

Article

A Wavelet-Guided Frequency–Spatial Decoupling Network for Visible–Infrared UAV Detection

Zeliang Dong ^{1,2,3}, Jiaxin Pan ^{1,2,3}, Xiangpeng Chen ^{1,2,3}, Wuxia Zhang ⁴  and Huinan Guo ^{1,3,*} 

¹ Xi'an Institute of Optics and Precision Mechanics (XIOPM), Chinese Academy of Sciences, Xi'an 710119, China; dongzeliang24@mailsucas.ac.cn (Z.D.); panjiaxin23@mailsucas.ac.cn (J.P.); chenxiangpeng23@mailsucas.ac.cn (X.C.)

² University of Chinese Academy of Sciences, Beijing 100049, China

³ Xi'an Key Laboratory of Aircraft Optical Imaging and Measurement Technology, Xi'an 710119, China

⁴ Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China; zhangwuxia@xupt.edu.cn

* Correspondence: guohuinan@opt.ac.cn

Highlights

What are the main findings?

- We present WFSD-Net, a novel frequency–spatial decoupling framework tailored for visible–infrared UAV detection. It combines wavelet-based spectral decomposition with lightweight axial contextual modeling to separate target signals from complex background clutter.
- WFSD-Net outperforms most state-of-the-art lightweight detectors on both Anti-UAV300 and MMFW-UAV datasets, achieving superior detection accuracy while maintaining a compact model size and low computational overhead.

What are the implications of the main finding?

- Our experiments confirm that frequency-decoupled visible–infrared fusion is effective for suppressing cross-modal noise and preserving complementary target cues, offering a practical direction for robust UAV detection under challenging conditions.
- The lightweight design of WFSD-Net enables efficient deployment on resource-constrained aerial platforms, providing a practical solution for real-time anti-UAV surveillance systems.

Abstract

Detecting unmanned aerial vehicles (UAVs) remains a difficult task, primarily due to their tiny size, rapid motion, and complex backgrounds. Fusing visible and infrared imagery offers complementary advantages for robust detection, yet existing methods rely on spatial feature aggregation that overlooks spectral disparities, coupling noise with textures. Moreover, the small scale and high dynamics of UAVs hinder standard convolution from decoupling target signals from background interference due to limited receptive fields. To solve these limitations, the Wavelet-guided Frequency–Spatial Decoupling Network (WFSD-Net) is designed for visible–infrared UAV detection. First, to tackle fusion noise, the Discrete Wavelet Band-Differentiated Fusion (DWBF) module is designed to explicitly decouple noise-dominant sub-bands from information-rich components by performing spectral decomposition. It aligns low-frequency distributions via adaptive spatial weighting and disentangles high-frequency details using physics-aware rules, achieving source-level noise suppression. Second, an Axial Strip Contextual Attention (ASCA) module is proposed. By utilizing anisotropic strip convolution via orthogonal decomposition, this module captures global contextual dependencies to effectively decouple weak target features from



Academic Editor: Carmine Serio

Received: 3 April 2026

Revised: 25 June 2026

Accepted: 26 June 2026

Published: 1 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 \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

background clutter, enhancing the spatial position encoding capability for weak targets. Finally, the proposed WFSD-Net method is validated on Anti-UAV300 and Multi-Sensor and Multi-View Fixed-Wing UAV (MMFW-UAV) datasets, and experiments demonstrate that the proposed method is superior to existing state-of-the-art (SOTA) methods.

Keywords: object detection; UAV detection; wavelet transform; visible–infrared image

1. Introduction

Object detection is a core task in computer vision, focusing on localizing and classifying instances in images and videos. Over the past decade, deep learning has brought about significant advances in object detection. These improvements have been demonstrated across diverse domains, including natural scene understanding, autonomous driving, and security surveillance. Within this context, detecting unmanned aerial vehicles (UAVs) has become a critical research hotspot. This task is characterized by unique challenges, such as the small spatial scale of targets, rapid and unpredictable motion patterns, and highly complex aerial backgrounds. Current detection schemes predominantly rely on single imaging modalities, which suffer from inherent limitations. Visible imagery provides texture and shape information but degrades under low-light conditions. Infrared imagery maintains stable detection in low light but lacks detail and is susceptible to thermal interference [1]. Given the inherent complementarity between these two modalities, visible and infrared fusion [2] has emerged as a reliable strategy for improving the robustness of UAV detection systems.

However, existing fusion methods face two major challenges. First, regarding fusion strategies, current methods fail to decouple spectral interference [3]. Existing approaches overwhelmingly focus on feature interaction within the spatial domain, including basic channel concatenation and element-wise addition [4]. These strategies overlook the intrinsic frequency-domain disparity in energy distributions [5] between visible and infrared modalities in the frequency domain. Second, regarding feature representation, CNN-based detectors face inherent limitations imposed by local receptive fields [6,7]. This limitation hinders the establishment of long-range geometric dependencies. As a result, weak targets in complex backgrounds are often insufficiently localized. Transformer-based architectures provide global modeling capability via self-attention operations. However, their high computational complexity makes real-time deployment difficult. Furthermore, their reliance on large-scale, high-quality paired data is difficult to satisfy in UAV detection, where annotated multi-modal samples are scarce [8,9].

To address the challenges of UAV detection, the Wavelet-guided Frequency–Spatial Decoupling Network (WFSD-Net) is proposed. The framework improves detection robustness through the integration of frequency-domain spectral decoupling and spatial-domain geometric modeling. Specifically, the Discrete Wavelet Band-Differentiated Fusion (DWBF) module performs spectral decomposition to suppress cross-modal noise interference, thereby improving fusion quality. In parallel, the Axial Strip Contextual Attention (ASCA) module employs anisotropic strip convolution with orthogonal decomposition to capture global contextual dependencies, effectively suppressing cross-modal noise interference [10,11].

The contributions of this article are as follows:

- We design a DWBF module, which solves the feature interference problem in spatial fusion through spectral decomposition and differentiated reconstruction strategies, achieving high-quality modal complementarity.

- We propose an ASCA module, which leverages anisotropic strip convolution to capture geometric dependencies and achieve spatial decoupling, significantly enhancing feature representation capabilities for dynamic small targets.
- We conduct extensive experiments on two public datasets, Anti-UAV300 and Multi-Sensor and Multi-View Fixed-Wing UAV (MMFW-UAV). Extensive experiments demonstrate that WFSD-Net consistently surpasses existing state-of-the-art (SOTA) methods in both accuracy and robustness.

2. Related Work

This section presents a structured review of prior work relevant to the proposed WFSD-Net, organized into four main themes. Section 2.1 introduces representative RGB–infrared benchmarks and UAV-oriented detection settings. Section 2.2 reviews frequency-domain analysis for infrared target detection. Section 2.3 summarizes representative fusion mechanisms for RGB–infrared object detection. Section 2.4 discusses context representation methods for small and weak UAV targets.

2.1. RGB–Infrared Benchmarks and UAV Detection Settings

The evolution of visible–infrared collaborative detection has been inextricably tied to the development of benchmark datasets from its very beginning. The Korea Advanced Institute of Science and Technology (KAIST) dataset introduced the pixel-aligned visible–thermal paradigm that has since become standard for multispectral pedestrian detection, providing a basis for fair cross-modal method comparison. The subsequent Low-Light Visible–Infrared Paired (LLVIP) benchmark [12,13] extended this research frontier to extreme low-illumination conditions, where visible light sensors suffer from catastrophic performance degradation. Not long after, the Multi-scenario Multi-modality Benchmark Dataset (M3FD), released alongside the Target-aware Dual Adversarial Learning (TarDAL) framework [14], expanded the task scope beyond pedestrian detection to general multi-class object detection, while also incorporating a much wider range of environmental conditions and lighting scenarios. Concurrently, a series of pioneering multispectral detectors began to analyze the complex interactions between the two modalities under degraded illumination. Xu et al. [15] investigated cross-modal representation transfer strategies. Li et al. [16] proposed a joint detection and segmentation pipeline. In a separate study, Li et al. [17] integrated illumination-aware weighting mechanisms into the Faster R-CNN architecture.

However, these initial efforts were designed for scenarios quite unlike anti-UAV detection, specifically in terms of target scale, scene layout, and typical imaging distance. Aerial RGB–infrared detection changes the nature of the problem more profoundly. The DroneVehicle dataset and Uncertainty-Aware Cross-Modality Vehicle Detection (UA-CMDet) [18] work demonstrated that fused RGB-IR imagery can deliver substantial performance gains for aerial object detection, but their focus remained exclusively on ground vehicles with relatively regular observation geometries and large target footprints. It was not until the release of the anti-UAV dataset [19] and the comprehensive benchmark study by Isaac-Medina et al. [20] that the unique challenges of UAV detection became fully apparent. When the target is a small UAV rather than a ground vehicle, four factors dominate detection errors: tiny target size due to long stand-off distances, rapid scale variation, low target saliency, and severe background clutter. The MMFW-UAV dataset [21] moved the field forward by focusing specifically on fixed-wing UAVs and air-to-air observation perspectives. Most recently, the Visible–Thermal Tiny Object Detection Benchmark (RGBT-Tiny) [11] conclusively demonstrated that tiny paired-target detection cannot be treated as a trivial extension of standard multispectral detection. When targets occupy fewer than 10 pixels, even minor sensor noise can completely obscure the target signal.

In all these UAV-specific detection scenarios, preserving faint target signatures is at least as critical as semantic classification. To tackle these challenges, many researchers have developed various tailored approaches for aerial multi-modal detection. For the widespread issue of cross-modal misalignment in aerial detection, Translation, Scale and Rotation Alignment Detection (TSRADet) [22] proposed to explicitly model three types of geometric deviations between RGB and infrared images: translation, scale, and rotation. Following this line of research, Offset-guided Adaptive Feature Alignment (OAFA) [23] further improved the alignment flexibility by introducing an offset-based mechanism to dynamically adjust the spatial positions of multi-modal features. More recently, Collaborative Compensation and Selective Fusion (CCSFuse) [4] and Dual-stream Hierarchical Interaction Network (DHANet) [5] shifted the research focus toward cross-stream compensation and hierarchical feature interaction, with the goal of preserving complementary information while minimizing the amplification of modality-specific noise. What remains less explored is the spectral structure of the fused features themselves. In most UAV-oriented detectors, visible and infrared responses are blended on full-band feature maps. This is problematic because weak UAV targets are often encoded by substantially different low-frequency and high-frequency cues across the two modalities.

