aerospace

MDPI

*Article*

# Lightweight Spatial–Frequency Constraint Propagation Framework for Satellite Detection in Space Surveillance

**Rui Hong [1], Jiahao Li [2], Han Pan [3] and Qian Wang [2,4,*]**

[1] Aviation Key Laboratory of Science and Technology on Aerospace Vehicle, Chengdu 610041, China; ruihong2026@sohu.com
[2] School of Telecommunication and Information Engineering, Xi'an University of Posts & Telecommunications, Xi'an 710121, China; ljh2003@stu.xupt.edu.cn
[3] School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai 200240, China; hanpan@sjtu.edu.cn
[4] Key Laboratory of Polar Science, Ministry of Natural Resources, Shanghai 200136, China
* Correspondence: wangqian57@xupt.edu.cn

**Abstract**

Satellite object detection in space surveillance is challenged by sparse and weak targets in large-scale, structured backgrounds (e.g., star fields, clouds, and streak noise). Such interference is not random but exhibits spatial correlation and frequency regularity, causing target responses to be overwhelmed and difficult to separate within a single representation space. To address this issue, we propose a lightweight framework, termed DRSS-Net, based on the key observation that target–background separability can be enhanced across complementary representation coordinate systems. Specifically, spatial modeling captures local structural consistency, while frequency-domain processing characterizes global energy distribution and structured patterns. By alternating between these domains, the proposed method enables constraint propagation, where predictable background patterns are suppressed, and structurally inconsistent target responses are emphasized. In the spatial domain, a mutual conditioning mechanism with asymmetric channel allocation enhances the consistency between localization and semantic responses. In the frequency domain, a coupled refinement module models the interaction between energy distribution and structural configuration to distinguish structured background from anomalous targets. In addition, a scale selection strategy retains stable intermediate representations for efficient detection. Experiments on two independent space target datasets demonstrate that DRSS-Net consistently achieves superior detection performance with a compact model size under diverse observation conditions, including variations in target appearance, illumination, and structured background interference.

**Keywords:** satellite object detection; space surveillance; cross-domain representation; constraint propagation; lightweight detection



Academic Editor: Vladimir S. Aslanov

Received: 28 April 2026
Revised: 6 June 2026
Accepted: 7 June 2026
Published: 9 June 2026



## 1. Introduction

Satellite object detection is a fundamental component of space surveillance systems, supporting tasks such as catalog maintenance, on-orbit monitoring, and collision risk assessment [1,2]. In contrast to conventional object detection scenarios, space-based observations are characterized by sparse targets and large-scale background interference, where weak target responses are often entangled with globally distributed patterns such as star fields

and cloud structures. This leads to a fundamental challenge in representation learning, where target and background become difficult to separate within a single feature space.

Recent advances in object detection have focused on both architectural efficiency and representation enhancement. CNN-based detectors, such as YOLO [3–7] and RT-MDet [8], continuously improve the trade-off between accuracy and efficiency through optimized backbone, neck, and detection-head designs. Meanwhile, Transformer-based approaches [9–12] introduce attention-driven end-to-end detection frameworks that eliminate hand-crafted post-processing components and achieve competitive performance.

To improve detection under complex backgrounds, several studies have explored enhanced feature representations and multi-scale modeling strategies. FFCA-YOLO [13] and FBRT-YOLO [14] improve small-object detection through feature calibration, attention mechanisms, and feature reuse strategies. CSFPR-RTDETR [15] further incorporates spatial-frequency modeling to capture structured background patterns and improve target discrimination. Beyond detector architecture, HRNet [16] demonstrates the importance of preserving high-resolution representations for accurate localization and fine-grained structural perception. These studies highlight the value of multi-scale feature preservation, cross-domain representation learning, and context-aware modeling for challenging detection scenarios.

Recently, frequency-domain representation learning has attracted increasing attention in visual recognition. Wang et al. proposed frequency-domain disentanglement for UAV object detection by separating domain-invariant and domain-specific spectral components [17]. Lin et al. introduced deep frequency filtering to enhance transferable frequency responses through learnable spectral attention [18]. Sun et al. further explored joint spatial-frequency representation learning via spatial-frequency entanglement mechanisms [19]. Other studies have investigated frequency perception and frequency-aware feature aggregation for improving feature discriminability [20]. Although these methods demonstrate the effectiveness of spectral representations, they primarily focus on feature enhancement or cross-domain interaction.

Despite recent progress, most detectors focus on architectural efficiency or single-domain feature enhancement, implicitly assuming sufficiently strong target responses and locally uncorrelated background noise. However, this assumption does not hold in satellite imagery, where sparse and weak targets are embedded in large-scale structured backgrounds with strong spatial correlation and frequency regularity. Under such conditions, the limitation lies not merely in feature extraction capacity, but in the inadequacy of representation itself. When target and background are modeled within a single representation space, weak targets are easily overwhelmed by structured interference, leading to poor feature separability. In contrast, background patterns exhibit predictable regularity in the frequency domain, while targets are characterized by local structural consistency in the spatial domain. This suggests that target–background separability can be enhanced when features are analyzed across complementary representation spaces, rather than within a single domain.

Unlike existing spatial-frequency or attention-based detectors that primarily aim to enhance feature representations, the proposed framework focuses on improving target–background separability through coordinated modeling across complementary representation spaces. We propose a lightweight spatial-frequency framework for satellite object detection. The core idea is to alternate between spatial and frequency representations, enabling constraint propagation across domains, where predictable background patterns are progressively suppressed and structurally inconsistent target responses are emphasized. Furthermore, we observe that feature reliability varies significantly across scales in satellite imagery, where shallow features are noise-dominated and deep features suffer from exces-

sive compression. This motivates a scale selection strategy that retains stable intermediate representations for robust and efficient detection. The main contributions of this work are summarized as follows:

- We introduce a spatial mutual conditioning (SMC) mechanism to address the entanglement between semantic responses and spatial structures. By decoupling features into semantic and structural components and modeling their where–what interaction, the proposed module improves feature discriminability and structural consistency under low-SNR conditions.
- We develop a spectral coupled refinement (SCR) module to model the interaction between energy distribution and structural configuration in the frequency domain. The proposed decoupling-refinement design enhances the stability and discriminability of weak target representations under noise and distortion.
- We present a lightweight spatial-frequency framework that jointly integrates complementary domain modeling and stability-guided scale selection. By coordinating representation across domains and scales, the proposed design achieves improved feature reliability while maintaining a compact architecture and efficient computation.

## 2. Methods

### *2.1. Overall Architecture*

The proposed framework, termed Decoupling-Refinement Spatial–Spectral Network (DRSS-Net), is built upon a YOLOv12-style detection pipeline [21], as illustrated in Figure 1. In large-scale space observation imagery, weak targets are easily overwhelmed by background interference during hierarchical feature extraction, leading to progressive target–background feature entanglement. To address this issue, we redesign the backbone using cascaded Spatial-Frequency Sequential Modeling (SFSM) stages, which explicitly perform feature disentanglement and refinement throughout the representation learning process. The resulting multi-scale features are subsequently processed by a scale-aware detection pipeline tailored for sparse target detection.

A key characteristic of the proposed framework is the alternating spatial-frequency modeling strategy implemented throughout the backbone. Specifically, each stage consists of a spatial mutual conditioning (SMC) module followed by a spectral coupled refinement (SCR) module. Rather than treating spatial and frequency representations as independent branches, the network progressively alternates between the two representation spaces across multiple stages. This design is motivated by the observation that target–background ambiguity cannot be fully resolved within a single representation domain. Spatial representations emphasize local structural consistency, whereas frequency representations capture global energy regularity. By repeatedly alternating between these complementary domains, discriminative cues extracted in one representation space are inherited and further refined in subsequent stages.

