

Article

Spectra-Net: Frequency-Aware Scale Adaptation for Small Object Detection

Yang Xu ¹, Kaiwang Wang ¹, Donglin Yang ², Guanqiu Qi ^{3,*} , Kunpeng Wang ⁴ and Shuang Li ^{2,*}

¹ College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; xuyang@cqupt.edu.cn (Y.X.); s240331003@stu.cqupt.edu.cn (K.W.)

² School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 2024212671@stu.cqupt.edu.cn

³ Computer Information Systems Department, State University of New York at Buffalo State, Buffalo, NY 14222, USA

⁴ School of Information and Control Engineering, Southwest University of Science and Technology, Mianyang 621010, China; kwang@swust.edu.cn

* Correspondence: qig@buffalostate.edu (G.Q.); d230201016@stu.cqupt.edu.cn (S.L.)

Abstract

Small object detection in real-world scenarios remains challenging due to two coupled factors: severe scale imbalance and progressive degradation of fine-grained cues during backbone downsampling. Under drastic scale variations, conventional detectors still rely on static backbones whose fixed convolutional responses cannot consistently accommodate the divergent spectral characteristics of large and tiny objects, leading to scale-mismatched representations. Meanwhile, repeated strided operations reduce the sampling rate of feature maps and tend to introduce aliasing, eroding the high-frequency details that are critical for tiny objects. To address these issues, we propose Spectra-Net, a novel frequency-aware detection framework that redesigns backbone feature encoding with explicit spectral control. At its core, we introduce Dynamic Fourier Alignment (DFA), which performs content-adaptive yet frequency-controllable modulation to reshape convolutional responses in the spectral domain, aligning representations across scales and amplifying discriminative cues for small objects. In addition, we develop Wavelet-Guided Spectral Downsampling (WGSD), which conducts explicit sub-band decomposition via Haar wavelets to suppress aliasing while selectively preserving informative high-frequency components during resolution reduction. Extensive experiments on VisDrone-2019 and TT100K, together with comprehensive ablations, demonstrate that Spectra-Net consistently improves small object detection performance under severe scale imbalance.

Keywords: small object detection; scale adaptation; Fourier alignment



Academic Editor: Omneya A. Attallah

Received: 10 June 2026

Revised: 19 July 2026

Accepted: 21 July 2026

Published: 27 July 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.

1. Introduction

Object detection has achieved remarkable progress and has become a cornerstone of modern computer vision, driven by classic frameworks such as Fast R-CNN [1] and YOLO [2], as well as comprehensive surveys of the field [3]. More broadly, dense visual perception tasks such as object detection and image segmentation have increasingly benefited from task-specific representation learning, including probability-map-guided feature modeling and physical-knowledge regularization for improving structural consistency and semantic discrimination [4,5]. However, small object detection remains substantially more challenging in real-world applications. In scenarios such as UAV-based aerial perception [6,7] and autonomous driving or road obstacle perception [8–10], targets often

occupy only a few pixels. These small objects provide limited discriminative cues, making them highly susceptible to background interference and localization uncertainty [11,12]. Similar degradation is also observed in domain-specific small-target detection tasks, where weak local structures and insufficient detail preservation often require dedicated detection heads or enhanced feature representations [13]. Contextual information and region proposal strategies have been widely explored to mitigate these issues [14,15], yet robust detection in complex environments remains an open problem. More importantly, practical scenarios often exhibit severe scale imbalance, where tiny distant targets coexist with large nearby structures. This imbalance amplifies the difficulty of learning consistent representations across scales [16], necessitating more sophisticated optimization strategies for anchor design [17] and feature encoding.

To tackle these challenges, most detectors enhance small-object representations via multi-scale aggregation, attention-based refinement, and cross-level feature fusion [9,10,18,19]. Approaches such as the Extended Feature Pyramid Network (EFPN) [20] and Cross-Layer Feature Pyramid Transformer [21] attempt to integrate multi-level semantics. Recent works like HS-FPN further emphasize the importance of high-frequency perception for tiny objects [22]. However, as illustrated in Figure 1, small objects are still frequently missed under severe scale imbalance, suggesting that scale-mismatched backbone encoding remains a key bottleneck. These strategies are largely built upon a static backbone, where convolution kernels have fixed receptive fields and largely invariant effective frequency responses once training is completed. From a spectral perspective, large objects are dominated by low-frequency structures, whereas small objects rely more on weak and sparse high-frequency cues [23]. A fixed convolution operator therefore behaves like a stationary filter and cannot consistently accommodate such divergent spectral characteristics across scales. Consequently, deep features tend to favor low-frequency patterns of large structures while progressively weakening high-frequency evidence critical for small objects, leading to scale-mismatched representations.

Intuitively, this problem can be understood through an analogy with audio equalization. In audio processing, an equalizer selectively amplifies or attenuates different frequency bands to achieve a balanced sound profile. Similarly, in visual feature extraction, large and small objects occupy different frequency bands of the feature spectrum: large objects correspond to low-frequency components that capture overall shape and context, while small objects correspond to high-frequency components that encode fine-grained details and boundaries. A static backbone acts like a fixed audio filter that cannot adapt its frequency response to the content—it uniformly amplifies certain frequencies regardless of whether the scene is dominated by large structures or tiny targets. Just as an adaptive equalizer dynamically adjusts frequency gains based on the audio content, a frequency-aware backbone should dynamically modulate its spectral response based on the visual content, amplifying high-frequency cues when small objects are present while preserving low-frequency context for larger structures. This analogy motivates our approach of explicitly controlling the frequency response of convolutional kernels.

This observation suggests that improving small-object detection requires not only multi-scale aggregation in the neck, but also an input-adaptive feature encoding mechanism that can dynamically modulate its spectral response. While dynamic convolution generates content-conditioned kernels to accommodate scale variations [24,25], directly predicting spatial-domain kernels offers limited explicit control over frequency preference. It is difficult to regulate whether the adaptive kernel should emphasize low-frequency structures or preserve high-frequency details. Recent studies have begun to explore spectral enhancement for tiny object detection [26,27], indicating the potential of frequency-domain analysis. Motivated by the interpretability of Fourier parameterization, we seek a mechanism where

frequency emphasis can be adjusted explicitly. Therefore, we introduce Dynamic Fourier Alignment (DFA) to realize content-adaptive yet frequency-controllable modulation, enabling the backbone to align cross-scale representations by steering its responses toward the spectral components that best preserve tiny-object cues.

As illustrated in Figure 1, under severe scale imbalance, static backbones produce scale-mismatched features that suppress weak small-object cues. Moreover, repeated downsampling without spectral control tends to cause aliasing and erode high-frequency details critical for tiny objects. Our Spectra-Net mitigates these issues via Dynamic Fourier Alignment (DFA) for content-adaptive spectral modulation and Wavelet-Guided Spectral Downsampling (WGSD) for detail-preserving anti-aliasing downsampling.

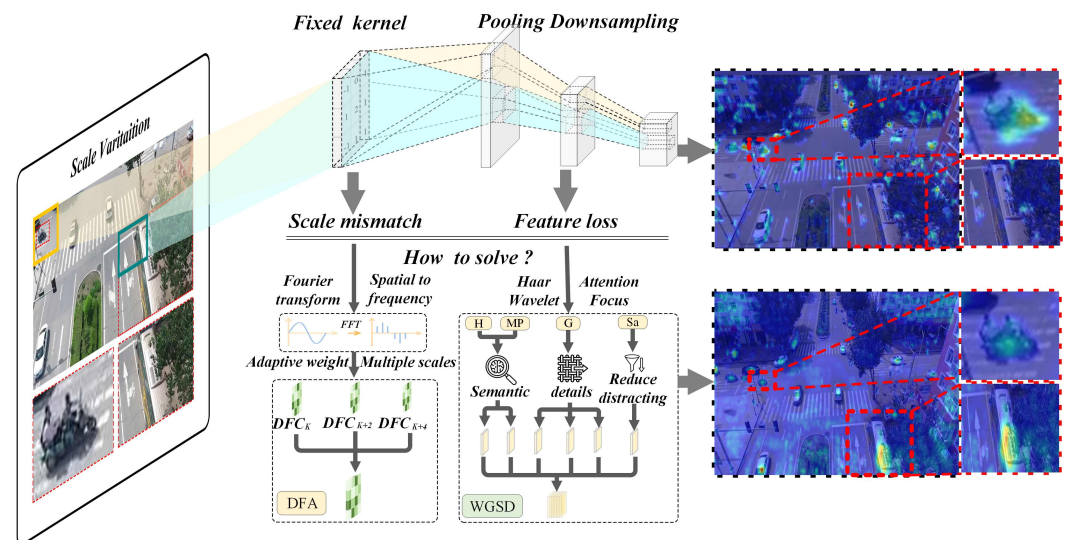


Figure 1. Scale mismatch and downsampling-induced aliasing in small object detection.

In addition to the scale mismatch induced by static convolutions, small-object cues are further degraded by ubiquitous spatial downsampling operations in backbones, such as strided convolutions and pooling [28]. Downsampling reduces the sampling rate of feature maps; without spectral control, high-frequency components associated with tiny objects can be distorted and folded into lower frequencies, resulting in aliasing and progressively corrupting fine textures. Conventional anti-aliasing relying on generic low-pass filtering may over-suppress high-frequency details indispensable for small objects. This calls for a frequency-aware downsampling operator that controls aliasing while retaining informative high-frequency components. Inspired by Haar wavelet downsampling in semantic segmentation [29], we propose to explicitly decompose features into wavelet sub-bands before resolution reduction. This enables selective sub-band retention to suppress aliasing without erasing tiny-object details.

To address the above limitations, we propose Spectra-Net, a frequency-aware detector that redesigns backbone feature encoding for robust small-object recognition under severe scale imbalance. The core component of Spectra-Net is the Dynamic Fourier Alignment (DFA) module, which enables input-adaptive spectral modulation to alleviate scale-mismatched representations. Specifically, DFA generates content-conditioned convolutional responses by leveraging global frequency statistics, allowing the backbone to dynamically adjust its effective frequency response. Complementary to DFA, we introduce a Wavelet-Guided Spectral Downsampling (WGSD) module to mitigate downsampling-induced degradation. WGSD performs Haar wavelet-based sub-band decomposition to retain informative high-frequency details while reducing spatial resolution, and combines it with spatially adaptive feature refinement to suppress background interference. By cou-

pling adaptive spectral alignment with anti-aliasing-oriented downsampling, Spectra-Net preserves and enhances small-object evidence throughout the backbone, leading to more reliable multi-scale fusion and improved detection performance in challenging real-world scenes such as VisDrone-2019 [30] and TT100K [31].

In summary, this paper makes the following contributions:

- We propose Spectra-Net, a novel frequency-aware detection framework for robust small-object recognition under severe scale imbalance.
- We introduce Dynamic Fourier Alignment (DFA), a content-adaptive and frequency-controllable modulation module that reshapes convolutional responses in the spectral domain to alleviate scale-mismatched representations.
- We further develop Wavelet-Guided Spectral Downsampling (WGSD) to alleviate downsampling-induced degradation by preserving informative wavelet sub-bands and suppressing background interference.
- Extensive experiments on VisDrone-2019 and TT100K, together with ablation studies, validate the effectiveness of Spectra-Net and demonstrate consistent improvements compared with state-of-the-art methods.

The rest of this paper is organized as follows: Section 2 reviews related work on small object detection and multi-scale feature learning. Section 3 describes the proposed Spectra-Net architecture and its key components in detail. Section 4 presents and analyzes the experimental results. Finally, Section 5 concludes the paper and discusses potential future research directions.

