DAFSDet: Dual-Attention Guided Few-Shot Object Detection in Remote Sensing Images

Guangshuai Gao * , Zhilin Zhang , Wei Zhang , Yunqi Shang , Yan Dong , Jiangtao Xi

Posted Date: 12 February 2026

doi: 10.20944/preprints202602.1028.v1

Keywords: remote sensing images; few-shot object detection; dual-attention guidance; content-aware strip pyramid; deformable attention region proposal network

Article

DAFSDet: Dual-Attention Guided Few-Shot Object Detection in Remote Sensing Images

Guangshuai Gao [1,*], Zhilin Zhang [1], Wei Zhang [1], Yunqi Shang [1], Yan Dong [1,2] and Jiangtao Xi [3]

[1] School of Information and Communication Engineering, Zhongyuan University of Technology, Zhengzhou 450007, China
[2] School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 610000, Sichuan Province, China
[3] School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Australia
* Correspondence: 6911@zut.edu.cn

Highlights

What are the main findings?

- Content- and direction-aware context aggregation is particularly effective for detecting objects with large scale variation and elongated shapes in remote sensing few-shot scenarios, while deeper networks or conventional multi-scale fusion bring limited gains.
- Adaptive spatial sampling and foreground-focused attention improve proposal localization in complex backgrounds, whereas fixed receptive-field designs struggle with geometric deformation and background clutter in few-shot detection.

What are the implications of the main findings?

- Few-shot object detection in remote sensing shows greater benefits from structure-aware and geometry-adaptive mechanisms than from conventional scale-agnostic feature fusion, providing a new perspective for designing data-efficient detection architectures.
- Observed improvements show that explicitly guiding proposals to better capture object regions at the proposal generation stage leads to enhanced region-level adaptability in few-shot detectors, rather than relying solely on classification refinement for improving performance.

Abstract

Few-shot object detection focuses on accurately identifying and locating novel categories with just a small set of annotated examples. However, accurate object detection in remote sensing images still faces challenges such as scale variations, small objects, and complex background interferences. To alleviate these issues, we propose DAFSDet, a dual attention guided few-shot object detection framework, to effectively mine discriminative key features. Specifically, we propose a Content-Aware Strip Pyramid (CASP) module that uses content-aware upsampling for spatial attention and bidirectional strip convolution for long-range context, forming a joint spatial–semantic attention mechanism. CASP produces robust multi-scale features, highlighting key regions while preserving semantic and contextual information for few-shot detection. Subsequently, we design a Deformable Attention Region Proposal Network (DA-RPN) to produce high-quality candidate regions from the enhanced features. Through the collaborative optimization of CASP and DA-RPN, our method significantly improves the accuracy and robustness of few-shot object detection in remote sensing images, particularly in complex scenarios with large number of small objects and messy backgrounds. Experimental results on two large-scale datasets, DIOR and NWPU VHR-10, demonstrate that the proposed model achieves strong performance and clear advantages.

Keywords: remote sensing images; few-shot object detection; dual-attention guidance; content-aware strip pyramid; deformable attention region proposal network

 doi:10.20944/preprints202602.1028.v1

1. Introduction

In recent years, object detection in remote sensing images has made significant progress through deep learning techniques, finding applications in areas such as land resource management [1–3], urban development [4–6], military reconnaissance [7–9], and disaster assessment [10–12]. Compared with natural scene images, remote sensing imagery often exhibits wide coverage, high spatial resolution, complex scene composition, and substantial variations in object scale [13]. These factors make precise object localization and category classification more challenging, while also considerably increasing the cost of acquiring high-quality annotated data. Against this backdrop, few-shot object detection in remote sensing images has gradually gained attention [14,15]. This research direction aims to achieve effective few-shot detection of novel object categories using only a limited number of annotated samples, thereby alleviating the challenges of high annotation costs and imbalanced sample distributions. In recent years, certain advances have been made in related studies [16–18]. Nevertheless, in complex remote sensing scenarios, achieving a balance between model performance and generalization when training with very few samples continues to be a difficult research problem.

In few-shot object detection for remote sensing imagery, current studies mainly focus on two methodological approaches: transfer learning and meta-learning. The latter [19,20] quickly adapts across tasks, providing advantages with very limited samples, but it requires complex training and can be sensitive to scale variations and background noise in remote sensing images. Transfer learning [21,22] typically first trains models using large-scale foundational classes and then fine-tunes them for novel categories, offering a simpler implementation and better compatibility with mainstream detection frameworks, which makes it the dominant approach. However, as shown in Figure 1, large variations in object scale and complex backgrounds in remote sensing images make feature modeling challenging. To handle scale differences [23–25], current approaches often employ feature pyramids, which can introduce redundancy and produce unstable representations in few-shot settings. For complex backgrounds [26,27], attention mechanisms help highlight discriminative regions but may introduce bias and amplify noise when samples are limited. These observations highlight the main challenges of the transfer learning paradigm in few-shot object detection for remote sensing, while also emphasizing the need for further improvements.

(a) (b)

Figure 1. Typical challenging cases in remote sensing object detection: (a) Large variations in object scale, including differences between classes (e.g., bridge vs. vehicle) and within the same class (e.g., vessels of different sizes). (b) Complex background interference, where small objects such as vehicles can be easily obscured by dense urban structures or cluttered surroundings.

Motivated by these challenges, we introduce DAFSDet, a dual-attention guided few-shot object detection model, designed for remote sensing images within the transfer learning framework. Specifically, we introduce the Content-Aware Strip Pyramid (CASP) network as the neck module,

 doi:10.20944/preprints202602.1028.v1

which replaces the conventional feature pyramid network. This module employs the Content-Aware ReAssembly of Features (CARAFE) [28] operator to selectively focus on spatial regions, enhancing the semantic coherence of critical areas. Simultaneously, bidirectional strip convolutions capture long-range, direction-sensitive spatial context, enabling combined spatial and semantic attention. The resulting multi-scale feature maps merge detailed spatial information with enriched semantic content, offering a robust basis for few-shot detection. For the detection head, we further design a Deformable Attention Region Proposal Network (DA-RPN) to supersede the traditional RPN. This module begins with deformable convolutions, which adaptively adjust the receptive field to better match the geometric shapes of targets and improve modeling for complex variations such as deformations and rotations. A spatial attention mechanism then redistributes weights across the feature maps, guiding the model to emphasize foreground target regions while suppressing background noise, thereby enhancing the quality of candidate proposals. Experiments on the DIOR and NWPU VHR-10 datasets demonstrate the model's effectiveness under various few-shot settings.

To summarize, this work presents three main contributions:

1. A dual-attention–guided detection model, DAFSDet, is proposed under the transfer learning paradigm. By combining two specialized modules for multi-scale target perception and background suppression, this approach improves feature learning from limited samples in remote sensing imagery.
2. The Content-Aware Strip Pyramid (CASP) module improves multi-scale feature representation. It combines content-aware upsampling with bidirectional strip convolutions to focus attention on relevant spatial regions and capture long-range contextual information, forming a spatial-semantic attention mechanism. As a result, CASP produces multi-scale features that blend semantic richness with spatial details, creating a solid basis for few-shot object detection.
3. The Deformable Attention Region Proposal Network (DA-RPN) enhances localization accuracy for targets in complex backgrounds. By combining deformable convolutions with a spatial attention mechanism, this network allows the receptive field to conform to target geometries and automatically attend to important foreground regions, effectively reducing background interference and improving the quality of candidate proposals.

The paper is structured as follows. Section II gives an overview of related work, Section III presents the proposed method in detail, Section IV reports and discusses the experimental results, and Section V concludes the study.

2. Related Work

2.1. Few-Shot Object Detection

Few-shot object detection (FSOD) seeks to allow models to correctly identify new categories using only limited annotated samples. Current mainstream approaches can be divided into two categories: meta-learning and transfer learning.