Let $P_l$ denote the input feature representation at the $(l)$ stage. The proposed backbone is composed of $L$ cascaded SFSM stages. The output of the ($l$)-th stage is defined as

$$P_l = \mathcal{F}_{\mathrm{SCR}}^{(l)}\left(\mathcal{F}_{\mathrm{SMC}}^{(l)}(P_{l-1})\right), \quad l = 2, \ldots, L, \tag{1}$$

where $\mathcal{F}_{\mathrm{SMC}}$ and $\mathcal{F}_{\mathrm{SCR}}$ denote the spatial mutual conditioning and spectral coupled refinement modules, respectively. Each stage inherits the refined representation generated by the previous stage and further performs structural disentanglement in the spatial domain followed by spectral refinement in the frequency domain. Consequently, structural constraints established at earlier stages are progressively preserved, propagated, and strengthened

throughout the backbone. We refer to this recursive refinement process as spatial-frequency constraint propagation.

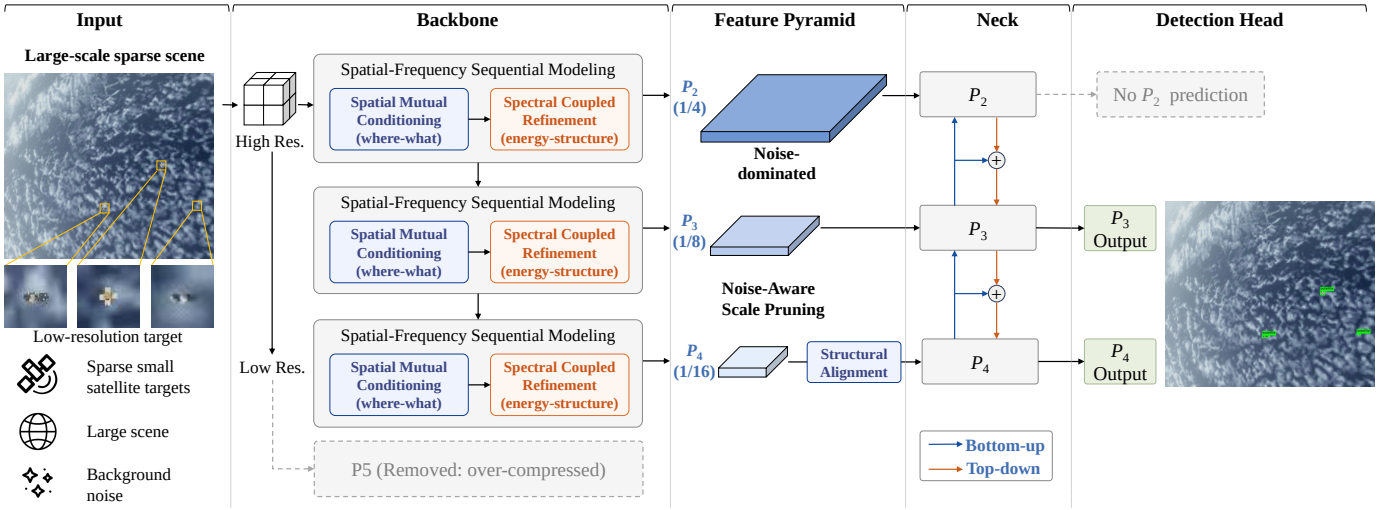


**Figure 1.** Overview of the proposed Decoupling-Refinement Spatial–Spectral Network (DRSS-Net) for satellite object detection. The framework is built upon a YOLOv12-style architecture with a redesigned backbone and scale-aware structural adaptation. Specifically, the deepest feature level ($P_5$) is removed to avoid excessive spatial compression, while the shallow level ($P_2$) is excluded from prediction due to noise dominance but still participates in feature aggregation through the neck. The backbone is composed of cascaded spatial–spectral modules, where spatial mutual conditioning improves feature identifiability by modeling where–what coupling, and spectral coupled refinement enhances structural consistency via frequency-domain interaction. The neck employs bidirectional feature aggregation, and the detection head produces predictions only at intermediate scales ($P_3$ and $P_4$), achieving a balance between robustness and efficiency.

### *2.2. Noise-Aware Scale Pruning for Feature Pyramid Optimization*

In large-scale space observation imagery, feature pyramid levels exhibit highly uneven information quality under low complex background noise conditions. Shallow features tend to be noise-dominated, while overly deep features suffer from excessive spatial compression. This motivates a noise-aware scale selection strategy that emphasizes informative intermediate representations.

Given an input image $I$, the backbone produces multi-level features:

$$\{P_2, P_3, P_4\} = \mathcal{F}_{\text{backbone}}(I), \tag{2}$$

where the deepest level $P_5$ is removed to avoid over-compressed semantics that are less informative for weak and sparse targets. The multi-scale features $(P_2, P_3, P_4)$ are generated from different depths of the cascaded SFSM backbone. Therefore, each pyramid level not only encodes features at a particular spatial scale but also reflects a different degree of spatial-frequency propagation. Deeper levels correspond to progressively refined representations obtained through repeated spatial–spectral interactions.

The multi-level features are fed into a standard bidirectional feature pyramid (FPN/PAN) for cross-scale aggregation. Different from conventional designs, we introduce an additional spatial refinement only at the highest available level:

$$\tilde{P}_4 = \mathcal{F}_{\text{SMC}}(P_4), \tag{3}$$

where $\mathcal{F}_{\text{SMC}}(\cdot)$ denotes the proposed spatial mutual conditioning (SMC) module. This operation serves as a structural alignment step, which reorganizes spatial responses and

mitigates the inconsistency introduced by cross-scale feature aggregation. $\tilde{P}_4$ is used to replace $P_4$ in the subsequent feature aggregation. This design is motivated by the need for structural alignment during multi-scale feature fusion. While the bidirectional aggregation enhances semantic richness, it also introduces structural inconsistency due to the mixing of features with different resolutions and signal qualities. As a result, the spatial configuration of weak targets can become distorted or misaligned across scales. The $P_4$ level provides a favorable trade-off between semantic abstraction and spatial resolution. After top-down aggregation, it carries relatively strong semantic cues but may still suffer from structural inconsistency due to weak target responses. Applying SMC at this level acts as a structural alignment mechanism, which reorganizes spatial responses and restores coherence between localization cues and semantic information before further propagation. In contrast, we do not apply spectral refinement at this stage. Since $P_4$ has already undergone substantial downsampling, its high-frequency content is limited, making additional frequency-domain processing less effective. Spatial refinement is therefore more suitable for consolidating structural information and achieving stable structural alignment at this level.

Given the multi-level features extracted from the backbone, the cross-scale feature aggregation is performed, producing refined feature representations:

$$\{P_2', P_3', P_4'\} = \mathcal{N}(\{P_2, P_3, \tilde{P}_4\}). \tag{4}$$

After feature aggregation, the detection head produces predictions only on the following:

$$Y = \mathcal{H}(\{P_3', P_4'\}), \tag{5}$$

while the shallow level $P_2$ is excluded from direct prediction due to its noise-dominated nature. Nevertheless, $P_2$ is retained within the pyramid to support feature propagation and enhance fine-grained information flow.

Overall, the proposed design performs noise-aware scale pruning by selectively suppressing ineffective scales (e.g., $P_5$ and $P_2$ for prediction), while reinforcing informative intermediate representations. The targeted refinement at $P_4$ further improves structural consistency, leading to more robust feature representations under complex background noise and sparse target conditions.

### *2.3. Structural Decoupling with Spatial Mutual Conditioning*

In large-scale satellite imagery under complex background noise conditions, feature representations are severely entangled, where target responses are mixed with background noise across both spatial structures and semantic responses. Direct feature enhancement tends to amplify noise together with signals, leading to unstable localization and degraded discriminability. To address this issue, we propose a structural decoupling strategy that separates feature representations into complementary components, followed by a mutual conditioning mechanism to restore their interaction, as illustrated in Figure 2.

Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, we first perform channel-wise partition to obtain two complementary subsets:

$$X = [X_s, X_t], \quad X_s \in \mathbb{R}^{\frac{1}{4}C\times H\times W},\ X_t \in \mathbb{R}^{\frac{3}{4}C\times H\times W}, \tag{6}$$

where $X_s$ denotes a lightweight semantic branch and $X_t$ denotes a structural branch. The asymmetric channel allocation is designed to preserve the distinct roles of the two branches. A larger proportion of channels is assigned to semantic representation learning, while a smaller subset is reserved for localization-oriented structural modeling. This design encourages complementary feature specialization before mutual conditioning.