2.2. Frequency-Domain Analysis for Infrared Target Detection

Frequency-domain techniques have been widely used in infrared target detection for many years. Weak thermal targets are hard to separate from cluttered backgrounds in the spatial domain alone. Early methods in this field mainly used local contrast enhancement, multi-scale decomposition and frequency-domain filtering. These techniques suppress slowly varying background components while enhancing target responses [24,25].

Frequency-domain representations offer more flexibility than direct spatial filtering. They better separate target signals from noise and background clutter. This advantage is critical for long-range infrared observation, where targets are often very small. The problem is more severe in infrared UAV detection. Targets here usually occupy only a few pixels and have very low thermal contrast. Low-frequency components capture global thermal distribution and background intensity variations. High-frequency components, on the other hand, preserve local target contours and structural details. Direct spatial-domain feature aggregation weakens discriminative target cues. It also carries background interference into later detection layers. For this reason, several recent studies have explored frequency-aware representation strategies. These methods improve both weak target detection and background suppression [3,26].

Frequency-domain modeling is not limited to infrared target detection. It also works well in other computer vision areas. These include image restoration, low-light enhancement, and image generation. For example, the Focal Frequency Loss approach [27] uses frequency-domain optimization to reduce reconstruction errors in image synthesis. Fourier-based low-light enhancement methods [28] support this view. They show that global illumination and local details can be better modeled in separate frequency subspaces. These findings show that frequency decomposition is more than a classical signal-processing tool. It provides an effective way to separate meaningful information from noise-prone components. Based on these studies, recent RGB–infrared detection methods have moved away from direct feature aggregation. They now adopt more refined cross-modal interaction and frequency-aware fusion strategies.

2.3. Fusion Mechanisms in RGB–Infrared Detection

Nearly all existing RGB–infrared object detectors adopt spatial-domain fusion as their core paradigm. Early works in this direction moved beyond simple channel concatenation

or element-wise addition. For example, the illumination-aware Faster R-CNN [17] adjusted fusion weights based on scene brightness, while the cross-modality interactive attention network modulated feature fusion according to inter-modal correlation [29]. Subsequent studies focused on diversifying feature interaction patterns, but most still retained the fundamental spatial-domain fusion foundation. Simultaneous detection and segmentation, one-stage attentive fusion, and feature interaction-driven designs all sought to make fusion more selective on the feature map [7,16,30]. Lightweight dual-branch detectors [1] and few-shot multispectral models [6] followed the same intuition in a more compact form, namely, that useful cross-modal interaction should be conditional rather than uniform. This line of work significantly improved the practical performance of RGB-T detection, but low-frequency thermal structure and high-frequency visible detail are still entangled in a single aggregation stage.

The rise of Transformer architectures opened up new possibilities for cross-modal fusion. The CFT framework was among the first to integrate both intra-modal feature refinement and cross-modal information exchange into a single self-attention block [31]. The cross-modality attention model of Lee et al. [8] and Calibrated and complementary transformer (C²former) [9] refined this direction by emphasizing explicit correspondence, calibration, and complementary modeling across modalities. Target-aware fusion for RGB-T pedestrian detection (TFDet) [2] is also instructive here. Rather than pursue ever denser feature exchange, it improved the contrast between target responses and noisy background regions after fusion. Collectively, these works reveal that the core challenge of RGB-T detection lies not in the amount of information exchanged, but in how effectively that information is filtered before being passed to the detector.

More recently, a growing subset of studies has taken a different approach. These works explore RGB-T fusion from alternative perspectives, including task coupling, training dynamics, and frequency band decomposition. TarDAL [14] tied fusion quality to downstream detection and showed that detector-oriented supervision changes what constitutes a useful fused representation. Multi-modal Object Detection from the Perspective of Mono-Modality Feature Learning (M²D-LIF) [32] argued that multi-modal joint training can weaken mono-modality feature learning and addressed this issue through distillation-assisted fusion. From the perspective of representation space, two recent works are particularly relevant to our research: Wavelet-Driven Mamba Fusion for RGB-Infrared Object Detection (WaveMamba) and Infrared and Visible Image Fusion Network Based on Discrete Wavelet Transform and Convolutional Neural Network (DCFNet) [3,26]. Unlike previous methods that treat the entire frequency spectrum as a uniform whole, both frameworks explicitly separate low-frequency and high-frequency components during the fusion process. For UAV detection, that distinction is difficult to ignore: a distant aircraft may appear as a faint structural contour in RGB, a small thermal response in IR, or an incomplete combination of the two. Once these cues are fused too early on full-band features, the detector can lose the detail needed for localization or inherit thermal clutter that is hard to suppress downstream. This is not just a minor architectural tweak. Frequency-decoupled fusion provides a principled way to explicitly regulate what information each modality contributes to weak target detection, rather than relying on implicit feature blending.

2.4. Context Representation for Small and Weak UAV Targets

The second thread closely related to UAV detection is small-target representation. As a classic structure for detection tasks, Feature Pyramid Network (FPN) [33] adopts a multi-scale feature fusion design, which mitigates semantic differences between shallow and deep feature layers. Building on this, Path Aggregation Network (PANet) [34] enhanced feature propagation by introducing an additional bottom-up path, while Scalable and Efficient

Object Detection (EfficientDet) [35] further optimized the architecture for better accuracy–efficiency trade-offs. These methods have become the de facto standard backbones for handling large-scale variations in modern detectors. These multi-scale methods play a critical role in aerial detection, yet merely optimizing feature routing mechanisms is insufficient to overcome the inherent difficulties associated with weak UAV targets. When an object occupies only a few pixels, the detector relies on very limited evidence. Motion blur, atmospheric effects, background clutter, and cross-modal inconsistency can each erase that evidence before classification and localization are attempted.

Attention mechanisms were later introduced to address this limitation by enhancing feature discriminability. Squeeze-and-Excitation Network (SENet) [36] and Convolutional Block Attention Module (CBAM) [36,37] are two representative lightweight attention modules that adaptively reweight important channels and spatial regions with minimal computational cost. Non-local blocks [38], on the other hand, expand the effective receptive field by modeling global pairwise relationships between features. While these attention mechanisms improve feature representation, their computational overhead and feature coupling characteristics require careful consideration in dual-branch RGB-T detection. A heavy global module may improve context aggregation, yet it can also make a lightweight UAV detector harder to deploy and more sensitive to irrelevant background interaction.

However, full two-dimensional global attention is computationally expensive, especially for high-resolution feature maps. Directional context modeling provides a more efficient alternative, striking a better balance between computational cost and contextual modeling capability. To address this issue, strip pooling [39] demonstrated that long-range dependencies can be efficiently captured using anisotropic strip pooling kernels instead of dense attention. Axial-DeepLab [40] reached a similar conclusion, showing that decomposing spatial interaction into separate horizontal and vertical axial operations can achieve comparable performance with much lower complexity. In tiny-object detection, HS-FPN [10] further indicated that high-frequency enhancement and spatial dependency modeling should be considered together rather than separately. These observations align well with what recent RGB–T UAV benchmarks reveal. Weak UAV targets are not recovered by fusion alone, nor by context alone. They require a detector that preserves small but discriminative cross-modal cues in the fusion stage and then stabilizes those cues with a sufficiently large yet efficient context model. Taken together, there remains a notable gap in the existing research. While RGB–infrared UAV detection has made significant progress in alignment, fusion, and small-object backbone design, relatively few studies have integrated frequency-decoupled fusion with lightweight directional context modeling into a unified framework.

3. Materials and Methods

3.1. Datasets

The performance of WFSD-Net is evaluated on two publicly available RGB-T UAV detection benchmarks: Anti-UAV300 [19,41] and MMFW-UAV [21,42]. Both datasets contain pixel-aligned visible and infrared image pairs, covering different UAV types and observation scenarios.

- **Anti-UAV300:** A widely used multi-modal benchmark consisting of 318 aligned RGB–T video pairs. It primarily focuses on rotary-wing UAVs in dynamic environments, covering diverse lighting conditions (day and night) and complex backgrounds (urban, cloud, and open fields). This dataset poses significant challenges such as thermal crossover and scale variation, serving as a core testbed for our frequency-decoupled fusion strategy.

- **MMFW-UAV:** A challenging dataset specifically designed for fixed-wing UAV detection with multi-sensor data. It contains 147,417 fixed-wing UAV images collected from three types of sensors, namely zoom visible, wide-angle visible, and thermal infrared cameras, and covers multiple observation views. Compared with the rotor-based Anti-UAV300, MMFW-UAV introduces additional difficulties related to long-range observation and viewpoint variation, as well as identifying tiny fixed-wing objects within large-scale scenes. This dataset is used to further verify the generalization capability of WFSD-Net on heterogeneous UAV targets beyond standard quadcopters.

In this study, UAV detection is formulated as a single-class object detection task, where the only foreground category is UAV and all non-UAV regions are treated as background.

3.2. Experimental Settings

For all experiments presented in this paper, all tests were conducted using a single GeForce RTX 4090 GPU. The code for the detector was implemented on the Ultralytics YOLO framework. During training, the SGD optimizer was used for 300 epochs, with a batch size set to 16 and an image input size of 640×640 pixels. Mosaic augmentation was disabled after the first 10 epochs, and all models were trained with 4-channel RGB-IR inputs to support multi-modal fusion. Logging was enabled throughout the training process to record both training metrics and error outputs for full reproducibility.

To ensure reliable and fair comparisons across all methods, independent training, validation, and test sets were used for all experiments. For the MMFW-UAV dataset, the samples were divided into three parts using a 6:2:2 ratio, yielding 7081 images for training, 2399 for validation, and 2399 for testing. Note that the original dataset contains 147,417 images; we constructed this representative subset via folder-level stratified sampling to preserve scenario diversity, and performed the split strictly at the folder level to eliminate any risk of data leakage. For the Anti-UAV300 dataset, which is derived from video frame extractions, a 7:1:2 split was adopted, resulting in 4984 training samples, 763 validation samples, and 1528 test samples.

During the training phase, the validation split was used exclusively for adjusting hyperparameters and selecting the optimal model configuration. All quantitative results presented in this paper were generated on the separate, held-out test set. The three subsets are completely disjoint. This protocol ensures that the reported mAP values reflect the generalization capability of the proposed detector rather than performance on seen samples.