Meta-learning methods simulate the "training-testing" process through constructed meta-tasks, allowing models to learn quickly from limited samples. Meta R-CNN [19] pioneers the application of meta-learning to the characteristics of the region-of-interests, laying the foundation for subsequent research. Following this, FSOD [29] addresses high annotation costs by leveraging knowledge from base categories to support novel classes with limited samples, facilitating fast adaptation and better generalization. FsDetView [30] employs multimodal features to strengthen category representation. To overcome insufficient prototype representation, ICPE [31] develops an information coupling and dynamic aggregation mechanism. VFA [32] additionally applies variational feature aggregation to reduce inter-class confusion and enhance detection stability via task decoupling. Although meta-learning methods demonstrate strong adaptation capabilities, they tend to exhibit substantial computational demands, sensitivity to task design, and reduced robustness in challenging scenarios like remote sensing imagery, where target scales vary greatly and backgrounds are cluttered. These drawbacks

 doi:10.20944/preprints202602.1028.v1

encourage the use of alternative strategies, such as transfer learning-based few-shot detection, which can utilize pre-trained representations more efficiently.

By comparison, transfer learning approaches initially undergo pre-training on base categories before adapting to novel classes via fine-tuning. TFA [21] shows that fine-tuning only the detection head yields strong results, forming a foundation for this research direction. Building on this, FSCE [33] introduces a contrastive loss to improve inter-class separability. DeFRCN [22] incorporates a gradient decoupled layer for multi-stage processing and a prototypical calibration block for multi-task adjustment, enhancing feature representation and classification in few-shot scenarios. PB-FSOD [34] modifies FSOD with a proposal balance refinement strategy that sequentially adjusts bounding boxes and exposes the RPN to novel classes, improving the quality and diversity of region proposals and supporting better detection of unseen categories. Overall, these transfer learning-based methods leverage large-scale base-class data to acquire robust feature representations and generalize effectively to novel classes with limited additional samples. Their simplicity and compatibility with mainstream detection frameworks contribute to their widespread adoption in few-shot object detection.

2.2. Few-Shot Object Detection in Remote Sensing Images

Few-shot detection has made significant progress in natural scenes. However, its direct application to remote sensing imagery still faces several challenges. These difficulties largely arise from the intrinsic properties of remote sensing data, including wide variations in object scales and complex background interference. To tackle these persistent issues, numerous studies have been conducted.

To tackle the issues caused by scale variations, several studies [35–37] have investigated different strategies. Their approaches include constructing feature pyramids [38], applying metric learning, and integrating attention mechanisms. FSODM [35] adapts YOLOv3 [39] for few-shot object detection in remote sensing images. The model produces multi-scale feature representations by using a meta-feature extractor combined with class-specific feature reweighting. It also employs multi-scale bounding box predictions to accommodate novel classes with very limited annotated samples, effectively mitigating scale variations. SAM-FSOD [36] introduces a shared attention mechanism. The attention maps learned at multiple scales during base-class training are reused during the fine-tuning stage to assist feature extraction. This design provides explicit spatial guidance for locating novel objects and improves detection performance on targets with diverse scales. Although existing approaches alleviate scale variation to some extent, they remain ineffective at precisely delineating large objects and often suffer from missed detections or incorrect merging in densely populated scenes. These issues reduce detection reliability when object scales are highly unbalanced. To address this limitation, we introduce a Content-Aware Strip Pyramid (CASP) module that strengthens multi-scale feature learning in few-shot object detection. The module integrates content-aware upsampling with bidirectional strip convolution to model spatial semantics while preserving long-range contextual dependencies. This process generates multi-scale features that integrate rich semantic information with spatial details. Compared to existing attention-based methods, CASP provides stronger robustness to scale variations.

To overcome interference from complex backgrounds, related studies [40–43] have primarily focused on designing powerful feature enhancement or contrastive learning strategies that strengthen discrimination between foreground objects and background clutter. Among these, MSOCL [40] integrates multi-scale object contrastive learning. Within a Siamese network structure to address background complexity, which designs a multi-scale instance feature module to capture contrastive information and leverages Contrastive Multi-Scale Proposals (CMSP) to fully exploit scale information, thereby effectively enhancing the model's feature representation capability and data adaptability. On the other hand, researchers propose the P-CNN [41], which generates category-aware prototypes via a Prototype Learning Network (PLN) and optimizes proposal generation through a Prototype-Guided Region Proposal Network (P-G RPN). While enhancing target-background discrimination, these methods remain limited in two aspects: difficulty in distinguishing highly similar backgrounds and restricted background suppression for extreme-scale or irregular objects due to rigid multi-scale and prototype designs. In contrast, in this work, we propose DA-RPN that integrates deformable convolutions with

 doi:10.20944/preprints202602.1028.v1

the spatial attention mechanism. This design enables dynamic adjustment of feature perception to match target geometry, while directing attention toward salient foreground regions, thereby suppressing background interference and improving the quality of candidate proposals. By combining learnable attention with deformable feature adaptation, our method provides greater flexibility and robustness in complex scenes compared to existing contrastive or prototype-based approaches.

3. Methods

3.1. Preliminaries

In the few-shot object detection (FSOD) task, it typically assums that there are two types of categories: base classes (C_b) and novel classes (C_n), satisfying $C_b \cap C_n = \emptyset$. Base classes are associated with abundant annotations and constitute the dataset $D_b = \{(x_i, y_i)\}_{i=1}^{N_b}$, while novel classes contain only a small number of annotated samples, forming the dataset $D_n = \{(x_j, y_j)\}_{j=1}^{N_n}$, where x refers to a remote sensing image, and y corresponds to its object-level annotations, including category labels and bounding box coordinates. This task focuses on learning a detection function $f_\theta(x)$ that performs object recognition and localization across both previously seen categories and newly introduced ones within a unified framework. To reduce overfitting under few-shot conditions, we employ a two-stage transfer learning scheme. The network is initially trained on a base-class dataset D_b to acquire generic visual representations, and is subsequently refined using a balanced dataset D_f that includes both base and novel categories, allowing fast adaptation to unseen classes while maintaining generalization ability. For each novel class, limited to K annotated samples and considering N novel classes, this setup corresponds to an N-way K-shot few-shot detection scenario. To reduce overfitting, the backbone network parameters are generally kept frozen during fine-tuning.

3.2. Overall Network Architecture

The proposed DAFSDet serves as a dual-attention guided few-shot object detection framework designed for remote sensing imagery. It adopts the classical backbone–neck–detection head architecture, with improvements in the neck and detection head aimed at enhancing feature representation and the quality of candidate regions under few-shot conditions.

- Backbone: Following [21,40], we adopt ResNet-101 to generate multi-level feature representations $\{C_2, C_3, C_4, C_5\}$, capturing hierarchical semantic information to serve as the foundation for detection.
- Neck: The CASP module combines content-aware upsampling with bidirectional strip convolution to form a collaborative attention mechanism across spatial and semantic dimensions. It improves multi-scale feature representations and strengthens the model's ability to capture long-range context. As a result, the network produces multi-scale features that are semantically rich and spatially detailed, providing a solid foundation for few-shot detection.
- Detection Head: The DA-RPN uses deformable convolutions to adjust to the geometric shapes of targets and applies spatial attention to dynamically emphasize key feature regions. This design effectively suppresses complex background interference and enhances the response of the foreground target, thus generating candidate target regions with more accurate localization and higher quality.

Training proceeds in two stages. During base training, the backbone, a Feature Pyramid Network (FPN) with CARAFE as the upsampling operator, and the RPN learns general representations from base-class data. For brevity, this configuration is hereafter referred to as FPN_CARAFE. During few-shot fine-tuning, the complete DAFSDet, which integrates CASP and DA-RPN, undergoes training on a limited set of novel-class samples to enhance object recognition and localization.

As shown in Figure 2, the input image is processed by the backbone, refined by CASP for multi-scale enhancement, and then passed through DA-RPN to produce high-quality region proposals. The following sub-sections detail the design of CASP and DA-RPN.

 doi:10.20944/preprints202602.1028.v1

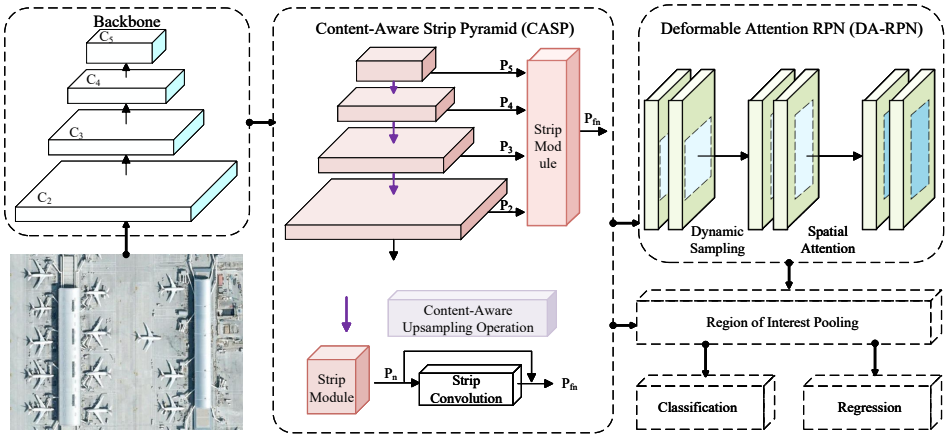

Figure 2. Overall framework of our proposed DA-FSDeT.