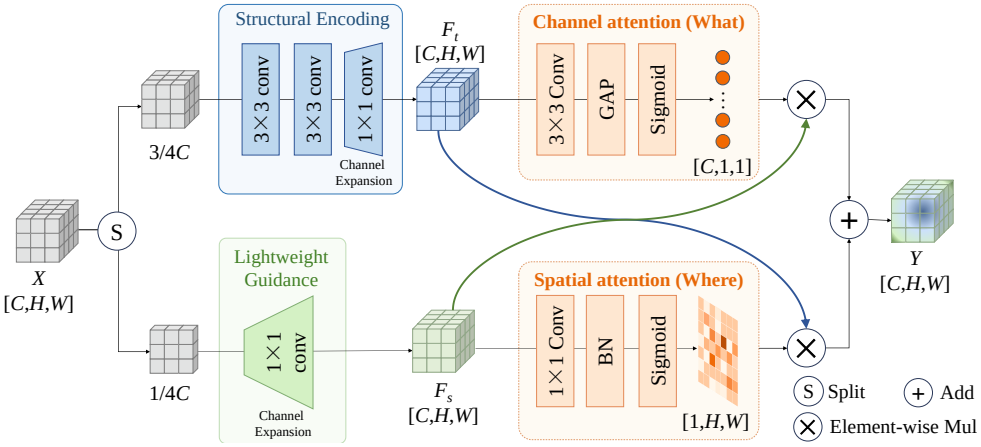


**Figure 2.** Illustration of the spatial component of DRSS-Net: spatial mutual conditioning (SMC). The module follows a decoupling-refinement paradigm, where where–what interaction is modeled through mutual conditioning to enhance feature discriminability and structural alignment.

The structural branch focuses on capturing local spatial patterns and structural responses:

$$F_t = \phi_{1\times1}(\phi_{3\times3}(\phi_{3\times3}(X_t))), \tag{7}$$

where $\phi_{k\times k}(\cdot)$ denotes a convolution with kernel size $k$, and $F_t \in \mathbb{R}^{C\times H\times W}$ is the structural feature. The semantic branch provides global semantic guidance with reduced complexity:

$$F_s = \psi_{1\times1}(X_s), \tag{8}$$

where $\psi_{1\times1}(\cdot)$ is a point-wise convolution and $F_s \in \mathbb{R}^{C\times H\times W}$.

After decoupling, the two branches are further refined through bidirectional conditioning, explicitly modeling the interaction between spatial localization (where) and semantic discrimination (what). First, the semantic branch generates a spatial attention map:

$$A_{spa} = \sigma(\mathrm{BN}(\mathrm{Conv}_{1\times1}(F_s))) \in \mathbb{R}^{1\times H\times W}, \tag{9}$$

which encodes spatial importance and acts as a where signal. This map is used to refine the structural feature:

$$\tilde{F}_t = F_t \odot A_{spa}, \tag{10}$$

where $\odot$ denotes element-wise multiplication. Conversely, the structural branch produces a channel attention vector:

$$A_{cha} = \sigma(\mathrm{GAP}(\mathrm{DWConv}_{3\times3}(F_t))) \in \mathbb{R}^{C\times1\times1}, \tag{11}$$

which captures channel-wise importance as a what signal. This is used to refine the semantic feature:

$$\tilde{F}_s = F_s \odot A_{cha}. \tag{12}$$

Finally, the refined features are fused to reconstruct a coherent representation:

$$Y = \tilde{F}_t + \tilde{F}_s. \tag{13}$$

The proposed SMC module follows a decoupling-refinement paradigm. First, feature channels are asymmetrically partitioned into localization-oriented and semantic-oriented groups, improving feature identifiability under complex background interference. Unlike conventional attention mechanisms that primarily perform feature re-weighting within

a single representation stream, SMC explicitly models the interaction between complementary feature groups through bidirectional mutual conditioning. Localization cues guide semantic refinement, while semantic context simultaneously constrains localization responses, establishing a structured where–what coupling. This design promotes feature consistency and discriminability, resulting in more robust representations for sparse weak-target detection in noise-dominated scenes.

### *2.4. Spectral Coupled Refinement Under Energy–Structure Consistency*

#### 2.4.1. Energy–Structure Consistency Analysis

To quantitatively justify the concept of energy–structure consistency, we conducted a detailed spectral analysis on patches extracted from satellite images. Patches were randomly cropped from the original images with size $128 \times 128$ (PATCH_SIZE = $128 \times 128$). Using annotations, a binary mask was generated for each patch: pixels inside object bounding boxes were labeled 1, and background pixels 0. Patches were selected according to target coverage thresholds: background patches contained less than 1% of target pixels (BACKGROUND_RATIO_TH = 0.01), and target patches contained more than 50% of target pixels (TARGET_RATIO_TH = 0.50). Each category included 200 patches, and patches with extremely low variance (MIN_PATCH_STD = 0.03) were discarded to avoid instability in frequency analysis caused by nearly uniform regions.

Each patch was transformed into the frequency domain using a 2D Fast Fourier Transform (FFT). To focus on informative local structures, we only considered mid- to high-frequency components within the normalized frequency range $[0.08, 0.60]$. For each patch, we computed the normalized mutual information (MI) between different spectral representations, including magnitude and phase components (energy–structure) as well as real and imaginary components (real–imaginary). Specifically, the MI was computed as

$$\mathrm{MI}(X, Y) = \sum_{x} \sum_{y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \tag{14}$$

where $p(x, y)$ denotes the joint probability of the two spectral components falling into discrete bins, and $p(x)$ and $p(y)$ are the corresponding marginal probabilities. Normalized MI was used to mitigate the influence of patch entropy, quantization, and dynamic range, allowing a fair comparison between background and target patches across different spectral representations.

Figure 3 presents the distribution of normalized mutual information for background and target patches under two spectral representations: energy–structure (magnitude–phase) and real–imaginary. The results reveal a clear separation between background and target distributions in both cases. Specifically, background regions consistently exhibit higher mutual information than target regions, indicating a stronger dependency between paired spectral components. This observation suggests that structured backgrounds preserve a more stable correspondence in the frequency domain, whereas weak targets tend to disrupt such regularity and exhibit weaker spectral coupling.

The analysis reveals that background regions exhibit significantly higher mutual information between spectral energy and structural components than target regions, indicating the existence of energy–structure consistency in structured backgrounds and energy–structure mismatch in weak targets.

Interestingly, the separation between background and target distributions is considerably larger under the real–imaginary representation than under the energy–structure representation. This suggests that the discriminative characteristics associated with target-induced spectral perturbations are more explicitly preserved in the complex spectrum. Although magnitude and phase provide a physically interpretable description of spectral

energy and structure, they are obtained through nonlinear transformations of the complex spectrum and require additional discretization during mutual-information estimation. These operations inevitably reduce part of the fine-grained spectral variations introduced by weak targets. In contrast, the real and imaginary components directly retain the original complex-valued spectral information, leading to a stronger distinction between regular background patterns and target-induced anomalies.

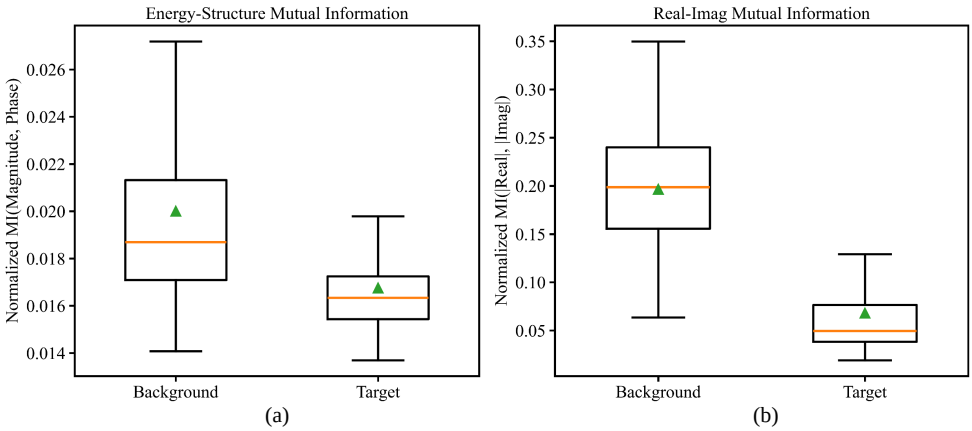


**Figure 3.** Mutual information analysis of background and target patches under different spectral representations. Box plots show the distributions of normalized mutual information for (**a**) energy–structure (magnitude–phase) and (**b**) real–imaginary representations. The orange line represents the median, while the green triangle indicates the mean.

