

Article

WaveUAV-YOLO: A Lightweight Architecture for UAV Object Detection with Frequency-Domain Edge Preservation and Stable Gradient Fusion

Qi Wang ¹, Shengqi Xu ^{1,*} and Yongji Chen ^{2,3}

¹ School of Computer Science and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China; 002086@nuist.edu.cn

² School of Data Science and Big Data Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China

³ School of Applied Mathematics, University of Reading, Reading RG6 6DX, UK

* Correspondence: 202512491192@nuist.edu.cn

Highlights

What are the main findings?

- A lightweight architecture, WaveUAV-YOLO, is proposed to effectively preserve sub-pixel edges and stabilize gradient fusion for UAV object detection.
- By integrating the WHFD, C2f_AMSB, and FFM_Concat modules, the model's detection accuracy for small targets is significantly improved against complex background interference.

What are the implications of the main findings?

- Extensive experiments on the VisDrone2019, DIOR, NWPU VHR-10, and HIT-UAV datasets confirm that the model achieves advanced overall performance and strong cross-modal generalization capabilities.
- WaveUAV-YOLO provides an effective, real-time solution for wide-area UAV surveillance, demonstrating great potential for practical applications such as 24/7 urban traffic monitoring and emergency rescue.

Abstract

Unmanned Aerial Vehicle (UAV) object detection is a core technology for cross-modal, wide-area surveillance. However, operating at high altitudes under complex flight conditions imposes extreme physical constraints, generating massive sub-pixel targets, elongated morphologies, and dense object clustering. These challenges are critically important because conventional lightweight detectors deployed on edge devices often suffer from severe high-frequency edge loss during spatial downsampling, morphological distortion of anisotropic targets, and the mathematical suppression of weak feature gradients during cross-scale fusion, ultimately leading to severe missed detections. To overcome these inherent bottlenecks, this paper proposes WaveUAV-YOLO, a lightweight architecture prioritizing frequency-domain edge preservation and stable gradient propagation. First, a Wavelet High-Frequency Downsampling (WHFD) module augments standard convolutions by utilizing Haar wavelet decomposition to explicitly capture and compensate for the lost boundary cues of sub-pixel targets. Second, an Asymmetric Multi-scale Bottleneck without Dimensionality Compression (C2f_AMSB) cancels forced channel compression and introduces asymmetric convolutions to effectively adapt to elongated targets. Third, a Mean-Normalized Feature Aggregation (FFM_Concat) prevents deep background features from suppressing weak shallow signals during fusion. Extensive experiments on VisDrone2019, DIOR, NWPU, and HIT-UAV demonstrate that WaveUAV-YOLO (11.70 M



Academic Editor: Xinghua Li

Received: 9 June 2026

Revised: 11 July 2026

Accepted: 17 July 2026

Published: 20 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

parameters) achieves competitive or superior detection precision against recent lightweight UAV detectors. Furthermore, it achieves an end-to-end inference speed of ≈ 50 FPS on a standard desktop GPU, validating a favorable balance between sub-pixel detection reliability and computational efficiency for adverse UAV environments.

Keywords: UAV object detection; discrete wavelet transform; mean-normalized fusion; mechanism coupling; remote sensing imagery

1. Introduction

With the rapid development of unmanned aerial vehicle (UAV) technology, real-time object detection from a top-down view has become a core challenge in computer vision [1]. Specifically, this work targets the typical application scenario of 24/7 high-altitude wide-area surveillance and reconnaissance (e.g., day-and-night urban traffic monitoring, security patrol, and emergency search and rescue) [2–4]. This specific scenario imposes three stringent constraints on detectors: (i) high-altitude perspectives induce extreme scale variations [5] and yield numerous sub-pixel small targets [6]; (ii) wide-area surveillance frequently involves severe background clutter and dense target clustering with significant anisotropy; (iii) 24/7 continuous operations demand robust cross-modal perception capabilities across heterogeneous imaging conditions, including daytime optical, high-resolution remote sensing, and night-time infrared imaging. Consequently, balancing real-time inference efficiency with reliable cross-modal feature extraction under such varied environmental constraints remains a critical and largely unresolved bottleneck.

While numerous advanced lightweight YOLO variants (e.g., YOLOv8 to YOLOv13 [7,8]) and UAV-specific detectors (e.g., ST-YOLO [9] and MBD-YOLO [10]) have been developed, a systematic analysis of recent advanced lightweight detectors reveals that they predominantly inherit standard fundamental operators that inherently degrade sub-pixel features under high-altitude views. For instance, ST-YOLO builds fine-grained operators using star operations but still relies on standard isotropic convolutions, which fails to model the extreme anisotropy in top-down views. Similarly, MBD-YOLO and CCAI-YOLO [11] optimize multi-branch topologies and cross-channel attention; however, they apply mandatory channel compression ($e = 0.5$) within their bottlenecks, which distorts the shape details of sub-pixel elongated targets. Moreover, recent general-purpose YOLO variants (e.g., YOLOv10 to YOLOv13) predominantly employ standard channel concatenation for multi-scale fusion. This makes them vulnerable to the suppression of shallow weak gradients by dominant deep background signals, a critical issue for dense small target detection. To explicitly resolve these persistent morphological and gradient challenges, this study proposes WaveUAV-YOLO, targeting three specific bottlenecks through corresponding architectural innovations:

- **Unresolved Limitation 1 (Loss of high-frequency edges):** Recent baselines rely on stride-2 convolutions or standard pooling for spatial downsampling. These operations act as lossy low-pass filters, tending to smooth out the delicate, sub-pixel high-frequency edges of tiny targets. Our Solution: We propose the Wavelet High-Frequency Downsampling (WHFD) module, which introduces a parallel Haar wavelet gating path to compensate for the information loss in stride-2 downsampling, effectively preserving boundary fidelity without heavy computation.
- **Unresolved Limitation 2 (Morphological distortion):** Existing detectors typically utilize isotropic square convolutions and impose mandatory channel compression (e.g., expansion ratio $e = 0.5$) within bottlenecks. This forced compression tends to distort

the anisotropic shapes of elongated targets (e.g., vehicles) common in UAV views. Our Solution: We design the Asymmetric Multi-scale Bottleneck (C2f_AMSB), which removes channel dimensionality compression ($e = 1.0$) and deploys asymmetric convolutions (1×3 and 3×1) to accurately model extreme aspect ratios.

- Unresolved Limitation 3 (Gradient suppression during fusion): During cross-level feature aggregation, current architectures mainly use standard channel concatenation. This direct fusion causes the strong activation gradients of deep backgrounds to mathematically suppress the weak gradients of shallow small targets during backpropagation. Our Solution: We introduce the Mean-Normalized Feature Aggregation (FFM_Concat) mechanism, a structural weighting scheme that maintains the magnitude of weak feature streams and defends their gradient space against background dominance.

Instead of blindly increasing network topological complexity, we draw inspiration from asymmetric multi-kernel strategies for slender tubular structures in medical image segmentation and frequency-domain mechanisms for low-SNR infrared perception [12]. We observe that anisotropic targets and weak signals in UAV images share two key properties: extreme aspect ratios and high-frequency edges. Therefore, we reconstruct the feature downsampling and bottleneck extraction modules specifically for UAV views to construct our WaveUAV-YOLO architecture.

The main contributions of this paper are threefold:

1. A Wavelet High-Frequency Downsampling (WHFD) module: WHFD uses Haar wavelet transform to extract high-frequency edge masks, preserving boundaries during downsampling with low computational cost.
2. An Asymmetric Multi-scale Bottleneck without Dimensionality Compression (C2f_AMSB): C2f_AMSB removes channel compression and uses asymmetric convolution kernels to better represent elongated targets under UAV views.
3. A Mean-Normalized Feature Aggregation mechanism (FFM_Concat): FFM_Concat uses a mean-constrained weighting scheme to prevent strong features from suppressing weak ones during cross-level concatenation.

We conduct extensive ablation and generalization experiments on four representative datasets: VisDrone2019, NWPU VHR-10, HIT-UAV, and DIOR. These datasets are deliberately selected to jointly validate the model's reliability in the targeted 24/7 wide-area surveillance scenario: VisDrone2019 represents typical daytime urban densely clustered scenes with extreme scale variation; DIOR and NWPU VHR-10 assess the model's robustness against complex backgrounds and viewpoint diversity from higher-altitude wide-area perspectives; and HIT-UAV evaluates the cross-modal perception capability under low signal-to-noise ratio (SNR) thermal conditions. Such cross-modality and cross-altitude evaluation ensures that the proposed architecture is not overfitted to a single imaging condition but generalizes coherently to the broader surveillance and reconnaissance scenario. The results show that our improvements significantly boost detection performance for dense small targets under complex backgrounds while maintaining high FPS, demonstrating good engineering potential.

The conceptual comparison between conventional mechanisms and our proposed WaveUAV-YOLO is intuitively illustrated in Figure 1. As depicted, conventional architectures typically suffer from high-frequency edge loss during spatial downsampling, morphological distortion of elongated targets, and gradient suppression of weak signals during feature fusion. In contrast, WaveUAV-YOLO systematically overcomes these bottlenecks through frequency-domain refinement (WHFD), anisotropic morphological modeling

(C2f_AMSB), and magnitude-safeguarding feature aggregation (FFM_Concat), establishing a robust and efficient paradigm for UAV top-down perception.

