



Article

RShDet: An Adaptive Spectral-Aware Network for Remote Sensing Object Detection Under Haze Corruption

Wei Zhang ^{1,†}, Yuantao Wang ^{2,†}, Haowei Yang ³ and Xuerui Mao ^{1,4,*}¹ School of Interdisciplinary Science, Beijing Institute of Technology, Beijing 100081, China² School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China; wangyt@mail.ncut.edu.cn³ School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China⁴ School of Interdisciplinary Science, Beijing Institute of Technology (Zhuhai), Zhuhai 519088, China

* Correspondence: xmao@bit.edu.cn

† These authors contributed equally to this work.

Highlights

What are the main findings?

- RShDet is a novel frequency-domain adaptive framework for haze-degraded remote sensing object detection that unifies image enhancement and object detection within a spectral-aware learning paradigm. By dynamically modeling object-relevant spectral characteristics and enabling adaptive cross-task interaction, the framework improves detection robustness and reliability under severe haze degradation.
- RShDet establishes a new state of the art in haze-degraded remote sensing object detection, enabling significantly more reliable object localization under severe atmospheric degradation. These gains are consistently validated across both synthetic and real-world hazy benchmarks, including a +2.16% mAP50 improvement over DHCNet on Hazy-DOTA.

What are the implications of the main findings?

- The findings demonstrate that frequency-domain joint optimization is a highly effective paradigm for haze-robust remote sensing object detection. By enabling adaptive spectral-aware learning and facilitating deeper integration between enhancement and detection processes, this approach provides a principled solution for mitigating atmospheric degradation and improving detection reliability under adverse visual conditions.
- The proposed Hazy-DOTA dataset, together with the frequency-aware enhancement and cross-task integration framework, establishes a reproducible benchmark and methodological foundation for future research on haze-robust object detection across diverse remote sensing platforms, including satellite and UAV systems.

Abstract

Remote sensing (RS) object detection faces intrinsic challenges arising from the overhead imaging paradigm and the diversity of climatic conditions. In particular, atmospheric phenomena such as clouds and haze cause severe visual degradation, making reliable object detection difficult. However, most existing detectors are developed under clear-weather conditions, which limits their generalization capability in realistic haze-degraded RS scenarios. To alleviate this issue, an adaptive spectral-aware network for RS object detection under haze interference is proposed, termed RShDet, which is designed to handle both high-altitude RS imagery and low-altitude Unmanned Aerial Vehicle (UAV) scenarios. Firstly, the Object-Centered Dynamic Enhancement (OCDE) module dynamically adjusts the spatial positions of key-value pairs through query-agnostic offsets, enabling the network to emphasize object-relevant regions while suppressing haze-induced background interference.



Academic Editor: Pedro Melo-Pinto

Received: 13 February 2026

Revised: 23 March 2026

Accepted: 27 March 2026

Published: 29 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Secondly, the Dynamic Multi-Spectral Perception and Filtering (DSPF) module introduces a multi-spectral attention mechanism that adaptively selects informative frequency components, thereby enhancing discriminative feature representations in hazy environments. Thirdly, the Frequency-Domain Multi-Feature Fusion (FDMF) module employs learnable weights to complementarily integrate amplitude and phase information in the frequency domain, enabling effective cross-task feature interaction between the enhancement and detection branches. Extensive experiments demonstrate that RShDet consistently achieves superior detection performance under hazy conditions across both synthetic and real-world benchmarks. Specifically, it achieves improvements of 2.4% mAP50 on Hazy-DOTA, 1.9% mAP on HazyDet, and 2.33% mAP on the real-world foggy dataset RTTS, surpassing existing state-of-the-art methods.

Keywords: remote sensing; object detection; multi-spectral attention; dynamic multi-spectral perception and filtering; haze robustness

1. Introduction

Haze-degraded remote sensing (RS) object detection refers to the task of automatically identifying and localizing objects of interest from aerial or satellite imagery affected by atmospheric haze or fog [1,2]. Unlike clear-weather RS object detection, this task must address the additional challenge of visibility degradation caused by atmospheric scattering. Although recent advances have significantly improved detection performance on clear RS imagery [3–5], their effectiveness is substantially reduced when applied to haze-degraded scenarios. This is because haze interference reduces image contrast and obscures structural details, making object features less distinguishable from complex backgrounds. Furthermore, RS imagery often contains small objects and large-scale scenes, small targets occupy few pixels and often have subtle edges easily masked by haze, which further amplifies the negative impact of haze degradation. In particular, small targets usually occupy only a few pixels and exhibit weak edge and texture cues, making them especially vulnerable to haze interference. Consequently, haze-degraded RS object detection remains a critical yet challenging problem in practical RS applications.

Notably, in natural scenes, extensive efforts have been devoted to improving the performance of object detection under degraded imaging conditions. Image adaptation methods restore image features according to degradation-related information, thereby benefiting subsequent object detection. More recent works [6–9] introduce differentiable image filters strategies that adaptively refine degraded features, leading to improved detection robustness. Domain adaptation methods [10,11] address degradation by reducing the domain discrepancy between clear images and degraded images, enabling the detector to learn domain-invariant feature representations. These methods address domain bias, facilitating the generalization of networks across different visual conditions. In addition, multi-task learning methods [12,13] jointly optimize image restoration and object detection within a unified framework, where shared feature representations facilitate mutual enhancement between restoration and detection tasks.

Despite the progress achieved in natural scene object detection under adverse weather conditions, research on haze-degraded object detection in the RS domain remains relatively limited. For example, DHC-Net [2] combines an adaptive dehazing module with centroid-guided contrastive learning to resolve the issues of low-quality images under haze conditions and class imbalance in RS datasets. By combining affine Lie group-driven dynamic convolution and weather-adaptive feature enhancement, ALGC-WFEMNet [14]

delivers robust and efficient road condition detection under rain, fog, and snow conditions. NIRNet [1] leverages noise perception and dual-path incentive calibration to mitigate cloud-induced feature ambiguity and improve detection robustness in RS imagery. While these methods achieve effective results, several significant challenges remain unaddressed. First, existing models often exhibit a pronounced frequency bias, primarily concentrating on low-frequency components or a restricted range of frequency bands. This limitation leads to the underutilization of vital high-frequency details, which are indispensable for delineating fine structures and object boundaries in RS imagery. Consequently, the lack of frequency-adaptive flexibility hinders the robustness of these models against complex hazy degradations. Furthermore, current frameworks frequently overlook the synergistic interaction between image enhancement and object detection in the frequency domain, thereby failing to fully exploit complementary cues under severe haze interference.

To address these limitations, we propose an adaptive spectral aware network for RS object detection under haze interference, termed RShDet, which integrates image enhancement with object detection in the frequency domain to improve detection performance under hazy conditions. During the extraction of informative frequency components, complex backgrounds in degraded images may interfere with the discriminability of foreground objects. To mitigate the negative impact caused by irrelevant backgrounds and ensure that the model focuses on object-relevant features, we develop an Object-Centered Dynamic Enhancement (OCDE) Module. Specifically, by dynamically adjusting sampling positions through learned offsets, OCDE focuses on important regions and effectively captures more informative features.

From the perspective of feature learning, we further analyze hazy images in the frequency domain to recover critical visual details that are often suppressed under haze conditions. Since images consist of diverse frequency components, both low- and high-frequency signals carry valuable information. However, CNNs are inherently more sensitive to low-frequency components, which leads to insufficient exploitation of high-frequency details. To compensate for this imbalance in feature learning, we design a Dynamic Multi-Spectral Perception and Filtering (DSPF) Module that selectively emphasizes informative frequency components, thereby improving feature representation in hazy environments. Specifically, the DSPF captures all frequency components generated by the Discrete Cosine Transform (DCT). The DCT coefficients of the same frequency are mapped to one channel, and each frequency selection switch is stacked on a specific frequency channel to choose the frequency components that flow into the model. Furthermore, we introduce a Frequency-Domain Multi-Feature Fusion (FDMF) module, which utilizes learnable weights to complementarily fuse amplitude and phase representations in the frequency domain, enabling effective cross-task feature interaction between the enhancement and detection branches.

In addition, to facilitate research on haze-degraded RS object detection, we construct a haze-corrupted RS dataset named Hazy-DOTA, which is derived from the widely used DOTA-v2.0 [15] benchmark by simulating realistic haze degradation with varying density levels. DOTA-v2.0 serves as a comprehensive, large-scale dataset for object detection in aerial images. Experimental results demonstrate that RShDet outperforms the baseline methods by 7.1% mAP50 on Hazy-DOTA, 3.1% mAP on HazyDet, and 6.63% mAP on RTTS. Compared with current state-of-the-art (SOTA) methods, it further achieves gains of 2.4% mAP50, 1.9% mAP, and 2.33% mAP on these datasets, respectively. As illustrated in Figure 1, YOLOv8, a widely used baseline for RS object detection, struggles to accurately detect objects in hazy images, which often leads to missed detections. By comparison, our method achieves more accurate detection results under hazy conditions.

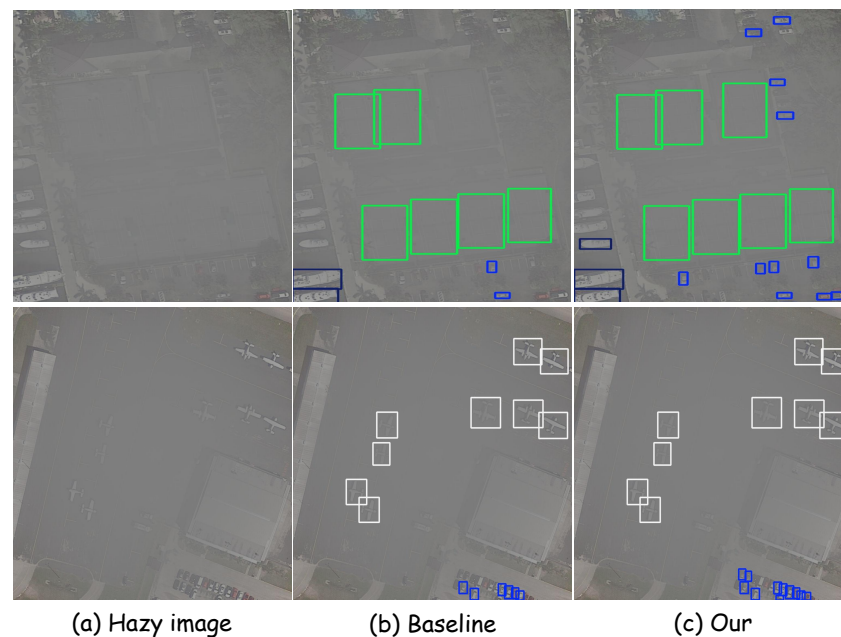


Figure 1. The proposed RShDet is an RS object detection model designed to handle hazy weather conditions. From left to right: haze-corrupted images from our Hazy-DOTA dataset, results from the baseline YOLOv8, and results from the proposed RShDet. The YOLOv8 detector has difficulty recognizing small vehicles and ships in hazy images. By comparison, RShDet effectively addresses these inaccuracies and achieves robust detection performance under haze degradation. Note: White bounding boxes for planes, blue for small vehicles, green for tennis courts, and dark blue for ships.

In brief, our contributions are summarized as follows.

- An Adaptive Spectral Aware Network, termed RShDet, is proposed for both high-altitude RS imagery and low-altitude UAV scenarios object detection under hazy weather. Motivated by theoretical insights into frequency utilization, RShDet integrates image restoration and object detection within a unified spectral-aware framework, exploiting complementary amplitude and phase information to improve detection performance under haze degradation.
- A dynamic object- and frequency-aware enhancement framework is designed that jointly improves feature quality and detection robustness. The first is OCDE module that enables the model to focus on object-relevant regions while suppressing background interference, and DSPF module that identifies and suppresses trivial frequency components based on DCT coefficients to mitigate haze-induced noise.
- We construct Hazy-DOTA, a new haze-degraded RS object detection dataset based on DOTA with realistically simulated haze effects, providing a dedicated benchmark for adverse weather evaluation. Extensive experiments on both synthetic and real-world haze datasets demonstrate that RShDet consistently outperforms recent SOTA methods, validating its effectiveness and generalization across diverse haze scenarios.

2. Related Works

2.1. Object Detection in Clear Weather

In the RS domain, Ren et al. [16] improved the Region Proposal Network (RPN) by generating anchors with more diverse scales and aspect ratios, and enhanced the feature extraction network to produce higher-resolution feature maps. Furthermore, they proposed a context-aware detection model that incorporates surrounding environmental information, significantly improving small object detection performance. Bai et al. [17] built upon Faster R-CNN by adopting a Dense Residual Network (DRNet) as the backbone and

introducing the ROI Align operation [18], achieving high-precision building detection in RS imagery. To address the challenges posed by object orientation in aerial imagery, Liu et al. [19] proposed RR-CNN, a rotated region-based CNN framework that accurately extracts features from arbitrarily oriented regions and precisely localizes rotated objects. Similarly, Zhang et al. [20] proposed the CFF-SDN ship detection network, which leverages multi-layer convolutional feature fusion to improve high-precision horizontal bounding box (HBB) ship detection. Han et al. [21] incorporated a capsule reasoning module into DETR and improved multi-scale object detection performance through enhanced feature aggregation. Li et al. [22] proposed a Strong Feature Extraction and Pre-screening Network (SFEPN) based on Deformable DETR, effectively alleviating scale variation and small object matching issues. In addition, Li et al. [23] developed a Transformer-based RS object detection framework by redesigning the Transformer architecture to better capture global spatial relationships and strengthen the interaction between feature embeddings and object queries.