#### 2.4.2. Spectral Coupled Refinement

Building on the analysis of energy–structure consistency, the proposed spectral coupled refinement (SCR) module leverages the real–imaginary representation of the feature spectrum to selectively enhance target discriminability. Given an input feature map

$$X \in \mathbb{R}^{C\times H\times W}, \tag{15}$$

we first compute its 2D Fourier Transform to obtain the complex spectrum:

$$\hat{X} = \mathcal{F}(X) = \hat{X}_r + j\hat{X}_i \in \mathbb{C}^{C\times U\times V}, \tag{16}$$

where $\hat{X}_r$ and $\hat{X}_i$ denote the real and imaginary components, jointly encoding spectral energy and structural phase information. FFT preserves the complete complex spectrum, enabling simultaneous modeling of both energy and structural information, and making structured background patterns more distinguishable from localized target responses.

The real and imaginary components are concatenated along the channel dimension to form

$$\hat{X}_c = \text{concat}(\hat{X}_r, \hat{X}_i) \in \mathbb{R}^{2C\times U\times V}, \tag{17}$$

followed by a $3 \times 3$ convolution in the frequency domain:

$$\tilde{Y} = \phi_{3\times 3}(\hat{X}_c). \tag{18}$$

This local convolution acts as a learnable operator that models cross-component interactions. Conceptually, it enforces energy–structure consistency by reinforcing spectral regions where real and imaginary components are coherent (corresponding to consistent background patterns), while simultaneously amplifying regions exhibiting energy–structure mismatch (corresponding to concealed or weak targets). In this way, the SCR module

selectively strengthens regular background structures and highlights target-induced spectral anomalies.

Finally, the refined complex spectrum is transformed back to the spatial domain via inverse FFT:

$$Y = \mathcal{F}^{-1}(\tilde{Y}_r + j\tilde{Y}_i) \in \mathbb{R}^{C\times H\times W}, \tag{19}$$

and a subsequent $1 \times 1$ convolution is applied for channel mixing:

$$Z = \phi_{1\times 1}(Y), \tag{20}$$

producing the output feature map with enhanced target discrimination while preserving structural consistency of the background. Figure 4 illustrates the workflow of the SCR module, highlighting its decoupling-refinement paradigm in the frequency domain.

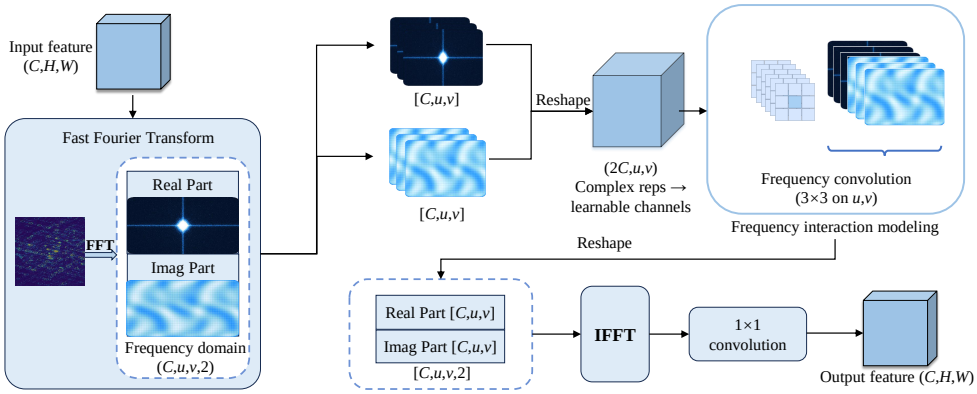


**Figure 4.** Illustration of the spectral component of DRSS-Net: spectral coupled refinement. The module follows a decoupling-refinement paradigm in the frequency domain, where energy–structure interaction is modeled through coupled refinement to enhance representation stability.

## 3. Experiments

### *3.1. Experimental Setup*

We evaluate the proposed method on the Satellite/Space Object Detection Dataset v2 (SODv2) [22,23] and the Large-Scale Synthetic Benchmark Dataset for Non-Cooperative Space target Perception (NCSTP) [24]. The SODv2 dataset contains satellite targets observed under diverse conditions, including variations in target brightness, star-field density, cloud interference, and structured background noise. These factors produce substantial appearance variations and make the dataset representative of practical space surveillance scenarios. The proposed model follows the standard YOLOv8 training configuration. The total loss is defined as

$$L = \lambda_{box}L_{IoU} + \lambda_{cls}L_{cls} + \lambda_{dfl}L_{DFL}, \tag{21}$$

where $L_{IoU}$ denotes the CIoU-based localization loss [25], $L_{DFL}$ denotes the distribution focal loss for fine-grained bounding box regression [26], and $L_{cls}$ represents the classification loss [5]. All models were trained for 300 epochs using a cosine annealing learning-rate schedule. Data augmentation was implemented using the Albumentations library [27].

All experiments are implemented using PyTorch 2.2.1 with CUDA 12.1 and conducted on a single NVIDIA RTX 3090 GPU. We adopt the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.01, momentum of 0.937, and weight decay of $5 \times 10^{-4}$ to train the models. The batch size is set to $16 \times 16$. Unless otherwise specified, all models are trained under the same settings to ensure fair comparisons. For SODv2, 650 images are used for training and 150 images are reserved for testing. For NCSTP, 1820 images are used for training, 260 images for validation, and 520 images for testing.

We evaluate the performance using standard object detection metrics, including precision, recall, and mean average precision ($mAP_{50}$ and $mAP_{50:95}$). Specifically, $mAP_{50}$ measures detection performance at an intersection-over-union (IoU) threshold of 0.5, while $mAP_{50:95}$ averages results over multiple IoU thresholds from 0.5 to 0.95, providing a more comprehensive evaluation of localization accuracy. In addition, model efficiency is assessed in terms of the number of parameters (Params) and floating point operations (FLOPs).

### *3.2. Comparison with State-of-the-Art Methods*

We compare the proposed method with a wide range of state-of-the-art object detectors, covering both general-purpose YOLO variants (e.g., YOLOv8, YOLOv9, YOLOv10, YOLOv11, YOLOv12, and YOLO26 with different model scales) and representative methods designed for small objects, including FFCA-YOLO [13], its lightweight variant L-FFCA-YOLO, FBRT-YOLO [14], CSFPR-RTDETR [15], RTMDet [8], RT-DETR [12], and HRNet [16]. The quantitative results on SODv2 are summarized in Table 1. The proposed method achieves the best overall trade-off between detection accuracy, model complexity, and inference efficiency. In particular, mAP@50 is adopted as the primary evaluation metric, as it more directly reflects target detection capability under sparse and weak signal conditions. The proposed method achieves the highest mAP@50 of 79.15%, outperforming all competing detectors, including larger YOLO variants and Transformer-based approaches. More importantly, it also attains a competitive mAP@50-95 of 28.47%, demonstrating that the performance gains are preserved under stricter localization requirements. These results indicate that the proposed framework improves not only target detectability but also bounding-box localization quality. Meanwhile, the model contains only 3.46 M parameters and requires 23.0 G FLOPs, which are substantially lower than most state-of-the-art detectors. In addition, the proposed method achieves 72.69 FPS, corresponding to an average inference latency of approximately 13.76 ms per image, demonstrating its suitability for real-time deployment in resource-constrained space surveillance systems.

