Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.086756

ARTICLE

FLD-RTDETR: A Frequency-Aware Lightweight Network for Small-Object Detection in Aerial Imagery

Yu Tong, Kaina Xiong, Jun Liu*, Xinyue Fan and Guixing Cao

College of Computer Science and Engineering, Northeastern University, Shenyang, China

*Corresponding Author: Jun Liu. Email: liujun@cse.neu.edu.cn

Received: 04 June 2026; Accepted: 24 August 2026; Published: 22 October 2026

ABSTRACT: Small-object detection in aerial remote-sensing imagery is constrained by the cascaded low-pass behaviour induced by consecutive spatial down-sampling, which progressively attenuates the high-frequency cues of weak targets. Recent detection transformers, whose self-attention operators behave—under commonly used temperature and rank regimes—as adaptive spatial low-pass filters, tend to inherit rather than mitigate this spectral bias, while existing remedies are typically obtained at the cost of considerable parameter and floating-point overhead. To bridge this gap, this paper proposes a Frequency-Aware Lightweight Network for Small-Object Detection (FLD-RTDETR), built upon the Real-Time Detection Transformer (RT-DETR), which augments this baseline along three orthogonal axes. A Partial Channel Bottleneck (PCB) routes a fraction of channels through identity bypass to preserve un-attenuated high-frequency channels while reducing redundant floating-point operations. A Local Spatial and Global Channel Attention (LSGCA) module factorizes self-attention into orthogonal spatial-covariance and channel-covariance branches, lowering complexity to local linearity. A Lightweight Multi-scale Fusion Neck (LMFN), centred on an Omni-Dimensional Frequency–Spatial Attention (ODFA) hub, injects explicit frequency-domain compensation at the multi-scale fusion junction. On the VisDrone2019 benchmark, FLD-RTDETR attains a mean Average Precision (mAP) of 50.2% (mAP_{50}) and 30.9% ($mAP_{50:95}$) at only 20.11M parameters and 86.4 giga floating-point operations (GFLOPs), improving over the RT-DETR-R18 baseline by 3.1 and 2.1 absolute points, respectively and reaching parity with RT-DETR-R50 at less than half its parameters and roughly two-thirds of its floating-point operations. Evaluation on the structurally different NWPU VHR-10 benchmark—on which the model is separately trained and tested—further indicates robustness across scale distributions and viewpoint geometries, providing a competitive accuracy/efficiency trade-off for high-fidelity small-object detection in aerial remote sensing.

KEYWORDS: Aerial imagery; small-object detection; frequency–spatial attention; lightweight detection

1 Introduction

Aerial imagery acquired by unmanned aerial vehicles (UAVs) and low-altitude platforms has emerged as a core data source for modern remote-sensing observation, supporting urban perception, traffic monitoring, infrastructure inspection, and emergency response [1]. Compared with near-nadir satellite views, low-altitude aerial imagery exhibits severe scale variance, dense spatial clustering, and pronounced illumination distortion, which together make weak-target detection markedly harder than on canonical natural-image benchmarks and push existing detectors toward their physical and computational limits.

The difficulty arises from two coupled physical factors. The first is per-target information scarcity: weak targets dominate aerial benchmarks, and because each instance occupies only a few pixels it carries very little

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

discriminative information, leaving a minimal margin for separation from the background. The second is the cascaded low-pass behaviour of repeated spatial down-sampling combined with non-ideal anti-aliasing in standard convolutional pipelines [2]: for an input of 640 × 640 pixels, five strided steps reduce the spatial extent to 20 × 20, so a target with an original footprint of roughly 20 × 20 pixels shrinks to sub-pixel size at the deepest layer, and its high-frequency geometric priors are progressively attenuated rather than merely under-learned [3]. The practical requirement that aerial detectors run within bounded parameter and floating-point budgets [4] further narrows the design space, ruling out indiscriminate capacity scaling as a remedy for this spectral attenuation.

Mainstream convolutional neural network (CNN) detectors have evolved along two divergent routes that each fall short under this regime. Single-stage architectures—including the YOLO family in its early [5–7], intermediate [8–11] and recent [12–15] generations, SSD [16], FCOS [17], RetinaNet [18], and YOLOX [19]—trade detection accuracy for throughput through dense prediction and decoupled heads, but their reliance on stacked depth and strided pooling to obtain global context directly compounds the cascaded low-pass attenuation that small targets cannot survive. Two-stage detectors such as Faster R-CNN [20] and Cascade R-CNN [21] mitigate missed detections through region proposals and feature resampling, yet their cascade inference and non-maximum suppression introduce latency that is difficult to reconcile with the real-time inference targets of aerial monitoring tasks.

Detection transformers were initially expected to dissolve this trade-off. Building on the vision-transformer paradigm [22], the DEtection TRansformer (DETR) [23] recast detection as bipartite-matching set prediction; Deformable DETR [24] broke the quadratic spatial complexity through learnable sparse sampling; and the Real-Time DETR (RT-DETR) [25] delivered the first real-time end-to-end transformer detector by decoupling intra-scale self-attention from cross-scale modulation. Subsequent variants such as OWRT-DETR [26] and RMT [27] further introduced multi-scale priors and retentive position embeddings.

A subtle but decisive observation, however, undercuts the assumption that transformers fully resolve the aerial small-target problem. Frequency-domain analyses of vision architectures [28–30] have shown that operators built on similarity-based weighted averaging—including soft-max self-attention under commonly used temperature and rank regimes—behave as adaptive spatial low-pass filters: their effective transfer functions concentrate energy near the DC component and attenuate high-frequency residuals. We do not claim that this equivalence is universal—in extreme low-temperature or high-rank regimes attention can locally amplify high-frequency residuals—but in the operating point of practical detectors the transformer tends to inherit, rather than counter-balance, the high-frequency attenuation of CNN backbones. When the two architectures are stacked, as in RT-DETR's hybrid encoder over a CNN backbone, their spectral biases compound, and the loss of high-frequency information for weak targets becomes more, not less, severe.

Two architecture-specific mismatches emerge when RT-DETR is transferred to aerial weak targets. First, its ResNet-style backbone [31] executes full-channel 3 × 3 convolutions in every bottleneck, despite the pronounced channel-wise collinear redundancy of deep aerial features; the resulting filtering wastes floating-point operations and reinforces edge smoothing on small targets. Second, the cross-channel feature modulation module of RT-DETR relies on a single 1 × 1 convolution for cross-channel mapping and offers no frequency-domain prior, leaving target topology—which concentrates in the high-frequency band—vulnerable to suppression by dominant low-frequency clutter.

The aim of this work is to design a frequency-aware lightweight detector that explicitly counteracts the cascaded spectral attenuation while keeping its parameter count comparable to the RT-DETR-R18 baseline and its floating-point cost well below the accuracy-matched RT-DETR-R50. To this end, we propose a Frequency-Aware Lightweight Network for Small-Object Detection (FLD-RTDETR), which augments RT-DETR along three orthogonal architectural axes. The principal contributions are as follows.

- A Partial Channel Bottleneck (PCB) is introduced to route a fraction of channels through identity bypass at every bottleneck. Building on the partial-channel computation paradigm of FasterNet [32], our contribution lies not in proposing partial convolution itself but in repositioning it as a spectral-preservation mechanism inside the RT-DETR backbone: the identity-bypass channels are interpreted, analyzed, and verified as carriers of un-attenuated high-frequency content that survives the cascaded down-sampling, while spatial floating-point operations are simultaneously reduced.
- A Local Spatial and Global Channel Attention (LSGCA) module is proposed to factorize self-attention into a windowed spatial-covariance branch and a global channel-covariance branch with an orthogonal-decoupling truncation at the interface. Relative to single-domain attention designs, such as window-based spatial attention [33] or channel-only attention [34], the contributions are twofold: a cross-layer attention-flow (CLAF) gate that propagates attention statistics across layers within each branch, and an explicit truncation rule that resets this flow at the spatial–channel interface to prevent inter-domain aliasing. The factorization converges attention complexity to local linearity.
- A Lightweight Multi-scale Fusion Neck (LMFN) is designed, centred on an Omni-Dimensional Frequency–Spatial Attention (ODFA) hub. ODFA performs two-dimensional fast Fourier transform-based spectral gating in parallel with an extra-large-kernel decomposed spatial branch. Compared with prior frequency-aware modules such as the Global Filter Network [35] and FcaNet [34], ODFA differs in two respects: (i) it is placed at the multi-scale fusion junction rather than at the backbone, where high- and low-frequency components from different scales must be re-aligned; and (ii) it adopts a CSP-style channel split so that only 25% of channels enter the spectral branch, keeping the parameter and floating-point cost of frequency processing bounded.
- Comprehensive experiments on the VisDrone2019 and NWPU VHR-10 datasets, including a complete 2^3 factorial module ablation, verify that the three mechanisms are individually beneficial and compose in a near-additive manner to yield consistent accuracy gains over the RT-DETR baseline at comparable parameter counts. Limitations, including the single-seed protocol and the absence of latency measurement beyond a single RTX 3090 GPU, are discussed explicitly in Section 5.

The remainder of this paper is organized as follows. Section 2 reviews related work on small-object detection in aerial imagery, detection transformers, and lightweight/frequency-aware feature learning. Section 3 details the proposed FLD-RTDETR architecture, including the PCB backbone, the LSGCA module, and the LMFN neck centred on the ODFA fusion hub. Section 4 reports comparative experiments against state-of-the-art detectors, module-level ablation studies, and second-benchmark evaluation on NWPU VHR-10. Section 5 discusses the empirical observations and identifies the physical boundaries of the current method. Section 6 concludes the paper and outlines future research directions.

2 Related Work

2.1 Small-Object Detection in Aerial Imagery

Small-object detection has emerged as a distinct subfield in recent years [1,36,37]. Three coupled difficulties characterize the regime. The first is representation-capacity depletion: because each target spans only a handful of pixels and thus carries little discriminative information, conventional convolutional features struggle to encode discriminative patterns. The second is anchor-assignment sparsity: standard intersection-over-union (IoU)-based label assignment becomes unstable when ground-truth boxes occupy only a handful of pixels, as discussed in tiny-object distribution-aware regression studies [38]. The third is feature-pyramid level mismatch: neither shallow high-resolution layers nor deep semantic layers alone are well-suited to such targets. Prior work has explored super-resolution-based detection for satellite imagery [39], degradation-reconstruction enhancement for tiny ship detection [40], Wasserstein-based regression for tiny boxes [38],

and high-resolution feature generation for small-ship detection in optical remote sensing imagery [41], but none of these directly addresses the spectral attenuation introduced by the architecture itself.

2.2 Detection Transformers

Building on the vision transformer paradigm [22], DETR [23] pioneered transformer-based detection as a set-prediction problem solved by bipartite matching. Deformable DETR [24] reduced the global quadratic complexity through learned sparse sampling. Swin Transformer [33] adopted shifted local windows for linear-time hierarchical features. RT-DETR [25] achieved real-time inference with an asymmetric hybrid encoder. Remote-sensing extensions such as OWRT-DETR [26] have transferred these architectures to small-object detection in optical UAV aerial imagery. Notwithstanding their global-context modelling capacity, all of these methods inherit the spectral bias of self-attention itself, leaving high-frequency small-target detail under-protected.

2.3 Lightweight Backbones and Multi-Scale Fusion

CSPNet [42] introduced cross-stage partial connections that split the feature input along the channel axis and apply non-linear transformation only to a subset, achieving roughly 20% floating-point operation reduction with improved gradient flow. FasterNet [32] further pushed partial-channel computation to the bottleneck level via the partial convolution (PConv) operator. ResNet [31] established the residual paradigm; the feature pyramid network (FPN) [3] popularized top-down lateral connections; and PANet [43] added bottom-up enhancement. A long-overlooked limitation of FPN/PAN is that fixed-kernel bilinear and nearest-neighbour up-sampling act as low-pass operators that over-smooth high-frequency detail and can suppress the response of micro-scale objects, together with the non-ideal anti-aliasing of strided convolutions [2]. Dynamic up-samplers offer point-wise learnable offsets; the proposed LMFN inherits this direction at the up-sampling nodes and combines it with the ODFA hub described below.

2.4 Frequency-Aware and Wavelet-Domain Feature Learning

Frequency-domain learning has recently emerged as a complementary direction for compensating the spectral bias of CNN/transformer pipelines. Xu et al. [28] learn in the JPEG DCT domain; Yao et al. [29] couple convolutional features with wavelet sub-bands; the Global Filter Network [35] replaces self-attention with a learnable Fourier-domain global filter; FcaNet [34] reformulates channel attention as discrete-cosine-spectrum selection; and wavelet-integrated CNNs explicitly preserve high-frequency sub-bands through lifting-based downsampling, a mechanism originally developed for noise-robust image classification [44]. Park and Kim [30] further showed that self-attention behaves predominantly as a low-pass operator under standard training regimes, motivating explicit high-frequency compensation. The proposed ODFA differs from these works in three respects. First, it is deployed at the multi-scale fusion junction rather than as a backbone-level replacement of attention, so it operates on already-aggregated multi-scale features whose spectra are most severely contaminated by cross-scale aliasing. Second, it couples a spectral-gating branch with an extra-large-kernel decomposed spatial branch (1×31, 31×1, 31×31, 1×1), so spatial macro-context and frequency-domain calibration are jointly optimized rather than sequentially stacked. Third, it adopts a CSP-style channel split with $e = 0.25$, so spectral processing is restricted to a small channel subset and the resulting parameter and floating-point overhead remain compatible with lightweight detectors. LSGCA, in turn, differs from Restormer-style channel-only attention and Swin-style window-only attention [33] in that the two branches are explicitly truncated at the interface to avoid mixing spatial-covariance and channel-covariance statistics across hierarchies.

