applied sciences

MDPI

Article

SFCD-Det: A Spatial–Frequency Collaborative Architecture for UAV Infrared Small-Object Detection

Yufeng Li, Yong He *, Lei Ji, Qianxu Ren and Dong Lv

College of Computer Science and Technology, Guizhou University, Guiyang 550025, China; gs.yfli24@gzu.edu.cn (Y.L.); gs.lji24@gzu.edu.cn (L.J.); gs.renqx24@gzu.edu.cn (Q.R.); gs.dlv24@gzu.edu.cn (D.L.)
* Correspondence: xiaoyongge666@163.com

Abstract

UAV infrared small-object detection is challenging because targets occupy few pixels, provide weak thermal contrast, and are easily confused with cluttered backgrounds. Although convolutional and Transformer-based detectors improve local representation and global context modeling, their predominantly spatial processing pipelines do not explicitly prevent fragile target responses from being attenuated during early encoding and multi-scale aggregation. To address this gap, we propose SFCD-Det, a spatial–frequency collaborative detector organized around a progressive preservation–purification–coordination methodology. A Wavelet-Transform Stem preserves low-frequency structures and localized high-frequency details before backbone encoding. A Feature Purification Layer regulates adjacent-scale interactions and suppresses clutter-dominated responses before aggregation, while a Spatial–Frequency Coordinated Feature Pyramid reconstructs multi-scale features using shallow spatial anchors, purified intermediate responses, and deep spatial–frequency priors. Experiments on four benchmarks show that SFCD-Det consistently outperforms the DEIM-N baseline. It achieves 62.6% mAP and 95.1% mAP_{50} on HIT-UAV, 41.4% and 86.4% on IRSTD-1K, 16.8% and 46.8% on RGBTDronePerson, and 34.1% and 88.0% on USOD, respectively. These results demonstrate that SFCD-Det strengthens weak-target representation in UAV infrared imagery. Its additional gains on USOD suggest that the frequency mechanism may also benefit weak and spatially localized responses in visible imagery under low illumination or shadow.

Keywords: UAV imagery; small-object detection; spatial–frequency collaboration; wavelet transform; feature purification; feature pyramid network; thermal imagery

Check for updates

Academic Editor: Atsushi Mase

Received: 4 July 2026
Revised: 27 July 2026
Accepted: 28 July 2026
Published: 30 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.

1. Introduction

Infrared object detection from UAV platforms has attracted increasing attention because of its importance in night-time surveillance, traffic monitoring, border patrol, and search-and-rescue applications. Compared with fixed ground-based sensors, UAV platforms provide flexible viewpoints, wide-area coverage, and rapid field application, making them highly suitable for complex outdoor missions. Reliable UAV operation requires both robust flight control under disturbances [1] and accurate aerial perception under complex outdoor conditions, which highlights the need for dependable infrared object detection from UAV platforms.

However, UAV-view infrared imagery differs substantially from ground-based or conventional visible-light images. Due to the elevated imaging perspective, objects are often observed from top-down or oblique angles, resulting in compressed shapes, limited

boundary details, and large variations in apparent scale. Thermal imaging further weakens texture and color information, so targets are mainly distinguished by radiation intensity and local structural contrast. In addition, platform motion, changing flight altitude, and complex ground materials can cause unstable target appearances across frames and scenes. These characteristics make UAV infrared detection highly dependent on the preservation of weak spatial structures and reliable multi-scale representation.

Modern detectors are mainly built on convolutional networks and vision Transformers [2], but differ in how they generate and refine object predictions. Proposal-based frameworks, represented by Faster R-CNN [3] and Cascade R-CNN [4], identify candidate regions before refining their categories and locations. The YOLO family [5–8] instead performs direct prediction and is widely adopted in time-sensitive detection tasks. Transformer-based detectors, including RT-DETR [9], RT-DETRv2 [10], DINO [11], and D-FINE [12], further improve contextual reasoning by establishing interactions between spatially separated image regions.

Within this end-to-end Transformer paradigm, DEIM [13] improves supervision for small and densely distributed objects through dense one-to-one matching and a matchability-aware loss. Its end-to-end prediction process also avoids erroneous suppression caused by conventional post-processing, which is beneficial for densely clustered targets in aerial imagery. These properties make DEIM a suitable Transformer-based baseline for investigating UAV infrared small-object detection.

Despite these advances, DEIM and other modern detectors remain limited when directly applied to UAV infrared imagery. In these thermal scenes, targets usually appear at extremely small scales and often exhibit weak contrast, blurred boundaries, and severe background interference. Whether based on convolutional backbones, one-stage YOLO-style direct prediction, or Transformer attention, existing pipelines still construct their representations mainly through spatial encoding and hierarchical downsampling. YOLO-style detectors are efficient for real-time prediction, but their spatial feature hierarchies may compress fragile thermal cues too early. Transformer attention is effective for modeling long-range dependencies, but it operates on features that have already passed through early spatial encoding and therefore cannot fully compensate for target cues weakened before global interaction.

This spatial representation process leads to three cascading issues. In other words, the difficulty is not caused by a single network layer, but by the cumulative effect of early compression, intermediate feature mixing, and final pyramid aggregation. A small thermal target may first lose boundary details, then be mixed with clutter responses at neighboring scales, and finally become diluted during repeated top-down or bottom-up fusion. First, during the initial network encoding phase, the fragile structural details of tiny targets are easily blurred and degraded by conventional spatial operations. Second, during hierarchical feature extraction, the lack of explicit noise filtering allows severe thermal background clutter to propagate across adjacent scales. Finally, traditional sequential multi-scale aggregation pathways easily dilute the already weak target signatures, failing to effectively coordinate deep spatial–frequency priors with shallow spatial features.

Consequently, the unresolved methodological gap is how to preserve fragile target responses during early encoding, purify them from adjacent-scale background interference, and coordinate them with contextual guidance before multi-scale reconstruction.

These observations motivate SFCD-Det, a progressive preservation–purification–coordination framework for UAV infrared small-object detection. Specifically, the proposed framework preserves fragile cues before backbone extraction, purifies adjacent-scale responses before aggregation, and reconstructs the feature pyramid through coordinated

spatial and frequency guidance. The main contributions of this work are summarized as follows:

1. Progressive Spatial–Frequency Methodology. We propose SFCD-Det based on a progressive preservation–purification–coordination methodology for UAV infrared small-object detection. Rather than applying frequency processing as an isolated enhancement operation, the proposed methodology incorporates it across early encoding, pre-aggregation purification, and multi-scale reconstruction to progressively protect weak target responses.
2. Early Spatial–Frequency Decoupled Encoding. We introduce an early spatial–frequency encoding strategy, implemented through WTStem, to reduce target-cue attenuation before backbone feature extraction. By combining spatial-to-depth rearrangement with high- and low-frequency decoupling, WTStem retains low-frequency structural information and high-frequency target details in the initial representation.
3. Pre-Aggregation Adjacent-Scale Purification. We propose a pre-aggregation purification strategy, implemented through FPL with Frequency-Dynamic Aligned Fusion (FDAF), to regulate information exchanged between adjacent feature levels. By modeling adjacent-scale interactions, FPL suppresses background clutter and refines target-related responses before multi-scale feature reconstruction.
4. Spatial–Frequency Coordinated Pyramid Reconstruction. We introduce a coordinated reconstruction strategy, implemented through SFCFP, that integrates shallow spatial anchors, purified intermediate responses, and deep spatial–frequency priors to construct discriminative multi-scale representations.

2. Related Work

2.1. UAV-Based Infrared Object Detection

Recent progress in deep learning has advanced aerial thermal perception, with architectures increasingly designed for weak contrast and cluttered backgrounds. Lightweight detectors such as ITD-YOLOv8 [14] and LRDS-YOLO [15] improve infrared target perception by redesigning backbone structures and feature interaction mechanisms for aerial platforms. For weak and small thermal targets, BDK-YOLOv8 [16] and YOLO-UIR [17] strengthen discriminative representation through enhanced feature modeling. Transformer-based detectors, including PHSI-RTDETR [18] and DISO-DETR [19], exploit long-range dependency modeling to capture richer contextual cues for small-object detection. More recently, RMT-YOLOv9s [20] improve aerial infrared small-object detection by combining a refined backbone with enhanced multi-scale fusion. Although these studies improve UAV infrared detection from backbone design, attention modeling, and feature fusion perspectives, most of them still operate after weak target cues have entered hierarchical encoding. Early protection of fragile thermal responses before strong spatial compression therefore remains insufficiently explored.

2.2. Spatial–Frequency Combination Methods

The challenge of background interference in the spatial domain has led to the promising development of spatial–frequency fusion strategies. WaveViT [21] demonstrates the effectiveness of unifying wavelet transforms with vision Transformers for multi-scale representation. Boundary-aware methods like SFLNet [22] and SFCANet [23] refine edge features, while GSFANet [24] and FM-Net [25] utilize spectral decomposition to distinguish target details from background clutter. Recent enhancement-oriented studies, such as the edge-prior guided dual-branch network reported by Pan et al. [26], further confirm that explicit structural guidance is beneficial for suppressing cluttered infrared backgrounds. Recent works like DHiF [27] and the frequency and spatial feature fusion method pro-

posed by Zhu et al. [28] further extend these capabilities to complex infrared scenarios. However, these methods primarily focus on frequency injection, structural enhancement, and multi-scale fusion, with less emphasis on statistical-adaptive feature purification. As a result, weak infrared small targets may still be insufficiently separated from cluttered backgrounds under severe thermal inconsistency.

2.3. Feature Pyramid Networks

Feature pyramid networks serve as the primary paradigm for cross-scale interaction and hierarchical feature reconstruction in object detection. Since the original FPN architecture [29], a series of variants have been proposed to strengthen information flow and alleviate feature degradation. These include path-aggregation architectures like BiFPN [30], asymptotic fusion designs like AFPN [31], as well as recent advanced routing and alignment mechanisms such as Gold-YOLO [32], HS-FPN [33], HyperACE [34], and A3FPN [35]. For aerial small-object detection, CFPT [36] further enhances cross-layer interaction by introducing an upsampler-free feature pyramid Transformer tailored to aerial imagery. While highly successful in generic optical scenarios, these spatial-domain progressive architectures struggle with thermal infrared small-object detection. Specifically, repeated spatial upsampling and top-down aggregation can dilute weak thermal signatures, while the lack of explicit cue preservation and clutter purification before pyramid reconstruction allows background interference to propagate across feature levels.

3. Methodologies

3.1. Overall Structure

The overall architecture of SFCD-Det is shown in Figure 1. Built upon DEIM-N, the detector retains the backbone–encoder–decoder pipeline and introduces three progressive components for aerial infrared small-object detection: WTStem, FPL, and SFCFP.

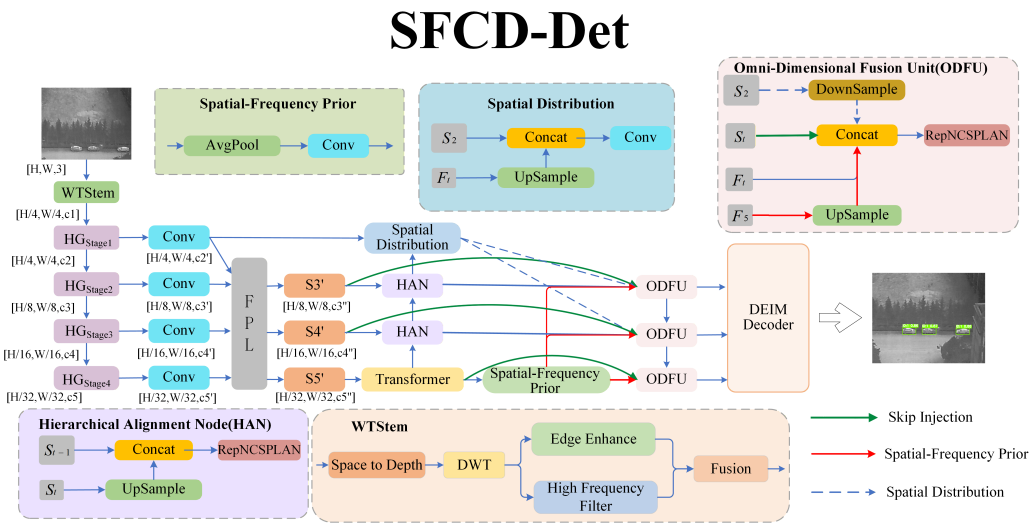

Figure 1. Overall structure of SFCD-Det. The proposed detector introduces WTStem for early spatial–frequency feature preservation, FPL for adjacent-scale feature purification, and SFCFP for coordinated multi-scale reconstruction.