**Table 1.** Comparison with state-of-the-art object detectors on the SODv2. The proposed method achieves the best trade-off across all evaluation metrics while maintaining a lightweight model size and computational cost.

| Scale | Model | Precision (%) | Recall (%) | mAP@50 (%) | mAP@50-95 (%) | Params | FLOPs | FPS |
|---|---|---|---|---|---|---|---|---|
| Small | YOLOv8-s | 76.77 | 72.32 | 71.76 | 26.82 | 11.13 M | 28.4 G | 101.65 |
| | YOLOv9-s | 68.33 | 69.10 | 69.51 | 25.92 | 7.17 M | 26.7 G | 40.99 |
| | YOLOv10-s | 69.12 | 62.66 | 69.40 | 25.93 | 7.22 M | 21.4 G | 83.94 |
| | YOLOv11-s | 73.05 | 70.97 | 70.67 | 25.69 | 9.41 M | 21.3 G | 85.14 |
| | YOLOv12-s | 75.69 | 73.82 | 73.01 | 27.64 | 9.23 M | 21.2 G | 50.94 |
| | YOLO26-s | 64.09 | 66.52 | 65.80 | 24.62 | 9.47 M | 20.5 G | 66.28 |
| | L-FFCA-YOLO | 67.20 | 66.50 | 61.70 | 19.50 | 5.04 M | 37.1G | 91.01 |
| | FBRT-YOLO | 74.10 | 74.20 | 74.00 | 27.30 | 7.36 M | 58.7 G | 111.00 |
| | FFCA-YOLO | 71.00 | 63.10 | 62.10 | 20.70 | 7.12 M | 51.2 G | 70.87 |
| Medium | YOLOv11-m | 72.07 | 70.87 | 69.55 | 27.51 | 20.03 M | 67.6 G | 59.43 |
| | YOLOv12-m | 68.49 | 72.75 | 72.21 | 26.93 | 20.11 M | 67.1 G | 38.00 |
| | YOLO26-m | 68.95 | 69.53 | 69.29 | 25.88 | 20.35 M | 67.8 G | 66.22 |
| | RTDETR-r18 | 72.71 | 72.02 | 71.95 | 26.01 | 19.87 M | 56.9 G | 33.22 |
| | CSFPR-RTDETR | 70.90 | 68.70 | 70.40 | 25.00 | 14.08 M | 63.8 G | 45.2 |

**Table 1.** *Cont.*

| Scale | Model | Precision (%) | Recall (%) | mAP@50 (%) | mAP@50-95 (%) | Params | FLOPs | FPS |
|---|---|---|---|---|---|---|---|---|
| Large | YOLOv11-l | 75.74 | 78.11 | 75.92 | 27.73 | 25.28 M | 86.6 G | 40.90 |
| | YOLOv12-l | 73.54 | 69.19 | 69.36 | 26.50 | 26.34 M | 88.5 G | 30.84 |
| | YOLO26-l | 72.98 | 64.38 | 67.17 | 24.20 | 24.75 M | 86.1 G | 36.64 |
| | RTMDet | 67.37 | 70.50 | 69.10 | 28.30 | 24.20 M | 39.2 G | – |
| | HRNet | 73.44 | 71.30 | 74.50 | 27.20 | 23.00 M | 38.3 G | – |
| Xlarge | YOLOv11-x | 71.56 | 73.45 | 73.01 | 27.64 | 56.83 M | 194.4 G | 40.42 |
| | YOLOv12-x | 79.58 | 76.92 | 76.44 | 29.83 | 59.04 M | 198.5 G | 26.69 |
| | YOLO26-x | 67.08 | 66.95 | 67.59 | 26.67 | 55.63 M | 193.4 G | 40.89 |
| – | Ours | 77.06 | 76.40 | 79.15 | 28.47 | 3.46 M | 23.0 G | 72.69 |

The quantitative results on the NCSTP are summarized in Table 2. The proposed method achieves the best overall trade-off, obtaining 99.38% mAP@50 and 83.19% mAP@50-95, while using only 3.46 M parameters. Compared with recent YOLO variants and remote-sensing-oriented detectors, our method consistently delivers higher detection accuracy with a substantially lighter model size. Notably, it surpasses the strongest competing methods in both mAP@50 and mAP@50-95, indicating that the proposed mechanism remains effective under different target categories and imaging conditions.

**Table 2.** Comparison with state-of-the-art object detectors on the NCSTP dataset. The proposed method achieves the best trade-off across all evaluation metrics while maintaining a lightweight model size and computational cost.

| Model | Precision (%) | Recall (%) | mAP@50 (%) | mAP@50-95 (%) | Params | FLOPs | FPS |
|---|---|---|---|---|---|---|---|
| YOLOv8-s | 98.55 | 96.92 | 98.66 | 82.27 | 11.13 M | 28.4 G | 115.80 |
| YOLOv9-s | 99.16 | 97.99 | 99.12 | 82.98 | 7.17 M | 26.7 G | 49.45 |
| YOLOv10-s | 96.24 | 97.35 | 98.60 | 81.20 | 7.22 M | 21.4 G | 83.94 |
| FBRT-YOLO-m | 98.71 | 98.11 | 99.05 | 82.54 | 7.36 M | 58.7 G | 90.32 |
| FFCA-YOLO | 97.60 | 97.20 | 99.10 | 79.60 | 7.12 M | 51.2 G | 52.73 |
| L-FFCA-YOLO | 98.50 | 97.70 | 98.50 | 76.50 | 5.04 M | 37.1 G | 46.83 |
| YOLOv11-s | 98.68 | 97.30 | 98.45 | 82.60 | 9.42 M | 21.3 G | 80.94 |
| YOLOv12-s | 97.15 | 96.80 | 98.68 | 81.87 | 9.08 M | 19.3 G | 51.92 |
| YOLOv13-s | 97.84 | 97.18 | 99.09 | 82.63 | 9.01 M | 20.1 G | 37.41 |
| YOLO26-s | 96.96 | 97.62 | 98.84 | 81.95 | 9.47 M | 20.5 G | 77.14 |
| Ours | 98.84 | 98.16 | 99.38 | 83.19 | 3.46 M | 23.0 G | 70.64 |

From a performance perspective, the improvement is particularly evident when compared to models of similar or even larger capacity. While larger models tend to rely on deeper architectures and higher computational complexity, they do not consistently yield better results in this task. This suggests that simply increasing model capacity is insufficient for handling large-scale sparse scenes with complex background noise. In contrast, our method explicitly addresses these challenges through noise-aware scale pruning and a decoupling-refinement design, leading to more effective feature representations.

To further illustrate the qualitative performance, Figure 5 presents visual comparisons on challenging cases. The examples are divided into two groups: (a) large-scale prominent targets and (b) small-scale weak targets. For large and prominent targets, most methods are able to produce correct detections, indicating that such scenarios are relatively less challenging. However, for small and weak targets, baseline methods often fail to detect objects

or produce incomplete detections due to insufficient feature representation under complex background noise conditions. In contrast, the proposed method is able to accurately localize these challenging targets with higher recall and more precise bounding boxes. These results demonstrate that the proposed method is particularly effective in handling small-scale and weak targets, where conventional detectors struggle. This improvement can be attributed to the proposed decoupling-refinement mechanism, which enhances feature identifiability in the spatial domain and stabilizes structural representation in the spectral domain.

### *3.3. Ablation Study*

#### 3.3.1. Component-Level Ablation

We conduct a detailed analysis of the effectiveness of each component, as shown in Table 3. Note that the first row corresponds to the baseline model, where the proposed modules are replaced by the original backbone blocks (e.g., C3k2 and A2C2f), resulting in higher parameter count and computational cost.

