learner from extracting discriminative prototype features or transformation rules. As a result, matching query samples frequently fails to achieve reliable feature enhancement or semantic alignment for sparse small targets, leading to notable degradation in few-shot detection and classification performance. Second, intra-class diversity in remote sensing caused by varying imaging conditions impedes feature alignment and knowledge transfer in meta-learning [9,10]. Factors such as illumination, season, atmosphere, and sensor parameters lead to substantial differences in color and morphology among same-category targets. The meta-learning paradigm typically assumes feature distribution consistency between support and query sets, yet such intra-class variation violates this assumption. Consequently, metric-based or prototype-alignment methods struggle to establish stable category representations, undermining generalization in few-shot scenarios and resulting in detection failures or misclassifications across different conditions.

In order to enhance detection accuracy for remote sensing targets under few-shot conditions, this paper introduces a Dynamic Relation Dual-stream Network (DRDS), which achieves dynamic semantic alignment between support and query sets by jointly modeling target evolutionary characteristics and relational features across samples. This mechanism substantially strengthens the model's ability to represent scarce query samples, addressing a key limitation of existing few-shot methods that rely on fixed feature matching or simple attention weighting, which often fail to adapt to intra-class feature variations caused by diverse imaging conditions in remote sensing scenarios. To further overcome feature distribution shifts and overfitting resulting from limited samples, we develop an Optimal Transport-based Generative Meta-Learner (OT-GML). This module conducts generative feature augmentation within a latent space structured by an optimal transport network, thereby expanding the representational capacity of the support set and enhancing cross-domain adaptability. Additionally, to mitigate interference from complex backgrounds and repetitive textures in high-resolution remote sensing imagery, an Orthogonal Frequency Decomposition Head (OFD) is proposed. This component adaptively decomposes query features into low-frequency structural components and high-frequency textural components, enabling effective foreground-background separation through frequency-domain decoupling. In contrast to previous approaches that primarily depend on spatial or channel attention mechanisms for feature recalibration, our framework integrates multi-level collaboration across feature evolution modeling, distribution-aligned augmentation, and frequency-aware decomposition. This integrated design systematically improves feature discriminability and generalization performance in few-shot remote sensing target detection.

To validate the effectiveness of the proposed generative meta-learning few-shot object detection algorithm, we conducted comprehensive training and evaluation on two satellite remote sensing datasets: DIOR and NWPU VHR-10. During the experimental phase, we partitioned the base classes and novel classes according to the few-shot task settings: in the DIOR dataset, 15 base classes and 5 novel classes were defined; in the NWPU VHR-10 dataset, 7 base classes and 3 novel classes were designated. Experiments were performed under 3-shot, 5-shot, and 10-shot configurations, respectively. After completing the dataset partitioning, we trained and tested the model based on the above few-shot tasks and compared it with mainstream few-shot detection methods such as FsDetView and Meta-RCNN. The experimental results show that on the DIOR dataset, our method achieved an improvement of 3% to 8% in average precision (AP50) over the second-best performing model in both 3-shot and 10-shot tasks. On the NWPU VHR-10 dataset, the model outperformed the FsDetView method by 1.73 and 2.45 percentage points under the 5-shot and 10-shot settings, respectively. Overall, compared to metric learning-based and optimization-based meta-learning methods, the proposed generative meta-learning framework demonstrates superior cross-task adaptability and generalization capability across different remote sensing datasets and backbone networks, while maintaining robust detection stability even with a limited number of few-shot support samples.

This study proposes a collaborative innovation framework to address three core challenges in few-shot object detection for remote sensing imagery: sparse features of small objects, intra-class distribution shifts, and complex background interference. Specifically, the framework achieves dynamic semantic alignment by modeling the evolutionary characteristics between samples through the Dynamic Relation Dual-stream Network (DRDS), overcoming the representation limitations imposed by sparse data. It performs structured feature augmentation in the latent space via the Optimal Transport-based Generative Meta-Learner (OT-GML), ensuring optimal alignment of generated samples from the distributional perspective. Furthermore, it extends feature processing to the frequency domain through the Orthogonal Frequency Decomposition Head (OFD), enabling effective foreground-background decoupling in the frequency dimension. This framework systematically addresses the bottlenecks of few-shot detection in remote sensing scenarios and provides a new paradigm for the application of meta-learning in non-natural image domains.

2. Related Works

Existing few-shot learning approaches are primarily categorized into transfer learning and meta-learning paradigms. Methods based on transfer learning aim to leverage knowledge acquired from sufficient data and transfer the learned knowledge by fine-tuning models with limited samples. Approaches based on meta-learning acquire "learning-to-learn" knowledge by designing and solving a series of few-shot object detection tasks, enabling models to adapt and generalize to novel data more effectively. Comparatively, both methods have their respective advantages and limitations. To integrate the strengths of both paradigms and overcome their inherent constraints, meta-transfer learning has emerged as a novel few-shot learning paradigm. It seeks to embed meta-learning mechanisms into the pre-training-fine-tuning framework of transfer learning, optimizing the model's initialization state or feature space during the meta-learning phase to achieve more efficient and robust knowledge transfer when fine-tuning with limited samples. Therefore, selecting an appropriate few-shot learning method for remote sensing object detection tasks is of particular importance.

2.1. Transfer Learning-Based Few-Shot Object Detection in Remote Sensing

Application of Transfer Learning in Few-Shot Object Detection first focuses on the core challenges at the data level. It addresses the issues of scarce samples and incomplete annotations through methods such as data augmentation and label calibration. To tackle the scarcity and low quality of remote sensing image data, Lv et al. [11] proposed a Deeply Supervised Generative Adversarial Network for remote sensing image data augmentation. This method can generate diverse, high-quality remote sensing images, reduces generation time, and provides data support for applications such as semantic interpretation of remote sensing images. To better extract target features from remote sensing images, Sun et al. [12] utilized intermediate convolutional layers of an Orthogonal Generative Adversarial Network combined with Selective Convolutional Descriptor Aggregation technology to more effectively obtain feature vectors of targets. Addressing the problem of incomplete or erroneous sample annotations, Qiao et al. [13] employed a confidence-based label calibration method to correct potential incorrect labels, thereby reducing the interference of mislabeling on model training. They also constructed a category center library to mitigate the negative impact of unlabeled targets. Similarly, Liu et al. [14] tackled the issue of unlabeled data by comparing the features of predicted target bounding boxes with a class-level feature library. They calculated the object detection score for each predicted target based on feature matching to determine whether a sample belongs to a known category. Details please check at Figure 1.

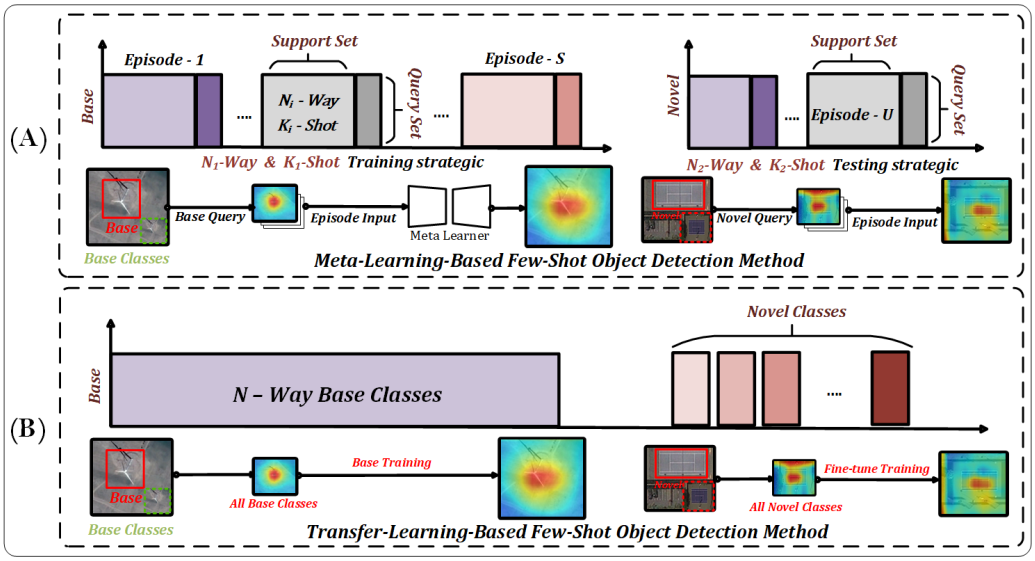

Figure 1. General workflow of meta-learning algorithms and transfer learning algorithms: (**A**) Meta-learning, (**B**) Transfer learning.

At the feature level, transfer learning methods enhance the discriminative power and representational capacity of target features while reducing inter-class confusion through techniques such as feature modulation, contrastive learning, and attention mechanisms. Su et al. [15] proposed a few-shot oriented object detection method, which employs a multi-orientation enhancement branch to generate multi-orientation and multi-scale positive samples, along with a context-aware module that enables the detector to effectively extract contextual information surrounding the target. To address the challenge of inaccurate reconstruction errors, Li et al. [16] introduced a feature map complementary reconstruction method, which effectively captures correlations across different feature map dimensions, providing a reliable basis for the model to learn more discriminative fine-grained features. To improve class distinguishability, Zhou et al. [17] designed a supervised contrastive learning module to tackle the problem of target misclassification. This module uses a

dynamically updated memory bank to store historical sample features and employs large-scale negative samples for contrastive training, thereby enhancing the model's ability to learn discriminative features for new classes and reducing inter-class confusion. Wu et al. [18] proposed an adaptive feature modulator, which utilizes features extracted from a small number of annotated samples to adjust the weights of target features, strengthening key features while suppressing redundant information to improve the class representational capacity of the features. Xiao et al. [19] conducted research on few-shot object detection and leveraged category-representative features extracted from different modalities of data to guide network predictions.