3.3. Content-Aware Strip Pyramid (CASP)

Remote sensing images exhibit significant object scale variations and complex backgrounds, posing substantial challenges for multi-scale feature modeling. Although existing detection frameworks based on Feature Pyramid Networks (FPNs) can partially alleviate scale discrepancies, their fixed upsampling operations lack content awareness, and the cross-level feature fusion process does not explicitly model spatial and semantic attention. As a result, discriminative features are easily weakened under few-shot settings. To address these limitations, we propose the Content-Aware Strip Pyramid (CASP) module, which enhances multi-scale feature representation by incorporating content-driven spatially selective attention and direction-sensitive contextual modeling, thereby improving robustness and discriminative ability for detecting objects under few-shot scenarios.

As shown in Figure 2, the CASP module first generates feature maps at different resolutions via content-aware upsampling, where reassembly kernel weights are adaptively predicted from local semantic content. This process incorporates spatially selective attention during feature reconstruction, helping to preserve object structures and boundary details. Building on this, CASP employs bidirectional strip convolutions to capture contextual information. Decomposed horizontal and vertical convolutions capture long-range directional dependencies, enabling the network to perceive linear and regular structures common in remote sensing images. The extracted features are normalized, combined, and integrated with the original features using residual connections, thereby providing direction-sensitive contextual attention while preserving semantic consistency. The feature transformation process of CASP is formally described in the following.

Given the set of multi-scale feature maps $\{C_2, C_3, C_4, C_5\}$ output by the backbone network, an initial multi-scale feature map is first generated through feature fusion and upsampling via the FPN_CARAFE backbone network $\{P_2, P_3, P_4, P_5\}$. To meet the need for additional scales in remote sensing object detection, P_6 is derived from P_5 using a convolutional layer with stride 2. This operation can be formally represented as:

$$\{P_l\} = \mathcal{F}_{\text{FPN\_CARAFE}}(\{C_l\}), \quad l \in \{2,3,4,5\} \tag{1}$$

where $\mathcal{F}_{\text{FPN\_CARAFE}}$ denotes the forward propagation function of the FPN_CARAFE network, and l represents the pyramid level.

Then, the feature map at each level $P_l \in \mathbb{R}^{C \times H_l \times W_l}$ is independently fed into the corresponding strip convolution module for enhancement. This module first performs parallel horizontal and vertical strip convolutions through bidirectional decomposed convolutions to capture spatial contextual information along orthogonal directions. The convolution outputs for the two directions, X_h^l and X_v^l, are respectively computed as:

$$X_h^l = \text{Conv}_{1\times K}(P_l), \quad X_v^l = \text{Conv}_{K\times 1}(P_l) \tag{2}$$

 doi:10.20944/preprints202602.1028.v1

where $\text{Conv}_{1\times K}$ and $\text{Conv}_{K\times 1}$ indicate convolution operations with kernel sizes of $1 \times K$ and $K \times 1$, respectively. X_h^l and X_v^l denote the output feature maps from the horizontal and vertical strip convolution branches at the l-th level.

After generating the bidirectional features, the module applies batch normalization and ReLU activation independently to the outputs of the two branches, enabling feature fusion while maintaining their directional specificity. In particular, following separate normalization of X_h^l and X_v^l, we obtain:

$$\hat{X}_h^l = \text{ReLU}(\text{BN}(X_h^l)) \tag{3}$$

$$\hat{X}_v^l = \text{ReLU}(\text{BN}(X_v^l)) \tag{4}$$

where $\text{BN}(\cdot)$ represents the batch normalization operation, it independently normalizes the feature maps from each branch, stabilizing their means and variances, which accelerates training convergence and improves model performance. This branch-wise normalization is essential because horizontal and vertical features may follow different distributions. By processing them separately, their directional characteristics are preserved, avoiding potential conflicts that could occur if the features were concatenated directly. $\text{ReLU}(\cdot)$ denotes the Rectified Linear Unit activation, which applies nonlinear transformations to the network and enhances its representational power.

Next, the feature maps from both branches are merged along the channel axis to produce the fused feature map.

$$X_{\text{cat}}^l = \text{Concat}(\hat{X}_h^l, \hat{X}_v^l) \tag{5}$$

where $\text{Concat}(\cdot)$ indicates concatenation along the channel axis. This operation combines the normalized and activated horizontal feature map $\hat{X}_h^l$ with the vertical feature map $\hat{X}_v^l$, producing a fused feature map X_{cat}^l that captures bidirectional spatial context. If each branch outputs C' channels, then X_{cat}^l has a total of $2C'$ channels.

Finally, a residual connection adds the fused feature X_{cat}^l to the input P_l of the strip convolution module, producing the final enhanced feature $\hat{P}_l$ for that level:

$$\hat{P}_l = X_{\text{cat}}^l + P_l \tag{6}$$

where P_l denotes the original feature map at the l-th level produced by FPN_CARAFE and fed into the strip convolution module.

This residual connection helps maintain stable gradient flow, reduces the vanishing gradient problem in deep networks, and performs identity mapping, ensuring that the original feature information is preserved during enhancement.

By integrating the spatial attention generated from content-aware upsampling with the contextual information captured by strip convolutions, CASP forms a combined spatial–semantic attention mechanism during multi-scale feature fusion. This mechanism adaptively focuses on important object regions, substantially improving the discriminability and robustness of multi-scale features in few-shot scenarios.

3.4. Deformable Attention Region Proposal Network (DA-RPN)

Although the CASP module effectively enhances multi-scale feature representations, candidate regions generated by the RPN can still be contaminated in complex backgrounds. Standard convolutions are sensitive to geometric deformations, and the diverse target shapes and cluttered backgrounds in remote sensing images often introduce substantial background noise, reducing feature representation and affecting subsequent classification and regression. To address this issue, existing studies typically attempt to suppress background interference by enhancing feature representations or incorporating attention mechanisms. For example, some methods [32,41] design multi-scale contrastive learning modules to improve feature discriminability. However, such pairwise contrastive learning-based strategies are often limited in complex scenes with dense or occluded targets. Other approaches [40,43]

 doi:10.20944/preprints202602.1028.v1

optimize region proposals using prototype-based methods, yet fixed prototypes exhibit insufficient adaptability to the diverse shapes of novel-class objects.

To address the aforementioned limitations, we design the Deformable Attention Region Proposal Network (DA-RPN), as illustrated in Figure 3. This module first employs the dynamic sampling mechanism of deformable convolutions to adaptively adjust convolutional sampling points according to the geometric deformations of targets, thereby enhancing feature extraction for objects with diverse or irregular shapes. Subsequently, a spatial attention subnetwork is applied to the geometrically enhanced feature maps, reweighting features along the spatial dimension to emphasize key foreground regions and suppress background interference. By combining geometric adaptability with spatial attention, this design produces more accurate and discriminative candidate region features, significantly improving classification and regression performance.