2.2. Object Detection in Hazy Weather

Object detection in hazy conditions has garnered significant attention in recent years due to its critical importance in RS research. Recent efforts have focused on leveraging adaptive enhancement methods to mitigate haze-induced degradation and strengthen detection performance. Li et al. [6,7] introduced a differentiable image processing framework designed to enhance images captured under challenging weather conditions, thereby facilitating object detection. Kalwar et al. [8] developed a gated differentiable image processing module that improves detection performance by adaptively weighting multiple image filters. Despite their effectiveness, both methods rely on image filters with fixed parameter settings applied uniformly across pixels, which constrains their robustness when handling spatially non-uniform degradation. To address this limitation, Xi et al. [9] proposed the DEDet framework, which predicts pixel-level exposure parameter maps for individual image filters and progressively refines degraded images through parameterized filtering. In a different vein, Zhang et al. [24] introduced a chain-of-thought (CoT) prompted adaptive enhancement strategy, where CoT prompts guide the model to adjust enhancement policies according to the inferred degradation type, leading to improved detection robustness in complex environments. Additionally, several studies [25–27] have incorporated Laplacian pyramid representations into image enhancement by decomposing images into low- and high-frequency components to separately address global degradation and fine-grained details. Nevertheless, although these approaches achieve certain performance gains, they generally neglect the feature-level interaction between restoration and detection, resulting in limited overall improvement.

A growing body of literature has explored domain adaptation techniques to alleviate the distribution discrepancy between the source and target domains [28–33]. Wang et al. [31] proposed Robust-YOLO, which introduces an image calibration approach that captures domain-invariant features through adversarial learning. Along a different line, Pang et al. [32] proposed MCNet, where a magnitude corrector is employed to suppress task-irrelevant components in the frequency domain. Despite their effectiveness, the performance of these domain adaptation methods remains highly dependent on the diversity and representativeness of the training data. Furthermore, most existing approaches primarily focus on reducing domain discrepancies while overlooking how image degradation directly affects feature extraction, which ultimately constrains their detection performance.

A further line of research integrates image enhancement and object detection within a unified framework, which has been reported to improve robustness on degraded imagery [12,13,34–36]. Huang et al. [13] linked restoration and detection by sharing early

convolutional representations, allowing the detector to directly benefit from the cleaner cues produced by the restoration branch under challenging conditions. Cui et al. [34] proposed MAET, which exploits degradation transformations to encourage the model to capture intrinsic structural patterns that are useful for detection in adverse weather. Wang et al. [12] introduced TogetherNet, where a dynamic transformer-based feature enhancement module coordinates the two subtasks and strengthens recognition under severe degradation. Wang et al. [37] further developed RDMNet by explicitly modeling degradation to steer both restoration and detection, improving generalization across diverse weather types.

In RS imagery, scenes are usually large in scale and contain numerous small objects, which makes object detection particularly challenging. Since these objects occupy only a few pixels, their feature representations are often weak. This issue becomes more severe under adverse weather conditions, where degradation factors such as haze and low contrast obscure object boundaries, suppress fine-grained details, and increase background interference, thus making small target detection particularly difficult. Wang et al. [38] developed MDD-ShipNet, which combines defogging enhancement based on multiple filters with a weighted bidirectional feature pyramid network to improve robustness to scale variation in hazy environments. Hu et al. [39] proposed HRAOD, a knowledge distillation framework that facilitates knowledge transfer from clean to hazy domains by leveraging gradient-guided feature imitation and degradation-weighted response distillation, thereby enabling the detector to learn degradation-resilient representations. Zhang et al. [40] proposed Q-Norm, a quality-adaptive normalization method that leverages image quality features to learn normalization parameters for implicit cross-domain feature alignment, while adaptively modulating the mean and variance through a quality assessment model. Feng et al. [41] established HazyDet, the first large-scale benchmark for drone view object detection under haze, and proposed DeCoDet, a depth conditioned detector that dynamically modulates feature representations to overcome severe visual degradation.

In domain adaptation, Zhao et al. [42] proposed MRT, a masked retraining teacher-student framework that integrates a masked autoencoder branch and a selective retraining mechanism to improve target-domain feature learning while mitigating the detrimental impact of low-quality pseudo labels. Chen et al. [43] proposed SF-TMAT, a statistically feedback-driven threshold and mask adjustment teacher-student framework, which enhances robustness to scale variations and background clutter through a dynamic feedback-based mask adjustment strategy for learning more robust domain-invariant feature representations.

In the multi-task setting, Liu et al. [44] introduced DFENet, a dual-branch detection model that leverages haze information to dynamically adjust feature weights, thereby guiding the fusion of restoration and detection branches to suppress noise generated during the dehazing process. Wan et al. [45] proposed MS-FODN, a multitask learning framework that achieves parameter interaction through a shared encoder and utilizes a multiscale attention fusion module to aggregate haze free features for robust detection in RS.

Despite significant advancements in RS object detection, two critical challenges remain unaddressed. First, many existing methods exhibit an over-reliance on low-frequency signals, which inherently limits their ability to capture the fine-grained spectral details essential for precise object localization and classification. Second, current frameworks often lack a synergistic interaction between image enhancement and object detection within the frequency domain, resulting in suboptimal feature integration. Consequently, a frequency-aware multi-task framework that enables cross-task fusion while adaptively exploiting multi-band spectral cues is needed. To this end, we propose RShDet, an RS object detector tailored to address these limitations.

3. Methodology

3.1. Overview

The overall architecture of RShDet is shown in Figure 2a. The proposed framework adopts a dual-branch design, consisting of a detection branch and an enhancement branch. In this framework, hazy images are fed into three distinct encoder networks to facilitate feature interaction. A degradation representation encoder is first employed to extract degradation cues from hazy images. The resulting degradation features are then leveraged to guide both the enhancement and detection encoders.

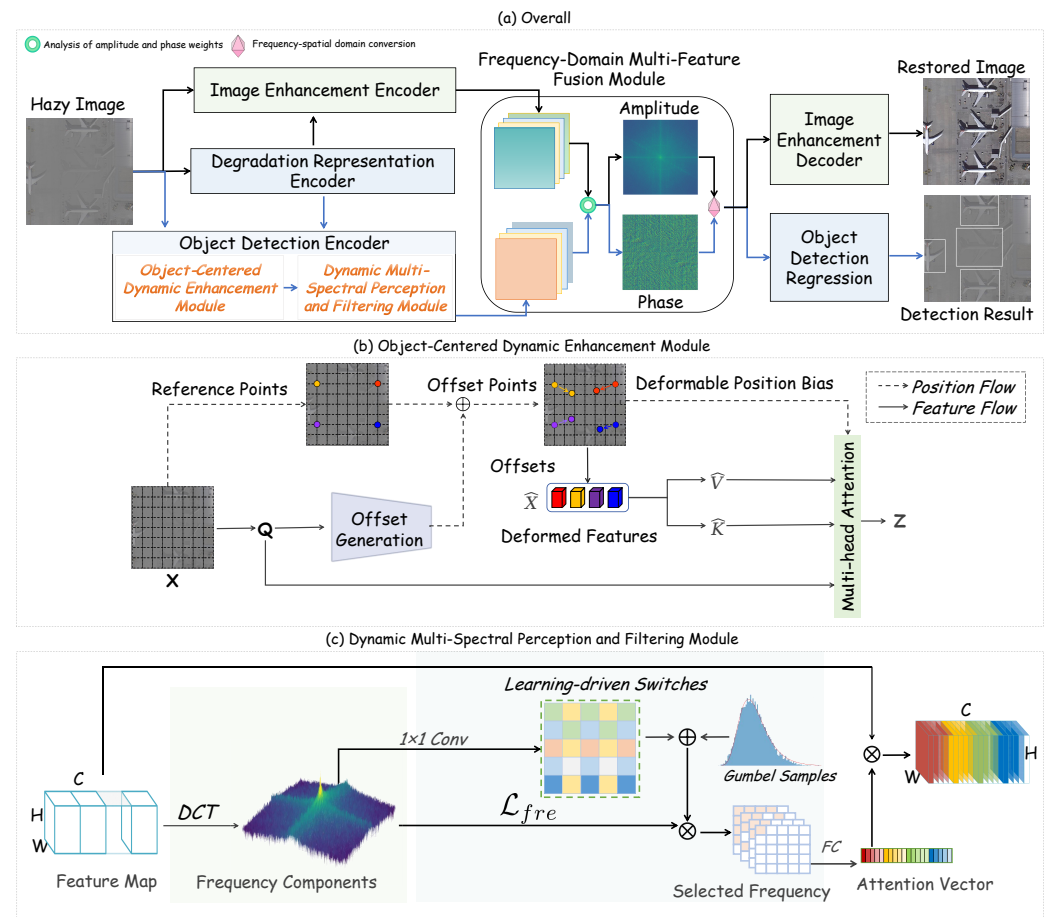


Figure 2. The architecture of RShDet. (a) Overall architecture of RShDet, where image enhancement and object detection are jointly optimized under hazy conditions; (b) Illustration of the proposed object-centered dynamic enhancement module, which enable the model to flexibly concentrate on important regions and extract richer features through offset sampling points; (c) Illustration of the proposed dynamic multi-spectral perception and filtering module, which utilize DCT and Learning-driven selection switches to compensate for the loss of high-frequency details. $\oplus$ denotes the element-wise sum operation, and $\otimes$ denotes the element-wise multiplication operation.

As depicted in Figure 2b, the OCDE module is first applied in the spatial domain to enhance object-relevant regions and suppress background interference through offset-based sampling. This step provides more object-centered feature representation for the subsequent frequency-domain analysis. As illustrated in Figure 2c, the DSPF module is then employed to perform frequency-domain modeling on these refined spatial features, so as to adaptively preserve informative spectral components while suppressing redundant ones. The FDMF module further fuses the features from the enhancement and detection branches in the amplitude and phase domains, so that degradation-aware appearance cues and object-related structural information can be jointly exploited, thereby strengthening the

network's representation capability for low-quality images. Therefore, OCDE, DSPF, and FDMF play complementary roles rather than overlapping ones. Finally, the fused features are subsequently delivered to an enhancement decoder and an object detection regression, which produce the restored images and the corresponding detection outputs, respectively.

During training, the entire network is optimized with a combination of multiple objectives, including detection, enhancement, contrastive, and frequency-based regularization. The total training loss is formulated as:

$$\mathcal{L}_{Total} = \lambda_1 \mathcal{L}_{Det} + \lambda_2 \mathcal{L}_{En} + \lambda_3 \mathcal{L}_{Fre} + \lambda_4 \mathcal{L}_{Ctr}, \quad (1)$$

where $\mathcal{L}_{Det}$ and $\mathcal{L}_{En}$ denote the object detection loss and the image enhancement loss, $\mathcal{L}_{Fre}$ and $\mathcal{L}_{Ctr}$ are the proposed frequency regularization term and the contrastive loss. Following previous works [37,46,47], we set the hyperparameters as $\lambda_1 = 0.2$, $\lambda_2 = 0.8$, $\lambda_3 = 0.1$, and $\lambda_4 = 0.1$, respectively. Specifically, the detection and enhancement losses are defined as:

$$\mathcal{L}_{Det} = \gamma \mathcal{L}_{IoU} + \mathcal{L}_{Cls} + \mathcal{L}_{Focal} \quad (2)$$

$$\mathcal{L}_{En} = \frac{1}{N} \sum_{i=1}^N \|\hat{Y}_i - Y_i\|_1 \quad (3)$$

where $\mathcal{L}_{IoU}$ and $\mathcal{L}_{Cls}$ denote the bounding box regression and classification losses, respectively. $\mathcal{L}_{Focal}$ is the focal loss [48], which mitigates the imbalance between positive and negative samples in object detection. The weight γ is set to 5 based on empirical evaluation. Here, Y_i is the enhanced output image, $\hat{Y}_i$ is the corresponding ground truth, and $\|\cdot\|_1$ indicates the L_1 norm.

3.2. Dynamic Multi-Spectral Perception and Filtering (DSPF) Module

The core of the designed DSPF is to extend conventional channel attention beyond the lowest-frequency response captured by GAP. Specifically, we use the multi-spectral attention mechanism based on the 2D discrete cosine transform (DCT) to capture comprehensive frequency information. Then, we leverage the learning-driven frequency filtering block to adaptively select informative frequency components and suppress trivial or noisy ones in a data-driven manner.

3.2.1. Image Frequency Utilization Analysis

To bolster the representational efficiency of Convolutional Neural Networks (CNNs), channel attention is frequently integrated to assign non-uniform significance to varying feature components via scalar weighting. In a conventional pipeline, GAP is employed as a computationally lean spatial descriptor to aggregate global statistics into channel-specific scalars, owing to its structural simplicity and minimal parameter overhead.

For an intermediate feature map $X \in \mathbb{R}^{C \times H \times W}$ characterized by C channels and spatial dimensions $H \times W$, the computation of the attention weight vector Att_{orig} is defined as:

$$\text{Att}_{\text{orig}} = \sigma(\text{Fc}(\text{GAP}(X))), \quad (4)$$

where $\text{Att}_{\text{orig}} \in \mathbb{R}^C$ represents the resulting channel-wise weight, and σ represents the sigmoid nonlinearity which constrains the weights to a specific range. Here, Fc signifies a transformation function—typically a fully connected layer—that captures inter-channel dependencies. The GAP operator effectively reduces the global spatial context of each individual channel into a single scalar value. Finally, the input feature map X is recalibrated by scaling each channel with its corresponding attention weight:

$$\hat{X}_i = \text{Att}_i X_i, \quad \text{where } i \in \{0, 1, \dots, C-1\}, \quad (5)$$

where $X_i \in \mathbb{R}^{H \times W}$ and $\hat{X}_i \in \mathbb{R}^{H \times W}$ denote the i -th channel of the input and recalibrated feature maps, respectively, and Att_i is the corresponding channel attention weight.