2.2. Meta-Learning-Based Few-Shot Object Detection in Remote Sensing

In the study of few-shot remote sensing object detection based on meta-learning, a series of methods focus on enhancing the model's discriminative capabilities by optimizing prototype generation and feature representation. Lin et al. [20] proposed a Noisy Query Prototype Network. This method utilizes a query mixing module to generate additional query samples with interference information for classification, thereby enhancing the model's feature learning ability. Building on this, Yang et al. [21] introduced a Discriminative Deep Nearest Neighbor Neural Network, which incorporates center loss and global pooling information to improve intra-class compactness and reduce background interference. To address the issue of ambiguous distribution in support samples, Guo et al. [22] proposed a Discriminative Prototype Learning Network to rectify the fuzzy distribution of support samples, thereby generating more representative category prototypes. Subsequently, Zhang et al. [23] tackled the asymmetry between the support set and the query set by proposing a targeted Dual-Forward Excitation Module. Zhao et al. [24] shortened the information transmission path through bottom-up information flow and leveraged the semantic information of low-level features to assist in target localization, thereby improving the detection capability for targets of varying scales. Meanwhile, Ma et al. [25] compared different optimal enhancement strategies within the network to adapt to the current dataset, selecting an effective data augmentation scheme for the model.

In order to further address deep-seated challenges such as feature misalignment, attention deviation, and inadequate utilization of semantic information in remote sensing scenarios, subsequent research has proposed more structured improvement strategies. Liu et al. [26] introduced a few-shot remote sensing object detector that employs a coarse-to-fine regression method combined with an enhanced feature extractor to improve the recall rate of the region proposal network. Wang et al. [27] enhanced the few-shot object detection method for remote sensing by integrating global features with contextual semantics. Li et al. [28] combined the IoU score between two bounding boxes with the foreground-background confidence score to filter redundant candidate regions. To leverage potential information from unlabeled samples, Zhang et al. [29] treated targets in base-class images that could potentially belong to novel classes but were unlabeled as "pseudo-labels," thereby utilizing information from unlabeled novel class objects to optimize the model. To address feature stability issues, Yang et al. [30] considered the significant variance fluctuations in feature vectors output by the support set and designed a stable feature extraction module to enhance the stability and category representational capacity of the feature vectors. Going a step further, Liu et al. [31] proposed a Transformation-Invariant Network (TINet), which ensures geometric invariance of features while explicitly aligning features between the query set and support set branches. Finally, Wang et al. [32] introduced a multi-level information interaction strategy to mitigate feature misalignment and attention deviation issues, while incorporating a semantic enhancement module to assist the strategy in extracting key semantic information.

2.3. Distributional and Generative Few-Shot Learning

Recent advances in few-shot learning have witnessed a growing interest in distributional and generative approaches to address data scarcity. In Generalized Zero-Shot Learning (GZSL), Wang et al. [33] proposed D^3GZSL, an end-to-end generative framework that mitigates the bias toward seen data by treating seen and synthesized unseen samples as in-distribution and out-of-distribution data, respectively, utilizing in-distribution dual space distillation (ID^2SD) and out-of-distribution batch distillation (O^2DBD) to enhance learning coherence and capture shared structures between categories. For few-shot image synthesis, Gupta et al. [34] introduced Conditional Distribution Modelling (CDM), a diffusion-based framework that models the latent distribution used to condition the diffusion process, leveraging training data statistics to better approximate unseen class distributions while employing an inversion-based optimization strategy to ensure sample fidelity. Abdollahzadeh et al. [35] provided a comprehensive survey on Generative Modeling under Data Constraint (GM-DC), examining limited-data, few-shot, and zero-shot settings while addressing challenges such as overfitting, frequency bias, and incompatible knowledge transfer through unified taxonomies of tasks and methodological approaches. Giannone [36] further advanced few-shot generative modeling with the Set-Context Hierarchical-Aggregation Variational Autoencoder (SCHA-VAE) for handling data scarcity and variability, alongside exploring generative applications in scientific and engineering domains including Text-ChemT5 for chemistry and natural language processing and DOM-TA leveraging diffusion models for engineering design. Collectively, these distributional and generative approaches demonstrate the expanding capabilities of deep generative models in data-constrained scenarios, providing robust solutions for few-shot recognition, synthesis, and cross-domain adaptation tasks.

3. Few-Shot Object Detection Algorithm Based on Generative Meta-Learner

This section presents the generative meta-learning algorithm for few-shot object detection in remote sensing proposed in this paper. It begins by outlining the overall architecture of the few-shot object detection model, followed by an explanation of its training strategy within the meta-learning framework and the corresponding loss function. The discussion then elaborates on two core components of the method: the generative meta-learner based on optimal transport, and the dynamic relational dual-stream collaborative network adapted for remote sensing targets. Finally, we introduce the orthogonal frequency-domain decomposition detection head, which is specifically designed to handle few-shot categories.

3.1. The Overall Architecture of the Network

This framework constructs a few-shot object detection system with multi-module collaboration, systematically addressing the core challenges in few-shot remote sensing object detection through three key components. The design philosophy of the framework stems from an in-depth analysis of the characteristics of remote sensing images and the difficulties inherent in few-shot learning. It fully accounts for the highly variable scales of targets, complex backgrounds, and significant intra-class variations in remote sensing imagery, as well as issues such as model overfitting and insufficient generalization capability caused by limited training data under few-shot conditions. The specific structure of the framework is illustrated in the Figure 2 below.

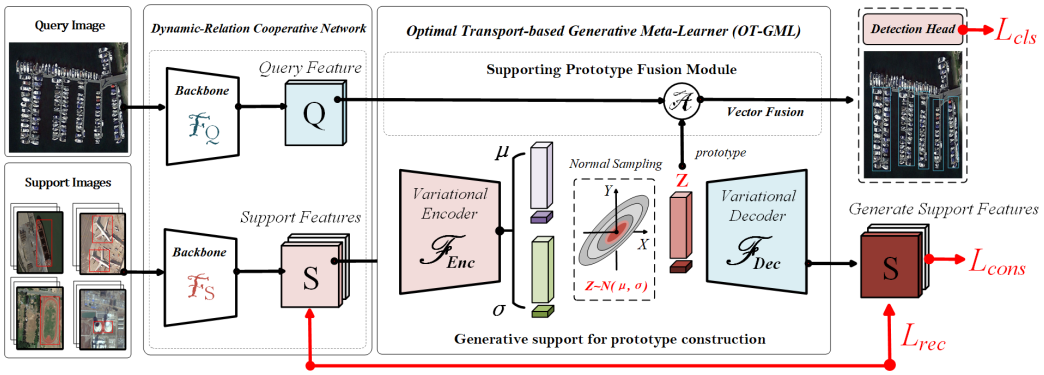

Figure 2. The overall architecture of a few-shot object detection framework based on generative meta-learning algorithms.

First, a dynamic relational dual-stream collaborative network is employed to extract feature information from query and support images and construct tasks. This network adopts a parallel dual-stream architecture that integrates spatial dependency modeling with global semantic reasoning: on one hand, it enhances spatial representation capabilities for multi-scale, irregular objects through improved non-causal Mamba units and 2D positional encoding; on the other hand, it combines a gated fusion mechanism to integrate the strengths of state-space models and self-attention, enabling long-range dependency and global context modeling while significantly reducing computational complexity. Further, generative feature enhancement and distributed prototype representations are introduced to effectively improve few-shot generalization performance, enabling the construction of discriminative feature spaces even under sample scarcity conditions.

Subsequently, an Optimal Transport-based Generative Meta-Learner (OT-GML) is utilized to enhance few-shot queries using support prototypes. The core concept of this approach is to represent categories as probability distributions rather than point estimates, measuring similarities between category distributions through optimal transport theory to provide more robust and accurate distance metrics for query samples. To accurately estimate category distributions under sample scarcity and avoid bias toward specific samples, OT-GML integrates generative models to distributionally augment the support set, offering richer and more reliable distribution estimates for optimal transport calculations.

Finally, an orthogonal frequency-domain decomposition detection head algorithm is applied to obtain more accurate category predictions. This algorithm incorporates lightweight frequency-domain processing at the end of the Feed-Forward Network (FFN), performing orthogonal frequency-domain analysis and filtering on output features only. This significantly enhances the discriminability of frequency-domain features while maintaining spatial computational efficiency. Compared to traditional methods that perform frequency-domain transformations at intermediate layers, this design greatly reduces redundant frequency-to-spatial conversion operations, effectively addressing issues such as frequency-domain feature degradation and sensitivity to noise under few-shot conditions. It enables robust frequency-aware feature reconstruction even with limited samples. In the following paragraphs, the specific implementation methods of these three components will be introduced.

3.2. Dynamic Relational Dual-Stream Collaborative Network for Remote Sensing

To address challenges such as high intra-class variance, sample scarcity, and complex spatial relationships in few-shot object detection for remote sensing images, this paper proposes a dynamic relational dual-stream collaborative network. This network integrates spatial dependency modeling with global semantic reasoning through a parallel dual-stream architecture. It employs improved non-causal Mamba units and 2D positional

encoding to enhance spatial representation capabilities for multi-scale, irregular objects, while incorporating a gated fusion mechanism to leverage the advantages of both state-space models and self-attention. This approach achieves long-range dependency and global context modeling while significantly reducing computational complexity. Through generative feature enhancement and distributed prototype representation, the network effectively improves few-shot generalization performance and constructs highly discriminative feature spaces even under conditions of sample scarcity. The structure of the dual-stream collaborative network is illustrated in the following Figure 3.

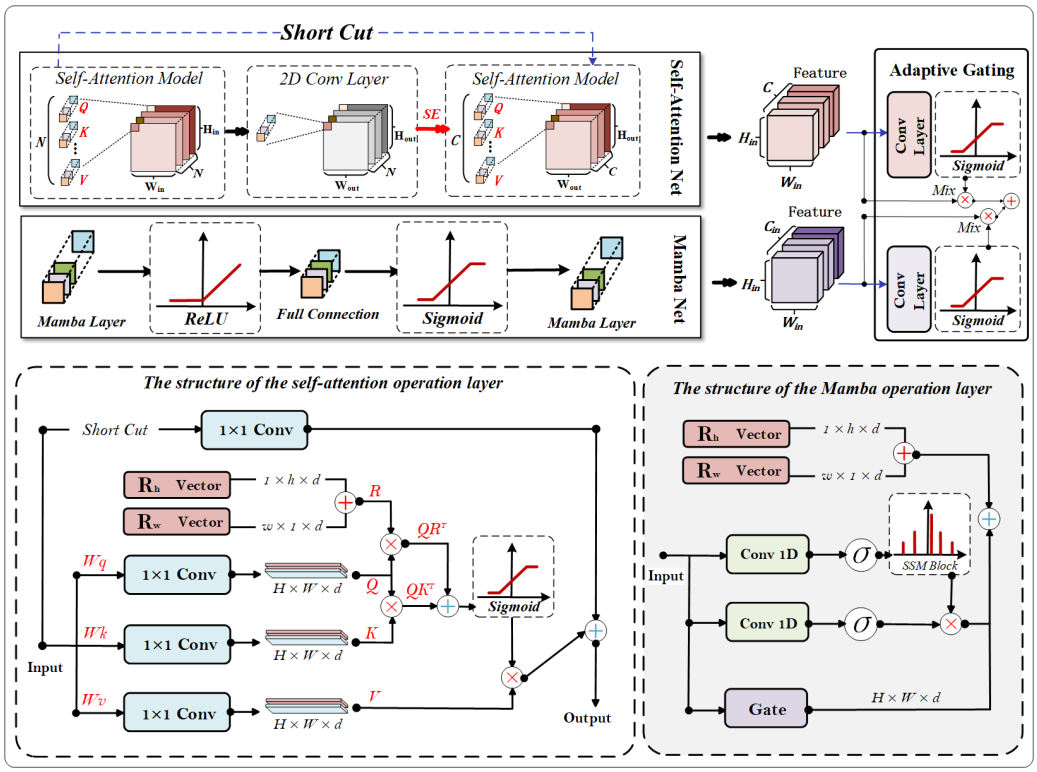

Figure 3. The overall structure of the Dynamic Relational Dual-Stream Collaborative Network.