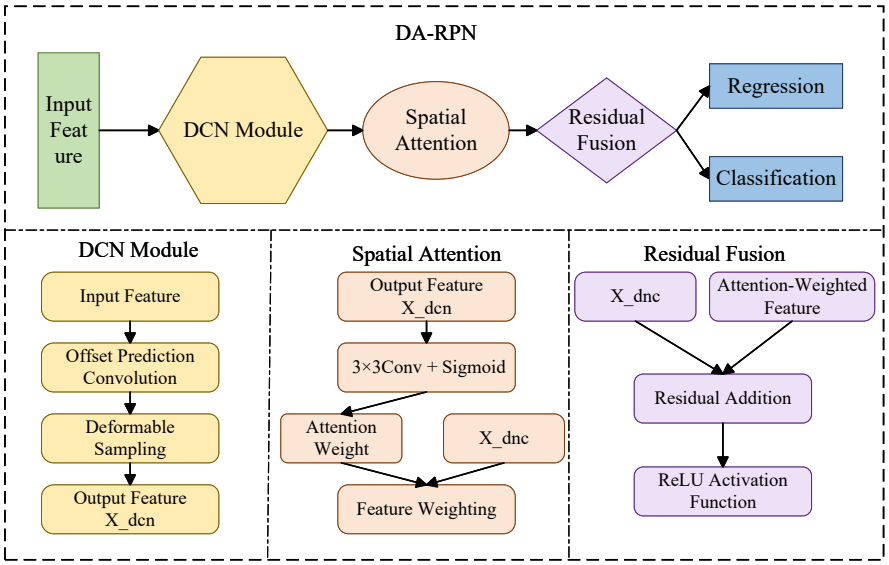

Figure 3. Detailed diagram of the DA-RPN module.

Conventional convolution uses fixed, regular sampling grids, which complicates adjusting to target geometries in imagery captured by remote sensing sensors. Given an input image, the output of standard convolution at position P_0 is:

$$y(P_0) = \sum_{P_n \in \mathcal{R}} W(P_n) \cdot X(P_0 + P_n) \tag{7}$$

where $\mathcal{R}$ defines the regular receptive field of the convolution kernel.

In deformable convolution, each sampling point P_n is assigned an adaptive offset ΔP_n and a modulation scalar ΔM_n, resulting in the following output:

$$y(P_0) = \sum_{P_n \in \mathcal{R}} W(P_n) \cdot X(P_0 + P_n + \Delta P_n) \cdot \Delta M_n \tag{8}$$

Both the offset ΔP_n and the modulation scalar ΔM_n are derived from the input feature map X using parallel convolution layers. This dynamic sampling allows the convolution kernel to modify its sampling positions according to the semantic information present in the key regions of the target. By determining suitable offsets, the network guarantees that sampling points consistently span the main regions of the target, thus markedly improving feature representation for geometrically deformed objects.

Building on the improved geometric awareness provided by deformable convolution, a spatial attention module is employed to emphasize foreground features while reducing background interference. This module dynamically selects features by adjusting the weights across spatial positions of the

 doi:10.20944/preprints202602.1028.v1

feature map. Given the deformable convolution output X_{dcn}, an element-wise spatial attention map A is constructed as follows:

$$A = \{a_{ij} \mid a_{ij} \in [0,1],\ i = 1,\ldots,H,\ j = 1,\ldots,W\} \tag{9}$$

where each attention weight a_{ij} is obtained through a 3×3 convolution, followed by a Sigmoid activation function:

$$a_{ij} = \sigma\Big(\text{Conv}_{3\times3}(X_{\text{dcn}})_{ij}\Big) \tag{10}$$

where X_{dcn} denotes the output feature produced by deformable convolution, and σ represents the Sigmoid function, constraining the attention weights $A \in [0,1]^{1\times H\times W}$.

Each element a_{ij} corresponds to the attention response at spatial position (i,j), reflecting how strongly the model attends to that location in the feature map. Higher values of a_{ij} indicate greater relevance to the detection task and guide the model to prioritize foreground regions over background areas in later stages.

The attention map is then applied to reweight spatial features, as formulated below:

$$X_{\text{attended}} = X_{\text{dcn}} \odot A \tag{11}$$

where $\odot$ denotes element-wise multiplication. The attention map A is expanded across channels, allowing each spatial weight a_{ij} to modulate all feature channels at position (i,j).

This operation functions as an adaptive feature filtering process, guiding the model toward feature responses that are most informative for candidate target regions. To maintain gradient stability and retain the initial feature information, we incorporate a residual connection in the feature enhancement module. The resulting fused feature X_{fused} is calculated as:

$$X_{\text{fused}} = X_{\text{dcn}} \odot (\mathbf{1} + A) \tag{12}$$

where $\mathbf{1}$ is an all-ones matrix with the same dimensions as A.Here, $\mathbf{1}$ denotes a matrix filled with ones, matching the dimensions of A.

This design allows the model to retain original feature information even in areas where attention weights are low, which helps preserve the completeness of feature representation while contributing to stable gradient propagation during training. The benefits of this design manifest in several ways. Firstly, even in regions where attention weights are minimal, the model preserves the original feature information, thereby maintaining the completeness and consistency of the feature representation. Second, during training, gradients can be directly backpropagated through the residual connection, effectively mitigating the vanishing gradient problem. Finally, this mechanism guides the model to capture the difference between input features and the desired outputs, thereby easing the optimization process.

4. Experimental Results and Analysis

In this section, we carry out a comprehensive evaluation of the DAFSDet model on two commonly used open-access remote sensing image datasets. To comprehensively validate the model's effectiveness in remote sensing few-shot scenarios, this paper conducts comparative experiments with both typical few-shot detectors and remote sensing few-shot detectors, along with ablation studies. Finally, the detection capability of the proposed method for novel classes in complex scenarios is further demonstrated through visualization analysis.

4.1. Datasets and Evaluation Metrics

DIOR [44] is a large-scale optical remote sensing image object detection benchmark proposed by research teams from Zhengzhou Institute of Surveying and Mapping and Northwestern Polytechnical University. The dataset includes a total of 23,463 images and 192,472 annotated instances. The training, validation, and test subsets consist of 5,862, 5,863, and 11,738 images, respectively. Each image

 doi:10.20944/preprints202602.1028.v1

measures 800 × 800 pixels, and the resolution varies between 0.5 and 30 meters. This dataset includes 20 object categories, each represented with a standardized abbreviation: airplane (AP), airport (AT), baseball field (BF), basketball court (BC), bridge (BR), chimney (CH), dam (DM), expressway service area (ESA), expressway toll station (ETS), golf course (GC), ground track field (GTF), harbor (HB), overpass (OP), ship (SH), stadium (ST), storage tank (STK), tennis court (TC), train station (TS), vehicle (VE), and windmill (WM).

NWPU VHR-10 [45] developed by Northwestern Polytechnical University, serves as a benchmark for studies on remote sensing object detection. Image resolutions vary between 800 × 800 and 2000 × 2000 pixels, which allows the dataset to support the recognition and categorization of several distinct target classes within high-resolution imagery. The NWPU VHR-10 dataset consists of ten object classes, including airplane, ship, car, bridge, building, road, parking lot, airport runway, storage yard, and harbor. Image dimensions vary roughly between 500 and 1200 pixels. The dataset is organized into two subsets: a positive subset comprising 650 images, each containing at least one annotated target, and a negative subset with 150 images that do not include any of the specified object classes.

For the DIOR dataset, the partitioning strictly follows the configuration described in [41], utilizing the four pre-defined base/novel class split schemes specified in that work. In each scheme, there are 15 base categories and 5 novel categories, and the exact composition of these classes is presented in Table 1. For the NWPU VHR-10 dataset, following the protocol in [46], three categories—airplane, baseball field, and tennis court—are treated as novel classes. The remaining categories are considered base classes to maintain consistency with prior work.

We assess the detection results using mean Average Precision (mAP), reporting metrics individually for base and novel classes. These are represented as AP_{base} and AP_{novel}, respectively. The metric AP_{base} evaluates how accurately the model detects base classes with sufficient samples, indicating its capability to preserve knowledge acquired from these classes. The metric AP_{novel} evaluates how effectively the model detects novel classes given limited annotated examples, and acts as an important measure of few-shot detection performance.

Table 1. Dataset Splits for Few-shot Object Detection.

Split	Novel Classes	Base
1	BF, BC, BR, CH, SH	Rest
2	AP, AT, ETS, HB, GTF	Rest
3	DM, GC, STK, TC, VE	Rest
4	ESA, OP, ST, TS, WM	Rest

4.2. Experimental Setup

The DAFSDet model is built upon the SAE-FSDet [46] framework, with a ResNet-101 [47] backbone pre-trained on large-scale data. During the initial training on base classes, DAFSDet is optimized for 18 epochs. The training begins with a learning rate of 0.002, following a step-wise schedule where the rate is reduced after the 12th and 16th epochs. Training is performed with a batch size of 2, using the Stochastic Gradient Descent (SGD) optimizer for parameter updates. In the fine-tuning stage, the model is optimized for 108 epochs starting with a learning rate of 0.0012, which is reduced using a step decay schedule at the 95th epoch. The training continues with a batch size of 2, employing the SGD optimizer for parameter updates. All experiments were carried out using PyTorch on one NVIDIA RTX 4090 GPU equipped with 24 GB of memory.