3.3. Evaluation Metrics

To evaluate the detector's performance, four widely adopted metrics are used: Precision (P), Recall (R), mAP@0.5, and mAP@0.5 : 0.95. Precision quantifies how many predicted positive samples are actual targets, while Recall reflects the model's capacity to find all ground-truth objects. mAP@0.5, which is the mean Average Precision computed at an IoU threshold of 0.50, serves as a standard measure for evaluating detection results under loose localization criteria.

For this work, mAP@0.5 : 0.95 is chosen as the primary performance metric. This metric computes the average AP over ten different IoU thresholds, spanning from 0.50 to 0.95 in steps of 0.05 increments. This choice is motivated by the unique characteristics of UAV detection: small targets require precise localization, and testing across various thresholds offers a more complete way to gauge the detector's overall robustness.

Specifically, for a given IoU threshold τ , the Average Precision (AP) is calculated as the integral of the corresponding Precision–Recall curve:

$$AP_{\tau} = \int_0^1 P_{\tau}(R) dR \quad (1)$$

Here, $P_{\tau}(R)$ represents the Precision–Recall function evaluated at IoU threshold τ . Following this formulation, $\text{mAP}@0.5 : 0.95$ is obtained by averaging the AP scores over all ten predefined thresholds:

$$\text{mAP}@0.5 : 0.95 = \frac{1}{10} \sum_{\tau \in \{0.50, 0.55, \dots, 0.95\}} \text{AP}_{\tau} \quad (2)$$

Beyond detection performance, the model’s computational efficiency is also evaluated using two key metrics: total parameters (Params) and giga floating-point operations (GFLOPs). Params represent the count of all learnable weights in the network, indicating the model’s size and memory footprint. GFLOPs measures the computational cost of a single forward inference pass, which indicates the arithmetic complexity of the model.

3.4. Overall Architecture

As a representative advanced model in the YOLO series, YOLOv11 excels in balancing high-precision feature representation with optimized computational efficiency. Building on the YOLOv11 framework, our WFSN-Net is designed to enhance the precision of UAV detection tasks. The overall structure of the WFSN-Net architecture is presented in Figure 1. This dual-path network consists of three core components: a backbone, a neck, and a detection head. First, the dual-branch feature extractor of YOLOv11 is leveraged to obtain multi-scale representations from both modalities, producing three distinct feature maps at different levels. Next, the neck stage introduces our DWBF module, which is specifically built to fuse cross-modal features. A detailed diagram of the DWBF design is provided in Figure 2. DWBF firstly decouples noise-dominant sub-bands from information-rich components through DWT. In the Low-Frequency Fusion Block (LFFB), it aligns low-frequency distributions via adaptive spatial weighting. In the High-Frequency Fusion Block (HFFB), it disentangles high-frequency details using physics-aware rules, thereby achieving source-level noise suppression. Thirdly, in the head stage, ASCA is proposed for capturing global contextual dependencies to effectively decouple weak target features from background clutter, enhancing the spatial position encoding capability for weak targets, as illustrated in Figure 3. It should be noted that the proposed DWBF and ASCA modules mainly enhance intermediate feature representation and cross-modal feature interaction, while the detection formulation itself remains unchanged. Specifically, WFSN-Net preserves the original YOLOv11 detection head, prediction format, and label assignment strategy. The model retains identical box regression and object classification tasks following the initial training supervision configuration. Therefore, adopting the original YOLOv11 loss function is both reasonable and beneficial for ensuring a fair comparison with the baseline detector. Finally, the detection results of the UAV are obtained. The loss function employed in WFSN-Net is consistent with that of YOLOv11.

3.5. DWBF Module

Traditional spatial-domain fusion methods (e.g., direct addition or channel concatenation) often ignore the frequency-domain disparity between RGB and infrared modalities, causing thermal noise to contaminate visible textures. To address this, the proposed DWBF module employs spectral decomposition for source-level denoising and feature complementarity. As shown in Figure 2, it consists of three steps: wavelet decomposition, band-differentiated fusion, and inverse reconstruction.

For input visible and infrared features, DWConv is first applied for pre-alignment. A two-dimensional discrete wavelet transform (2D-DWT) then decouples them into four sub-bands: LL, LH, HL, and HH. LL contains the main energy and structure, while LH, HL,

and HH capture horizontal, vertical, and diagonal details. This decomposition separates infrared noise from visible textures.

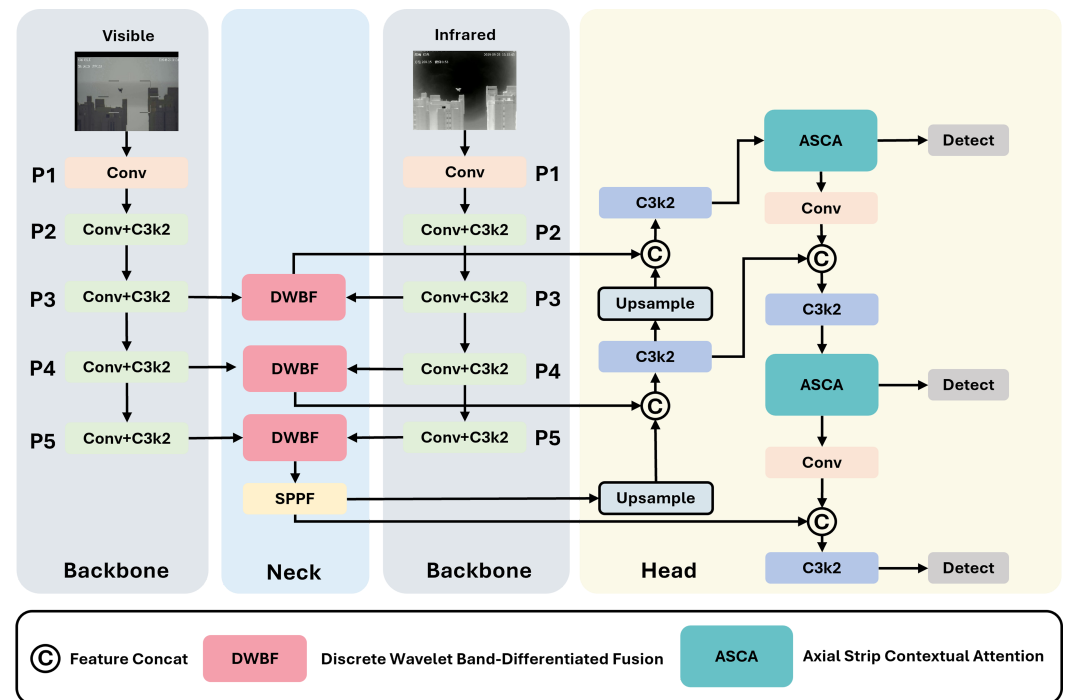


Figure 1. Overall architecture of the proposed WFSD-Net.

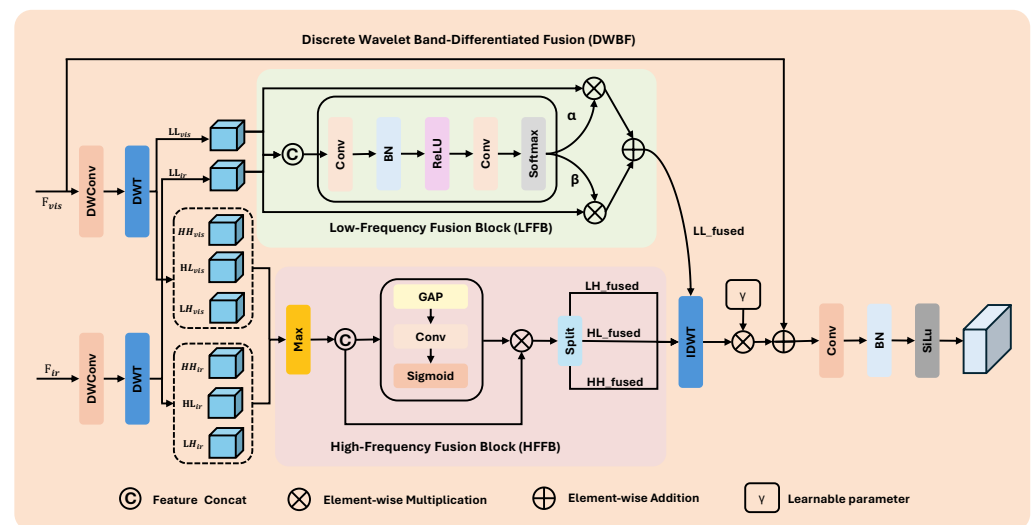


Figure 2. Structure of the proposed DWBF module.

The low-frequency component LL determines brightness and contours but often includes redundant thermal background. The pixel-level adaptive weighting strategy is designed to suppress interference. Visible and infrared LL components are concatenated and processed by a lightweight network (1×1 convolution, BN, ReLU, 1×1 convolution, Softmax), producing spatial weight maps. The fused LL is obtained as:

$$LL_{fused} = \alpha \odot LL_{vis} + \beta \odot LL_{ir} \quad (3)$$

where α and β are spatial weights, and $\odot$ denotes element-wise multiplication. LL_{vis} and LL_{ir} denote the low-frequency sub-bands of the visible and infrared features, respectively,

and LL_{fused} denotes the fused low-frequency component. This design reduces infrared noise and illumination effects.