Based on digital signal processing principles, the 2D Discrete Cosine Transform (DCT) provides a spectral perspective to analyze feature map compression. The basis functions for the 2D DCT are defined as:

$$D_{h,w}^{i,j} = \cos\left(\frac{\pi h}{H}\left(i + \frac{1}{2}\right)\right) \cos\left(\frac{\pi w}{W}\left(j + \frac{1}{2}\right)\right). \quad (6)$$

The forward transformation, which maps an input feature map $x^{2d} \in \mathbb{R}^{H \times W}$ into the frequency spectrum $f^{2d} \in \mathbb{R}^{H \times W}$, is formulated as:

$$f_{h,w}^{2d} = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j}^{2d} D_{h,w}^{i,j} \quad \text{where } h \in \{0, 1, \dots, H-1\}, \quad w \in \{0, 1, \dots, W-1\}. \quad (7)$$

A key insight is revealed when $h = 0$ and $w = 0$. In this case, Equation (7) can be simplified as follows:

$$\begin{aligned} f_{0,0}^{2d} &= \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j}^{2d} \cos\left(\frac{0}{H}\left(i + \frac{1}{2}\right)\right) \cos\left(\frac{0}{W}\left(j + \frac{1}{2}\right)\right) \\ &= \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j}^{2d} \\ &= \text{GAP}(x^{2d})HW, \end{aligned} \quad (8)$$

where $f_{0,0}^{2d}$ represents the lowest frequency component of the 2D DCT. This derivation implies that the DCT serves as a generalized weighted combination of features, with the conventional GAP acting as a specific instance of this spectral representation. Consequently, the two-dimensional inverse discrete cosine transform can be expanded as:

$$\begin{aligned} x_{i,j}^{2d} &= \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} f_{h,w}^{2d} D_{h,w}^{i,j} \\ &= f_{0,0}^{2d} D_{0,0}^{i,j} + f_{0,1}^{2d} D_{0,1}^{i,j} + \dots + f_{H-1,W-1}^{2d} D_{H-1,W-1}^{i,j} \\ &= \text{GAP}(x^{2d})HW D_{0,0}^{i,j} + f_{0,1}^{2d} D_{0,1}^{i,j} + \dots + f_{H-1,W-1}^{2d} D_{H-1,W-1}^{i,j}, \end{aligned} \quad (9)$$

where $i \in \{0, 1, \dots, H-1\}, \quad j \in \{0, 1, \dots, W-1\},$

where $D_{0,0}^{i,j}$ is a constant. From Equation (9), it is evident that input image features are composed of diverse frequency components. However, standard channel attention as defined in Equation (4) relies exclusively on GAP, meaning the feature reconstruction x^{2d} effectively becomes:

$$x^{2d} = \underbrace{\text{GAP}(x^{2d})HW D_{0,0}^{i,j}}_{\text{utilized}} + \underbrace{f_{0,1}^{2d} D_{0,1}^{i,j} + \dots + f_{H-1,W-1}^{2d} D_{H-1,W-1}^{i,j}}_{\text{discarded}}. \quad (10)$$

This theoretical decomposition highlights that GAP-based attention only retains the lowest frequency information, while discarding mid- and high-frequency components that are often critical for robust feature extraction.

Even when multiple predefined frequency components are selected in some works [49,50], the fixed nature of the selection set still poses significant limitations. Let $\mathcal{I}_{\text{fixed}} \subseteq \{(h, w)\}$ denote the manually chosen set of frequency indices. The filtered reconstruction in such cases is given by:

$$x_{\text{fixed}}^{2d} = \sum_{(h,w) \in \mathcal{I}_{\text{fixed}}} f_{h,w}^{2d} \cdot D_{h,w}^{ij}. \quad (11)$$

Since $\mathcal{I}_{\text{fixed}}$ is handcrafted and invariant to image content or degradation type, informative frequency components outside this set are irreversibly discarded. This static strategy lacks flexibility and fails to adapt to complex hazy scenarios, limiting generalization and robustness.

From a spectral perspective, we uncover two intrinsic bottlenecks in current feature representation paradigms: First, reliance on GAP can inadvertently suppress frequency cues such as edges and textures. These components are indispensable for object discriminability, particularly in imagery degraded by hazy weather. Second, fixed frequency selection strategies are unable to adapt to varying image content and degradation conditions, which limits flexibility and robustness in complex hazy scenarios. Motivated by these observations, we introduce two corresponding improvements, detailed in the subsequent parts.

3.2.2. Multi-Spectral Attention Mechanism

To enrich the information captured by the model and strengthen feature representation, we extend GAP to incorporate a broader range of 2D DCT frequency components, thereby enabling more comprehensive multi-spectral feature modeling.

Given an input feature map $X \in \mathbb{R}^{C \times H \times W}$, we first divide it uniformly along the channel dimension into n disjoint groups, denoted as $[X^0, X^1, \dots, X^{n-1}]$, where each group $X^i \in \mathbb{R}^{C' \times H \times W}$ contains $C' = C/n$ channels. DSPF employs a learning-driven frequency filtering block to adaptively select informative 2D DCT frequency components. The selected frequency components are then assigned to different channel groups for subsequent spectral compression. Let $\mathcal{I}_{\text{sel}} = \{(u_i, v_i)\}_{i=0}^{n-1}$ denote the selected frequency index set, where each selected frequency component is assigned to one channel group. Accordingly, for the i -th group, the spectral descriptor is computed as:

$$\begin{aligned} \text{Freq}^i &= \text{DCT}^{(u_i, v_i)}(X^i) \\ &= \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \mathbf{x}_{h,w}^i D_{h,w}^{u_i, v_i}, \end{aligned} \quad (12)$$

$$\text{where } i \in \{0, 1, \dots, n-1\}. \quad (13)$$

where $\mathbf{x}_{h,w}^i \in \mathbb{R}^{C'}$ denotes the channel vector of the i -th group at spatial location (h, w) , (u_i, v_i) denotes the selected 2D frequency index assigned to X^i , and $\text{Freq}^i \in \mathbb{R}^{C'}$ is the corresponding C' -dimensional spectral descriptor. The multi-frequency representation $\text{Freq} \in \mathbb{R}^C$ is formed by concatenating:

$$\text{Freq} = \text{Concat}\left(\left[\text{Freq}^0, \text{Freq}^1, \dots, \text{Freq}^{n-1}\right]\right). \quad (14)$$

Finally, the multi-spectral attention vector is computed as:

$$\text{Att}_{\text{ms}} = \sigma(\text{Fc}(\text{Freq})), \quad (15)$$

As a result, our approach generalizes the original GAP algorithm to multi-spectral components, thereby significantly enhancing the feature representation. The overall workflow of DSPF is illustrated in Figure 3.

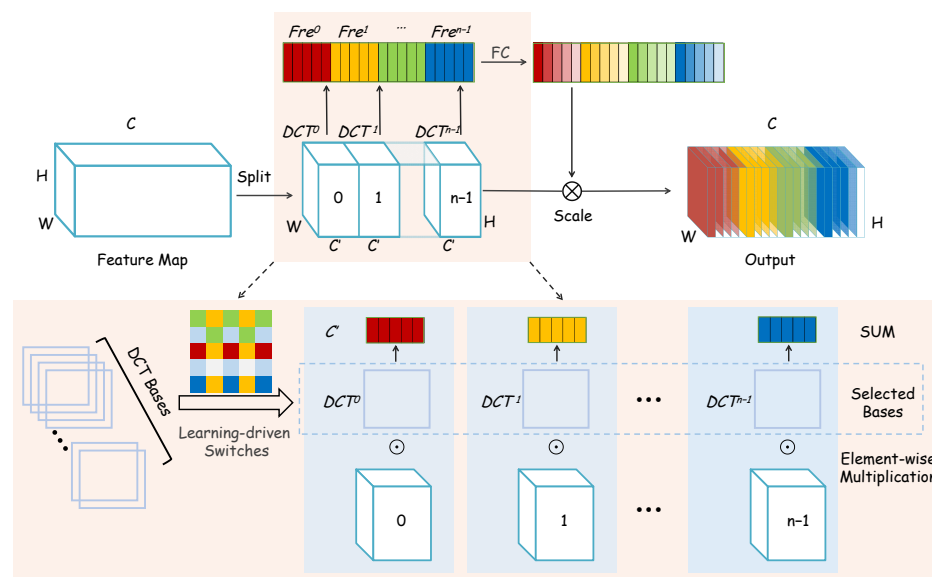


Figure 3. Detailed workflow of the DSPF module.

3.2.3. Learning-Driven Frequency Filtering Block

Feature representations in visual perception inherently span a wide range of frequency components; however, their contributions to object detection are highly unbalanced. In particular, certain frequency components may introduce redundant or noisy information, which degrades detection robustness. To address this issue, we propose a Learning-driven Frequency Filtering Block that explicitly evaluates the importance of different frequency components and selectively preserves the informative ones. The core idea is to introduce learnable binary switches that assign each frequency component a selection indicator, such that useful frequencies are retained while less relevant ones are suppressed.

As depicted in Figure 2c, the complete DCT frequency tensor is first fed into a 1×1 convolutional layer that predicts two logits for each candidate frequency component, corresponding to the retain and suppress states, respectively. These logits are normalized through a softmax function to model binary selection probabilities. During training, we adopt the Gumbel-Softmax approximation [51] to enable differentiable optimization of the discrete frequency selection process. During inference, the soft approximation is replaced by a hard binary decision: a frequency component is retained if the retain logit is larger than the suppress logit; otherwise, it is discarded.

Formally, let $f_{h,w}^{2d}$ denote the DCT coefficient at frequency index (h, w) . During inference, the learning-driven frequency filtering block is represented by a binary selection function S , where

$$S(f_{h,w}^{2d}) = \begin{cases} 1, & z_{h,w}^{\text{ret}} > z_{h,w}^{\text{sup}} \\ 0, & \text{otherwise,} \end{cases} \quad (16)$$

where $z_{h,w}^{\text{ret}}$ and $z_{h,w}^{\text{sup}}$ denote the logits corresponding to the retain and suppress states, respectively. The selected frequency index set can therefore be written as

$$\mathcal{I}_{\text{sel}} = \{(h, w) \mid S(f_{h,w}^{2d}) = 1\}. \quad (17)$$

Only the selected frequency components in $\mathcal{I}_{\text{sel}}$ are retained and subsequently assigned to different channel groups for group-wise spectral compression. To encourage balanced frequency usage and reduce undesirable preference toward particular bands, we impose an additional regularizer $\mathcal{L}_{\text{Fre}}$ [46], defined as

$$\mathcal{L}_{Fre} = \log \left(\sum_{h=0}^{H-1} \sum_{w=0}^{W-1} S(f_{h,w}^{2d}) \right), \quad (18)$$

where the logarithmic operation smooths the discrete selection responses, mitigating the stepped distribution induced by binary switching and improving training stability.

3.3. Object-Centered Dynamic Enhancement (OCDE) Module

In the dynamic multi-spectral perception and filtering, the frequency components of images can be extensively explored. However, background information in degraded images is also extracted, which may interfere with object discrimination. Standard convolution performs feature aggregation over fixed and content-invariant local neighborhoods, which makes its sampling pattern insensitive to object structure and degradation distribution. However, in foggy images, object boundaries are often blurred, local contrast is weakened, and object responses are easily entangled with haze-contaminated background regions. Under such conditions, fixed convolution tends to aggregate mixed foreground-background responses, making it difficult to effectively suppress irrelevant background interference.

To overcome this limitation, OCDE adopts offset-based deformable sampling instead of fixed local aggregation, as illustrated in Figure 2b. Different from standard convolution, the sampling positions in OCDE are dynamically predicted from the input features, allowing the module to shift its receptive field toward object-relevant regions and capture more informative context. Such content-adaptive sampling is particularly beneficial in foggy scenes, where discriminative object structures are often spatially ambiguous. Therefore, the advantage of OCDE does not simply lie in enlarging the receptive field, but in enabling object-centered feature extraction under degraded visual conditions.

The key aspect of the OCDE module is its ability to effectively model relationships among tokens under the guidance of the important regions in the image feature. These regions of interest are defined by several groups of offset sampling points, which are generated from the queries via an offset generation block. Deformed features are derived by sampling feature maps through bilinear interpolation, and these features are subsequently input into key and value projections to produce the deformed keys and values. Following this, multi-head attention is employed to align queries with the sampled keys and aggregate information from the deformed values. Moreover, the locations of offset points contribute to a more robust deformable position bias, improving the model's learning capabilities.