2. Related Work

This section reviews representative small object detection methods from four perspectives: data augmentation, anchor design, multi-scale feature learning, and context modeling.

2.1. Data Augmentation for Small Object Detection

Data augmentation is a simple yet effective strategy to enlarge the training set and improve data diversity, which often benefits small object detection by increasing the frequency of hard and rare samples [32,33]. Recent studies have explored tailored augmentation schemes for challenging scenarios. For example, Wang et al. proposed an image enhancement network (IENet) [34] with an adaptive joint filtering strategy to improve detection under complex dynamic conditions. To explicitly generate difficult instances, Chen et al. focused on hard-sample augmentation by randomly erasing pixels within object regions [35], which can improve robustness against partial occlusion and appearance corruption. Considering weather variations in outdoor scenes, Zhang et al. designed an image enhancement module with exposure [36], hue, and brightness filters to enhance generalization under diverse illumination and weather conditions. Overall, these augmentation methods improve sample diversity and complexity, helping detectors better cope with real-world disturbances and thus improving performance on small objects.

2.2. Anchor Design for Small Object Detection

For small objects, minor localization errors may cause a substantial drop in overlap due to their limited spatial extent, making anchor matching and regression particularly sensitive [12]. To improve the reliability of matching and post-processing, many works optimize anchor-related components and localization metrics. Li et al. proposed SA-NMS (Soft-Aware Non-Maximum Suppression) to reduce false positives and improve final detection quality [37]. Zhou and Zhu adopted distribution similarity [38] (e.g., Kullback–Leibler divergence, KLD) as a localization measure to better reflect subtle shifts in small bounding boxes, leading to improved localization stability. Chen et al. dynamically adjusted

the number of candidate positive anchors for each ground truth by jointly considering classification and regression losses during positive–negative assignment [39], alleviating sample imbalance and improving detection accuracy. These anchor optimization strategies enhance robustness for tiny objects, but they are often coupled with the representation quality produced by the backbone features.

2.3. Multi-Scale Feature Learning for Small Object Detection

To mitigate the progressive loss of small-object cues across network stages, multi-scale learning typically combines deep semantic information with shallow spatial details. For instance, Zhao and Zhu designed an efficient feature extraction module using multi-kernel convolutions [40] to parse features at different scales, improving scale coverage. Huang et al. proposed a scale-aware weighted loss [41] that dynamically adjusts supervision weights based on scale statistics, strengthening the learning of small objects. Li et al. utilized a bidirectional feature pyramid network [42] to enhance cross-scale information flow and significantly improve small object detection accuracy. Nevertheless, the gain from multi-scale fusion still hinges on whether backbone representations across resolutions are properly aligned and can retain the weak details of tiny objects.

2.4. Context Modeling for Small Object Detection

Since small objects often exhibit weak and ambiguous appearance cues, incorporating surrounding context can substantially improve detection. With CNN-based detectors, many studies integrate contextual cues around objects to enhance feature discrimination. Fu et al. proposed I2D-Net [43], which aggregates semantic, spatial, and contextual information to improve small defect detection in cluttered backgrounds. To preserve more spatial information for small targets, Ran et al. used upsampled images as inputs to enhance detector sensitivity to tiny objects [44]. Ma et al. designed an adaptive architecture to capture contextual information [45] at sparse locations, strengthening the representation of small objects. These results suggest that context modeling is effective, especially when local appearance is insufficient.

2.5. Discussion and Limitations

While small object detection has seen notable progress, it still faces significant hurdles [3]. However, most existing methods focus on the feature fusion neck or training strategies, while backbone encoding remains largely unaddressed and static. This leads to scale-mismatched representations, especially under drastic scale variations, where strong low-frequency signals from large objects suppress the weak high-frequency details of tiny targets [15]. Moreover, standard downsampling without spectral awareness often causes aliasing, further eroding the critical cues of small objects. Recent advances in 2025–2026 have seen growing interest in frequency-domain methods for detection. For instance, multi-spectral object detection has been explored for frequency-aware aerial detection [27], and dynamic kernel networks have been proposed for adaptive feature extraction [25]. Lightweight spectral transformers for UAV-based detection [19] and environment-aware Mamba architectures for road perception [46] further demonstrate the increasing trend toward spectral and adaptive modeling in detection pipelines. However, these methods primarily focus on the neck or attention mechanisms, while backbone-level spectral modulation remains underexplored. To bridge these gaps, it is essential to develop a backbone that can explicitly modulate its spectral response and preserve high-frequency details, which serves as the primary motivation for our Spectra-Net.

3. The Proposed Methods

3.1. Overall Architecture of Spectra-Net

The overall architecture of the proposed frequency-aware detector, Spectra-Net, is illustrated in Figure 2. Designed specifically to tackle the challenge of high-frequency information loss in small object detection, Spectra-Net follows a streamlined single-stage detection paradigm. The framework comprises three primary components: a frequency-enhanced backbone for feature extraction, a multi-scale feature fusion neck, and decoupled detection heads.

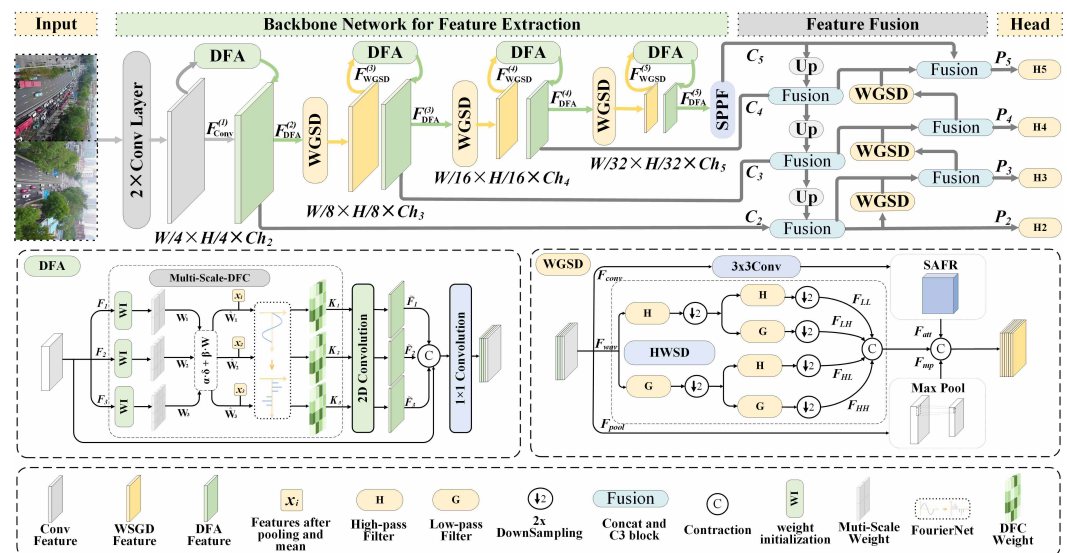


Figure 2. The overall framework of the proposed Spectra-Net. It consists of three main parts: the backbone embedded with Dynamic Fourier Alignment (DFA) and Wavelet-Guided Spectral Downsampling (WGSD), a feature fusion neck, and detection heads.

Unlike traditional architectures that rely on static convolutions and lossy pooling, we re-engineer the backbone by integrating two core spectral modules at each stage. Specifically, the Dynamic Fourier Alignment (DFA) module is employed to replace standard static convolutions. It introduces a frequency-driven dynamic mechanism to align feature representations across drastically different scales, thereby ensuring global context awareness. Complementing this, the Wavelet-Guided Spectral Downsampling (WGSD) module substitutes standard spatial downsampling operations. By combining Haar wavelet analysis with spatial attention, WGSD effectively separates and preserves high-frequency details, which are crucial for small objects but typically discarded by standard max pooling.

In the forward pass, given an input image $I \in \mathbb{R}^{H \times W \times 3}$, the backbone progressively processes the visual data through these modules. This process generates a hierarchy of spectrally enriched feature maps, denoted as $\{C_k\}_{k=2}^5$, where C_k represents the output feature map at the k -th stage with a stride of 2^k relative to the input image. These multi-scale features are then fed into a feature-pyramid neck $\mathcal{N}(\cdot)$, which performs top-down and lateral fusion to produce semantically strong features $\{P_k\}_{k=2}^5$. Finally, detection heads $\mathcal{H}(\cdot)$ are attached to each pyramid level P_k to predict category scores and bounding box coordinates. This unified spectral design allows Spectra-Net to maintain the efficiency of one-stage detectors while significantly enhancing robustness against scale variations and feature degradation.

3.2. Dynamic Fourier Alignment

Standard convolutional neural networks rely on static kernels with fixed spatial receptive fields and invariant frequency responses. While effective for local pattern extraction,

this design becomes problematic when handling objects with drastic scale variations. From a spectral perspective, large objects primarily occupy low-frequency components, whereas small objects are dominated by high-frequency details. A fixed kernel thus acts as a stationary filter that cannot simultaneously accommodate such divergent spectral characteristics. As a result, deep layers tend to over-smooth high-frequency cues essential for small objects while favoring larger structures, leading to the scale mismatch problem. To address this issue, a feature extractor should be capable of dynamically modulating its frequency response based on the input content.

Motivated by this, we propose the Dynamic Fourier Alignment (DFA) module. Unlike conventional layers, DFA integrates adaptive multi-scale alignment in the spatial domain with dynamic spectral modeling. By constructing convolution kernels conditioned on the global frequency content of the input, DFA explicitly strengthens high-frequency responses associated with small targets while suppressing background noise, ensuring robust feature alignment across scales. Specifically, DFA decomposes the input feature into multiple scale-aware branches and performs content-adaptive spectral modulation followed by self-aligned fusion.

As illustrated in the bottom-left part of Figure 2, the DFA module performs frequency-aware alignment through a multi-branch architecture, where each branch is responsible for modeling spectral patterns at a specific spatial scale. For clarity, we describe the formulation for a generic stage and omit the stage index l in the following notation. Given an input feature $F_{\text{in}} \in \mathbb{R}^{H \times W \times C}$, we first project it via a 1×1 convolution and split it channel-wise into three scale-aware branches, denoted as $\{F_1, F_2, F_3\}$, where $F_i \in \mathbb{R}^{H \times W \times \frac{C}{3}}$. To capture hierarchical frequency patterns, each branch is assigned a distinct spatial receptive field by using convolution kernels of sizes $k_i \in \{3, 5, 7\}$ for $i = 1, 2, 3$, respectively. This design allows the network to simultaneously model high-frequency local details with small kernels and low-frequency global structures with large kernels, forming the basis for cross-scale feature alignment.

Rather than adopting standard static convolutions, each branch is implemented with a content-adaptive convolution operator, termed Dynamic Fourier Convolution (DFC), which enables input-dependent frequency modulation and will be detailed in the following. After spectral modulation in each branch, the resulting multi-scale features are aggregated to form a unified representation. The fusion process is performed by channel-wise concatenation followed by a linear projection, as formulated below:

$$F_{\text{DFA}} = \text{Conv}_{1 \times 1} (\text{Concat}(\tilde{F}_1, \tilde{F}_2, \tilde{F}_3, F_{\text{in}})), \quad (1)$$

where $\tilde{F}_i$ denotes the output feature of the i -th scale-aware branch after dynamic spectral modulation, obtained by applying the Dynamic Fourier Convolution operator to the corresponding input feature, i.e., $\tilde{F}_i = \text{DFC}(F_i)$. $\text{Concat}(\cdot)$ denotes channel-wise concatenation of multi-scale features together with the original input feature F_{in} , enabling joint aggregation of dynamically modulated and original representations. Finally, $\text{Conv}_{1 \times 1}$ is applied to fuse these aggregated representations and project the concatenated channels to the target output dimension C_{out} , producing the refined feature map F_{DFA} .