Given an infrared image, WTStem preserves informative target cues before the backbone feature extraction stage through spatial-to-depth rearrangement and high- and low-frequency decoupling. The backbone then produces hierarchical features $\{\mathbf{S}_2, \mathbf{S}_3, \mathbf{S}_4, \mathbf{S}_5\}$.

Before pyramid reconstruction, FPL refines adjacent-scale representations through frequency-dynamic aligned fusion. Specifically, it operates on $(\mathbf{S}_2, \mathbf{S}_3)$, $(\mathbf{S}_3, \mathbf{S}_4)$, and $(\mathbf{S}_4, \mathbf{S}_5)$

to generate purified layers $\mathbf{S}'_3$, $\mathbf{S}'_4$, and $\mathbf{S}'_5$, respectively. In this stage, spatial-detail guidance and spatial–frequency modulation jointly enhance target responses while suppressing background clutter.

Based on $\{\mathbf{S}_2, \mathbf{S}'_3, \mathbf{S}'_4, \mathbf{S}'_5\}$, SFCFP reconstructs the final feature pyramid by integrating shallow spatial anchors via an S2 spatial distribution path, spatial–frequency prior guidance via an S5 spatial–frequency prior broadcasting path, and purified intermediate responses via skip injections. The resulting coordinated multi-scale features are then fed into the decoder for object classification and localization.

In this way, SFCD-Det does not rely on a single enhancement operation but distributes the preservation, purification, and reconstruction of weak target cues across different stages of the detector.

3.2. Wavelet-Transform Stem

From a signal-processing perspective, low-frequency components describe coarse thermal structures, whereas high-frequency components capture local variations and boundary details [37,38]. Unlike global transforms such as FFT and DCT, wavelet decomposition retains spatial localization while separating frequency components, making it suitable for small targets embedded in nonuniform infrared backgrounds.

WTStem is designed to reduce the irreversible loss of fragile thermal signatures during early downsampling in conventional spatial stems, as illustrated in Figure 2. By embedding discrete wavelet decomposition at the input stage, WTStem reformulates initial feature extraction as a spatial–frequency decoupled encoding process, aiming to retain low-frequency structural continuity and high-frequency target-related variations before the features enter the backbone.

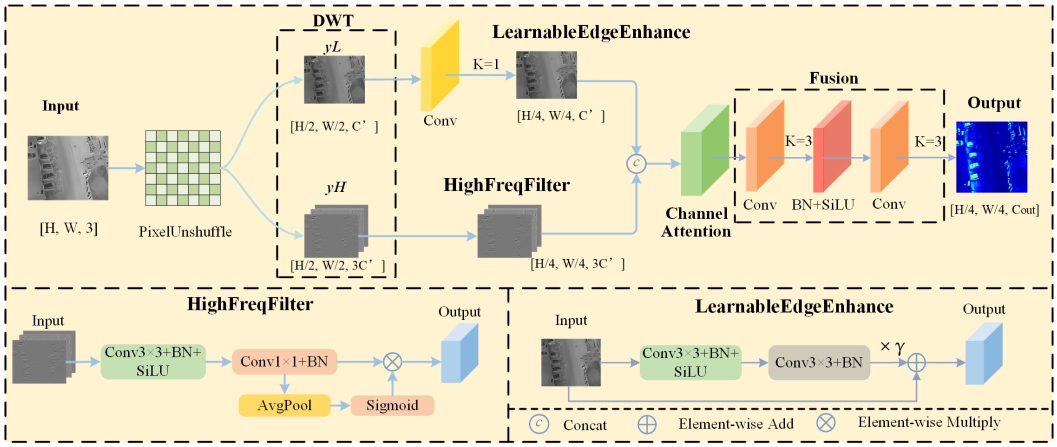

Figure 2. Structure of the proposed WTStem. The input image is decoupled into orthogonal frequency sub-bands, followed by a dual-branch adaptive enhancement strategy for initial signal preservation.

Given an infrared image $\mathbf{X} \in \mathbb{R}^{C_0 \times H \times W}$, a space-to-depth mapping $\mathcal{R}_{\mathrm{s}\to\mathrm{d}}(\cdot)$ first compresses the spatial resolution by rearranging local pixels into the channel dimension, yielding $\mathbf{X}_{\mathrm{s2d}} = \mathcal{R}_{\mathrm{s}\to\mathrm{d}}(\mathbf{X})$. A 2D Haar discrete wavelet transform is then applied to decouple the rearranged representation into four orthogonal frequency sub-bands. Haar wavelets are selected for their compact support and orthogonality, which limit spatial spreading and preserve localized responses more effectively than longer-support families during early downsampling.

$$\mathbf{Y}_{LL}, \mathbf{Y}_{LH}, \mathbf{Y}_{HL}, \mathbf{Y}_{HH} = \mathcal{W}_{\mathrm{haar}}(\mathbf{X}_{\mathrm{s2d}}), \tag{1}$$

where $\mathcal{W}_{\mathrm{haar}}(\cdot)$ denotes a one-level 2D Haar discrete wavelet transform. $\mathbf{Y}_{LL}$ is the low-frequency approximation component, while $\mathbf{Y}_{LH}$, $\mathbf{Y}_{HL}$, and $\mathbf{Y}_{HH}$ are the three high-

frequency directional components. After the space-to-depth rearrangement and one-level Haar decomposition, each sub-band has a spatial size of $H/4 \times W/4$, while local information is transferred into channel and frequency sub-band representations.

Directly mixing these components may weaken fragile target responses during early representation learning. Therefore, WTStem adopts a dual-branch strategy to process low-frequency structures and high-frequency details separately. In the low-frequency branch, $\mathbf{Y}_{LL}$ is first mapped by a channel projection $\mathcal{P}(\cdot)$ and then refined by a residual edge enhancement unit $\mathcal{E}(\cdot)$:

$$\mathbf{L} = \mathcal{P}(\mathbf{Y}_{LL}), \quad \mathbf{F}_L = \mathbf{L} + \gamma \mathcal{E}(\mathbf{L}) \tag{2}$$

where γ is a learnable scalar, and $\mathbf{F}_L$ denotes the enhanced low-frequency representation.

In the high-frequency branch, the three directional components are calibrated by learnable sub-band-wise coefficients α_h, α_v, and α_d, which are shared across channels and spatial positions. The weighted components are concatenated and processed by the high-frequency filtering block $\mathcal{H}(\cdot)$:

$$\widetilde{\mathbf{Y}}_H = \alpha_h \mathbf{Y}_{LH} \parallel \alpha_v \mathbf{Y}_{HL} \parallel \alpha_d \mathbf{Y}_{HH}, \quad \mathbf{F}_H = \mathcal{H}(\widetilde{\mathbf{Y}}_H) \tag{3}$$

where α_h, α_v, and α_d are learnable scalar coefficients for the horizontal, vertical, and diagonal sub-bands, respectively. The notation $\parallel$ represents channel-wise concatenation, and $\mathbf{F}_H$ is the filtered high-frequency representation.

The low- and high-frequency responses are integrated as $\mathbf{Z} = \mathbf{F}_L \parallel \mathbf{F}_H$. A channel attention mechanism then recalibrates $\mathbf{Z}$ using dynamic weights:

$$\mathbf{W}_{ca} = \sigma\big(\mathbf{W}_e \delta\big(\mathbf{W}_r \mathcal{P}_{\text{gap}}(\mathbf{Z})\big)\big) \tag{4}$$

where $\mathcal{P}_{\text{gap}}(\cdot)$ denotes global average pooling, $\delta(\cdot)$ and $\sigma(\cdot)$ denote the ReLU and Sigmoid functions, respectively, and $\mathbf{W}_r$ and $\mathbf{W}_e$ are the channel reduction and expansion projections. Finally, the preserved stem representation is obtained via a fusion mapping $f_{\text{fuse}}(\cdot)$ composed of successive 3×3 convolutions:

$$\mathbf{F}_{\text{stem}} = f_{\text{fuse}}(\mathbf{W}_{ca} \odot \mathbf{Z}) \tag{5}$$

where $\odot$ denotes the Hadamard product. The resulting $\mathbf{F}_{\text{stem}}$ provides a detail-preserved initial representation for subsequent hierarchical encoding. Since the high- and low-frequency components have been separately processed before fusion, the backbone receives features with clearer structural continuity and more explicit edge details. This improves the initial feature basis without changing the overall backbone hierarchy.

3.3. Frequency-Dynamic Aligned Fusion

The Frequency-Dynamic Aligned Fusion (FDAF) module, illustrated in Figure 3, is designed as a basic adjacent-scale interaction unit used in FPL. Its role is to model complementary spatial and frequency responses between two neighboring feature levels, rather than acting as an independent purification stage. Given adjacent feature maps $\mathbf{X}_0$ and $\mathbf{X}_1$, FDAF first performs spatial alignment and concatenates them along the channel dimension to form a unified representation $\mathbf{X}_{cat} \in \mathbb{R}^{2C \times H \times W}$.

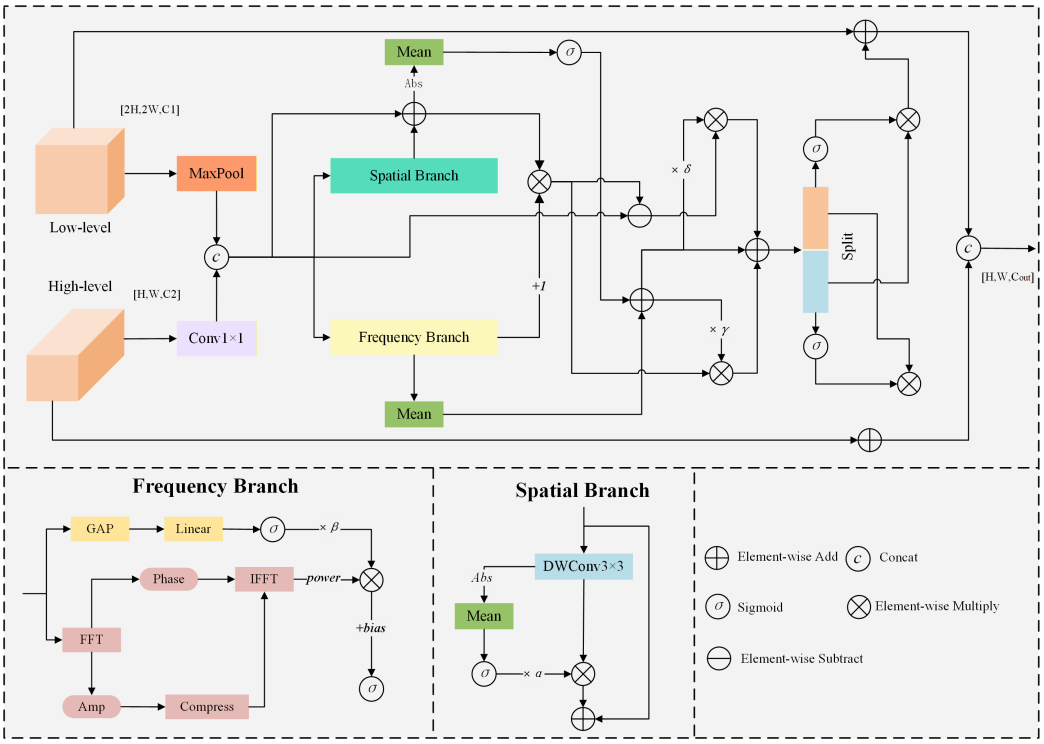

Figure 3. Architecture of the FDAF module. FDAF models adjacent-scale feature interaction through spatial response encoding and spatial–frequency dynamic modulation.