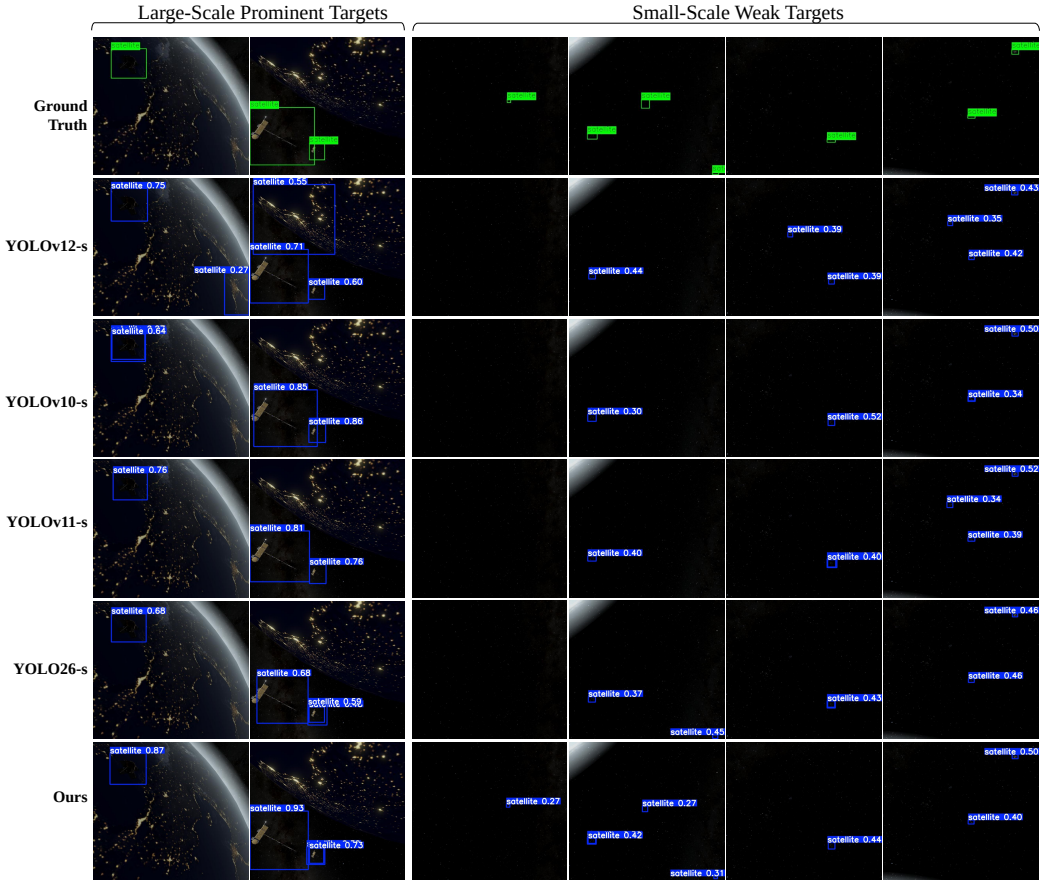


**Figure 5.** Qualitative comparison of challenging scenarios. The first two columns show large-scale prominent targets, where most methods perform comparably. The last four columns present small-scale weak targets under complex backgrounds, where baseline methods suffer from missed detections and false alarms, while the proposed method achieves more accurate and complete detection. The blue boxes represent the detected targets, while the green boxes indicate the ground-truth annotations.

Introducing the spatial module together with the $P_4$ structural alignment leads to a notable improvement in mAP@50-95 (from 26.23 to 28.90), while significantly reducing model complexity (5.25 M $\rightarrow$ 2.94 M parameters). This indicates that the proposed spatial decoupling effectively enhances feature identifiability under low-SNR conditions, while the structural alignment at $P_4$ further stabilizes spatial responses after cross-scale fusion. The improvement in recall suggests better localization of weak targets, whereas the slight

variation in precision reflects a more balanced trade-off between detection sensitivity and noise suppression.

**Table 3.** Ablation study on the effectiveness of different components. "Spatial (Backbone)" and "Spectral" denote the proposed spatial and spectral modules embedded in the backbone, respectively, while "Structural Alignment ($P_4$ Neck)" denotes the additional spatial refinement applied at the $P_4$ level in the neck.

| Spatial (Backbone) | Spectral (Backbone) | Structural Alignment ($P_4$ Neck) | Precision (%) | Recall (%) | mAP@50 (%) | mAP@50-95 (%) | Params | FLOPs |
|---|---|---|---|---|---|---|---|---|
| × | × | × | 73.80 | 73.82 | 75.56 | 26.23 | 5.25 M | 35.9 G |
| ✓ | × | ✓ | 74.26 | 73.06 | 75.45 | 28.90 | 2.94 M | 20.4 G |
| ✓ | ✓ | × | 74.20 | 74.96 | 75.37 | 27.18 | 3.25 M | 22.4 G |
| ✓ | ✓ | ✓ | 77.06 | 76.40 | 79.15 | 28.47 | 3.46 M | 23.0 G |

*Note:* ✓ denotes the presence of the corresponding module in the model configuration, whereas × denotes its absence.

When the spectral module is introduced without structural alignment, recall increases (73.82 → 75.54), but precision and mAP degrade. This phenomenon suggests that spectral modeling enhances sensitivity to global structural patterns (improving detection of weak signals), but may also amplify noise or introduce structural inconsistency if not properly constrained in the spatial domain. This highlights that spectral refinement alone is insufficient and requires complementary spatial alignment. By jointly integrating spatial decoupling, spectral coupled refinement, and structural alignment, the model achieves the best trade-off across all metrics. This demonstrates a clear complementary mechanism: spatial decoupling improves feature separability, spectral refinement enhances structural expressiveness, and structural alignment stabilizes cross-scale representations. The significant gain in both precision and recall indicates that the proposed design effectively balances noise suppression and target sensitivity.

Compared to the baseline, all variants with the proposed modules achieve substantial reductions in parameters and FLOPs. This confirms that the performance gain is not due to increased model capacity, but rather more efficient feature representation enabled by the proposed design. The final model achieves superior accuracy with significantly lower complexity, highlighting its suitability for large-scale space observation scenarios.

To further understand the effectiveness of the proposed design, we visualize feature responses at different stages of the network. As shown in Figure 6, the features after spatial-frequency modeling already exhibit enhanced target responses, indicating that complementary domain modeling helps suppress interference. After introducing structural alignment, the target regions become more concentrated and spatially consistent, while irrelevant background patterns are further attenuated. This progressive refinement demonstrates that the proposed framework improves feature discriminability through coordinated modeling across domains and scales, rather than relying on a single-stage enhancement.

#### 3.3.2. Scale-Level Ablation

We further analyze the impact of the proposed noise-aware scale pruning strategy by examining the roles of the shallow level ($P_2$) and the deepest level ($P_5$). The results are presented in Table 4. By removing $P_5$ and excluding $P_2$ from prediction, the proposed method achieves the best trade-off. This confirms that effective scale selection is crucial for this task. Instead of uniformly utilizing all pyramid levels, the proposed strategy selectively suppresses unreliable scales while preserving informative intermediate representations, leading to more robust detection results.