3 Method

3.1 Overall Architecture

FLD-RTDETR retains the backbone–encoder–decoder structure of RT-DETR, in which the proposed LMFN neck replaces the cross-scale fusion of the original hybrid encoder. A stem of standard convolutions and max-pooling produces P_2 at 1/4 resolution. Four cascaded PCB stages then yield the multi-level set

$$\{P_2, P_3, P_4, P_5\}, \qquad \text{strides} = \{4, 8, 16, 32\}. \tag{1}$$

The highest-level feature P_5 is enhanced by the LSGCA module, and the LMFN neck explicitly incorporates the high-resolution shallow feature P_2 via a space-to-depth convolution (S2DConv), aligning multi-level features through dynamic up-sampling (DySample), phase-shift convolution (PSConv), and the ODFA fusion hub. The resulting pyramid $\{F_3, F_4, F_5\}$ is fed to the RT-DETR decoder; specific decoder hyperparameters are reported in Section 4. The overall architecture is illustrated in Fig. 1.

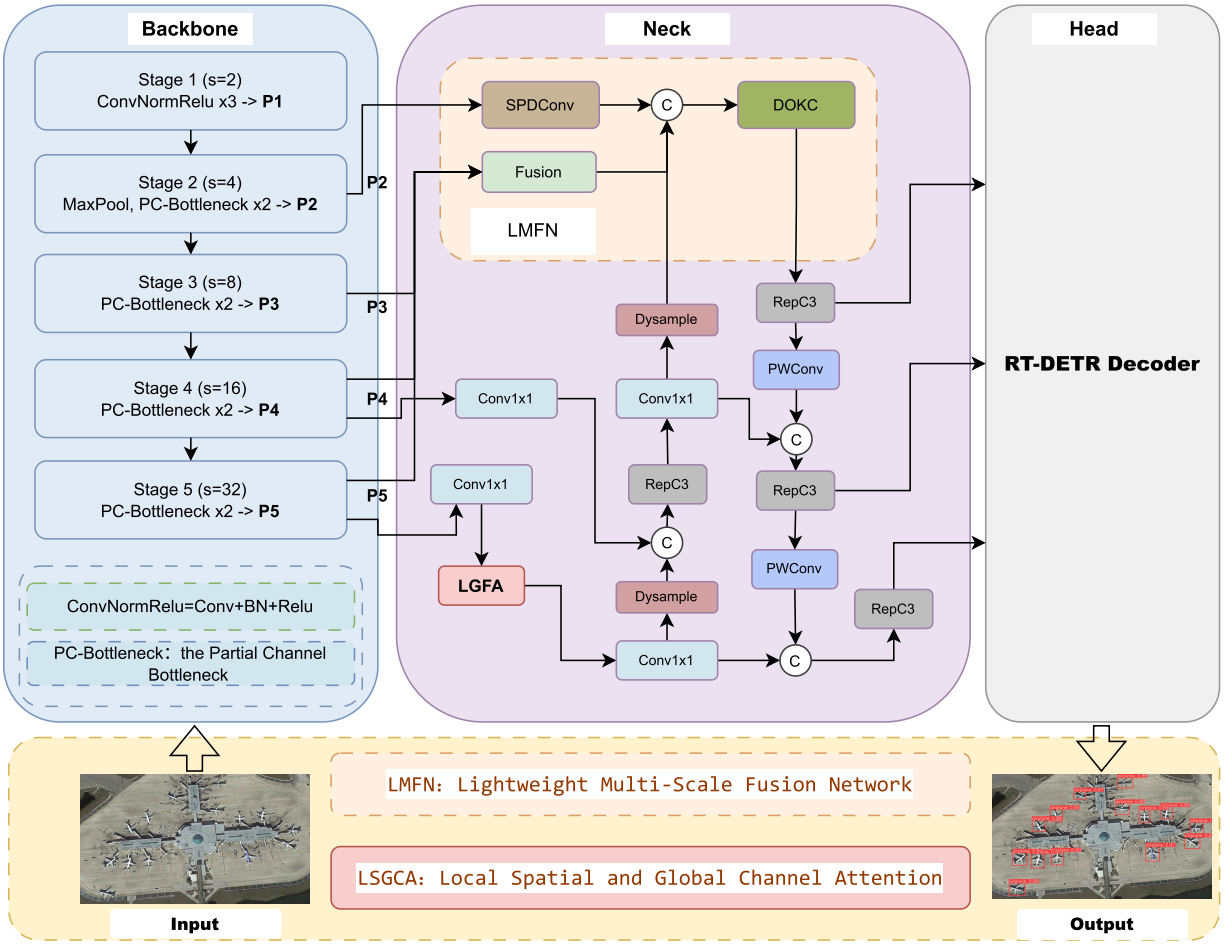


Figure 1: Overall architecture of the proposed FLD-RTDETR algorithm.

The data flow proceeds as follows. The backbone produces P_2 through three convolution blocks and max-pooling, and four PCB stages produce P_3, P_4 and P_5. On P_5, a 1×1 convolution projects to 256 channels and LSGCA performs long-range reasoning; another 1×1 convolution forms the lateral feature Y_5. In the top-down path, Y_5 is up-sampled via DySample and concatenated with the projected P_4, then refined through

a RepC3 block to yield Y_4; Y_4 is further up-sampled. In the multi-scale fusion, the shallow P_2 is processed by S2DConv while P_3, P_4 and the projected P_5 are fused via an S3F module. The outputs of S2DConv, S3F and the up-sampled feature are concatenated, processed by ODFA and a RepC3 block, and yield the high-resolution F_3. In the bottom-up path, F_3 is down-sampled via PSConv, fused with Y_4, and refined to F_4; similarly, F_4 is down-sampled and fused with Y_5 to yield F_5. Finally, $\{F_3, F_4, F_5\}$ are sent to the RT-DETR decoder.

For clarity, the auxiliary operators used above are summarized as follows. *S2DConv* (space-to-depth convolution) is the standard lossless down-sampling operator that rearranges a $H \times W \times C$ tensor into $(H/2) \times (W/2) \times 4C$ by pixel-shuffle and then applies a 1×1 convolution, so that high-frequency content of the shallow P_2 is repacked into channels rather than discarded by strided convolution. *S3F* (scale-sequence feature fusion) denotes a lightweight cross-scale aggregator that projects P_3, P_4 and the projected P_5 to a common channel width, aligns their spatial resolution with bilinear interpolation, and fuses them by element-wise summation followed by a 3×3 convolution; it is used here only as a sequence-context generator and contains no learnable spectral component. *DySample* is the dynamic up-sampler that predicts point-wise sampling offsets $\mathcal{O}$ from deep semantics and resamples through a grid sampler. *PSConv* (phase-shift convolution) implements stride-2 down-sampling by decomposing the input into four phase-shifted sub-grids, applying a shared 3×3 depthwise convolution to each, and recomposing them via a 1×1 pointwise convolution; its multi-phase mapping preserves the polygonal boundaries of small targets that would otherwise be erased by a single strided 3×3 kernel. *RepC3* is the re-parameterizable three-branch convolution block introduced by RT-DETR, fused into a single 3×3 convolution at inference time. None of these auxiliary operators carries the spectral-compensation role of ODFA, which is described in Section 3.4.

3.2 Partial Channel Bottleneck (PCB)

Deep features in aerial scenes exhibit substantial channel-wise collinear redundancy. Inspired by the partial-channel computation paradigm of FasterNet [32], we design the Partial Channel Bottleneck (PCB) that adapts this idea into the RT-DETR backbone via the partial channel 3×3 (PC3) operator (Fig. 2). Our contribution here is not the partial convolution operator itself, which is inherited from FasterNet, but its re-positioning inside a detection-transformer backbone as a spectral-preservation mechanism: the bypassed channels are retained specifically to keep an un-attenuated high-frequency band alive through the cascaded down-sampling, a role that FasterNet, designed purely for efficiency, does not consider. Given an input tensor $\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$, PC3 partitions $\mathbf{X}$ along the channel dimension into two disjoint subsets,

$$\mathbf{X} = [\mathbf{X}^{(c)}, \mathbf{X}^{(u)}],\ \mathbf{X}^{(c)} \in \mathbb{R}^{B \times C_c \times H \times W},\ \mathbf{X}^{(u)} \in \mathbb{R}^{B \times C_u \times H \times W}, \tag{2}$$

with $C_c = \lfloor C/n_{\text{div}} \rfloor$ and $C_u = C - C_c$. Only $\mathbf{X}^{(c)}$ is filtered:

$$\widetilde{\mathbf{X}}^{(c)} = \text{Conv}_{3\times3}(\mathbf{X}^{(c)}). \tag{3}$$

$\mathbf{X}^{(u)}$ passes through identity bypass; the final output is

$$\widetilde{\mathbf{X}} = \text{Concat}_{\text{ch}}(\widetilde{\mathbf{X}}^{(c)}, \mathbf{X}^{(u)}) \in \mathbb{R}^{B \times C \times H \times W}. \tag{4}$$

Because the 3×3 convolution is applied only to the $C_c = \lfloor C/n_{\text{div}} \rfloor$ selected channels (mapping $C_c \to C_c$), its spatial floating-point cost is approximately $1/n_{\text{div}}^2$ of a full $C \to C$ convolution; n_{div} is fixed to 4 to mirror the partial ratio recommended by FasterNet [32]. We interpret the identity-bypass path as a spectral-preservation channel: the bypassed subset $\mathbf{X}^{(u)}$ inherits the input's high-frequency residuals unchanged, so whatever high-frequency content it carries is forwarded intact rather than being re-filtered. We do not claim that the learned

3×3 filter on $\mathbf{X}^{(c)}$ is necessarily low-pass—a trained kernel may equally sharpen edges—but the identity bypass guarantees, by construction, that a fraction of the input high-frequency band survives every stage regardless of what the filtered branch learns.

Given input $\mathbf{X} \in \mathbb{R}^{B \times C_{\text{in}} \times H \times W}$ and target output channels C_{out}, the PCB block first applies a 1×1 projection $\mathbf{X}' = \text{Conv}_{1\times1}(\mathbf{X})$, then optionally down-samples

$$\mathbf{X}'' = \begin{cases} \mathbf{X}', & s = 1, \\ \text{MaxPool}_{2\times2}(\mathbf{X}'), & s = 2, \end{cases} \tag{5}$$

and the main path is $\mathbf{Z} = \sigma(\text{BN}(\text{PC3}(\mathbf{X}''))$ with output $\mathbf{Y} = \mathbf{X}_{\text{sc}} + \mathbf{Z}$. To keep the residual addition dimensionally consistent, the shortcut is $\mathbf{X}_{\text{sc}} = \mathbf{X}$ only when $C_{\text{in}} = C_{\text{out}}$ and $s = 1$; when the channel count changes it applies a 1×1 projection, and when $s = 2$ it additionally applies 2×2 average pooling before the projection, so that $\mathbf{X}_{\text{sc}}$ matches $\mathbf{Z}$ in both spatial resolution and channel dimension.

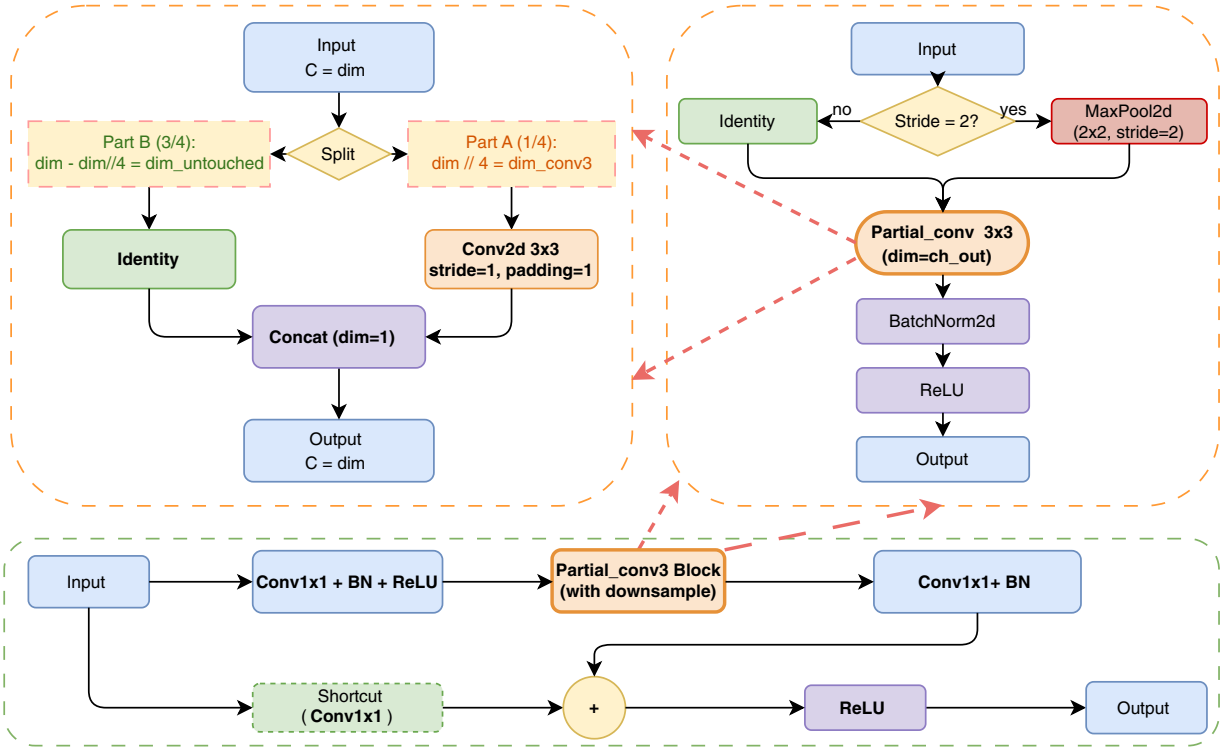


Figure 2: Architecture of the Partial Channel Bottleneck (PCB) module.