In the spatial branch, local geometric responses $\mathbf{X}_{loc}$ are extracted by depth-wise convolution, and the spatial representation $\mathbf{X}_{sp}$ is obtained through residual attention:

$$\mathbf{X}_{sp} = \mathbf{X}_{cat} + \alpha \odot (\mathbf{X}_{loc} \odot \sigma(\mu_c(|\mathbf{X}_{loc}|))) \tag{6}$$

where α is a learnable channel-wise scaling vector, and $\mu_c(\cdot)$ denotes channel-wise averaging. This branch preserves local continuity and positional responses during adjacent-scale interaction.

The frequency branch then converts $\mathbf{X}_{cat}$ to the spectral domain with a 2D FFT $\mathcal{F}(\cdot)$. Its amplitude $\mathcal{A}$ is adjusted by the learnable exponent η, whereas the phase ϕ is kept unchanged to retain structural layout.

$$\mathbf{X}_{fr} = \mathcal{F}_r^{-1}\left((\mathcal{A} + \epsilon)^{\eta} e^{j\phi}\right) \tag{7}$$

The global channel prior $\mathbf{P}_{ch}$ is generated from the globally pooled $\mathbf{X}_{cat}$ through a learnable linear projection and Sigmoid normalization:

$$\mathbf{P}_{ch} = \sigma\left(\mathbf{W}_p \mathcal{P}_{\text{gap}}(\mathbf{X}_{cat}) + \mathbf{b}_p\right) \tag{8}$$

where $\mathcal{P}_{\text{gap}}(\cdot)$ denotes global average pooling, $\mathbf{W}_p$ is a learnable channel projection matrix, and $\mathbf{b}_p$ is the corresponding bias term. A dynamic frequency gate $\mathbf{G}_{fr}$ is then generated by combining frequency energy with this global channel prior:

$$\mathbf{G}_{fr} = \sigma\left(\beta \odot (\mathbf{P}_{ch} \odot \mathbf{X}_{fr}^2) + \mathbf{b}_f\right) \tag{9}$$

where β and $\mathbf{b}_f$ are learnable channel-wise parameters. Because high-frequency energy may also arise from clutter and sensor noise, $\mathbf{G}_{fr}$ conditions frequency responses on global channel statistics rather than enhancing them indiscriminately, while the residual branch retains complementary information. This design is consistent with recent frequency-

domain noise-suppression studies in infrared small-target detection, where purifying high-frequency components is used to reduce noise interference and false alarms [39].

Guided by $\mathbf{G}_{fr}$, FDAF constructs a frequency-enhanced signal component $\mathbf{X}_{sig}$ alongside a complementary residual component $\mathbf{X}_{res}$:

$$\mathbf{X}_{sig} = \mathbf{X}_{sp} \odot \Big(1 + \mathbf{M} \odot \mathbf{G}_{fr}\Big), \quad \mathbf{X}_{res} = \mathbf{X}_{cat} - \mathbf{X}_{sig}, \tag{10}$$

where $\mathbf{M}$ is a learnable channel-wise modulation tensor.

To adaptively balance the two components, a spatial–frequency response map $\mathbf{P}_{fs} \in \mathbb{R}^{1\times H\times W}$ is estimated from both spatial and frequency responses:

$$\mathbf{P}_{fs} = \sigma\Big(k\Big(\sigma\big(\mu_c(|\mathbf{X}_{sp}|)\big) + \mu_c(\mathbf{G}_{fr})\Big) + b_{fs}\Big), \tag{11}$$

where k and b_{fs} are learnable scalar parameters. The interacted representation is then synthesized by the following equation:

$$\mathbf{X}_{int} = \mathbf{X}_{sig} \odot (1 + \theta\mathbf{P}_{fs}) + \mathbf{X}_{res} \odot (1 - \lambda\mathbf{P}_{fs}), \tag{12}$$

where θ and λ are bounded learnable scalar coefficients for the signal and residual branches, respectively.

Finally, $\mathbf{X}_{int}$ is split along the channel dimension into modulation masks $\mathbf{W}_0$ and $\mathbf{W}_1$ to enhance the original inputs bidirectionally:

$$\widetilde{\mathbf{X}}_0 = \mathbf{X}_0 + \mathbf{X}_1 \odot \sigma(\mathbf{W}_0), \quad \widetilde{\mathbf{X}}_1 = \mathbf{X}_1 + \mathbf{X}_0 \odot \sigma(\mathbf{W}_1). \tag{13}$$

The final FDAF output is obtained by concatenating the enhanced features and applying an output projection $f_{\text{out}}(\cdot)$:

$$\mathbf{Y} = f_{\text{out}}\Big(\widetilde{\mathbf{X}}_0 \parallel \widetilde{\mathbf{X}}_1\Big). \tag{14}$$

Through reciprocal modulation, FDAF serves as the adjacent-scale interaction unit embedded in FPL.

3.4. Feature Purification Layer

Based on FDAF, the Feature Purification Layer (FPL) acts as a pre-fusion hub between the encoder projections and the final multi-scale reconstruction module. Unlike FDAF, which focuses on adjacent-scale interaction, FPL emphasizes feature purification by organizing multiple FDAF units to refine hierarchical representations before pyramid aggregation.

As illustrated in Figure 4, FPL deploys three FDAF units on adjacent feature pairs $(\mathbf{S}_2, \mathbf{S}_3)$, $(\mathbf{S}_3, \mathbf{S}_4)$, and $(\mathbf{S}_4, \mathbf{S}_5)$. In each pair, the high-resolution feature provides spatial-detail guidance, while the lower-resolution feature contributes stronger contextual responses. The purified feature maps are generated as follows:

$$\mathbf{S}'_l = \text{FDAF}(\mathbf{S}_{l-1}, \mathbf{S}_l), \quad l \in \{3, 4, 5\}. \tag{15}$$

Unlike conventional methods that directly aggregate multi-scale features, FPL performs pre-reconstruction purification to reduce background clutter and cross-scale inconsistencies before large-scale feature propagation. The resulting purified features $\{\mathbf{S}'_3, \mathbf{S}'_4, \mathbf{S}'_5\}$ provide cleaner and target-sensitive inputs for SFCFP. By filtering adjacent-scale representations before large-scale aggregation, FPL reduces the risk that background clutter is propagated across multiple levels. Meanwhile, the adjacent interaction preserves comple-

mentary details from neighboring resolutions, which helps maintain the continuity of weak target responses.

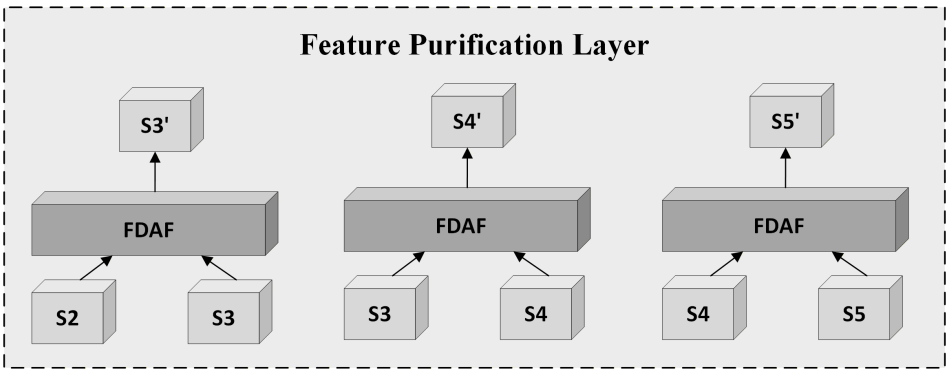

Figure 4. Structure of the Feature Purification Layer (FPL). The layer integrates FDAF units to perform adjacent-scale feature purification.

FPL operates on adjacent feature levels because neighboring representations have relatively close spatial resolutions and stronger semantic correspondence, which facilitates reliable alignment during purification. In contrast, direct interaction between non-adjacent levels may introduce excessive scale mismatch and mix weak target responses with semantically inconsistent background features. Frequency modulation is performed before pyramid reconstruction so that target-related responses are refined and clutter-dominated components are suppressed before cross-scale propagation; otherwise, weak cues may already be diluted during repeated fusion and upsampling.

3.5. Spatial–Frequency Coordinated Feature Pyramid (SFCFP)

Following the feature purification stage, the Spatial–Frequency Coordinated Feature Pyramid (SFCFP) is constructed to perform the final multi-scale reconstruction. As shown in Figure 5, SFCFP takes $\mathbf{S}_2$, $\mathbf{S}'_3$, $\mathbf{S}'_4$, and $\mathbf{S}'_5$ as inputs, where $\mathbf{S}_2$ denotes the high-resolution shallow feature, and $\mathbf{S}'_3$, $\mathbf{S}'_4$, and $\mathbf{S}'_5$ denote the purified intermediate features generated by FPL. Unlike conventional sequential feature pyramids, which often suffer from signal dilution during layer-by-layer propagation, SFCFP establishes a parallel and coordinated reconstruction paradigm. Specifically, it introduces three coordinated pathways: S2 spatial distribution, S5 spatial–frequency prior broadcasting, and purified intermediate skip injection. These pathways jointly preserve fragile thermal signatures and improve localization reliability.

Let $f_{\text{up}}(\cdot)$ and $f_{\text{down}}(\cdot)$ denote spatial upsampling and downsampling operations, respectively. To establish a robust semantic baseline, the deepest purified feature $\mathbf{S}'_5$ is first enhanced by a Transformer encoder, denoted as $f_{\text{trans}}(\cdot)$. In infrared imagery, where local contrast is weak, capturing long-range contextual dependencies is important for distinguishing real targets from complex thermal backgrounds. The enhanced deep features are then propagated through a top-down pathway:

$$\mathbf{T}_5 = f_{\text{trans}}(\mathbf{S}'_5), \quad \mathbf{T}_l = \Phi_l\big(f_{\text{up}}(\mathbf{T}_{l+1}) \parallel \mathbf{S}'_l\big), \quad l \in \{3,4\}, \tag{16}$$

where $\Phi_l(\cdot)$ denotes the top-down fusion block. This pathway provides foundational semantic guidance for the intermediate scales.

To preserve geometric localization cues for extremely small objects, SFCFP constructs a parallel S2 spatial distribution path. In conventional architectures, fine-grained target boundaries can be weakened after repeated multi-scale convolutions. To alleviate this issue, the shallowest feature $\mathbf{S}_2$, which contains rich spatial gradients, is fused with the top-down

feature $\mathbf{T}_3$ to form a high-resolution spatial hub $\mathbf{H}_2$. This hub is then redistributed to deeper levels via direct downsampling:

$$\mathbf{H}_2 = f^{s}_{\text{proj}}\big(f_{\text{up}}(\mathbf{T}_3) \parallel \mathbf{S}_2\big), \quad \mathbf{H}_l = f^{(l-2)}_{\text{down}}(\mathbf{H}_2), \quad l \in \{3,4\}. \tag{17}$$

where f^{s}_{proj} denotes a scale-specific 1×1 projection for the P2 spatial hub, and $f^{(l-2)}_{\text{down}}(\cdot)$ redistributes $\mathbf{H}_2$ to level l.

This shortcut design injects high-resolution spatial anchors into deeper pyramid levels, effectively alleviating the localization degradation caused by sequential downsampling.