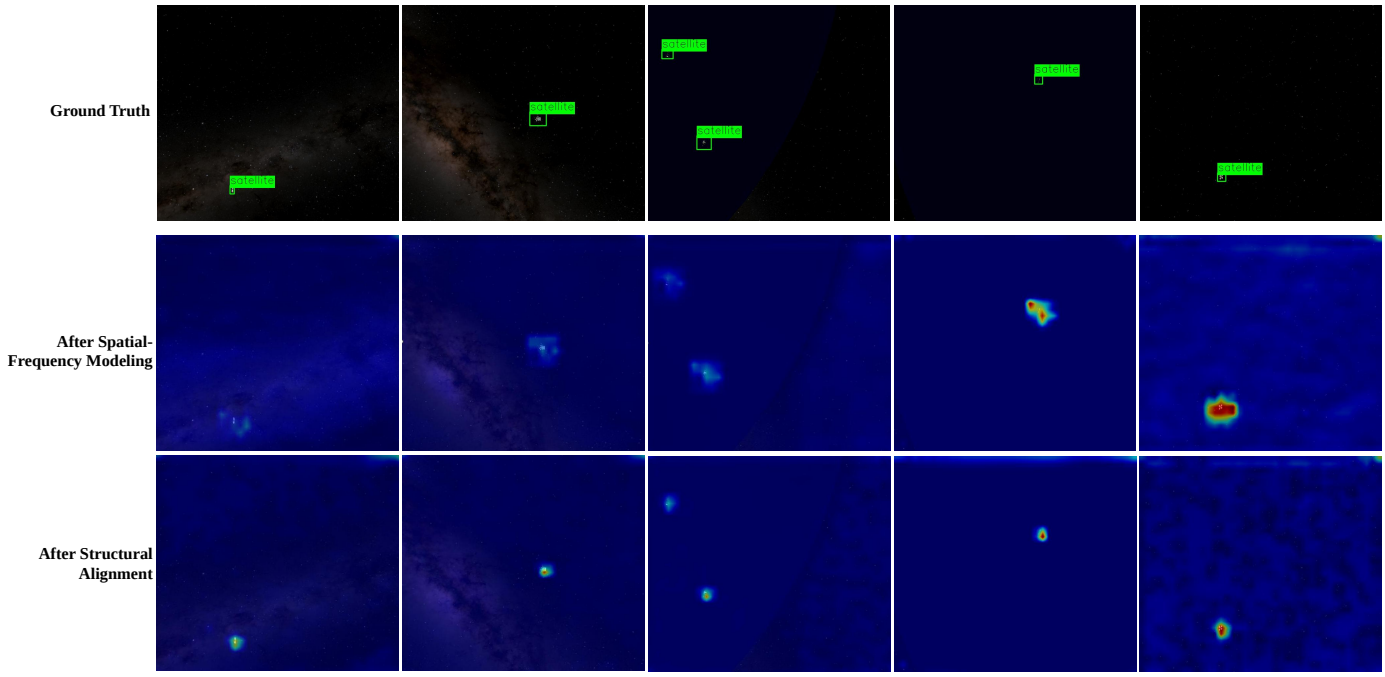


**Figure 6.** Visualization of feature evolution at different stages. From top to bottom: ground truth, features after spatial-frequency modeling, and features after structural alignment. It can be observed that target responses become progressively more salient, while background interference is gradually suppressed, leading to improved localization consistency. The color map represents the feature intensity, with warmer colors indicating higher activation values and cooler colors indicating lower values.

Overall, these results validate that both component-level design and scale-level selection are essential for handling large-scale sparse scenes under complex background noise conditions. The proposed framework achieves improved performance by jointly optimizing feature representation and scale utilization.

**Table 4.** Ablation study on noise-aware scale pruning. "P2 Head" indicates whether predictions are made on the $P_2$ level, and "P5 Layer" denotes whether the deepest feature level is retained in the backbone.

| P5 Layer | P2 Head | Precision (%) | Recall (%) | mAP@50 (%) | mAP@50-95 (%) | Params | FLOPs |
|---|---|---|---|---|---|---|---|
| ✓ | × | 74.48 | 74.68 | 74.30 | 26.87 | 9.91 M | 19.2 G |
| × | ✓ | 77.18 | 71.24 | 74.68 | 27.85 | 3.62 M | 30.7 G |
| × | × | 77.06 | 76.40 | 79.15 | 28.47 | 3.46 M | 23.0 G |

*Note:* ✓ denotes the presence of the corresponding module in the model configuration, whereas × denotes its absence.

### 3.3.3. Effect of Frequency Representation

To investigate the impact of different frequency representations, we replace the FFT-based spectral branch with a wavelet-based counterpart while keeping the remaining architecture unchanged. For the wavelet-based variant, a one-level discrete wavelet transform (DWT) with the Haar wavelet basis is employed. The input feature map is decomposed into four sub-bands (LL, LH, HL, and HH), corresponding to low-frequency and high-frequency components. Symmetric boundary extension is adopted to mitigate boundary artifacts. To ensure a fair comparison, the resulting wavelet features are processed using the same network architecture and training configuration as the proposed FFT-based implementation.

The quantitative results are reported in Table 5. As shown, the FFT-based representation achieves superior performance across all evaluation metrics while maintaining lower computational complexity. This suggests that the global spectral characteristics captured by FFT are more effective for distinguishing weak targets from background interference in satellite imagery.

**Table 5.** Comparison of different frequency representations within the spectral refinement module.

| Method | Precision | Recall (%) | mAP@50 (%) | mAP@50-95 (%) | Params | FLOPs | FPS |
|---|---|---|---|---|---|---|---|
| Wavelet-based | 75.23 | 72.10 | 72.46 | 26.75 | 5.79 M | 25.8 G | 53.91 |
| FFT-based (Ours) | 77.06 | 76.40 | 79.15 | 28.47 | 3.46 M | 23.0 G | 72.69 |

#### 3.3.4. Effect of Spectral Kernel Size

As shown in Table 6, the proposed $3 \times 3$ spectral kernel achieves the highest detection accuracy while maintaining the lowest computational complexity. Increasing the kernel size leads to a noticeable decline in mAP@50, despite the larger receptive field. This suggests that local spectral interactions are sufficient for capturing the structured frequency patterns relevant to weak target detection, whereas larger kernels tend to over-smooth discriminative spectral responses and introduce unnecessary computational overhead.

**Table 6.** Effect of spectral kernel size in the SCR module.

| Kernel | Precision | Recall | mAP@50 | mAP@50-95 | Params | FLOPs | FPS |
|---|---|---|---|---|---|---|---|
| $3 \times 3$ | 77.06 | 76.40 | 79.15 | 28.47 | 3.46 M | 23.0 G | 72.69 |
| $5 \times 5$ | 82.01 | 75.54 | 77.15 | 27.95 | 4.84 M | 28.2 G | 64.01 |
| $7 \times 7$ | 75.03 | 73.51 | 71.80 | 28.01 | 6.91 M | 36.0 G | 53.48 |

### *3.4. Robustness Analysis*

To evaluate the stability and reproducibility of the proposed framework, we conducted three independent training runs using different random seeds under the same experimental settings on the SODv2 dataset. The quantitative results are summarized in Table 7. The proposed method exhibits consistently stable performance across different random initializations. In particular, the standard deviation of mAP@50 is only 0.49%, while the standard deviation of mAP@50-95 is 0.24%. Similar low variances are observed for precision and recall. These results indicate that the performance gains achieved by DRSS-Net are reproducible and not attributable to favorable random initialization. The small variance across multiple runs further demonstrates the robustness and training stability of the proposed framework.

**Table 7.** Robustness analysis of the proposed method across three independent runs on the SODv2 dataset.

| Metric | Run 1 | Run 2 | Run 3 | Mean $\pm$ Std |
|---|---|---|---|---|
| Precision (%) | 77.06 | 76.88 | 77.84 | $77.26 \pm 0.42$ |
| Recall (%) | 76.40 | 75.18 | 76.09 | $75.89 \pm 0.52$ |
| mAP@50 (%) | 79.15 | 78.06 | 78.17 | $78.46 \pm 0.49$ |
| mAP@50-95 (%) | 28.47 | 28.16 | 27.88 | $28.17 \pm 0.24$ |

To further demonstrate the robustness of DRSS-Net, we conducted three independent runs on the SODv2 dataset. We performed Welch's $t$-tests against the strongest baseline (YOLOv12). As shown in Table 8, all improvements are statistically significant ($p < 0.05$),

confirming that the proposed method consistently outperforms the baseline and is robust to random initialization.

**Table 8.** Robustness analysis and statistical significance of the proposed method compared to YOLOv12 on SODv2 across three independent runs.

| Metric | Baseline (Mean) | Ours (Mean) | Improvement | Welch *t*-Test *p*-Value | Significant ($p < 0.05$) |
|---|---|---|---|---|---|
| Precision (%) | 74.77 | 77.26 | 2.49 | 0.0261 | Yes |
| Recall (%) | 73.15 | 75.89 | 2.74 | 0.0095 | Yes |
| mAP@50 (%) | 74.24 | 78.46 | 4.22 | 0.0093 | Yes |
| mAP@50-95 (%) | 27.14 | 28.17 | 1.03 | 0.0496 | Yes |

Challenging Detection Scenario

Figure 7 presents a representative challenging scenario under extreme bright-star interference. In this case, strong stellar emissions generate highly concentrated responses that dominate the surrounding feature representation. Although the proposed framework effectively suppresses most structured background patterns, the target signal becomes partially submerged within the intense stellar response, resulting in a missed detection. This observation suggests that the proposed method remains sensitive to cases where the target-to-background contrast is extremely low. Nevertheless, such scenarios are relatively uncommon in the evaluated datasets, and the overall results demonstrate strong robustness across a wide range of observation conditions.