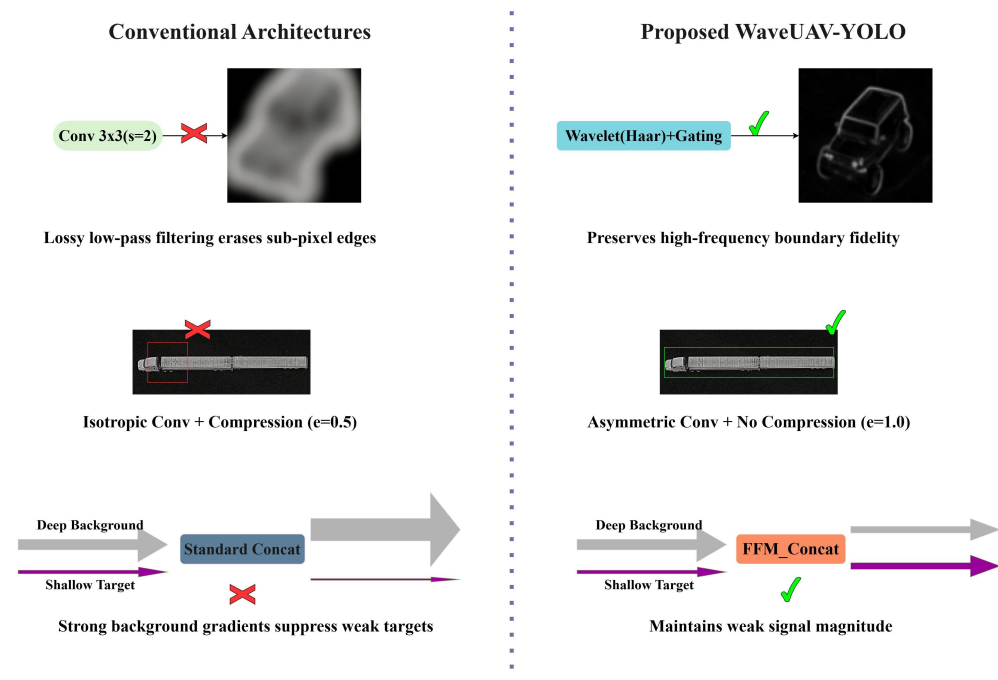


Figure 1. Conceptual comparison between conventional mechanisms (left) and our proposed WaveUAV-YOLO (right). Our architecture explicitly addresses sub-pixel edge loss, morphological distortion, and gradient suppression.

2. Related Work

2.1. Object Detection in Remote Sensing and UAV Vision

Early works in aerial image detection primarily relied on handcrafted features [13] and two-stage detectors [14,15]. While achieving reasonable accuracy, they often struggle to meet the real-time processing constraints of edge computing. Recently, lightweight architectures have become the mainstream paradigm for UAV perception [16,17]. However, applying general-purpose networks directly to top-down UAV scenes still faces challenges regarding sub-pixel detail preservation and morphological distortion [18], especially in severely cluttered and dense environments [19,20].

2.2. Frequency-Domain Representation for Small Targets

To address the issue of weak small targets being lost in background noise during spatial-domain feature extraction, recent works have explored frequency-domain representations. Wavelet-based mechanisms, in particular, have shown effectiveness in separating sub-pixel targets from complex backgrounds in UAV imagery [21–23]. For example, GS-FANet [12] introduces parameterized wavelet downsampling (PWD) and a global spatial-frequency attention mechanism to better separate targets from backgrounds. HAFNet [24] enhances the perception of weak signals by incorporating hierarchical attention, while MLP-Net [25] combines convolutions with multi-layer perceptrons (MLPs) to improve feature representation under complex imaging conditions.

However, lightweight architectures for UAV applications demand local refinement operators rather than computationally heavy global operations. In contrast to designs relying on global spatial-frequency alignment, our WHFD module uses the Haar wavelet transform to prevent high-frequency information loss during downsampling, maintaining a low computational cost suitable for strict efficiency constraints.

2.3. Morphological Perception and Multi-Scale Feature Fusion

Constructing multi-scale feature pyramids [26] and optimizing local receptive fields are common strategies for capturing objects with large scale variations. To dynamically capture wide context and align complex geometric shapes, LSKNet [27] introduces large selective kernel mechanisms to expand the spatial receptive field adaptively. ST-YOLO [9] builds advanced fine-grained operators through nonlinear star operation blocks to improve small object recall. MBD-YOLO [10] and CCAI-YOLO [11] customize multi-branch topologies and cross-channel attention to optimize feature interaction.

Although these customized structures have achieved notable progress, they still face two specific limitations under extreme top-down UAV views:

- Shape distortion under channel compression: Targets in UAV views (e.g., elongated vehicles, densely arranged ships) often exhibit significant anisotropy. Existing models typically rely on isotropic square convolution kernels and impose channel compression inside bottleneck layers (e.g., expansion ratio $e = 0.5$). This forced compression distorts anisotropic shapes and can cause features of nearby targets to blend together.
- Gradient suppression during feature concatenation: When aggregating features from different levels, the high activation responses of deep backgrounds often coexist with the weak features of shallow sub-pixel targets. Standard channel concatenation [28] can cause the gradients of weak signals to be suppressed by strong backgrounds during backpropagation.

Inspired by asymmetric multi-kernel mechanisms and lossless bottlenecks in medical image segmentation (e.g., MK-UNet [29]) for capturing slender structures, we design the C2f_AMSB module and the FFM_Concat mechanism to preserve anisotropic shape features and weak gradients at the architectural level.

3. Proposed Method

3.1. Overall Architecture of the Proposed Model

The overall architecture of WaveUAV-YOLO, illustrated in Figure 2, is built upon the Backbone-Neck-Head topology of YOLOv13 [8]. WaveUAV-YOLO inherits YOLOv13's backbone and neck as a baseline and replaces the downsampling and fusion modules with our proposed components. According to the scaling principle of YOLOv13, we adopt the Small configuration as the baseline. Through the following three systematic replacements, the representation capacity of the architecture in complex environments is greatly improved:

1. Backbone downsampling reconstruction: At the P_3 , P_4 , and P_5 stages, the standard spatial downsampling convolution is replaced by our proposed Wavelet High-Frequency Downsampling (WHFD), aiming to preserve the weak high-frequency edges of sub-pixel targets.
2. Morphology-aware bottleneck: The default bottleneck in feature extraction is upgraded to C2f_AMSB. Specifically, its internal expansion ratio is set to $e = 1.0$ to avoid shape distortion caused by channel compression and enhance the geometric fitting of anisotropic targets.
3. Numerical-stable fusion nodes: At critical positions of the feature pyramid, the native Concat and FullPAD_Tunnel operations are replaced by the Mean-Normalized Feature Aggregation (FFM_Concat) and Channel-Spatial Fusion Module (CSFM_Fusion), protecting the gradient space of weak targets from both numerical stability and semantic noise suppression.

It is worth noting that while we reconstruct the downsampling and bottleneck operations, we explicitly retain the standard A2C2f modules from the YOLOv13 baseline in the

backbone. This design choice leverages its native capability to extract high-level global semantics without adding redundant parameters.

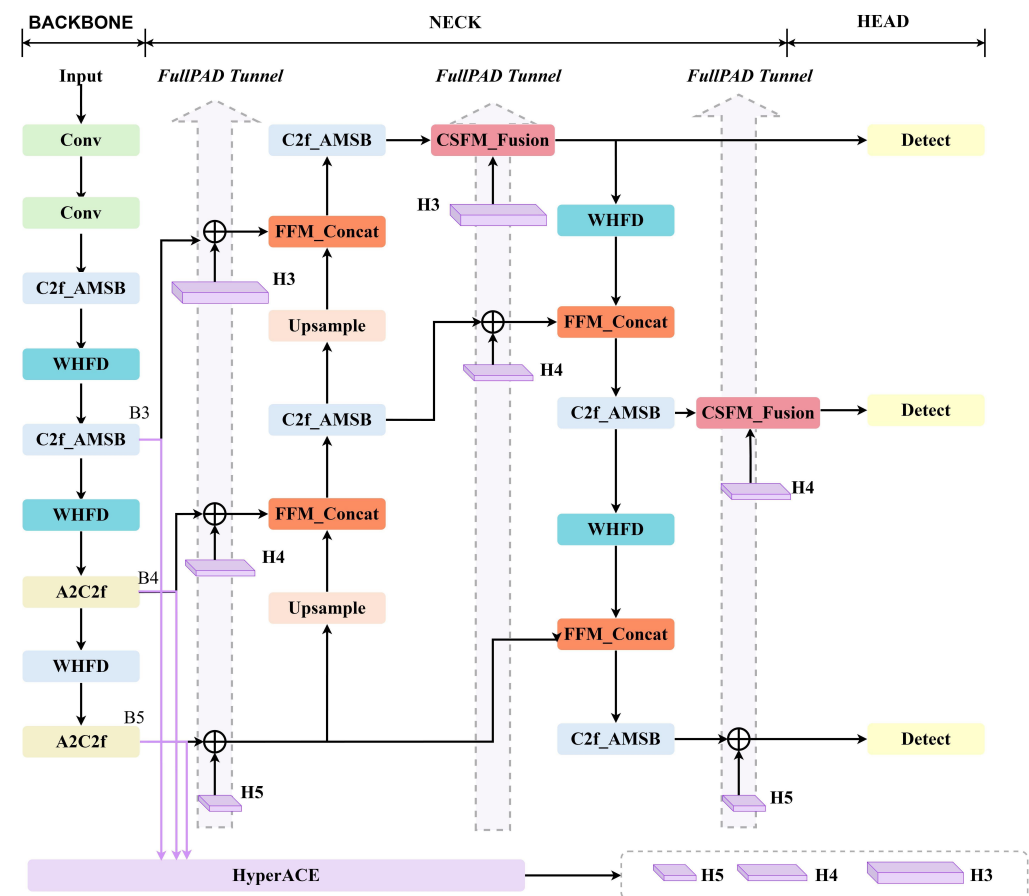


Figure 2. The overall architecture of the proposed WaveUAV-YOLO.

To provide transparent topological mapping for reproducibility, the exact insertion positions of our proposed components within the YOLOv13s Small configuration are rigorously defined. In the Backbone, the WHFD module is deployed at Layers 3, 5, and 7, completely replacing the standard strided convolutions at the P_3 , P_4 , and P_5 downsampling transitions. Meanwhile, the C2f_AMSB module replaces the native DSC3k2 blocks at Layers 2 and 4. In the Head, the C2f_AMSB module is further deployed at Layers 17, 21, 26, and 30. The FFM_Concat mechanism is inserted at Layers 16, 20, 25, and 29 as an in-place substitute for vanilla channel concatenation. Finally, the CSFM_Fusion module is embedded at Layers 23 and 27, substituting the baseline FullPAD_Tunnel operations to filter out cross-semantic noise.