4.3. Experimental Results and Comparisons

4.3.1. Experimental Results on the DIOR Dataset

A thorough evaluation of the model was performed by comparing it against multiple few-shot object detection methods, such as Meta R-CNN [19], FsDetView [30], TFA [21], and P-CNN [41].

 doi:10.20944/preprints202602.1028.v1

Table 2 presents the results, where DAFSDet attains the highest accuracy in the majority of settings. Across various shot numbers in Splits 2, 3, and 4, it consistently surpasses competing approaches, indicating reliable robustness and good generalization across different class partitions. When the number of shots ranges from 3 to 20, all methods show gradual performance growth. DAFSDet, in particular, demonstrates consistent and notable gains. The model performs especially well on categories containing small or visually complex objects, reflecting its ability to leverage scarce annotated samples while retaining distinguishing features across various classes.

Table 2. Performance Comparison of Methods Across Different Splits and Shots.

Split	Method	3-shot	5-shot	10-shot	20-shot
1	Meta RCNN[19]	12.02	13.09	14.07	14.45
	FsDetView[30]	13.19	14.29	18.02	18.01
	TFA w/cos[21]	16.07	15.36	16.45	18.93
	P-CNN[41]	18.00	22.80	27.60	29.60
	FSOD[29]	15.94	20.27	24.22	28.16
	FSCE[33]	27.91	28.60	33.05	37.55
	MSOCL[40]	24.97	27.27	33.37	39.22
	ICPE[31]	11.68	12.34	12.95	14.33
	VFA[32]	21.94	21.27	23.32	24.28
	SAE-FSDT[46]	**28.80**	**32.40**	**37.09**	**42.46**
	SAE-FSDT*[46]	25.08	28.91	35.57	41.77
	DA-FSDeT(Ours)	27.22	29.54	33.86	39.15
2	Meta RCNN[19]	8.84	10.88	14.90	16.71
	FsDetView[30]	10.83	9.63	13.57	14.76
	TFA w/cos[21]	6.81	7.53	8.93	11.05
	P-CNN[41]	14.50	14.90	18.90	22.80
	FSOD[29]	9.35	9.73	14.84	16.20
	FSCE[33]	13.17	14.07	15.79	20.93
	MSOCL[40]	13.31	13.40	15.00	18.15
	ICPE[31]	10.92	10.56	12.39	13.18
	VFA[32]	12.10	12.70	14.72	15.47
	SAE-FSDT[46]	13.99	15.65	17.41	21.34
	SAE-FSDT*[46]	12.85	14.04	14.53	20.78
	DA-FSDeT(Ours)	**15.83**	**18.57**	**22.07**	**25.52**
3	Meta RCNN[19]	9.10	12.29	11.96	16.14
	FsDetView[30]	7.49	12.61	11.49	17.02
	TFA w/cos[21]	8.73	9.31	12.19	16.97
	P-CNN[41]	16.50	18.80	23.30	28.80
	FSOD[29]	10.40	10.74	12.26	11.52
	FSCE[33]	15.59	16.24	23.75	28.89
	MSOCL[40]	13.11	15.07	23.39	27.44
	ICPE[31]	10.56	11.21	12.38	13.08
	VFA[32]	11.97	13.19	15.45	17.61
	SAE-FSDT[46]	16.74	19.07	28.44	29.88
	SAE-FSDT*[46]	14.04	16.48	26.65	28.42
	DA-FSDeT(Ours)	**18.21**	**20.78**	**29.76**	**31.44**
4	Meta RCNN[19]	13.94	15.84	15.07	18.17
	FsDetView[30]	14.28	15.95	15.37	16.96
	TFA w/cos[21]	9.54	13.82	13.82	16.61
	P-CNN[41]	15.20	17.50	18.90	25.70
	FSOD[29]	11.84	12.98	17.17	18.46
	FSCE[33]	17.45	20.42	22.22	24.96
	MSOCL[40]	10.40	12.29	16.64	22.67
	ICPE[31]	14.45	14.52	15.95	15.61
	VFA[32]	15.52	17.76	18.62	20.05
	SAE-FSDT[46]	17.27	20.48	22.69	26.75
	SAE-FSDT*[46]	14.87	16.92	20.21	24.96
	DA-FSDeT(Ours)	**17.82**	**22.63**	**28.15**	**30.51**

* All results are obtained from our own reimplementation under identical experimental configurations.

 doi:10.20944/preprints202602.1028.v1

Among all experimental splits, Split 1 shows a different trend: DAFSDet performs marginally worse than the baseline at 10-shot and 20-shot levels, yet it remains competitive in very low-shot scenarios like 3-shot and 5-shot. This discrepancy is mainly linked to specific novel categories, including ships and basketball courts. In remote sensing images, ships tend to appear in dense clusters and are usually very small. As a result, they are prone to interference from surrounding objects or the background. Basketball courts generally span large and uniform areas with minimal texture, making them hard to differentiate from other human-made structures. When more training samples are available, the baseline model can adjust the category decision boundaries more accurately in both the classification head and the region proposal stage, resulting in improved performance at higher-shot levels. DAFSDet, on the other hand, focuses on enhancing features through attention mechanisms while suppressing background noise, which makes it particularly robust when training data are very limited.

To provide a clearer comparison with the baseline, we present the detection results of DAFSDet on the DIOR test set, highlighting its improved performance. Figure 4 provides a qualitative visualization of detection results obtained by the baseline model (top row) and DAFSDet (bottom row) in Split 2 with 20 annotated samples per class. In this setting, the performance difference between the two methods becomes more evident, allowing the detection advantages of the proposed approach to be more clearly observed. Figure 4 (a) and (b) illustrate that DAFSDet alleviates redundant detections and yields tighter localization results relative to the baseline approach. In addition, the examples in Figure 4 (c) indicate that objects previously assigned to incorrect categories can be correctly recognized by the proposed method, whereas Figure 4 (d) highlights its ability to recover targets that are missed by the baseline detector. Overall, these visualization results indicate that DAFSDet improves detection accuracy and completeness by jointly enhancing localization precision, classification reliability, and robustness to missed detections.

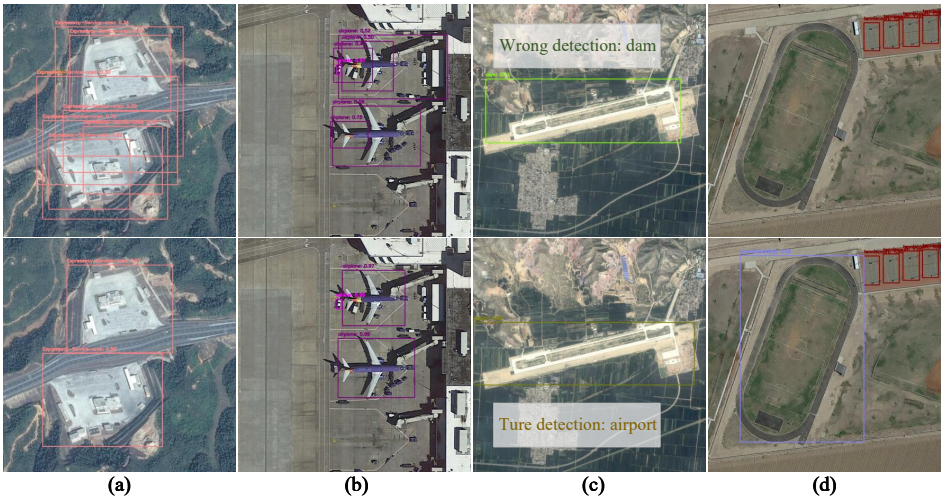

Figure 4. Visualization of the compared results between the baseline (top row) and DAFSDet (bottom row) on the DIOR test set in the split 2 under a 20-shot setting.