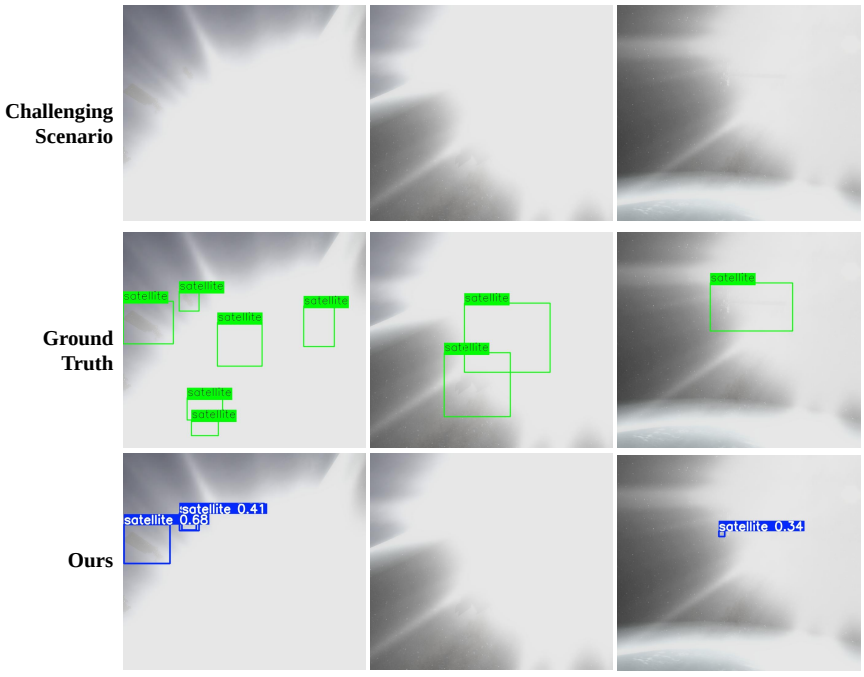


**Figure 7.** Detection results under extreme bright-star interference. The target response is partially overwhelmed by strong stellar emissions, resulting in a missed detection. The blue boxes represent the detected targets, while the green boxes indicate the ground-truth annotations.

## 4. Conclusions

In this paper, we investigated the problem of satellite object detection from a representation perspective, focusing on the intrinsic difficulty of separating weak targets from background interference in large-scale space observation scenarios. We argued that this challenge arises from the limitation of single-domain feature representations, where target and background become entangled and difficult to distinguish.

To address this issue, we proposed a lightweight spatial-frequency framework that improves feature discriminability through coordinated modeling across complementary representation spaces. By alternately exploiting spatial structural consistency and frequency-domain regularity, the proposed design enables more reliable suppression of structured background patterns and enhancement of target responses. In addition, we introduced a stability-guided scale selection strategy that focuses on informative intermediate feature levels, providing a better balance between structural preservation and semantic abstraction. Extensive experiments demonstrate that the proposed framework achieves strong detection performance while maintaining a compact model size and efficient computation. These results highlight the importance of considering representation reliability, rather than solely increasing model complexity, for satellite object detection in space surveillance scenarios.

Although the proposed framework demonstrates strong performance across diverse observation conditions, extremely severe motion blur may remain challenging for frequency-domain representation learning. In such cases, motion-induced spectral distortion can weaken the discriminability of frequency features and reduce the effectiveness of the proposed SFC-Block. Future work will investigate adaptive enhancement mechanisms for extremely high-intensity interference regions and motion-aware modeling and temporal information integration to improve robustness under high-speed imaging conditions.

**Author Contributions:** Conceptualization, R.H. and Q.W.; methodology, H.P.; software, J.L.; validation, H.P., R.H., and J.L.; formal analysis, Q.W.; investigation, H.P.; resources, H.P.; data curation, R.H.; writing original draft preparation, R.H.; review and editing, Q.W.; visualization, J.L.; supervision, Q.W.; project administration, H.P.; funding acquisition, H.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by Aviation Key Laboratory of Science and Technology on Aerospace Vehicle (No. J2025-STAV-03-001) and Key Laboratory of Polar Science, Ministry of Natural Resources (No. KP202603).

**Institutional Review Board Statement:** Not applicable.

**Data Availability Statement:** The data used in this study are publicly available. The satellite object detection dataset can be accessed at https://github.com/AEL-Lab/satellite-object-detection-dataset-v2, accessed on 3 January 2026. No new datasets were generated in this work.

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

## References

1. Du, J.; Lei, X.; Sang, J. A space surveillance satellite for cataloging high-altitude small debris. *Acta Astronaut.* **2019**, *157*, 268–275. [CrossRef]
2. Tao, J.; Cao, Y.; Ding, M. SDebrisNet: A spatial-temporal saliency network for space debris detection. *Appl. Sci.* **2023**, *13*, 4955. [CrossRef]
3. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016*; IEEE: New York, NY, USA, 2016; pp. 779–788.
4. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*; IEEE: New York, NY, USA, 2017; pp. 7263–7271.
5. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
6. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
7. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023*; IEEE: New York, NY, USA, 2023; pp. 7464–7475.
8. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. RTMDet: An empirical study of designing real-time object detectors. *arXiv* **2022**, arXiv:2212.07784. [CrossRef]

9. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2020; pp. 213–229.
10. Zheng, D.; Dong, W.; Hu, H.; Chen, X.; Wang, Y. Less is more: Focus attention for efficient DETR. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023*; IEEE: New York, NY, USA, 2023; pp. 6674–6683.
11. Dai, X.; Chen, Y.; Yang, J.; Zhang, P.; Yuan, L.; Zhang, L. Dynamic DETR: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021*; IEEE: New York, NY, USA, 2021; pp. 2988–2997.
12. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024*; IEEE: New York, NY, USA, 2024; pp. 16965–16974.
13. Rabbi, J.; Ray, N.; Schubert, M.; Chowdhury, S.; Chao, D. Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network. *Remote Sens.* **2020**, *12*, 1432. [CrossRef]
14. Xiao, Y.; Xu, T.; Xin, Y.; Li, J. FBRT-YOLO: Faster and better for real-time aerial image detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2025; Volume 39, pp. 8673–8681.
15. Hu, L.; Yuan, J.; Cheng, B.; Xu, Q. CSFPR-RTDETR: Real-time small object detection network for UAV images based on cross spatial-frequency domain and position relation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5638219. [CrossRef]
16. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef] [PubMed]
17. Wang, K.; Fu, X.; Huang, Y.; Cao, C.; Shi, G.; Zha, Z.J. Generalized UAV Object Detection via Frequency Domain Disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023*; IEEE: New York, NY, USA, 2023; pp. 1064–1073.
18. Lin, S.; Zhang, Z.; Huang, Z.; Lu, Y.; Lan, C.; Chu, P.; You, Q.; Wang, J.; Liu, Z.; Parulkar, A.; et al. Deep Frequency Filtering for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023*; IEEE: New York, NY, USA, 2023; pp. 11797–11807.
19. Sun, Y.; Xu, C.; Yang, J.; Xuan, H.; Luo, L. Frequency-Spatial Entanglement Learning for Camouflaged Object Detection. In *European Conference on Computer Vision*; Springer: Cham Switzerland, 2024; pp. 343–360.
20. Zhou, H.; Tian, C.; Zhang, Z.; Li, C.; Xie, Y.; Li, Z. Frequency-aware Feature Aggregation Network with Dual-task Consistency for RGB-T Salient Object Detection. *Pattern Recognit.* **2024**, *146*, 110043.
21. Tian, Y.; Ye, Q.; Doermann, D. Yolov12: Attention-centric real-time object detectors. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 78433–78457.
22. Zhang, W.; Hu, P. Toward onboard ai-enabled solutions to space object detection for space sustainability. *arXiv* **2025**, arXiv:2505.01650. [CrossRef]
23. Zhang, W.; Hu, P. An Edge AI Solution for Space Object Detection. *arXiv* **2025**, arXiv:2505.13468. [CrossRef]
24. Liu, Y.; Bian, C.; Nie, H.; Chen, S.; Yang, Z. A Large-Scale Synthetic Benchmark Dataset for Non-Cooperative Space Target Perception. *Sci. Data* **2025**, *12*, 1780. [CrossRef] [PubMed]
25. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2020; pp. 12993–13000.
26. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
27. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and flexible image augmentations. *Information* **2020**, *11*, 125. [CrossRef]