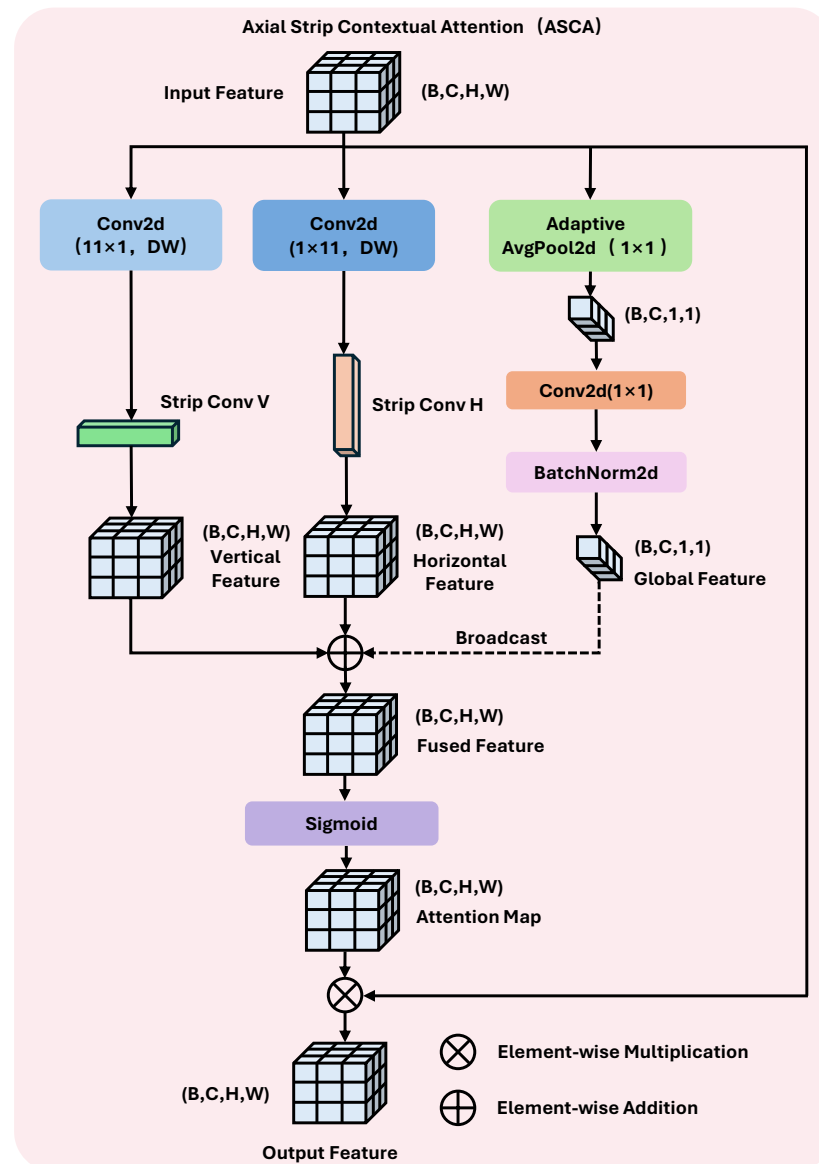


Figure 3. Structure of the proposed ASCA module.

High-frequency sub-bands LH, HL, HH contain edges and fine textures critical for UAV detection. To retain salient details, a physics-aware rule is employed, applying maximum absolute selection to each sub-band:

$$H_{\text{pre}}^k = \text{MaxAbs}(H_{\text{vis}}^k, H_{\text{ir}}^k), \quad k \in \{\text{LH}, \text{HL}, \text{HH}\} \quad (4)$$

where $\text{MaxAbs}(\cdot)$ selects the element with maximum absolute value from visible and infrared sub-bands. H_{vis}^k and H_{ir}^k denote the k -th high-frequency sub-bands of the visible and infrared features in sequence. The three pre-fused sub-bands are concatenated:

$$H_{\text{concat}} = [H_{\text{pre}}^{\text{LH}}, H_{\text{pre}}^{\text{HL}}, H_{\text{pre}}^{\text{HH}}] \quad (5)$$

where $[\cdot]$ denotes channel-wise concatenation.

Channel attention is then applied. Global average pooling, convolution, and Sigmoid generate channel-wise weights:

$$w = \sigma(\text{Conv}(\text{GAP}(H_{\text{concat}}))) \quad (6)$$

The size of the convolution kernel is dynamically adjusted according to the channel count through a non-linear mapping function:

$$k = \left\lfloor \frac{\log_2(C) + b}{\eta} \right\rfloor_{\text{odd}} \quad (7)$$

where C represents the concatenated channels of the high-frequency sub-bands (LH, HL, HH), making $C = 3 \times C_{in}$. The hyperparameters η and b are determined empirically to be 2 and 1 in that order. The notation $\lfloor \cdot \rfloor_{\text{odd}}$ denotes the nearest odd integer operation, ensuring a symmetric convolution window to preserve the feature sequence length.

Weighted fusion and split back into three sub-bands is performed as:

$$H_{\text{weighted}} = w \odot H_{\text{concat}} \quad (8)$$

$$[H_{\text{fused}}^{\text{LH}}, H_{\text{fused}}^{\text{HL}}, H_{\text{fused}}^{\text{HH}}] = \text{Split}(H_{\text{weighted}}) \quad (9)$$

$\text{Split}(\cdot)$ denotes channel-wise partition of H_{weighted} into three groups corresponding to the fused LH, HL , and HH sub-bands.

This highlights critical textures and improves discriminability. The fused sub-bands are reconstructed via IDWT. A residual fusion unit combines the reconstruction with the original input shortcut using a learnable parameter γ :

$$F_{\text{raw}} = \text{IDWT}(LL_{\text{fused}}, H_{\text{fused}}^{\text{LH}}, H_{\text{fused}}^{\text{HL}}, H_{\text{fused}}^{\text{HH}}) \quad (10)$$

$$F_{\text{out}} = F_{\text{shortcut}} + \gamma \cdot F_{\text{raw}} \quad (11)$$

Finally, the output passes through a feature adaptation module (1×1 convolution, BN, SiLU) to unify distributions and enhance representation:

$$F_{\text{fused}} = \text{Adapter}(F_{\text{out}}) \quad (12)$$

3.6. ASCA Module

To reduce computational burden while preserving contextual information, the input feature $X \in R^{C \times H \times W}$ is decomposed into orthogonal strip paths. Specifically, depthwise convolutions with kernel shapes $1 \times K$ and $K \times 1$ are applied to extract horizontal and vertical contextual features:

$$Y_h = \text{DWConv}_{1 \times K}(X), \quad Y_v = \text{DWConv}_{K \times 1}(X) \quad (13)$$

where $Y_h, Y_v \in R^{C \times H \times W}$ represent horizontal and vertical strip features. This orthogonal decomposition captures anisotropic contextual cues across various spatial orientations, offering specific advantages for distinguishing small UAV targets from complex backgrounds. Global semantic information is aggregated by compressing the input feature through global average pooling, followed by a point-wise convolution to enhance channel relevance:

$$Y_{\text{global}} = \text{BN}(\text{Conv}_{1 \times 1}(\text{AvgPool}(X))) \quad (14)$$

The global feature is broadcast to match spatial dimensions and combined with horizontal and vertical strip features. A Sigmoid activation generates the spatial attention map:

$$A = \sigma(Y_{\text{global}} \oplus Y_h \oplus Y_v) \quad (15)$$

Here, $\oplus$ represents the broadcasting addition operation, while $\sigma(\cdot)$ is the Sigmoid activation function. In the final step, the attention map is applied element-wise to the initial input feature, refining target regions and reducing background noise:

$$X_{\text{out}} = A \odot X \quad (16)$$

This mechanism enables the ASCA module to effectively separate weak target signals from complex background noise, thereby improving localization accuracy for small UAVs.

4. Results

4.1. Comparison Experiments

WFSD-Net is compared with a series of representative detectors, including YOLOv7 [43], YOLOv8 [44], YOLOX [45], YOLOv11 [46], YOLOv12 [47], YOLOv13 [48], Hyper-YOLO [49], PP-YOLOE [50], PP-PicoDet [51], RT-DETR [52], ICAFusion [53] and CMFADet [54]. To ensure a fair comparison, all benchmark models are adapted to a multi-modal configuration by concatenating visible and infrared images at the input stage, allowing them to process the same integrated spectral information as WFSD-Net. Numerical results are reported in Tables 1 and 2, and include visual comparisons in Figures 4 and 5.

Table 1. Performance comparison with different methods on the Anti-UAV300 dataset. The top performance on each dataset is marked in bold.

Method	P	R	mAP@0.5	mAP@0.5 : 0.95	Params (M)	GFLOPs
YOLOv7 [43]	0.97	0.91	0.95	0.561	43.5	130.40
YOLOv8 [44]	0.97	0.91	0.95	0.566	4.3	11.34
YOLOX [45]	0.96	0.81	0.86	0.470	3.3	9.33
YOLOv11 [46]	0.97	0.90	0.95	0.556	4.6	9.30
YOLOv12 [47]	0.96	0.91	0.94	0.559	4.0	9.71
YOLOv13 [48]	0.97	0.89	0.94	0.567	6.7	15.40
Hyper-YOLO [49]	0.97	0.91	0.95	0.573	5.7	15.04
PP-YOLOE [50]	0.97	0.91	0.95	0.567	12.7	27.89
PP-PicoDet [51]	0.97	0.91	0.95	0.570	4.7	56.82
RT-DETR [52]	0.97	0.90	0.93	0.541	66.3	194.00
ICAFusion [53]	0.97	0.91	0.94	0.570	23.2	14.88
CMFADet [54]	0.89	0.81	0.85	0.440	3.5	3.5
WFSD-Net (Ours)	0.97	0.91	0.96	0.583	3.5	9.57

On Anti-UAV300, WFSD-Net achieves scores of $P = 0.97$, $R = 0.91$, $mAP@0.5 = 0.96$, and $mAP@0.5 : 0.95 = 0.583$, surpassing all compared methods. The $mAP@0.5$ of 0.96 represents the highest value among all evaluated detectors, indicating that WFSD-Net produces fewer missed detections and false alarms under relaxed localization constraints. The improvement in $mAP@0.5 : 0.95$ from 0.573 (Hyper-YOLO) [49] to 0.583 further demonstrates that the proposed method achieves more precise bounding-box localization. This precision is critical for small UAV targets, as localization quality directly affects downstream tracking and interception tasks. Notably, WFSD-Net attains this performance with only 3.5 M parameters and 9.57 GFLOPs. It is substantially more compact than RT-DETR (66.3 M parameters, 194.00 GFLOPs) [52] and ICAFusion (23.2 M parameters) [53]. This confirms that frequency-

decoupled fusion and lightweight contextual attention can achieve competitive accuracy without sacrificing deployment efficiency.