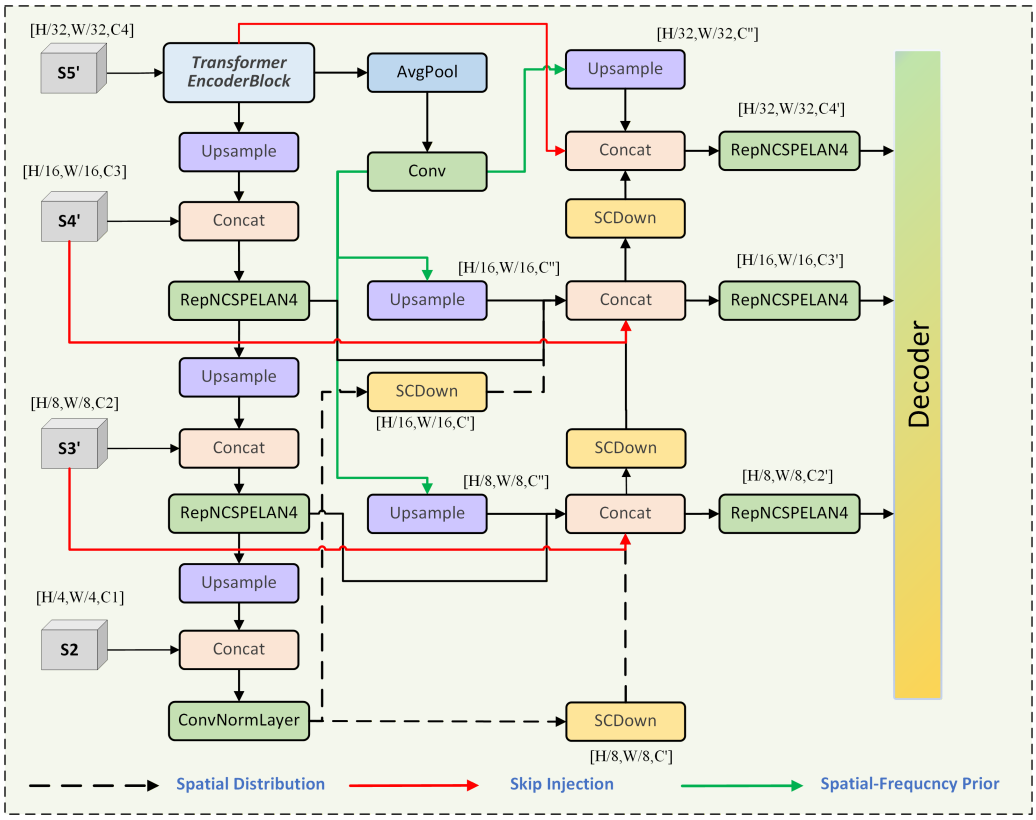

Figure 5. Structure of the Spatial–Frequency Coordinated Feature Pyramid (SFCFP). SFCFP receives $\mathbf{S}_2$ and the purified features $\mathbf{S}'_3$, $\mathbf{S}'_4$, and $\mathbf{S}'_5$ and reconstructs discriminative multi-scale features through S2 spatial distribution, S5 spatial–frequency prior broadcasting, and purified intermediate skip injection.

In parallel with spatial distribution, SFCFP introduces an S5 deep spatial–frequency prior broadcasting path. The deepest semantic feature $\mathbf{T}_5$ is selected because it has the smallest spatial scale after repeated downsampling and convolutional filtering, providing the largest receptive field, richest semantic abstraction, and less local noise than shallower features. Because small thermal targets often lack distinct morphological features and are easily confused with background clutter, global scene statistics are valuable for suppressing false alarms. A spatial–frequency prior is extracted from the deepest semantic feature $\mathbf{T}_5$ through global average pooling and channel projection:

$$\mathbf{q}_5 = f^{g}_{\text{proj}}\big(\mathcal{P}_{\text{gap}}(\mathbf{T}_5)\big), \quad \mathbf{G}_l = f^{(l)}_{\text{exp}}(\mathbf{q}_5), \quad l \in \{3,4,5\}, \tag{18}$$

where f^{g}_{proj} denotes the global-context channel projection, and $f^{(l)}_{\text{exp}}(\cdot)$ expands the global prior vector $\mathbf{q}_5$ to match the spatial resolution of level l. This broadcasting mechanism enables all pyramid levels to share spatial–frequency prior, reducing over-response to local

background noise. Finally, the purified skip features, spatial anchors, top-down semantics, and spatial–frequency priors are integrated during the bottom-up reconstruction phase. The final reconstructed pyramid features are denoted as $\mathbf{F}_3$, $\mathbf{F}_4$, and $\mathbf{F}_5$:

$$\mathbf{F}_3 = \Psi_3(\mathbf{H}_3 \parallel \mathbf{T}_3 \parallel \mathbf{S}'_3 \parallel \mathbf{G}_3), \tag{19}$$

$$\mathbf{F}_4 = \Psi_4(f_{\text{down}}(\mathbf{F}_3) \parallel \mathbf{T}_4 \parallel \mathbf{S}'_4 \parallel \mathbf{H}_4 \parallel \mathbf{G}_4), \tag{20}$$

$$\mathbf{F}_5 = \Psi_5(f_{\text{down}}(\mathbf{F}_4) \parallel \mathbf{T}_5 \parallel \mathbf{G}_5), \tag{21}$$

where $\Psi_l(\cdot)$ denotes the final aggregation block at level l. In this formulation, Ψ_l functions as a coordinated fusion center. For instance, at level 3, it jointly integrates the spatial anchors from $\mathbf{H}_3$, local semantics from $\mathbf{T}_3$, frequency-refined intermediate features from $\mathbf{S}'_3$, and spatial–frequency prior guidance from $\mathbf{G}_3$. Through the cooperation of these decoupled pathways, SFCFP reconstructs a robust multi-scale representation with precise localization cues and strong semantic discrimination for UAV infrared small-object detection.

4. Experiments and Results

4.1. Experimental Dataset Description

(1) HIT-UAV:HIT-UAV [40] provides thermal imagery collected from high-altitude UAV viewpoints for aerial object detection. Its aerial scenes contain several categories and exhibit substantial variations in target scale and thermal appearance. To alleviate class imbalance, we discard the "Dontcare" annotations and consolidate "Car" and "Other Vehicle" into a single "Vehicle" category. The processed samples are partitioned into training, validation, and test subsets using a 7:1:2 ratio.

(2) IRSTD-1K: IRSTD-1K [41] is an infrared small-target benchmark with diverse target appearances and complex background interference. It serves as a representative testbed for small-target detection under weak responses, low signal-to-clutter ratios, and cluttered thermal scenes. In this study, IRSTD-1K further assesses the proposed method under extremely small target scales, limited structural information, low local contrast, and challenging infrared scenes.

(3) RGBTDronePerson: RGBTDronePerson [42] provides spatially aligned RGB and thermal observations of densely distributed pedestrians viewed from different UAV altitudes and perspectives. In our experiments, only the thermal stream is utilized, omitting any cross-modal fusion with the visible spectrum. Following [42], samples labeled as "Uncertained" are excluded to avoid ambiguity during evaluation.

(4) USOD: USOD [43] is a visible-light remote sensing small-object detection dataset built from UNICORN2008. It contains 3000 images and 43,378 annotated vehicle instances. The dataset is dominated by extremely small objects, with 96.3% of targets smaller than 16×16 pixels and 99.9% smaller than 32×32 pixels. Many vehicle instances appear under low illumination or shadow occlusion, resulting in weak and spatially localized responses. USOD therefore provides a complementary evaluation of the frequency mechanism beyond thermal imagery, while UAV infrared detection remains the primary scope of this study.

As shown in Figure 6, the target width–height distributions of HIT-UAV, IRSTD-1K, RGBTDronePerson, and USOD are mainly concentrated in the small-object region marked by the 32-pixel boundary, indicating that these datasets are dominated by small targets. In particular, IRSTD-1K, RGBTDronePerson, and USOD are highly concentrated at very small scales, while HIT-UAV shows wider scale variation but still maintains its main density in the small-target range. These observations indicate that weak responses, limited spatial extent, scale sensitivity, and insufficient visual detail are common challenges across the four benchmarks.

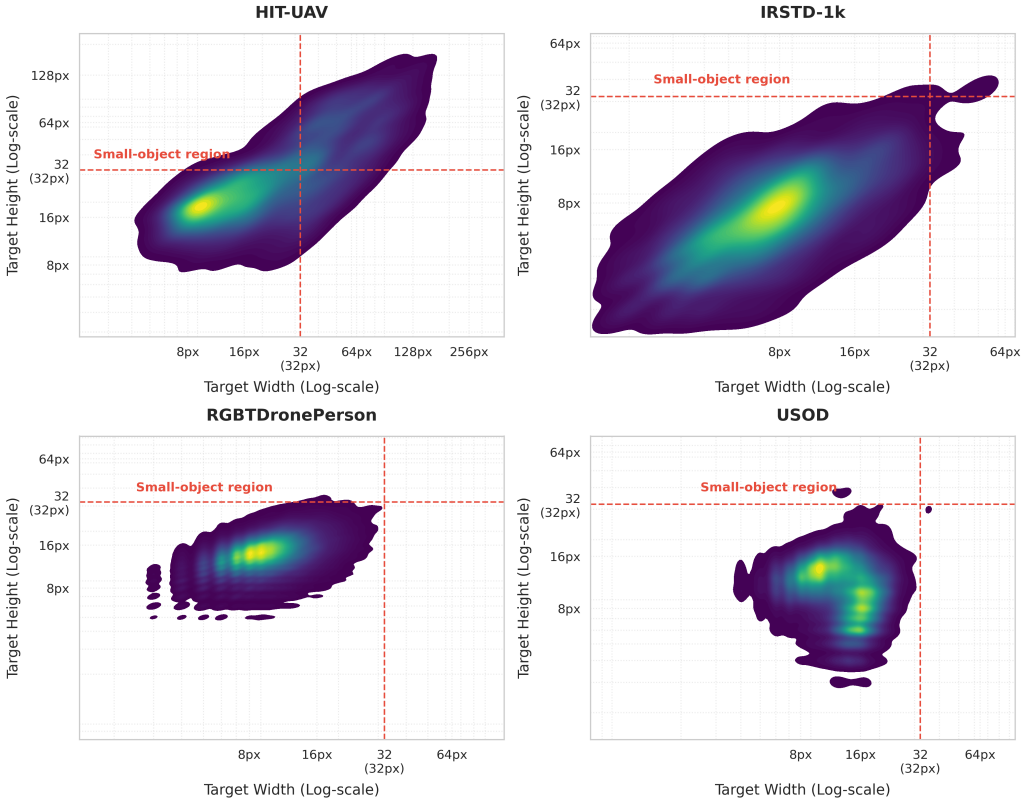

Figure 6. Target width–height distributions of HIT-UAV, IRSTD-1K, RGBTDronePerson, and USOD in logarithmic coordinate space.

4.2. Training and Experimental Comparison Platform

Table 1 summarizes the implementation settings used in all experiments. The proposed method is implemented in PyTorch 2.3.0 with Python 3.10 and trained on an NVIDIA A100 GPU with an AMD EPYC 7742 64-Core Processor. All input images are resized to 640×640. The batch size is set to 8, and AdamW is used with an initial learning rate of 4×10^{-4} and a weight decay of 1×10^{-4}.

Table 1. Implementation settings used for model training and evaluation.

Item	Setting
GPU	NVIDIA A100
CPU	AMD EPYC 7742 64-Core Processor
Framework	PyTorch 2.3.0
Python Version	Python 3.10
Batch Size	8
Optimizer	AdamW
Initial Learning Rate	4×10^{-4}
Weight Decay	1×10^{-4}
Resize	640×640

4.3. Evaluation Metrics

We evaluate detection performance using precision, recall, AP, mAP_{50}, and $\text{mAP}_{50:95}$. In this paper, mAP refers to $\text{mAP}_{50:95}$ unless otherwise specified. Precision (P) and recall (R) are defined as

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{22}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{23}$$

where TP, FP, and FN represent correctly detected objects, false alarms, and missed objects, respectively.

For category i at IoU threshold τ, AP is computed from the precision–recall curve:

$$\text{AP}_i^{\tau} = \int_0^1 P_i^{\tau}(R)\, dR, \tag{24}$$

where $P_i^{\tau}(R)$ denotes the precision–recall curve of category i.

The two mAP metrics are calculated as

$$\text{mAP}_{50} = \frac{1}{C} \sum_{i=1}^{C} \text{AP}_i^{0.50}, \tag{25}$$

$$\text{mAP} = \frac{1}{10} \sum_{\tau \in \{0.50, 0.55, \ldots, 0.95\}} \left(\frac{1}{C} \sum_{i=1}^{C} \text{AP}_i^{\tau} \right) \tag{26}$$

where C is the number of object categories, and τ ranges from 0.50 to 0.95 in increments of 0.05.

For multi-class datasets, mAP_{50} and mAP are averaged over all categories. For single-class datasets, they are equivalent to the AP of the target category. Precision and recall are also reported to provide a more complete evaluation of detection quality in complex scenes.

4.4. Comparison with Advanced Detectors

To comprehensively evaluate the proposed SFCD-Det, we compare it with representative one-stage, Transformer-based detectors across four benchmarks. The evaluation covers UAV infrared object detection, infrared small-object detection, UAV thermal tiny-person detection, and visible-light remote sensing small-object detection.