Dynamic Fourier Convolution. At the core of DFA lies a content-adaptive convolution operation, which enables each branch to dynamically adjust its spectral response according to the input. As shown in Figure 3, each branch is equipped with a content-adaptive convolution mechanism to accommodate the distinct frequency characteristics of objects at different scales. Instead of applying a fixed convolutional kernel, the convolution operation in each branch is governed by a dynamically generated kernel whose response is conditioned on both the input feature and learnable convolutional weights. This design allows the convolution to flexibly emphasize high-frequency details or suppress low-

frequency background components according to the input content. Formally, for the i -th branch, the output feature $\tilde{F}_i$ is obtained by applying a content-adaptive convolution with kernel $K_i \in \mathbb{R}^{\frac{C}{3} \times \frac{C}{3} \times k_i \times k_i}$ to the branch feature F_i :

$$\tilde{F}_i = F_i \otimes K_i(F_i, \mathbf{W}_i), \quad (2)$$

where $\otimes$ denotes the convolution operation, K_i is dynamically generated based on the branch feature, and $\mathbf{W}_i$ denotes a learnable base convolutional weight.

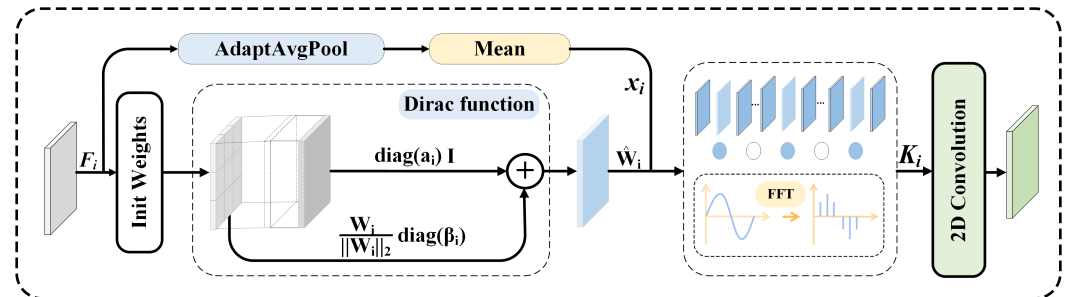


Figure 3. Structure of the Dynamic Fourier Convolution (DFC).

To realize such a content-adaptive kernel in a stable and controllable manner, the key challenge lies in incorporating input-dependent information into the kernel generation process without introducing training instability. Directly conditioning convolution kernels on input features often leads to unstable optimization, as unconstrained kernel variations can cause abrupt changes in the effective filter response, resulting in gradient explosion and inconsistent feature transformations across samples. This instability not only hampers convergence but also undermines the structural inductive bias of convolution, degrading generalization performance.

To address this issue, we construct the dynamic kernel upon a well-conditioned base kernel that preserves the structural properties of conventional convolution, while allowing controlled input adaptivity. Specifically, input-dependent information is introduced only through smooth amplitude-frequency modulation in the spectral domain, which constrains kernel variations to be bounded and continuous, thereby ensuring stable optimization. Under this formulation, we re-parameterize the convolutional weights using a Dirac-residual scheme to obtain a stabilized base kernel:

$$\hat{\mathbf{W}}_i = \text{diag}(\alpha_i)\mathbf{I} + \text{diag}(\beta_i) \frac{\mathbf{W}_i}{\|\mathbf{W}_i\|_2}, \quad (3)$$

where $\mathbf{I}$ denotes the Dirac identity kernel, and α_i and β_i are learnable scaling vectors whose diagonal forms $\text{diag}(\alpha_i)$ and $\text{diag}(\beta_i)$ control the contribution of the identity and normalized convolution kernels, respectively. The scaling vectors are initialized to 1.0 and 0.1.

Given the stabilized base kernel $\hat{\mathbf{W}}_i$, we further introduce input-dependent modulation to generate the final dynamic convolution kernel. Specifically, for each scale-aware branch, a compact global descriptor x_i is extracted from the corresponding feature F_i via adaptive average pooling followed by a mean operation, which summarizes the global statistics of F_i under a fixed receptive field and provides a stable conditioning signal for kernel modulation:

$$x_i = \text{Mean}(\text{AAP}_{k_i}(F_i)), \quad k_i \in \{3, 5, 7\}, \quad i = 1, 2, 3. \quad (4)$$

Here, $\text{AAP}_{k_i}(\cdot)$ denotes adaptive average pooling with receptive field size k_i , and $\text{Mean}(\cdot)$ computes the channel-wise mean.

To enhance the expressive power and fitting capability of the convolution kernel, the proposed Fourier convolution adopts a network structure based on multiplicative filters. Let u denote a generic input variable. Following the formulation of Multiplicative Filter Networks (MFN) [47], all nonlinearities are introduced through sinusoidal filter functions applied directly to the input. The filter function at layer i is defined as

$$g(u; \theta^{(i)}) = \sin(\omega^{(i)}u + \phi^{(i)}), \quad (5)$$

where $\theta^{(i)} = \{\omega^{(i)}, \phi^{(i)}\}$ denote the frequency and phase parameters.

Given the stabilized base kernel $\hat{\mathbf{W}}_i$, the multiplicative modulation process is initialized as

$$\mathbf{y}_i^{(1)} = g(\hat{\mathbf{W}}_i; \theta^{(1)}). \quad (6)$$

At each subsequent layer $j = 1, \dots, k-1$, the intermediate representation is updated via element-wise multiplicative interactions:

$$\mathbf{y}_i^{(j+1)} = \sigma(\mathbf{W}^{(j)}\mathbf{y}_i^{(j)} + \mathbf{b}^{(j)}) \circ g(\hat{\mathbf{W}}_i; \theta^{(j+1)}) \circ g(x_i; \tilde{\theta}^{(j+1)}), \quad (7)$$

where $\circ$ denotes element-wise multiplication, $\sigma(\cdot)$ is a nonlinear activation function, and $\mathbf{y}_i^{(j)}$ denotes the hidden representation at the j -th layer of the network.

After k layers of multiplicative modulation, the dynamic convolution kernel is obtained by a linear projection:

$$K_i = \hat{\mathbf{W}}_i^{(k)}\mathbf{y}_i^{(k)} + \mathbf{b}^{(k)}. \quad (8)$$

Since the element-wise product of sinusoidal functions can be rewritten as a linear combination of sinusoidal bases, the above recursive formulation admits an explicit Fourier expansion. Consequently, the resulting dynamic kernel can be equivalently expressed as

$$K_i = \sum_{t=1}^T a_t \left(\sin(\omega_t x_i + \phi_t) \circ \sin(\tilde{\omega}_t \hat{\mathbf{W}}_i + \tilde{\phi}_t) \right) + b, \quad (9)$$

where $\{\omega_t, \phi_t, \tilde{\omega}_t, \tilde{\phi}_t\}$ denote the effective frequencies and phase offsets induced by the multiplicative expansion, a_t are the corresponding amplitude coefficients, and T is the number of resulting sinusoidal basis components.

By using DFC as the fundamental unit of DFA, we introduce MFN-style Fourier multiplicative modulation on a Dirac-residual-stabilized base kernel, with the frequency and phase parameters conditioned on a global descriptor to generate content-adaptive convolution kernels. This constrained frequency-domain modulation mitigates weight drift and scale sensitivity, enabling cross-scale feature alignment while enhancing high-frequency details of small objects and suppressing low-frequency background interference.

Novelty of DFA Compared with Existing Approaches. It is worth highlighting the key differences between DFA and existing frequency-domain or dynamic convolution methods. Standard dynamic convolution [24] generates input-conditioned kernels by recombining static weight bases, but it does not explicitly control the frequency response of the generated kernel; the adaptivity is confined to the spatial domain without spectral regularization. The Spectral Enhancement for Tiny Object Detection (SET) [26] applies spectral enhancement at the feature level by reweighting frequency components of the feature map, but the convolution kernel itself remains static. In contrast, DFA uniquely combines content-adaptive kernel generation with explicit frequency control via multiplicative filter networks: the kernel is dynamically generated while its spectral characteristics are simultaneously constrained through sinusoidal basis functions. This dual mechanism

ensures that the modulation is both input-adaptive and frequency-controllable, which is fundamentally different from merely reweighting features or recombining spatial kernels.

3.3. Wavelet-Guided Spectral Downsampling

Spatial downsampling is a fundamental operation in convolutional neural networks, essential for expanding receptive fields and reducing computational complexity. However, standard pooling mechanisms impose critical limitations when processing small objects. From a signal processing perspective, downsampling effectively reduces the sampling rate of the feature map. According to the Nyquist–Shannon sampling theorem, if high-frequency components characterizing small objects are not properly handled, the reduction in sampling rate leads to aliasing. Specifically, Average Pooling functions as a low-pass filter that indiscriminately blurs high-frequency details, causing small targets to fade into the background. Conversely, Max Pooling performs non-linear sub-sampling by selecting peak values. While this preserves strong activations, it disrupts the spatial continuity of signals and induces severe aliasing artifacts, leading to structural distortion and the loss of weaker but distinct spatial cues. Since small objects are inherently sparse high-frequency impulses, these destructive operations result in irreversible feature loss in deeper layers.

To address this dilemma, we propose the Wavelet-Guided Spectral Downsampling (WGSD) module, as illustrated in the bottom-right part of Figure 2. Departing from single-operator pooling, WGSD introduces a hybrid downsampling strategy centered on Haar wavelet spectral decomposition [29] and complemented by spatial attention. This approach allows the network to explicitly separate and preserve critical high-frequency textures via orthogonal wavelet transformation while actively highlighting salient object regions via attention, thereby achieving information-preserving reduction suitable for small object detection.

Let F_{DFA} denote the input feature map from the preceding DFA module. To enable efficient multi-branch processing, we split F_{DFA} along the channel dimension into three parts, i.e., F_{wav} , F_{pool} , and F_{conv} , written as $F_{\text{wav}}, F_{\text{conv}}, F_{\text{pool}} = \text{Split}(F_{\text{DFA}})$. Specifically, F_{pool} preserves half of the channels to retain stable structural cues, while F_{wav} and F_{conv} each take one quarter for spectral-detail and spatial refinement, respectively.

Haar Wavelet Sub-band Decomposition. We employ the 2D discrete Haar wavelet transform (DWT) on the first sub-feature F_{wav} to capture frequency-domain characteristics. Unlike standard pooling which discards pixels, DWT performs a reversible orthogonal transformation based on the low-pass scaling function ϕ and the high-pass wavelet function ψ . As a result, F_{wav} is decomposed into four complementary spectral components:

$$\begin{aligned} F_{LL} &= (F_{\text{wav}} * \phi_r) * \phi_c, & F_{LH} &= (F_{\text{wav}} * \phi_r) * \psi_c, \\ F_{HL} &= (F_{\text{wav}} * \psi_r) * \phi_c, & F_{HH} &= (F_{\text{wav}} * \psi_r) * \psi_c, \end{aligned} \quad (10)$$

where $*$ denotes the filtering operation via strided convolution with fixed Haar kernels. Intuitively, F_{LL} corresponds to the low-frequency approximation, whereas the set $\{F_{LH}, F_{HL}, F_{HH}\}$ captures high-frequency components along the horizontal, vertical, and diagonal directions, explicitly preserving fine-grained boundary cues.

Spatially Adaptive Feature Refinement. Although the Haar decomposition in Equation (10) preserves rich sub-band details, it may also retain undesired high-frequency clutter arising from complex backgrounds. To suppress irrelevant responses while emphasizing spatially salient regions during downsampling, we introduce a Spatially Adaptive Feature Refinement mechanism. Notably, this mechanism operates directly on F_{conv} obtained from channel partitioning, rather than on the wavelet sub-bands, thereby providing complementary spatial selectivity that is decoupled from the spectral decomposition.