This dynamic dual-stream collaborative network primarily consists of three components: a non-causal Mamba unit, 2D multi-head self-attention, and an adaptive gated fusion mechanism. First, the causal constraint of the original Mamba severely limits its performance in few-shot remote sensing tasks. The model enforces chain-like unidirectional dependencies between tokens, making it difficult to establish associations between distant regions in the image, leading to attenuation of long-range feature dependencies. Additionally, it fails to effectively capture the complex spatial layouts of ground objects, resulting in the loss of spatial structural information and reduced responsiveness to multi-scale targets. In few-shot remote sensing scenarios, these structural shortcomings are significantly amplified, severely constraining the model's representational capacity and generalization performance. To address this, our method improves the spatial perception mechanism of the Mamba unit by introducing a gating mechanism to adaptively modulate features, thereby enhancing the model's backbone perception capability for few-shot remote sensing targets.

Given an input feature tensor $X \in R^{H*W*C}$, it is first reshaped into a sequential representation $X_{seq} = Reshape(x) \in R^{B*(H*W)*C}$. To encode spatial structural priors, learnable relative position embeddings are introduced separately along the height and width dimensions $P = R_H + R_W$, $P \in R^{1*(H*W)*C}$, where R_H and R_W are trainable parameters capturing

positional biases in the vertical and horizontal directions respectively. The positional information is then injected into the sequential features $\hat{X}_{seq} = X_{seq} + P^T$.

Subsequently, a spatially context-aware gating mechanism is constructed through 1D convolution and Sigmoid activation function, which can adaptively regulate feature response intensity. Meanwhile, the sequence is processed by the non-causal Mamba block **M(·)** to capture long-range contextual dependencies. Finally, the output of Mamba is element-wise fused with the gating signal as described in the following formula, and the output sequence is reshaped back to the spatial dimension of in the end.

$$\hat{\mathbf{X}}_{mamba} = \mathbf{X}_{mamba} \odot G = \mathcal{M}[(x_{seq}^T] \odot \text{Sigmoid}[Conv(x_{seq}^T)] \tag{1}$$

where $\hat{\mathbf{X}}_{mamba}$ denotes the output feature after Mamba processing with gating modulation, $\mathbf{X}_{mamba}$ represents the output of the non-causal Mamba block $\mathcal{M}(\cdot)$, and G is the gating signal generated by a 1D convolution followed by Sigmoid activation. Specifically, x_{seq}^T indicates the transposed sequential input features, $Conv(\cdot)$ refers to the 1D convolution operation that captures local context along the sequence dimension, and Sigmoid$(\cdot)$ produces normalized gating weights between 0 and 1. The symbol $\odot$ denotes element-wise multiplication, which adaptively modulates the feature responses based on the learned spatial context. This mechanism enables the model to selectively emphasize informative regions while suppressing less relevant ones, thereby enhancing the representational capacity of the Mamba branch.

To address the limitations of the Mamba structure in global semantic modeling, a two-dimensional multi-head self-attention mechanism is introduced on top of this architecture, jointly forming a dual-stream collaborative structure with Mamba. By explicitly modeling the interactions among all spatial positions, the complete global context of features is captured. Given an input feature, it is transformed into queries, keys and values, and the attention weights are computed as follows:

$$\textbf{Attention}(\mathbf{Z}_{w_j}) = \text{Softmax}\left(\frac{\mathbf{Q}'_{w_j}(\mathbf{K}'_{w_j})^\top}{\sqrt{d}}\right)\mathbf{V}'_{w_j} \tag{2}$$

where Attention(Z_{w_j}) denotes the output of the multi-head self-attention mechanism for the j-th window or head, Z_{w_j} represents the input features within the j-th window. $\mathbf{Q}'_{w_j}$, $\mathbf{K}'_{w_j}$, and $\mathbf{V}'_{w_j}$ are the query, key, and value matrices derived from the input features after linear projections, respectively. The term $(\mathbf{K}'_{w_j})^\top$ is the transpose of the key matrix, and $\sqrt{d}$ is the scaling factor, where d denotes the dimension of the query/key vectors, used to prevent the dot products from growing too large in magnitude. The Softmax function is applied along the appropriate dimension to compute the attention weights, which are then used to aggregate the value vectors. This formulation enables the model to capture global contextual dependencies by explicitly modeling interactions among all spatial positions within each attention head.

To preserve spatial structural information of features in self-attention networks, a two-dimensional relative position encoding is further introduced into this architecture, enabling the attention mechanism to perceive relative geometric relationships between positions. The incorporation of the two-dimensional self-attention mechanism significantly enhances the model's ability to capture complex spatial arrangements and semantic dependencies of ground objects in remote sensing imagery.

To fully integrate the efficiency advantages of Mamba with the global modeling capability of self-attention in dynamic dual-stream networks, this method proposes a gated adaptive fusion mechanism. This dynamic gating mechanism splits the input features equally along the channel dimension, feeding them into the Mamba and self-attention

branches respectively, and subsequently fuses the outputs of the two branches via learnable gated fusion, as described in the following formula, where and represent learnable projection weights.

$$\mathbf{X}_{\text{fused}} = \mathbf{Gate}_m \cdot \mathbf{W}_m * \mathbf{X}_{\text{mamba}} + \mathbf{Gate}_h \cdot \mathbf{W}_h * \mathbf{X}_{\text{attention}} \tag{3}$$

$$\begin{cases} \mathbf{Gate}_m = \text{Sigmoid}[\mathbf{W}_m * \mathbf{X}_{\text{mamba}}] \\ \mathbf{Gate}_h = \text{Sigmoid}[\mathbf{W}_h * \mathbf{X}_{\text{attention}}] \end{cases} \tag{4}$$

where $Xfused$ denotes the fused output feature that integrates information from both the Mamba and self-attention branches. $\mathbf{X}mamba$ and $\mathbf{X}_{attention}$ represent the output features from the Mamba branch and the self-attention branch, respectively. $\mathbf{W}m$ and $\mathbf{W}h$ are learnable projection weight matrices applied to the Mamba and attention features, respectively, enabling feature transformation before fusion. The symbol $*$ denotes the convolution operation (or matrix multiplication, depending on implementation) that applies the projection weights to the features. $\mathbf{Gate}m$ and $\mathbf{Gate}h$ are dynamic gating signals generated by passing the projected features through a Sigmoid activation function, which produces values in the range $[0, 1]$. These gates adaptively control the contribution of each branch to the final fused output based on the input content. The operator $\cdot$ represents element-wise multiplication between the gating signals and the projected features. This gated adaptive fusion mechanism enables the model to dynamically balance local detail modeling from Mamba and global semantic modeling from self-attention, achieving collaborative feature enhancement while maintaining computational efficiency.

This strategy enables the model to dynamically adjust the contributions of the two branches based on the input content, thereby achieving collaborative modeling of local details and global semantics while maintaining low computational overhead. This adaptive dual-stream structure demonstrates excellent balance between performance and accuracy in remote sensing object detection tasks, particularly suitable for remote sensing target recognition scenarios under conditions of sample scarcity.

3.3. Optimal Transport-Based Generative Meta-Learner

Few-Shot Learning (FSL) faces the core challenge of learning a model with strong generalization capabilities from extremely limited samples. Among existing approaches, meta-learning algorithms, which employ a "learning-to-learn" strategy to extract meta-knowledge from a large number of related tasks, have become one of the mainstream paradigms for addressing few-shot problems. However, existing methods, particularly metric-based meta-learning approaches such as Prototypical Networks, typically compress all sample information of each support category into a single prototype vector, such as the class centroid. In application scenarios like few-shot remote sensing images, which exhibit high intra-class variance and complex data distributions, this approach can lead to prototype estimation bias, thereby reducing the model's detection accuracy.

To overcome this limitation, this chapter proposes an Optimal Transport-based Generative Meta-Learner (OT-GML). The core concept of this meta-learner is to treat each category as a probability distribution rather than a point estimate. The chapter introduces Optimal Transport (OT) theory to measure the similarity between category distributions, thereby providing more robust and precise distance metrics for query samples. Additionally, to better transform support features into category distributions under conditions of sample scarcity and ensure that the estimated distribution is not biased toward specific samples, this meta-learner integrates generative models. By performing generative modeling and enhancement on the support set distribution, it provides richer and more reliable

distribution estimates for optimal transport calculations. The framework structure of the OT-GML module is illustrated in the following Figure 4:

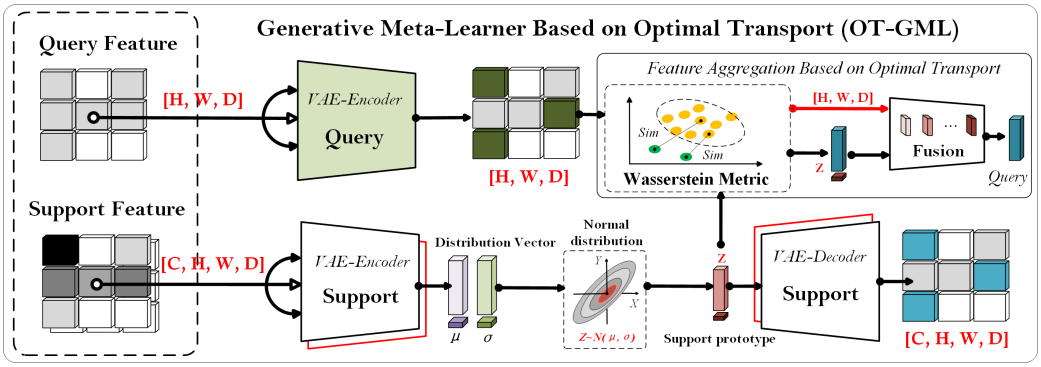

Figure 4. The overall structure of the Optimal Transport-based Generative Meta-Learner.