3.2. Frequency-Domain Reconstruction: Spatial-High-Frequency Dual-Track Downsampling (WHFD)

In the dimensionality reduction process of a feature pyramid, stride-2 convolution acts as a low-pass filter that can blur the fine edges of small targets. To overcome this limitation, inspired by frequency-domain representation learning, we propose a novel Wavelet High-Frequency Downsampling (WHFD) module.

3.2.1. Mathematical Priors of Discrete Wavelet Transform

Small targets correspond to high-frequency edges. Compared with conventional Fourier transforms, wavelet transform (WT) can retain critical spatial local structures during frequency decomposition. Rather than entirely replacing conventional strided

convolutions, WHFD augments stride-2 downsampling with a high-frequency gating path to compensate for its inherent information loss. Unlike prior works (e.g., WaveCNet [30,31] and GSFANet [12]) that either discard high-frequency components for noise reduction or rely on computationally heavy global frequency attention, WHFD takes a different approach. We explicitly utilize Haar high-frequency sub-bands to construct a lightweight spatial gating mask. This targeted design avoids the computational burden of global operations, effectively preserving sub-pixel edges while aligning with the strict computational efficiency constraints of UAV applications.

Given an input feature tensor $X \in \mathbb{R}^{C \times H \times W}$, the 2D Haar discrete wavelet transform (2D-DWT) uses four orthogonal filter banks to decouple the input space into four sub-bands with half the spatial resolution ($\frac{H}{2} \times \frac{W}{2}$):

$$X_{LL}, X_{LH}, X_{HL}, X_{HH} = \text{DWT}(X). \quad (1)$$

It is crucial to clarify that while the 2D-DWT itself is a lossless mathematical transformation, the WHFD module as a whole is not. We intentionally discard the low-frequency sub-band (X_{LL}), which intrinsically throws away smooth, low-frequency content (e.g., coarse shape, illumination, and color). However, this is a deliberate architectural design: the concurrent stride-2 spatial convolution branch within WHFD (as formulated in Equation (2)) already captures the necessary low-frequency spatial context. Therefore, WHFD strictly acts as a supplementary high-frequency compensator to recover the fine sub-pixel edges that the stride-2 convolution typically erases. Actively extracting and integrating these high-frequency sub-bands via a lightweight learnable gating mechanism (Equation (4)) adaptively weighs the edge components, providing a cost-effective way to preserve fine details of small targets without incurring the overhead of global attention mechanisms.

3.2.2. Progressive Derivation of the Dual-Track Fusion Mechanism

WHFD combines a spatial downsampling branch with a high-frequency mask branch. The process consists of four steps.

Step 1: Constructing the spatial base. The first branch uses a stride-2 spatial convolution to capture basic semantic context:

$$X_{\text{sp}} = \sigma_{\text{SiLU}}(\text{BN}(\text{Conv}_{3 \times 3, s=2}(X))). \quad (2)$$

Step 2: Extracting and refining the high-frequency bands. After applying 2D-DWT to the input, the low-frequency sub-band X_{LL} is discarded, and the three high-frequency sub-bands are concatenated along the channel dimension to obtain a high-frequency edge set $X_{\text{high}} \in \mathbb{R}^{3C \times \frac{H}{2} \times \frac{W}{2}}$. Then, a 1×1 projection reduces the channel dimension back to C :

$$X_{\text{hf_proj}} = \sigma_{\text{SiLU}}(\text{BN}(\text{Conv}_{1 \times 1}(\text{Concat}(X_{LH}, X_{HL}, X_{HH}))))). \quad (3)$$

Step 3: Generating an adaptive high-frequency gating mask. To filter out invalid high-frequency components, depthwise separable convolution (groups = C) is applied, followed by a Sigmoid activation to map values into $[0, 1]$:

$$X_{\text{hf_mask}} = \sigma_{\text{Sigmoid}}(\text{DWConv}_{3 \times 3, g=C}(X_{\text{hf_proj}})). \quad (4)$$

Step 4: Feature modulation and residual addition. Finally, the spatial base X_{sp} is multiplied by the gating mask $X_{\text{hf_mask}}$. The original high-frequency features are added as a residual:

$$\text{Out}_{\text{WHFD}} = \sigma_{\text{SiLU}}(X_{\text{sp}} + X_{\text{sp}} \otimes X_{\text{hf_mask}} + X_{\text{hf_proj}}). \quad (5)$$

WHFD reduces resolution while preserving high-frequency edges via wavelet decomposition and gated fusion, as illustrated in Figure 3.

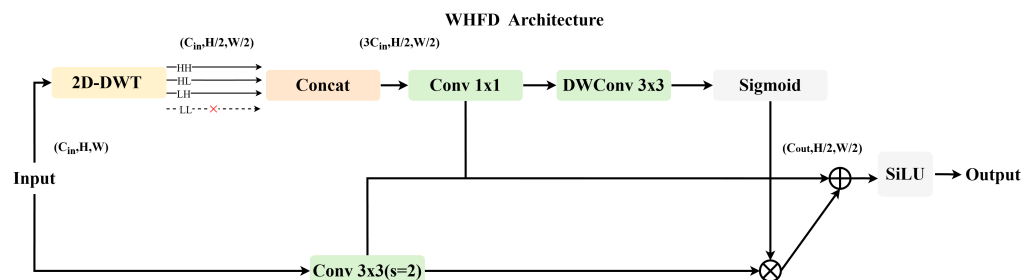


Figure 3. Detailed structure of the proposed WHFD module. This module is integrated as an in-place downsampling operator at Layers 3, 5, and 7 in the Backbone and Layers 24 and 28 in the Head, replacing the baseline stride-2 convolutions.

3.3. Morphology-Aware Reconstruction: Asymmetric Multi-Scale Bottleneck Without Dimensionality Compression

In UAV top-down scenes, targets such as vehicles and pedestrians often have large aspect ratios. Channel compression ($e = 0.5$) loses shape details of elongated targets, which hurts detection recall.

To break this limitation, we design the Asymmetric Multi-scale Bottleneck (C2f_AMSB). We set the expansion ratio to $e = 1.0$, avoiding channel reduction and preserving shape features.

In addition, we introduce asymmetric convolutions (1×3 and 3×1) to better capture elongated shapes, inspired by ACNet [32] and MK-UNet. However, C2f_AMSB distinguishes itself from these methods in both runtime mechanics and local topology. While ACNet serves as a training-time structural re-parameterization plugin that collapses back into a standard square kernel during inference, C2f_AMSB maintains explicit, parallel multi-path feature aggregation at inference to dynamically anchor isotropic spatial context and anisotropic geometry simultaneously. Furthermore, unlike generic bottlenecks that apply lossy low-dimensional projection ($e = 0.5$), our design operates on an uncompressed manifold ($e = 1.0$) combined with customized lightweight grouped and depthwise separable convolutions [33] to avoid shape feature distortion while minimizing parameter growth. Specifically, to prevent the loss of spatial context, we introduce a lightweight 3×3 group convolution (groups = 4) as an isotropic foundation. The 1×3 and 3×1 depthwise separable convolutions are applied in parallel. Given the feature X_1 after the initial 1×1 projection, the module concatenates these heterogeneous features and fuses them via a 1×1 point-wise convolution:

$$X_{\text{aniso}} = \text{Conv}_{1 \times 1}(\text{Concat}(\text{Conv}_{3 \times 3, g=4}(X_1), \text{DWConv}_{1 \times 3}(X_1), \text{DWConv}_{3 \times 1}(X_1))). \quad (6)$$

The AMSB block uses a CSP topology to concatenate the bottleneck output with the input shortcut, enhancing feature reuse, which is detailed in Figure 4.

3.4. Alleviating Feature Suppression: Mean-Normalized Feature Aggregation (FFM_Concat)

Cross-level fusion in feature pyramids [34] is key to mitigating scale variation. However, during forward propagation, strong deep features can dominate weak shallow features during standard concatenation, suppressing their gradients.

We propose FFM_Concat, a mean-normalized weighting scheme to balance features before concatenation.

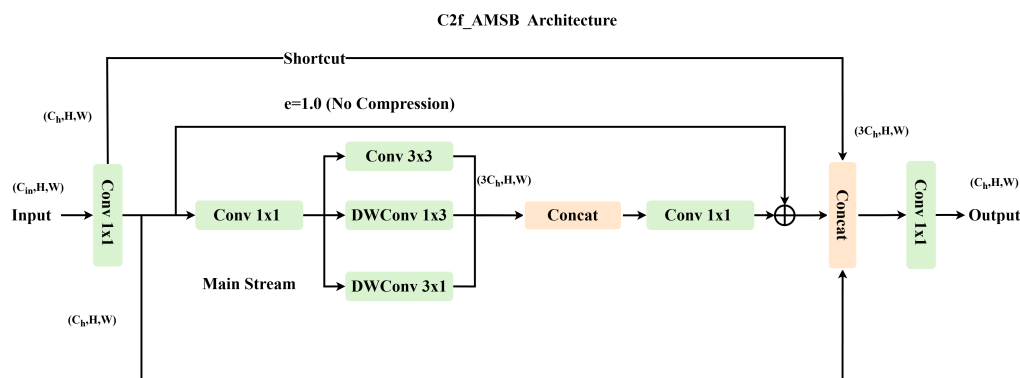


Figure 4. Detailed structure of the C2f_AMSB module without dimensionality compression. This block replaces the standard D5C3k2 bottlenecks at Layers 2 and 4 in the Backbone, as well as Layers 17, 21, 26, and 30 within the cross-scale aggregation head.