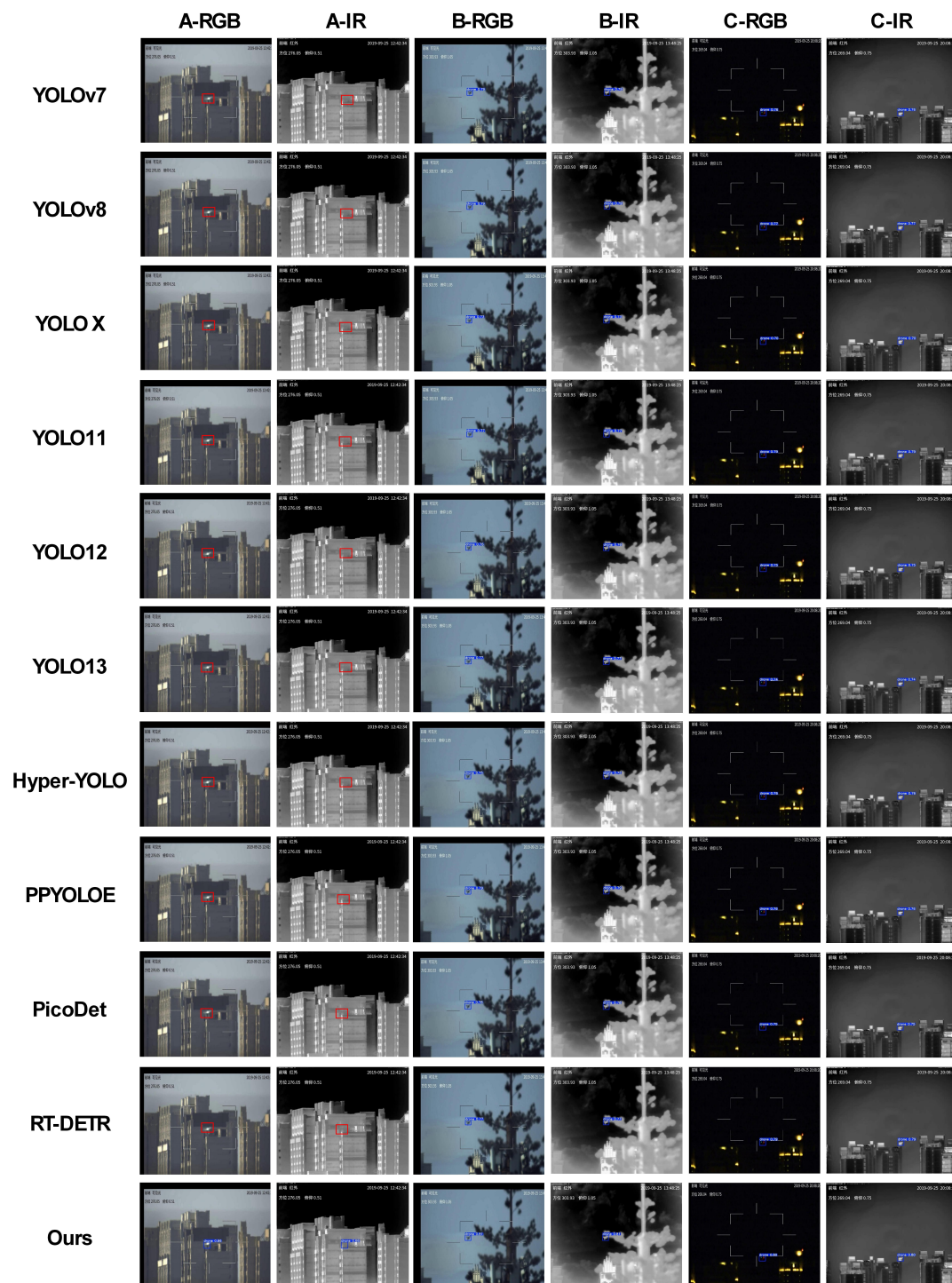


Figure 4. Comparison of detection results on the Anti-UAV300 dataset. Blue boxes with confidence scores indicate successful detections produced by the corresponding detectors, while red boxes denote manually marked target locations for missed detections.

On MMFW-UAV, WFS-Net achieves scores of $P = 0.98$, $R = 0.96$, $mAP@0.5 = 0.98$, and $mAP@0.5 : 0.95 = 0.864$, outperforming contemporary lightweight models of a similar scale (e.g., YOLOX [45], YOLOv11 [46], YOLOv12 [47]; 3.3 M–4.6 M parameters). $mAP@0.5 : 0.95$ of 0.864 exceeds PP-PicoDet (0.861) [51] and Hyper-YOLO (0.858) [49], both of which carry

substantially higher parameter counts or computational costs. The proposed method achieves the highest performance compared to all competing approaches.

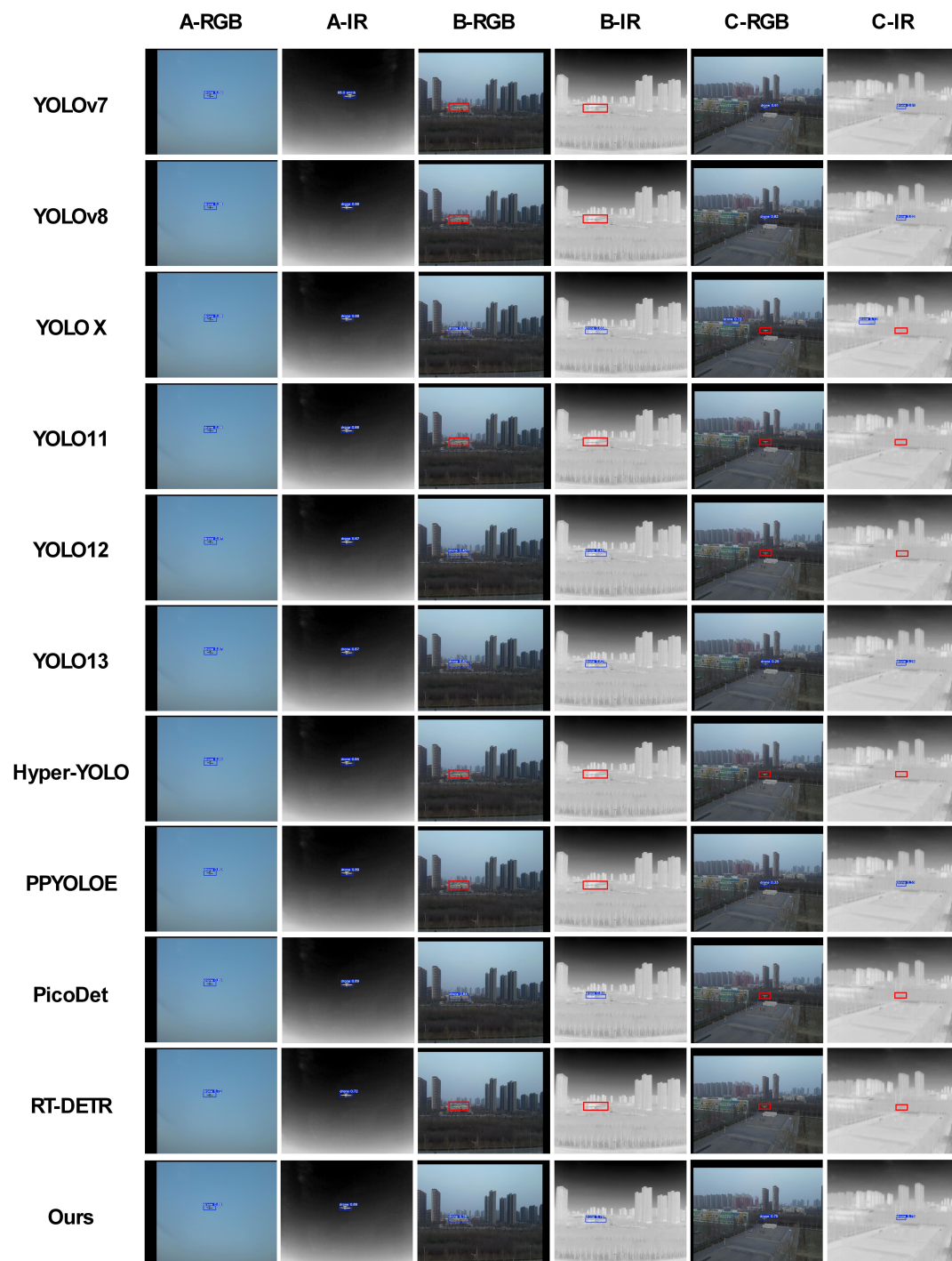


Figure 5. Comparison of detection results on the MMFW-UAV dataset. Blue boxes with confidence scores indicate successful detections produced by the corresponding detectors, while red boxes denote manually marked target locations for missed detections.

Qualitative comparisons further confirm these findings. As shown in Figures 4 and 5, our detector produces more precise bounding boxes and fewer false alarms across both RGB and infrared modalities. Under difficult conditions, including low-light urban environments and complex backgrounds, our method effectively suppresses infrared noise while preserving visible textures, enabling accurate UAV localization.

Table 2. Performance comparison with different methods on the MMFW-UAV dataset. The top performance on each dataset is marked in bold.

Method	P	R	mAP@0.5	mAP@0.5 : 0.95	Params (M)	GFLOPs
YOLOv7 [43]	0.98	0.96	0.97	0.847	43.5	130.40
YOLOv8 [44]	0.95	0.96	0.98	0.854	4.3	11.34
YOLOX [45]	0.94	0.86	0.94	0.787	3.3	9.33
YOLOv11 [46]	0.98	0.95	0.98	0.852	4.6	9.30
YOLOv12 [47]	0.96	0.96	0.98	0.848	4.0	9.71
YOLOv13 [48]	0.97	0.95	0.98	0.842	6.7	15.40
Hyper-YOLO [49]	0.97	0.95	0.97	0.858	6.6	17.62
PP-YOLOE [50]	0.97	0.96	0.98	0.835	12.7	27.89
PP-PicoDet [51]	0.97	0.95	0.97	0.861	4.7	56.82
RT-DETR [52]	0.92	0.93	0.95	0.695	66.3	194.00
ICAFusion [53]	0.97	0.96	0.97	0.750	23.2	14.88
CMFADet [54]	0.95	0.86	0.89	0.550	3.5	3.5
WFSD-Net (Ours)	0.98	0.96	0.98	0.864	3.5	9.57

4.1.1. Qualitative Comparison on Anti-UAV300

Figure 4 presents qualitative comparisons on Anti-UAV300, a benchmark characterized by ground-to-air observation, structured urban backgrounds, foreground occlusion, and nighttime low-illumination conditions.