Specifically, we squeeze channel information of F_{conv} using average and max pooling to form two spatial descriptors, which are concatenated to predict an attention map M_s :

$$M_s = \sigma(\text{Conv}_{7 \times 7}([\text{Avg}_c(F_{\text{conv}}), \text{Max}_c(F_{\text{conv}})])), \quad (11)$$

$$F_{\text{att}} = \text{Conv}_{3 \times 3}(F_{\text{conv}} \circ M_s), \quad (12)$$

where $\text{Avg}_c(\cdot)$ and $\text{Max}_c(\cdot)$ denote pooling along the channel dimension, $\circ$ is element-wise multiplication, and $\sigma(\cdot)$ is the Sigmoid function. The stride-2 convolution aligns the output resolution with other downsampling branches.

Complementary to the spatially adaptive refinement, the F_{pool} partition is processed via a standard max-pooling operator to capture the most prominent local activations and ensure translation-invariant cues, yielding $F_{\text{mp}} = \text{MaxPool}_{2 \times 2}(F_{\text{pool}})$. Unlike the attention-based path that performs soft, spatially selective reweighting, max pooling functions as a deterministic “winner-take-all” operator that preserves peak signal intensities regardless of the learned attention weights. This mechanism is particularly effective when small objects manifest as sparse but high-magnitude impulses that might otherwise be smoothed. Finally, the WGSD module synthesizes a comprehensive representation by concatenating the wavelet sub-bands, the attention-refined features, and the max-pooled salient responses along the channel dimension:

$$F_{\text{WGSD}} = \text{Concat}(F_{LL}, F_{LH}, F_{HL}, F_{HH}, F_{\text{att}}, F_{\text{mp}}). \quad (13)$$

By unifying spectral decomposition with Spatially Adaptive Feature Refinement, WGSD effectively mitigates feature degradation, ensuring that subsequent layers receive a feature representation that is both compact and rich in high-frequency details essential for small object recognition.

Novelty of WGSD Compared with Conventional Wavelet Downsampling. While Haar wavelet downsampling has been previously explored for semantic segmentation [29], WGSD differs from it in several key aspects. First, the conventional approach [29] retains only the low-pass approximation (F_{LL}) and discards high-frequency sub-bands as “noise,” which is acceptable for segmentation where boundary precision is less critical but is detrimental for small object detection where high-frequency cues are indispensable. In contrast, WGSD explicitly retains all four wavelet sub-bands ($F_{LL}, F_{LH}, F_{HL}, F_{HH}$), preserving both low-frequency context and high-frequency boundary information. Second, WGSD further integrates a Spatially Adaptive Feature Refinement mechanism and a max-pooling branch, which work complementarily to suppress background clutter and highlight salient object regions during downsampling. This hybrid design is specifically tailored for small object detection and goes beyond simple wavelet-based resolution reduction.

3.4. Optimization Objective

Small object detection poses unique challenges, primarily due to the high sensitivity of bounding box regression to minor offsets and the extreme imbalance between foreground and background samples. To tackle these issues effectively, we train Spectra-Net using a composite loss function that synergizes CIoU for robust localization and Varifocal Loss (VFL) for quality-aware classification. For bounding box regression, we leverage the CIoU loss $\mathcal{L}_{\text{CIoU}}$ [48] to enforce precise spatial alignment. It is defined as:

$$\mathcal{L}_{\text{CIoU}} = 1 - \text{IoU} + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} + \alpha v, \quad (14)$$

where $\mathbf{b}$ and $\mathbf{b}^{gt}$ denote the predicted and ground-truth boxes, respectively. $\rho(\cdot)$ represents the Euclidean distance between their center points, while c is the diagonal length of the

smallest enclosing box. The term v , which serves to penalize aspect-ratio inconsistencies, is defined as:

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2 \quad (15)$$

where it is modulated by the trade-off factor α .

In parallel, to mitigate class imbalance while prioritizing high-quality predictions, we utilize VFL $\mathcal{L}_{\text{VFL}}$ [49]. By defining p as the predicted score and q as the target IoU quality, this loss is expressed as:

$$\mathcal{L}_{\text{VFL}}(p, q) = \begin{cases} -q(q \log(p) + (1 - q) \log(1 - p)) & \text{if } q > 0, \\ -\alpha_{\text{neg}} p^\gamma \log(1 - p) & \text{if } q = 0, \end{cases} \quad (16)$$

where the hyperparameters are set to $\gamma = 2.0$ and $\alpha_{\text{neg}} = 0.75$. Consequently, the total objective function is formulated as a weighted combination of these two components:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{VFL}} + \lambda_2 \mathcal{L}_{\text{CIoU}}, \quad (17)$$

where the balancing coefficients are set to $\lambda_1 = 1.0$ and $\lambda_2 = 2.5$.

4. Experiments

This section details the adopted datasets and their corresponding evaluation metrics. The results of ablation and comparison experiments are presented, analysed and discussed. Visual illustrations are also provided to demonstrate the experimental outcomes intuitively.

TT-100K [31] traffic dataset consists of 9170 images of which 6105 images are used for training and 3065 images are used for validation. The images have resolutions 2048×2048 and cover a variety of lighting and weather conditions. The traffic signs range in size from 8×8 to 400×400 and occupy approximately 0.001% to 4% of the entire image. To ensure that each category has enough data, categories with fewer than 100 instances are disregarded following the operations in [31], resulting in a total of 45 categories in the TT-100K dataset. The final dataset contains 45 categories, as shown in Figure 4.

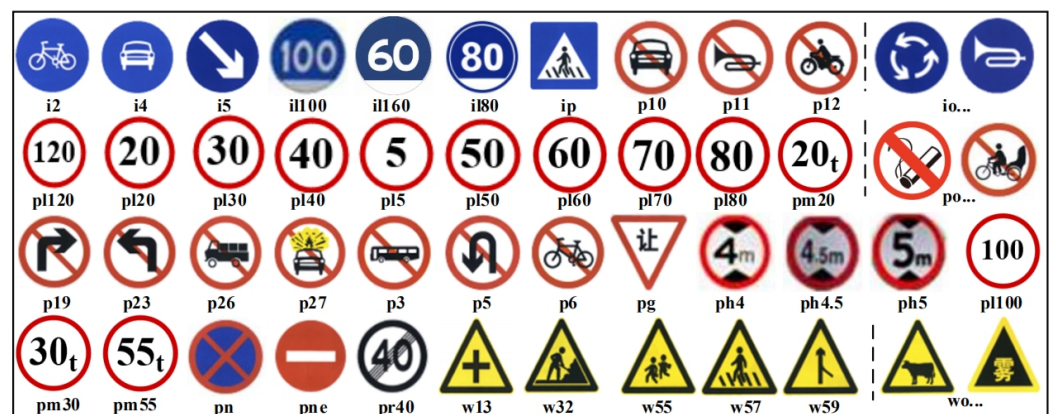


Figure 4. Forty-five traffic sign categories in the TT-100K dataset, each with at least 100 instances collected under diverse road conditions. pg: make way, wo...(the second sign): fog.

VisDrone-2019 [30] dataset consists of 8599 images captured by drone platforms at various locations and altitudes. The dataset is characterized by small object sizes, dense distributions, and potential partial obstructions. Additionally, different scenes exhibit varying lighting and perspective effects. This dataset includes over 540,000 annotated bounding boxes, divided into ten predefined categories: pedestrians, people, bicycles, cars, vans, trucks, tricycles, awning tricycles, buses, and motorcycles. The training and

validation subsets consist of 6471 and 548 images, respectively, from different locations but set in similar environmental settings, as shown in Figure 5. Evaluation Metrics: Following the COCO protocol, we report two primary metrics: (1) mAP_{50} , the mean Average Precision averaged over all categories at a fixed IoU threshold of 0.5; and (2) $AP_{50:95}$, the standard COCO metric computed as the average of AP over ten IoU thresholds from 0.5 to 0.95 with a step size of 0.05. Size-specific metrics AP_s , AP_m , and AP_l are also reported following the same COCO protocol (averaged over IoU 0.50:0.95) for small ($<32 \times 32$), medium ($32\text{--}96$ pixels), and large ($>96 \times 96$) objects, respectively. For brevity, we use “mAP” to denote mAP_{50} and “AP” to denote $AP_{50:95}$ in all tables.



Figure 5. Ten object categories in the VisDrone-2019 dataset, featuring small, dense targets captured by UAVs in complex urban scenes.

4.1. Datasets

Small Target Range: In the COCO (Common Objects in Context) dataset [50], the definition of small targets is classified according to the area size of the target. Specifically, the COCO dataset classifies targets into three size categories: small, medium, and large. Among them, targets with an area smaller than 32×32 pixels are defined as small targets, targets with an area smaller than 96×96 pixels are defined as medium targets, and the remaining targets are defined as large targets. This classification is useful for evaluating the detection performance of targets of different sizes, because targets of different sizes usually show different detection difficulties in detection algorithms.

4.2. Implementation Details

The module framework in this paper is implemented using PyTorch 2.12.0, with extensive reliance on its deep learning libraries for efficient computation. The dataset images are uniformly resized to 640×640 to ensure consistency in the input dimensions. A batch size of 6 is selected, balancing GPU memory utilization and computational speed. The training process is conducted on two NVIDIA TITAN RTX GPUs with 24 GB of VRAM each (NVIDIA, Santa Clara, CA, USA), enabling parallel processing and speeding up model convergence. For optimization, we use the Stochastic Gradient Descent (SGD) algorithm with a momentum of 0.93 to accelerate gradient vectors in the correct direction and improve convergence rates. All other optimizer parameters, such as dampening and weight decay, are kept at their default settings in PyTorch. The initial learning rate is set to 1×10^{-4} , gradually decaying over time, and a weight decay of 1×10^{-5} is applied to prevent overfitting by penalizing large weight values. The model is trained over 200 epochs, with early stopping based on validation loss to prevent overfitting. Data augmentation techniques, such as random horizontal flipping, scaling, and color jittering, are applied during preprocessing to improve the model’s generalization ability.

4.3. Comparison with State-of-the-Art Methods

To verify the effectiveness of this method in detecting small objects, it was compared with several state-of-the-art recognition methods benchmarked on the VisDrone-2019 remote sensing dataset and the TT-100K traffic dataset. The cutting-edge small object detection methods can be divided into four main categories: (i) Anchor-based small object detection methods: DTSS [39], KLDet [38], SA-NMS [37]. (ii) Data augmentation-based small object detection methods: TDOF [36], VATSD [34], CDFE [32], TSRA [35]. (iii) Multi-scale learning-based small object detection methods: TS-DETR [51], CFPT [21], SOD [42], FGE [41], CBPH-net [40], Yolo-cs [52], GMDRA [53], FBRT-YOLO [54]. (iv) Context learning-based small object detection methods: CEASC [55], SDM [45], HRGiraffe-Det [44], I2D-Net [43]. In addition, we include two recent 2026 methods for comparison: SPAR-Det [56], a segmentation-guided and prior-aided routing detector, and CA-INR [57], a coordinate-aware implicit neural representation method for UAV small object detection. Figure 6 illustrates the overall mAP₅₀ performance of the proposed method on TT-100K and VisDrone-2019 datasets.