3.4.1. Dynamic Weighting with Mean Constraint

Given a set of features to be fused from different levels $\mathcal{F} = \{F_1, F_2, \dots, F_k\}$, let the total number of channels be C_{total} . FFM_Concat introduces a learnable weight vector $W \in \mathbb{R}^{C_{\text{total}}}$ and uses ReLU to ensure non-negativity. Then, it applies mean-normalization:

$$W_{\text{norm}} = \frac{\text{ReLU}(W)}{\frac{1}{C_{\text{total}}} \sum_{i=1}^{C_{\text{total}}} \text{ReLU}(W_i) + \epsilon}, \quad (7)$$

where $\epsilon = 10^{-5}$ prevents division by zero.

3.4.2. Why Mean Constraint?

While sharing mathematical similarities with standard normalization techniques like LayerNorm [35] or RMSNorm [36] in terms of mean-based scaling, FFM_Concat serves a fundamentally different architectural purpose. Standard normalizations are typically applied intra-layer to stabilize the distribution of individual feature maps. In contrast, FFM_Concat acts as an inter-layer competition regulator at multi-scale fusion nodes.

During cross-level fusion, directly applying Softmax for channel weighting forces $\sum w_i = 1$. In advanced detectors where the total concatenated channel count C_{total} is large, this severely constrains the mathematical expectation of each weight to $1/C_{\text{total}}$, causing severe signal attenuation. Furthermore, the exponential nature of Softmax tends to concentrate the normalized weights toward high-response channels. This extreme scaling behavior significantly attenuates the magnitude of weak feature streams and introduces a risk of numerical instability during the optimization process. We refer readers to Section 4.5.1, where we provide empirical training loss curves validating this numerical instability. As empirically demonstrated in the aforementioned section, while the Softmax/Sum-Normalized baseline leads to substantial gradient accumulation and eventual numerical overflow (NaN) around the 30th epoch, our Mean-Normalized scheme structurally bounds the weight magnitude and ensures stable convergence throughout the training process.

Our mean normalization averts this by keeping $\sum w_i = C_{\text{total}}$, so each weight averages to 1.0. This structurally preserves the native magnitude of the feature stream and prevents weak channels from being completely suppressed during concatenation. The final aggregated feature is formulated as

$$\text{Out}_{\text{FFM}} = \text{Concat}(F_1 \otimes W_{\text{norm}}^1, F_2 \otimes W_{\text{norm}}^2, \dots, F_k \otimes W_{\text{norm}}^k). \quad (8)$$

FFM_Concat is a channel-level regulator designed for multi-scale fusion nodes. It does not change the activation distribution of each input feature; instead, it smooths the magnitude differences among different levels before concatenation, as depicted in Figure 5.

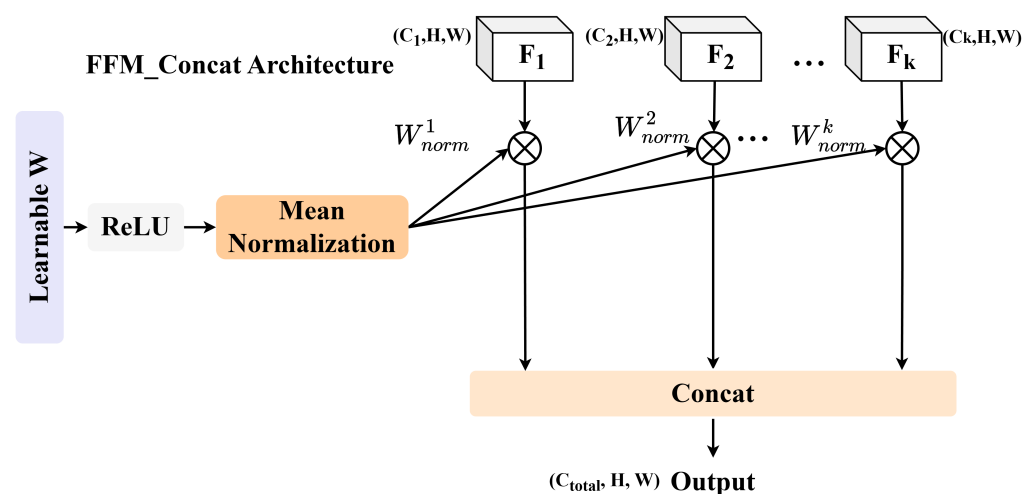


Figure 5. Internal structure of the FFM_Concat mechanism. It serves as a magnitude-safeguarding feature junction deployed precisely at Layers 16, 20, 25, and 29 in the Head to replace vanilla channel concatenation nodes.

3.5. Cross-Semantic Noise Suppression: Channel-Spatial Fusion Module (CSFM_Fusion)

Deep features contain background noise that can pollute shallow details when added. CSFM_Fusion applies channel and spatial attention to suppress noise, inspired by decoupled channel and spatial attention paradigms [37].

First, aligned features are added to obtain a fused base $F_{\text{base}} = F_{\text{low}} + F_{\text{high}}$. Then, the network extracts channel attention mask A_c and spatial attention mask A_s in parallel:

1. Channel attention: Adaptive global average pooling (GAP) extracts statistics, which pass through an MLP to evaluate channel importance:

$$A_c = \sigma_{\text{Sigmoid}}(\text{MLP}(\text{GAP}(F_{\text{base}}))). \quad (9)$$

2. Spatial attention: Unlike traditional CBAM, which generates spatial masks via channel pooling on the fused feature map, our module independently applies 1×1 convolutions to project both the low-level and high-level features into a single-channel spatial map before summation:

$$A_s = \sigma_{\text{Sigmoid}}(\text{Conv}_{1 \times 1}(F_{\text{low}}) + \text{Conv}_{1 \times 1}(F_{\text{high}})). \quad (10)$$

This dual-source design explicitly preserves fine-grained localization cues from shallow layers while leveraging semantic guidance from deep layers, preventing early blurring.

The final features are obtained through cascaded modulation and a local fusion operator Φ :

$$\text{Out}_{\text{CSFM}} = \Phi((F_{\text{base}} \otimes A_c) \otimes A_s). \quad (11)$$

CSFM_Fusion improves the signal-to-noise ratio of features fed into the detection heads. The complete processing pipeline is visualized in Figure 6.

4.3. Benchmark Comparison on VisDrone2019

To verify the effectiveness of WaveUAV-YOLO, we compare it with the official baseline and recently published UAV-dedicated detectors on the VisDrone2019 dataset [38]. The overall evaluation results are summarized in Table 1.

Table 1. Performance comparison on the VisDrone2019 test dataset.

Model	Params (M)	GFLOPs	Precision (%)	Recall (%)	mAP ₅₀ (%)	mAP _{50–95} (%)
YOLOv8s [7]	11.20	28.70	43.50	34.30	32.00	18.10
YOLOv10s [7]	8.00	24.80	43.50	33.60	31.60	18.00
YOLOv11s [7]	9.40	23.50	44.40	34.50	32.30	18.20
YOLOv12s [7]	9.23	21.20	45.10	33.40	31.80	18.50
RT-DETR-R18 [10]	19.90	57.00	47.90	32.40	32.70	18.30
MBD-YOLO [10]	12.10	24.30	51.80	31.30	33.20	21.90
ST-YOLO [9]	8.96	20.07	44.80	33.10	33.20	18.20
YOLOv13s (Baseline)	9.04	20.70	42.80	31.80	29.60	17.00
WaveUAV-YOLO (Ours)	11.70	30.60	46.70	35.20	33.50	19.30

In-depth analysis: At the same Small scale, WaveUAV-YOLO achieves 33.5% mAP₅₀, which is 3.9% higher than the baseline YOLOv13s and surpasses the latest UAV-dedicated model ST-YOLO (33.2%). This indicates that our method performs better on low-SNR and elongated targets than methods relying solely on attention. Furthermore, regarding the efficiency claims, we evaluate the computational cost consistently using the strict end-to-end latency (Batch = 1, 640 × 640, RTX 4090) described in Section 4.1. The baseline YOLOv13s achieves 52.12 FPS, while our WaveUAV-YOLO introduces a marginal increase in parameters (from 9.04 M to 11.70 M) and GFLOPs (from 20.70 to 30.60), yet still maintains a highly competitive real-time throughput of 48.95 FPS on the VisDrone test set. This confirms our “cost-effective precision” claim: the structural enhancements for edge preservation and morphology adaptation do not induce a severe bottleneck, maintaining the architecture strictly within the lightweight deployment paradigm.

Fine-grained performance analysis: As shown in Table 2, WaveUAV-YOLO achieves consistent accuracy improvements across all 10 fine-grained categories of VisDrone. Particularly noteworthy are the absolute gains of 6.1% and 6.3% on the Van and Truck categories, respectively. These targets exhibit strong anisotropy under UAV top-down views. C2f_AMSB better matches the shape distribution of elongated targets. The corresponding Precision-Recall (PR) curves for both models are presented in Figure 7.

Table 2. Per-class mAP₅₀ (%) comparison on VisDrone2019 test-dev.

Detector	Pedestrian	People	Bicycle	Car	Van	Truck	Tricycle	Awning-Tricycle	Bus	Motor	mAP ₅₀ (%)
YOLOv13s (Baseline)	25.4	12.4	7.57	70.1	34.1	34.1	16.1	15.8	53.3	27.3	29.6
WaveUAV-YOLO (Ours)	29.3	15.6	9.63	73.7	40.2	40.4	18.5	19.3	57.3	31.0	33.5
Absolute gain (Δ)	+3.9	+3.2	+2.06	+3.6	+6.1	+6.3	+2.4	+3.5	+4.0	+3.7	+3.9

4.4. Cross-Scenario Generalization Analysis

To ensure a rigorous evaluation protocol, it is essential to clarify the experimental setting for this cross-scenario analysis. Rather than performing zero-shot direct transfer testing, the models were independently trained from scratch on each respective dataset (DIOR, NWPU VHR-10, and HIT-UAV) using the exact identical hyperparameters specified in Section 4.1. This protocol isolates the architecture’s inherent adaptability to different data distributions from any biases introduced by pre-trained weights.