4.3.2. Experimental Results on the NWPU VHR-10 Dataset

We also conduct comparative experiments on the NWPU VHR-10 dataset, as illustrated in Table 3, which can be observed that the proposed DAFSDet method outperforms all comparative methods across all shot settings. It reveals that whether in extremely limited number of samples with 3-shot or relative common ones with 20-shot, the proposed model can obtain satisfactory detection results. Furthermore, we also visualize the detection results of DAFSDet on the NWPU VHR-10 dataset, as shown in Figure 5, it further demonstrates the extraordinary localization ability in extreme complex scenarios.

 doi:10.20944/preprints202602.1028.v1

Table 3. Performance Comparison on the NWPU VHR-10 Benchmark.

Method	3-shot	5-shot	10-shot	20-shot
Meta R-CNN[19]	20.51	21.77	26.98	28.24
FsDetView[30]	24.56	29.55	31.77	32.73
TFA w/cos[21]	16.17	20.49	21.22	21.57
P-CNN[41]	41.80	49.17	63.29	66.83
FSOD[29]	41.80	49.17	63.29	66.83
FSCE[33]	10.95	15.13	16.23	17.11
MSOCL[40]	41.63	48.80	59.97	79.60
ICPE[31]	6.10	9.10	12.00	12.20
VFA[32]	13.14	15.08	13.89	20.18
SAE-FSDT[46]	57.96	59.40	71.02	85.08
DA-FSDeT(Ours)	**60.05**	**60.80**	**72.50**	**85.56**

Figure 5. Visualization of the detection results on NWPU VHR-10 in the split 2 under a 20-shot setting.

4.4. Ablation Study

To thoroughly manifest the effectiveness of the designed modules in our proposed model, we design and conduct a series of ablation experiments on the DIOR dataset with Split 2 setting.

4.4.1. Effect of CASP

The CASP module is designed to enhance multi-scale feature representations by integrating content-aware upsampling with strip-based contextual modeling. As shown in Table 4, introducing CASP into the baseline consistently improves detection performance in all shot settings. Specifically, the mAP increases by 2.91 and 2.47 points in the 3-shot and 5-shot settings, respectively, while larger gains of 5.64 and 4.03 points are observed in the 10-shot and 20-shot settings. These findings indicate that CASP can effectively ease the difficulties arising from pronounced scale discrepancies and limited contextual representation in remote sensing images, and its benefits become increasingly evident as the amount of training data grows.

Table 4. Detection Performance Across Datasets Under Varying Shot Numbers.

CASP	DA-RPN	3-shot	5-shot	10-shot	20-shot	Params	FLOPs
-	-	12.85	14.04	14.53	20.78	60.89	181.13
√	-	15.76	16.51	20.17	24.81	68.65	202.02
-	√	13.62	15.36	18.03	23.08	60.75	175.79
√	√	**15.83**	**18.57**	**22.07**	**25.52**	68.65	202.15

4.4.2. Effect of DA-RPN

The DA-RPN is designed to enhance region proposal quality by incorporating geometric adaptability together with spatial attention mechanisms. When applied as an independent component, it

 doi:10.20944/preprints202602.1028.v1

brings stable improvements compared with the baseline across all shot settings. Notably, the mAP increases by 3.50 points under the 10-shot setting and by 2.30 points under the 20-shot setting. While the overall gain is somewhat smaller than that of CASP, the findings suggest that DA-RPN successfully mitigates background interference and refines proposal localization, resulting in more consistent performance, particularly in the 10-shot setting.

4.4.3. Combined Effect

Combining CASP with DA-RPN allows the model to achieve superior performance across different shot settings. Relative to the baseline, the mAP increases by 2.98, 4.53, 7.54, and 4.74 points for the 3-shot, 5-shot, 10-shot, and 20-shot settings, respectively. The findings confirm that the two modules complement each other: CASP enhances multi-scale feature discriminability, whereas DA-RPN improves the precision of region proposals. The combined effect of these modules markedly improves the model's resilience to variations in object scale and complex backgrounds, addressing key challenges in limited-sample remote sensing detection tasks.

Table 4 illustrates that DA-RPN adds negligible computational cost while steadily enhancing detection performance, demonstrating its efficiency. The CASP module results in a modest rise in parameters and FLOPs because of the enriched multi-scale contextual modeling. When used together, the total computational cost is largely unchanged relative to CASP alone, while achieving the highest performance, indicating a beneficial balance between accuracy and efficiency.

4.5. Failure Case Analysis

While DAFSDet demonstrates strong performance in most remote sensing scenarios, some challenges persist in highly complex cases. As shown in Figure 6, the failure instances can be grouped into two primary categories.

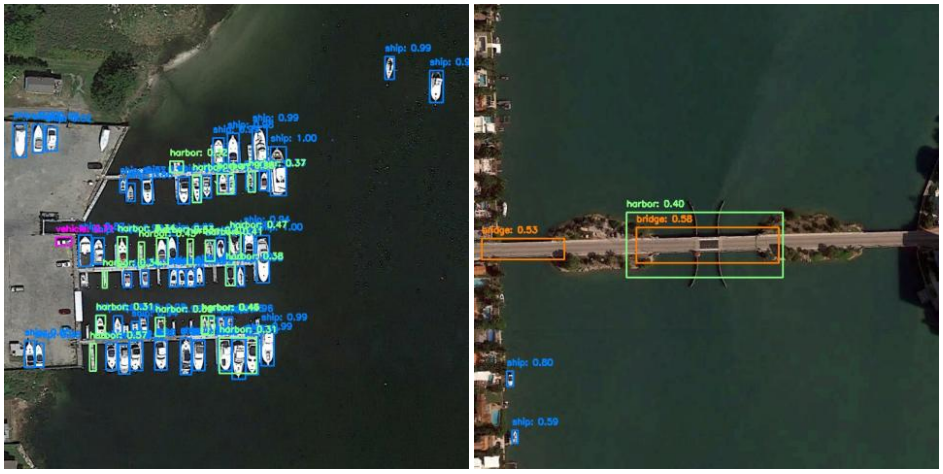

Figure 6. Visualization of failure cases on the DIOR test set. The figure displays four object categories: ships (blue), harbors (green), bridges (orange), and vehicles (purple). Each bounding box includes the predicted category along with its confidence score (ranging from 0 to 1). Overlapping boxes indicate multiple predictions for the same region; missed detections correspond to objects without any bounding box, while false detections occur when the predicted category does not match the ground truth.

- **Performance degradation in densely distributed small-object scenes.** In typical remote sensing scenarios, targets such as ships often exhibit high density, small scale, and compact spatial arrangements. Due to extremely close distances between objects and frequent boundary overlaps, the proposed method may still suffer from missed detections or redundant predictions during region proposal generation and subsequent classification. In particularly crowded areas, the discriminative features of small targets are easily overwhelmed by neighboring objects or background clutter, leading to lower detection confidence or missed detections.

 doi:10.20944/preprints202602.1028.v1

- **Confusion among visually similar categories.** Objects such as bridges, harbors, and overpasses share similar geometric structures and texture patterns when observed from an aerial perspective. In few-shot settings, the scarcity of category-specific samples hinders the model from acquiring adequately discriminative features, resulting in confusion between classes. As shown in Figure 6, some bridge regions are misclassified as harbors or overpasses, indicating that fine-grained category discrimination remains challenging when visual differences between classes are subtle.

These failure cases suggest that while the CASP and DA-RPN effectively enhance multi-scale feature representation and highlight foreground regions, the model still faces challenges in extremely dense target distributions and high inter-class similarity. Future work may explore incorporating instance-level relational modeling or stronger category-discriminative constraints to further improve robustness in complex remote sensing scenarios.

5. Conclusions

To tackle the main difficulties in limited-sample remote sensing object detection—such as multi-scale variations, geometric deformations, and complex backgrounds—this paper introduces a dual-attention-guided model, DAFSDet. The model consists of two core modules: CASP and DA-RPN. CASP replaces the traditional feature pyramid network and generates robust multi-scale feature maps. It integrates content-aware upsampling and bidirectional strip convolutions with spatial and semantic attention, preserving key regions and capturing long-range contextual dependencies to provide a strong feature representation for few-shot object detection. DA-RPN enhances candidate region quality by combining deformable convolutions for adaptive geometric perception with spatial attention to focus on foreground targets while suppressing background noise. Extensive experiments on the DIOR and NWPU VHR-10 datasets show that DAFSDet achieves state-of-the-art performance in most few-shot settings, with remarkable advantages in scenarios with extremely limited samples. Ablation studies confirm the individual and collaborative effectiveness of the two designed modules. Future work will focus on addressing failure cases, such as dense small objects and visually similar categories, by incorporating prototype-aware, class-discriminative, or context-enhancing modules to further improve robustness and discriminative ability in complex scenes.