The first example is particularly challenging because the UAV appears against a building-dominated background with strong structural interference. Under this condition, all competing detectors miss the target, while WFSD-Net successfully localizes it in both the RGB and infrared images. This behavior is consistent with the design of the proposed wavelet-guided fusion scheme. By decoupling low- and high-frequency components during multi-modal interaction, DWBF is better able to suppress background-dominant responses and retain the weak target evidence required for detection in cluttered urban scenes.

A different pattern is observed in case B. Here, all methods are able to respond to the target, so the more informative distinction lies in detection confidence rather than in box placement. WFSD-Net maintains comparatively higher and more balanced confidence across the two modalities, despite the presence of foreground branches and complex surrounding structures. WFSD-Net maintains higher and more balanced confidence scores across both modalities compared to all competing methods.

The third example comes from a nighttime scene, where the target is weakly salient and further disturbed by artificial light sources. Although all detectors generate visualized predictions, their confidence levels differ substantially. WFSD-Net yields the highest confidence in both modalities, indicating that its advantage in this case lies not in recovering an otherwise missed target, but in sustaining a stronger and more stable response under weak illumination and light-source interference. Taken together, the qualitative evidence in Figure 4 shows that the proposed method improves Anti-UAV300 detection in three complementary ways: it reduces missed detections in cluttered urban scenes, maintains stronger confidence under foreground interference, and preserves reliable cross-modal responses at night. These observations agree well with the quantitative results in Table 1.

4.1.2. Qualitative Comparison on MMFW-UAV

Figure 5 reports qualitative results on MMFW-UAV, which differs from Anti-UAV300 in that it mainly contains fixed-wing UAVs under long-range observation and large-scene backgrounds. In this setting, the main challenge arises from the relatively small target scale, elongated target geometry, and the heightened likelihood of both missed targets and false positives in wide-angle scenes.

In case A, the background is relatively simple and most detectors are able to produce valid predictions. The distinction therefore lies primarily in the confidence scores. Compared with the competing methods, WFSD-Net produces consistent detection results with higher confidence scores in both RGB and infrared modalities, while other methods exhibit lower or fluctuating confidence values.

The second example is more demanding because the target is embedded in a long-range urban scene with substantially richer background content. Under these conditions, several competing methods fail to detect the UAV, whereas WFSD-Net still produces correct detections in both modalities. Rather than reflect a minor difference in localization, this example highlights a more fundamental difference in target recoverability: the proposed detector is less likely to lose the weak fixed-wing target in the presence of large-scale scene interference.

Case C further reveals the difficulty of MMFW-UAV. YOLOX outputs a detection box in this example, but the predicted location does not correspond to the true target and should therefore be regarded as a false alarm. WFSD-Net, by contrast, correctly identifies the UAV in both modalities. This comparison is significant as it demonstrates that the advantages of our approach extend beyond recovering missed targets; it also improves discrimination between the real UAV and background structures that can attract erroneous responses. Such behavior is consistent with the joint effect of DWBF and ASCA, where frequency-decoupled fusion reduces misleading background information and axial contextual aggregation helps stabilize the representation of the true target.

The visual comparisons on MMFW-UAV show that WFSD-Net offers more reliable detection of small fixed-wing UAVs, especially in scenes where background complexity increases the likelihood of missed targets and false positives. This qualitative trend aligns with the quantitative outcomes in Table 2, further verifying the generalization capability of our approach across diverse UAV detection scenarios.

Overall, these findings confirm that the proposed DWBF-based detector delivers state-of-the-art accuracy on Anti-UAV300 and competitive performance on MMFW-UAV, while retaining a lightweight structure. This validates the effectiveness of frequency-domain fusion and contextual attention in multi-modal UAV detection.

4.2. Ablation Experiments

4.2.1. Effectiveness of Different Input Modalities and the Proposed Modules

To confirm the impact of each component, ablation studies are carried out on the Anti-UAV300 and MMFW-UAV datasets, with the findings detailed in Table 3 and 4.

Table 3. Performance comparison of different input modalities on the Anti-UAV300 and MMFW-UAV datasets. The top performance on each dataset is marked in bold.

Method	Anti-UAV300 (mAP@0.5 : 0.95)	MMFW-UAV (mAP@0.5 : 0.95)
RGB only	0.439	0.650
Infrared only	0.472	0.798
Baseline RGB-IR fusion	0.556	0.852

Effectiveness of Different Modalities: As shown in Table 3, using only RGB inputs yields limited performance (0.439 mAP@0.5 : 0.95 on Anti-UAV300 and 0.650 on MMFW-UAV). Infrared inputs provide better results (0.472 and 0.798), reflecting their robustness under low-light conditions. However, the baseline multi-modal fusion achieves significantly higher accuracy (0.556 and 0.852), confirming the necessity of RGB-IR integration for UAV detection.

Effectiveness of Proposed Modules: Table 4 shows the contribution of our proposed DWBF (Module A) and ASCA (Module B). Adding module A improves the baseline from 0.556 to 0.579 on Anti-UAV300 and from 0.852 to 0.855 on MMFW-UAV. Module B also enhances performance (0.571 and 0.858). Combining both modules produces the optimal results, achieving 0.583 on Anti-UAV300 and 0.864 on MMFW-UAV.

Table 4. The efficacy of our proposed modules across the Anti-UAV300 and MMFW-UAV benchmark datasets. The top performance on each dataset is marked in bold.

Method	Anti-UAV300 (mAP@0.5 : 0.95)	MMFW-UAV (mAP@0.5 : 0.95)
Baseline	0.556	0.852
Baseline + A	0.579	0.855
Baseline + B	0.571	0.858
Baseline + A + B	0.583	0.864

These ablation results demonstrate that multi-modal fusion is essential for UAV detection, and both DWBF and contextual attention modules contribute complementary gains. Their joint integration consistently yields the highest accuracy, validating the effectiveness of our design.

4.2.2. Analysis of DWBF Components and Hyperparameters on the Anti-UAV300 Dataset

The contributions of the LFFB and HFFB within the DWBF module are evaluated. As presented in Table 5, the HFFB-only configuration attains an mAP@0.5 : 0.95 of 0.573, while the LFFB-only configuration reaches 0.568. The full DWBF module, which integrates both branches, obtains the highest mAP@0.5 : 0.95 of 0.583.

Table 5. Ablation study on different frequency components within the DWBF module on the Anti-UAV300 dataset. The best performance is highlighted in bold.

Component	mAP@0.5	mAP@0.5 : 0.95	Params (M)	GFLOPs
HFFB only	0.95	0.573	3.4	9.57
LFFB only	0.95	0.568	3.5	9.57
Full DWBF (LFFB + HFFB)	0.96	0.583	3.5	9.57

Further analysis is conducted on the sensitivity of the DWBF module to hyperparameters η and b , which jointly determine the adaptive kernel size in the high-frequency channel attention. Table 6 summarizes the results across several representative combinations. The configuration $\eta = 2, b = 1$ yields the best mAP@0.5 : 0.95 of 0.583. When η is increased to 3, the performance drops to 0.564. The remaining configurations produce mAP@0.5 : 0.95 values between 0.571 and 0.574.

Table 6. Hyperparameter sensitivity analysis of the DWBF module. The top performance on each dataset is marked in bold.

η	b	P	R	mAP@0.5	mAP@0.5 : 0.95	Params (M)	GFLOPs
1	1	0.97	0.91	0.95	0.571	3.5	9.57
2	0	0.97	0.91	0.95	0.574	3.5	9.57
2	2	0.97	0.91	0.95	0.574	3.5	9.57
3	1	0.97	0.91	0.95	0.564	3.5	9.57
2	1	0.97	0.91	0.96	0.583	3.5	9.57

4.2.3. Analysis of ASCA Components on the Anti-UAV300 Dataset

The contribution of each sub-component within the ASCA module is analyzed on the Anti-UAV300 dataset. Table 7 reports the results for three configurations. The strip-only variant, which omits the global context branch, yields an mAP@0.5 : 0.95 of 0.573. Replacing the orthogonal strip convolutions with a standard 3×3 depthwise convolution leads to a slightly lower score of 0.565. The complete ASCA module, combining the global branch and the orthogonal strip features, achieves the highest mAP@0.5 : 0.95 of 0.583, with 3.5 M parameters and 9.57 GFLOPs.

Table 7. Ablation study on different components within the ASCA module on the Anti-UAV300 dataset. The top performance on each dataset is marked in bold.

Component	mAP@0.5	mAP@0.5 : 0.95	Params (M)	GFLOPs
Strip-only	0.95	0.573	3.4	9.15
Global + 3×3 DWConv	0.95	0.565	3.4	9.57
Full ASCA (Global + Strip)	0.96	0.583	3.5	9.57

5. Discussion

5.1. Effectiveness of Frequency–Spatial Decoupling for Multi-Modal UAV Detection

The quantitative results on both Anti-UAV300 and MMFW-UAV consistently demonstrate that WFSD-Net outperforms competing detectors across all reported metrics, including Precision, Recall, mAP@0.5, and mAP@0.5 : 0.95. These gains are not attributable to a single design choice but reflect the joint effect of two complementary mechanisms operating at different representation levels.

The DWBF module addresses a fundamental limitation of conventional spatial-domain fusion: when visible and infrared features are aggregated on full-band feature maps, low-frequency thermal background and high-frequency visible noise are inevitably entangled in the fused representation. WaveMamba [3] and DCFNet [26] demonstrate that frequency decomposition effectively separates target signals from cross-modal noise. Our work further shows that band-differentiated fusion is particularly critical for UAV detection. By performing wavelet decomposition prior to fusion, DWBF separates the spectral content into sub-bands with distinct physical interpretations. The low-frequency branch applies adaptive spatial weighting to align thermal and visible intensity distributions, suppressing slowly varying background components that would otherwise dominate the fused response. The high-frequency branch retains the most salient edge and contour information from both modalities via maximum absolute selection. Subsequent channel attention further enhances discriminative structural details.

The ablation results in Table 5 confirm that both branches contribute independently, and their combination yields the highest mAP@0.5 : 0.95 of 0.583 on Anti-UAV300, compared with 0.573 for HFFB alone and 0.568 for LFFB alone. This pattern suggests that low-frequency alignment and high-frequency detail preservation address complementary aspects of the fusion problem and are most effective when applied jointly. This advantage is particularly pronounced for long-range fixed-wing UAV detection, where targets often occupy only a few pixels and provide limited visual evidence. As reported in the RGBT-Tiny benchmark [11], tiny paired-target detection requires preserving discriminative frequency cues from both modalities, which is exactly what our frequency-decoupled fusion strategy achieves.