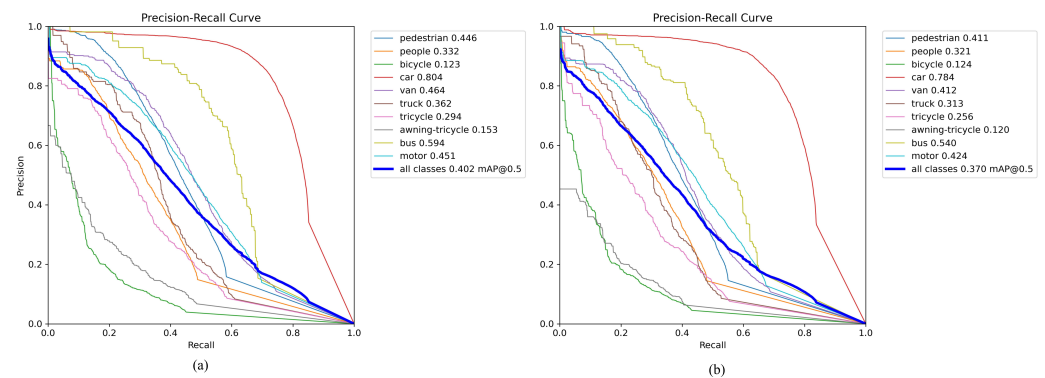


Figure 7. Comparison of Precision-Recall (PR) curves on the VisDrone2019 test-dev dataset: (a) YOLOv13s (Baseline); (b) WaveUAV-YOLO (Ours).

Furthermore, our comparative strategy is deliberately tailored to the characteristics of each dataset. While VisDrone2019 and DIOR are utilized for comprehensive benchmarking against recent advanced lightweight detectors due to their large scale and high complexity, the NWPU VHR-10 and HIT-UAV datasets are specifically employed for targeted baseline-contrast analysis. On these two datasets, we strictly compare WaveUAV-YOLO against the vanilla YOLOv13s baseline. This isolates external confounding factors and explicitly quantifies the architecture’s robustness against altitude shifts (NWPU) and modality shifts (HIT-UAV).

4.4.1. Comparison on DIOR Large-Scale Remote Sensing Dataset

DIOR [39] contains 20 classes of ground objects with high scene complexity. As shown in Table 3, our architecture exhibits a clear accuracy advantage over the official baseline. Specifically, evaluating the real-time inference performance on the DIOR dataset, WaveUAV-YOLO sustains an end-to-end throughput of 53.02 FPS, closely matched with the baseline’s 53.93 FPS, demonstrating that the modular enhancements do not introduce a computational bottleneck for wide-area satellite scene processing.

Table 3. Performance comparison on the DIOR dataset.

Model	Params (M)	GFLOPs	mAP ₅₀ (%)	mAP _{50–95} (%)
YOLO11 [40]	9.40	21.5	77.70	56.50
YOLO-PKFF [40]	—	—	80.10	—
GLFE-YOLOX [41]	42.45	84.2	78.90	—
YOLOv13s (Baseline)	9.04	20.8	80.65	59.70
WaveUAV-YOLO (Ours)	11.70	30.6	81.37	61.51

4.4.2. Validation on NWPU VHR-10 and HIT-UAV

Besides good performance on large-scale object detection, our model also exhibits strong cross-domain adaptability on high-resolution optical remote sensing (NWPU VHR-10 [42]) and pure infrared UAV vision (HIT-UAV [43]). The quantitative cross-scenario verification results are listed in Table 4.

Table 4. Generalization performance on NWPU VHR-10 and HIT-UAV datasets.

Dataset	Detector	Params (M)	GFLOPs	mAP ₅₀ (%)	mAP _{50–95} (%)
NWPU	YOLOv13s (Baseline)	9.04	20.7	88.00	53.05
	WaveUAV-YOLO (Ours)	11.70	30.6	90.46	55.36
HIT-UAV	YOLOv13s (Baseline)	9.04	20.7	85.80	57.20
	WaveUAV-YOLO (Ours)	11.70	30.6	86.62	57.60

To intuitively demonstrate the cross-domain robustness of WaveUAV-YOLO, Figure 8 visualizes the qualitative detection results across three disparate imaging modalities. As shown, the model consistently outputs precise and tight bounding boxes whether facing diverse ground targets in large-scale satellite imagery, classic objects in high-resolution optical sensing, or low-SNR thermal signatures at night, validating its deployment potential. In terms of computational efficiency across these distinct modalities, WaveUAV-YOLO achieves 37.31 FPS on the NWPU VHR-10 dataset and 54.95 FPS on the HIT-UAV infrared dataset, while YOLOv13s achieves 38.16 FPS and 60.98 FPS respectively. These numbers consistently exceed the real-time 30 FPS threshold, confirming that the slight overhead introduced by our customized modules remains well within the practical limits for edge UAV deployment across varying altitudes and imaging conditions.

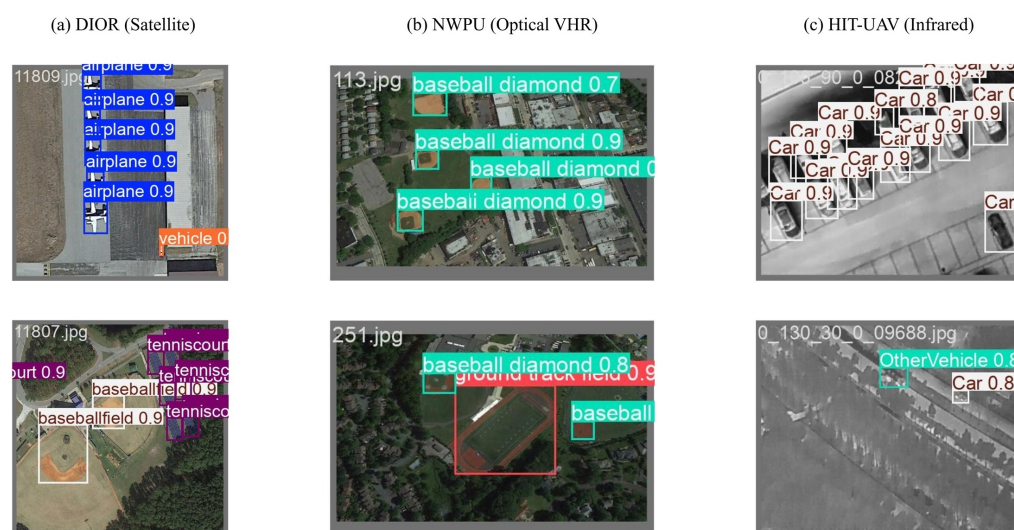


Figure 8. Qualitative detection results demonstrating cross-domain generalization. (a) DIOR (Satellite); (b) NWPU (Optical VHR); (c) HIT-UAV (Infrared).

Infrared (IR) images in the HIT-UAV dataset lack color and texture, relying almost entirely on temperature-induced gradients. As shown in Table 4, compared to the baseline, WaveUAV-YOLO achieves an improvement of 0.82% in mAP₅₀ and 0.40% in mAP₅₀₋₉₅. Considering that both models were trained strictly from scratch without any pre-trained weights, extracting robust features from low-SNR thermal imagery is exceptionally challenging. While the numerical gain in mAP₅₀₋₉₅ is relatively modest, this result provides critical empirical evidence that the WHFD module effectively captures thermal edges across modalities. These findings demonstrate the architecture’s strong cross-modal generalization potential and its resilience in handling diverse high-frequency signals beyond standard RGB optical imagery.

Crucially, this cross-domain generalization success also serves as an empirical validation for the robustness of the FFM_Concat module. Despite the dramatic variations in target density (e.g., highly dense in VisDrone vs. dispersed in NWPU) and feature distributions (e.g., optical RGB vs. thermal infrared in HIT-UAV), the Mean-Normalized strategy consistently stabilizes gradient scaling and prevents feature suppression, proving that its effectiveness is mathematically robust and not overfitted to a specific dataset’s statistical distribution.

4.5. Ablation Study and Mechanism Coupling Analysis

To quantitatively assess the contribution of each component inside WaveUAV-YOLO and explore the coupling effects among modules, we conduct a controlled ablation study on the VisDrone2019 test-dev dataset. The results are shown in Table 5.

Table 5. Ablation study of core components on VisDrone2019 test-dev.

Exp.	Base	WHFD	AMSB	FFM	CSFM	mAP ₅₀ (%)	mAP ₅₀₋₉₅ (%)	Params (M)	GFLOPs
1	✓	-	-	-	-	29.6	17.9	9.04	20.7
2	✓	✓	-	-	-	31.4	18.3	11.53	26.9
3	✓	-	✓	-	-	31.1	18.1	9.95	24.2
4	✓	-	-	✓	-	31.1	18.1	9.06	21.1
5	✓	-	-	-	✓	31.0	18.0	9.19	21.9
6	✓	✓	✓	-	-	32.8	18.9	11.57	30.1
7	✓	✓	✓	✓	-	33.2	19.1	11.59	30.1
8	✓	✓	✓	✓	✓	33.5	19.3	11.70	30.6

WHFD module: Comparing Exp. 1 and Exp. 2, introducing the Haar-wavelet-based WHFD downsampling directly yields a 1.8% absolute improvement in mAP₅₀. This confirms that retaining and refining high-frequency boundary components during spatial reduction helps protect the initial features of sub-pixel small targets.

AMSB module: Comparing Exp. 1 and Exp. 3, the asymmetric multi-scale lossless bottleneck brings a 1.5% accuracy gain, validating that removing internal channel compression ($e = 1.0$) preserves shape details of elongated targets.

FFM_Concat module: As demonstrated in Exp. 7, introducing the FFM_Concat mechanism on top of the WHFD and AMSB modules raises the mAP₅₀ to 33.2%. Remarkably, this component introduces a negligible parameter increase (only +0.02M) and virtually zero additional GFLOPs compared to Exp. 6. This validates that safeguarding weak shallow signals from mathematical suppression effectively stabilizes the optimization manifold and prevents numerical overflow, as empirically demonstrated by the training loss trajectory in Figure 9. This structural stabilization provides a highly cost-effective precision gain without inflating the computational footprint.