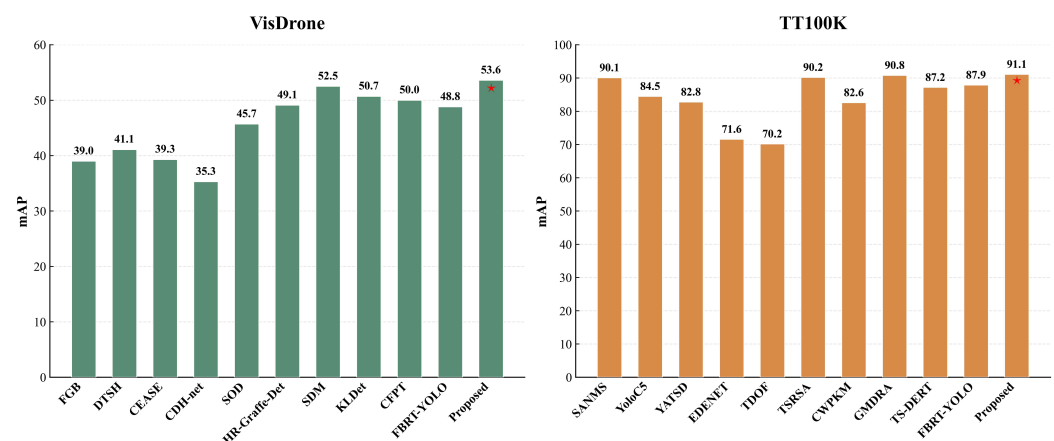


Figure 6. Performance evaluation of different small object detection methods on standard dataset based on mAP₅₀ metrics.

According to the results reported in Table 1, the method proposed in this paper is more competitive compared to other methods. Specifically, it achieves an mAP₅₀ of 53.6% on the VisDrone-2019 benchmark, which is 1.4%, 1.8%, 4.8%, 3.6%, 0.9%, 1.1%, 4.5%, 7.9%, 18.3%, 14.3%, 12.2% and 14.6% higher than other methods respectively. Notably, our method outperforms the most recent 2026 methods (SPAR-Det and CA-INR) by 1.8% and 1.4% in mAP₅₀, demonstrating that the proposed frequency-aware backbone design remains competitive against the latest approaches. Additionally, its AP_{50:95} metrics also surpass other comparative methods. Focusing on different scales of target categories, it is evident that for the crucial metric APs for small objects, this method achieves the best results. Particularly, this method shows the smallest differences among various scales of targets, reflecting its high adaptability to multi-scale targets.

To further validate the effectiveness and universality of this method, experiments were conducted on the TT100K dataset. Given the larger and more similar categories included in the TT-100K dataset, this extended experiment posed greater challenges and practicality. The recognition accuracy results for different methods on the TT-100K traffic dataset are shown in Table 2. This method is more competitive than others. Specifically, it achieves an mAP₅₀ of 91.1% on the TT-100K benchmark, which is 0.2%, 0.6%, 3.2%, 0.3%, 0.8%, 0.9%, 20.9%, 19.5%, 8.3%, 6.6%, 3.9% and 1.0% higher than other methods respectively. The results confirm that our method also outperforms the latest 2026 approaches (SPAR-Det and CA-INR) on this dataset. Additionally, to further demonstrate the effectiveness of

the proposed method, the latest small object detection network GMDRA was chosen for comparison with the detection results.

Table 1. Performance evaluation of different small object detection methods on the standard VisDrone-2019 dataset. Bold fonts indicate the best results.

Method	mAP ₅₀	AP _{50:95}	AP _s	AP _m	AP _l	Year
FGE [41]	39.0	23.4	13.0	35.6	41.4	2024
DTSS [39]	41.4	25.5	18.6	34.3	41.2	2024
CEASC [55]	39.3	20.7	11.4	31.1	39.5	2024
CBPH-net [40]	35.3	-	12.4	32.8	43.4	2023
SOD [42]	45.7	26.3	15.6	37.6	47.6	2024
HRGiraffe-Det [44]	49.1	30.2	21.3	42.0	43.3	2023
SDM [45]	52.5	30.2	22.6	39.6	39.8	2023
KLDet [38]	52.7	28.5	20.0	31.9	41.7	2023
CFPT [21]	50.0	29.7	19.7	41.9	48.0	2025
FBRT-YOLO [54]	48.8	30.7	19.4	41.7	45.3	2025
SPAR-Det [56]	51.8	32.1	21.5	42.5	47.0	2026
CA-INR [57]	52.2	32.5	21.9	43.0	47.2	2026
Proposed	53.6	33.8	22.8	43.7	48.1	-

Table 2. Performance evaluation of different small object detection methods on the standard TT-100K dataset based on mAP₅₀ metrics. Bolded fonts represent the best detection results.

Method	Year	mAP ₅₀ (%)
SA-NMS [37]	2024	90.1
YOLO-CS [52]	2024	84.5
VATSD [34]	2024	82.8
I2D-Net [43]	2023	71.6
TDOF [36]	2024	70.2
TSRSA [35]	2022	90.2
CDFF [32]	2023	90.3
GMDRA [53]	2024	90.8
TS-DETR [51]	2025	87.2
FBRT-YOLO [54]	2025	87.9
SPAR-Det [56]	2026	90.5
CA-INR [57]	2026	90.9
Proposed	-	91.1

As shown in Figure 7, when using GMDRA for detection, some instances of missed and false detections occur. In contrast, our method can effectively avoid such missed detections. Using our approach, traffic signs can be accurately identified. These results demonstrate that our method can effectively detect distant, tiny traffic signs, thereby ensuring that autonomous driving systems can accurately plan traffic routes.

Evaluation on the DIOR Dataset. To further validate the effectiveness of Spectra-Net, we conduct experiments on the DIOR dataset [58], a large-scale optical remote sensing benchmark for object detection. The DIOR dataset is chosen as an additional evaluation benchmark because it provides diverse object categories and scales in remote sensing scenarios, complementing the UAV-based VisDrone-2019 and the traffic-sign-focused TT-100K datasets. Following the standard split, we train both the baseline and our method on the DIOR training set and evaluate on the validation set. As shown in Table 3, Spectra-Net achieves an mAP₅₀ of 83.6%, outperforming the baseline by 2.1%. In terms of object scale,

our method improves AP_s , AP_m , and AP_l by 1.8%, 4.0%, and 2.0%, respectively. The most prominent performance gain is observed on medium-sized objects, while consistent improvements are also achieved on both small and large objects. This result demonstrates that the frequency-aware backbone effectively enhances feature representations across all scales: it delivers a particularly strong boosting effect on medium-scale targets, and meanwhile effectively refines fine-grained feature representations critical for small object detection, while maintaining robust performance on large objects in remote sensing scenarios.

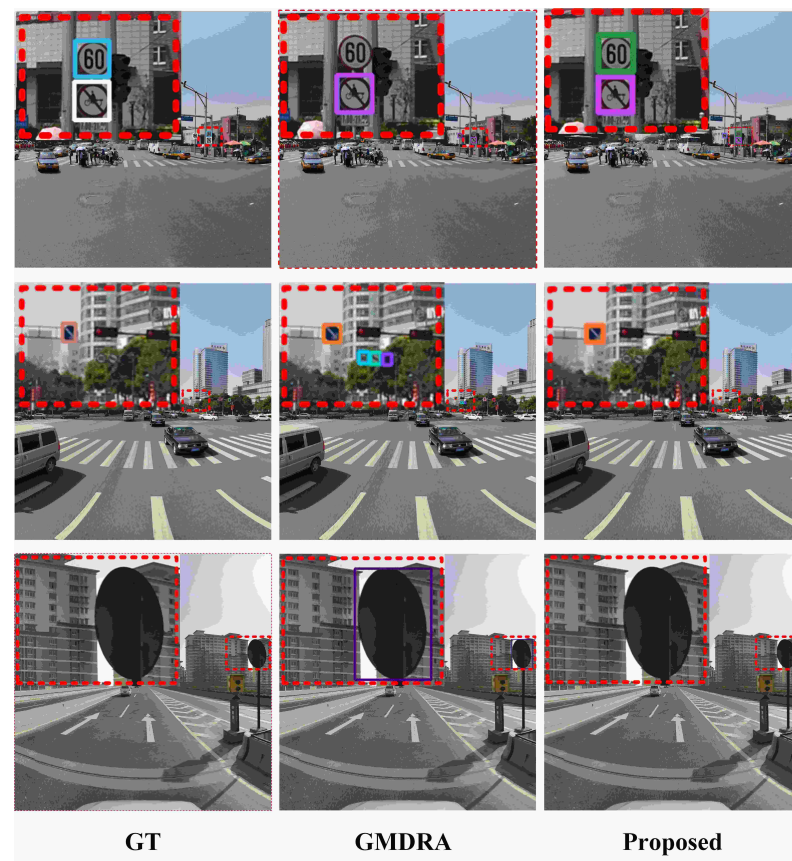


Figure 7. The practical detection performance comparison between the proposed method and GMDRA. For better observation, a magnified image of the detection region is shown in the upper left corner.

Table 3. Performance comparison on the DIOR validation set. The best results are marked in bold.

Method	mAP ₅₀	AP _s	AP _m	AP _l
Baseline	81.5	10.8	39.2	75.8
Spectra-Net (Ours)	83.6	12.6	43.2	77.8
Gain	+2.1	+1.8	+4.0	+2.0

Comparison with Frequency-Aware and Dynamic Convolution Methods. To further demonstrate the advantage of our DFA module, we compare it with two representative approaches: SET [26], a recent spectral enhancement method for tiny object detection, and DyConv [24], a standard dynamic convolution method. Specifically, SET [26] operates at the feature level by reweighting frequency components of the feature map, while DyConv [24] generates input-conditioned kernels by recombining static weight bases in the spatial domain. As shown in Table 4, all three methods improve detection performance over the baseline, but DFA achieves the best results. Specifically, SET improves AP_{50:95}

by 0.9% and AP_s by 0.5% over the baseline by reweighting frequency components at the feature level, but it does not modify the convolution kernel itself. DyConv improves $AP_{50:95}$ by 0.6% and AP_s by 0.3% through input-conditioned kernel recombination, but lacks explicit frequency control. In contrast, our DFA achieves a $AP_{50:95}$ improvement of 1.1% and AP_s improvement of 1.0% over the baseline, outperforming both SET and DyConv. This is because DFA uniquely combines content-adaptive kernel generation with explicit frequency control, enabling the backbone to dynamically steer its spectral response toward high-frequency cues critical for small objects.

Table 4. Comparison of DFA with SET and DyConv on VisDrone-2019. The best results are marked in bold.

Method	$AP_{50:95}$	AP_s	AP_m	AP_l
Baseline	32.4	21.8	42.0	45.4
Baseline + SET [26]	33.3	22.3	42.8	46.5
Baseline + DyConv [24]	33.0	22.1	42.5	46.2
Baseline + DFA (Ours)	33.5	22.8	43.2	47.3

4.4. Ablation Study

To further verify the effectiveness of the main components of the proposed method, including the DFA module, and WGSD module, this section conducts an ablation study. This phase utilizes five metrics— $AP_{50:95}$, mAP_{50} , AP_s , AP_m , and AP_l —to comprehensively evaluate each module. The results of the ablation experiments are shown in Table 5. This experiment uses the standard single-stage target detection network Yolo as the baseline module. Different components are then gradually added to verify their effectiveness. Specifically, the performance of the following five modules was compared:

Table 5. Performance of mAP_{50} and $AP_{50:95}$ metrics for ablation experiments on the VisDrone-2019 benchmark dataset. The best results are marked in bold.