As shown in Table 2 and Figure 7, the proposed method reaches 62.6% mAP and 95.1% mAP_{50} on the HIT-UAV dataset. Compared with the DEIM-N baseline, our method improves mAP from 59.0% to 62.6% and mAP_{50} from 93.4% to 95.1%, with gains of 3.6% and 1.7%, respectively. This improvement indicates that the proposed spatial–frequency collaborative design effectively enhances the representation of small infrared targets in UAV scenes. In addition, our method surpasses the larger DEIM-S model, which attains 62.2% mAP and 94.4% mAP_{50}. Compared with DEIM-N, the improvement is obtained with a small parameter change, from 3.72 M to 3.81 M, and a FLOP increase from 7.1 G to 9.9 G. The trade-off visualization in Figure 7 further supports this observation, where our method occupies a favorable region with high detection accuracy and relatively low parameter and FLOP costs. These results indicate that the proposed detector improves small-target discrimination over DEIM-N while remaining competitive among the compared compact models.

Among the compared methods on HIT-UAV, YOLO-ViT [44] achieves 98.1% vehicle AP and 91.3% recall, but it requires 17.30M parameters and 33.1G FLOPs. Compact detectors such as FBRT-YOLO [45] and BDK-YOLO [16] reach 61.7% and 61.6% mAP, respectively, showing competitive accuracy with relatively small model sizes. In comparison, SFCD-Det records 62.6% mAP, 95.1% mAP_{50}, 94.7% person AP, and 92.7% bicycle AP with only 3.81 M parameters. Relative to DEIM-N, the improvement is reflected not only in overall mAP but also in the Person and Bicycle categories, which are more sensitive to scale variation, partial occlusion, and weak thermal contrast. These results indicate that SFCD-Det offers a favorable trade-off between accuracy, parameter count, and FLOPs.

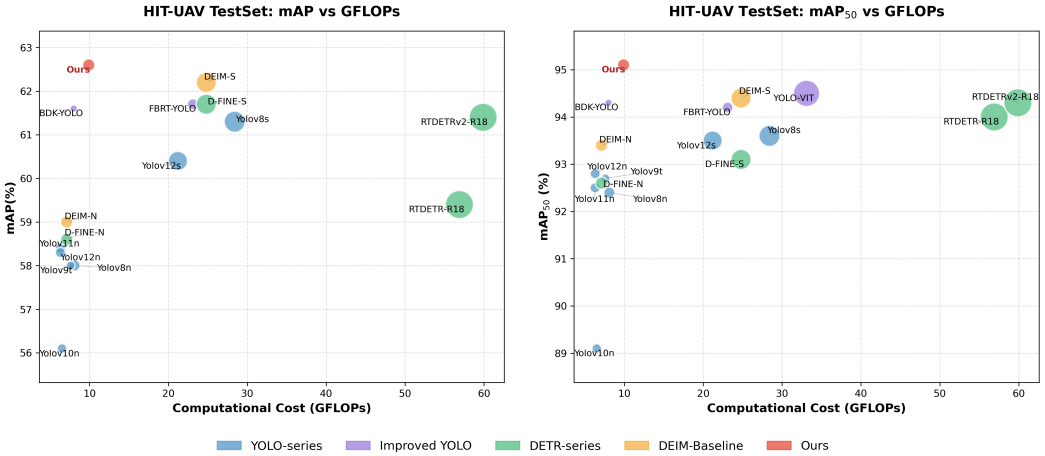

Figure 7. Accuracy vs. computational cost on the HIT-UAV test set. The bubble area indicates the parameter count. The top-left region indicates models with higher detection accuracy and lower reported parameter and FLOP costs.

Table 2. Detection performance of advanced detectors on the HIT-UAV dataset. 'Per.', 'Veh.', and 'Bic.' denote the Person, Vehicle, and Bicycle categories. Bold and underline indicate the best and second-best results. The downward arrow indicates that lower values are better.

Method	Per. (%)	Veh. (%)	Bic. (%)	mAP (%)	mAP_{50} (%)	P (%)	R (%)	Params (M) ↓	FLOPs (G) ↓
YOLOv8-N	91.0	97.5	88.6	58.0	92.4	90.4	87.1	3.01	8.1
YOLOv8-S	92.2	97.1	91.5	61.3	93.6	91.8	89.2	11.13	28.4
YOLOv9-T	90.4	97.5	90.1	58.0	92.7	89.6	88.0	<u>1.97</u>	7.6
YOLOv10-N	88.4	95.4	83.5	56.1	89.1	83.2	83.3	2.27	<u>6.5</u>
YOLOv11-N	91.3	97.7	88.6	58.4	92.5	89.9	86.9	2.58	**6.3**
YOLOv12-N	91.0	97.6	89.8	58.3	92.8	90.7	87.2	2.56	**6.3**
YOLOv12-S	92.5	96.4	91.5	60.4	93.5	<u>92.0</u>	88.3	9.23	21.2
RT-DETR-R18	<u>93.9</u>	96.5	91.5	59.4	94.0	**93.0**	89.4	19.87	56.9
RT-DETRv2-R18	93.7	97.6	91.7	61.4	94.3	90.5	89.5	19.88	59.9
D-FINE-N	91.6	96.5	89.6	58.6	92.6	89.4	87.1	3.72	7.1
D-FINE-S	93.1	97.0	89.3	61.7	93.1	90.3	87.1	10.18	24.8
DEIM-N	92.5	97.1	90.6	59.0	93.4	89.7	88.1	3.72	7.1
DEIM-S	<u>93.9</u>	97.0	92.4	<u>62.2</u>	94.4	91.0	90.2	10.18	24.8
FBRT-YOLO	92.4	<u>97.8</u>	<u>92.5</u>	61.7	94.2	91.6	90.0	2.90	23.1
YOLO-VIT	93.3	**98.1**	92.0	-	<u>94.5</u>	90.0	**91.3**	17.30	33.1
BDK-YOLO	-	-	-	61.6	94.3	90.5	90.3	**1.35**	8.0
Ours	**94.7**	<u>97.8</u>	**92.7**	**62.6**	**95.1**	91.2	<u>90.9</u>	3.81	9.9

To further evaluate the ability to detect extremely small infrared targets, we conduct experiments on the IRSTD-1K dataset. As shown in Table 3, the proposed method achieves 41.4% mAP and 86.4% mAP_{50}. It outperforms DEIM-N by 2.4% points in mAP and 2.3 points in recall, showing stronger recovery of weak targets in cluttered infrared scenes. Compared with DEIM-S, the proposed method also yields better accuracy while using far fewer parameters and FLOPs. Against IRMSD-YOLO [46], it attains slightly higher mAP, mAP_{50}, and recall, while requiring lower computation. In addition, compared with RT-DETR-R18, the proposed method obtains higher accuracy with a much lighter model size and computation budget, which further confirms its efficiency on extremely small infrared targets.

Table 3. Detection performance of advanced detectors on the IRSTD-1K dataset. Bold and underline indicate the best and second-best results. The downward arrow indicates that lower values are better.

Method	mAP (%)	mAP_{50} (%)	P (%)	R (%)	Params (M) ↓	FLOPs (G) ↓
YOLOv8-N	38.0	83.4	86.7	78.1	3.01	8.1
YOLOv9-T	35.2	80.5	85.0	77.9	**1.97**	7.6
YOLOv11-N	37.8	81.8	<u>87.8</u>	76.4	2.58	**6.3**
YOLOv12-N	37.2	82.3	84.7	80.8	<u>2.56</u>	**6.3**
RT-DETR-R18	40.2	85.3	87.0	80.8	19.87	56.9
D-FINE-N	38.7	83.1	84.3	81.1	3.72	<u>7.1</u>
D-FINE-S	39.4	83.9	83.1	82.9	10.18	24.8
DEIM-N	39.0	83.7	83.4	82.7	3.72	<u>7.1</u>
DEIM-S	40.8	84.6	85.3	<u>83.2</u>	10.18	24.8
FBRT-YOLO	39.1	82.0	85.6	75.0	2.90	23.1
IRMSD-YOLO	<u>41.2</u>	<u>86.2</u>	**89.1**	83.1	8.35	37.1
Ours	**41.4**	**86.4**	83.5	**85.0**	3.81	9.9

To further validate the efficacy of SFCD-Det in UAV thermal pedestrian detection, we conduct evaluations on the RGBTDronePerson benchmark. As shown in Table 4, only the thermal modality is used in this experiment. Compared with the DEIM-N baseline, our method improves mAP from 15.8% to 16.8%, mAP_{50} from 44.3% to 46.8%, precision from 50.8% to 53.2%, and recall from 49.5% to 53.6%. These consistent improvements indicate that the proposed spatial–frequency modeling strategy remains effective when targets appear as tiny thermal responses with severe scale variation. Compared with D-FINE-S, our method also yields 0.3% higher mAP while utilizing 6.37M fewer parameters. Furthermore, compared with the recent YOLOv12-M, our method attains 1.2% higher mAP and 3.7% higher mAP_{50} while significantly reducing the parameter count by 16.29M and FLOPs by 57.2G. Thus, SFCD-Det still shows a clear performance advantage in UAV thermal imagery, where targets are very small, densely packed, and often occluded, indicating robustness under challenging infrared conditions.

Table 4. Detection performance of advanced detectors on the RGBTDronePerson dataset. Bold and underline indicate the best and second-best results. The downward arrow indicates that lower values are better.

Method	mAP (%)	mAP_{50} (%)	P (%)	R (%)	Params (M) ↓	FLOPs (G) ↓
YOLOv8-S	13.0	37.9	48.7	45.5	11.13	28.4
YOLOv11-S	14.3	40.8	<u>52.3</u>	44.9	9.41	21.3
YOLOv12-M	15.6	43.1	51.9	45.5	20.1	67.1
FBRT-YOLO	13.0	35.5	44.7	40.6	**2.90**	23.1
D-FINE-S	<u>16.5</u>	<u>45.8</u>	50.7	<u>52.9</u>	10.18	24.8
DEIM-N	15.8	44.3	50.8	49.5	<u>3.72</u>	**7.1**
Ours	**16.8**	**46.8**	**53.2**	**53.6**	3.81	<u>9.9</u>

Finally, USOD is used as a complementary visible-light benchmark to examine whether the proposed design remains effective beyond infrared imagery. As shown in Table 5, our method reaches 34.1% mAP and 88.0% mAP_{50}. Compared with DEIM-N, it increases mAP from 32.7% to 34.1% and mAP_{50} from 85.1% to 88.0%, while the parameter count increases from 3.72 M to 3.81 M and FLOPs from 7.1 G to 9.9 G. Although YOLOv11-S obtains higher precision and recall of 89.4% and 82.5%, respectively, it reports 9.41 M parameters and 21.3 G FLOPs, compared with 3.81 M parameters and 9.9 G FLOPs for our method. These results suggest that the observed improvement is associated with weak and spatially localized target cues rather than thermal-specific degradation alone. Therefore,

the frequency mechanism may also benefit visible-light small-object detection when target evidence is weakened by low illumination or shadow occlusion. Overall, the quantitative comparisons across four datasets support the effectiveness and cross-modal robustness of the proposed method, especially under small-scale, low-contrast, and degraded imaging scenarios.

Table 5. Detection performance of advanced detectors on the USOD dataset. Bold and underline indicate the best and second-best results. The downward arrow indicates that lower values are better.