The overall framework of the OT-GML module follows a meta-learning training mechanism, addressing few-shot object detection in optical remote sensing images through distribution modeling and optimal transport. The framework consists of three core components: first, the generative enhancement module forms synthetic features for specified categories based on a generative network to augment the support set; subsequently, by merging the original support features with the generated features, each category is represented as a probability distribution, explicitly modeling intra-class variance; finally, the optimal transport classifier measures the Wasserstein distance between the feature distribution of query samples and each category distribution. This distance effectively captures geometric relationships between complex distributions, thereby enhancing classification robustness and accuracy under few-shot conditions in the subsequent structure.

The generative enhancement module in this method is implemented based on the Variational Autoencoder (VAE) framework but introduces significant improvements in objective formulation and implementation. The core idea of this module is to probabilistically map given support features to a structured distribution in the latent space. Through sampling operations, semantically rich variational features are obtained and subsequently used in the feature aggregation process, significantly enhancing the model's adaptability and generalization performance for scarce samples. Unlike classical generative models that primarily focus on reconstructing input features, our method shifts emphasis toward effective sampling of latent variables and semantic enhancement. This shift enables the model not only to learn the underlying distribution characteristics of few-shot remote sensing support data but also to generate highly discriminative feature prototype representations, better addressing intra-class distribution variation in few-shot scenarios.

In terms of specific implementation, the prior distribution $p(z)$ is set as a standard multivariate Gaussian distribution $N(0, I)$, while the approximate posterior distribution $q(z|S)$ is modeled as a multivariate Gaussian distribution $N(\mu, \sigma)$ with a diagonal covariance matrix. The parameters of these distributions, μ and σ, are generated by a specially designed feature encoder F_{enc}. To achieve a differentiable sampling process, the reparameterization trick is subsequently employed to generate variational features:

$$z = \mu + \sigma \cdot \epsilon, \epsilon \sim N(0, I) \tag{5}$$

Two loss terms with clear physical meanings are formulated as the evidence lower bound, namely the reconstruction loss and the KL divergence loss. The reconstruction loss is defined as the L2 distance between the input features and the features reconstructed through the decoder:

$$\mathrm{L}_{rec} = ||S - F_{dec}(z)|| \tag{6}$$

Meanwhile, the KL divergence loss is expressed as:

$$\mathrm{L}_{KL} = D_{KL}[q(z|S)||p(z)] \tag{7}$$

This loss term ensures that the variational features z closely adhere to the assumption of a standard normal distribution, thereby preserving the regularity and interpretability of the latent space. To further enhance the class-discriminative ability of the latent features z, this chapter introduces a novel consistency loss L_{cons}. This loss function is defined as the cross-entropy between the output of a lightweight linear classifier F_{cls} applied to the reconstructed features S' and the ground-truth class labels c:

$$\mathrm{L}_{cons} = L_{CE}[F_{cls}(S'), c] \tag{8}$$

By jointly optimizing this loss function, each support feature S_i is driven to approximate the parametric distribution $N(\mu_i, \sigma_i)$ of its corresponding category z, thereby ensuring that the sampled features z not only retain the semantic information of the original features but also exhibit enhanced inter-class discriminability. Specifically, the meta-learner adopts a class-agnostic aggregation strategy. Given a query feature Q_i from category i and support features S_j from category j, this chapter first precisely estimates their respective category distributions $N(\mu_j, \sigma_j)$, then obtains variational features $z_j = \mu_j + \sigma_j \cdot \epsilon$ through sampling, and finally employs the following aggregation scheme:

$$\mathrm{Q}_{ij} = A(Q_i, z_j) = Q_i \odot Sig(z_j) \tag{9}$$

where Sig denotes the sigmoid activation function. This aggregation approach maintains semantic consistency of features through channel-wise multiplication while introducing nonlinear transformation capabilities.

Furthermore, to more accurately measure the distribution discrepancy between query samples and support set categories, an Optimal Transport-based Distance Metric is incorporated into the generative meta-learner. This metric assesses class similarity by computing the Wasserstein distance between distributions. Specifically, let the query feature $h^q = f_\phi(x^q)$ be a Dirac distribution δ_{h^q}, and the i-th support category be a uniform empirical distribution $\hat{P}i = \frac{1}{L}\sum j = 1^L \delta_{h_i^j}$ over its L support samples $h_i^j j = 1^L$. The 2-Wasserstein distance between δh^q and $\hat{P}_i$ is:

$$W_2(\delta_{h^q}, \hat{P}i)^2 = \min \gamma \in \Pi(\delta_{h^q}, \hat{P}i) \int |h^q - h|^2 d\gamma(h^q, h) = \frac{1}{L}\sum j = 1^L |h^q - h_i^j|^2 \tag{10}$$

where $\Pi(\cdot, \cdot)$ denotes the set of couplings with given marginals, and $|\cdot|$ is the Euclidean distance. The predicted category is:

$$\hat{y} = \underset{i \in 1,\ldots,N}{\operatorname{argmin}}; W_2(\delta_{h^q}, \hat{P}_i)^2 \tag{11}$$

In summary, this section proposes an optimal transport-based generative meta-learning module that models categories as probability distributions rather than point estimates. By augmenting the support set through generative methods and measuring distribution similarity using the Wasserstein distance, the module significantly improves classification accuracy in few-shot remote sensing scenarios.

3.4. Orthogonal Frequency-Domain Decomposition Detection Head Algorithm

To effectively enhance feature discriminability and mitigate overfitting in few-shot remote sensing object detection, this paper proposes an orthogonal frequency-domain

decomposition detection head algorithm. This algorithm places the frequency-domain processing stage at the end of the traditional Feed-Forward Network (FFN), performing only lightweight orthogonal frequency-domain analysis and filtering on the output features. This design significantly enhances the discriminability of frequency-domain features while preserving the computational efficiency of the spatial-domain backbone. Compared to traditional methods that apply frequency-domain transformations at intermediate layers, this approach substantially reduces redundant conversions between frequency and spatial domains, effectively addressing issues such as frequency-domain feature degradation and sensitivity to noise under few-shot conditions. It enables robust frequency-aware feature reconstruction even with a limited number of training samples. The structure of this algorithm is illustrated in the following Figure 5:

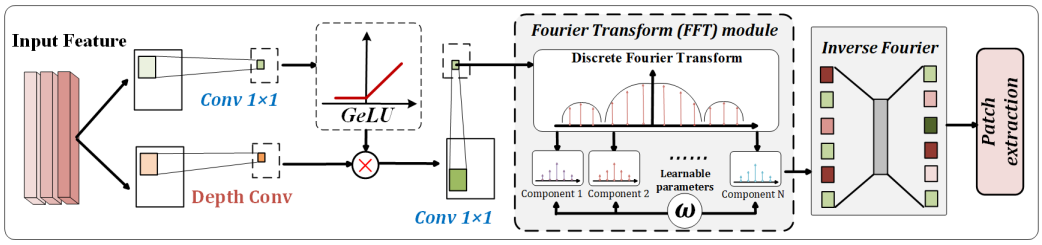

Figure 5. The overall structure of the orthogonal frequency domain decomposition detection head module.

Specifically, the operation of this detection head comprises three stages. First, the channel dimension is expanded via pointwise convolution, and spatial feature extraction is performed using depthwise separable convolution. Subsequently, the feature map is equally split along the channel dimension into two branches: one branch passes through a GELU activation function, while the other retains linear features. Finally, the two branches are element-wise multiplied to achieve efficient gated nonlinear fusion, followed by a pointwise convolution to restore the channel count. Secondly, the spatial feature map is divided into 8×8 pixel blocks, and the Fast Fourier Transform (FFT) is applied to each image block to convert it into the frequency domain. A learnable frequency-domain filter is used to adaptively weight the spectrum, achieving selective enhancement and suppression of frequency components through element-wise multiplication.

Finally, the Inverse Fast Fourier Transform (IFFT) is performed on the filtered spectrum to convert the processed frequency-domain information back into spatial image blocks. All processed image blocks are reassembled into a complete feature map, and the padded boundary portions are removed, ultimately outputting an enhanced feature map with the same dimensions as the input.

This method effectively enhances few-shot object detection performance through a collaborative spatial-frequency domain mechanism: frequency-domain filtering strengthens the discriminability of key target features while suppressing noise, and its lightweight design significantly reduces the risk of overfitting. The block-based processing balances multi-scale adaptability with computational efficiency, and the terminal frequency-domain operation minimizes overhead while preserving spatial perception. Together, these strategies address core challenges in few-shot scenarios, such as inadequate feature representation, weak generalization, and unstable optimization.

4. Experimental Analysis and Results

In this section, a comprehensive evaluation of the proposed method will be conducted. Firstly, a quantitative experimental comparison will be performed to analyze its performance differences in detection accuracy compared with current mainstream few-shot remote sensing object detection models based on meta-learning algorithms. The strengths

and weaknesses of different methods under data scarcity conditions will be discussed in depth. Subsequently, a complete set of ablation experiments will be carried out to verify the effectiveness of key modules in the proposed method layer by layer, illustrating the impact of each improvement on the model's final performance. Finally, to more intuitively demonstrate the detection capability of the model, several few-shot categories are selected, and their detection results are visualized. Combined with qualitative analysis, the model's performance in practical remote sensing object detection scenarios is discussed.