Next, we elaborate on the structure of the proposed OCDE. As shown in Figure 2b, Starting from $X \in \mathbb{R}^{H \times W \times C}$, OCDE first initializes a regular reference lattice $p \in \mathbb{R}^{H_g \times W_g \times 2}$. Specifically, the lattice resolution is reduced from the original feature size by a factor r , where $H_g = H/r$ and $W_g = W/r$. The reference points are arranged on a regular 2D lattice with coordinates (u, v) , where $u \in \{0, 1, \dots, W_g - 1\}$ and $v \in \{0, 1, \dots, H_g - 1\}$, and are then normalized to the range $[-1, 1]$ according to the grid size (H_g, W_g) , where $(-1, -1)$ and $(1, 1)$ correspond to the top-left and bottom-right corners, respectively. The query tokens are produced via a linear mapping $Q = XW_Q$, which are subsequently used to predict the offsets for the reference points. These query tokens are then input into an offset generation block $\theta_{offset}(\cdot)$, which adaptively computes the offsets. To ensure stable training and avoid excessively large displacements, the predicted offsets are bounded as

$$\Delta p = s \tanh(\theta_{offset}(Q)), \quad (19)$$

where s is a scaling factor controlling the maximum offset range. Since $\tanh(\cdot) \in (-1, 1)$, each offset value is constrained within $[-s, s]$. The final sampling locations are therefore given by

$$\hat{p} = p + \Delta p. \quad (20)$$

Subsequently, features are sampled at the offset point locations, which serve as keys and values, followed by further processing through projection matrices:

$$Q = XW_Q, \hat{K} = \hat{X}W_K, \hat{V} = \hat{X}W_V, \quad (21)$$

$$\text{with } \hat{X} = \phi(X; \hat{p}), \quad (22)$$

where $\hat{K}$ denotes the feature of deformed key, $\hat{V}$ denotes the feature of deformed value. W_Q, W_K, W_V are the projection matrices, and $\phi(\cdot; \cdot)$ is the sampling function implemented through bilinear interpolation:

$$\phi(Z; (p_x, p_y)) = \sum_{(r_x, r_y)} G(p_x, r_x) G(p_y, r_y) Z[r_y, r_x, :], \quad (23)$$

where $G(a, b) = \max(0, 1 - |a - b|)$ and (r_x, r_y) indexes all the locations on $Z \in \mathbb{R}^{H \times W \times C}$. Similar to common transformer-based methods, we apply multi-head attention to $Q, \hat{K}$, and $\hat{V}$, incorporating relative displacements R . The result for a single attention head is given by:

$$Z^{(m)} = \sigma\left(Q^{(m)} \hat{K}^{(m)\top} / \sqrt{d} + \phi(\tilde{B}; R)\right) \hat{V}^{(m)}, \quad (24)$$

where $\sigma(\cdot)$ represents the softmax function, and $\phi(\tilde{B}; R) \in \mathbb{R}^{HW \times H_s W_s}$ denotes the position embedding. $\tilde{B}$ is the relative position bias table. d denotes the dimension of each head. The features from all attention heads are concatenated and passed through a projection matrix W_O to produce the final output:

$$Z = \text{Concat}\left(Z^{(1)}, \dots, Z^{(M)}\right) W_O. \quad (25)$$

Since each reference point corresponds to an $S \times S$ neighborhood (with S representing the maximum offset), the offset generation block needs to perceive local features to produce appropriate offsets. To achieve this, the offset generation block is designed using two convolution modules with a nonlinear activation. Initially, the input features are processed through a 5×5 depthwise convolution to extract local feature information. This is followed by GELU activation and a 1×1 convolution to generate the 2D offsets.

The deformable position bias encodes relative positions between queries and keys, thereby enhancing the model's ability to capture spatial information. For a feature map of the shape of $H \times W$, its relative coordinate displacement range of $[-H, H]$ for the height dimension and $[-W, W]$ for the width dimension. We maintain a learnable bias table $\tilde{B} \in \mathbb{R}^{(2H-1)(2W-1)}$ [52]; however, the key positions in the OCDE are continuous and may not perfectly align with the discrete positions in the bias table. To address this, the relative coordinate displacements are normalized to the range $[-1, 1]$. The function $\phi(\tilde{B}; R)$ then interpolates the value in the bias table $\tilde{B}$ according to the continuous relative displacements R , ensuring that the model can handle all possible offset values.

In particular, we introduce the OCDE module in the final stage of the backbone network to further enhance the extracted features. Since the keys and values in earlier backbone stages have much larger spatial resolutions, applying deformable attention at those stages would significantly increase the computational cost of dot-product attention and bilinear interpolation. Therefore, to achieve a better trade-off between model capacity and computational burden, we place the OCDE module only in the final stage of the

backbone network. The proposed module enables global context aggregation over object-relevant regions, thereby improving feature representation and the detection capability of RShDet.

3.4. Frequency-Domain Multi-Feature Fusion (FDMF) Module

The Frequency-Domain Multi-Feature Fusion (FDMF) module aims to combine features extracted from the enhancement encoder branches with those from the detection encoder branches, thereby improving the model's ability to represent objects in corrupted images. As illustrated in Figure 2, the enhancement feature F_{enh} and the detection feature F_{det} undergo a transformation into the frequency domain during the fusion process. This design is motivated by both signal processing theory and recent computer vision studies. In Fourier analysis, the amplitude spectrum mainly characterizes the distribution of spectral energy across frequencies, which is closely related to contrast, intensity statistics, and degradation-related variations, whereas the phase spectrum preserves the spatial arrangement of structures, such as edges, contours, and object layouts. Such complementary roles have long been recognized in signal processing and have also been exploited in recent vision works on degraded image understanding and restoration [53–56]. For example, WF-Diff [55] and FrDiff [56] likewise exploit amplitude–phase properties for image restoration and dehazing. Therefore, decomposing features into amplitude and phase components provides a principled way to decouple and subsequently integrate appearance-related degradation cues and structure-related object information.

$$F_1 = \mathcal{F}(F_{det}), \quad (26)$$

$$F_2 = \mathcal{F}(F_{enh}), \quad (27)$$

where $\mathcal{F}(\cdot)$ denotes the 2D FFT operation. This transformation decomposes the features into their respective real and imaginary components, from which the amplitude spectrum $|F_i|$ and phase spectrum ϕ_i are derived:

$$|F_i| = \sqrt{\text{Re}(F_i)^2 + \text{Im}(F_i)^2}, \quad (28)$$

$$\phi_i = \arg(F_i), \quad (29)$$

where $i \in \{1, 2\}$, and $\text{Re}(\cdot)$ and $\text{Im}(\cdot)$ represent the real and imaginary parts of the frequency spectrum, respectively. For degraded remote sensing imagery, the enhancement branch provides degradation-aware appearance cues that are useful for compensating corrupted responses, while the detection branch preserves object-oriented semantic and structural information. Their interaction in the frequency domain therefore enables the model to jointly exploit appearance-related and structure-related clues in a more interpretable manner. Learnable channel-wise weights are employed to control the fusion process. The fused amplitude $|F_{com}|$ and phase ϕ_{com} are computed as follows:

$$|F_{com}| = \alpha_1 |F_1| + \alpha_2 |F_2|, \quad (30)$$

$$\phi_{com} = \arg(\beta_1 e^{j\phi_1} + \beta_2 e^{j\phi_2}), \quad (31)$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{R}^{C \times 1 \times 1}$ are learnable parameters that adaptively scale the importance of each feature branch at the channel level. Subsequently, the two spectral components are recombined to synthesize a complex-valued frequency representation:

$$F_{fus} = |F_{com}| \cdot e^{j\phi_{com}}. \quad (32)$$

In this formulation, j denotes the imaginary unit. The term $e^{j\theta}$ integrates the magnitude and directional (phase) information into a unified spectral tensor, where the phase angle θ corresponds to ϕ_{com} . Finally, the fused frequency-domain signal is mapped back to the spatial domain via the inverse FFT:

$$X_{out} = \text{Re}\left(\mathcal{F}^{-1}(F_{fus})\right) \in \mathbb{R}^{B \times C \times H \times W}. \quad (33)$$

where $\text{Re}(\cdot)$ denotes the real-part operator. The final output X_{out} represents the refined spatial-domain representation, effectively blending information from the enhancement and detection branches while mitigating the adverse effects of degradation.

3.5. Encoders

3.5.1. Degradation Representation Encoder

In image processing, the degradation information of hazy images is critical. To effectively capture this information, we design a degradation representation encoder that enhances RShDet's ability to perceive image degradation. Given a hazy image x , it is first passed through a 3×3 convolution layer to obtain the feature map x_l . This is then processed by five Residual Blocks (RBs). Each RB contains two 3×3 convolutional layers, along with normalization and activation functions, which can be expressed as:

$$F_{RB} = \sigma(T_{Cb1}(T_{Cbs}(x_l)) + T_{Cb2}(x_l)) \quad (34)$$

where F_{RB} denotes the output of a residual block, σ represents the SiLU activation function, T_{Cb1} and T_{Cb2} are convolution-batch normalization operations, and T_{Cbs} additionally includes a SiLU activation.

The output from the last residual block is downsampled to 1/32 of the original input size, followed by GAP and two fully connected layers to produce a 256-dimensional degradation representation vector r :

$$r = T_{linear}\left(\sigma(T_{linear}(GAP(F_{RB})))\right) \quad (35)$$

where T_{linear} denotes a fully connected layer. To ensure the encoder focuses on capturing degradation information, a contrastive loss is employed. Specifically, for a degraded image x , two patches are randomly cropped: one serving as the query Q and the other as the positive sample P . Patches from other images with different degradation levels are used as negative samples N . These patches are encoded to obtain one-dimensional feature vectors r_Q , r_P , and r_N . Since contrastive learning benefits from a large number of negative samples, a negative sample queue k_i^- is maintained to provide a substantial pool of negatives. Using these representations, the contrastive loss is defined as:

$$\mathcal{L}_{Ctr} = -\log \frac{\exp(r_Q \cdot r_P / \tau)}{\sum_{i=1}^K \exp(r_Q \cdot k_i^- / \tau)} \quad (36)$$

where K is the length of the negative sample queue and τ is a temperature hyperparameter, set to 0.07, consistent with prior works such as RestorNet [47].

To leverage degradation information effectively for guiding feature transformations in both the object detection and enhancement encoder branches, RShDet integrates a degradation representation encoder to learn multi-scale degradation features. These features are produced at five spatial resolutions—1/2, 1/4, 1/8, 1/16, and 1/32—corresponding to the feature scales in the encoder branches. As illustrated in Figure 4, the encoder feature F_{IN} is combined with the degradation representation F_D through pixel-wise addition.

The result is then processed by a 1×1 convolution followed by a Sigmoid activation to generate degradation guidance weights. These weights are applied to F_{IN} via pixel-wise multiplication to produce the final output feature F_{OUT} :

$$F_{OUT} = \sigma(T_{Conv1}(F_{IN} + F_D)) \otimes F_{IN} \quad (37)$$

where T_{Conv1} represents a 1×1 convolution layer, and $\otimes$ indicates element-wise multiplication. This degradation-guided fusion mechanism enables RShDet to adapt to various degradation levels, improving the encoder's ability to extract meaningful features from corrupted inputs.

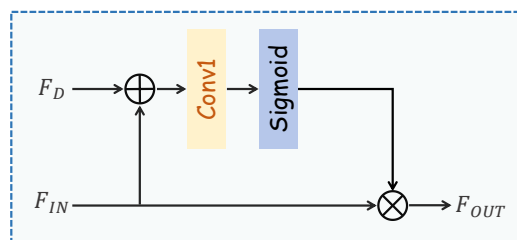


Figure 4. The degradation-guided fuse process.

3.5.2. Image Enhancement Encoder

The purpose of the image enhancement encoder is to capture features relevant to image enhancement. In this work, the encoder is built using the channel attention block from RDMNet [37], which serves as the core building unit. Given an hazy image, the encoder initially captures shallow features through a Focus operation. Deeper features are then extracted across five hierarchical scales by consecutively applying five channel attention blocks, with bilinear interpolation used to downsample between scales. Each channel attention block comprises two 3×3 convolution layers along with a channel attention mechanism, facilitating the model to concentrate on meaningful features while attenuating less relevant ones.

4. Experiments and Analysis

We first introduce the implementation details and the datasets. We then demonstrate the superiority of RShDet in handling challenging haze conditions by comparing it with state-of-the-art (SOTA) methods on both real-world and synthetic datasets. In addition, a series of ablation studies is conducted to validate the effectiveness of the proposed components.

4.1. Implementation Details

RShDet is implemented using PyTorch 2.2.0 and optimized using stochastic gradient descent (SGD) with an initial learning rate of 1×10^{-2} . The model is trained on an NVIDIA GeForce RTX 4090 GPU with a batch size of 4 for 100 epochs. During both training and testing, all input images are resized to 640×640 . Different from standard detectors that learn solely from degraded inputs, RShDet exploits paired hazy and clear image samples during training to strengthen the feature representation capability of the enhancement branch, which in turn facilitates more robust object detection. To quantitatively assess the performance of RShDet under hazy conditions, mean Average Precision (mAP) is adopted as the primary evaluation metric, which is widely used in object detection benchmarks.

4.2. Datasets

Existing public object detection datasets in the RS domain rarely include hazy scenarios. Therefore, we construct a synthetic haze dataset, termed Hazy-DOTA, based on DOTA

v2.0 [15]. In addition, to further evaluate the performance of RShDet under haze conditions, two publicly available haze object detection datasets, RTTS [57] and HazyDet [41], are employed.