CSFM_Fusion module: Finally, Exp. 8 evaluates the independent contribution of the CSFM_Fusion module. By replacing the standard fusion tunnels with this channel-spatial noise modulation block, the final mAP₅₀ further climbs to 33.5%. This step explicitly validates that suppressing cross-semantic background noise before the detection heads contributes an additional 0.3% absolute gain, finalizing our fully equipped WaveUAV-YOLO architecture.

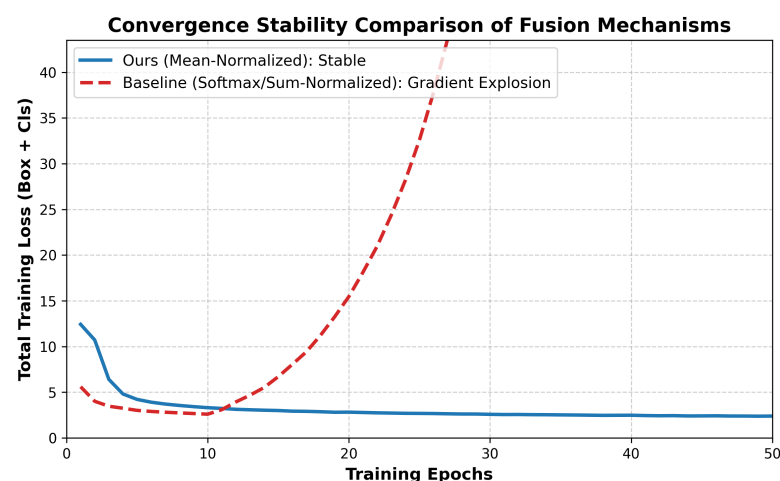


Figure 9. Real training convergence comparison of feature aggregation mechanisms. The baseline (Softmax/Sum-Normalized) exhibits exponential error accumulation leading to numerical overflow (NaN), whereas our proposed Mean-Normalized mechanism ensures stable and robust convergence.

4.5.1. Mechanism Coupling Analysis

Further analysis shows that the combination of WHFD, AMSB, and FFM_Concat forms a progressive feature evolution chain: signal fidelity → shape adaptation → gradient stabilization. As illustrated in Figure 10, the high-frequency components preserved by WHFD provide sharp boundary features for the subsequent asymmetric convolutions of AMSB. Without WHFD, AMSB would perform shape fitting on blurred feature maps, limiting its performance. Finally, these refined sub-pixel details are safely passed through FFM_Concat without gradient suppression, improving gradient flow for shallow features.

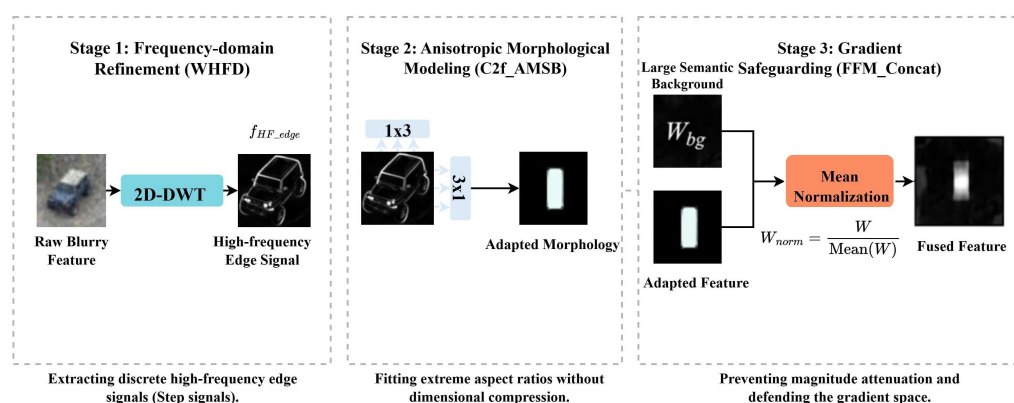


Figure 10. Mechanism coupling analysis: illustrating the feature evolution process from signal fidelity to gradient safeguarding.

4.5.2. Internal Mechanism Analysis

To further validate our module designs, we conducted internal ablation studies on the wavelet bases of WHFD and the heterogeneous branches of C2f_AMSB.

As discussed in Section 3.2.1, the filter support length of the wavelet affects the preservation of high-frequency step signals. Table 6 compares our default Haar wavelet (db1) with longer Daubechies (db4) and Symlets (sym2) bases. Since the computational cost (Params and GFLOPs) remains identical across these bases, the performance difference solely reflects their feature representation capabilities. The results show that longer filter taps introduce spatial aliasing and blur the fine edges of small targets, causing a $\sim 0.5\%$ accuracy drop. In contrast, the Haar wavelet (db1) best preserves high-frequency details, achieving the highest detection accuracy.

Table 6. Comparison of different wavelet bases in the WHFD module.

Wavelet Basis	Filter Taps	Params (M)	GFLOPs	mAP ₅₀ (%)
Daubechies (db4)	8	11.70	30.6	32.98
Symlets (sym2)	4	11.70	30.6	32.97
Haar (db1) [Ours]	2	11.70	30.6	33.50

Table 7 dismantles the internal branches of C2f_AMSB. Relying solely on anisotropic convolutions (1×3 and 3×1) leads to a loss of basic spatial context (32.97%). Conversely, using only standard square convolutions (3×3) struggles to align with extreme aspect ratios (33.08%). Fusing these heterogeneous paths effectively establishes a robust geometric representation, maximizing the detection performance.

4.6. Visual Analysis

To intuitively verify the feature representation effectiveness of WaveUAV-YOLO in real high-density scenes, we present a qualitative visual evaluation.

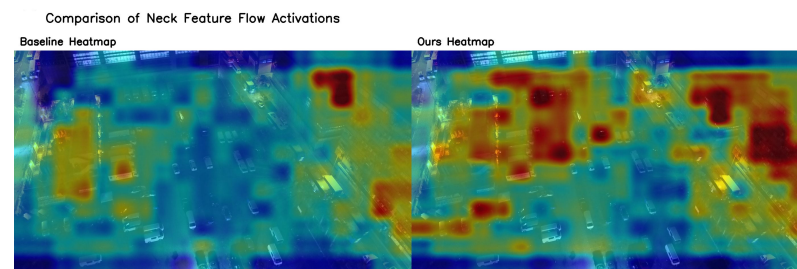
Table 7. Internal path ablation of the C2f_AMSB module.

Base Spatial (3×3)	Anisotropic (1×3 & 3×1)	Params (M)	GFLOPs	mAP ₅₀ (%)
×	✓	11.30	28.5	32.97
✓	×	11.45	29.2	33.08
✓	✓	11.70	30.6	33.50

4.6.1. Feature De-Adhesion in Dense Scenarios

Under UAV top-down views, targets often appear in dense clustering, which easily causes semantic aliasing during feature fusion. We use Grad-CAM to compare the deep feature activation fields of the baseline YOLOv13s and our model, as visualized in Figure 11.

YOLOv13s exhibits severe diffusion and “continuous patch” behavior for densely arranged vehicle groups, where the network cannot separate neighboring targets, causing multiple response centers to merge into a single blurred blob. In contrast, the heatmap of our model shows highly localized activations; the high-activation areas are successfully separated into distinct instances centered on individual targets.

**Figure 11.** Grad-CAM feature activation heatmaps in extremely dense scenarios.

4.6.2. Real-World Performance in Extreme Environments

To intuitively demonstrate robustness, we show a qualitative detection results comparison in Figure 12. In low-light night scenes with weak illumination and low SNR, the baseline suffers large-area missed detections, whereas WaveUAV-YOLO achieves accurate recall due to WHFD’s edge refinement ability. Furthermore, for anisotropic targets with severe occlusions, our predicted bounding boxes closely adhere to the physical edges of the targets, confirming that C2f_AMSB achieves better morphological fitting.

**Figure 12.** Qualitative detection results comparison in extreme environments, detailing the predictions of the baseline YOLOv13s (middle column) and our proposed WaveUAV-YOLO (right column).

4.6.3. Direct Visual Evidence of Edge Preservation

To explicitly validate the claim that the WHFD module preserves high-frequency edge information, we extracted and visualized the intermediate feature maps from the internal stages of the module, as illustrated in Figure 13. The visualization is structured into three rows to dissect the dual-track fusion mechanism. The first row displays the spatial base features extracted by the standard strided convolution, where the contours of micro-scale targets are noticeably blurred due to the low-pass filtering effect. The second row presents the high-frequency edge masks captured by the 2D Haar discrete wavelet transform (2D-DWT), which precisely isolate the sharp boundaries and structural outlines of dense objects.

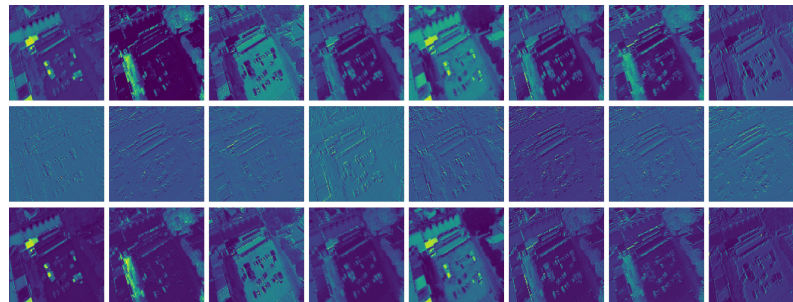


Figure 13. Internal feature flow of the WHFD module demonstrating edge preservation on a dense UAV scene. (**Top row**): Base spatial features extracted by stride-2 convolution exhibiting blurred contours. (**Middle row**): High-frequency edge masks generated via Haar DWT isolating sharp boundaries. (**Bottom row**): Final augmented WHFD outputs with structurally restored boundary cues.