Method	mAP (%)	mAP_{50} (%)	P (%)	R (%)	Params (M) ↓	FLOPs (G) ↓
YOLOv8-S	30.1	85.1	87.3	81.5	11.13	28.4
YOLOv11-S	32.3	<u>87.5</u>	**89.4**	**82.5**	9.41	21.3
FBRT-YOLO	28.3	83.1	85.5	80.4	**2.90**	23.1
DEIM-N	<u>32.7</u>	85.1	85.8	78.1	<u>3.72</u>	**7.1**
Ours	**34.1**	**88.0**	<u>87.8</u>	<u>81.7</u>	3.81	<u>9.9</u>

4.5. Ablation Study

4.5.1. Overall Ablation Study

Table 6 isolates the effect of WTStem, FPL, and SFCFP. With WTStem, the mAP of DEIM-N increases from 59.0% to 59.8%, indicating that preserving early spatial–frequency cues improves the subsequent encoding process. Introducing FPL further raises mAP to 60.1%, which shows the benefit of removing clutter-dominated adjacent-scale responses before pyramid construction. Among the single-module variants, SFCFP contributes the most obvious improvement and reaches 61.2% mAP, highlighting the role of coordinated multi-scale reconstruction. The combined variants further show that these modules are complementary rather than independent additions. Pairing WTStem with SFCFP increases mAP to 61.8%, whereas the FPL+SFCFP configuration reaches 62.0%. This comparison indicates that purified adjacent-scale features provide more reliable inputs for the pyramid. The full configuration with WTStem, FPL, and SFCFP obtains the best performance, reaching 62.6% mAP, 95.1% mAP_{50}, 91.2% precision, and 90.9% recall. Relative to the baseline, the complete model improves mAP by 3.6% and recall by 2.8%. Beyond the numerical improvements, the ablation results reveal a progressive working pattern: WTStem strengthens the initial representation, FPL refines neighboring-scale responses before aggregation, and SFCFP reorganizes the final multi-scale features. This progression shows that the three components form a connected optimization path for weak object detection.

Table 6. Ablation study of the proposed components on the HIT-UAV dataset.

Variant	WTStem	FPL	SFCFP	mAP (%)	mAP_{50} (%)	P (%)	R (%)
Baseline	-	-	-	59.0	93.4	89.7	88.1
Exp. 1	✓	-	-	59.8	93.9	90.1	89.6
Exp. 2	-	✓	-	60.1	94.1	90.6	89.4
Exp. 3	-	-	✓	61.2	93.8	90.0	89.1
Exp. 4	✓	-	✓	61.8	94.3	91.1	90.0
Exp. 5	-	✓	✓	62.0	94.6	90.3	90.4
Exp. 6	✓	✓	✓	62.6	95.1	91.2	90.9

4.5.2. Effectiveness of Different Stem Designs

Table 7 compares various stem designs built upon the DEIM-N baseline. RepStem [47] obtains 58.5% mAP and 88.6% recall, which are lower than those of the other stem variants. LOGStem [48] reaches 59.4% mAP but requires 15.4G FLOPs. SRFD [49] achieves the highest precision of 90.3%, whereas its mAP, mAP_{50}, and recall remain below those of

WTStem. Overall, WTStem achieves the highest mAP, mAP_{50}, and recall, reaching 59.8%, 93.9%, and 89.6%, respectively, while maintaining the second-highest precision of 90.1%.

Table 7. Accuracy and computational cost of different stem designs. Bold and underline indicate the best and second-best results. The downward arrow indicates that lower values are better.

Method	mAP (%)	mAP_{50} (%)	P (%)	R (%)	Params (M) ↓	FLOPs (G) ↓
DEIM-N	59.0	93.4	89.7	88.1	**3.72**	<u>7.1</u>
SRFD	<u>59.4</u>	<u>93.8</u>	**90.3**	<u>89.2</u>	**3.72**	7.3
LOGStem	<u>59.4</u>	93.5	89.8	89.1	3.76	15.4
RepStem	58.5	93.2	89.7	88.6	**3.72**	**6.7**
WTStem	**59.8**	**93.9**	<u>90.1</u>	**89.6**	<u>3.74</u>	7.5

Feature maps produced by the baseline stem and WTStem are visualized to examine their ability to retain target cues during initial encoding. As shown in Figure 8, the baseline feature maps exhibit blurred object contours, indicating that fragile thermal structures are partially degraded during early encoding. In contrast, WTStem preserves more continuous structural morphology of tiny targets through spatial–frequency decoupling. The difference maps show that the improvements are mainly concentrated on high-frequency details and edge textures. This visual evidence indicates that combining spatial-to-depth rearrangement with high- and low-frequency decoupling at the input stage mitigates early information decay and retains sharper edge representations for subsequent feature extraction.

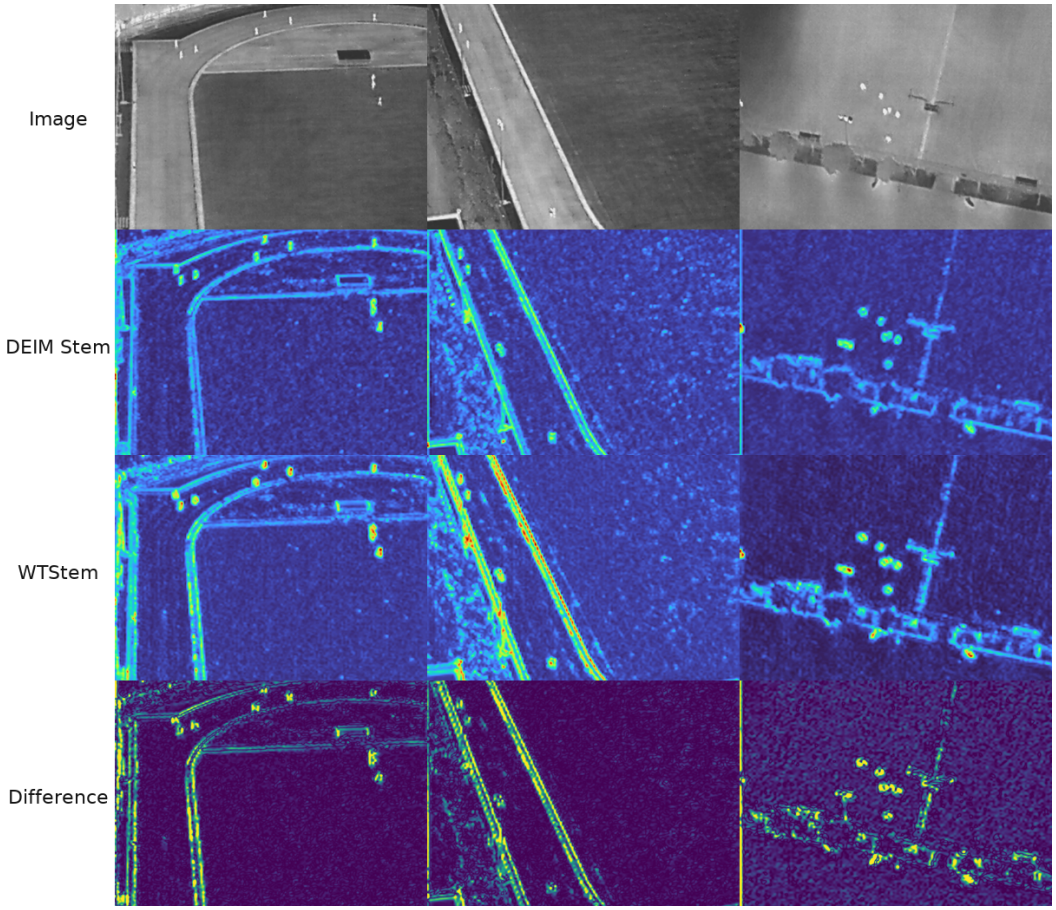

Figure 8. Visual comparison of initial signal preservation between the baseline stem and WTStem.

4.5.3. Feature Pyramid Network Comparative Experiment

As shown in Table 8, the proposed SFCFP achieves the best overall detection performance among the compared feature fusion modules. Compared with the DEIM-N baseline, SFCFP improves mAP from 59.0% to 61.2%, corresponding to a gain of 2.2% points, while mAP_{50} increases from 93.4% to 93.8%. These gains demonstrate that SFCFP improves detection accuracy across different IoU thresholds. Among the compared feature pyramid variants, SFCFP records the highest mAP. Although GoldYOLO reaches the same mAP_{50} and recall, its mAP is 1.4% points lower than that of SFCFP. HS-FPN and HyperACE also remain below SFCFP in mAP. Overall, these results demonstrate that spatial–frequency coordination strengthens multi-scale feature representation and improves detection accuracy.

Table 8. Accuracy and computational cost of different feature pyramid network designs. Bold and underline indicate the best and second-best results. The downward arrow indicates that lower values are better.

Method	mAP (%)	mAP_{50} (%)	P (%)	R (%)	Params (M) ↓	FLOPs (G) ↓
DEIM-N	59.0	<u>93.4</u>	89.7	88.1	<u>3.72</u>	**7.1**
HS-FPN	<u>59.9</u>	93.2	88.9	88.6	5.36	10.7
GoldYOLO	59.8	**93.8**	<u>90.0</u>	**89.1**	4.17	<u>7.2</u>
HyperACE	59.3	<u>93.4</u>	89.9	88.5	4.08	8.1
A3FPN	59.2	93.1	**90.5**	<u>88.7</u>	3.73	**7.1**
SFCFP	**61.2**	**93.8**	<u>90.0</u>	**89.1**	**3.60**	9.3

To more directly clarify the contribution of the S2 spatial distribution path, we further conduct an internal ablation study of SFCFP, as shown in Table 9. Introducing the S2 path alone improves mAP from 59.0% to 60.1%, indicating that redistributing high-resolution spatial cues is beneficial for small-target localization. When the S5 spatial–frequency prior is further incorporated, mAP increases to 60.8%, suggesting that spatial–frequency prior guidance complements local spatial anchors by reducing background ambiguity. After adding the skip injection branch, the full SFCFP achieves the best performance, with 61.2% mAP and 89.1% recall. These results show that the S2 path contributes directly to localization-sensitive representation, while the S5 prior and skip branch provide complementary semantic guidance and intermediate feature reinforcement.

Table 9. Internal ablation study of different pathways in SFCFP on the HIT-UAV dataset.

Method	S2 Path	S5 Prior	Skip Injection	mAP (%)	mAP_{50} (%)	P (%)	R (%)
DEIM-N	-	-	-	59.0	93.4	89.7	88.1
SFCFP w/ S2	✓	-	-	60.1	93.6	89.5	88.6
SFCFP w/ S2+S5	✓	✓	-	60.8	93.8	89.8	88.9
Full SFCFP	✓	✓	✓	61.2	93.8	90.0	89.1

4.5.4. Visual Analysis of Module Interactions

Figure 9 visualizes the feature responses before and after SFCFP under different module combinations. Compared with the baseline, SFCFP alone increases responses around small targets, but some activations are still scattered over background structures. When WTStem is combined with SFCFP, more target-related responses are retained at the SFCFP input, indicating that early spatial–frequency encoding provides richer weak-object cues for subsequent pyramid reconstruction. In contrast, FPL+SFCFP produces cleaner responses with reduced background activation, showing that adjacent-scale purification suppresses clutter before feature aggregation. The full model further combines these advantages: target regions become more continuous and concentrated after SFCFP, while

irrelevant background responses are weakened. These visual comparisons support the progressive preservation–purification–coordination mechanism of SFCD-Det and explain why the complete model achieves the best performance in the ablation study.

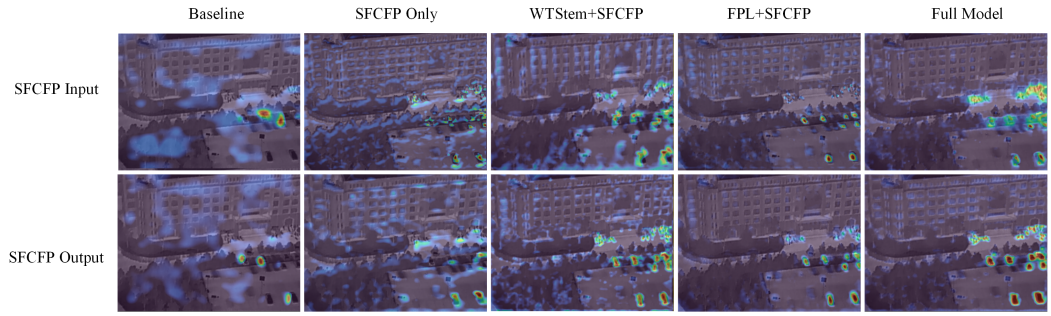

Figure 9. Visual comparison of SFCFP input and output activations under different module combinations.

4.6. Error Analysis

As illustrated in Figure 10, the proposed method consistently improves the correct classification ratios across target categories compared with the DEIM-N baseline. For challenging small targets, the ratio increases from 0.93 to 0.95 for the Person category and from 0.92 to 0.94 for the Bicycle category. The model also reduces missed detections, which correspond to true target classes predicted as the Background category in the rightmost column. Specifically, the false negative ratio of the Person category decreases from 0.07 to 0.04, while that of the Bicycle category decreases from 0.06 to 0.04. This reduction indicates that the spatial–frequency collaborative mechanism better preserves weak infrared structural features and alleviates missed detections in cluttered aerial thermal scenes.