Modules			Detection Metrics (%)				
Baseline	DFA	WGSD	$AP_{50:95}$	mAP_{50}	AP_s	AP_m	AP_l
✓	✗	✗	32.4	52.0	21.8	42.0	45.4
✓	✓	✗	33.5 (+1.1)	53.3 (+1.3)	22.8 (+1.0)	43.2 (+1.2)	47.3 (+1.9)
✓	✗	✓	33.3 (+0.9)	53.1 (+1.1)	22.5 (+0.7)	43.0 (+1.0)	47.7 (+2.3)
✓	✓	✓	33.8 (+1.4)	53.6 (+1.6)	22.8 (+1.0)	43.7 (+1.7)	48.1 (+2.7)

Baseline. This involves using a trained single-object detection network for detection, serving as the benchmark model. As shown in Table 5, the baseline model applied directly to the dataset produces the poorest test results with $AP_{50:95}$ and mAP_{50} values of 32.4% and 52.0%, respectively. Focusing on different object scales, the results are AP_s , AP_m , and AP_l at 21.8%, 42.0%, and 45.4%, respectively.

Analysis of DFA. To assess the impact of the DFA module, we first integrate it into the baseline. As shown in Table 5, compared to the baseline and baseline with DFA, $AP_{50:95}$ and mAP_{50} increase by 1.1% and 1.3% respectively. AP_s , AP_m , and AP_l improve by 1.0%, 1.2%, and 1.9%, respectively, demonstrating the effectiveness of DFA. This is due to the global perspective of DFC, which improves scale invariance when processing objects of different sizes. **Analysis of WGSD.** We then evaluate the effect of integrating WGSD into the baseline. As observed in Table 5, compared to the baseline and baseline with WGSD, $AP_{50:95}$ and mAP_{50} increase by 0.9% and 1.1%, respectively. AP_s , AP_m , and AP_l improve by

0.7%, 1.0%, and 2.3%, respectively, verifying the effectiveness of the WGSD. This is due to the module's ability to retain as much image information as possible during downsampling and to enhance the model's ability to capture small objects through Spatially Adaptive Feature Refinement. Joint Effect of DFA and WGSD. This represents the complete model proposed. As shown in Table 5, compared to the baseline, $AP_{50:95}$ and mAP_{50} increase by 1.4% and 1.6%, respectively. AP_s , AP_m , and AP_l improve by 1.0%, 1.7%, and 2.7%, respectively, confirming the effectiveness of the designed small object detection network. The design of the self-aligning module promotes effective fusion of features at different scales, combined with the global perspective of Fourier convolution, enhancing scale invariance when processing objects of different sizes. Combined with the performance of the Haar downsampling, this scale invariance and adaptability to complex scenes enable the model to maintain stable performance in varying environments. Frequency-Domain Interpretation of Ablation Results. To better understand why DFA and WGSD individually and jointly improve small-object detection, we interpret the ablation results from a frequency perspective. DFA enhances high-frequency responses associated with small objects while suppressing low-frequency background components, which is directly reflected in the AP_s gain (+1.0% over baseline). By dynamically steering the convolutional spectral response toward high-frequency cues, DFA ensures that tiny targets retain discriminative features even in deep layers. WGSD, on the other hand, preserves aliasing-free high-frequency details during downsampling through wavelet sub-band decomposition, leading to better localization of small objects (AP_s +0.7%). Without spectral control, standard downsampling folds high-frequency components into lower frequencies, corrupting the spatial structure of tiny targets. The joint effect of DFA and WGSD is complementary rather than redundant: DFA aligns cross-scale feature representations before downsampling, while WGSD protects these aligned features from further degradation during resolution reduction. This sequential protection pipeline explains why the combined model achieves the highest AP_s (22.8%) and the largest overall $AP_{50:95}$ improvement (+1.4%), demonstrating that spectral alignment and anti-aliasing downsampling address different stages of the same frequency degradation problem.

Ablation Study of DFA. As DFA is a core component of Spectra-Net, we first conduct an ablation study to analyze the contributions of its interaction part and the filtering block (DFC). In the first experiment, we replace the filtering block with a standard convolution, denoted as DFA (Conv), while keeping the interaction part fixed. As shown in Table 6, DFA (Conv) improves $AP_{50:95}$ from 32.4% to 32.8% and mAP_{50} from 52.0% to 52.6%. For different object scales, AP_s , AP_m , and AP_l increase by 0.3%, 0.2%, and 1.0%, respectively, indicating the contribution of the interaction part. When the standard convolution is replaced with the proposed DFC (DFA (Ours)), mAP_{50} and $AP_{50:95}$ further increase to 53.3% and 33.5%, respectively. Additionally, AP_s , AP_m , and AP_l improve to 22.8%, 43.2%, and 47.3%, respectively, demonstrating that DFC provides the primary performance gain by enhancing scale invariance through its global perspective on object sizes.

Table 6. Component-wise ablation of DFA. “Interaction” denotes the fixed interaction part inside DFA. The best results are marked in bold.

Variant	Interaction	DFC	$AP_{50:95}$	mAP_{50}	AP_s	AP_m	AP_l
Baseline	✗	✗	32.4	52.0	21.8	42.0	45.4
DFA (Conv)	✓	✗	32.8 (+0.4)	52.6 (+0.6)	22.1 (+0.3)	42.2 (+0.2)	46.4 (+1.0)
DFA (Ours)	✓	✓	33.5 (+1.1)	53.3 (+1.3)	22.8 (+1.0)	43.2 (+1.2)	47.3 (+1.9)

To provide a more intuitive visualization of the ablation results, Figure 8 presents the mAP_{50} and AP_s improvements from both the main ablation study (Table 5) and the DFA component ablation (Table 6) as bar charts. The visual comparison clearly shows that the full model (DFA + WGSD) achieves the highest performance across both metrics and that the DFC filtering block contributes the primary gain within DFA.

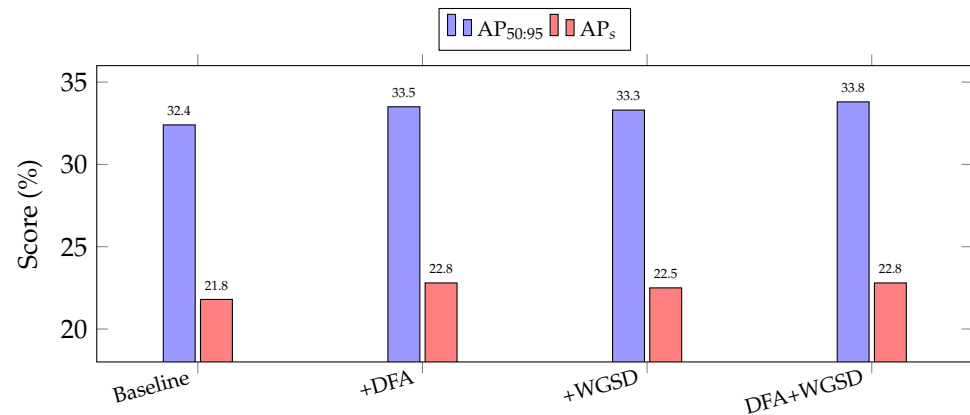


Figure 8. Bar chart visualization of the main ablation results ($AP_{50:95}$ and AP_s) on VisDrone-2019. The full model with both DFA and WGSD achieves the best performance.

Depth Study of DFC. Building on the above findings, we further investigate the internal design of DFC by varying its network depth while keeping all other settings constant. We use $AP_{50:95}$ and mAP_{50} as evaluation metrics and report the results in Table 7. The performance of the single-layer network steadily improves during training, with $AP_{50:95}$ increasing from 28.4% to 32.6% and mAP_{50} from 46.6% to 52.3%. Despite its simplicity and limited capacity, this design is easy to optimize and helps reduce overfitting on smaller datasets. The double-layer network follows a similar trend, with mAP_{50} increasing from 45.6% to 52.0%, but it performs slightly worse than the single-layer variant at 200 epochs. This suggests that while increasing the model capacity, the double-layer network does not significantly enhance feature capture in this setup. The triple-layer network achieves the best performance, with mAP_{50} increasing from 46.0% to 53.2% and $AP_{50:95}$ from 28.0% to 33.4%. This demonstrates that a moderate network depth balances representation power and generalization, allowing the model to capture more complex features without overfitting. In contrast, the quadruple-layer network performs the worst, with mAP_{50} only reaching 49.9%. This indicates that excessive depth leads to overfitting and reduces training efficiency, likely due to the model learning noise rather than useful features. In conclusion, the triple-layer design strikes the best balance between complexity and generalization and is therefore adopted as the default configuration for DFC in DFA.

Table 7. Performance metrics at different epochs for different network layers. The best results are marked in bold.

Epoch	Layer 1		Layer 2		Layer 3		Layer 4	
	$AP_{50:95}$	mAP_{50}	$AP_{50:95}$	mAP_{50}	$AP_{50:95}$	mAP_{50}	$AP_{50:95}$	mAP_{50}
50	28.4	46.6	27.7	45.6	28.0	46.0	26.8	43.8
100	31.6	50.9	31.5	50.7	31.6	50.9	29.8	48.3
150	32.2	51.8	32.2	51.8	32.8	52.5	30.8	49.6
200	32.6	52.3	32.3	52.0	33.4	53.2	31.0	49.9

4.5. Visualization Analysis

In this section, we utilize visualizations to qualitatively compare the detection results with the baseline. As shown in Figure 9, our method can identify difficult objects more effectively than the baseline. Since most objects in the VisDrone2019 [47] images are typically small, we have enlarged some samples for clearer display. The results demonstrate that, facing different complex scenes, this method is relatively more effective and robust compared to the baseline. For example, as shown in the enlarged part of the first row, the baseline model overlooks the largest bus, but our method perfectly adapts to this scale change, paying attention to the bus while ensuring the accuracy of other target identifications. In the second row, the baseline incorrectly identifies a van, whereas our model accurately recognizes the type of vehicle. The baseline neglects pedestrians on the roadside, while the proposed method also gives sufficient attention to smaller, more peripheral targets. These visual results prove that the proposed method enables the detector to better face challenges brought by scale changes and retains enough feature information to help the detector better understand difficult targets, thereby improving detection performance.

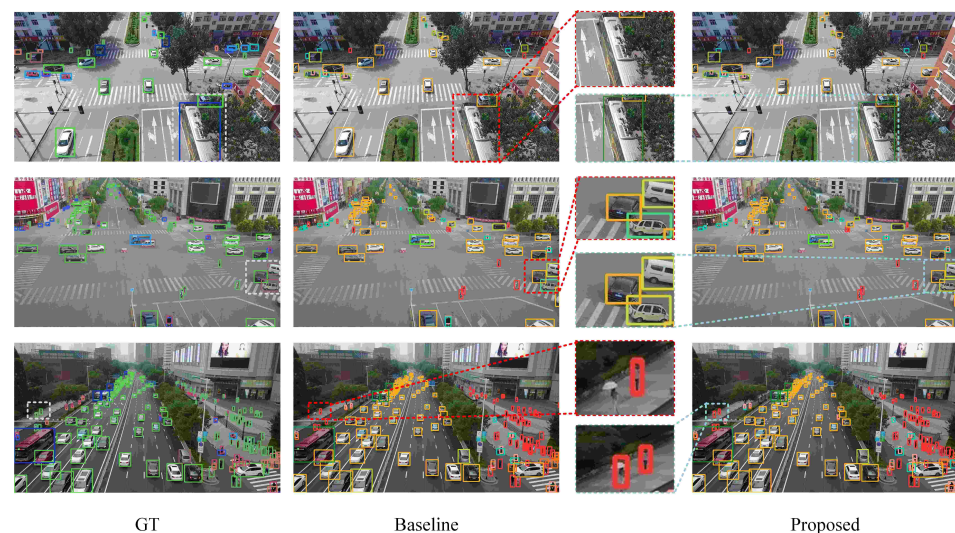


Figure 9. Some result comparison between the proposed and the baseline method on the validation set of VisDrone-2019. Dotted boxes indicate that our method exhibits outstanding performance in challenging areas, adapts better to scale variations, and can precisely locate and identify difficult objects.