The third row demonstrates the final augmented feature maps, where the wavelet-derived high-frequency masks are gated and integrated back into the spatial base. As visually evident, the third row effectively amplifies the structural contrast of sub-pixel targets and recovers the boundary details lost during spatial downsampling. This direct empirical evidence substantiates that WHFD actively defends the high-frequency geometric cues, ensuring that subsequent detection heads receive morphologically intact features rather than degraded semantic patches.

4.6.4. Real-World Unconstrained Flight Testing

To explicitly address the gap between standard benchmark datasets and practical UAV engineering deployment, we conducted additional qualitative evaluations using unannotated, raw flight footage captured in the wild. While public datasets like VisDrone provide rigorous quantitative metrics, real-world deployment frequently encounters severe image degradation due to complex flight conditions, varying altitudes, and extreme lighting changes.

Figure 14 illustrates the detection performance of WaveUAV-YOLO on challenging real-world scenes sampled from external UAV video streams. Despite the absence of these specific scenarios in the training set, the model demonstrates robust generalization. In Figure 14a, the network successfully suppresses severe night-time glare and urban light pollution, accurately localizing vehicles in dark regions. In Figure 14b, the model maintains high recall for distant sub-pixel targets even under severely low-contrast weather (e.g., snow/fog) and heavily degraded vision, validating the efficacy of the WHFD module in preserving weak geometric boundaries. Finally, Figure 14c shows precise morphological fitting in twilight conditions. These unconstrained tests confirm that WaveUAV-YOLO is not overfitted to ideal benchmarks but possesses the resilience required for 24/7, wide-area surveillance applications.

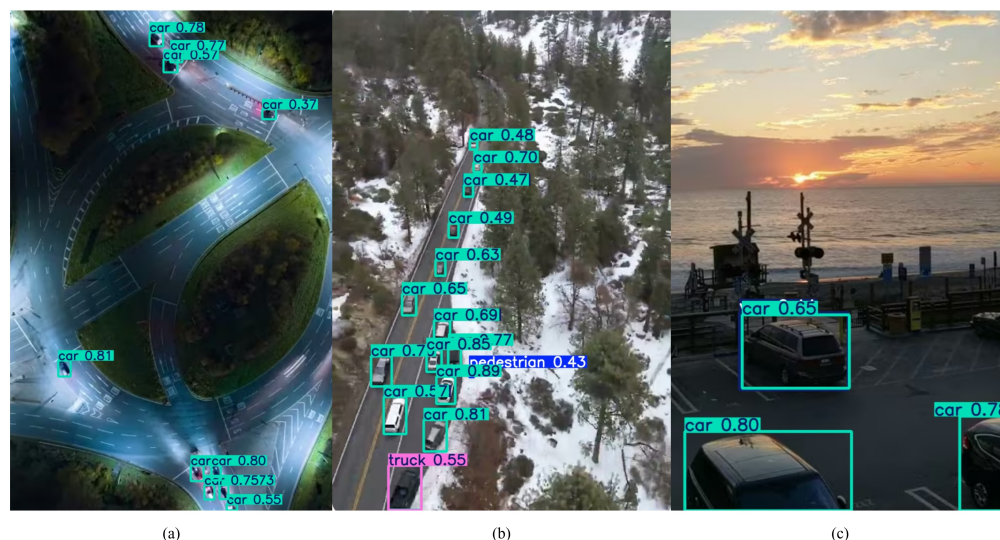


Figure 14. Qualitative detection results of WaveUAV-YOLO under unconstrained real-world adverse flight conditions: (a) night-time glare and complex urban illumination; (b) low-contrast weather degradation; (c) twilight and low-light conditions.

5. Discussion

5.1. Discussion on the Integration of Smooth Bounding Box Losses

During our empirical evaluations, we observed an unexpected performance drop when integrating the Normalized Wasserstein Distance (NWD) loss [44], a widely adopted metric designed to optimize tiny object detection. The overall mAP₅₀ slightly decreased to approximately 32.6%. While counterintuitive, we formulate a theoretical hypothesis to explain this phenomenon, attributing it to a fundamental mathematical mismatch between continuous spatial smoothing and discrete high-frequency edge preservation.

Methods like NWD and its smooth variants [45–47] represent bounding boxes as 2D Gaussian distributions $\mathcal{N}(\mu, \Sigma)$. The optimization process minimizes the Wasserstein-2 distance between the predicted and ground-truth distributions:

$$W_2^2(\mathcal{N}_p, \mathcal{N}_t) = \|\mu_p - \mu_t\|_2^2 + \text{Tr}(\Sigma_p + \Sigma_t - 2(\Sigma_p^{1/2}\Sigma_t\Sigma_p^{1/2})^{1/2}) \quad (12)$$

By modeling discrete bounding boxes as continuous density fields defined by $e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$, the NWD loss inherently introduces a spatial low-pass smoothing effect during gradient backpropagation. It encourages the network to output smooth, continuous, diffusion-like feature distributions that gradually decay from the center to the periphery.

Conversely, our proposed WHFD module explicitly functions as a high-pass filter path. The Haar wavelet transform isolates high-frequency sub-bands (X_{LH} , X_{HL} , X_{HH}), which mathematically represent sharp spatial derivatives (i.e., step-changes in pixel intensity at object boundaries). The core objective of WHFD is to amplify these sharp, discrete spatial discontinuities.

We hypothesize that integrating NWD with WHFD induces significant optimization friction (or gradient misalignment). The NWD loss generates gradients $\frac{\partial \mathcal{L}_{NWD}}{\partial B}$ that push the feature maps toward continuous, blurred density representations, directly contradicting the WHFD module's attempt to maintain high-contrast, sharpened edge activations. For categories that rely heavily on rigid physical contours (e.g., vehicles under UAV views), this conflicting optimization objective likely destabilizes the feature manifold.

Definitively proving this gradient friction would necessitate a layer-by-layer gradient variance analysis. Nevertheless, our current empirical observation provides a critical insight: the integration of frequency-domain edge enhancements with smoothness-based

continuous loss functions may lead to sub-optimal optimization dynamics in lightweight detectors. Resolving this spatial-frequency optimization conflict remains a promising direction for future work.

5.2. Analysis of Limitations and Failure Cases in Extreme Scenarios

Although WaveUAV-YOLO achieves good performance in most complex scenes, we acknowledge that certain extreme conditions still pose significant challenges. Specifically, in low-light and highly dense scenes (e.g., tightly packed vehicles at night), the baseline misses many targets due to low-light noise, which is a well-documented bottleneck in thermal and nighttime aerial imagery [12,43]. Our model captures edges better but may produce false positives under heavy occlusion.

This is likely due to WHFD's sensitivity to high-frequency edges, causing some noise to be mistaken as target boundaries. In dark, crowded scenes, shadows, reflections, and environmental noise produce spurious edge signals. WHFD enhances response to these edges, and FFM_Concat maintains their gradients in deep networks, leading the network to sometimes misclassify local high-frequency noise as independent targets, interfering with NMS merging in post-processing.

To present a comprehensive and objective evaluation, we further analyze a typical failure case encountered during our testing, as illustrated in Figure 15. In this scenario, characterized by ultra-dense object clustering combined with heavy shadows and low ambient light, two primary modes of detection failure emerge:

1. **False Negatives (Missed Detections) under Severe Shadowing:** In the heavily occluded regions (e.g., targets concealed by thick tree canopies or building shadows), the model fails to recall certain sub-pixel vehicles. At a 640×640 input resolution, the high-frequency edge gradients of these heavily shadowed targets decay to near-zero levels. Consequently, the WHFD module lacks sufficient signal strength to extract structural boundaries, reaching the physical perception limit of the sensor and the spatial resolution constraint.
2. **Bounding Box Redundancy (False Positives) in Dense Clusters:** In regions of extreme density, we observe occasional instances of redundant or overlapping bounding boxes that survive the Non-Maximum Suppression (NMS) post-processing. This is an unintended side effect of our frequency-domain enhancement: the WHFD module is highly sensitive to high-frequency signals. In chaotic, dense environments, strong specular reflections (e.g., from windshields or complex lighting) can produce spurious, sharp geometric contours. The network occasionally misinterprets these high-frequency artifacts as independent vehicle boundaries.

These failure cases highlight the inherent trade-off between maximizing edge sensitivity for sub-pixel targets and maintaining noise resilience in heavily degraded environments. Balancing edge preservation and noise suppression remains an open problem. In future work, we plan to introduce context-aware adaptive frequency denoising thresholds or combine multi-modal feature fusion strategies to further improve detection in extreme environments.

Furthermore, regarding practical engineering deployment, a major bottleneck lies in the transition from high-performance desktop GPUs to resource-constrained UAV edge devices (e.g., NVIDIA Jetson Nano). While WaveUAV-YOLO achieves real-time throughput on the desktop architecture, operations such as the 2D Haar wavelet transform and multi-branch asymmetric convolutions employed in the WHFD and AMSB modules may encounter operator compatibility issues during TensorRT or ONNX conversion. Additionally, the fundamental INT8 quantization adopted for ultra-low-power deployment often degrades the delicate high-frequency edge signals carefully preserved by WHFD,

potentially leading to a drop in practical precision. Overcoming these hardware-aware deployment barriers, potentially through the development of custom CUDA kernels and quantization-aware training (QAT), remains a critical direction for our future engineering implementations.



Figure 15. Visualization of a typical failure case in an extremely dense and shadowed environment. The model occasionally exhibits missed detections (false negatives) in heavy shadows where edge gradients decay and redundant bounding boxes (false positives) triggered by complex reflections and high-frequency noise interference.

6. Conclusions

To address the challenges of feature loss and shape distortion for small target detection in UAV top-down scenes, this paper proposes a lightweight architecture named WaveUAV-YOLO that integrates frequency-domain feature refinement and stable gradient fusion. By systematically modifying the baseline YOLOv13s, the proposed model improves the perception accuracy of dense sub-pixel targets while maintaining real-time inference speed.