4.1. Dataset Introduction and Few-Shot Category Grouping

The DIOR few-shot dataset used in this experiment is constructed by grouping and subsetting the complete large-scale DIOR dataset. The original DIOR dataset consists of 23,463 remote sensing images, each with a uniform size of 800 × 800 pixels. The spatial resolution covered by the dataset ranges widely from 0.5 m to 30 m, making it suitable for multi-scale object detection and recognition tasks. The dataset provides a total of 192,472 finely annotated foreground instances, which are divided into 20 different categories according to a predefined semantic category system. These categories cover various typical ground objects and targets, such as vehicles, buildings, airports, and ports. Each instance is annotated with an accurate bounding box and category label, offering reliable supervision information for few-shot learning. The overall composition and instance distribution of the dataset are illustrated in the Figure 6 below.

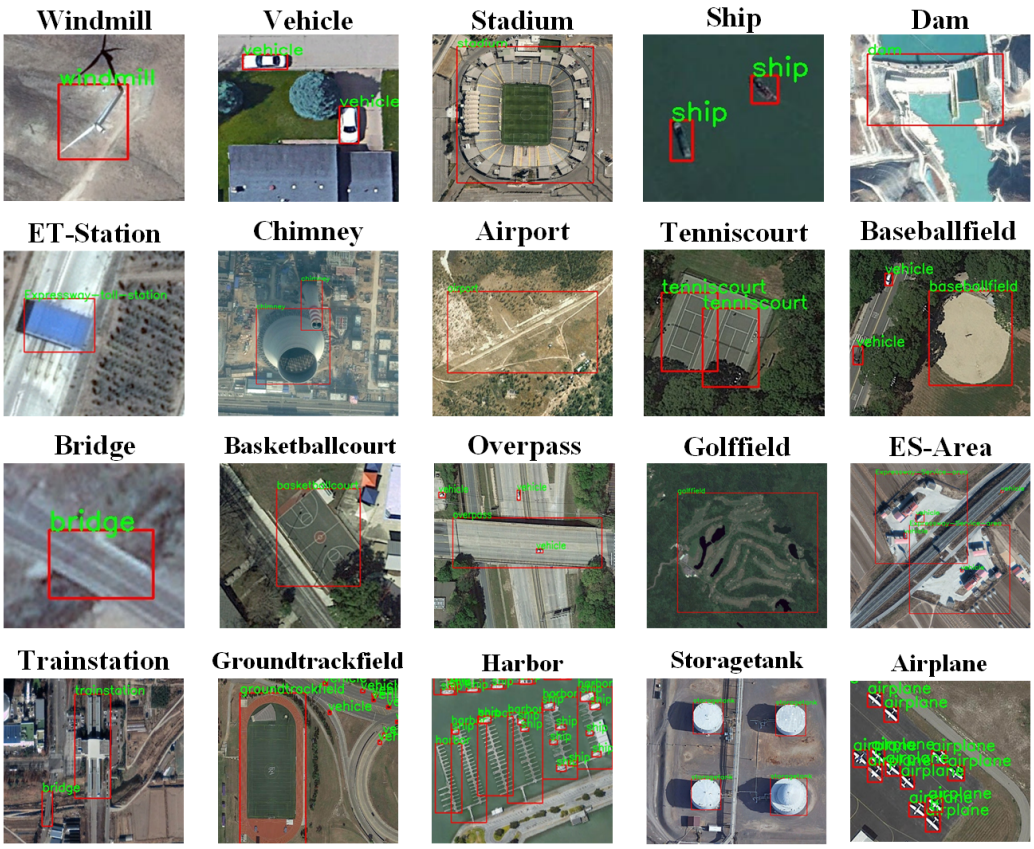

Figure 6. Statistical Distribution of Object Categories in the DIOR Dataset.

Based on previous research paradigms in few-shot object detection, in order to apply the DIOR dataset to few-shot tasks, the dataset needs to be divided into two experimental splits according to different categories. In each split, 5 categories are selected as novel classes, while the remaining 15 categories serve as base classes. The base classes retain their originally abundant samples, while for the extracted 5 novel classes, the number of samples

(shots) is set to 1, 2, 3, 5, and 10, respectively, thereby constructing a few-shot experimental environment. To ensure consistency with previous studies on few-shot learning, this paper uses the official validation set consisting of 5863 images to evaluate the model's performance across various aspects on the DIOR dataset. The few-shot class groupings corresponding to the DIOR dataset are shown in the Table 1 below:

Table 1. Base-Class and Novel-Class Splits on the DIOR Dataset.

Split Configuration	Split-1	Split-2
Novel Classes	airplane, airport, bridge, dam, ship	vehicle, overpass, harbor, dam, ship
Base Classes	Others	

The NWPU VHR-10 dataset adopted in this study is a medium-scale, high-resolution Earth observation dataset focused on remote sensing image analysis. It comprises 800 remote sensing images, of which 715 are positive sample images and 85 are negative background images. Training models on this dataset requires careful consideration of the distinctive characteristics of remote sensing imagery. The dataset includes a total of 10 finely delineated object categories, thus placing higher demands on the model's multi-scale object perception capability in complex geographical environments. The spatial resolution of the images in the NWPU VHR-10 dataset ranges from 0.5 to 2 m, and while the image sizes vary, they are generally large, typically falling between 1000 × 1000 pixels and 2000 × 2000 pixels. The dataset contains a total of 3651 high-quality instance annotations corresponding to the image data. Details please check at Figure 7.

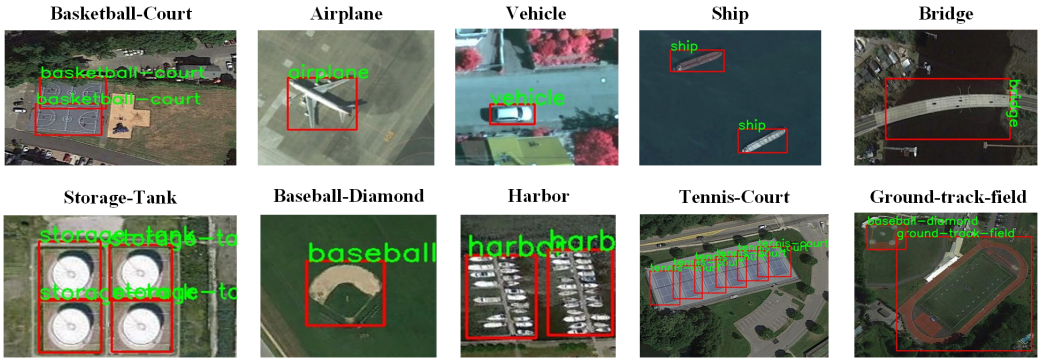

Figure 7. Statistical Distribution of Object Categories in the NWPU-VHR-10 Dataset.

Existing research on few-shot detection in remote sensing typically divides the 10 categories of the NWPU VHR-10 dataset into 7 base classes and 3 novel classes based on the number of instances and category characteristics. For the extracted 3 novel classes, the number of samples (shot) is set to 1, 3, 5, and 10, respectively. The selection of novel classes usually considers categories with fewer instances, such as bridges, tennis courts, and the like. To ensure the reliability of the comparative experiments, this paper adopts a widely recognized data partitioning protocol in the academic community: 60% of the positive sample images (429 images) are used as the training set, while the remaining 40% of the positive sample images (286 images) and all 85 negative sample images are used as the test set, thereby comprehensively evaluating model performance. As shown in the table below, the experimental grouping for the NWPU VHR-10 dataset is as follows Table 2:

Table 2. Base-Class and Novel-Class Splits on the NWPU VHR-10 Dataset.

Split Configuration	Split-1
Novel Classes	bridge, baseball-diamond, airplane
Base Classes	basketball-court, ground-track-field, harbor, ship, storage-tank, tennis-court, vehicle

In summary, through scientific category division and controlled sample settings (such as the separation of base/novel classes and the construction of learning scenarios with different shots), the DIOR and NWPU VHR-10 datasets together form a well-structured and clearly defined experimental benchmark for few-shot detection. This benchmark not only covers remote sensing object detection scenarios at different scales and resolutions but also provides a reliable and consistent experimental foundation for subsequent performance validation, comparative analysis, and reproducible research of related methods.

4.2. Introduction to the Experimental Setup

4.2.1. Experimental Runtime Environment

The experimental environment in this paper is built upon the PyTorch deep learning framework, with all code executed under the Ubuntu operating system. To accelerate the model training and inference processes, the experiments utilize NVIDIA Tesla A800 GPUs, along with the CUDA 11 parallel computing platform for efficient parallel computation optimization. In terms of algorithm implementation, the model construction, training, and evaluation primarily rely on the MMFewShot few-shot detection toolkit, which provides standardized workflow support and a rich set of benchmark implementations for few-shot learning tasks. Furthermore, the experiments integrate a range of commonly used Python scientific computing and image processing libraries, including NumPy for numerical operations and OpenCV for image preprocessing and augmentation, collectively forming a comprehensive toolchain for algorithm development and experimental validation.

To ensure the fairness and reproducibility of the comparative experiments, all baseline models and our proposed method were evaluated under strictly identical conditions. Specifically, all methods were trained and tested using the same hardware configuration (NVIDIA Tesla A800 GPUs), software environment (Ubuntu operating system, PyTorch framework, CUDA 11), and toolkit versions (MMFewShot). The data partitioning strategies, input image sizes, batch sizes, and optimization hyperparameters (e.g., learning rate, momentum, weight decay) were kept consistent across all compared methods. For baseline models that require specific tuning, we adopted the optimal hyperparameter settings reported in their original papers and fine-tuned them within the same validation set to ensure a fair comparison. All experiments were conducted with fixed random seeds to minimize stochastic variability. Through this rigorous control of extraneous variables, we guarantee that the performance differences observed in our experiments can be attributed solely to the methodological innovations rather than inconsistencies in experimental conditions.

4.2.2. Evaluation Metrics for Experimental Results

To ensure a fair comparison with previous studies, this paper adopts the mean Average Precision at an Intersection over Union (IoU) threshold of 0.5 (mAP50) across all categories of the two datasets as the core performance evaluation metric. Furthermore, for a multi-dimensional assessment of model performance, ablation experiments introduce three supplementary metrics: mean Average Precision at an IoU threshold of 0.75 (mAP75) across all categories, average precision for base-class samples (bAP), and average precision for novel-class samples (nAP).

4.3. Analysis of Comparative Experimental Results

To quantitatively demonstrate the accuracy performance of the model across different remote sensing application scenarios, it is necessary to conduct a horizontal accuracy comparison with other approaches in the same domain using identical datasets. This provides a quantifiable reference standard for the research presented in this paper. This chapter adopts two preprocessed few-shot optical remote sensing image datasets, DIOR and NWPU VHR-10, as the evaluation benchmarks for comparative experiments.