Heatmaps are an important visualization tool mainly used to represent the spatial distribution of specific types of information in images. They utilize different colors to indicate varying intensities of signals, aiding in the understanding and analysis of image data. As shown in Figure 10, we present a heatmap comparison between our method and the baseline. We have also enlarged some samples for clearer display. The results show that, facing different complex scenes, this method has stronger effectiveness and robustness compared to the baseline. For instance, as shown in the enlarged part of the first row, the baseline model overlooks the largest bus because it does not give sufficient attention to it. This is reflected in the heatmap, where the bus receives less attention compared to smaller targets like cars, but our method perfectly adapts to this scale change, giving it sufficient attention in the heatmap. In the second row, the baseline overlooks pedestrians on a distant overpass, while our model accurately focuses on these targets. The third row compares the recognition of occluded targets; unlike the baseline, our method recognizes motorcycles obscured by trees, demonstrating its superiority in identifying occluded targets. These heatmap visualizations prove that the proposed method allows

the detector to dynamically focus on targets of various scales and locations, adapting to challenges brought by scale changes.

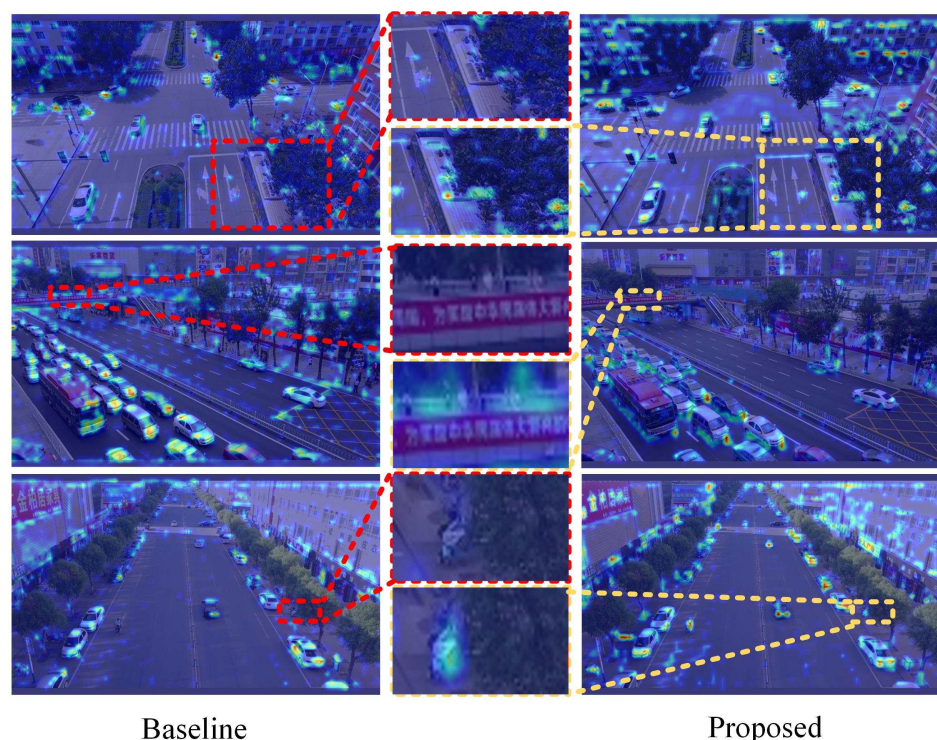


Figure 10. Heatmap visualizations of some challenging areas reveal why the proposed method adapts better to scale variations and accurately identifies potential targets.

4.6. Computational Cost Analysis

Since the proposed frequency-aware operations (DFA and WGSD) introduce additional computational modules, it is important to evaluate their impact on model complexity and inference efficiency. Table 8 reports the number of parameters, FLOPs (floating-point operations per second), and inference FPS (frames per second) for the baseline and the full Spectra-Net model, measured on a single NVIDIA TITAN RTX GPU with an input resolution of 640×640 .

Table 8. Computational cost comparison between the baseline and the proposed Spectra-Net. The best results are marked in bold.

Model	Params (M)	FLOPs (G)	FPS
Baseline	26.08	113.0	68
Spectra-Net (Ours)	30.65	130.8	56
Change	+17.5%	+15.8%	−17.6%

As shown in Table 8, Spectra-Net introduces a moderate computational overhead compared with the baseline, increasing the number of parameters and FLOPs by 17.5% and 15.8%, respectively. Meanwhile, the inference speed decreases from 68 FPS to 56 FPS, corresponding to a 17.6% reduction. Despite this additional cost, Spectra-Net still maintains real-time inference capability while providing improved detection accuracy.

5. Conclusions

We presented Spectra-Net, a frequency-aware framework for small object detection under severe scale imbalance. By redesigning backbone feature encoding from a spectral perspective, Spectra-Net strengthens high-frequency cues critical to tiny objects via Dynamic Fourier Alignment (DFA) and alleviates downsampling-induced aliasing through Wavelet-Guided Spectral Downsampling (WGSD) with wavelet sub-band decomposition and selective retention of informative high-frequency components. Experiments on VisDrone-2019 and TT100K demonstrate consistent improvements over strong baselines, achieving 53.6% mAP₅₀ and 91.1% mAP₅₀, respectively, with clear gains on small-object subsets.

Technical Findings. Our key technical findings are threefold. First, static convolutional backbones inherently suffer from scale-mismatched representations because their fixed frequency responses cannot accommodate the divergent spectral characteristics of large and small objects simultaneously. By introducing DFA, which generates content-adaptive kernels with explicit frequency control via multiplicative filter networks, the backbone can dynamically align its spectral response to preserve high-frequency cues for small objects (+1.0% AP_s on VisDrone-2019). Second, conventional downsampling operations (e.g., max pooling, average pooling) cause aliasing that irreversibly corrupts fine-grained details of tiny targets; our WGSD module addresses this by retaining all wavelet sub-bands and refining them with spatial attention, yielding aliasing-free high-frequency preservation (+0.7% AP_s). Third, DFA and WGSD are complementary: DFA aligns features before downsampling, while WGSD protects the aligned representations from further degradation, and their joint application achieves the best performance (+1.4% AP_{50:95}).

Practical Implications. The consistent improvements on VisDrone-2019 and TT100K, which represent challenging UAV-based aerial perception and autonomous driving scenarios, respectively, demonstrate that Spectra-Net is well-suited for real-world deployment. The model runs at approximately 56 FPS on a single TITAN RTX GPU, meeting real-time requirements for traffic sign recognition and aerial surveillance. The frequency-aware design is particularly beneficial in environments where small objects are prevalent and scale variations are extreme, such as urban traffic monitoring, drone-based search and rescue, and infrastructure inspection.

Limitations. Despite the encouraging results, Spectra-Net has several limitations. First, while the moderate computational overhead (+17.5% parameters, +15.8% FLOPs) is acceptable for GPU-based deployment, it may still be too heavy for resource-constrained edge devices. Second, the model's performance degrades under extreme occlusion scenarios where small objects are almost entirely obscured, as the spectral cues become too weak to be recovered. Third, very low-resolution inputs (e.g., below 32 × 32 pixels) pose challenges even for the frequency-aware modules, as the high-frequency information is inherently limited at such resolutions.

Future Directions. Several promising directions remain for future research. First, we plan to extend Spectra-Net to video-based small object detection, where temporal information can complement the spectral cues and help recover occluded targets. Second, model compression techniques such as knowledge distillation and neural architecture search will be explored to reduce the computational overhead and enable deployment on lightweight platforms. Third, integrating Spectra-Net with multi-object tracking frameworks could enable end-to-end perception pipelines for autonomous systems. Finally, evaluating the framework on additional datasets and domains (e.g., satellite imagery, underwater detection) would further validate its generalizability.

Author Contributions: Conceptualization, Y.X. and S.L.; methodology, Y.X. and G.Q.; software, Y.X.; validation, Y.X., K.W. (Kaiwang Wang) and K.W. (Kunpeng Wang); formal analysis, K.W. (Kaiwang

Wang) and D.Y.; investigation, Y.X. and K.W. (Kunpeng Wang); resources, S.L.; data curation, Y.X. and K.W. (Kaiwang Wang); writing—original draft preparation, Y.X.; writing—review and editing, G.Q., D.Y. and S.L.; visualization, Y.X. and D.Y.; supervision, S.L.; project administration, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Chongqing Municipal Education Commission: KJZD-M202500604; Science and Technology Innovation Key R&D Program of Chongqing: CSTB2023TIAD-STX0019; Sichuan Science and Technology Program (2026NSFSC0450); Central Government Program for Guiding Local Regional Innovation System Development Project: 202502ZYDF029.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available: TT-100K (<https://cg.cs.tsinghua.edu.cn/traffic-sign/> (accessed on 19 July 2026)), VisDrone-2019 (<https://visdrone.github.io/> (accessed on 19 July 2026)), and DIOR (<https://data.mendeley.com/datasets/vvrhgbr643> (accessed on 19 July 2026)).

Acknowledgments: The authors would like to thank the editors and anonymous reviewers for their valuable comments and suggestions to improve this paper. We also appreciate the guidance from our research team. Meanwhile, we sincerely thank the developers of the TT-100K, VisDrone-2019 and COCO datasets for providing the public data resources used in this work.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

mAP ₅₀	Mean Average Precision at IoU = 0.5
AP _{50:95}	Average Precision averaged over IoU 0.50:0.95
IoU	Intersection over Union
UAV	Unmanned Aerial Vehicle
DFA	Dynamic Fourier Alignment
WGSD	Wavelet-Guided Spectral Downsampling
DFC	Dynamic Fourier Convolution

References

1. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 1440–1448.
2. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
3. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. <https://doi.org/10.1109/JPROC.2023.3238524>.
4. Zhu, Z.; Zhang, Z.; Qi, G.; Li, Y.; Yang, P.; Liu, Y. Probability Map-Guided Network for 3D Volumetric Medical Image Segmentation. *IEEE Trans. Image Process.* **2025**, *34*, 7222–7234. <https://doi.org/10.1109/TIP.2025.3623259>.
5. Ding, Y.; Li, S.; Li, H.; Qi, G.; Cong, B.; Gong, Y.; Zhu, Z. Physical Regularization Loss: Integrating Physical Knowledge to Image Segmentation. *Int. J. Comput. Vis.* **2026**, *134*, 137.
6. Zhang, L.; Xiong, N.; Pan, X.; Yue, X.; Wu, P.; Guo, C. Improved Object Detection Method Utilizing YOLOv7-Tiny for Unmanned Aerial Vehicle Photographic Imagery. *Algorithms* **2023**, *16*, 520. <https://doi.org/10.3390/a16110520>.
7. Wang, S.; Zhang, Z.; Chao, Q.; Yu, T. AFE-YOLOv8: A Novel Object Detection Model for Unmanned Aerial Vehicle Scenes with Adaptive Feature Enhancement. *Algorithms* **2024**, *17*, 276. <https://doi.org/10.3390/a17070276>.
8. Tahir, N.U.A.; Zhang, Z.; Asim, M.; Chen, J.; ELAffendi, M. Object Detection in Autonomous Vehicles under Adverse Weather: A Review of Traditional and Deep Learning Approaches. *Algorithms* **2024**, *17*, 103. <https://doi.org/10.3390/a17030103>.