Hazy-DOTA: We propose a new hazy RS object detection dataset, termed Hazy-DOTA, which is designed to evaluate and enhance the robustness of object detection models under haze-degraded conditions. Specifically, synthetic haze effects with varying densities are incorporated into the DOTA v2.0 dataset to generate hazy images at different degradation levels. DOTA v2.0 is a large-scale benchmark dataset for RS object detection released by Wuhan University in 2021. It collects images from diverse sources, including Google Earth, the GF-2 satellite, and aerial platforms. The benchmark comprises 11,268 images, within which 1,793,658 objects are annotated over 18 categories. The haze synthesis process is formulated as

$$I(x) = J(x)t(x) + A(1 - t(x)), \quad (38)$$

$$t(x) = e^{-\beta d(x)}, \quad (39)$$

$$d(x) = -0.04\rho(x) + \sqrt{\max(H, W)}. \quad (40)$$

where $I(x)$ denotes the synthesized hazy image and $J(x)$ denotes the corresponding clear image. A is the global atmospheric light, β is the atmospheric scattering coefficient, $t(x)$ is the transmission map, and $d(x)$ is the scene depth proxy used for haze synthesis. H and W denote the image height and width, respectively.

We construct a deterministic depth proxy map using the analytic formulation in Equation (40). Specifically, for an image of size $H \times W$, the image center is defined as $c = (H/2, W/2)$. For each pixel $x = (u, v)$, we compute its Euclidean distance to the image center as

$$\rho(x) = \sqrt{\left(u - \frac{H}{2}\right)^2 + \left(v - \frac{W}{2}\right)^2}, \quad (41)$$

and then assign the depth value by Equation (40) to obtain a full-resolution depth map $d(\cdot)$.

In our implementation, the atmospheric light is set to $A = 0.5$, and the scattering coefficient is defined as $\beta = 0.01i + 0.05$, where i is a random integer sampled from the range $[0, 9]$, enabling the synthesis of haze with varying densities across different images. Figure 5 presents visual examples of the original images and our synthetic hazy samples. To better approximate the variations observed under real-world weather conditions, a mixed sampling strategy is adopted in the constructed dataset, where hazy and clear images are combined at a ratio of 2:1.

HazyDet: HazyDet [41] is a benchmark object detection dataset captured from a UAV perspective under hazy conditions. It contains a total of 11,600 images with 383,000 annotated object instances. The dataset comprises 600 real-world hazy images and 11,000 synthetic hazy images, and includes three object categories: car, truck, and bus. HazyDet is organized into a synthetic hazy training set, a synthetic hazy test set, and a real-world hazy test set (Rw-Test).

RTTS: The RTTS dataset [57] is a real-world foggy dataset consisting of 4322 foggy images annotated with five distinct object categories. It is used as a test set to evaluate the detection performance of RShDet under hazy conditions. In contrast, we create a synthetic hazy dataset, VOC_Fog (VFog), from the VOC dataset [58]. To synthesize these hazy images, we employ the atmospheric scattering model to simulate the effects of haze on clear imagery. Building upon the original VOC subsets (VNorm-train and VNorm-test), we create VFog-train and VFog-test by randomly applying fog processing.



Figure 5. Visual comparison between the original images and our synthetic hazy samples. The first row displays original RS images from the DOTA-v2.0 dataset, while the second row presents the corresponding haze-corrupted images from our proposed Hazy-DOTA dataset. The synthetic samples simulate realistic atmospheric degradation across varying density levels to evaluate model robustness in hazy scenarios.

4.3. Comparative Experiments

4.3.1. Experiments on Hazy-DOTA

To thoroughly assess the potential and generalization capabilities of RShDet for object detection under hazy conditions in RS, we evaluate RShDet on the synthetic Hazy-DOTA benchmark and benchmark it against SOTA methods. For Domain Adaptation, we apply DAYOLO [59], and SF-TMAT [43]. For Image Adaptation, we include IA-YOLO [6], GDIP [8], Q-Norm [40] BAD-Net [60] and DHC-Net [2]. In addition, for frequency enhancement, we further include FcaNet and D3 [61] as representative methods.

Table 1 reports that RShDet outperforms other compared methods on the Hazy-DOTA test set. Specifically, RShDet achieves a 2.4% improvement in mAP50 over the recently proposed DHCNet, demonstrating its superior detection performance under hazy conditions. Beyond these numerical results, we further provide visual comparisons on representative hazy samples, as shown in Figure 6. The visual evidence shows that RShDet achieves superior localization accuracy and identifies a greater number of targets than existing methods. Even in dense-fog conditions, our approach maintains robust detection performance where other models fail.

Table 1. Evaluation results on the proposed Hazy-DOTA dataset.

Method	Type	Precision (%)	Recall (%)	mAP50 (%)
YOLOv8 [62]	Baseline	70.8	55.3	57.3
FcaNet [49]	Frequency Enhancement	71.5	55.5	59.2
D3 [61]	Frequency Enhancement	72.1	55.9	60.7
DAYolo [59]	Domain Adaptation	66.8	46.7	49.2
SF-TMAT [43]	Domain Adaptation	72.7	56.1	61.8
RDMNet [37]	Multi-task	71.4	55.4	58.5
IA-YOLO [6]	Image Adaptation	69.3	46.6	50.5
GDIP [8]	Image Adaptation	70.2	48.2	52.1
BAD-Net [60]	Image Adaptation	71.2	55.3	57.6
Q-Norm [40]	Image Adaptation	71.9	55.4	60.3
DHC-Net [2]	Image Adaptation	72.8	56.0	62.0
RShDet	Multi-task	74.5	58.1	64.4

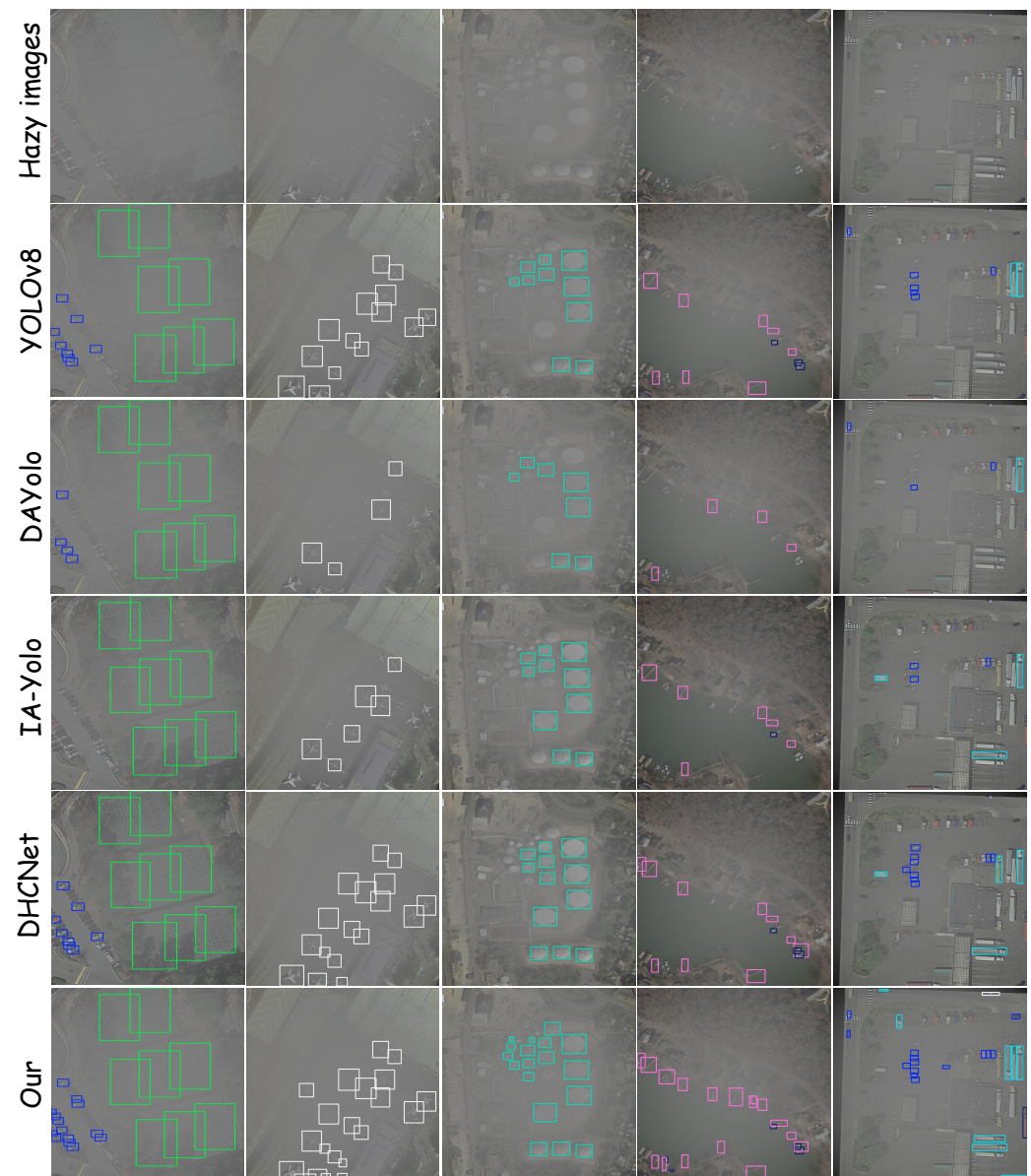


Figure 6. Visual comparison of detection results under hazy conditions. RShDet outperforms SOTA methods by producing more complete detections and reducing missed targets. Note: Bounding box colors denote object categories: white for plane, green for tennis court, blue for small vehicle, light green for storage tank, pink for bridge, dark blue for ship, and light blue for large vehicle.

4.3.2. Experiments on HazyDet

To comprehensively evaluate the capability of RShDet, we further extend our experiments to UAV-based object detection under hazy conditions. For the “restore + detect” paradigm, we employ GridDehaze [63], MixDehazeNet [64], DSANet [65], FFA [66], DehazeFormer [67], C2PNet [68], and RIDCP [69] as preprocessing methods for haze removal. These dehazing models are applied prior to object detection using a Faster R-CNN [70] detector trained on clear UAV images. Moreover, comparisons are conducted with joint restoration-detection methods, namely IA-YOLO, GDIP, DeCoDet [41], DHCNet, Q-Norm, and TogetherNet [12]. For domain adaptation approaches, MRT [42], RST, and SF-TMAT are also included in the comparison.

As reported in Table 2, RShDet consistently outperforms all competing methods under UAV hazy conditions. Specifically, RShDet achieves mAP improvements of 3.1% and 1.9% over YOLOv8 and DHCNet on the test set, respectively. Furthermore, on the

Rw-Test dataset, RShDet(T) yields a substantial performance gain, improving the mAP of DHCNet by 2.2%. These results collectively highlight the superior detection performance and robustness of RShDet in UAV-based hazy scene object detection.

Table 2. Detection performance (mAP) comparison on the HazyDet dataset. The best results are highlighted in bold, and “–” indicates unavailable results.

Method	Type	Test-Set				Rw-Test			
		Car	Truck	Bus	mAP	Car	Truck	Bus	mAP
GridDehaze [63]	Defog	48.7	21.2	46.8	38.9	37.6	8.0	13.3	19.6
MixDehazeNet [64]	Defog	49.5	22.0	48.4	39.9	40.8	8.3	14.4	21.2
DSANet [65]	Defog	50.1	23.0	49.2	40.8	41.8	9.7	15.8	22.4
FFA [66]	Defog	50.2	23.5	49.9	41.2	41.1	9.4	15.5	22.0
DehazeFormer [67]	Defog	51.3	24.7	51.5	42.5	41.2	9.3	15.2	21.9
C2PNet [68]	Defog	51.5	25.4	51.7	42.9	41.8	9.5	16.0	22.4
RIDCP [69]	Defog	52.9	26.1	55.4	44.8	43.9	9.7	19.0	24.2
IA-YOLO	Image Adaptation	44.1	22.2	48.6	38.3	41.9	8.0	17.3	22.4
GDIP	Image Adaptation	47.0	22.7	50.6	40.1	46.6	8.5	17.8	24.3
TogetherNet [12]	Multi-task	53.4	25.4	55.0	44.6	48.2	11.3	16.1	25.2
FCOS-DeCoDet [41]	Image Adaptation	55.9	28.6	57.6	47.4	48.1	11.1	17.8	24.3
VNet-DeCoDet [41]	Image Adaptation	58.3	33.7	62.5	51.5	49.0	9.0	19.7	25.9
Q-Norm	Image Adaptation	81.4	35.7	74.1	63.7	70.3	9.1	21.7	33.7
MRT	Domain Adaptation	77.8	37.5	77.5	64.2	–	–	–	–
RST [71]	Domain Adaptation	82.1	41.3	70.4	64.6	–	–	–	–
SF-TMAT	Domain Adaptation	77.9	44.9	78.1	66.9	67.6	12.9	24.3	34.9
YOLOv8	Baseline	85.8	36.6	75.6	65.9	71.9	7.6	22.7	34.1
DHC-Net	Image Adaptation	86.6	37.1	77.5	67.1	74.9	10.4	23.0	36.1
RShDet	Multi-task	88.4	38.9	79.6	69.0	76.6	12.7	25.5	38.3

4.3.3. Experiments on RTTS

To verify the generality of RShDet for object detection under hazy weather conditions, we conduct extensive comparisons with representative object detection methods on clear images (VNorm-test), synthetic hazy images (VFog-test), and real-world hazy images (RTTS). RShDet and all compared methods are trained on Hybrid datasets, in which the ratio of hazy images (VFog-train) to clean images (VNorm-train) is set to 2:1. We select YOLOv8 [62] as the baseline detector. For clarity, we refer to the network trained on Hybrid samples as YOLOv8, and its counterpart trained on normal samples as YOLOv8*.