The quantitative comparative experimental results of the method proposed in this paper are presented in the two tables that follow. The accuracy data in these tables correspond to the DIOR and NWPU VHR-10 datasets, respectively. Overall, the method proposed in this paper achieves significant improvements across all accuracy metrics. By conducting a horizontal quantitative comparison with the accuracy performance of other methods, it can be observed that the model proposed in this paper surpasses most current meta-learning-based few-shot object detection methods in remote sensing. First, the accuracy comparison experiment is conducted on the DIOR dataset. In this paper, mAP50 (mean Average Precision at an IoU threshold of 0.5) is used as the evaluation metric for remote sensing object detection accuracy across all categories. The detection accuracy of the method proposed in this paper is compared with some meta-learning-based few-shot object detection models, as shown in the Table 3 below:

Table 3. Comparison of Detection Model Accuracy on the DIOR Dataset.

Method	Split 1					Split 2				
	1 Shot	**2 Shot**	**3 Shot**	**5 Shot**	**10 Shot**	**1 Shot**	**2 Shot**	**3 Shot**	**5 Shot**	**10 Shot**
Meta R-CNN	–	–	12.02	13.09	14.07	–	–	9.10	12.29	11.96
FsDetView	–	–	13.19	14.29	18.02	–	–	7.49	12.61	11.49
TFA	–	–	16.07	15.36	16.45	–	–	8.73	9.31	12.19
FSOD	–	–	15.94	20.27	24.22	–	–	10.40	10.74	12.26
FSCE	–	–	27.91	28.60	33.05	–	–	15.59	16.24	23.75
MSOCL	–	–	24.97	27.27	33.37	–	–	13.11	15.07	23.30
ICPE	–	–	11.68	12.34	12.95	–	–	10.56	11.21	12.38
P-CNN [37]	–	–	18.00	22.80	27.60	–	–	16.50	18.80	23.30
TINet (R-50) [31]	–	–	–	21.70	24.01	–	–	–	19.20	21.55
TINet (R-101) [31]	–	–	–	22.80	25.30	–	–	–	21.98	22.38
FSODM [38]	17.28	19.81	22.37	24.49	29.33	13.79	14.28	15.21	15.97	20.83
Ours	**19.21** ±0.6%	**21.83** ±0.3%	**25.81** ±1.1%	**29.37** ±0.6%	**35.11** ±0.5%	**11.27** ±0.5%	**13.59** ±0.9%	**17.26** ±0.7%	**18.91** ±1.0%	**25.81** ±0.9%

The experimental results show that the few-shot object detection algorithm based on generative meta-learning achieved excellent performance across both few-shot splits of this remote sensing dataset. Specifically, under the same evaluation criteria, the method proposed in this paper was compared with eight high-performing few-shot object detection methods in remote sensing. In most accuracy comparisons, the proposed method outperformed these eight approaches. Experiments on the Split1 few-shot split demonstrate that the proposed method achieves the best accuracy among all transfer learning methods. Compared to the second-highest ranked model, it achieves an accuracy improvement of 3% to 8% in both 1-shot and 10-shot few-shot experiments. On the Split2 few-shot split, the proposed model also surpasses the accuracy of most comparative methods, achieving the best performance in most cases with an improvement of approximately 2% to 3%. In a few specific sub-branch accuracy comparisons, it slightly underperforms compared to the ICPE model. Therefore, based on the above results, it can be concluded that the few-shot object detection algorithm based on generative meta-learning performs well in terms of the

mAP50 metric across all optical remote sensing categories. A comprehensive evaluation of the model's performance indicates that it not only achieves effective generalization to few-shot remote sensing object instances but also hardly compromises the detection effectiveness for base-class instances, thereby successfully facilitating the transfer of knowledge from base classes to novel classes.

To further validate the superiority of our method in few-shot object detection tasks for remote sensing imagery, we compared the proposed framework with three remote sensing-specific few-shot detection methods: P-CNN [37], TINet [31] (with ResNet-50/101 backbones), and FSODM [38]. The experimental results demonstrate that our method significantly outperforms all compared approaches across all shot settings. Specifically, compared with P-CNN, our method achieves improvements of 7.81, 6.57, and 7.51 percentage points on the 3-shot, 5-shot, and 10-shot tasks of Split 1, respectively. Against TINet (ResNet-101), our method surpasses it by 6.57 and 9.81 percentage points on the 5-shot and 10-shot settings, respectively. Compared with FSODM, our method achieves performance gains ranging from 1.93 to 5.78 percentage points across the comprehensive evaluation from 1-shot to 10-shot settings.

Subsequently, similar comparative experiments were conducted on the NWPU VHR-10 dataset. In these experiments, the few-shot settings primarily adopted were 10-shot and 30-shot, using nAP50 and nAP75 (with IoU thresholds of 0.5 and 0.75, respectively) for novel classes as the evaluation metrics. The results show that, compared to other meta-learning methods, the proposed approach in this paper achieved a certain degree of improvement in average precision for few-shot novel class detection, as illustrated in Table 4 below:

Table 4. Comparative Experiment of Algorithm Accuracy on the NWPU VHR-10 Dataset.

Method	Novel 5-Shot		Novel 10-Shot	
	nAP50	**nAP75**	**nAP50**	**nAP75**
Meta R-CNN	21.77	12.38	26.98	14.81
FsDetView	29.55	14.92	31.77	16.27
TFA	20.49	11.88	21.22	11.93
FSOD	15.13	8.21	16.23	10.10
FSCE	15.08	8.93	13.89	8.24
ICPE	9.10	7.11	12.00	8.37
Ours (50)	**31.28** $\pm$ 0.8%	**16.17** $\pm$ 0.6%	**34.22** $\pm$ 1.2%	**18.29** $\pm$ 0.4%
Ours (101)	**32.19** $\pm$ 0.7%	**16.97** $\pm$ 0.7%	**35.19** $\pm$ 0.9%	**18.92** $\pm$ 0.3%

The table above presents the evaluation results of the object detection models on the NWPU VHR-10 dataset, assessing the mean detection accuracy for novel classes under the settings of Shot = 5 and Shot = 10. To achieve a more comprehensive evaluation on the NWPU VHR-10 dataset, the criterion of intersection over union (IoU) greater than or equal to 0.75 (AP75) was additionally adopted. This metric, along with the previously mentioned AP50, was used to jointly calculate the average precision for novel classes. Compared to mainstream methods based on transfer learning and meta-learning, the generative meta-learning-based model delivered superior detection results. Specifically, when compared to the FsDetView meta-learning framework, the method proposed in this paper achieved an improvement of 2.45 percentage points in the 10-shot setting and 1.73 percentage points in the 5-shot setting. These results indicate that the model achieved optimal performance across different benchmark comparisons, demonstrating its strong generalization capability in few-shot remote sensing object detection tasks.

To enhance the statistical reliability of the experimental results, we conducted multiple independent rounds of training and testing for each experimental setting using five different random seeds. This approach mitigates the bias caused by stochastic factors such as parameter initialization or data loading order in single-run experiments. Based on the outcomes of these multiple runs, we calculated the mean and 95% confidence interval (CI) for each accuracy metric, providing an illustration of the performance variability and stability: a narrower interval indicates stronger robustness of the model against random factors, while non-overlapping intervals can, to a certain extent, suggest that the performance differences between methods are statistically significant. The supplemented experimental results are presented in Tables 3 and 4. As shown, the proposed method not only achieves superior mean performance compared to the baseline methods, but its confidence intervals are either non-overlapping or exhibit minimal overlap with those of the baselines under most experimental settings. This further validates the reliability and statistical significance of the performance improvements achieved by our method. Additionally, the inclusion of error bars (±X%) offers readers an intuitive understanding of the model's stability.

4.4. Visualization Experiment Analysis

To intuitively demonstrate the performance of the proposed method in multi-scenario remote sensing object detection, Figure 8 presents the visualization results of few-shot object detection on the DIOR and NWPU-VHR-10 datasets, respectively. As shown in the figures, the model can accurately locate and identify typical remote sensing targets (e.g., vehicles, ships, aircraft) across varying scales and background complexities, exhibiting high bounding box regression accuracy while maintaining robust detection capability under dense arrangements and partial occlusions.

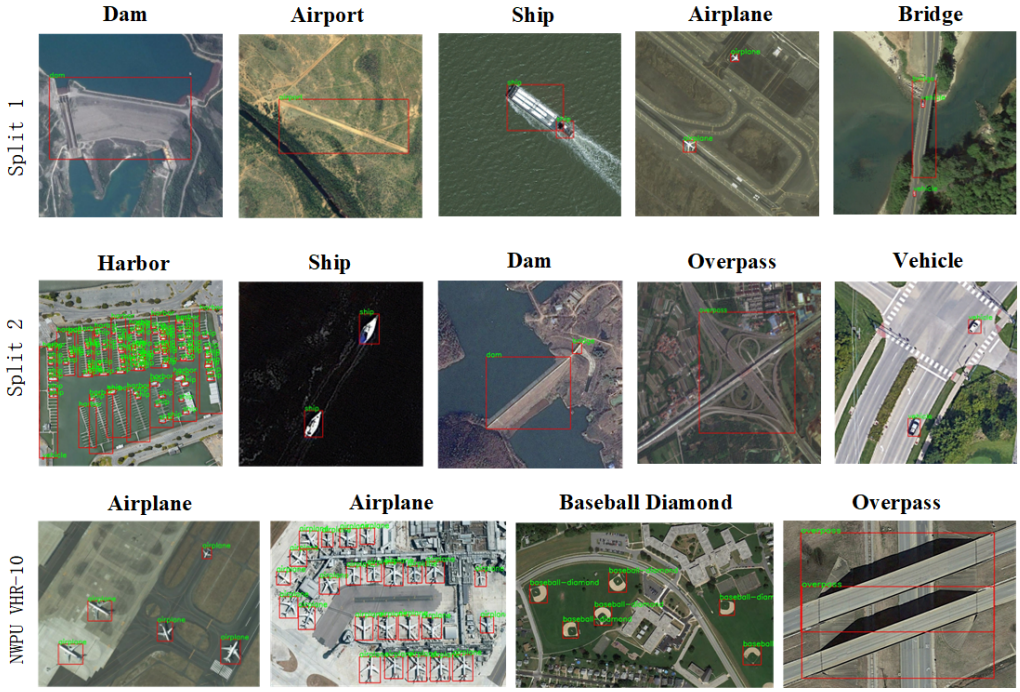

Figure 8. Visualizationresults of detected objects on DIOR and NWPU-VHR-10 datasets.