The main conclusions are summarized as follows:

1. Superiority of WHFD: It preserves high-frequency edges better than standard down-sampling, which benefits later morphological processing.
2. Efficacy of FFM_Concat: It mathematically prevents weak features from being suppressed during fusion, which is more cost-effective than adding complex spatial modules.
3. Interplay with smoothness losses: We observed that smoothness-based losses like NWD can inadvertently hurt performance when the network architecture already preserves sharp, high-frequency edges.

Extensive experiments on four datasets (VisDrone2019, NWPU VHR-10, HIT-UAV, and DIOR) demonstrate that WaveUAV-YOLO achieves a good balance between accuracy and speed. Future work will explore applying the architecture to multi-spectral fusion and real-time onboard tracking tasks to further verify its cross-modal generalization potential. Additionally, although the current evaluation is primarily conducted on high-performance desktop environments, we plan to focus on hardware-aware quantization and deploying the architecture onto physical UAV edge devices to validate its real-world real-time capabilities.

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

Funding: This research was funded by the Postgraduate Research & Practice Innovation Program of Jiangsu Province, grant number SJCX25_0525.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The VisDrone2019 dataset is available at <https://github.com/VisDrone/VisDrone-Dataset> (accessed on 16 July 2026). The DIOR and NWPU VHR-10 datasets can be accessed via <https://gcheng-nwpu.github.io/#Datasets> (accessed on 16 July 2026). The HIT-UAV dataset is publicly available at <https://github.com/suojiaashun/HIT-UAV-Infrared-Thermal-Dataset> (accessed on 16 July 2026). The complete source code and pre-trained weights of WaveUAV-YOLO are publicly available at <https://github.com/mrqx/WaveUAV-YOLO> (accessed on 16 July 2026).

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

References

1. Srivastava, S.; Narayan, S.; Mittal, S. A Survey of Deep Learning Techniques for Vehicle Detection from UAV Images. *J. Syst. Archit.* **2021**, *117*, 102152. [CrossRef]
2. Manaswini, V.N.S.; Kamatchi, K.; Nigam, C.; Ali, S.S.; Niranjana, R.; Suman. Real-Time Object Detection in Drone Surveillance Using YOLOv5. In *Proceedings of the 2025 3rd International Conference on IoT, Communication and Automation Technology (ICICAT)*; IEEE: Gorakhpur, India, 2025; pp. 1–6. [CrossRef]
3. Hasan, M.M.; Wang, Z.; Fan, H.; Fatima, K.; Hussain, M.A.I.; Shaha, R.; Habib, T.M.A. BDNet: A lightweight YOLOv12-based vehicle detection framework for smart urban traffic monitoring. *Smart Cities* **2026**, *9*, 33. [CrossRef]
4. Bashir, M.H.; Ahmad, M.; Rizvi, D.R.; El-Latif, A.A.A. Efficient CNN-based disaster events classification using UAV-aided images for emergency response application. *Neural Comput. Appl.* **2024**, *36*, 10599–10612. [CrossRef]
5. Liu, Y.; Li, Q.; Yuan, Y.; Du, Q.; Wang, Q. ABNet: Adaptive Balanced Network for Multiscale Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
6. Hua, W.; Chen, Q. A survey of small object detection based on deep learning in aerial images. *Artif. Intell. Rev.* **2025**, *58*, 162. [CrossRef]
7. Wang, Q.; Cang, M.; Chen, Y. ECP-YOLO: Integrating Edge-Aware Attention and Contextual Refinement for UAV Object Detection. *Electronics* **2026**, *15*, 2067. [CrossRef]
8. 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.
9. Yan, H.; Kong, X.; Wang, J.; Tomiyama, H. ST-YOLO: An Enhanced Detector of Small Objects in Unmanned Aerial Vehicle Imagery. *Drones* **2025**, *9*, 338. [CrossRef]
10. Xu, B.; Cai, D.; Sui, K.; Wang, Z.; Liu, C.; Pei, X. MBD-YOLO: An Improved Lightweight Multi-Scale Small-Object Detection Model for UAVs Based on YOLOv8. *Appl. Sci.* **2025**, *15*, 10877. [CrossRef]
11. Liu, H.; Dong, H.; Shi, H.; Li, F. CCAI-YOLO: A High-Precision Synthetic Aperture Radar Ship Detection Model Based on YOLOv8n Algorithm. *Remote Sens.* **2026**, *18*, 145. [CrossRef]
12. Deng, C.; Zhao, Z.; Xu, X.; Xia, Y.; Li, J.; Plaza, A. GSFANet: Global Spatial-Frequency Attention Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5007017. [CrossRef]
13. Cheng, G.; Zhou, P.; Yao, X.; Yao, C.; Zhang, Y.; Han, J. Object detection in VHR optical remote sensing images via learning rotation-invariant HOG feature. In *Proceedings of the 2016 4th International Workshop on Earth Observation and Remote Sensing Applications (EORSA)*, Guangzhou, China, 4–6 July 2016; pp. 433–436.
14. Ren, Y.; Zhu, C.; Xiao, S. Small object detection in optical remote sensing images via modified faster R-CNN. *Appl. Sci.* **2018**, *8*, 813. [CrossRef]
15. Ding, J.; Xue, N.; Xia, G.S.; Bai, X.; Yang, W.; Yang, M.Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7778–7796. [CrossRef]
16. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
17. Wang, C.Y.; Liao, H.Y.M. YOLOv9: Learning Clean Filters and What to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.
18. Han, W.; Dong, S.; Zhang, Y. Multispectral Small Object Detection for UAV Remote Sensing: A Comprehensive Review. *Remote Sens.* **2025**, *17*, 249. [CrossRef]
19. Yang, F.; Fan, H.; Chu, P.; Blasch, E.; Ling, H. Clustered object detection in aerial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8311–8320.
20. Ye, T.; Qin, W.; Li, Y.; Wang, S.; Zhang, J.; Zhao, Z. Dense and small object detection in UAV-vision based on a global-local feature enhanced network. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–13. [CrossRef]

21. Jin, B.; Yin, F.; Cai, W.; Li, H.; Zhu, H.; Huang, W.; Wu, Q.; Chen, H.; Sun, Z. HWANet: A Haar Wavelet-based Attention Network for remote sensing object detection. *PLoS ONE* **2025**, *20*, e0330759. [[CrossRef](#)] [[PubMed](#)]
22. Chen, J.; Liu, N.; Sun, H.; Wang, Y. Freq-DETR: Frequency-aware transformer for real-time small object detection in unmanned aerial vehicle imagery. *Expert Syst. Appl.* **2025**, *298*, 129710.
23. Yang, Y.; Yuan, G.; Li, J. SFFNet: A Wavelet-Based Spatial and Frequency Domain Fusion Network for Remote Sensing Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17. [[CrossRef](#)]
24. Zhang, Y.; Bao, W.; Yang, Y.; Wan, W.; Xiao, Q.; Zou, X. HAFNet: Hierarchical Attention Fusion Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5007316. [[CrossRef](#)]
25. Wang, Z.; Wang, C.; Li, X.; Xia, C.; Xu, J. MLP-Net: Multilayer Perceptron Fusion Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5601313. [[CrossRef](#)]
26. 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)*; IEEE: Piscataway, NJ, USA, 2020; pp. 10778–10787. [[CrossRef](#)]
27. Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.M.; Yang, J.; Li, X. Large Selective Kernel Network for Remote Sensing Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 1–6 October 2023; pp. 16740–16750.
28. Huang, G.; Liu, Z.; Weinberger, K.Q. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
29. Rahman, M.M.; Marculescu, R. MK-UNet: Multi-kernel Lightweight CNN for Medical Image Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Honolulu, HI, USA, 19–20 October 2025; pp. 979–988.
30. 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*, 7074–7089. [[CrossRef](#)] [[PubMed](#)]
31. Williams, T.; Li, R.Y. Wavelet Pooling for Convolutional Neural Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, 30 April–3 May 2018.
32. Ding, X.; Guo, Y.; Ding, G.; Han, J. ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convolution Blocks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1911–1920.
33. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
34. Lin, T.Y.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. 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.
35. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
36. Zhang, B.; Sennrich, R. Root Mean Square Layer Normalization. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 12360–12371.
37. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 8–14 September 2018; pp. 3–19.
38. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q.; Zheng, J.; Peng, T.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*; IEEE: Piscataway, NJ, USA, 2019; pp. 213–226. [[CrossRef](#)]
39. Li, K.; Wan, G.; Cheng, G.; Cao, L.; Han, J. DIOR: A Large-Scale Benchmark Dataset for Object Detection in Optical Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *164*, 13–23.
40. Liu, J.; Huang, B.; Lv, J.Y. YOLO-PKFF: Remote Sensing Object Detection Enhanced with Poly Kernel Inception and Attentional Cross-Level Feature Fusion. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 6010805. [[CrossRef](#)]
41. Gu, Q.; Huang, H.; Han, Z.; Fan, Q.; Li, Y. GLFE-YOLOX: Global and Local Feature Enhanced YOLOX for Remote Sensing Images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 2516112. [[CrossRef](#)]
42. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [[CrossRef](#)]
43. Suo, J.; Wang, T.; Zhang, X.; Chen, H.; Zhou, W.; Shi, W. HIT-UAV: A High-altitude Infrared Thermal Dataset for Unmanned Aerial Vehicle-based Object Detection. *Sci. Data* **2023**, *10*, 227. [[CrossRef](#)] [[PubMed](#)]
44. Wang, J.; Xu, C.; Yang, W.; Yu, L. A Normalized Gaussian Wasserstein Distance for Tiny Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 10–17 October 2021; pp. 764–773.
45. Xu, C.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.S. Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 79–93. [[CrossRef](#)]

46. Wang, G.; Zhao, H.; Lyu, S.; Cheng, G.; Chang, Q.; Feng, W.; Zhao, Q.; Shi, Z. SWIN-TOD: Smooth wasserstein distance and instance-level neighboring enhancement for remote sensing tiny object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15. [[CrossRef](#)]
47. Zhang, F.; Zhou, S.; Wang, Y.; Wang, X.; Hou, Y. Label assignment matters: A gaussian assignment strategy for tiny object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–12. [[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.