3.3 Local Spatial and Global Channel Attention (LSGCA)

LSGCA is deployed only at the deepest semantic level P_5 to limit overhead. To suppress cropping artefacts common in aerial imagery, the module uses reflection padding for size alignment (Fig. 3).

Each LSGCA block stacks N_1 Local Spatial Attention (LSA) layers followed by N_2 Global Channel Attention (GCA) layers; a single attention layer is computed as

$$\widehat{\mathbf{X}}^{(l)} = \mathbf{X}^{(l-1)} + \text{DropPath}\big(\text{Attn}(\text{GN}(\mathbf{X}^{(l-1)}), \mathbf{A}^{(l-1)})\big), \tag{6}$$

$$\mathbf{X}^{(l)} = \widehat{\mathbf{X}}^{(l)} + \text{DropPath}\big(\text{LeFFN}(\text{GN}(\widehat{\mathbf{X}}^{(l)}))\big), \tag{7}$$

where LeFFN inserts a 3 × 3 depthwise convolution after the up-projection to re-introduce a local inductive bias. A 5 × 5 depthwise convolution precedes the $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ projection,

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathrm{Split}(\mathrm{DWConv}_{5\times5}(\mathrm{Conv}_{1\times1}(\mathbf{X}))) \tag{8}$$

An L_2-normalized cosine attention with learnable temperature τ is adopted; the affinities are passed through a ReLU kernel rather than a row-wise soft-max, so the resulting attention map is not row-stochastic. Its magnitude is instead bounded implicitly by the L_2 normalization of $\mathbf{Q}$ and $\mathbf{K}$ and by the temperature τ, which together mitigate gradient vanishing and attention collapse without an explicit normalizing denominator.

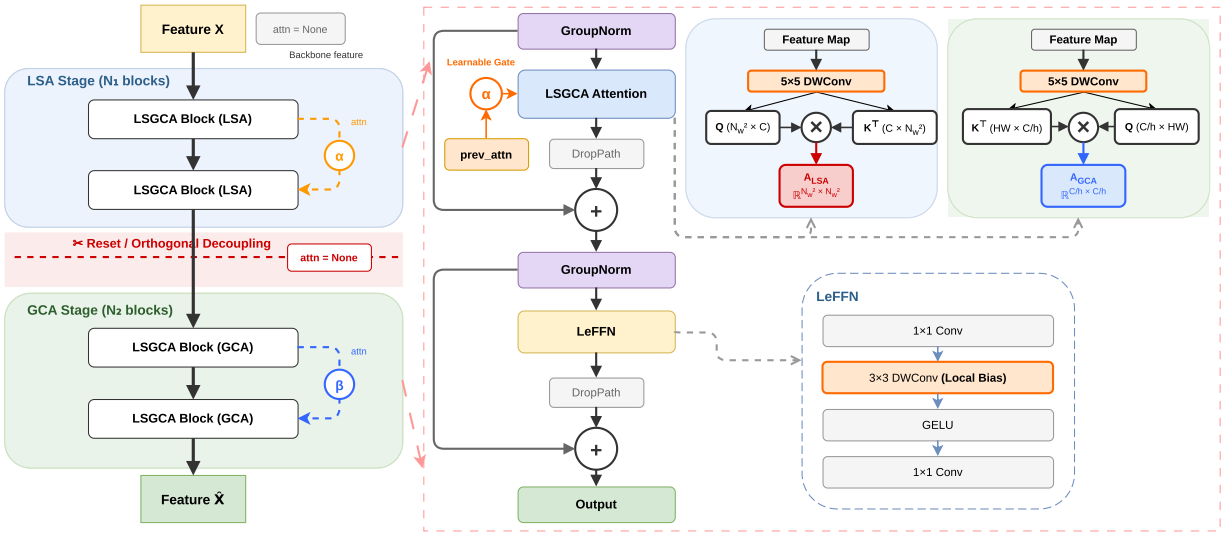


Figure 3: Block diagram of the Local Spatial and Global Channel Attention (LSGCA) module.

LSA partitions features into $w_s \times w_s$ windows (with $w_s = 4$) and includes a Cross-Layer Attention Flow (CLAF) gate, which adds a learnable-gated fraction α of the previous layer's attention map *before* the non-linearity, so that long-range affinities established in earlier layers are not re-estimated from scratch at every block:

$$\mathbf{A}_{\mathrm{LSA}}^{(l)} = \mathrm{ReLU}\left(\frac{\mathbf{Q}_w}{\|\mathbf{Q}_w\|_2}\left(\frac{\mathbf{K}_w}{\|\mathbf{K}_w\|_2}\right)^{\top}\tau + \alpha\,\mathbf{A}_{\mathrm{LSA}}^{(l-1)}\right). \tag{9}$$

GCA computes a cross-channel covariance matrix along the spatial dimension with complexity $\mathcal{O}(C^2HW)$, in the spirit of channel-only attention [34]:

$$\mathbf{A}_{\mathrm{GCA}}^{(l)} = \mathrm{ReLU}\left(\frac{\mathbf{Q}_c}{\|\mathbf{Q}_c\|_2}\left(\frac{\mathbf{K}_c}{\|\mathbf{K}_c\|_2}\right)^{\top}\tau + \beta\,\mathbf{A}_{\mathrm{GCA}}^{(l-1)}\right). \tag{10}$$

A key design choice concerns the LSA–GCA interface. The two attention maps reside in different ambient spaces: $\mathbf{A}_{\mathrm{LSA}}^{(l)} \in \mathbb{R}^{(w_s^2)\times(w_s^2)}$ encodes spatial-position affinities within a window, whereas $\mathbf{A}_{\mathrm{GCA}}^{(l)} \in \mathbb{R}^{(C/h)\times(C/h)}$, computed per head over $h = 8$ heads (i.e., 32 × 32 at $C = 256$), encodes inter-channel affinities. Concretely, an entry of $\mathbf{A}_{\mathrm{LSA}}$ has the physical unit "pixel/pixel", while an entry of $\mathbf{A}_{\mathrm{GCA}}$ has the unit "channel/channel"; the two affinity scores are not commensurable, and propagating one as a prior for the

other through the residual path would amount to mixing two unrelated statistical hierarchies. We therefore reset the cross-layer attention-flow state to zero when entering the GCA branch, $\mathbf{A}_{\mathrm{GCA}}^{(0)} = \mathbf{0}$. We refer to this rule as an *orthogonal-decoupling truncation*: it preserves the within-branch temporal smoothing introduced by CLAF while preventing inter-domain aliasing across the interface, and keeps the representation pure at each hierarchy. The same truncation also applies in reverse when an LSGCA block is repeated, so that each branch starts from its own zero state at the next instantiation.

3.4 Lightweight Multi-Scale Fusion Neck (LMFN)

LMFN reconstructs the FPN/PAN paradigm along three coordinated axes: dynamic semantic up-sampling, high-fidelity spatial down-sampling, and the ODFA fusion hub.

Dynamic semantic up-sampling. Static bilinear or nearest-neighbour up-sampling over-smooths and aliases micro-scale edges, attenuating the high-frequency detail on which small targets depend. Every up-sampling node in the top-down pathway of LMFN is therefore replaced with DySample, which predicts point-wise spatial offsets from deep semantics; the internal scale alignment inside the S3F sequence generator retains lightweight interpolation and is not a DySample node:

$$\mathcal{O} = \Delta\big(\mathrm{Conv}_{3\times3}(P_{i+1})\big), \quad P_i = \mathrm{GridSample}\big(P_{i+1}, \mathcal{G} + \mathcal{O}\big), \tag{11}$$

where $\Delta(\cdot)$ is a lightweight offset network and $\mathcal{G}$ is the standard up-sampling grid.

High-fidelity spatial down-sampling. Strided 3×3 convolution in the bottom-up pathway wipes out the micro-structural detail of small targets. It is replaced with the phase-shift convolution (PSConv):

$$F_{i+1} = \mathrm{PSConv}_{s=2}(F_i), \tag{12}$$

whose multi-scale phase mapping preserves polygonal boundaries during dimensionality reduction.

Omni-Dimensional Frequency–Spatial Attention (ODFA). ODFA is the core fusion hub at the highest-resolution junction. Its input concatenates three heterogeneous streams: the high-fidelity shallow P_2 processed by S2DConv, the dynamically up-sampled feature from the top-down pathway, and the cross-scale sequence context from P_3–P_5 extracted by an S3F module. Let the concatenated tensor be $X \in \mathbb{R}^{B\times C\times H\times W}$. A CSP-style split routes only 25% of channels through the core processing branch:

$$X_{\mathrm{ok}},\ X_{\mathrm{id}} = \mathrm{Split}\big(\mathrm{Conv}_{1\times1}(X)\big),\ e = 0.25. \tag{13}$$

The core branch performs macro-scale spatial calibration via a decomposed extra-large-kernel depth-wise convolution group, in which a horizontal stripe (1×31), a vertical stripe (31×1), and a full 31×31 kernel are combined with a depthwise 1×1 branch and a parameter-free identity term:

$$\begin{aligned} F_{\mathrm{spa}} = {} & X_{\mathrm{ok}} + \mathrm{DW}_{1\times31}(X_{\mathrm{ok}}) + \mathrm{DW}_{31\times1}(X_{\mathrm{ok}}) \\ & + \mathrm{DW}_{31\times31}(X_{\mathrm{ok}}) + \mathrm{DW}_{1\times1}(X_{\mathrm{ok}}). \end{aligned} \tag{14}$$

All four convolutional terms are depthwise (their group count equals the channel number). They model complementary spatial priors: the 1×31 and 31×1 stripes capture horizontal and vertical macro-context typical of aerial roads and runways; the 31×31 kernel supplies isotropic global context; the depthwise 1×1 term applies a per-channel learnable scaling and therefore does not mix information across channels; and the identity term anchors the bypass so that the sum behaves as a learnable spatial residual rather than a fully replacing transformation.

In parallel, a frequency-domain branch performs channel attention directly on the spectrum. A per-channel modulation vector $\mathbf{w} \in \mathbb{R}^{B\times C\times 1\times 1}$ is produced by global average pooling followed by a 1×1 convolution, multiplied onto the two-dimensional fast Fourier transform of X_{ok}, and the inverse transform is taken, whose modulus yields a real-valued response:

$$\mathbf{w} = \text{Conv}_{1\times 1}(\text{GAP}(X_{\text{ok}})), \qquad F_{\text{freq}} = \left|\mathcal{F}^{-1}\left(\mathbf{w} \odot \mathcal{F}(X_{\text{ok}})\right)\right|. \tag{15}$$

Since $\mathbf{w}$ is a real, per-channel scalar broadcast over the spatial frequencies, the product is well defined on the complex spectrum, and taking the modulus after the inverse transform resolves its complex-valued nature. This constitutes a frequency-domain channel attention in the spirit of FcaNet [34] and the Global Filter Network [35], but with a per-sample, content-dependent weight rather than a static filter. The frequency response is further refined by a lightweight spatial–channel gate and a Fourier gating module; the two branches are then combined and activated as $F_{\text{omni}} = \text{ReLU}(F_{\text{spa}} + F_{\text{freq}})$. The processed subset is projected and concatenated with the untouched identity bypass through 1×1 convolutions:

$$Y = \text{Conv}_{1\times 1}\left(\text{Concat}(\text{Conv}_{1\times 1}(F_{\text{omni}}),\, X_{\text{id}})\right). \tag{16}$$

3.5 Complexity Considerations

Relative to the RT-DETR baseline, the per-stage 3×3 spatial convolution cost in the backbone is reduced to approximately $1/n_{\text{div}}^2$ via PCB. LSGCA is placed only at the lowest-resolution semantic stage, capping its overhead. DySample is computationally comparable with bilinear interpolation; PSConv adds only a small parameter increment relative to a strided convolution; and ODFA executes the frequency–spatial joint processing on a 25% channel subset under a CSP architecture, which bounds the cost of its large-kernel and FFT operations. Section 4 reports the integrated parameter count and floating-point cost on the VisDrone protocol. To verify that the FFT and large-kernel branches do not incur a hidden memory cost that the theoretical giga floating-point operations (GFLOPs) would miss, we also measured peak GPU memory at a batch size of one and 640×640 input: FLD-RTDETR peaks at 2234 MiB, only 98 MiB (4.6%) above the RT-DETR-R18 baseline (2136 MiB) despite its added FFT and 31×31 large-kernel operations, and 96 MiB below the accuracy-parity RT-DETR-R50 (2330 MiB). We emphasize, however, that GFLOPs is an upper bound on theoretical arithmetic cost and is not a direct surrogate for wall-clock latency on hardware other than the RTX 3090 used in our experiments, since FFT and windowed-attention kernels are typically memory-bound; we therefore frame the efficiency claim of this paper at the algorithmic level (parameters, GFLOPs, and peak memory) and leave hardware-specific latency characterization out of scope.

4 Experiments and Results

4.1 Datasets

The proposed FLD-RTDETR is evaluated on two complementary remote-sensing detection benchmarks: VisDrone2019 (low-altitude UAV) and NWPU VHR-10 (high-altitude satellite). Representative scenes are shown in Figs. 4 and 5.