9. Cheng, S.; Chen, L.; Yang, K. DGSS-YOLOv8s: A Real-Time Model for Small and Complex Object Detection in Autonomous Vehicles. *Algorithms* **2025**, *18*, 358. <https://doi.org/10.3390/a18060358>.
10. Tan, P.; Wang, Z.; Chang, X. Road Obstacle Detection Method Based on Improved YOLOv5. *Algorithms* **2025**, *18*, 300. <https://doi.org/10.3390/a18060300>.
11. Tong, K.; Wu, Y.; Zhou, F. Recent advances in small object detection based on deep learning: A review. *Image Vis. Comput.* **2020**, *97*, 103910.
12. Liu, Y.; Sun, P.; Wergeles, N.; Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert Syst. Appl.* **2021**, *172*, 114602.
13. Zhen, Q.; Wu, L.; Liu, G. An Oracle Bone Inscriptions Detection Algorithm Based on Improved YOLOv8. *Algorithms* **2024**, *17*, 174. <https://doi.org/10.3390/a17050174>.
14. Chen, G.; Wang, H.; Chen, K.; Li, Z.; Song, Z.; Liu, Y.; Chen, W.; Knoll, A. A Survey of the Four Pillars for Small Object Detection: Multiscale Representation, Contextual Information, Super-Resolution, and Region Proposal. *IEEE Trans. Syst. Man Cybern. Syst.* **2022**, *52*, 936–953. <https://doi.org/10.1109/TSMC.2020.3005231>.
15. Liang, X.; Zhang, J.; Zhuo, L.; Li, Y.; Tian, Q. Small Object Detection in Unmanned Aerial Vehicle Images Using Feature Fusion and Scaling-Based Single Shot Detector With Spatial Context Analysis. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 1758–1770. <https://doi.org/10.1109/TCSVT.2019.2905881>.
16. Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; Han, J. Towards Large-Scale Small Object Detection: Survey and Benchmarks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13467–13488. <https://doi.org/10.1109/TPAMI.2023.3290594>.
17. Zhong, Y.; Wang, J.; Peng, J.; Zhang, L. Anchor Box Optimization for Object Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 2–5 March 2020; pp. 1286–1294.
18. Ding, X.; Zhang, R.; Liu, Q.; Yang, Y. Real-time small object detection using adaptive weighted fusion of efficient positional features. *Pattern Recognit.* **2025**, *167*, 111717. <https://doi.org/10.1016/j.patcog.2025.111717>.
19. Li, Z.; Yuan, J.; Zhang, Z. SolarIR-DETR: A lightweight transformer for UAV-assisted thermal hotspot detection in solar photovoltaic panels. *Sci. Rep.* **2026**, *66*, 17172. <https://doi.org/10.1038/s41598-026-50451-z>.
20. Deng, C.; Wang, M.; Liu, L.; Liu, Y.; Jiang, Y. Extended Feature Pyramid Network for Small Object Detection. *IEEE Trans. Multimed.* **2022**, *24*, 1968–1979. <https://doi.org/10.1109/TMM.2021.3074273>.
21. Du, Z.; Hu, Z.; Zhao, G.; Jin, Y.; Ma, H. Cross-Layer Feature Pyramid Transformer for Small Object Detection in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5625714. <https://doi.org/10.1109/TGRS.2025.3572706>.
22. Shi, Z.; Hu, J.; Ren, J.; Ye, H.; Yuan, X.; Ouyang, Y.; He, J.; Ji, B.; Guo, J. HS-FPN: High Frequency and Spatial Perception FPN for Tiny Object Detection. In Proceedings of the AAAI Conference on Artificial Intelligence; Wulsh, T., Shah, J., Koiter, Z., Eds.; PKP Publishing Services: Philadelphia, PA, USA, 2025; Volume 39, pp. 6896–6904. <https://doi.org/10.1609/aaai.v39i7.32740>.
23. Tong, K.; Wu, Y. Deep learning-based detection from the perspective of small or tiny objects: A survey. *Image Vis. Comput.* **2022**, *123*, 104471.
24. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic convolution: Attention over convolution kernels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11030–11039.
25. Liu, C.; Zhao, Y.; Wang, J. Dynamic Kernel Network for Adaptive Feature Extraction in Object Detection. *Pattern Recognit.* **2026**, *177*, 111820. <https://doi.org/10.1016/j.patcog.2026.111820>.
26. Sun, H.; Wang, R.; Li, Y.; Yang, L.; Lin, S.; Cao, X.; Zhang, B. SET: Spectral Enhancement for Tiny Object Detection. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 4713–4723. <https://doi.org/10.1109/CVPR52734.2025.00444>.
27. Peng, S.; Xue, R.; Tong, Y.; Wang, Z.; Yang, H. Multispectral Object Detection via Edge-Enhanced and Frequency-Aware Fusion Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 5078–5090. <https://doi.org/10.1109/JSTARS.2025.3648007>.
28. Wu, Y.H.; Liu, Y.; Zhang, L.; Cheng, M.M.; Ren, B. EDN: Salient Object Detection via Extremely-Downsampled Network. *IEEE Trans. Image Process.* **2022**, *31*, 3125–3136. <https://doi.org/10.1109/TIP.2022.3164550>.
29. Xu, G.; Liao, W.; Zhang, X.; Li, C.; He, X.; Wu, X. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recognit.* **2023**, *143*, 109819.
30. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.
31. Zhu, Z.; Liang, D.; Zhang, S.; Huang, X.; Li, B.; Hu, S. Traffic-Sign Detection and Classification in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2110–2119.
32. Wu, Y.; Li, Z.; Chen, Y.; Nai, K.; Yuan, J. Real-time traffic sign detection and classification towards real traffic scene. *Multimed. Tools Appl.* **2020**, *79*, 18201–18219.

33. Bosquet, B.; Cores, D.; Seidenari, L.; Brea, V.M.; Mucientes, M.; Bimbo, A.D. A full data augmentation pipeline for small object detection based on generative adversarial networks. *Pattern Recognit.* **2023**, *133*, 108998. <https://doi.org/10.1016/j.patcog.2022.108998>.
34. Wang, J.; Chen, Y.; Ji, X.; Dong, Z.; Gao, M.; Lai, C.S. Vehicle-Mounted Adaptive Traffic Sign Detector for Small-Sized Signs in Multiple Working Conditions. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 710–724. <https://doi.org/10.1109/TITS.2023.3309644>.
35. Chen, J.; Jia, K.; Chen, W.; Lv, Z.; Zhang, R. A real-time and high-precision method for small traffic-signs recognition. *Neural Comput. Appl.* **2022**, *34*, 2233–2245.
36. Zhang, J.; Lv, Y.; Tao, J.; Huang, F.; Zhang, J. A Robust Real-Time Anchor-Free Traffic Sign Detector With One-Level Feature. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 1437–1451. <https://doi.org/10.1109/TETCI.2024.3349464>.
37. Li, Z.; Chen, H.; Biggio, B.; He, Y.; Cai, H.; Roli, F.; Xie, L. Toward Effective Traffic Sign Detection via Two-Stage Fusion Neural Networks. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 8283–8294. <https://doi.org/10.1109/TITS.2024.3373793>.
38. Zhou, Z.; Zhu, Y. KLDet: Detecting Tiny Objects in Remote Sensing Images via Kullback–Leibler Divergence. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. <https://doi.org/10.1109/TGRS.2024.3382099>.
39. Chen, L.; Liu, C.; Li, W.; Xu, Q.; Deng, H. DTSSNet: Dynamic Training Sample Selection Network for UAV Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. <https://doi.org/10.1109/TGRS.2023.3348555>.
40. Zhao, J.; Zhu, H. CBPH-Net: A Small Object Detector for Behavior Recognition in Classroom Scenarios. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–12. <https://doi.org/10.1109/TIM.2023.3296124>.
41. Huang, S.; Ren, S.; Wu, W.; Liu, Q. Discriminative features enhancement for low-altitude UAV object detection. *Pattern Recognit.* **2024**, *147*, 110041.
42. Li, Y.; Wang, Y.; Ma, Z.; Wang, X.; Tang, Y. Sod-Uav: Small Object Detection For Unmanned Aerial Vehicle Images Via Improved Yolov7. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 7610–7614. <https://doi.org/10.1109/ICASSP48485.2024.10448458>.
43. Fu, Q.; Liu, J.; Zhang, X.; Zhang, Y.; Ou, Y.; Jiao, R.; Li, C.; Mazzanti, G. A Small-Sized Defect Detection Method for Overhead Transmission Lines Based on Convolutional Neural Networks. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–12. <https://doi.org/10.1109/TIM.2023.3298424>.
44. Ran, Q.; Zhang, C.; Wei, W.; Zhang, L. Efficient and Accurate Giraffe-Det for UAV Image Based Object Detection. In Proceedings of the IGARSS 2023—2023 IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 16–21 July 2023; pp. 6190–6193. <https://doi.org/10.1109/IGARSS52108.2023.10282585>.
45. Ma, Y.; Chai, L.; Jin, L. Scale Decoupled Pyramid for Object Detection in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. <https://doi.org/10.1109/TGRS.2023.3298852>.
46. Chen, Y.; Fan, B.; Liu, N.; Yang, Y.; Tang, J. EEnvA-Mamba: Effective and environment-aware adaptive Mamba for road object detection in adverse weather scenes. *Pattern Recognit.* **2026**, *175*, 113127. <https://doi.org/10.1016/j.patcog.2026.113127>.
47. Fathony, R.; Sahu, A.K.; Willmott, D.; Kolter, J.Z. Multiplicative filter networks. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
48. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. *IEEE Trans. Cybern.* **2022**, *52*, 8574–8586. <https://doi.org/10.1109/TCYB.2021.3095305>.
49. Zhang, H.; Wang, Y.; Dayoub, F.; Sunderhauf, N. VarifocalNet: An IoU-Aware Dense Object Detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 8514–8523.
50. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014, Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Proceedings, Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
51. Cui, Y.; Han, Y.; Guo, D. TS-DETR: Multi-scale DETR for traffic sign detection and recognition. *Pattern Recognit. Lett.* **2025**, *190*, 147–152. <https://doi.org/10.1016/j.patrec.2025.01.027>.
52. Yang, C.; Lin, Z.; Lan, Z.; Chen, R.; Wei, L.; Liu, Y. Evolutionary channel pruning for real-time object detection. *Knowl.-Based Syst.* **2024**, *287*, 111432.
53. Zhu, Z.; Zheng, R.; Qi, G.; Li, S.; Li, Y.; Gao, X. Small Object Detection Method Based on Global Multi-level Perception and Dynamic Region Aggregation. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 10011–10022. <https://doi.org/10.1109/TCSVT.2024.3402097>.
54. Xiao, Y.; Xu, T.; Xin, Y.; Li, J. FBRT-YOLO: Faster and Better for Real-Time Aerial Image Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 8673–8681. <https://doi.org/10.1609/aaai.v39i8.32937>.
55. Du, B.; Huang, Y.; Chen, J.; Huang, D. Adaptive Sparse Convolutional Networks with Global Context Enhancement for Faster Object Detection on Drone Images. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 13435–13444. <https://doi.org/10.1109/CVPR52729.2023.01291>.

56. Kwon, S.; Lim, G.; Han, Y. SPAR-Det: Segmentation-guided and Prior-Aided Routing for Small Object Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 5–9 January 2026; pp. 2146–2155.
57. Yang, Y.; Wan, T.; Guo, L.; Zhang, M. Coordinate aware implicit neural representation for UAV small object detection. *PLoS ONE* **2026**, *21*, e0350990. <https://doi.org/10.1371/journal.pone.0350990>.
58. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. <https://doi.org/10.1016/j.isprsjprs.2019.11.023>.

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.