Furthermore, Figure 9 displays the Grad-CAM heatmap visualization results based on the DIOR dataset. By visualizing the key regions focused on during the model's decision-making process, this figure reveals the network's attention mechanism toward target semantic features. The heatmap distribution indicates that the model's attention is highly concentrated on the main structure of the targets rather than redundant background information, validating the effectiveness of the proposed method in feature selection and semantic perception.

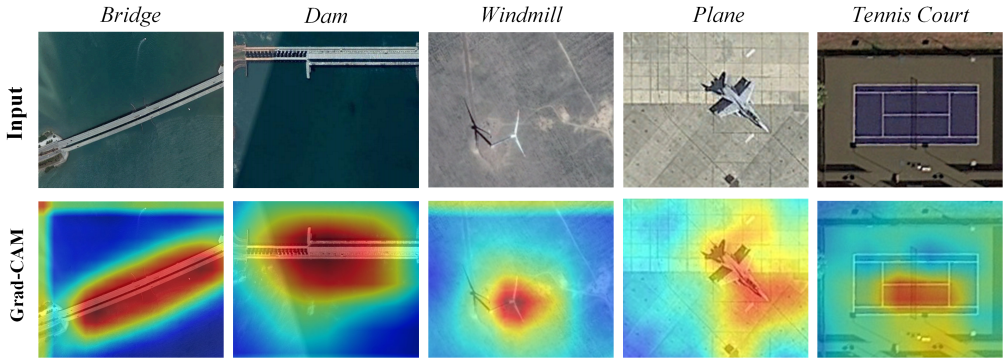

Figure 9. Grad-CAM heatmap visualization for the DIOR dataset.

The comprehensive visualization results demonstrate that the proposed method achieves consistent detection performance and interpretability across different remote sensing datasets. Its attention mechanism and feature fusion strategy contribute to strong discriminative ability and localization accuracy for remote sensing targets.

4.5. Ablation Experiment Analysis

To quantitatively evaluate the impact of each internal component in the method proposed in this chapter on its performance, an ablation experiment was designed to verify the effectiveness of the various improvements. The Split1 of the DIOR dataset was adopted as the benchmark for the ablation study, and the experimental results are shown in Tables 5–7. The results of the ablation experiment assess the influence of three main improvement components on the model's performance from both quantitative and qualitative perspectives. Through systematic analysis of the quantitative results, the following key conclusions can be drawn:

First, by comparing the data in the first and second columns of Tables 5–7, it can be observed that after introducing OT-GML, the model shows a significant improvement in the average precision across all categories under few-shot settings with sample sizes of 1, 2, 3, and 10. The evaluation metrics for the novel categories exhibit a similar distribution pattern, indicating that this measurement method effectively mitigates confusion between novel categories and other categories. Notably, a decrease in accuracy occurs at the 5-shot stage. In-depth analysis reveals that this specific sample size for novel categories may amplify distance errors in similarity measurements between novel and base classes, thereby weakening the model's detection performance for novel classes.

Second, the results in the third column of Tables 5–7 show that after introducing the OFD method, the average precision across all categories under various experimental conditions significantly improved. This confirms that OFD not only enhances the model's learning ability for new samples but also stabilizes the similarity metrics between different categories under critical conditions, thereby achieving more robust performance improvements.

Table 5. Ablation Study Results of Model Improvement Schemes.

Sample	Model Configuration				
	Baseline	**+OT-GML**	**+OFD**	**+DRDS**	**Full Model**
DRDS	–	–	–	✓	✓
OT-GML	–	✓	✓	✓	✓
OFD	–	–	✓	–	✓
1-Shot Evaluation					
AP50	15.31	16.28	18.29	17.41	**19.21**
AP75	7.22	7.89	8.22	8.01	**9.33**
2-Shot Evaluation					
AP50	19.04	19.83	21.01	20.97	**21.83**
AP75	9.81	10.21	10.59	9.30	**11.09**
3-Shot Evaluation					
AP50	22.39	23.28	24.07	23.97	**25.81**
AP75	10.27	10.82	11.81	11.29	**12.57**
10-Shot Evaluation					
AP50	30.79	31.51	34.29	33.97	**35.11**
	11.61	12.29	13.29	13.37	**14.22**

Table 6. Comparison of Model Complexity and Inference Efficiency.

Method	**Params (M)**	**FLOPs (G)**	**Speed (FPS)**	**mAP (%)**
Baseline	58	292	7.11	24.29
+DRDS	69	345	6.58	26.42
+OT-GML	71	352	5.92	28.15
+OFD Head	76	378	5.43	29.37
Full Framework	**76**	**378**	**5.43**	**29.37**

Table 7. Ablation study: Impact of different block sizes on OFD module performance.

Block Size	**GFLOPs**	**mAP@50**	**Speed (FPS)**
4×4	14.2	21.79	7.89
8×8	18.8	23.11	7.21
16×16 (Ours)	**26.3**	**29.37**	**5.43**
32×32	39.2	31.29	4.29

Further analysis of the data in the fourth and fifth columns of the table reveals that after introducing DRDS, the model's overall accuracy metrics improved markedly. More importantly, as DRDS was progressively integrated, the model's accuracy continued to show significant enhancements. On one hand, this demonstrates that DRDS effectively strengthens the model's ability to perceive query information. On the other hand, it verifies that the two earlier improvement modules maintain strong adaptability within the new feature extraction framework, enabling them to achieve stable accuracy gains.

4.6. Model Efficiency Analysis

To comprehensively evaluate the computational cost of the proposed detection framework, we conducted a quantitative analysis of model complexity and efficiency. Specifically, under identical experimental conditions, we compared the baseline model with the complete DRDS framework (incorporating the DRDS, OT-GML, and OFD Head modules)

in terms of the number of parameters, FLOPs, and average inference time per image. The experimental results are presented in Table 6.

As can be seen from the table, the parameter count and computational complexity gradually increase with the sequential introduction of the DRDS, OT-GML, and OFD Head modules. The complete framework exhibits **76 M** parameters, **378 G** FLOPs, and an inference speed of **5.43 FPS** (corresponding to approximately **184 ms per image**). Compared to the baseline model, the complete framework shows an increase of **31.0%** in parameters (from 58 M to 76 M), **29.5%** in computational cost (from 292 G to 378 G), and a **23.6%** reduction in inference speed (from 7.11 FPS to 5.43 FPS). However, this increase in computational overhead yields a substantial improvement in detection accuracy, achieving a gain of **5.08** percentage points in mAP (from 24.29% to 29.37%, see Section X for details). This demonstrates that although model complexity increases to a certain extent, the overall computational overhead remains within an acceptable range through rational module design and feature reuse mechanisms, while attaining considerable performance gains. Throughout the design process, we have fully considered the trade-off between accuracy and efficiency to ensure the feasibility of the proposed method in practical application scenarios such as industrial inspection.

4.7. Efficiency and Sensitivity Analysis of the Relevant Modules Within the Model

To empirically validate the effectiveness of the proposed Orthogonal Frequency-domain Decomposition (OFD) module and investigate the impact of its key hyperparameters, we conduct a series of ablation studies. Specifically, we analyze the sensitivity of the module to different block size configurations, as this parameter governs the granularity of frequency-domain decomposition and directly influences both feature representation quality and computational efficiency.

Table 7 presents ablation results for different block sizes in the OFD module. As block size increases from 4 × 4 to 32 × 32, both computational complexity and detection accuracy rise. mAP@50 improves from 21.79 (4 × 4) to 31.29 (32 × 32), indicating that larger blocks capture broader context and finer frequency resolution, enhancing feature discriminability. However, this gain comes at the cost of increased overhead: GFLOPs increase from 14.2 to 39.2, and inference speed drops from 7.89 to 4.29 FPS. We adopt square blocks because FFT is highly optimized for square matrices, maximizing efficiency and preserving the lightweight design. Square partitioning also simplifies reconstruction after frequency filtering, eliminating complex alignment and ensuring end-to-end trainability. Thus, we constrain block shape to square, control decomposition scale via side length n, and systematically evaluate its impact through ablation.

Table 8 presents the computational cost and efficiency comparison of different modules. Compared to the baseline, our DRDS module increases GFLOPs from 292 to 345, while the Transformer and Pure Mamba counterparts raise it to 319 and 324, respectively. Although DRDS incurs slightly higher computational overhead (+53 GFLOPs vs. baseline), it achieves the best detection performance with 26.42 mAP@50, surpassing Transformer (24.97) and Pure Mamba (25.09) by +1.45 and +1.33, respectively. In terms of inference speed, DRDS achieves 6.58 FPS, which is marginally lower than baseline (7.11) but comparable to Transformer (6.97) and Pure Mamba (7.05). This slight reduction in speed is a reasonable trade-off for the significant performance gain. The results demonstrate that our DRDS module effectively balances computational efficiency and detection accuracy, achieving superior few-shot generalization with acceptable overhead.

Table 8. The computational cost and efficiency analysis of the DRDS module.

Method	GFLOPs	Speed (FPS)	mAP@50
Baseline	292	7.11	24.29
+Transformer	319	6.97	24.97
+Pure Mamba	324	7.05	25.09
+DRDS (Ours)	345	6.58	26.42

Table 9 presents the detailed architecture and hyperparameter configuration of the VAE-based generative module. Regarding the network structure, the encoder F_{enc} adopts a three-layer fully connected network ($256 \rightarrow 512 \rightarrow 128$) with ReLU activation and batch normalization, outputting the mean μ and log variance $\log \sigma^2$ of the Gaussian distribution. The decoder F_{dec} employs a symmetric structure ($128 \rightarrow 512 \rightarrow 256$) also with ReLU activation. The latent variable z is set to 128 dimensions, and the lightweight classifier F_{cls} is a single linear layer ($256 \rightarrow 5$). For loss function weighting, the KL divergence weight α is set to 0.1, and the consistency loss weight β is set to 0.5. In terms of sampling strategy, $L_{train} = 5$ variational features are sampled per support sample during training, increasing to $L_{test} = 10$ during inference. The optimizer uses Adam with an initial learning rate of 1×10^{-3}. This configuration comprehensively presents the implementation details of the generative module, providing the foundation for the reproducibility of the proposed method.