The dehaze-then-detect methods first leverage restoration algorithms to enhance the visual quality of hazy images, followed by object detection using a YOLOv8 detector trained on clean images. Specifically, MSBDN [72] and GridDehaze [63] are combined with YOLOv8*, resulting in two cascaded models: MSBDN-YOLOv8* and GridDehaze-YOLOv8*. To ensure data consistency with the RTTS [57] dataset, which focuses on outdoor scenes, both restoration networks are trained using the OTS dataset [57].

In contrast, end-to-end approaches unify image enhancement and object detection within a single framework, enabling direct object detection tasks in hazy images without intermediate preprocessing. This category includes multi-task learning methods (e.g., DSNet [13], DR-YOLO [73], and RDMNet), domain adaptation methods (e.g., DAYOLO and SF-TMAT), and image adaptation methods (e.g., IA-YOLO, DHC-Net, GDIP and its variants), which are compared with RShDet.

Table 3 summarizes the quantitative comparison between RShDet and competing methods under different testing conditions. On the VNorm-test dataset, RShDet achieves the best performance among all methods, yielding a 1.91% mAP improvement over the recent DHC-Net. On the VFog-test dataset, RShDet exhibits substantial performance gains, outperforming YOLOv8, SF-TMAT, and DHC-Net by 5.04%, 2.79%, and 2.23% in mAP, respectively. These results demonstrate the robustness of RShDet under synthetic hazy conditions. For the RTTS dataset, RShDet achieves an mAP of 55.73%, surpassing all

compared methods. In particular, RShDet outperforms YOLOv8, SF-TMAT, and DHCNet by 6.63%, 2.61%, and 2.33%, respectively, indicating strong generalization ability in real-world hazy scenarios.

Overall, these results indicate that RShDet consistently achieves superior performance across clear, synthetic hazy, and real-world hazy datasets. Furthermore, it is observed that end-to-end detection approaches generally outperform the “restore-and-detect” methods, highlighting the effectiveness of unified optimization for hazy scene object detection.

Table 3. Performance comparison (mAP) between models evaluated on synthetic hazy images and real haze images. The best results are highlighted in bold, and “–” indicates unavailable results.

Method	Type	Train Data	VNorm-Test	VFog-Test	RTTS
YOLOv8	Baseline	Hybrid	78.69	78.09	49.10
MSBDN [72]	Defog	VNorm-train	–	57.38	30.30
GridDehaze [63]	Defog	VNorm-train	–	58.23	31.42
DAYOLO	Domain Adaptation	Hybrid	56.51	55.11	29.93
SF-TMAT	Domain Adaptation	Hybrid	80.96	80.34	53.12
DR-YOLO [73]	Multi-task	Hybrid	77.65	78.55	51.63
RDMNet [37]	Multi-task	Hybrid	79.44	78.41	52.34
DSNet [13]	Multi-task	Hybrid	53.29	67.40	28.91
IA-YOLO	Image Adaptation	Hybrid	73.23	72.03	37.08
GDIP-YOLO	Image Adaptation	Hybrid	73.70	71.92	42.42
MGDIP-YOLO	Image Adaptation	Hybrid	75.36	73.77	42.84
GDIP Regularizer	Image Adaptation	Hybrid	73.17	72.77	39.52
DHCNet	Image Adaptation	Hybrid	81.50	80.90	53.40
RShDet	Multi-task	Hybrid	83.41	83.13	55.73

4.3.4. Experiments on DOTA-v2.0

To validate the effectiveness of the proposed method on RS imagery, we evaluate RShDet on the DOTA-v2.0 dataset. RShDet is compared with several representative and well-established RS detection methods. For consistency, RShDet and all comparison approaches are trained and tested on the same DOTA-v2.0 dataset under identical experimental settings.

As reported in Table 4, RShDet consistently outperforms all compared methods in terms of mAP50. Notably, RShDet surpasses PDACL and DHCNet by 2.07% and 1.28% mAP50, even though both are strong contenders on this benchmark. These improvements indicate that RShDet exhibits strong generalization ability when applied to complex RS scenes. To provide a clearer comparison of the mAP50 performance across different methods, we further visualize the results using a histogram, as shown in Figure 7. The histogram clearly illustrates that RShDet achieves the highest mAP50 value among all methods, further highlighting its superior detection performance in RS scenarios. Thus, the proposed model maintains superior performance under both hazy and clear-weather conditions, demonstrating generalization capability across different weather conditions.

Table 4. Evaluation results on DOTA-v2.0 dataset. The best results are highlighted in bold, and “–” indicates unavailable results.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	CC	AP	HL	mAP50
BFM [74]	80.12	54.12	50.07	65.68	43.98	60.07	67.85	79.11	64.38	60.56	45.98	58.26	58.31	64.82	69.84	32.78	89.37	11.07	58.69
DFDet [75]	75.44	52.17	42.28	60.17	48.80	53.36	62.67	78.15	56.85	66.62	40.78	53.05	48.42	59.23	51.40	25.47	66.29	16.38	53.19
Oriente R-CNN [76]	78.65	51.80	47.15	65.78	43.25	58.29	60.89	82.83	63.51	59.50	43.40	55.57	52.90	56.18	54.13	27.55	66.24	5.22	54.06
OAN [77]	79.18	51.57	47.50	66.61	43.30	58.07	60.73	82.85	64.47	59.62	44.31	56.66	52.71	56.73	53.04	26.01	66.41	14.42	54.64
DG-YOLO [78]	86.20	70.80	49.90	57.81	60.00	76.90	79.20	93.30	58.40	70.90	79.20	93.30	58.40	70.90	42.50	61.30	67.50	64.80	57.91
RHINO [79]	79.32	58.33	46.06	65.79	57.25	58.38	68.53	80.46	63.27	73.01	48.69	57.80	56.21	60.04	59.06	42.70	78.96	12.84	59.26
YOLOv8	91.89	72.87	43.42	60.95	62.13	79.11	85.79	94.51	60.78	75.22	42.36	56.53	75.99	68.35	48.51	2.24	33.82	0.15	58.51
RQFormer [80]	77.58	49.96	38.60	53.82	54.97	57.12	68.93	77.88	59.59	71.92	40.17	51.22	53.00	57.12	49.66	25.42	66.92	5.17	53.28
FCOS-O [81]	74.84	47.53	40.83	57.41	43.89	47.72	55.66	78.61	57.86	63.00	38.02	52.38	41.91	53.24	40.22	7.15	65.51	7.42	48.51
DACL [82]	83.08	68.53	44.31	58.33	63.04	79.12	88.18	93.87	58.51	72.95	54.01	54.84	73.21	57.80	40.70	3.05	61.41	49.67	61.21
O-Rep [83]	73.02	46.68	42.37	63.05	47.06	50.28	58.64	78.84	57.12	66.77	35.21	50.76	48.77	51.62	34.23	6.17	64.66	5.87	48.95
DCFL [84]	79.49	55.97	50.15	61.59	49.01	55.33	59.31	81.81	66.52	60.06	52.87	56.71	57.83	58.13	60.35	35.66	78.65	13.03	57.66
OASL [85]	76.65	55.46	46.33	62.49	53.18	56.62	66.16	80.75	63.07	67.03	44.89	55.68	54.24	59.04	59.15	35.00	77.81	14.65	57.18
H2RBox-v2 [86]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	50.00
GRA [87]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	57.95
MTP [88]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	58.96
DHCNet	84.10	65.40	45.10	61.00	62.90	79.20	88.50	95.04	63.40	76.00	51.60	57.50	76.20	54.10	31.80	9.87	67.50	49.50	62.00
RShDet	86.24	67.45	44.91	61.26	64.78	78.73	87.63	95.46	66.27	76.53	53.52	58.49	75.81	58.43	33.60	11.36	67.22	51.26	63.28

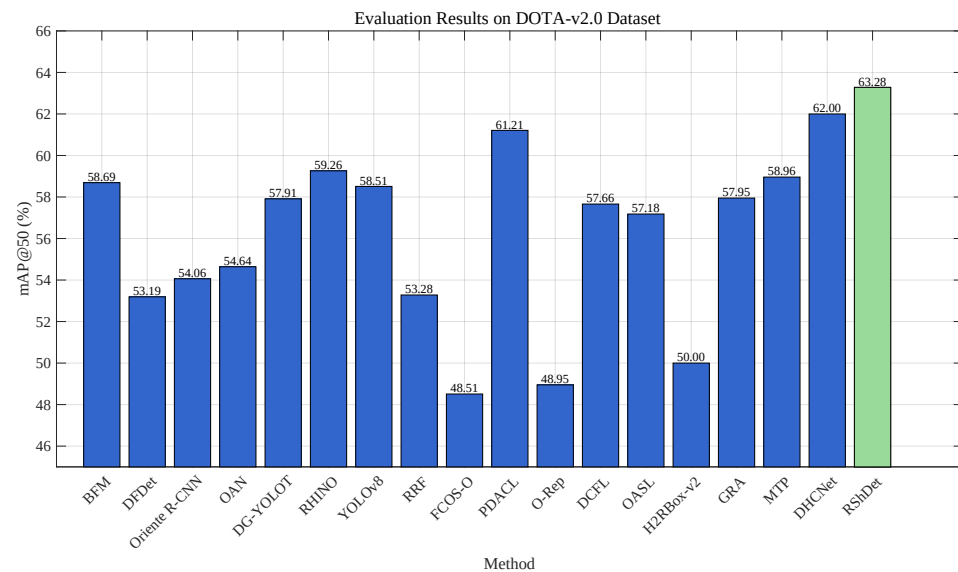


Figure 7. Bar chart visualization of mAP50 (%) results on the DOTA-v2.0 dataset. The green bars denote the proposed RShDet, while the blue bars represent the compared methods.

4.4. Computational Complexity

In this section, we analyze the computational complexity of RShDet in comparison with other multi-task learning and image adaptation methods. Params and FLOPs are reported to reflect the intrinsic model complexity, while frame rate (FPS) and time cost are adopted to evaluate the actual inference efficiency of each model. To ensure a fair comparison, all methods in Table 5 are evaluated under the same experimental environment and inference settings. Specifically, all models are tested using a single NVIDIA GeForce RTX 3090 GPU, and the input images are resized to 640×640 .

As shown in Table 5, baseline detectors generally achieve the highest inference speed. In terms of model complexity, RShDet contains only 9.24 M parameters and 10.9 G FLOPs, which are substantially lower than those of most compared methods, indicating that the proposed framework remains computationally efficient. Multi-task learning approaches provide a better balance between efficiency and detection performance, especially under hazy conditions. Although RShDet is slightly slower than YOLOv8, it still maintains competitive inference efficiency and is faster than DHCNet. Meanwhile, RShDet also achieves improved detection accuracy while maintaining competitive inference efficiency.

Table 5. Computational complexity comparison of RShDet and other methods.

Method	Type	Params (M)	FLOPs (G)	FPS	Speed (ms)
YOLOv8	Baseline	3.2	8.7	310.56	3.22
IA-YOLO	Image Adaptation	61.7	154.7	17.44	57.35
GDIP-YOLO	Image Adaptation	196.9	405.9	6.77	147.63
MGDIP-YOLO	Image Adaptation	138.2	351.3	7.79	128.32
GDIP as regularizer	Image Adaptation	61.5	154.5	17.49	57.16
DHCNet	Image Adaptation	11.3	28.9	92.5	10.81
SF-TMAT	Domain Adaptation	40.0	70.3	38.79	25.78
DR-YOLO	Multi-task	43.6	104.6	26.24	38.11
RShDet	Multi-task	9.24	10.9	305.0	3.28

4.5. Ablation Studies

4.5.1. Module Ablation

To investigate the necessity and complementarity of different components, we conduct ablation experiments on FDMF, OCDE, and DSPF. All ablation models are trained and tested on the Hazy-DOTA dataset. As shown in Table 6, the full model equipped with all

three modules achieves the best performance, demonstrating that these components are complementary rather than redundant.

Table 6. Ablation study results on model performance with different components. In the DSPF module, DCT denotes the multi-spectral attention mechanism and Switches denotes the learning-driven frequency filtering block.

FDMF	OCDE	DCT	Switches	Precision (%)	Recall(%)	mAP50(%)	mAP95(%)
×	×	×	×	70.80	55.30	57.30	38.70
✓	×	×	×	71.59	55.91	59.01	39.01
✓	✓	×	×	73.06	56.65	61.03	39.94
✓	✓	✓	×	73.78	57.31	62.75	40.68
✓	✓	✓	✓	74.50	58.10	64.40	41.30

Specifically, integrating the FDMF module improves the mAP50 by 1.71%, which validates the effectiveness of frequency-domain fusion between the enhancement and detection branches. By jointly integrating amplitude and phase information from the two branches, FDMF facilitates complementary information exchange between the enhancement and detection features.

After further introducing OCDE, the mAP50 increases by another 2.02%. This result indicates that OCDE effectively enhances feature extraction by suppressing irrelevant background interference in the spatial domain and aggregating context over target-relevant regions. Such spatial-domain focusing provides more discriminative features for the subsequent frequency-domain analysis. Notably, we observe that the model using the learned subset of frequency components outperforms the variant using all frequency components, achieving a 1.65% increase in mAP50. This observation shows that, even after spatial background suppression, the retained object features still contain mixed spectral responses and haze-related distortions. Therefore, frequency-domain selection remains necessary to further suppress trivial or noisy frequency components and preserve the most discriminative spectral cues.