VisDrone2019 [45] contains 10,209 high-resolution UAV images (6471 training, 548 validation, 1610 testing, with 1580 unlabelled). It covers 10 object categories (pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus and motor). Approximately 70% of instances are smaller than 32×32 pixels, and the dataset features dense occlusion, severe scale variation, and complex urban background.

NWPU VHR-10 [46] is a high-resolution optical remote-sensing benchmark (0.5–2 m ground sampling distance, acquired by both airborne and spaceborne sensors) covering 10 target categories on 800 images (650 positive, 150 negative). Since the dataset does not ship an official train/validation split, we follow the

70/30 partition that is most widely adopted in the remote-sensing detection literature: the 650 positive images are randomly split into 455 training and 195 validation images with a fixed random seed of 0, and the same split is reused for every method so that comparisons share identical evaluation samples; the 150 negative (target-free) images are used by neither training nor evaluation, following common practice on this benchmark. Instances are very small (down to 20 × 20 pixels), densely distributed, and embedded in cluttered backgrounds.

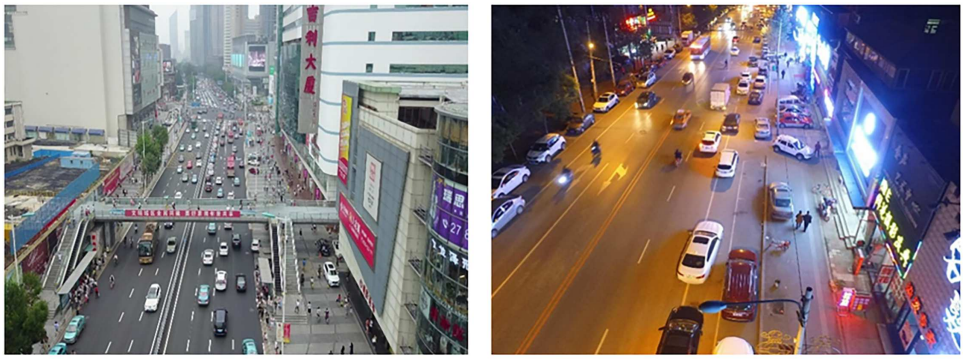


Figure 4: Typical scene samples from the VisDrone2019 dataset. (**a**) Dense urban traffic scene. (**b**) Crowded pedestrian scene.

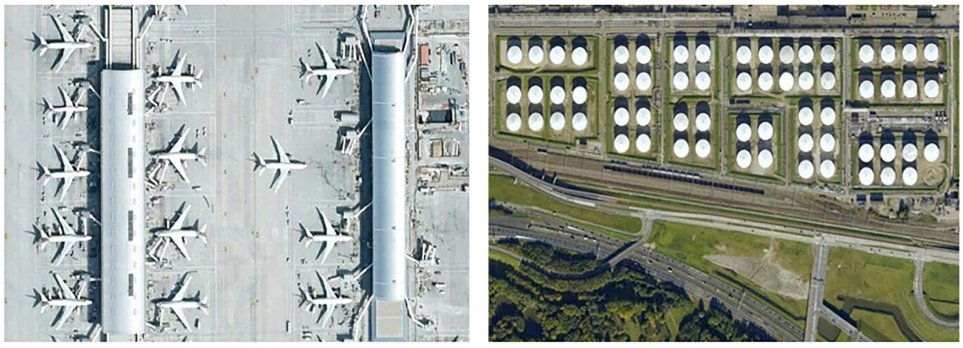


Figure 5: Typical scene samples from the NWPU VHR-10 dataset. (**a**) Densely arranged storage tanks. (**b**) Multi-scale aircraft and vehicles.

The scale distribution of both datasets is quantified at the input resolution of 640 × 640 pixels, following the MS-COCO scale convention: small (area < 32^2), medium (32^2 ≤ area < 96^2) and large (area ≥ 96^2). The instance counts, per-class scale ratios, and global distributions are presented in Figs. 6 and 7.

The VisDrone2019 distribution exhibits an extreme small-target dominance: among 457,066 annotated instances, small targets account for 85.7%, while large targets occupy less than 0.5%. The average pixel areas of the pedestrian and people categories are about 188 px^2 and 153 px^2, respectively; even the car category has nearly 80% small instances. NWPU VHR-10 instead displays a medium-dominated profile (74.3% medium, 15.5% small, 10.2% large among 3896 instances), but with significant inter-class scale variance: the ground-track-field category contributes most large targets, while the storage-tank and vehicle categories dominate the small-target population. These complementary distributions enable joint evaluation of frequency-aware small-target retention and multi-scale generalization.

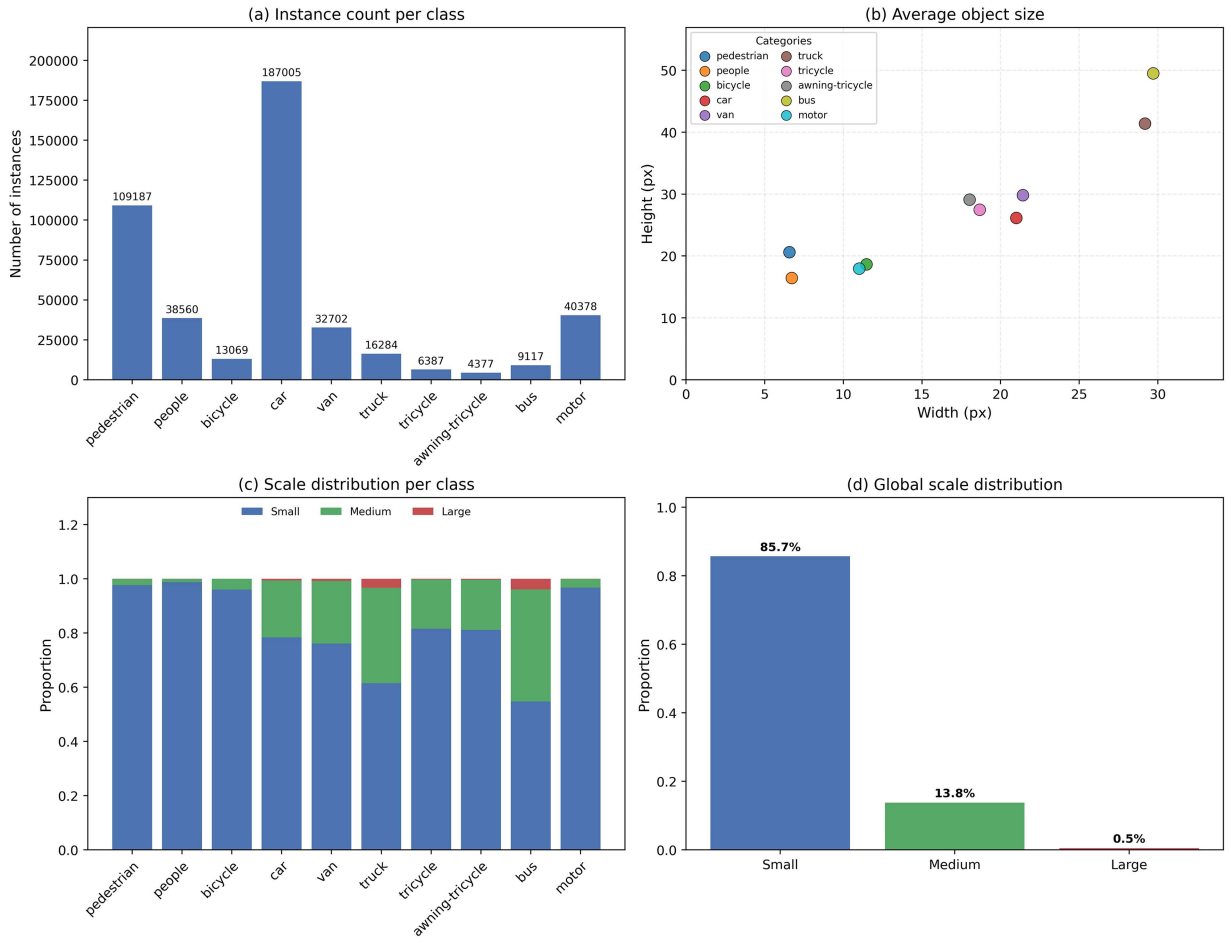


Figure 6: Statistical analysis of instance counts and multi-scale distributions in the VisDrone2019 dataset.

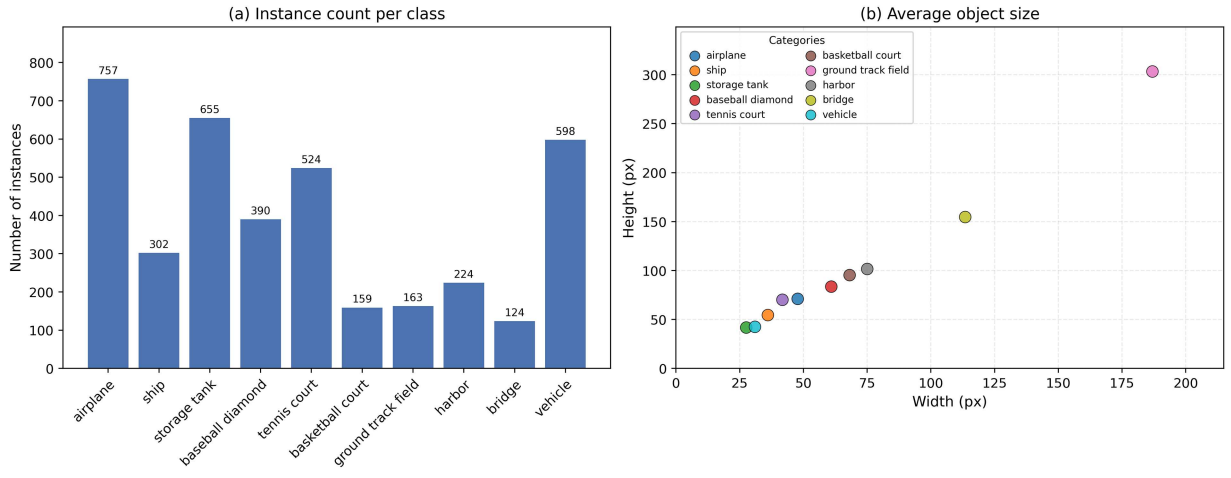


Figure 7: (Continued)

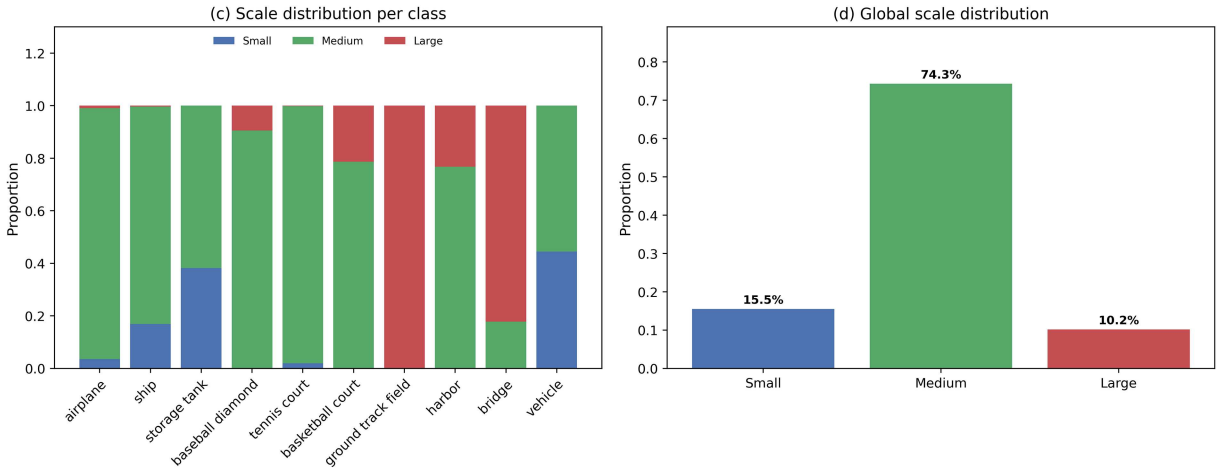


Figure 7: Statistical analysis of instance counts and multi-scale distributions in the NWPU VHR-10 dataset.

4.2 Experimental Environment

All experiments share the hardware and software platform listed in Table 1. Models are trained for 300 epochs with a batch size of 8 and an input resolution of 640 × 640 pixels. Data augmentation follows the default RT-DETR pipeline, namely random horizontal flipping, RandomResizedCrop, photometric distortion and the standard RT-DETR multi-scale jitter; no Mosaic or MixUp is used so that the spectral comparison against the baseline is not confounded by augmentation strategy. The AdamW optimizer is used with an initial learning rate of 1×10^{-4}, first- and second-moment coefficients $(\beta_1, \beta_2) = (0.9, 0.999)$, and weight decay of 1×10^{-4}; backbone parameters use a 0.1× learning-rate multiplier consistent with the RT-DETR official protocol. The PCB split factor is set to $n_{\text{div}} = 4$, mirroring the ratio recommended by FasterNet [32]. LSGCA stacks $N_1 = 2$ LSA layers (window size $w_s = 4$) and $N_2 = 2$ GCA layers; the temperature τ is initialized to $\log 10$. The ODFA channel-split ratio is fixed to $e = 0.25$, mirroring CSPNet's empirical sweet spot [42]; a full sensitivity sweep over n_{div}, e, w_s and (N_1, N_2) is left for future work due to compute budget. The detection head and bipartite-matching loss follow RT-DETR; the number of object queries is 300 and the decoder has four layers. For a strictly fair comparison, every compared detector in the comparison tables is re-trained by us from its official open-source implementation under the identical protocol described above (same 300 epochs, batch size, 640 × 640 resolution, optimizer, augmentation, and random seed); no number is copied from the original papers, so all rows share the same training and evaluation conditions. Unless otherwise noted, all reported numbers correspond to a single training run with the random seed fixed to 0.