Table 9. Architecture and hyperparameter configuration of the VAE-based generative module.

Component	Configuration
Encoder F_{enc}	$256 \rightarrow 512 \rightarrow 128$, ReLU + BN, output μ / $\log \sigma^2$
Decoder F_{dec}	$128 \rightarrow 512 \rightarrow 256$, ReLU
Latent variable z	Dimension: 128
Classifier F_{cls}	$256 \rightarrow 5$ (linear)
Component	**Loss Weights**
KL weight α	0.1
Consistency weight β	0.5
Component	**Sampling Strategy**
Training L_{train}	5 per support
Inference L_{test}	10 per support
Component	**Optimization**
Optimizer	Adam, learning rate 1×10^{-3}

Table 10 presents the hyperparameter sensitivity analysis of the VAE-based generative module in OT-GML, evaluated using mAP (%). For the KL divergence weight α, experiments were conducted across a range from 0.01 to 1.0. The results show that optimal performance (27.64%) is achieved at $\alpha = 0.1$, while values that are too small (0.01) or too large (1.0) lead to performance degradation (25.32% and 24.93%, respectively). This indicates that appropriate KL regularization is crucial for balancing latent space structure and feature reconstruction quality. For the consistency loss weight β, comparative experiments demonstrate that introducing this loss ($\beta = 0.5$) improves mAP from 24.86% to 27.97%, validating its essential role in enhancing the inter-class discriminability of generated features. For the number of inference samples L, values ranging from 3 to 20 were tested. The results show that $L = 10$ yields the best performance (28.15%), while $L = 1$ results in lower performance (25.41%) due to excessive sampling randomness, and $L = 20$ leads to a slight performance decrease (27.89%), indicating stable model performance within the

range of 5–20. This analysis identifies the optimal configuration for each hyperparameter, providing experimental support for the effectiveness of the VAE-based generative module.

Table 10. Hyperparameter sensitivity analysis of OT-GML (mAP%).

Hyperparameter	Value	mAP (%)
KL weight α	0.01	25.32
	0.05	26.78
	0.1	**27.64**
	0.5	26.15
	1.0	24.93
Consistency weight β	0.0	24.86
	0.5	**27.97**
Inference samples L	3	25.41
	5	27.32
	10	**28.15**
	20	27.89

5. Conclusions

In summary, the three improvements proposed in this paper can significantly enhance the performance of few-shot object detection models. They not only improve the detection accuracy for new categories but also effectively maintain performance on base categories, fully validating the effectiveness of the strategies. These modules exhibit strong synergistic effects within a complete framework, and their combined use leads to superior performance gains. However, models based on deep neural networks typically involve a large number of parameters and high computational complexity. Therefore, it is necessary to design lightweight models to facilitate deployment on edge devices with limited computational resources, such as satellite-based or airborne platforms.

Author Contributions: Conceptualization: X.S.; Methodology: X.S., S.L., and Q.H.; Software: X.S. and S.L.; Validation: X.S., S.L., Q.H. and C.L.; Formal analysis: X.S., S.L., and Q.H.; Investigation: X.S. and S.L.; Resources: X.S. and Y.D.; Data curation: X.S. and Y.D.; Writing—original draft: X.S., S.L., Q.H., and Y.D.; Writing—review & editing: X.S., S.L., Q.H., and Y.D.; Visualization: X.S. and Y.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Henan Provincial Science and Technology Research Project under Grant 242102211006, the Universities in Henan Province under Grant 25A620001, the Shanxi Province Basic Research Program (Free Exploration for Youth Science Research Project) under Grant 202303021212387 and the General International Scientific and Technological Cooperation Project of Henan Province under Grant No. 252102521049.

Data Availability Statement: The DIOR dataset used in this study is publicly available in the opendatalab repository: https://opendatalab.org.cn/OpenDataLab/DIOR (accessed on 1 July 2025). The NWPU-VHR-10 dataset is publicly available in the GitHub repository: https://github.com/Gaoshuaikun/NWPU-VHR-10 (accessed on 19 July 2025).

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

References

1. Huang, G.; Laradji, I.; Vázquez, D.; Lacoste-Julien, S.; Rodríguez, P. A Survey of Self-Supervised and Few-Shot Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4071–4089. [CrossRef]
2. Wei, J.; Lan, Y.; Tang, T.; Liu, T. A Survey on Transfer Reinforcement Learning. In Proceedings of the 2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE), Shanghai, China, 21–23 March 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 2511–2518.

3. Zhang, P.; Liu, C.; Chang, X.; Li, Y.; Li, M. Metric-based Meta-Learning Model for Few-Shot PolSAR Image Terrain Classification. In Proceedings of the 2021 CIE International Conference on Radar (Radar), Haikou, China, 15–19 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 2529–2533.
4. Gaikwad, M.; Doke, A. Survey on Meta Learning Algorithms for Few Shot Learning. In Proceedings of the 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 25–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1876–1879.
5. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
6. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
7. Liu, S.; You, Y.; Su, H.; Meng, G.; Yang, W.; Liu, F. Few-Shot Object Detection in Remote Sensing Image Interpretation: Opportunities and Challenges. *Remote Sens.* **2022**, *14*, 4435. [CrossRef]
8. Zhang, X.; Zhang, T.; Wang, G.; Zhu, P.; Tang, X.; Jia, X.; Jiao, L. Remote Sensing Object Detection Meets Deep Learning: A Metareview of Challenges and Advances. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 8–44. [CrossRef]
9. 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. [CrossRef]
10. Sun, X.; Wang, P.; Wang, C.; Liu, Y.; Fu, K. PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 50–65. [CrossRef]
11. Lv, N.; Ma, H.; Chen, C.; Pei, Q.; Zhou, Y.; Xiao, F.; Li, J. Remote sensing data augmentation through adversarial training. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *14*, 9318–9333. [CrossRef]
12. Sun, S.; Mu, L.; Wang, L.; Liu, P.; Liu, X.; Zhang, Y. Semantic segmentation for buildings of large intra-class variation in remote sensing images with o-gan. *Remote Sens.* **2021**, *13*, 475. [CrossRef]
13. Qiao, B.; Zhou, H.; Yang, L.; Xie, X. Few shot object detection with incompletely annotated samples. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 18–23 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–7.
14. Liu, N.; Xu, X.; Gao, Y.; Zhao, Y.; Li, H.-C. Semi-supervised object detection with uncurated unlabeled data for remote sensing images. *Int. J. Appl. Earth Observ. Geoinf.* **2024**, *129*, 103814. [CrossRef]
15. Su, H.Z.; You, Y.A.; Liu, S.X. Multi-oriented enhancement branch and context-aware module for few-shot oriented object detection in remote sensing images. *Remote Sens.* **2023**, *15*, 3544. [CrossRef]
16. Li, Y.; Chen, L.; Li, W.; Wang, N. Few-shot fine-grained classification with rotation-invariant feature map complementary reconstruction network. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5608312. [CrossRef]
17. Zhou, J.W.; Li, W.Z.; Cao, Y.; Li, X. Few-shot oriented object detection with memorable contrastive learning in remote sensing images. *arXiv* **2024**, arXiv:2403.13375. [CrossRef]
18. Wu, W.; Jiang, C.; Yang, L.; Wang, W.; Chen, Q.; Zhang, J.; Yang, H.; Chen, Z. Arbitrary oriented few-shot object detection in remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2024**, *17*, 17930–17944. [CrossRef]
19. Xiao, Y.; Lepetit, V.; Marlet, R. Few-shot object detection and viewpoint estimation for objects in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 3090–3106.
20. Lin, W.; Tang, X.; Han, X.; Ma, J.; Zhang, X.; Jiao, L. Nq-protonet: Noisy query prototypical network for few-shot remote sensing scene classification. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 3620–3623.
21. Yang, X.; Nan, X.T.; Song, B. D2n4: A discriminative deep nearest neighbor neural network for few-shot space target recognition. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 3667–3676. [CrossRef]
22. Guo, M.K.; You, Y.A.; Liu, F. Discriminative prototype learning for few-shot object detection in remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5624313. [CrossRef]
23. Zhang, Y.C.; Zhang, B.; Wang, B. Few-shot object detection with self-adaptive global similarity and two-way foreground stimulator in remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2022**, *15*, 7263–7276. [CrossRef]
24. Zhao, Z.; Tang, P.; Zhao, L.; Zhang, Z. Few-shot object detection of remote sensing images via two-stage fine-tuning. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8021805. [CrossRef]
25. 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*, 5615211. [CrossRef]
26. 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*, 5612114. [CrossRef]
27. 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. [CrossRef]

28. 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*, 5617011. [CrossRef]
29. 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*, 5603514. [CrossRef]
30. Yang, Z.; Guan, W.; Xiao, L.; Chen, H. Few-shot object detection in remote sensing images via data clearing and stationary meta-learning. *Sensors* **2024**, *24*, 3882. [CrossRef]
31. 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*, 5625314. [CrossRef]
32. 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*, 5628014. [CrossRef]
33. Wang, Y.; Hong, M.; Huangfu, L.; Huang, S. Data Distribution Distilled Generative Model for Generalized Zero-Shot Recognition. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 5695–5703. [CrossRef]
34. Gupta, P.; Hayat, M.; Dhall, A.; Do, T.T. Conditional Distribution Modelling for Few-Shot Image Synthesis with Diffusion Models. In Proceedings of the Asian Conference on Computer Vision, Hanoi, Vietnam, 8–12 December 2024; Springer: Berlin/Heidelberg, Germany, 2024; Volume 1, pp. 818–834.
35. Abdollahzadeh, M.; Liu, G.; Malekzadeh, T.; Teo, C.T.H.; Chandrasegaran, K.; Cheung, N.-M. A survey on generative modeling with limited data, few shots, and zero shot. *arXiv* **2023**, arXiv:2307.14397. [CrossRef]
36. Giannone, G. Few-Shot Generative Models: Learning Generative Models with Limited Data. Ph.D. Thesis, Technical University of Denmark, Kongens Lyngby, Denmark, 2024.
37. 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*, 5604610. [CrossRef]
38. Li, X.; Deng, J.; Fang, Y. Few-shot object detection on remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5601614. [CrossRef]

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