5.2. Role of Contextual Attention in Weak Target Localization

The ASCA module addresses a different but equally important challenge: once cross-modal features have been fused, the detector must still distinguish the weak UAV response from residual background interference across the spatial domain. Standard convolution has inherent limitations in this context. Its local receptive field cannot capture the long-range geometric dependencies required to contextualize small targets in large scenes.

Strip pooling [39] and Axial-DeepLab [40] show that decomposing spatial interaction into orthogonal horizontal and vertical operations captures long-range dependencies with significantly lower computational cost than full 2D attention. Building on this insight, ASCA resolves this tension by decomposing spatial context into orthogonal strip operations along horizontal and vertical axes, supplemented by a global average pooling branch that captures scene-level semantic information.

The ablation results in Table 7 show that the strip-only configuration achieves an mAP@0.5 : 0.95 of 0.573, while replacing the orthogonal strips with a standard 3×3 depth-wise convolution reduces performance to 0.565. The full ASCA module, which combines both the global branch and the orthogonal strip features, reaches 0.583. The marginal but consistent improvement from the global branch shows that scene-level context effectively complements directional strip features. This is especially important in scenes where the UAV occupies only a small fraction of the image and its local appearance offers limited discriminative information. In particular, qualitative results on Anti-UAV300 show that ASCA improves target response stability by strengthening contextual support around small UAVs, even under partial foreground occlusion.

5.3. Spectral Characterization of Wavelet Sub-Bands Across Modalities

To further substantiate the rationale behind the band-differentiated fusion strategy in DWBF, the spectral energy distribution of wavelet sub-bands across the two modalities is analyzed. Figure 6 presents the averaged normalized information entropy of each DWT sub-band computed over representative images from the Anti-UAV300 dataset.

The entropy values reveal a consistent and physically interpretable pattern. For both RGB and infrared images, the LL sub-band carries substantially higher entropy than the high-frequency sub-bands, reflecting the dominant role of low-frequency components in encoding global intensity distribution and coarse structural content. Across modalities, however, the LL entropy of the infrared channel is notably higher than that of the visible channel, indicating that infrared images exhibit a more diffuse and spatially uniform low-frequency energy distribution. This is consistent with the thermal imaging principle: infrared sensors respond to emitted radiation rather than reflected light, producing smoother intensity gradients that spread energy more broadly across the low-frequency spectrum. In contrast, visible images concentrate low-frequency energy more selectively around illuminated structures, resulting in lower overall LL entropy.

The high-frequency sub-bands exhibit the opposite tendency. Their entropy values are higher in the visible modality than in the infrared, reflecting the richer edge and texture content captured by optical sensors under adequate illumination. This spectral disparity has also been observed in the RGBT-Tiny benchmark [11], which reported that tiny targets exhibit distinct frequency signatures across visible and infrared modalities.

These observations provide direct empirical support for the design choices in DWBF. The elevated and diffuse low-frequency entropy of infrared images motivates the use of adaptive spatial weighting in LFFB. A uniform aggregation of the two LL sub-bands would allow the thermally dominant infrared background to suppress the more structured visible low-frequency content. This would degrade the quality of the fused representation. The higher high-frequency entropy of visible images, by contrast, justifies the

maximum absolute selection strategy in HFFB, which preferentially retains the modality with the stronger local response at each spatial location. Taken together, the entropy analysis confirms that the two modalities carry fundamentally different spectral signatures across frequency sub-bands. It also shows that treating them with a uniform fusion rule is suboptimal for UAV detection. This is exactly the implicit approach used in spatial-domain concatenation.

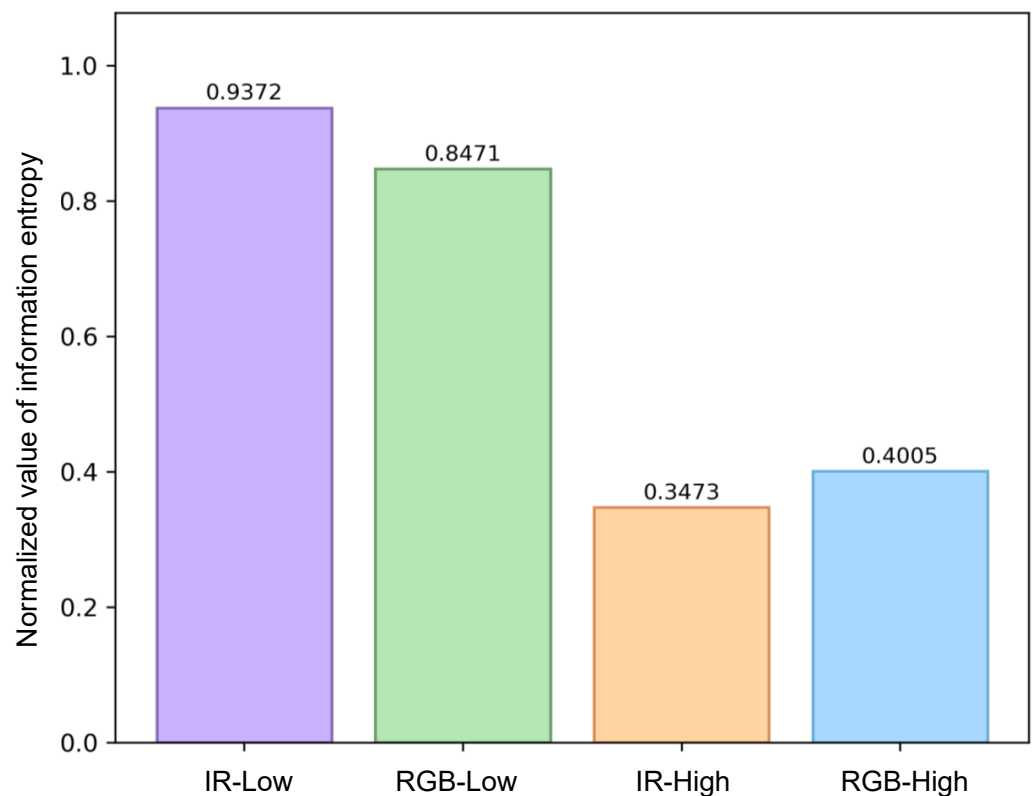


Figure 6. Averaged normalized information entropy of DWT sub-bands for RGB and infrared images.

5.4. Sensitivity Analysis and Hyperparameter Robustness

The hyperparameter sensitivity analysis in Table 6 reveals that the DWBF module is relatively robust to the choice of η and b within a reasonable range. The configuration $\eta = 2$, $b = 1$ yields the best mAP@0.5 : 0.95 of 0.583, while most alternative settings produce values between 0.571 and 0.574. The only notable exception is $\eta = 3$, $b = 1$, which reduces performance to 0.564. This degradation can be attributed to the larger kernel size produced under this setting, which may introduce excessive smoothing in the channel attention operation and reduce the selectivity of the high-frequency weighting. The relatively narrow performance range across other configurations suggests that the adaptive kernel size formula in Equation (7) provides a stable and principled mechanism for adjusting attention granularity to the channel dimension, without requiring exhaustive manual tuning.

5.5. Overall Assessment and Design Implications

These performance improvements are primarily driven by the proposed innovations, which achieve a favorable accuracy–efficiency balance in complex multi-modal scenarios. The DWBF module mitigates spatial fusion noise via spectral decomposition. It explicitly decouples noise-dominant sub-bands from information-rich components. This enables accurate modal complementarity while avoiding heavy parameter overhead. The ASCA module addresses the limited receptive fields of standard convolutions by leveraging anisotropic

strip convolutions to capture global contextual dependencies, effectively separating weak, highly dynamic UAV targets from severe background clutter.

Notably, WFSD-Net achieves these gains with only 3.5 M parameters and 9.57 GFLOPs. Its footprint is comparable to single-modality lightweight detectors. This suggests that the improvements stem from more effective use of cross-modal information rather than increased model capacity. Compared with recent state-of-the-art UAV detectors that rely on spatial-domain feature interaction, such as CCSFuse [4] and DHANet [5], our method achieves higher detection accuracy with similar computational cost. Through the integration of these lightweight spectral and spatial strategies, WFSD-Net establishes a robust baseline for dual-modal aerial surveillance, confirming that jointly decoupling noise and targets across both frequency and spatial domains represents an effective paradigm for visible–infrared UAV detection.

6. Conclusions

This article introduces WFSD-Net, a wavelet-guided frequency–spatial decoupling network for visible–infrared UAV detection. The proposed DWBF module performs spectral decomposition to separate noise-dominant sub-bands from informative components, achieving source-level denoising and enhancing modal complementarity. In parallel, the ASCA module employs anisotropic strip convolution to capture long-range dependencies, effectively decoupling weak target signals from complex backgrounds and improving spatial perception of small UAVs.

Extensive experiments on Anti-UAV300 and MMFW-UAV datasets demonstrate that WFSD-Net achieves superior accuracy compared with state-of-the-art detectors while maintaining a lightweight architecture. Ablation studies further confirm the necessity of multi-modal fusion and the complementary contributions of DWBF and ASCA. Overall, the proposed framework provides a robust and efficient solution for UAV detection, offering resilience against spectral interference in the frequency domain and enhanced target–background separation in the spatial domain.

Future work will extend WFSD-Net to broader spectral combinations and explore adaptive fusion strategies for real-time UAV monitoring under diverse operational conditions.

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

Funding: This work was supported in part by the National Natural Science Foundation of China General Program under Grant 62471389, as well as the Shaanxi Provincial Department of Education Special Program for Service to Local Region under Grant 25JE014.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed in this study are publicly available benchmark datasets, including Anti-UAV300 and MMFW-UAV.

Acknowledgments: During the preparation of this work, the authors used DeepSeek solely for text editing and language polishing. The authors have reviewed and edited the content as needed and take full responsibility for the content of the publication.