Evaluation follows the MS-COCO protocol [47]. The IoU between a predicted box B_p and the ground-truth box B_{gt} is defined as

$$\text{IoU} = \frac{\text{Area}(B_p \cap B_{gt})}{\text{Area}(B_p \cup B_{gt})}. \tag{17}$$

With true positives (TP), false positives (FP) and false negatives (FN),

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad R = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{AP} = \int_0^1 P(R)\, \mathrm{d}R. \tag{18}$$

The reported metrics are mAP_{50} (IoU threshold 0.5), $mAP_{50:95}$ (IoU swept from 0.5 to 0.95 with step 0.05), the parameter count in millions (M), and the computational cost in giga floating-point operations (GFLOPs); per-class AP is also reported.

Table 1: Hardware and software configuration of the experimental platform.

Component	Version/Specification
Operating system	Linux x86_64 Ubuntu 22.04.2 LTS
Programming language	Python 3.10.15
Deep-learning framework	PyTorch 2.3.0
Parallel-computing framework	CUDA 12.1
CPU	AMD EPYC 9354 (32 cores/64 threads, 16 cores allocated)
GPU	NVIDIA GeForce RTX 3090 (24 GB)
RAM	128 GB

4.3 Comparison with Advanced Detectors

The proposed FLD-RTDETR is compared against a unified set of 12 representative detectors—nine members of the YOLO family (YOLOv3-tiny, YOLOv5m, YOLOv6s, YOLOv7-AF, YOLOv8m, YOLOv9m, YOLOv10m, YOLOv11m, YOLOv12m) and the full RT-DETR series (R18/R34/R50)—on the VisDrone2019 validation set. The same detector set is used across all quantitative tables and qualitative comparisons in this paper. The quantitative results are presented in Table 2.

Table 2: Quantitative comparison on the VisDrone2019 validation set. All per-class AP, **mAP_{50}** and **$mAP_{50:95}$** values are reported in percentage (%); Params are in millions (M) and computational cost is in GFLOPs. Best results in each column are in **bold**; when two methods are within 0.1 absolute point (i.e., 0.001 on the unit scale) on **mAP_{50}** or **$mAP_{50:95}$** we treat them as statistically tied and do not bold either, since the gap falls inside the typical single-seed run-to-run variance.

Method	Ped.	People	Bicycle	Car	Van	Truck	Tricycle	Awning	Bus	Motor	mAP_{50}	$mAP_{50:95}$	Params (M)	GFLOPs
YOLOv3-tiny [6]	19.3	14.2	4.0	60.7	29.4	23.4	16.5	6.8	39.3	22.0	23.6	13.1	12.13	18.9
YOLOv5m [7]	44.8	34.5	16.0	80.4	47.3	42.3	31.9	16.7	61.0	46.5	42.1	25.5	25.05	64.0
YOLOv6s [8]	39.7	31.8	10.0	78.8	43.2	35.0	24.6	14.0	54.1	41.2	37.2	22.3	16.30	44.0
YOLOv7-AF [9]	52.0	39.7	21.7	83.8	52.6	**48.7**	**39.0**	**21.2**	**69.8**	54.3	48.3	30.1	43.51	130.2
YOLOv8m [10]	47.6	36.4	15.7	81.6	47.0	38.1	31.3	17.3	60.3	48.3	42.4	25.8	25.84	78.7
YOLOv9m [11]	48.3	37.2	18.5	82.6	51.9	44.9	35.9	20.8	65.8	51.4	45.7	28.3	32.56	130.8
YOLOv10m [12]	45.2	35.5	17.1	81.2	47.1	40.8	32.2	16.2	61.5	48.1	42.5	26.1	15.31	58.9
YOLOv11m [13]	48.8	36.4	17.5	82.0	48.9	42.7	33.7	18.4	63.3	50.0	44.2	27.0	20.03	67.7
YOLOv12m [14]	45.6	34.4	16.7	78.9	48.2	40.6	31.7	18.9	62.0	48.0	42.5	25.9	20.11	67.2
RT-DETR-R18 [25]	56.6	49.1	20.6	85.4	49.6	38.1	33.5	16.8	60.9	60.6	47.1	28.8	19.88	57.0
RT-DETR-R34 [25]	56.2	49.5	22.2	85.5	49.0	40.0	36.9	19.2	61.8	61.5	48.2	29.5	31.12	88.8
RT-DETR-R50 [25]	**58.7**	**51.0**	**26.1**	85.6	50.5	42.5	37.4	20.5	68.9	61.5	**50.3**	30.8	41.97	129.6
FLD-RTDETR (Ours)	58.0	50.8	25.6	**86.4**	**52.9**	41.9	38.5	19.8	66.4	**61.7**	50.2	**30.9**	**20.11**	**86.4**

FLD-RTDETR strikes a competitive balance between accuracy and efficiency, reaching 50.2% mAP_{50} and 30.9% $mAP_{50:95}$. On mAP_{50}, the proposed method is statistically tied with the much larger RT-DETR-R50 (50.3%, gap 0.1 absolute point); on the more stringent $mAP_{50:95}$ metric, which directly reflects

bounding-box localization quality, FLD-RTDETR is on par with RT-DETR-R50 (30.9% vs. 30.8%, gap 0.1 absolute point) and clearly outperforms compactly sized contenders such as YOLOv12m (25.9%). We attribute the localization-side improvement to ODFA's frequency-domain calibration, which strengthens object boundaries that would otherwise be progressively attenuated by cascaded down-sampling, rather than to capacity scaling.

In terms of efficiency, FLD-RTDETR achieves accuracy parity with RT-DETR-R50 (mAP_{50} within 0.1 absolute point, $mAP_{50:95}$ within 0.1 absolute point) while reducing parameters by 52.1% (from 41.97M to 20.11M) and floating-point operations by 33.3% (from 129.6 to 86.4 G). Against the similarly sized YOLOv12m (20.11M parameters), FLD-RTDETR improves mAP_{50} by 7.7 absolute points, validating the importance of frequency-aware fusion at equal parameter budget. On dense aerial categories such as car, van, and motor, FLD-RTDETR records the best per-class AP at 86.4%, 52.9%, and 61.7%, respectively, indicating that the proposed mechanism is effective at separating mutually interfering small targets in clustered scenes. We caution that the proposed method's edge over RT-DETR-R50 on the overall metrics is within single-seed run-to-run variance, so the principal practical claim is parameter/FLOPs efficiency at equivalent accuracy rather than absolute accuracy superiority.

Qualitative comparisons on VisDrone2019 across multiple complex scenes are provided in Figs. 8 and 9. CNN-only baselines miss many densely parked vehicles and pedestrians under uneven illumination, while transformer-based methods (RT-DETR series and FLD-RTDETR) remain more stable. FLD-RTDETR matches RT-DETR-R50 in feature purification yet produces tighter bounding boxes on overlapping rigid vehicles, reducing redundant false positives.



(a) Ground truth

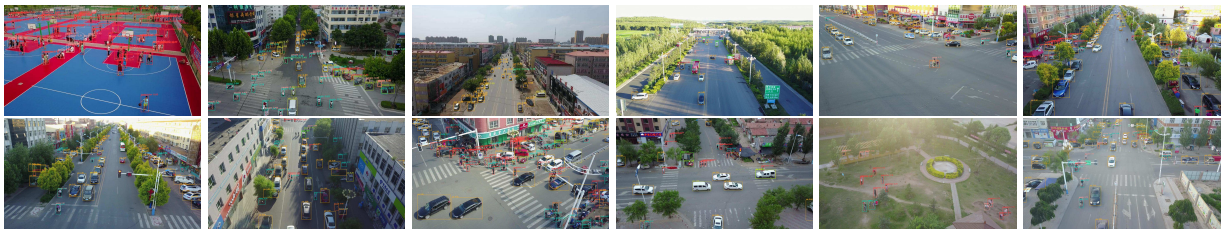


(b) YOLOv6s

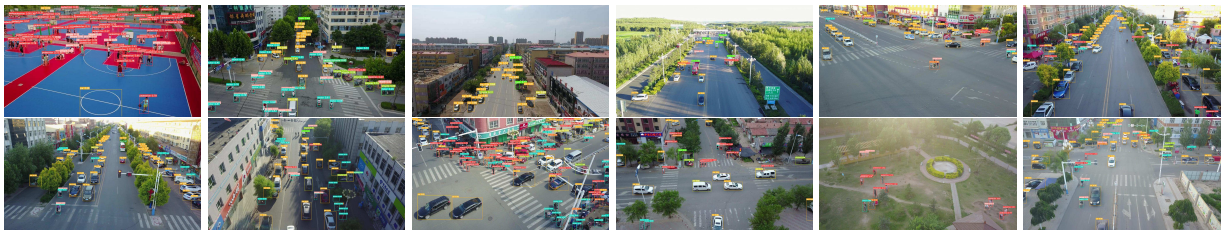


(c) YOLOv7-AF

Figure 8: (Continued)

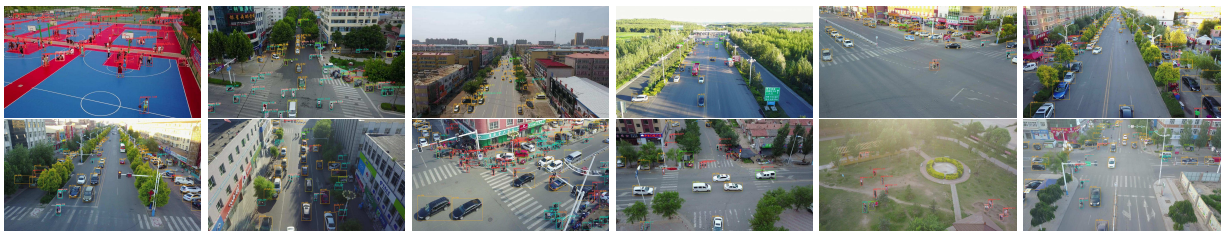


(d) YOLOv8m



(e) YOLOv9m

Figure 8: Qualitative comparison across multiple scenes on the VisDrone2019 validation set (Part I): (**a**) ground truth; (**b**) YOLOv6s; (**c**) YOLOv7-AF; (**d**) YOLOv8m; (**e**) YOLOv9m.



(a) YOLOv10m



(b) YOLOv11m



(c) YOLOv12m

Figure 9: (Continued)



(d) RT-DETR-R50



(e) FLD-RTDETR (Ours)

Figure 9: Qualitative comparison across multiple scenes on the VisDrone2019 validation set (Part II): (**a**) YOLOv10m; (**b**) YOLOv11m; (**c**) YOLOv12m; (**d**) RT-DETR-R50; (**e**) FLD-RTDETR (Ours). Our method demonstrates strong robustness against missed detections and false alarms in dense small-object scenarios.

4.4 Ablation Studies

Module-level ablation is conducted on VisDrone2019 with RT-DETR-R18 as the baseline. We report the complete $2^3 = 8$ factorial design—the baseline, the three single-module variants, all three two-module combinations, and the full model—so that the contribution of each module and their pairwise interactions can be read off directly. All variants are trained under the identical protocol and the same LSGCA configuration. The results are summarized in Table 3.

Table 3: Module-level ablation on the VisDrone2019 validation set. All per-class AP, **mAP$_{50}$** and **mAP$_{50:95}$** values are reported in percentage (%); Params are in millions (M) and computational cost is in GFLOPs. Module rows follow the architectural insertion order—backbone (PCB), multi-scale neck (LMFN), and deepest semantic stage (LSGCA)—which is the same order used in Fig. 10. Best per-column results are in **bold**; ties within 0.1 absolute point (i.e., 0.001 on the unit scale) on **mAP$_{50}$** or **mAP$_{50:95}$** are not bolded. A ✓ marks the modules that are enabled in each configuration.

Modules			AP by Class										Metrics			
PCB	LMFN	LSGCA	Ped.	People	Bicycle	Car	Van	Truck	Tricycle	Awning	Bus	Motor	mAP$_{50}$	mAP$_{50:95}$	Params	GFLOPs
			56.6	49.1	20.6	85.4	49.6	38.1	33.5	16.8	60.9	60.6	47.1	28.8	19.88M	57.0
✓			56.4	48.6	22.8	85.2	50.4	39.5	34.5	18.0	63.5	60.8	48.0	29.1	16.99M	53.0
	✓		58.0	50.4	25.4	85.9	49.8	40.6	36.0	18.7	66.3	**61.9**	49.3	30.2	21.57M	77.9
		✓	56.1	50.0	22.1	85.7	50.4	36.5	34.8	**20.4**	62.7	59.4	47.8	29.2	20.34M	57.7
✓	✓		**58.3**	49.5	**25.7**	86.1	50.4	41.8	36.1	16.9	65.6	61.5	49.2	30.1	19.76M	85.9
✓		✓	55.5	49.1	23.5	85.2	48.9	40.5	35.0	18.5	65.1	60.8	48.2	29.6	17.80M	54.0
	✓	✓	57.8	49.4	23.5	85.8	49.8	39.2	34.5	17.7	**67.8**	60.5	48.6	29.7	21.96M	78.3
✓	✓	✓	58.0	**50.8**	25.6	**86.4**	**52.9**	**41.9**	**38.5**	19.8	66.4	61.7	**50.2**	**30.9**	20.11M	86.4

Introducing PCB alone improves mAP$_{50}$ by 0.9 absolute points while reducing parameters by 2.89M and floating-point operations by 4.0 G, confirming that partial channel routing simultaneously reduces redundancy and preserves high-frequency content. Adding LMFN alone produces the largest single-module gain, +2.2 absolute points in mAP$_{50}$, driven primarily by ODFA's frequency-domain compensation at

the multi-scale junction. LSGCA alone contributes a +0.7 absolute-point gain at the cost of only 0.46M parameters, exhibiting the highest parameter efficiency among the three modules.