4.5.2. Qualitative Analysis of DSPF

To qualitatively verify the effectiveness of DSPF, we visualize intermediate feature responses produced by RShDet. Figure 8 shows the feature maps obtained with the DCT-based spectral representation and the learned Switches. Specifically, Figure 8a presents the input images, while Figure 8b and Figure 8c correspond to the feature maps generated using all frequency components and the learned frequency subset, respectively. As can be seen, the maps in Figure 8c exhibit more pronounced activations than those in Figure 8b, suggesting that the learned subset better preserves discriminative cues.

4.5.3. Effect of Amplitude and Phase in FDMF

To further investigate the contribution of each component in the FDMF module, we conduct an ablation study on the amplitude and phase branches, and report the results in Table 7. The results show that both branches improve detection performance when used individually. Moreover, jointly enabling the two branches yields the best performance, with gains of 1.66% in mAP50 and 0.51% in mAP95 over the baseline without amplitude or phase fusion. These findings verify that amplitude and phase play complementary roles in the proposed FDMF module, and that their joint fusion is crucial for robust object detection under degraded conditions.

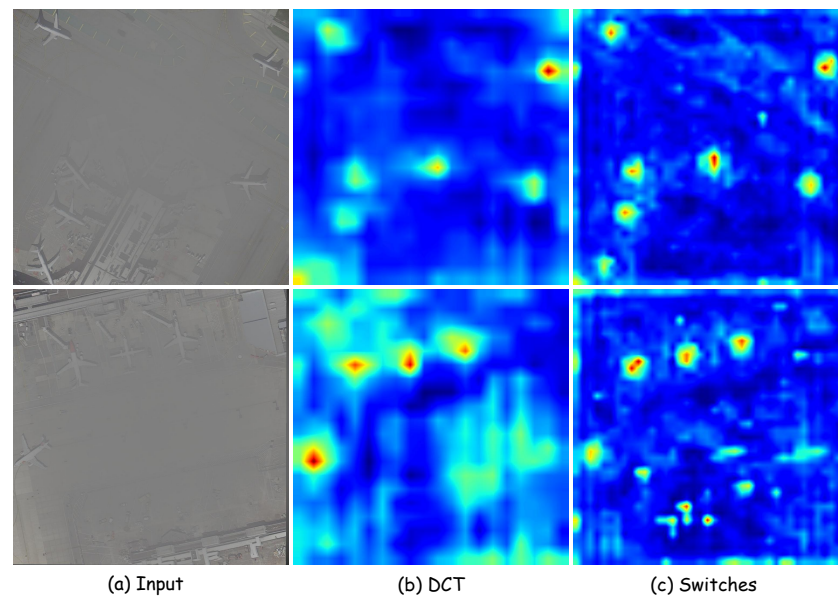


Figure 8. Qualitative comparison of RShDet feature responses with DSPF: all-frequency features versus features selected by learned Switches. Warmer colors indicate stronger responses, while cooler colors indicate weaker responses.

Table 7. Ablation study results on FDMF. ✓ denotes that the corresponding component is included, while × denotes that it is excluded.

Amplitude	Phase	Precision (%)	Recall (%)	mAP50 (%)	mAP95 (%)
×	×	73.87	57.43	62.74	40.79
✓	×	74.26	57.91	63.87	41.15
×	✓	74.14	57.62	63.36	40.99
✓	✓	74.50	58.10	64.40	41.30

4.5.4. Effect of Loss Weights

To enhance the detection capability of RShDet in hazy environments, the network is trained with a unified objective that jointly combines the object detection loss $\mathcal{L}_{Det}$, the image enhancement loss $\mathcal{L}_{En}$, the frequency regularization term $\mathcal{L}_{Fre}$, and the contrastive loss $\mathcal{L}_{Ctr}$. Accordingly, four weighting coefficients, λ_1 , λ_2 , λ_3 , and λ_4 , are introduced to control the relative contribution of each term. To determine suitable values for these coefficients, we perform extensive experiments on the Hazy-DOTA dataset, and the corresponding results are summarized in Table 8. The results indicate that the balance between the detection loss and enhancement loss has a noticeable influence on the final detection performance. The setting $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.2, 0.8, 0.1, 0.1)$ achieves the best mAP50 of 64.4%, indicating that the adopted weight configuration provides the best trade-off among detection supervision, enhancement supervision, and the two auxiliary regularization terms. The results further show that excessively small or excessively large values of λ_3 and λ_4 are less favorable, while moderate settings lead to better performance.

Table 8. Sensitivity analysis of the loss weights in the total objective.

λ_1	λ_2	λ_3	λ_4	mAP50 (%)
0.1	0.9	0.1	0.1	63.5
0.2	0.8	0.1	0.1	64.4
0.3	0.7	0.1	0.1	63.3
0.4	0.6	0.1	0.1	62.9
0.2	0.8	0.05	0.1	63.9
0.2	0.8	0.15	0.1	64.1
0.2	0.8	0.20	0.1	63.7
0.2	0.8	0.1	0.05	63.7
0.2	0.8	0.1	0.15	63.9
0.2	0.8	0.1	0.20	63.4

5. Conclusions

In this paper, we propose RShDet, an Adaptive Spectral Aware Network designed to address the challenges of RS object detection under haze interference. Unlike traditional cascading methods, RShDet does not simply treat restoration and detection as sequential steps. Instead, it synergizes both tasks in the frequency domain, where the restoration task serves to reclaim essential object details that facilitate more accurate detection. Specifically, the OCDE module employs query-agnostic offsets to adjust the sampling point positions, enabling the module to focus on important regions. Guided by our spectral analysis of hazy imagery, we develop DSPF, which attenuates irrelevant information while adaptively retaining discriminative frequency components. Furthermore, the FDMF module achieves deep cross-task synergy by integrating amplitude and phase components across branches. To facilitate research in this field, we established the Hazy-DOTA benchmark, which simulates realistic haze degradation on RS images. Extensive experiments demonstrate the effectiveness of RShDet for object detection in hazy RS images. It improves the mAP50 of the YOLOv8 baseline on Hazy-DOTA from 57.3% to 64.4%, and achieves consistent gains across multiple datasets, including 7.1% mAP50 on Hazy-DOTA, 3.1% mAP on HazyDet, and 6.63% mAP on RTTS. Compared with SOTA methods, it further yields improvements of 2.4% mAP50, 1.9% mAP, and 2.33% mAP on these datasets, respectively, demonstrating strong robustness and generalization capability.

While our method achieves promising results, it currently relies on the assumption that paired hazy and clear images are available during the training phase. Such supervision facilitates the learning of haze-aware representations; however, in many real-world RS scenarios, collecting perfectly aligned paired data is both time-consuming and often impractical. Consequently, the dependence on paired datasets may limit the scalability and broader applicability of the proposed approach, particularly in situations where only unpaired or limited data are available.

In addition, the current RShDet framework is primarily tailored for hazy weather conditions, with a specific focus on mitigating the impact of atmospheric haze on detection performance. In practical RS scenarios, however, images are often affected by a wider range of degradations, such as rain, snow, fog, and low-light conditions, each introducing different types of visual distortions. These factors are not explicitly considered in the current framework.

In future work, we will investigate unsupervised restoration strategies to exploit degraded images without relying on paired supervision, so as to mine more useful information for object detection. Meanwhile, we will further explore a unified detection framework for multiple adverse conditions, with the goal of achieving robust object detection in rainy, hazy, snowy, and dark scenes.

Author Contributions: Conceptualization, W.Z. and Y.W.; methodology, W.Z. and Y.W.; software, H.Y.; validation, Y.W.; formal analysis, Y.W.; investigation, Y.W.; resources, H.Y.; data curation, Y.W.; writing—original draft preparation, W.Z.; writing—review and editing, X.M.; visualization, W.Z.; supervision, X.M.; project administration, W.Z. 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 grant number 92152109.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

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

References

1. Zhang, P.; Cheng, G.; Lang, C.; Xie, X.; Han, J. NIRNet: Noise Incentive Robust Network in Remote Sensing Object Detection under Cloud Corruption. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5629713. [\[CrossRef\]](#)
2. Li, Y.; Ling, Q.; An, Y.; Yin, H.; Gao, X.; Zhu, Z.; Han, P. DHC-Net: A Remote Sensing Object Detection Under Haze and Class Imbalance. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4703016. [\[CrossRef\]](#)
3. Tang, D.; Cao, X.; Wu, X.; Li, J.; Yao, J.; Bai, X.; Jiang, D.; Li, Y.; Meng, D. AeroGen: Enhancing Remote Sensing Object Detection with Diffusion-Driven Data Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 3614–3624.
4. Gao, T.; Xia, S.; Liu, M.; Zhang, J.; Chen, T.; Li, Z. MSNet: Multi-Scale Network for Object Detection in Remote Sensing Images. *Pattern Recognit.* **2025**, *158*, 110983. [\[CrossRef\]](#)
5. Yan, L.; He, Z.; Zhang, Z.; Xie, G. LS-MambaNet: Integrating Large Strip Convolution and Mamba Network for Remote Sensing Object Detection. *Remote Sens.* **2025**, *17*, 1721. [\[CrossRef\]](#)
6. Liu, W.; Ren, G.; Yu, R.; Guo, S.; Zhu, J.; Zhang, L. Image-adaptive YOLO for object detection in adverse weather conditions. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; pp. 1792–1800.
7. Liu, W.; Li, W.; Zhu, J.; Cui, M.; Xie, X.; Zhang, L. Improving Nighttime Driving-Scene Segmentation via Dual Image-Adaptive Learnable Filters. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 5855–5867. [\[CrossRef\]](#)
8. Kalwar, S.; Patel, D.; Aanegola, A.; Konda, K.R.; Garg, S.; Krishna, K.M. GDIP: Gated Differentiable Image Processing for Object Detection in Adverse Conditions. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 7083–7089. [\[CrossRef\]](#)
9. Xi, Y.; Jia, W.; Miao, Q.; Feng, J.; Ren, J.; Luo, H. Detection-Driven Exposure-Correction Network for Nighttime Drone-View Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5605014. [\[CrossRef\]](#)
10. Oza, P.; Sindagi, V.A.; VS, V.; Patel, V.M. Unsupervised Domain Adaptation of Object Detectors: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 4018–4040. [\[CrossRef\]](#)
11. Li, G.; Ji, Z.; Qu, X.; Zhou, R.; Cao, D. Cross-Domain Object Detection for Autonomous Driving: A Stepwise Domain Adaptative YOLO Approach. *IEEE Trans. Intell. Veh.* **2022**, *7*, 603–615. [\[CrossRef\]](#)
12. Wang, Y.; Yan, X.; Zhang, K.; Gong, L.; Xie, H.; Wang, F.L.; Wei, M. TogetherNet: Bridging Image Restoration and Object Detection Together via Dynamic Enhancement Learning. *Comput. Graph. Forum* **2022**, *41*, 465–476. [\[CrossRef\]](#)
13. Huang, S.C.; Le, T.H.; Jaw, D.W. DSNet: Joint Semantic Learning for Object Detection in Inclement Weather Conditions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 2623–2633. [\[CrossRef\]](#)
14. Liu, J.; Zhan, J.; Guo, Y.; Li, T.; Zhao, Y.; Zhang, J.; Tang, L.; Li, Y.; Wei, Y.; Cai, W. Learning Geometric-Aware and Weather-Adaptive Semantics in Remote Sensing: Affine Lie Group Enhanced Detector for UAV Road Scenes. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5605820. [\[CrossRef\]](#)
15. Ding, J.; Xue, N.; Xia, G.S.; Bai, X.; Yang, W.; Yang, M.Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7778–7796. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Ren, Y.; Zhu, C.; Xiao, S. Small Object Detection in Optical Remote Sensing Images via Modified Faster R-CNN. *Appl. Sci.* **2018**, *8*, 813. [\[CrossRef\]](#)
17. Bai, T.; Pang, Y.; Wang, J.; Han, K.; Luo, J.; Wang, H.; Lin, J.; Wu, J.; Zhang, H. An Optimized Faster R-CNN Method Based on DRNet and RoI Align for Building Detection in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 762. [\[CrossRef\]](#)
18. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
19. Liu, Z.; Hu, J.; Weng, L.; Yang, Y. Rotated region based CNN for ship detection. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 900–904.