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

References

- Huang, L.; Peng, Z.; Chen, F.; Dai, S.; He, Z.; Liu, K. Cross-modality interaction for few-shot multispectral object detection with semantic knowledge. *Neural Netw.* **2024**, *173*, 106156. [[PubMed](#)]
- Zhang, X.; Zhang, X.; Wang, J.; Ying, J.; Sheng, Z.; Yu, H.; Li, C.; Shen, H.L. TFDet: Target-aware fusion for RGB-T pedestrian detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 13276–13290.
- Zhu, H.; Dong, W.; Yang, L.; Li, H.; Yang, Y.; Ren, Y.; Zhu, Q.; Feng, Z.; Li, C.; Lin, S.; et al. WaveMamba: Wavelet-Driven Mamba Fusion for RGB-Infrared Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025; pp. 11219–11229.
- Zhang, T.; Lu, R.; Yang, X.; Zhang, D.; Li, Y.; Xie, X.; Li, Y. CCSFuse: Collaborative Compensation and Selective Fusion for UAV-Based RGB-IR Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *64*, 5400314.
- Wu, X.; Wang, L.; Guan, J.; Ji, H.; Xu, L.; Hou, Y.; Fei, A. DHANet: Dual-stream Hierarchical Interaction Networks for Multimodal Drone Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5406013. [[CrossRef](#)]
- Tian, D.; Yan, X.; Zhou, D.; Wang, C.; Zhang, W. IV-YOLO: A lightweight Dual-Branch object detection network. *Sensors* **2024**, *24*, 6181. [[PubMed](#)]
- Xie, Y.; Zhang, L.; Yu, X.; Xie, W. YOLO-MS: Multispectral object detection via feature interaction and self-attention guided fusion. *IEEE Trans. Cogn. Dev. Syst.* **2023**, *15*, 2132–2143.
- Lee, W.Y.; Jovanov, L.; Philips, W. Cross-modality attention and multimodal fusion transformer for pedestrian detection. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 608–623.
- Yuan, M.; Wei, X. C²former: Calibrated and complementary transformer for rgb-infrared object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5403712.
- 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*; AAAI Press: Philadelphia, PA, USA, 2025; Volume 39, pp. 6896–6904.
- Ying, X.; Xiao, C.; An, W.; Li, R.; He, X.; Li, B.; Cao, X.; Li, Z.; Wang, Y.; Hu, M.; et al. Visible-thermal tiny object detection: A benchmark dataset and baselines. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 6088–6096. [[CrossRef](#)] [[PubMed](#)]
- Hwang, S.; Park, J.; Kim, N.; Choi, Y.; Kweon, I.S. Multispectral Pedestrian Detection: Benchmark Dataset and Baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1037–1045. [[CrossRef](#)]
- Jia, X.; Zhu, C.; Li, M.; Tang, W.; Zhou, W. LLVIP: A Visible-Infrared Paired Dataset for Low-Light Vision. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 11–17 October 2021; pp. 3496–3504.
- Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; Luo, Z. Target-Aware Dual Adversarial Learning and a Multi-Scenario Multi-Modality Benchmark To Fuse Infrared and Visible for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5802–5811.
- Xu, D.; Ouyang, W.; Ricci, E.; Wang, X.; Sebe, N. Learning Cross-Modal Deep Representations for Robust Pedestrian Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4236–4244. [[CrossRef](#)]
- Li, C.; Song, D.; Tong, R.; Tang, M. Multispectral Pedestrian Detection via Simultaneous Detection and Segmentation. In Proceedings of the British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018.
- Li, C.; Song, D.; Tong, R.; Tang, M. Illumination-aware Faster R-CNN for Robust Multispectral Pedestrian Detection. *Pattern Recognit.* **2019**, *85*, 161–171. [[CrossRef](#)]
- Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-Based RGB-Infrared Cross-Modality Vehicle Detection via Uncertainty-Aware Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700–6713. [[CrossRef](#)]
- Jiang, N.; Wang, K.; Peng, X.; Yu, X.; Wang, Q.; Xing, J.; Li, G.; Zhao, J.; Guo, G.; Han, Z. Anti-UAV: A Large Multi-Modal Benchmark for UAV Tracking. *arXiv* **2021**, arXiv:2101.08466.
- Isaac-Medina, B.K.S.; Poyser, M.; Organisciak, D.; Willcocks, C.G.; Breckon, T.P.; Shum, H.P.H. Unmanned Aerial Vehicle Visual Detection and Tracking Using Deep Neural Networks: A Performance Benchmark. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 11–17 October 2021; pp. 1223–1232.
- Liu, Y.; Sun, Z.; Xi, L.; Zhang, L.; Dong, W.; Chen, C.; Lu, M.; Fu, H.; Deng, F. MMFW-UAV Dataset: Multi-Sensor and Multi-View Fixed-Wing UAV Dataset for Air-to-Air Vision Tasks. *Sci. Data* **2025**, *12*, 185. [[CrossRef](#)] [[PubMed](#)]
- Yuan, M.; Wang, Y.; Wei, X. Translation, Scale and Rotation: Cross-Modal Alignment Meets RGB-Infrared Vehicle Detection. In *Proceedings of the Computer Vision—ECCV 2022*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 509–525. [[CrossRef](#)]
- Chen, C.; Qi, J.; Liu, X.; Bin, K.; Fu, R.; Hu, X.; Zhong, P. Weakly Misalignment-free Adaptive Feature Alignment for UAVs-based Multimodal Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 26836–26845.

24. Deshpande, S.D.; Meng, H.; Venkateswarlu, R.; Luo, P. Max-mean and max-median filters for detection of small targets. *Signal Data Process. Small Targets* **1999**, *3809*, 74–83.
25. Gao, C.; Meng, D.; Yang, Y.; Wang, Y.; Zhou, X.; Hauptmann, A.G. Infrared small target detection using multiscale patch-based contrast measure. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 4241–4254.
26. Wu, D.; Wang, Y.; Wang, H.; Wang, F.; Gao, G. DCFNet: Infrared and Visible Image Fusion Network Based on Discrete Wavelet Transform and Convolutional Neural Network. *Sensors* **2024**, *24*, 4065. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Jiang, L.; Dai, B.; Wu, W.; Loy, C.C. Focal Frequency Loss for Image Reconstruction and Synthesis. In Proceedings of the ICCV, Montreal, QC, Canada, 10–17 October 2021.
28. Wang, C.; Wu, H.; Jin, Z. FourLLIE: Boosting Low-Light Image Enhancement by Fourier Frequency Information. In Proceedings of the ACM Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023.
29. Zhang, L.; Liu, Z.; Zhang, S.; Yang, X.; Qiao, H.; Huang, K.; Hussain, A. Cross-modality Interactive Attention Network for Multispectral Pedestrian Detection. *Inf. Fusion* **2019**, *50*, 20–29. [\[CrossRef\]](#)
30. Cao, Z.; Yang, H.; Zhao, J.; Guo, S.; Li, L. Attention Fusion for One-Stage Multispectral Pedestrian Detection. *Sensors* **2021**, *21*, 4184. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Fang, Q.; Han, D.; Wang, Z. Cross-Modality Fusion Transformer for Multispectral Object Detection. *arXiv* **2021**, arXiv:2111.00273.
32. Zhao, T.; Liu, B.; Gao, Y.; Sun, Y.; Yuan, M.; Wei, X. Rethinking Multi-modal Object Detection from the Perspective of Mono-Modality Feature Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–20 October 2025; pp. 6364–6373.
33. Lin, T.Y.; 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), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [\[CrossRef\]](#)
34. 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), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768. [\[CrossRef\]](#)
35. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 10781–10790. [\[CrossRef\]](#)
36. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [\[CrossRef\]](#)
37. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the Computer Vision–ECCV 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–19. [\[CrossRef\]](#)
38. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-Local Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
39. Hou, Q.; Zhang, L.; Cheng, M.M.; Feng, J. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4003–4012. [\[CrossRef\]](#)
40. Wang, H.; Zhu, Y.; Green, B.; Adam, H.; Yuille, A.; Chen, L.C. Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. In *Proceedings of the Computer Vision–ECCV 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 108–126. [\[CrossRef\]](#)
41. Zhao, J.; Li, J.; Jin, L.; Chu, J.; Wang, J.; Xia, J.; Wang, K.; Liu, Y.; Gulshad, S.; Zhao, J.; et al. The 3rd Anti-UAV Workshop & Challenge: Methods and Results. *arXiv* **2023**, arXiv:2305.07290.
42. Jahan, I.; Majid, M.E.; Murugappan, M.; Chowdhury, M.E.H.; Prakash, N.B.; Kashem, S.B.A.; Balusamy, B.; Khandakar, A. Alignment-Aware and Reliability-Gated Multimodal Fusion for Unmanned Aerial Vehicle Detection Across Heterogeneous Thermal-Visual Sensors. *arXiv* **2026**, arXiv:2603.08208.
43. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
44. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO. GitHub Repository, 2023. GitHub Repository. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 1 January 2023).
45. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
46. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725.
47. Tian, Y.; Ye, Q.; Doermann, D. Yolov12: Attention-centric real-time object detectors. *arXiv* **2025**, arXiv:2502.12524.
48. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733.
49. Feng, Y.; Huang, J.; Du, S.; Ying, S.; Yong, J.H.; Li, Y.; Ding, G.; Ji, R.; Gao, Y. Hyper-yolo: When visual object detection meets hypergraph computation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 2388–2401. [\[CrossRef\]](#)

50. Xu, S.; Wang, X.; Lv, W.; Chang, Q.; Cui, C.; Deng, K.; Wang, G.; Dang, Q.; Wei, S.; Du, Y.; et al. PP-YOLOE: An evolved version of YOLO. *arXiv* **2022**, arXiv:2203.16250.
51. Yu, G.; Chang, Q.; Lv, W.; Xu, C.; Cui, C.; Ji, W.; Dang, Q.; Deng, K.; Wang, G.; Du, Y.; et al. PP-PicoDet: A better real-time object detector on mobile devices. *arXiv* **2021**, arXiv:2111.00902.
52. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
53. Shen, J.; Chen, Y.; Liu, Y.; Zuo, X.; Fan, H.; Yang, W. ICAFusion: Iterative Cross-Attention Guided Feature Fusion for Multispectral Object Detection. *Pattern Recognit.* **2024**, *145*, 109913. [[CrossRef](#)]
54. Yu, C.; Shin, Y. A Cross-Modality Feature Adaptive Interaction Approach for RGB-Infrared Object Detection in Aerial Imagery. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5401315. [[CrossRef](#)]

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