Across the full factorial grid, every single-module variant improves mAP_{50} over the baseline (PCB +0.9, LSGCA +0.7, LMFN +2.2), and all three two-module combinations also remain above the baseline (48.2, 49.2 and 48.6). The three-module model is the strongest overall, reaching 50.2% mAP_{50} and 30.9% $mAP_{50:95}$—absolute gains of 3.1 and 2.1 points over the baseline—with a parameter count essentially identical to the baseline (20.11M vs. 19.88M). We describe the composition as *near-additive* rather than strictly additive: LMFN is the strongest single module, and placing a second module on top of it changes mAP_{50} only marginally (e.g., LMFN 49.3 vs. PCB+LMFN 49.2 vs. LSGCA+LMFN 48.6), a fluctuation within single-seed run-to-run variance, and the full gain is realized only when all three modules act jointly. The full model also achieves the best AP on dense aerial categories such as car, van, and motor, consistent with the three mechanisms addressing the same underlying spectral attenuation at orthogonal architectural locations.

To trace the physical evolution of deep features, Fig. 10 presents GradCAM++ heatmaps across the ablation sequence on challenging aerial scenes. The heatmap columns correspond, in order, to (b) the RT-DETR-R18 baseline, (c) +PCB, (d) +LSGCA, (e) +LMFN, and (f) the full FLD-RTDETR. These columns visualize the baseline, each single-module variant, and the full model; they are a subset of the ablation configurations in Table 3 (the two-module combinations are omitted for visual clarity) and are intended only as a qualitative illustration rather than a one-to-one match to the table rows. The baseline's high-response regions are diffuse and partly absorbed by road textures and shadows, yielding missed detections and false positives. With PCB introduced, the ineffective activation area contracts; with LSGCA added, high-energy responses concentrate on semantic entities (cars, pedestrians); with LMFN added, distant micro-targets begin to activate. The full FLD-RTDETR produces the most precise, target-centred activations with strongly suppressed background noise.

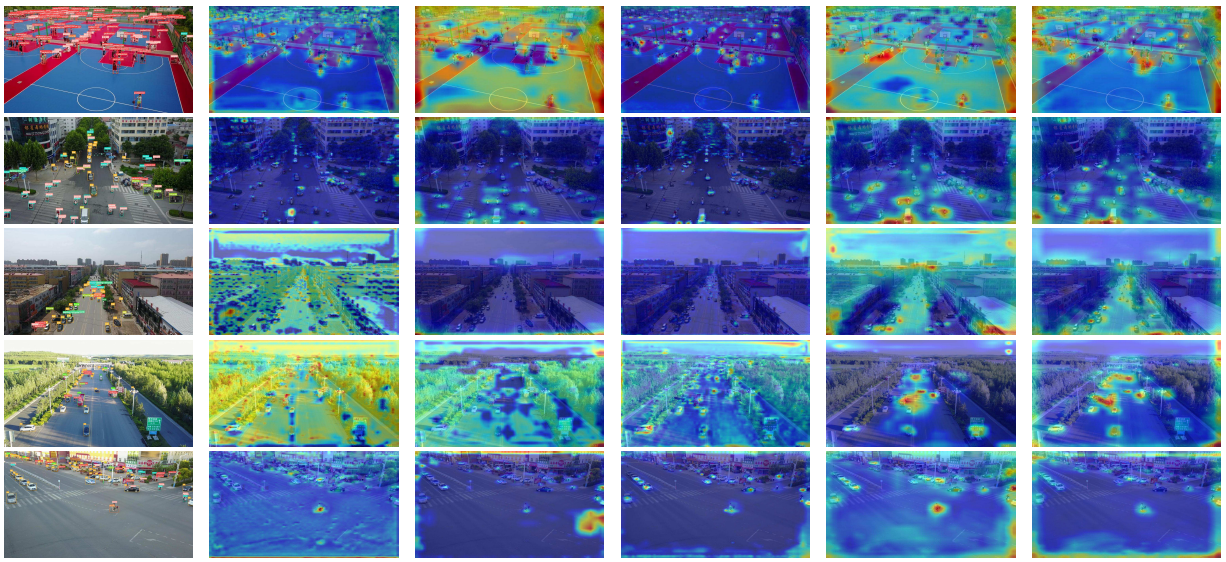


Figure 10: (Continued)

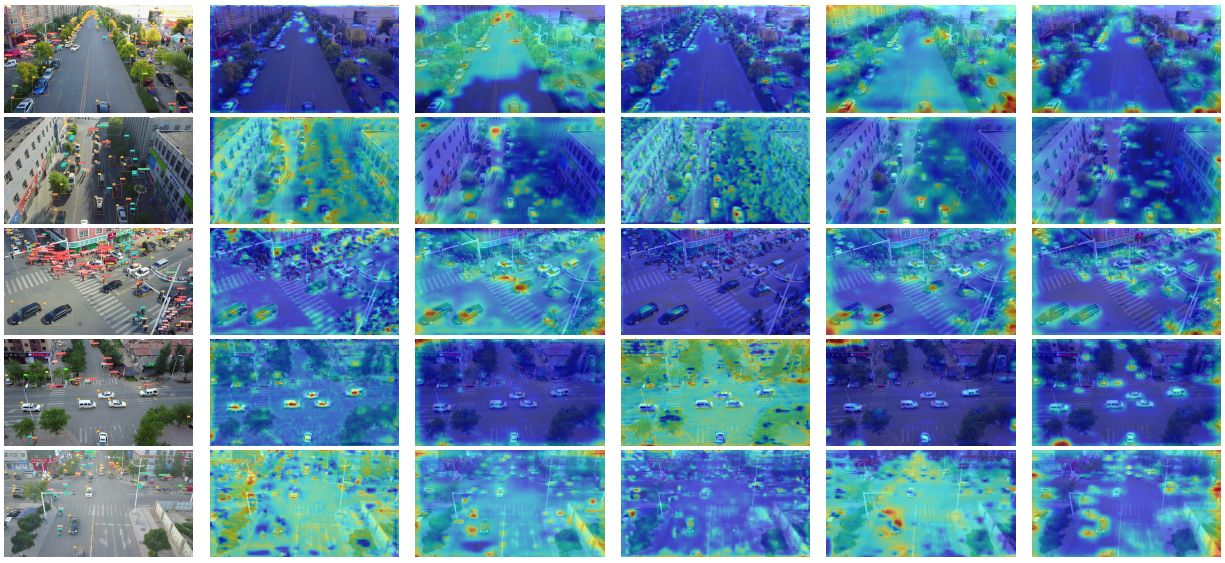


Figure 10: Feature-attention heatmap evolution across ablation stages on the VisDrone2019 validation set. Columns left to right: (**a**) ground-truth annotation; (**b**) baseline RT-DETR-R18; (**c**) +PCB; (**d**) +LSGCA; (**e**) +LMFN; (**f**) full FLD-RTDETR (Ours). High-response (red) regions gradually converge to ground-truth targets and background noise is progressively suppressed.

4.5 Spectral Validation of the Frequency-Aware Design

To provide spectral evidence for the design rationale of Section 3, we measured the high-frequency energy ratio—the fraction of two-dimensional-FFT spectral power that lies outside the central low-frequency window—of the feature maps entering and leaving each proposed module, averaged over 100 VisDrone2019 validation images (Fig. 11). The measurements are consistent with the motivating analysis. First, the LSGCA attention module lowers the high-frequency ratio (0.208 → 0.049), consistent with self-attention acting as a low-pass filter; at the deepest 1/32 semantic stage this attenuation is acceptable, since global context rather than fine detail dominates there, and the module still yields a net AP_S gain in the ablation of Section 4. Second, the PCB backbone slightly raises the high-frequency ratio (0.341 → 0.364), consistent with its identity-bypass channels forwarding un-attenuated high-frequency content. Third, the ODFA fusion hub raises the high-frequency ratio (0.092 → 0.216) at the fusion junction, consistent with the frequency-domain compensation for which it was designed.

4.6 Comparison against a Generic Frequency-Domain Fusion Module

A natural question is whether the benefit of the ODFA hub stems from its specific spatial–spectral design and its placement at the high-resolution fusion junction, or merely from performing some form of frequency-domain processing with additional capacity. To isolate this factor, we keep the entire network fixed and replace *only* the ODFA hub with a generic Fourier-based mixer—a C2f-style fast Fourier convolution mixer (FFCM) of substantially larger capacity—and re-train on NWPU VHR-10 under the identical protocol. Because a single frequency block introduces run-to-run variance, the generic mixer is evaluated over two random seeds. As Table 4 shows, ODFA attains a higher mAP_{50} (91.9%) than *either* seed of the generic mixer (91.0%–91.3%), i.e., above the seed-to-seed spread of the generic block, at a comparable $mAP_{50:95}$, while using only about one third of the fusion-hub parameters (1.14M vs. 3.56M). This indicates that the advantage of ODFA originates from its targeted design and fusion-junction placement rather than from generic frequency processing or added capacity.

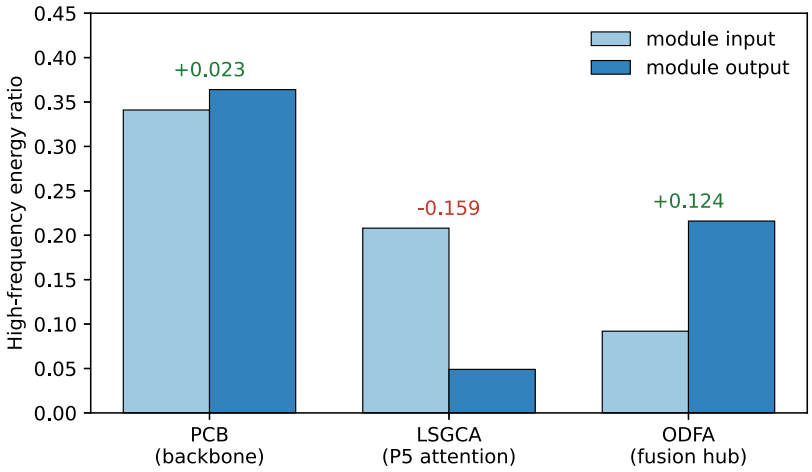


Figure 11: High-frequency energy ratio of the feature maps entering vs. leaving each proposed module (mean over 100 VisDrone2019 validation images). Self-attention (LSGCA) attenuates high frequency (−0.159), whereas PCB preserves (+0.023) and ODFA restores (+0.123) it, in line with the frequency-aware design.

Table 4: Effect of replacing the ODFA fusion hub with a generic Fourier mixer (FFCM) of larger capacity, on NWPU VHR-10 under the identical training protocol (only the fusion hub is changed). Best per column in **bold**. The generic mixer is reported for two random seeds. ODFA achieves the best mAP_{50} at roughly one third of the hub parameters.

Fusion Hub	mAP_{50} (%)	$mAP_{50:95}$ (%)	Hub Params (M)
Generic Fourier mixer (FFCM), seed 1	91.3	**58.7**	3.56
Generic Fourier mixer (FFCM), seed 2	91.0	58.0	3.56
ODFA (Ours)	**91.9**	58.4	**1.14**

4.7 Scale-Stratified Evaluation on Small Objects

Because the headline mAP aggregates over all object scales, it cannot by itself confirm that the reported gains originate from small objects—the regime this work targets. We therefore re-evaluate the full detector set on VisDrone2019 under the MS-COCO scale convention (small: area $< 32^2$; medium: 32^2–96^2; large: $>96^2$ pixels) using the pycocotools COCOeval interface with maxDets = 300. VisDrone is used for this analysis because its small-object-dominated distribution (85.7% small instances) makes the small-object metric statistically stable; NWPU VHR-10, by contrast, contains too few small instances for a reliable per-scale comparison, and its scale-stratified numbers are therefore omitted here (discussed in Section 5). Table 5 reports the scale-stratified AP and AR. The COCOeval AP runs a few points below the training-framework protocol used for the headline mAP on this dense benchmark, so these values are internally comparable but should not be cross-compared against Table 2.

Two observations stand out. First, FLD-RTDETR attains the highest small-object precision of all thirteen detectors (AP_S = 20.6%), ahead of RT-DETR-R50 (20.1%) at less than half its parameters and of every YOLO variant—directly confirming that the reported gains originate in the small-object regime. Second, on the medium/large and recall-oriented columns the leader is instead the much heavier YOLOv7-AF (43.5M parameters, 130.2 GFLOPs), which tops AP_M, AP_L and all three AR metrics; FLD-RTDETR concedes a modest amount of medium/large recall in exchange for a far smaller parameter budget while still leading on small-object precision. This is the intended behaviour of the frequency-aware design, which is tailored to the high-frequency cues of small targets rather than to large-object coverage.

Table 5: Scale-stratified evaluation on the VisDrone2019 validation set (%), under the MS-COCO area convention (small < 32^2, medium 32^2–96^2, large > 96^2 px), computed with pycocotools COCOeval (maxDets = 300). Best per column in **bold**. Values are comparable within this table only, not against the headline-mAP Table 2.