20. Zhang, Y.; Guo, L.; Wang, Z.; Yu, Y.; Liu, X.; Xu, F. Intelligent Ship Detection in Remote Sensing Images Based on Multi-Layer Convolutional Feature Fusion. *Remote Sens.* **2020**, *12*, 3316. [\[CrossRef\]](#)
21. Han, Y.; Meng, W.; Tang, W. Capsule-inferenced Object Detection for Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 5260–5270. [\[CrossRef\]](#)
22. Li, M.; Cao, C.; Feng, Z.; Xu, X.; Wu, Z.; Ye, S.; Yong, J. Remote Sensing Object Detection Based on Strong Feature Extraction and Prescreening Network. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 8000505. [\[CrossRef\]](#)
23. Li, Q.; Chen, Y.; Zeng, Y. Transformer With Transfer CNN for Remote-Sensing-Image Object Detection. *Remote Sens.* **2022**, *14*, 984. [\[CrossRef\]](#)
24. Zhang, Y.; Wu, Y.; Liu, Y.; Peng, X. CPA-Enhancer: Chain-of-Thought Prompted Adaptive Enhancer for Downstream Vision Tasks Under Unknown Degradations. In Proceedings of the ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; pp. 1–5.
25. Yin, X.; Yu, Z.; Fei, Z.; Lv, W.; Gao, X. PE-YOLO: Pyramid Enhancement Network for Dark Object Detection. In Proceedings of the International Conference on Artificial Neural Networks (ICANN), Heraklion, Greece, 26–29 September 2023; pp. 163–174.
26. Qin, Q.; Chang, K.; Huang, M.; Li, G. DENet: Detection-Driven Enhancement Network for Object Detection under Adverse Weather Conditions. In Proceedings of the Asian Conference on Computer Vision (ACCV), Macao, China, 4–8 December 2022; pp. 2813–2829.
27. Xue, R.; Duan, J.; Du, Z. MPE-DETR: A Multiscale Pyramid Enhancement Network for Object Detection in Low-Light Images. *Image Vis. Comput.* **2024**, *150*, 105202. [\[CrossRef\]](#)
28. Zhang, J.; Huang, J.; Tian, Z.; Lu, S. Spectral Unsupervised Domain Adaptation for Visual Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 9829–9840.
29. Li, J.; Xu, R.; Ma, J.; Zou, Q.; Ma, J.; Yu, H. Domain Adaptive Object Detection for Autonomous Driving under Foggy Weather. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–7 January 2023; pp. 612–622.
30. Kennerley, M.; Wang, J.G.; Veeravalli, B.; Tan, R.T. 2PCNet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 11484–11493.
31. Wang, L.; Qin, H.; Zhou, X.; Lu, X.; Zhang, F. R-YOLO: A Robust Object Detector in Adverse Weather. *IEEE Trans. Instrum. Meas.* **2022**, *72*, 5000511. [\[CrossRef\]](#)
32. Pang, J. MCNet: Magnitude Consistency Network for Domain Adaptive Object Detection under Inclement Environments. *Pattern Recognit.* **2024**, *145*, 109947. [\[CrossRef\]](#)
33. Huang, J.; Guan, D.; Xiao, A.; Lu, S. FSDR: Frequency Space Domain Randomization for Domain Generalization. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 6887–6898.
34. Cui, Z.; Qi, G.-J.; Gu, L.; You, S.; Zhang, Z.; Harada, T. Multitask AET With Orthogonal Tangent Regularity for Dark Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Online, 11–17 October 2021; pp. 2553–2562.
35. Xi, Y.; Jia, W.; Miao, Q.; Feng, J.; Liu, X.; Li, F. CoDerainNet: Collaborative Deraining Network for Drone-View Object Detection in Rainy Weather Conditions. *Remote Sens.* **2023**, *15*, 1487. [\[CrossRef\]](#)
36. Xiao, A.; Huang, J.; Xuan, W.; Ren, R.; Liu, K.; Guan, D.; El Saddik, A.; Lu, S.; Xing, E.P. 3D Semantic Segmentation in the Wild: Learning Generalized Models for Adverse-Condition Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 19–24 June 2023; pp. 9382–9392.
37. Wang, X.; Liu, X.; Yang, H.; Wang, Z.; Wen, X.; He, X.; Qing, L.; Chen, H. Degradation Modeling for Restoration-Enhanced Object Detection in Adverse Weather Scenes. *IEEE Trans. Intell. Veh.* **2025**, *10*, 2064–2079. [\[CrossRef\]](#)
38. Wang, N.; Wang, Y.; Feng, Y.; Wei, Y. MDD-ShipNet: Math-data integrated defogging for fog-occlusion ship detection. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 15040–15052. [\[CrossRef\]](#)
39. Hu, Q.; Zhang, Y.; Zhang, R.; Xu, F.; Yang, W. Beyond Dehazing: Learning Intrinsic Hazy Robustness for Aerial Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5645614. [\[CrossRef\]](#)
40. Zhang, L.; Zhou, Y.; Gao, F.; Li, Z.; Qiao, M.; Xu, J.; Wang, N. Q-Norm: Robust Representation Learning via Quality-Adaptive Normalization. In Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025; pp. 13901–13911.
41. Feng, C.; Chen, Z.; Li, X.; Wang, C.; Yang, J.; Cheng, M.M.; Dai, Y.; Fu, Q. HazyDet: Open-Source Benchmark for Drone-View Object Detection with Depth-Cues in Hazy Scenes. *arXiv* **2024**, arXiv:2409.19833.
42. Zhao, Z.; Wei, S.; Chen, Q.; Li, D.; Yang, Y.; Peng, Y.; Liu, Y. Masked Retraining Teacher-Student Framework for Domain Adaptive Object Detection. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 19039–19049.

43. Chen, H.; Liu, J.; Wang, Y.; Zhu, J.; Feng, D.; Xie, Y. Teaching in Adverse Scenes: A Statistically Feedback-Driven Threshold and Mask Adjustment Teacher-Student Framework for Object Detection in UAV Images under Adverse Scenes. *ISPRS J. Photogramm. Remote Sens.* **2025**, *227*, 332–348. [\[CrossRef\]](#)
44. Liu, B.; Chen, S.-B.; Wang, J.-X.; Tang, J.; Luo, B. An Oriented Object Detector for Hazy Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1001711. [\[CrossRef\]](#)
45. Wan, Y.; Li, J.; Lin, L.; Yuan, Q.; Shen, H. Collaboration of Dehazing and Object Detection Tasks: A Multi-Task Learning Framework for Foggy Image. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5600815. [\[CrossRef\]](#)
46. Zhang, D.; Lu, Y.; Li, Y.; Ding, W.; Zhang, B.; Xiao, J. Frequency Learning Attention Networks Based on Deep Learning for Automatic Modulation Classification in Wireless Communication. *Pattern Recognit.* **2023**, *137*, 109345. [\[CrossRef\]](#)
47. Wang, X.; Chen, H.; Gou, H.; He, J.; Wang, Z.; He, X.; Qing, L.; Sheriff, R.E. RestorNet: An efficient network for multiple degradation image restoration. *Knowl. Based Syst.* **2023**, *282*, 111116. [\[CrossRef\]](#)
48. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
49. Qin, Z.; Zhang, P.; Wu, F.; Li, X. FcaNet: Frequency Channel Attention Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Online, 11–17 October 2021; pp. 783–792.
50. Xie, Z.; Wang, S.; Xu, K.; Zhang, Z.; Tan, X.; Xie, Y.; Ma, L. Boosting Night-Time Scene Parsing with Learnable Frequency. *IEEE Trans. Image Process.* **2023**, *32*, 2386–2398. [\[CrossRef\]](#) [\[PubMed\]](#)
51. Jang, E.; Gu, S.; Poole, B. Categorical Reparametrization with Gumbel-Softmax. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
52. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Online, 11–17 October 2021; pp. 10012–10022.
53. Oppenheim, A.V.; Lim, J.S. The Importance of Phase in Signals. *Proc. IEEE* **1981**, *69*, 529–541. [\[CrossRef\]](#)
54. Bi, Q.; You, S.; Gevers, T. Generalized Foggy-Scene Semantic Segmentation by Frequency Decoupling. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 16–22 June 2024; pp. 1389–1399.
55. Zhao, C.; Cai, W.; Dong, C.; Hu, C. Wavelet-Based Fourier Information Interaction with Frequency Diffusion Adjustment for Underwater Image Restoration. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 8281–8291.
56. Liu, C.; Qi, L.; Pan, J.; Qian, X.; Yang, M.-H. Frequency Domain-Based Diffusion Model for Unpaired Image Dehazing. In Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025; pp. 7538–7547.
57. Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; Wang, Z. Benchmarking Single-Image Dehazing and Beyond. *IEEE Trans. Image Process.* **2018**, *28*, 492–505. [\[CrossRef\]](#)
58. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [\[CrossRef\]](#)
59. Hnewa, M.; Radha, H. Multiscale Domain Adaptive YOLO for Cross-Domain Object Detection. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 3323–3327.
60. Li, C.; Zhou, H.; Liu, Y.; Yang, C.; Xie, Y.; Li, Z.; Zhu, L. Detection-Friendly Dehazing: Object Detection in Real-World Hazy Scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 8284–8295. [\[CrossRef\]](#)
61. Si, Y.; Xu, H.; Zhu, X.; Liu, R.; Li, H. Attention-Modulated Frequency-Aware Pooling via Spatial Guidance. *Neurocomputing* **2025**, *625*, 129507. [\[CrossRef\]](#)
62. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 10 February 2026).
63. Liu, X.; Ma, Y.; Shi, Z.; Chen, J. GridDehazeNet: Attention-based multi-scale network for image dehazing. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7314–7323.
64. Lu, L.; Xiong, Q.; Xu, B.; Chu, D. MixDehazeNet: Mix structure block for image dehazing network. In Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 30 June–5 July 2024; pp. 1–10.
65. Cui, Y.; Knoll, A. Dual-domain strip attention for image restoration. *Neural Netw.* **2024**, *171*, 429–439. [\[CrossRef\]](#)
66. Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; Jia, H. FFA-Net: Feature Fusion Attention Network for Single Image Dehazing. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020; pp. 11908–11915.
67. Song, Y.; He, Z.; Qian, H.; Du, X. Vision transformers for single image dehazing. *IEEE Trans. Image Process.* **2023**, *32*, 1927–1941. [\[CrossRef\]](#)

68. Zheng, Y.; Zhan, J.; He, S.; Dong, J.; Du, Y. Curricular Contrastive Regularization for Physics-Aware Single Image Dehazing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 5785–5794.
69. Wu, R.Q.; Duan, Z.P.; Guo, C.L.; Chai, Z.; Li, C. RIDCP: Revitalizing Real Image Dehazing via High-Quality Codebook Priors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 22282–22291.
70. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [[CrossRef](#)]
71. Han, J.; Yang, W.; Wang, Y.; Chen, L.; Luo, Z. Remote Sensing Teacher: Cross-Domain Detection Transformer With Learnable Frequency-Enhanced Feature Alignment in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5619814. [[CrossRef](#)]
72. Dong, H.; Pan, J.; Xiang, L.; Hu, Z.; Zhang, X.; Wang, F.; Yang, M.H. Multi-scale boosted dehazing network with dense feature fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 14–19 June 2020; pp. 2157–2167.
73. Zhong, F.; Shen, W.; Yu, H.; Wang, G.; Hu, J. Dehazing & Reasoning YOLO: Prior Knowledge-Guided Network for Object Detection in Foggy Weather. *Pattern Recognit.* **2024**, *156*, 110756. [[CrossRef](#)]
74. Cha, K.; Seo, J.; Lee, T. A Billion-scale Foundation Model for Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, 1–17. . [[CrossRef](#)]
75. Xie, X.; Cheng, G.; Rao, C.; Lang, C.; Han, J. Oriented Object Detection via Contextual Dependence Mining and Penalty-Incentive Allocation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5618010. [[CrossRef](#)]
76. Xie, X.; Cheng, G.; Wang, J.; Li, K.; Yao, X.; Han, J. Oriented R-CNN and Beyond. *Int. J. Comput. Vis.* **2024**, *132*, 2420–2442. [[CrossRef](#)]
77. Xie, X.; Cheng, G.; Li, Q.; Miao, S.; Li, K.; Han, J. Fewer Is More: Efficient Object Detection in Large Aerial Images. *Sci. China Inf. Sci.* **2024**, *67*, 112106. [[CrossRef](#)]
78. Saini, N.; Chattopadhyay, C.; Das, D. DG-YOLO: A Lightweight Density Guided YOLO-transformer for Remote Sensing Object Detection. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Pasadena, CA, USA, 16–21 July 2023; pp. 6490–6493.
79. Lee, H.; Song, M.; Koo, J.; Seo, J. Hausdorff Distance Matching with Adaptive Query Denoising for Rotated Detection Transformer. In Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 28 February–4 March 2025; pp. 1872–1882.
80. Zhao, J.; Ding, Z.; Zhou, Y.; Zhu, H.; Du, W.-L.; Yao, R.; El Saddik, A. RQFormer: Rotated Query Transformer for End-to-End Oriented Object Detection. *Expert Syst. Appl.* **2025**, *266*, 126034. [[CrossRef](#)]
81. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: A simple and strong anchor-free object detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1922–1933. [[CrossRef](#)]
82. Biswas, D.; Tešić, J. Domain Adaptation With Contrastive Learning for Object Detection in Satellite Imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5620615. [[CrossRef](#)]
83. Li, W.; Chen, Y.; Hu, K.; Zhu, J. Oriented RepPoints for aerial object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 1829–1838.
84. Xu, C.; Ding, J.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.S. Dynamic coarse-to-fine learning for oriented tiny object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 7318–7328.
85. Zhao, Z.; Li, S. OASL: Orientation-aware adaptive sampling learning for arbitrary oriented object detection. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *128*, 103740. [[CrossRef](#)]
86. Yu, Y.; Yang, X.; Li, Q.; Zhou, Y.; Da, F.; Yan, J. H2RBox-v2: Incorporating symmetry for boosting horizontal box supervised oriented object detection. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 10–16 December 2023; pp. 59137–59150.
87. Wang, J.; Pu, Y.; Han, Y.; Guo, J.; Wang, Y.; Li, X.; Huang, G. GRA: Detecting oriented objects through group-wise rotating and attention. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 298–315.
88. Wang, D.; Zhang, J.; Xu, M.; Liu, L.; Wang, D.; Gao, E.; Han, C.; Guo, H.; Du, B.; Tao, D.; et al. MTP: Advancing Remote Sensing Foundation Model via Multitask Pretraining. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 11632–11654. [[CrossRef](#)]

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