Author Contributions: Conceptualization, G.G. and Z.Z.; Methodology, G.G. and Z.Z.; Software, G.G., Z.Z. and W.Z.; Validation, G.G. and Z.Z.; Formal analysis, G.G. and Y.S.; Investigation, G.G., Z.Z. and W.Z.; Resources, Y.D. and J.X.; Data curation, Z.Z. and W.Z.; Writing—original draft, Z.Z.; Writing—review and editing, G.G., Z.Z. and Y.S.; Visualization, Z.Z.; Supervision, G.G.; Project administration, G.G.; Funding acquisition, G.G. 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) (Nos. 62301623, 62472463, and 61873293), the Henan Key Research and Development Project (No. 241111220700), the Incubation Program for Young Master Supervisors of Zhongyuan University of Technology (No. SD202416), the General International Scientific and Technological Cooperation Project of Henan Province (No. 252102521049), the Key Scientific Research Project of Colleges and Universities in Henan Province (No. 25A620001), and the Key Project of the Natural Science Foundation of Zhongyuan University of Technology (No. K2026ZD021).

Data Availability Statement: Researchers can obtain the DIOR dataset through the Mendeley Data repository: https://data.mendeley.com/datasets/vvrhgbr643/1 . Meanwhile, the NWPU VHR-10 dataset (COCO-formatted version) is accessible on GitHub: https://github.com/chaozhong2010/VHR-10_dataset_coco.

Acknowledgments: The authors sincerely thank the editors and anonymous reviewers for their valuable time, thoughtful comments, and helpful suggestions, which have greatly enhanced the quality of this manuscript.

Conflicts of Interest: The authors report that there are no conflicts of interest related to this work.

 doi:10.20944/preprints202602.1028.v1

Abbreviations

The abbreviations employed throughout this manuscript are listed as follows:

RPN	Region Proposal Network
FPN	Feature Pyramid Network
FSOD	Few-shot object detection
mAP	Mean Average Precision
SGD	Stochastic Gradient Descent
GPU	Graphics Processing Unit

Appendix A. Task Setting and Dataset Characteristics of DIOR

Appendix A.1. Detection Task

This work focuses on detecting objects in aerial remote sensing images. For a given input image, the task involves detecting all objects that belong to the predefined categories and marking them with bounding boxes, assigning each object its respective category label. This approach works in a single-modality framework, depending only on visual cues and not incorporating any textual data or extra semantic guidance. All experiments are performed in a limited-sample learning scenario, with annotated examples provided only for the novel categories.

Appendix A.2. Image Properties and Scene Diversity

The DIOR dataset includes high-resolution satellite and aerial imagery, typically sized around 800×800 pixels. Because of differences in imaging platforms and conditions, the ground sampling distance varies considerably, leading to significant variation in object sizes.

In the DIOR dataset, objects frequently exhibit random orientations and irregular spatial arrangements. Numerous scenes include tightly packed object configurations, intricate backgrounds, areas with shadows, and recurring textures. Typical examples cover airports and runways, highway networks along with their supporting facilities, seaports and coastal zones, major industrial sites, as well as urban and suburban regions that contain sports complexes. In these challenging scenarios, accurate object detection often relies more on detailed geometric patterns and contextual information than on overall appearance features alone.

Appendix A.3. Category Structure

To facilitate category organization and experimental evaluation, the DIOR object classes are grouped into several broad categories according to their functional and structural traits.

- Aerial transport: covering airplanes and airport facilities as seen from above.
- Maritime transport: including vessels and port facilities located along coastal and inland waterways.
- Road transport and associated infrastructure: including vehicles and highway-linked structures, for example, service zones, toll plazas, bridges, and flyovers.
- Rail transport: including railway stations and rail hubs with extended track arrangements.

The categories in each group show significant variation in form, size, and spatial arrangement, creating further difficulties for precise detection when only a few labeled examples are available.

Appendix A.4. Summary

This appendix summarizes the task formulation, scene properties, and category structure of the DIOR dataset adopted in this work. This supplementary description is intended to enhance clarity and ensure experimental reproducibility. No extra annotations, multimodal signals, or external forms of supervision are involved.

 doi:10.20944/preprints202602.1028.v1

Appendix B. Task Setting and Dataset Characteristics of NWPU VHR-10

Appendix B.1. Detection Task

On the NWPU VHR-10 dataset, the detection task is defined in a manner consistent with that used for DIOR. Each very-high-resolution overhead image is processed to locate all instances from predefined categories with bounding boxes, while assigning the corresponding class labels. The proposed method operates solely on visual inputs and excludes the use of textual prompts, semantic priors, or other multimodal cues. Its performance is examined in few-shot scenarios to evaluate robustness when only a small number of training samples are available.

Appendix B.2. Image Properties and Scene Diversity

NWPU VHR-10 consists of very-high-resolution imagery acquired using multiple sensing platforms. Relative to larger benchmark datasets, this collection includes fewer categories, yet shows substantial diversity in object scale, visual appearance, and surrounding context. The objects span a wide spectrum, from compact vehicles to large-scale facilities, and frequently occur in cluttered scenes where foreground structures are visually similar to the background.

The NWPU VHR-10 dataset includes scenes spanning urban areas, transportation infrastructures, port regions, industrial sites, as well as open terrains. In practice, effective detection depends not only on local visual cues but also on modeling geometric patterns and overall spatial organization.

Appendix B.3. Category Structure

NWPU VHR-10 defines ten object classes corresponding to frequently observed artificial structures in overhead imagery. For clarity, the categories are summarized below.

- Transportation-related objects include airplanes, ships, vehicles, bridges, and harbors, which appear in scenes structured by runways, road systems, or water routes.
- Infrastructure and functional facilities are represented by storage tanks and ground track fields, characterized by regular geometry and extensive spatial coverage.
- Sports and recreational facilities include baseball fields, basketball courts, and tennis courts, which are distinguished by clear markings and highly symmetric layouts.

Variations in object scale, geometric structure, and surrounding context among categories render NWPU VHR-10 a standard dataset for assessing few-shot detection capabilities.

Appendix B.4. Summary

This appendix summarizes the task formulation, scene characteristics, and category composition of the NWPU VHR-10 dataset used in this study. The description is provided to enhance clarity and reproducibility and does not introduce any multimodal information, semantic prompts, or additional supervision.

References

1. Li, B.; Zhou, Z.; Wu, T.; Luo, J. Fine-Grained Land Use Remote Sensing Mapping in Karst Mountain Areas Using Deep Learning with Geographical Zoning and Stratified Object Extraction. *Remote Sens.* **2025**, *17*, 2368, doi:10.3390/rs17142368. 368.
2. Wang, Z.; Liang, Y.; He, Y.; Cui, Y.; Zhang, X. Refined Land Use Classification for Urban Core Area from Remote Sensing Imagery by the EfficientNetV2 Model. *Appl. Sci.* **2024**, *14*, 7235, doi:10.3390/app14167235.
3. Zhang, R.; Tang, X.; You, S.; Duan, K.; Xiang, H.; Luo, H. A Novel Feature-Level Fusion Framework Using Optical and SAR Remote Sensing Images for Land Use/Land Cover (LULC) Classification in Cloudy Mountainous Area. *Appl. Sci.* **2020**, *10*, 2928, doi:10.3390/app10082928.
4. Ragab, M.; Abdushkour, H. A.; Khadidos, A. O.; Alshareef, A. M.; Alyoubi, K. H.; Khadidos, A. O. Improved Deep Learning-Based Vehicle Detection for Urban Applications Using Remote Sensing Imagery, *Remote Sens.* **2023**, *15*, 4747. doi:10.3390/rs15194747.