Method	AP_S	AP_M	AP_L	AR_S	AR_M	AR_L
YOLOv3-tiny [6]	4.0	20.1	36.6	10.8	37.1	48.3
YOLOv5m [7]	14.5	36.2	42.8	27.7	53.0	55.8
YOLOv6s [8]	11.0	32.6	42.2	23.2	50.0	54.2
YOLOv7-AF [9]	18.5	**43.0**	**52.6**	**33.5**	**59.2**	**70.9**
YOLOv8m [10]	14.9	35.7	44.5	28.9	51.9	56.6
YOLOv9m [11]	16.3	40.2	51.4	30.1	57.1	69.1
YOLOv10m [12]	14.7	36.7	43.8	28.8	53.2	55.8
YOLOv11m [13]	16.3	37.0	44.6	30.2	53.0	57.8
YOLOv12m [14]	15.8	37.6	45.1	29.1	54.0	58.7
RT-DETR-R18 [25]	18.7	36.5	39.4	28.2	48.0	51.9
RT-DETR-R34 [25]	18.8	38.2	43.8	28.5	50.2	53.5
RT-DETR-R50 [25]	20.1	39.5	41.3	29.9	50.3	54.7
FLD-RTDETR (Ours)	**20.6**	38.4	41.2	30.4	49.6	54.1

To verify that this small-object advantage is built up by the three modules jointly rather than by a single dominant component, Table 6 reports the scale-stratified module ablation on VisDrone, whose small-object-dominated distribution (85.7% small instances) makes AP_S statistically stable. Every single module lifts AP_S above the baseline (18.7 → 19.0/19.7/19.2 for PCB/LMFN/LSGCA), each two-module combination stays at or above 19.2, and the full model reaches the highest AP_S (20.6%) and AR_S (30.4%). This near-monotone progression on AP_S indicates that the three modules contribute complementarily to small-object retention. We report this scale-stratified ablation only on VisDrone: NWPU VHR-10 contains too few small instances (about 15.5%) for AP_S to be statistically reliable at this granularity.

Table 6: Scale-stratified module ablation on the VisDrone2019 validation set (%), computed with pycocotools COCOeval (maxDets = 300). VisDrone is small-object dominant, so AP_S is statistically stable here. Best per column in **bold**. A ✓ marks the modules that are enabled in each configuration.

Modules			Scale-Stratified Metrics					
PCB	LMFN	LSGCA	AP_S	AP_M	AP_L	AR_S	AR_M	AR_L
			18.7	36.5	39.4	28.2	48.0	51.9
✓			19.0	36.3	40.5	28.9	47.6	48.3
	✓		19.7	38.1	**43.8**	29.5	49.6	51.1
		✓	19.2	35.9	40.3	28.9	47.5	50.5
✓	✓		19.9	37.6	40.4	29.6	49.5	48.2
✓		✓	19.9	**38.8**	42.8	29.6	**50.1**	51.5
	✓	✓	19.2	37.4	39.6	29.1	49.1	51.8
✓	✓	✓	**20.6**	38.4	41.2	**30.4**	49.6	**54.1**

4.8 Generalization Analysis

To assess FLD-RTDETR under a structurally different application scenario and imaging viewpoint, the model is separately trained and evaluated on the NWPU VHR-10 high-resolution optical remote-sensing dataset. We stress that this is an evaluation on a second, independently trained benchmark rather than zero-shot cross-dataset transfer, since the two datasets share no object categories. Unlike the oblique low-altitude views of VisDrone, NWPU VHR-10 is acquired from a near-nadir satellite viewpoint, covers geographic targets such as airplanes, ships, and storage tanks, and exhibits a broader scale range and more complex background. The quantitative comparison is presented in Table 7.

Table 7: Comparison of generalization performance on the NWPU VHR-10 remote-sensing dataset. All per-class AP, **mAP$_{50}$** and **mAP$_{50:95}$** values are reported in percentage (%). Best results in each column are in **bold**; ties within 0.1 absolute point on **mAP$_{50}$** or **mAP$_{50:95}$** are not bolded.

Method	Airplane	Ship	Storage	Baseball	Tennis	Basketball	Track Field	Harbor	Bridge	Vehicle	mAP$_{50}$	mAP$_{50:95}$
YOLOv3-tiny [6]	99.5	81.3	95.1	**99.4**	92.9	80.3	99.3	96.9	89.0	70.1	90.4	55.9
YOLOv5m [7]	99.5	85.6	98.1	99.0	90.6	68.7	99.4	98.3	86.8	86.9	91.3	58.9
YOLOv6s [8]	99.5	84.8	97.1	99.1	87.9	60.0	99.4	98.4	88.4	79.0	89.3	57.4
YOLOv7-AF [9]	99.5	**91.0**	97.4	98.4	95.4	88.2	99.5	98.8	87.3	88.8	**94.4**	61.2
YOLOv8m [10]	99.5	85.4	97.9	99.2	91.2	71.3	99.2	98.6	86.7	85.0	91.4	60.0
YOLOv9m [11]	99.5	88.8	98.0	98.7	93.7	74.3	99.4	**99.2**	88.3	85.7	92.5	61.3
YOLOv10m [12]	99.3	80.9	97.1	98.1	90.9	72.8	97.3	96.2	65.1	79.8	87.7	56.9
YOLOv11m [13]	99.5	85.9	97.6	99.0	91.3	73.4	99.5	97.4	86.7	83.0	91.3	59.4
YOLOv12m [14]	99.5	85.6	98.3	97.5	90.9	69.8	99.4	94.6	78.0	85.2	89.9	57.4
RT-DETR-R18 [25]	99.5	83.1	98.0	97.9	92.0	67.3	99.5	97.4	87.2	81.7	90.3	58.6
RT-DETR-R34 [25]	99.4	87.0	98.1	98.0	90.9	80.5	99.5	97.3	87.2	78.9	91.7	58.6
RT-DETR-R50 [25]	99.5	89.0	**98.5**	98.4	**95.6**	83.1	99.5	98.6	86.1	**89.7**	93.8	**61.7**
FLD-RTDETR (Ours)	99.5	86.7	98.2	98.6	94.8	**88.9**	99.5	98.0	**89.2**	86.6	94.0	61.5

FLD-RTDETR achieves 94.0% mAP$_{50}$ and 61.5% mAP$_{50:95}$. Among the compared detectors it is second only to the much heavier YOLOv7-AF (94.4% mAP$_{50}$ at 43.5M parameters and 130.2 GFLOPs), and it is statistically tied with the 2.1× larger RT-DETR-R50 (93.8%/61.7%; gaps $\leq$ 0.2 absolute points, within typical single-seed run-to-run variance) at less than half the parameters. In other words, on this second benchmark FLD-RTDETR is the strongest detector in its efficiency class rather than the single best model overall. Per-class results indicate that FLD-RTDETR delivers the leading scores on basketball court (88.9%) and bridge (89.2%)—the two categories with the most complex background—while remaining competitive elsewhere. These results suggest that the proposed frequency–spatial fusion mechanism transfers across scale distributions and viewpoint geometries, rather than over-fitting to the VisDrone-specific oblique view; a more rigorous claim of generalization would require multi-seed averaging and is left for future work.

Fig. 12 compares the detection results on representative NWPU VHR-10 scenes. Lightweight detectors such as RT-DETR-R18 and YOLOv12m suffer from missed detections when ships or vehicles are tightly packed, and from false positives on rigid macro-structures such as bridges. RT-DETR-R50 partly remedies this through greater depth. FLD-RTDETR, leveraging frequency–spatial fusion and contextual awareness, simultaneously regresses tight bounding boxes for large entities and extracts micro-vehicle features under strong clutter at only ~20.1M parameters.

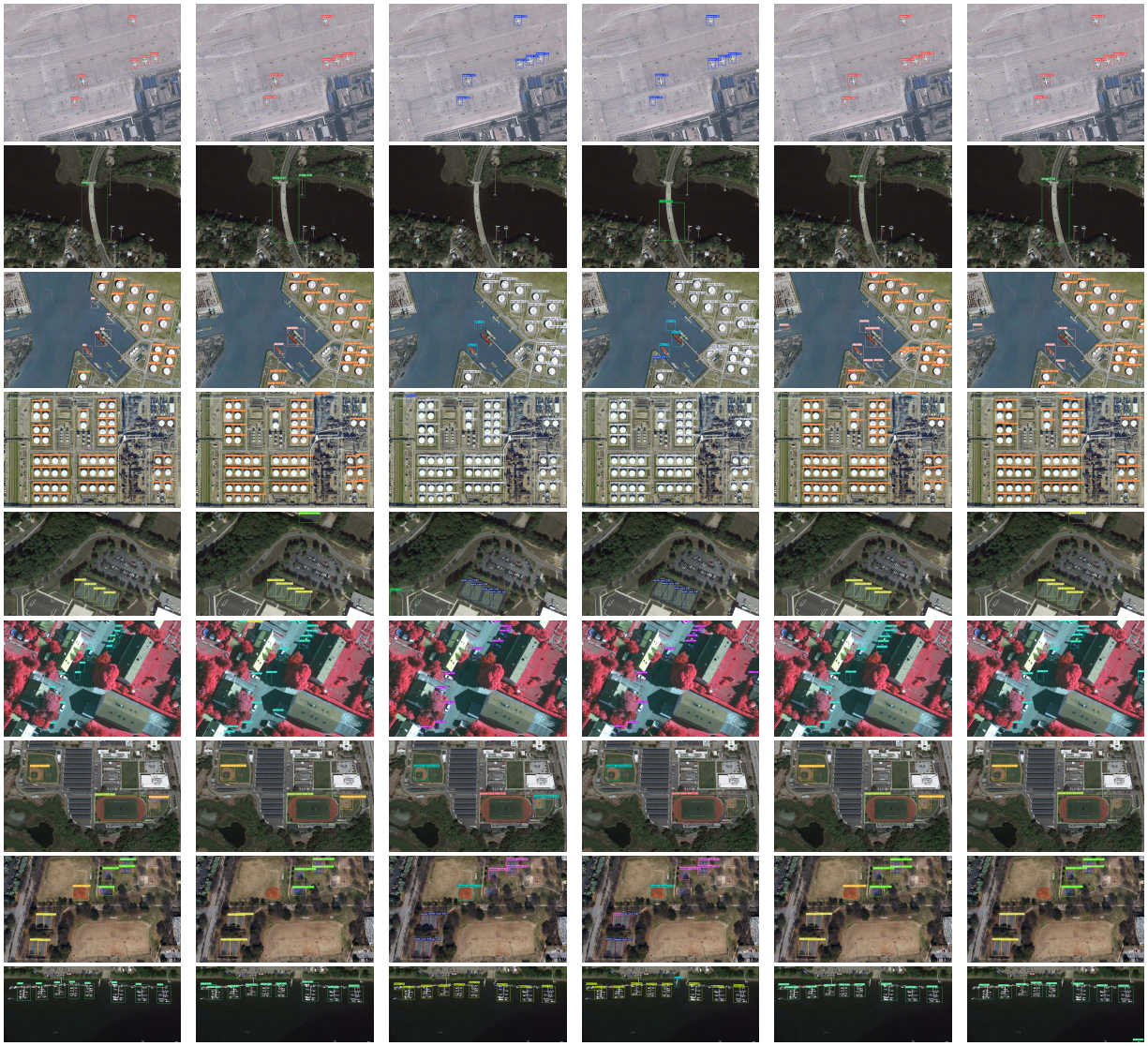


Figure 12: Qualitative comparison on the NWPU VHR-10 dataset. Columns left to right: (**a**) ground truth; (**b**) baseline RT-DETR-R18; (**c**) YOLOv8m; (**d**) YOLOv12m; (**e**) RT-DETR-R50; (**f**) FLD-RTDETR (Ours). FLD-RTDETR exhibits superior anti-interference capability under cluttered backgrounds.

5 Discussion

The experimental results suggest that FLD-RTDETR partially counteracts the spectral attenuation caused by consecutive spatial down-sampling, while leaving several open questions that we discuss honestly below. We organize the discussion into three observations, three methodological limitations, and five explicit directions for future work, to distinguish what the present paper has established from what remains open.

Observation 1: orthogonality and near-additive composition of the three modules. The three proposed modules act at architecturally orthogonal locations—PCB at the backbone, LMFN at the multi-scale fusion junction, and LSGCA at the deepest semantic stage. The complete $2^3 = 8$ factorial ablation (Table 3) shows that every single-module and two-module variant stays above the baseline and that the full model

is strongest, lifting mAP_{50} by 3.1 absolute points while keeping the parameter count essentially unchanged (20.11M vs. 19.88M). We characterize the composition as near-additive rather than strictly additive: because LMFN alone is already the dominant single module, some two-module combinations move mAP_{50} by less than a tenth of a point relative to LMFN alone—within single-seed run-to-run variance—and the full benefit emerges only when all three modules are combined. A multi-seed confirmation of these small interaction effects is left for future work.

Observation 2: parity with a 2.1× larger backbone, not absolute superiority. Compared with RT-DETR-R50, FLD-RTDETR reaches parameter/FLOPs parity at less than half the parameter count (−52.1%) and roughly two-thirds of the floating-point operations (−33.3%). On both mAP_{50} and $mAP_{50:95}$ the two methods are within 0.1–0.2 absolute points, which we explicitly treat as statistically tied. The principal practical claim is therefore *efficiency at equivalent accuracy*, not absolute accuracy superiority. We attribute the efficiency gain to ODFA's explicit frequency-domain calibration, which compensates the cumulative low-pass attenuation directly rather than through capacity scaling, and to PCB's spectral-preservation bypass that reduces the redundant filtering of channel-wise collinear features.