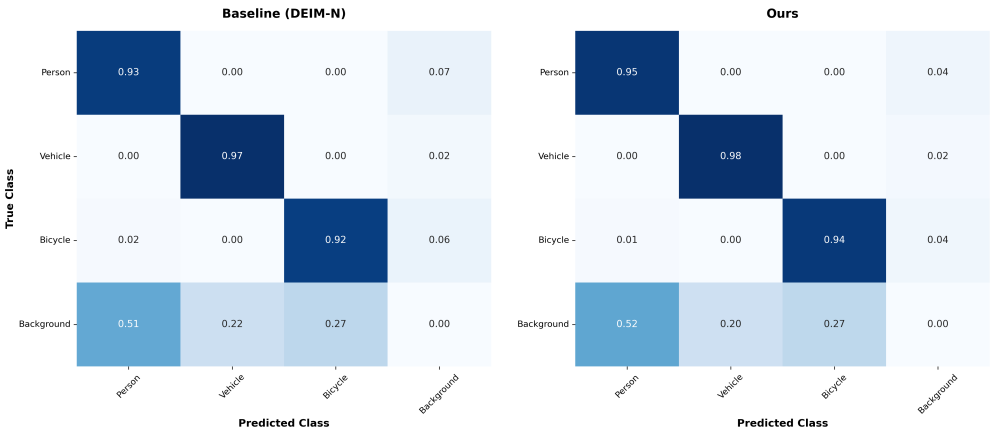

Figure 10. Confusion matrix comparison between the baseline DEIM-N and our proposed method.

Figure 11 presents the TIDE error analysis, where a lower potential AP gain indicates fewer remaining errors of the corresponding type. For the DEIM-N baseline, localization and background errors constitute the primary performance bottlenecks. Localization and background errors decrease from 2.46 to 1.71 and from 1.69 to 1.35, respectively, after incorporating the proposed modules. The overall error summary further reveals that the false negative rate drops from 2.03 to 1.18. These reductions indicate that the proposed architecture improves bounding box localization and helps distinguish fragile infrared targets from complex background clutter.

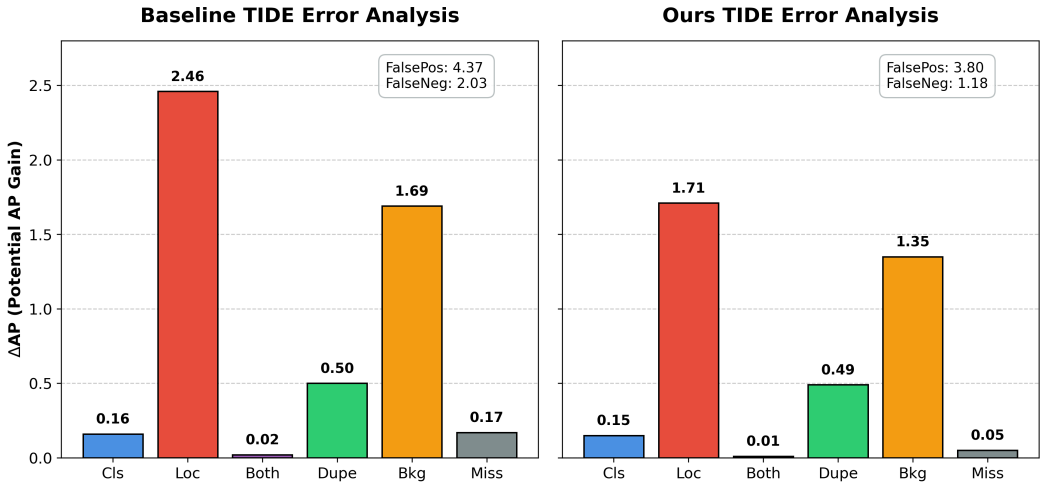

Figure 11. TIDE error analysis comparison between the baseline DEIM-N and our proposed method.

4.7. Statistical Consistency Analysis

To examine the consistency of the experimental results, each model was independently trained and evaluated five times with different random seeds under identical settings. Table 10 presents the mean and standard deviation of mAP, mAP_{50}, precision, and recall. SFCD-Det achieves $62.60 \pm 0.16\%$ mAP on HIT-UAV and $16.76 \pm 0.09\%$ on RGBTDronePerson, exceeding DEIM-S and D-FINE-S, respectively. Its higher mean values across all four metrics on both datasets demonstrate robust detection performance in UAV infrared scenes, including the challenging high-density RGBTDronePerson benchmark.

Table 10. Five-seed stability analysis, reported as mean ± standard deviation.

Dataset	Method	mAP (%)	mAP_{50} (%)	P (%)	R (%)
HIT-UAV	DEIM-S	62.22 ± 0.19	94.44 ± 0.17	90.96 ± 0.27	90.22 ± 0.20
HIT-UAV	SFCD-Det	62.60 ± 0.16	95.06 ± 0.17	91.18 ± 0.31	90.94 ± 0.18
RGBTDronePerson	D-FINE-S	16.52 ± 0.08	45.78 ± 0.19	50.70 ± 0.23	52.90 ± 0.14
RGBTDronePerson	SFCD-Det	16.76 ± 0.09	46.76 ± 0.15	53.16 ± 0.24	53.62 ± 0.16

4.8. Visual Analysis

To intuitively examine detection behavior under challenging scenes, we present qualitative results on four benchmarks. These visual comparisons offer a direct view of the detection differences among methods under representative challenging cases. Figure 12 compares detection performance on the HIT-UAV dataset across six representative challenging scenarios. As highlighted by red dashed circles, baseline models like YOLOv8-N, FBRT-YOLO, and DEIM-S struggle with severe occlusions and extremely small scales, leading to frequent missed detections in dense crowds and distant views. Furthermore, DEIM-N and D-FINE-N occasionally generate false positives by misclassifying background artifacts, denoted by yellow dashed circles. In contrast, our proposed method reduces missed detections in heavily clustered areas and suppresses false alarms, showing robust localization across diverse aerial scenes. These visual results are consistent with the quantitative improvements in recall and category-level AP, especially for small person and bicycle targets.

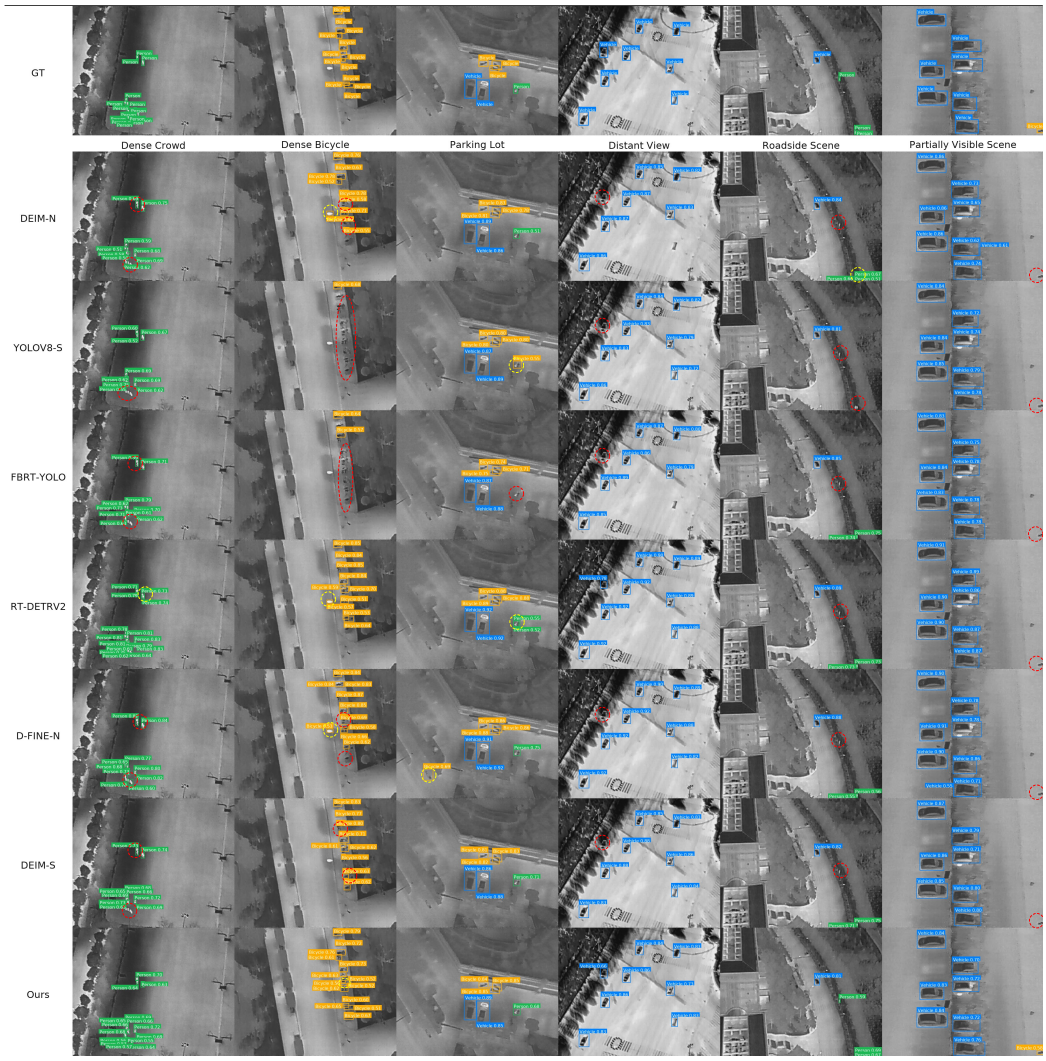

Figure 12. Visual detection examples on HIT-UAV under typical UAV infrared scenarios. Red circles indicate visually identifiable missed detections, while yellow circles denote false predictions.

Figure 13 evaluates the detection performance on the IRSTD-1K dataset, using red and yellow dashed circles to indicate missed targets and false positives, respectively. Baseline models like FBRT-YOLO and DEIM-N exhibit noticeable missed detections when targets are submerged in complex backgrounds with low thermal contrast. Conversely, models such as D-FINE-N and DEIM-S are prone to generating false positives by misclassifying local clutter. By integrating early signal preservation, intermediate feature purification, and coordinated multi-scale reconstruction, our proposed SFCD-Det effectively reduces both errors, yielding more reliable localization. This observation indicates that the three spatial–frequency modules jointly help separate weak infrared targets from severe background interference.

Figure 14 assesses the RGBTDronePerson dataset, characterized by extremely dense clustering and very tiny pedestrians. Using true positive and false positive metrics for scene-level comparison, we observe that YOLOv8-S and DEIM-N suffer from substantial missed detections in heavily populated scenarios where ground truth instances approach a hundred. Conversely, our method separates closely packed instances more effectively, increasing true positives and outperforming the baselines in highly crowded scenes. Furthermore, in sparse but challenging environments, our architecture consistently recalls genuine targets while maintaining near-zero false positives, confirming its capability to handle severe instance clutter. The visual comparisons indicate that the proposed method

improves detection not only by reducing missed targets and false alarms, but also by maintaining clearer target structures across dense, tiny, and low-contrast scenes.