5. Ye, F.; Ai, T.; Wang, J.; Yao, Y.; Zhou, Z. A Method for Classifying Complex Features in Urban Areas Using Video Satellite Remote Sensing Data. *Remote Sens.* **2022**, *14*, 2324, doi:10.3390/rs14102324.
6. Wang, M.; Ding, W.; Wang, F.; Song, Y.; Chen, X.; Liu, Z. A Novel Bayes Approach to Impervious Surface Extraction from High-Resolution Remote Sensing Images. *Sensors* **2022**, *22*, 3924, doi:10.3390/s22103924.
7. Du, X.; Song, L.; Lv, Y.; Qiu, S. A Lightweight Military Target Detection Algorithm Based on Improved YOLOv5. *Electronics* **2022**, *11*, 3263, doi:10.3390/electronics11203263.
8. Zeng, B.; Gao, S.; Xu, Y.; Zhang, Z.; Li, F.; Wang, C. Detection of Military Targets on Ground and Sea by UAVs with Low-Altitude Oblique Perspective. *Remote Sens.* **2024**, *16*, 1288, doi:10.3390/rs16071288.
9. Sun, L.; Chen, J.; Feng, D.; Xing, M. Parallel Ensemble Deep Learning for Real-Time Remote Sensing Video Multi-Target Detection. *Remote Sens.* **2021**, *13*, 4377, doi:10.3390/rs13214377.
10. Liu, J.; Luo, Y.; Chen, S.; Wu, J.; Wang, Y. BDHE-Net: A Novel Building Damage Heterogeneity Enhancement Network for Accurate and Efficient Post-Earthquake Assessment Using Aerial and Remote Sensing Data. *Appl. Sci.* **2024**, *14*, 3964, doi:10.3390/app14103964.
11. Xie, Z.; Zhou, Z.; He, X.; Fu, Y.; Gu, J.; Zhang, J. Methodology for Object-Level Change Detection in Post-Earthquake Building Damage Assessment Based on Remote Sensing Images: OCD-BDA. *Remote Sens.* **2024**, *16*, 4263, doi:10.3390/rs16224263.
12. Da, Y.; Ji, Z.; Zhou, Y. Building Damage Assessment Based on Siamese Hierarchical Transformer Framework. *Mathematics* **2022**, *10*, 1898, doi:10.3390/math10111898.
13. Wang, D.; Yan, Z.; Liu, P. Fine-Grained Interpretation of Remote Sensing Image: A Review. *Remote Sens.* **2025**, *17*, 3887, doi:10.3390/rs17233887.
14. 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, doi:10.3390/rs14184435.
15. Nguyen, K.; Huynh, N. T.; Le, D. T.; et al. A Comprehensive Review of Few-Shot Object Detection on Aerial Imagery. *Comput. Sci. Rev.* **2025**, *57*, 100760, doi:10.1016/j.cosrev.2025.100760.
16. 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.
17. 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.
18. Zhang, Y.; Lyu, X.; Li, X.; Zhou, S.; Fang, Y.; Ding, C.; Gao, S.; Chen, J. Complementary Local–Global Optimization for Few-Shot Object Detection in Remote Sensing. *Remote Sens.* **2025**, *17*, 2136, doi:10.3390/rs17132136.
19. Wu, X.; Sahoo, D.; Hoi, S. Meta-RCNN: Meta Learning for Few-Shot Object Detection. In Proceedings of the 28th ACM International Conference on Multimedia (ACM MM); ACM: Seattle, WA, USA, October 2020; pp. 1679–1687.
20. Han, G.; Huang, S.; Ma, J.; He, Y.; Chang, S.-F. Meta Faster R-CNN: Towards Accurate Few-Shot Object Detection with Attentive Feature Alignment. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI); AAAI Press: Virtual Event, 2022; Vol. 36, No. 1, pp. 780–789.
21. Wang, X.; Huang, T. E.; Darrell, T.; Gonzalez, J. E.; Yu, F. Frustratingly Simple Few-Shot Object Detection. *arXiv* **2020**, arXiv:2003.06957, doi:10.48550/arXiv.2003.06957.
22. 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 IEEE/CVF International Conference on Computer Vision (ICCV); IEEE: Montreal, QC, Canada, October 2021; pp. 8681–8690.
23. Lin, H.; Li, N.; Yao, P.; Dong, K.; Guo, Y.; Hong, D. Generalization-Enhanced Few-Shot Object Detection in Remote Sensing. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, doi:10.1109/TCSVT.2025.3528262.
24. Gao, H.; Wu, S.; Wang, Y.; Kim, J. Y.; Xu, Y. FSOD4RSI: Few-Shot Object Detection for Remote Sensing Images via Features Aggregation and Scale Attention. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2024**, *17*, 4784–4796. doi:10.1109/JSTARS.2024.3362748.
25. Su, H.; You, Y.; Meng, G. Multi-Scale Context-Aware R-CNN for Few-Shot Object Detection in Remote Sensing Images. In Proceedings of the IGARSS 2022 – 2022 IEEE International Geoscience and Remote Sensing Symposium (IGARSS); IEEE: Kuala Lumpur, Malaysia, July 2022; pp. 1908–1911.
26. Li, X.; Hu, X.; Yang, J. Spatial Group-Wise Enhance: Improving Semantic Feature Learning in Convolutional Networks. *arXiv* **2019**, arXiv:1905.09646, doi:10.48550/arXiv.1905.09646.
27. Qin, A.; Chen, F.; Li, Q.; Song, T.; Zhao, Y.; Gao, C. Few-Shot Remote Sensing Scene Classification via Subspace Based on Multiscale Feature Learning. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2024**, doi:10.1109/JSTARS.2024.3432895.

 doi:10.20944/preprints202602.1028.v1

28. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. CARAFE: Content-Aware ReAssembly of Features. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); IEEE: Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3007–3016.
29. Fan, Q.; Zhuo, W.; Tang, C.; Tai, Y. Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE, Seattle, WA, USA, June 2020; pp. 4013–4022.
30. Xiao, Y.; Lepetit, V.; Marlet, R. Few-Shot Object Detection and Viewpoint Estimation for Objects in the Wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 3090–3106, doi:10.1109/TPAMI.2022.3174072.
31. Lu, X.; Diao, W.; Mao, Y.; Li, J.; Wang, P.; Sun, X.; Fu, K. Breaking Immutable: Information-Coupled Prototype Elaboration for Few-Shot Object Detection. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI); AAAI Press, Virtual Event, 2023; Vol. 37, No. 2, pp. 1844–1852.
32. Han, J.; Ren, Y.; Ding, J.; Yan, K.; Xia, G. Few-Shot Object Detection via Variational Feature Aggregation. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI); AAAI Press, Virtual Event, 2023; Vol. 37, No. 1, pp. 755–763.
33. Sun, B.; Li, B.; Cai, S.; Yuan, Y.; Zhang, C. FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE, Virtual Event, June 2021; pp. 7352–7362.
34. Kim, S.; Nam, W.; Lee, S. Few-Shot Object Detection with Proposal Balance Refinement. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR); IEEE, Montreal, QC, Canada, August 2022; pp. 4700–4707.
35. Li, X.; Deng, J.; Fang, Y. Few-Shot Object Detection on Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14, doi:10.1109/TGRS.2021.3051383.
36. Huang, X.; He, B.; Tong, M.; Wang, D.; He, C. Few-Shot Object Detection on Remote Sensing Images via Shared Attention Module and Balanced Fine-Tuning Strategy. *Remote Sens.* **2021**, *13*, 3816, doi:10.3390/rs13193816.
37. Zhou, Z.; Li, S.; Guo, W.; Gu, Y. Few-Shot Aircraft Detection in Satellite Videos Based on Feature Scale Selection Pyramid and Proposal Contrastive Learning. *Remote Sens.* **2022**, *14*, 4581, doi:10.3390/rs14184581.
38. Lin, T.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE, Honolulu, HI, USA, July 2017; pp. 2117–2125.
39. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv preprint* arXiv:1804.02767, 2018, doi:10.48550/arXiv.1804.02767.
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. Cheng, G.; Yan, B.; Shi, P.; Li, K.; Yao, X.; Guo, L. Prototype-CNN for Few-Shot Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–10, doi:10.1109/TGRS.2021.3078507.
42. Lu, X.; Sun, X.; Diao, W.; Mao, Y.; Li, J.; Zhang, Y. Few-Shot Object Detection in Aerial Imagery Guided by Text-Modal Knowledge. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–19, doi:10.1109/TGRS.2023.3250448.
43. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A. Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE, Virtual Event, June 2020; pp. 4096–4105.
44. 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.
45. 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.
46. Liu, Y.; Pan, Z.; Yang, J.; Zhang, B.; Zhou, G.; Hu, Y. 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.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE, Las Vegas, NV, USA, June 2016; pp. 770–778.

 doi:10.20944/preprints202602.1028.v1