Observation 3: behaviour on the second, medium-dominated benchmark. On NWPU VHR-10 the proposed design remains competitive under a large viewpoint and scale-distribution shift, delivering the leading per-class scores on bridge and basketball court—categories where high-frequency target boundaries are easily confused with background texture. Two caveats must, however, be stated honestly. First, FLD-RTDETR is not the single best model on this benchmark: the much heavier YOLOv7-AF attains a higher overall mAP_{50} (94.4% vs. 94.0% at more than twice the parameters), so our edge here is one of efficiency rather than absolute accuracy. Second, and more important for the central thesis of this paper, the small-object advantage that is clear on VisDrone does not manifest on VHR-10: with only about 15.5% small instances, its scale-stratified small-object metric is neither stable nor decisive (indeed several YOLO variants score higher small-object AP there under this noise). NWPU VHR-10 should therefore be read as a medium/large-object generalization check rather than as additional evidence for small-object superiority.

Limitation 1: efficiency is reported only at the algorithmic level. All experiments are conducted on a single RTX 3090 GPU, and the efficiency advantage claimed in this paper is restricted to parameter count and theoretical GFLOPs. We do not measure wall-clock latency or throughput, and we do not extrapolate to hardware different from the RTX 3090 used here. In particular, the actual GPU latency of FFT-based ODFA and windowed-attention LSGCA may be dominated by memory-bandwidth and kernel-launch overhead rather than by raw FLOPs, so a fixed GFLOPs reduction is not guaranteed to translate into a proportional latency reduction. Hardware-specific latency profiling and the corresponding speed–accuracy trade-off analysis are therefore explicitly outside the scope of this paper.

Limitation 2: modality-specific design assumptions. PCB's identity bypass exploits channel-wise collinear redundancy and may not transfer cleanly to physically channel-independent modalities such as hyperspectral imagery, where each band carries near-independent radiometric semantics. ODFA's frequency gate rests on the empirical prior that targets occupy the high-frequency band and backgrounds the low-frequency band; in scenes dominated by high-frequency clutter such as regular urban grids, snow, or sand, the gate may amplify spurious responses. LSGCA's window size is fixed to $w_s = 4$ and is sensitive to extreme scale variance; it currently relies on the multi-scale neck for indirect compensation rather than on a scale-adaptive window mechanism.

Limitation 3: single-seed evaluation and incomplete hyperparameter sweeps. The reported numbers correspond to a single training run with random seed 0. Multi-seed averaging would be required to make stronger claims of statistical significance, particularly because the gap to RT-DETR-R50 on both datasets is within typical single-seed variance. Likewise, the key hyperparameters of the three modules—$n_{\mathrm{div}} = 4$ (PCB),

$e = 0.25$ (ODFA), $(N_1, N_2) = (2, 2)$ and $w_s = 4$ (LSGCA)—are fixed at values inherited from or adapted from the underlying paradigms (FasterNet, CSPNet, and windowed attention) without a dedicated sensitivity study. We discuss this transparency to make the reproducibility envelope of the present results explicit.

These limitations indicate the five directions for future work that we develop in the Conclusion.

6 Conclusions

This paper presented FLD-RTDETR, a frequency-aware lightweight detection transformer for small-object detection in aerial remote-sensing imagery. The contribution is not a single new operator, but a coordinated re-positioning of three known principles—partial-channel computation, dual-domain attention, and frequency-aware fusion—into three architecturally orthogonal locations of an RT-DETR detector: PCB at the backbone, LSGCA at the deepest semantic stage, and an LMFN neck centred on the ODFA spectral-spatial hub at the multi-scale fusion junction. On the VisDrone2019 benchmark, FLD-RTDETR attains 50.2% mAP_{50} and 30.9% $mAP_{50:95}$ at only 20.11M parameters and 86.4 GFLOPs; on the NWPU VHR-10 benchmark it attains 94.0% mAP_{50} and 61.5% $mAP_{50:95}$. On both benchmarks it reaches parity with detectors more than twice its size under a single-seed protocol. As discussed in Section 5, the principal practical claim is parameter/FLOPs efficiency at equivalent accuracy, not absolute accuracy superiority.

Building on the limitations explicitly enumerated in Section 5, future work will pursue five directions: (i) multi-seed averaging of the full factorial ablation, in order to confirm the small pairwise interaction effects at statistical significance; (ii) sensitivity studies over n_{div}, e, w_s and (N_1, N_2), together with learnable frequency-domain bases that adapt to scene statistics so that the high/low-frequency assignment is data-driven rather than empirical; (iii) a systematic wall-clock latency and quantization study on the same GPU class as well as a range of other hardware, to translate the algorithmic efficiency reported in this paper into hardware-level speed–accuracy trade-offs; (iv) a broader validation of the small-object advantage on additional scale-diverse benchmarks that contain a sufficient population of small instances (e.g., AI-TOD, DOTA), since NWPU VHR-10 proved too medium-dominated for a decisive small-object comparison; and (v) extension of the framework to multi-modal small-object detection in which complementary sensors (SAR, infrared, multispectral) supply the spectral content that a single optical band cannot recover.

Acknowledgement: None.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grant 61701100.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Yu Tong, Guixing Cao and Jun Liu; methodology, Yu Tong; software, Yu Tong and Kaina Xiong; validation, Yu Tong, Xinyue Fan and Kaina Xiong; formal analysis, Yu Tong; investigation, Yu Tong and Kaina Xiong; resources, Yu Tong and Kaina Xiong; data curation, Yu Tong, Xinyue Fan and Kaina Xiong; writing—original draft preparation, Yu Tong; writing—review and editing, Yu Tong, Guixing Cao and Jun Liu; visualization, Yu Tong, Xinyue Fan and Kaina Xiong; supervision, Guixing Cao and Jun Liu; project administration, Yu Tong, Guixing Cao and Jun Liu; funding acquisition, Jun Liu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Publicly available datasets were analyzed in this paper. The VisDrone2019 dataset is available at http://aiskyeye.com/ (accessed on 10 June 2025). The NWPU VHR-10 dataset is available at https://github.com/Gaoshuaikun/NWPU-VHR-10 (accessed on 10 June 2025).

Ethics Approval: Not applicable.

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

References

1. Xiao Y, Xu T, Yu X, Fang Y, Li J. A lightweight fusion strategy with enhanced interlayer feature correlation for small object detection. IEEE Trans Geosci Remote Sens. 2024;62:1–15. doi:10.1109/tgrs.2024.3457155.
2. Zhang R. Making convolutional networks shift-invariant again. Proc Mach Learn Res. 2019;97:7324–34. doi:10.48550/arxiv.1904.11486.
3. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 936–44.
4. Bahl G. Deep learning architectures for onboard satellite image analysis [dissertation]. Sophia Antipolis, France: Université Côte d'Azur; 2022.
5. 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); 2016 Jun 27–30; Las Vegas, NV, USA. p. 779–88.
6. Redmon J, Farhadi A. YOLOv3: an incremental improvement. arXiv:1804.02767. 2018.
7. Jocher G, Ultralytics Team. YOLOv5 by Ultralytics (v7.0). 2022 [cited 2026 Jan 1]. Available from: https://github.com/ultralytics/yolov5.
8. Li C, Li L, Jiang H, Weng K, Geng Y, Li L, et al. YOLOv6: a single-stage object detection framework for industrial applications. arXiv:2209.02976. 2022.
9. Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 18–23; Vancouver, BC, Canada. p. 7464–75.
10. Jocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8. 2023 [cited 2026 Jan 1]. Available from: https://github.com/ultralytics/ultralytics.
11. Wang CY, Yeh IH, Liao HYM. YOLOv9: learning what you want to learn using programmable gradient information. In: Proceedings of the European Conference on Computer Vision (ECCV); 2024 Sep 29–Oct 4; Milan, Italy. p. 1–21.
12. Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. YOLOv10: real-time end-to-end object detection. arXiv:2405.14458. 2024.
13. Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancements. arXiv:2410.17725. 2024.
14. Tian Y, Ye Q, Doermann D. YOLOv12: attention-centric real-time object detectors. arXiv:2502.12524. 2025.
15. Chen D, Zhang L. SL-YOLO: a stronger and lighter drone target detection model. arXiv:2411.11477. 2024.
16. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot MultiBox detector. In: Proceedings of the European Conference on Computer Vision (ECCV); 2016 Oct 11–14; Amsterdam, The Netherlands. p. 21–37.
17. Tian Z, Shen C, Chen H, He T. FCOS: a simple and strong anchor-free object detector. IEEE Trans Pattern Anal Mach Intell. 2022;44(4):1922–33.
18. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):318–27. doi:10.1109/tpami.2018.2858826.
19. Ge Z, Liu S, Wang F, Li Z, Sun J. YOLOX: exceeding yolo series in 2021. arXiv:2107.08430. 2021.
20. 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. 2017;39(6):1137–49. doi:10.1109/tpami.2016.2577031.
21. Cai Z, Vasconcelos N. Cascade R-CNN: high quality object detection and instance segmentation. IEEE Trans Pattern Anal Mach Intell. 2021;43(5):1483–98. doi:10.1109/tpami.2019.2956516.
22. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
23. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision (ECCV); 2020 Aug 23–28; Online. p. 213–29.

24. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: deformable transformers for end-to-end object detection. In: Proceedings of the International Conference on Learning Representations (ICLR); 2021 May 3–7; Virtual.
25. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. DETRs beat YOLOs on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 17–21; Seattle, WA, USA. p. 16965–74.
26. Ma S, Zhang Y, Peng L, Sun C, Ding B, Zhu Y. OWRT-DETR: a novel real-time transformer network for small-object detection in open-water search and rescue from UAV aerial imagery. IEEE Trans Geosci Remote Sens. 2025;63:1–13.
27. Fan Q, Huang H, Chen M, Liu H, He R. RMT: retentive networks meet vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 17–21; Seattle, WA, USA. p. 5641–51.
28. Xu K, Qin M, Sun F, Wang Y, Chen YK, Ren F. Learning in the frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 14–19; Seattle, WA, USA. p. 1740–9.
29. Yao T, Pan Y, Li Y, Ngo CW, Me T. Wave-ViT: unifying wavelet and transformers for visual representation learning. In: Proceedings of the European Conference on Computer Vision (ECCV); 2022 Oct 23–27; Tel Aviv, Israel. p. 328–45.
30. Park N, Kim S. How do vision transformers work? In: Proceedings of the International Conference on Learning Representations (ICLR); 2022 Apr 25–29; Virtual.
31. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8.
32. Chen J, Sh K, He H, Zhuo W, Wen S, Lee CH, et al. Run, don't walk: chasing higher FLOPS for faster neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 12021–31.
33. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 10012–22.
34. 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); 2021 Oct 10–17; Montreal, QC, Canada. p. 783–92.
35. Rao Y, Zhao W, Zhu Z, Lu J, Zhou J. Global filter networks for image classification. Adv Neural Inf Process Syst. 2021;34:980–93.
36. Dong R, Xu D, Zhao J, Jiao L, An J. Sig-NMS-based faster R-CNN combining transfer learning for small target detection in VHR optical remote sensing imagery. IEEE Trans Geosci Remote Sens. 2019;57(11):8534–45. doi:10.1109/tgrs.2019.2921396.
37. Yu L, Zhi X, Zhang S, Jiang S, Hu J, Zhang W, et al. A method for detecting aircraft small targets in remote sensing images by using CNNs fused with handcrafted features. IEEE Geosci Remote Sens Lett. 2024;21:1–5. doi:10.1109/lgrs.2024.3403548.
38. Wang J, Xu C, Yang W, Yu L. A normalized Gaussian Wasserstein distance for tiny object detection. ISPRS J Photogramm Remote Sens. 2022;187:340–53.
39. Shermeyer J, Van Etten A. The effects of super-resolution on object detection performance in satellite imagery. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2019 Jun 16–17; Long Beach, CA, USA. p. 1432–41.
40. Chen J, Chen K, Chen H, Zou Z, Shi Z. A degraded reconstruction enhancement-based method for tiny ship detection in remote sensing images with a new large-scale dataset. IEEE Trans Geosci Remote Sens. 2022;60:1–14. doi:10.1109/tgrs.2022.3180894.
41. Zhang H, Wen S, Wei Z, Chen Z. High-resolution feature generator for small-ship detection in optical remote sensing images. IEEE Trans Geosci Remote Sens. 2024;62(5):1–11. doi:10.1109/tgrs.2024.3377999.

42. Wang CY, Mark Liao HY, Wu YH, Chen PY, Hsieh JW, Yeh IH. CSPNet: a new backbone that can enhance learning capability of CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2020 Jun 14–19; Seattle, WA, USA. p. 390–1.
43. Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018 Jun 18–23; Salt Lake City, UT, USA. p. 8759–68.
44. Li Q, Shen L, Guo S, Lai Z. WaveCNet: wavelet integrated CNNs to suppress aliasing effect for noise-robust image classification. IEEE Trans Image Process. 2021;30:7074–89.
45. Zhu P, Wen L, Du D, Bian X, Fan H, Hu Q, et al. Detection and tracking meet drones challenge. IEEE Trans Pattern Anal Mach Intell. 2021;44(11):7380–99. doi:10.1109/tpami.2021.3119563.
46. Cheng G, Zhou P, Han J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Trans Geosci Remote Sens. 2016;54(12):7405–15. doi:10.1109/tgrs.2016.2601622.
47. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV); 2014 Sep 6–12; Zurich, Switzerland. p. 740–55.