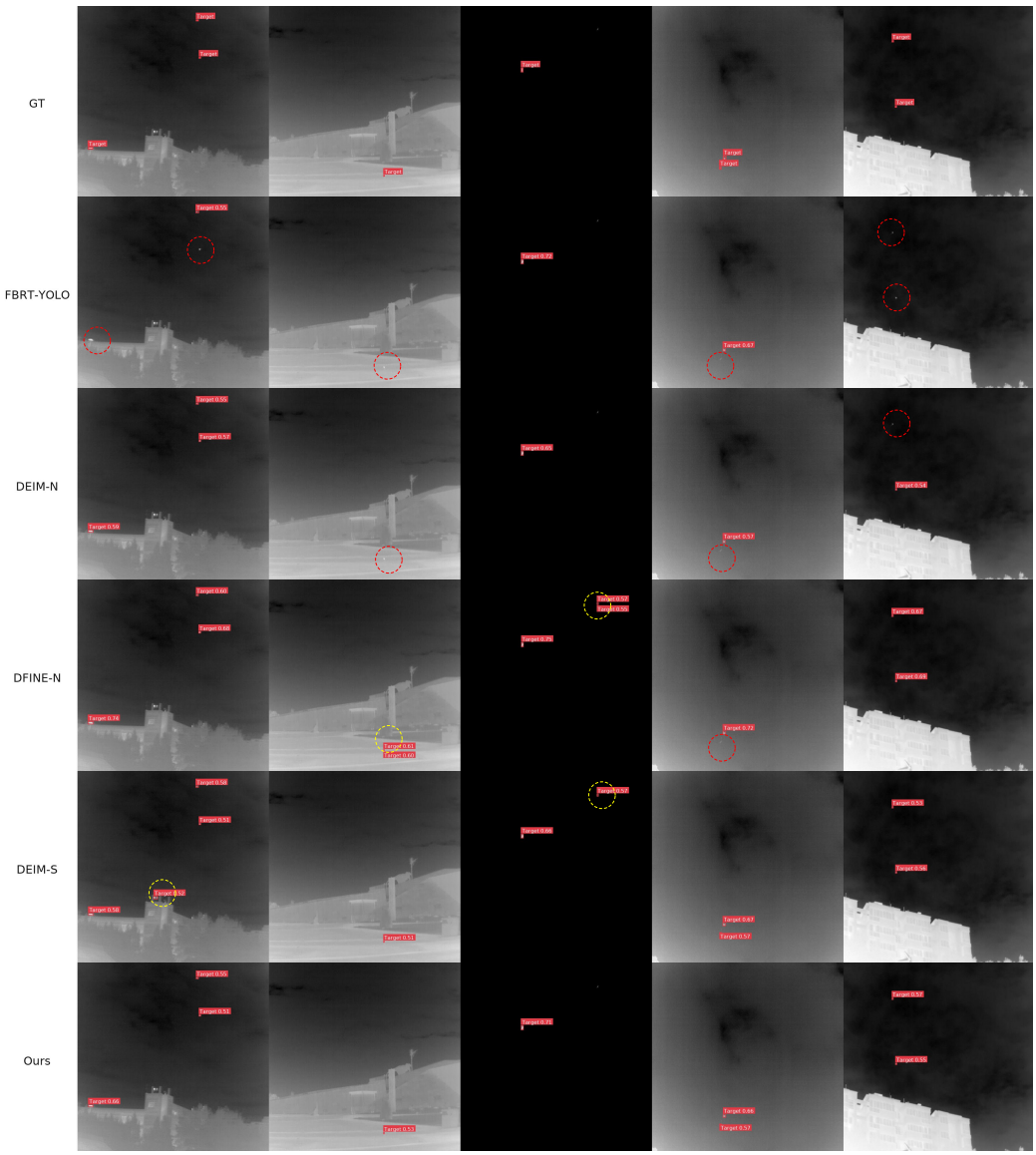

Figure 13. Visual detection examples on IRSTD-1K under cluttered infrared backgrounds. Red circles indicate visually identifiable missed detections, yellow circles denote false predictions, and white circles indicate detected objects whose predicted bounding boxes do not fully cover the entire object.

Figure 15 provides complementary evidence on the visible-light USOD dataset. Although USOD is not an infrared benchmark, many targets are extremely small and appear under low illumination or shadow occlusion, resulting in weak and spatially localized responses similar to those in degraded thermal scenes. As highlighted by the red dashed circles, both DEIM-N and SFCD-Det still miss some vehicles near dark backgrounds, but SFCD-Det recalls more weak targets and produces more robust detection results. This observation suggests that the proposed frequency mechanism is not only related to the lack of texture and color in infrared imagery, but may also help preserve weak localized cues when visible-light target evidence is degraded. Nevertheless, this work centers on UAV infrared detection, with USOD included only to examine cross-modal behavior.

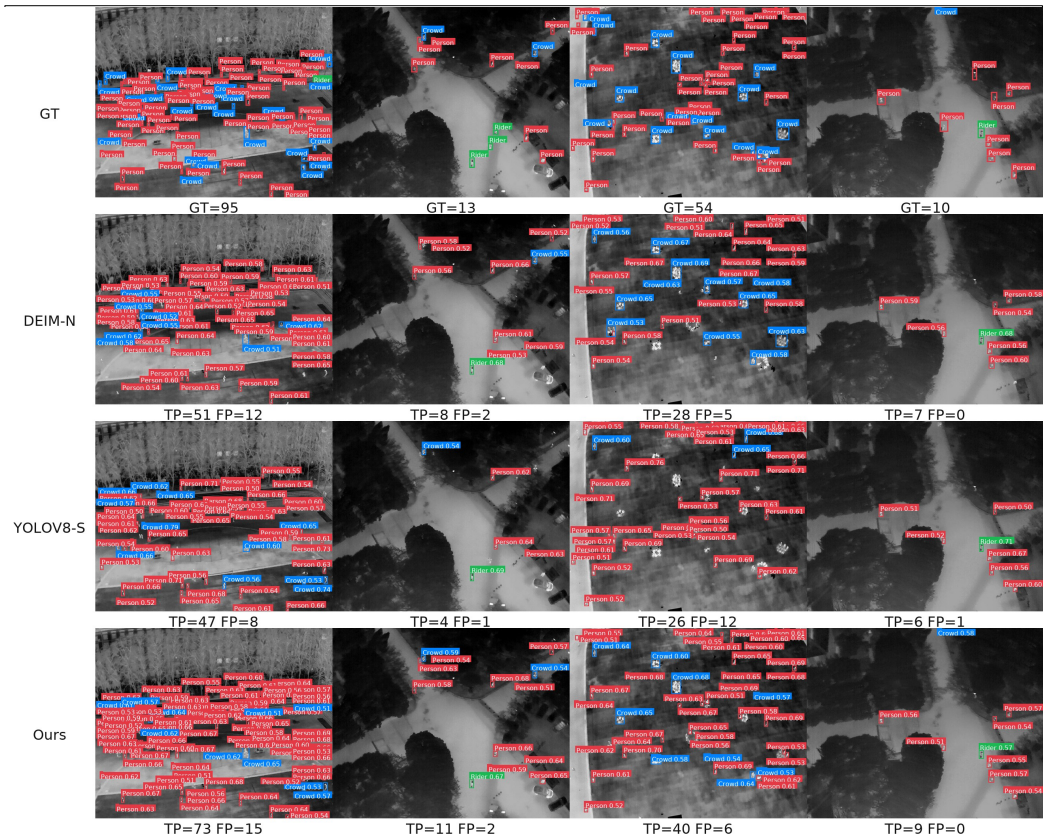

Figure 14. Visual detection examples on the RGBTDronePerson dataset are provided to intuitively analyze model predictions in terms of TP and FP detection cases.

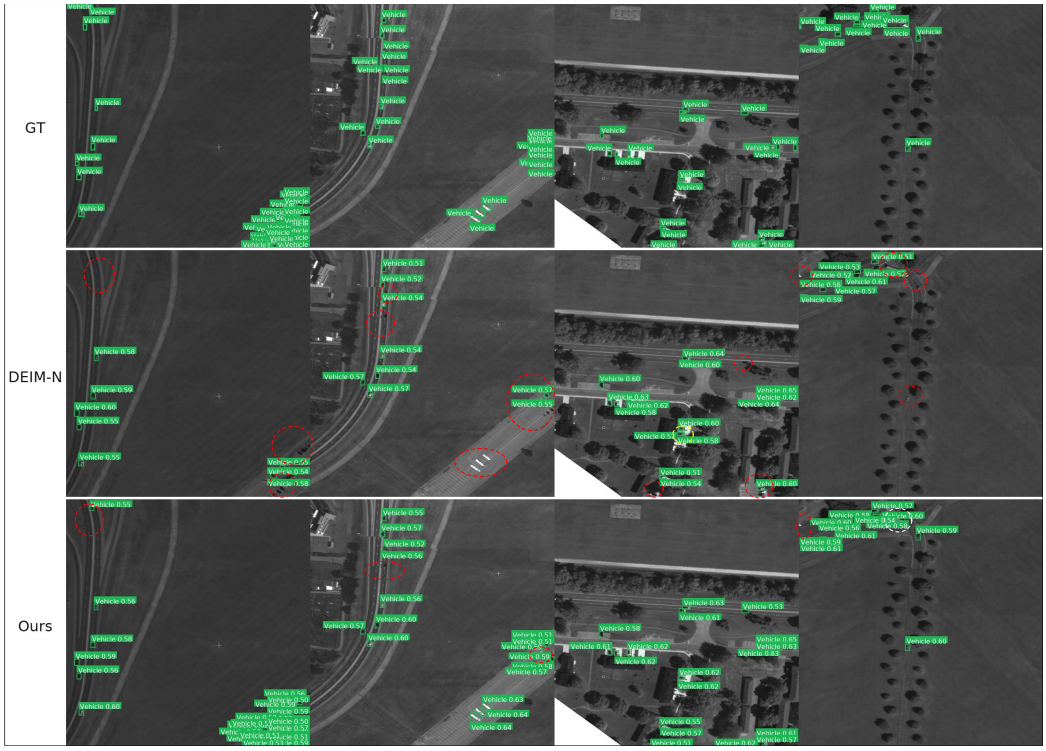

Figure 15. Visual detection examples on USOD under low illumination and shadow occlusion. Red circles indicate visually identifiable missed detections, yellow circles denote false predictions, and white circles indicate detected objects whose predicted bounding boxes do not fully cover the entire object.

5. Conclusions

This paper presents SFCD-Det, a spatial–frequency collaborative architecture for UAV infrared small-object detection. The proposed framework integrates WTStem for early spatial–frequency preservation, FPL for adjacent-scale feature purification, and SFCFP for coordinated multi-scale reconstruction. Experiments on HIT-UAV, IRSTD-1K, and RGBTDronePerson show that SFCD-Det improves weak-target representation in UAV-view infrared scenes, extremely small infrared target detection, and dense thermal pedestrian detection. The evaluation on USOD further suggests that the proposed frequency mechanism is not restricted to thermal degradation, but may also benefit weak and localized visible-light responses under low illumination or shadow. Overall, these results indicate that spatial–frequency collaboration is effective for small targets with weak contrast, cluttered backgrounds, and limited spatial evidence, while UAV infrared detection remains the primary focus of this study.

Despite these promising results, several limitations remain. Since SFCD-Det addresses fragile cue attenuation through a progressive three-stage pipeline rather than an end-to-end single module, its structural complexity is slightly increased. WTStem aims to reduce cue loss at the source by preserving spatial–frequency details before backbone encoding. FPL then suppresses clutter-dominated adjacent-scale responses before pyramid reconstruction, reducing the risk that background interference is propagated during aggregation. SFCFP further compensates for the remaining attenuation by reorganizing shallow spatial anchors, purified intermediate responses, and spatial–frequency priors during multi-scale reconstruction. Therefore, the method may still struggle when the original target response is extremely weak or severely submerged in clutter. In addition, the robustness analysis is currently based on selected benchmarks and representative comparison models, and broader validation across more datasets, detector scales, and deployment settings remains necessary. The computational analysis is also mainly based on parameter counts and FLOPs, which do not fully reflect hardware-dependent inference efficiency. Future work will include broader robustness evaluations, report practical inference latency, and extend the framework to oriented object detection.

Author Contributions: Conceptualization, Y.L.; Methodology, Y.L.; Software, Y.L.; Writing—original draft, Y.L.; Visualization, Y.L.; Supervision, Y.H.; Project administration, Y.H.; Funding acquisition, Y.H.; Validation, L.J.; Investigation, L.J.; Data curation, Q.R.; Formal analysis, Q.R. and D.L.; Resources, D.L.; Writing—review and editing, D.L.; All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by the Science and Technology Support Plan Project of Guizhou (QKHZC [2022] 267).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed in this study are publicly available. The HIT-UAV dataset can be accessed at https://github.com/suojiashun/HIT-UAV-Infrared-Thermal-Dataset (accessed on 21 August 2025). The RGBTDronePerson dataset is available at https://github.com/NNNNerd/mmdet-rgbtdroneperson (accessed on 22 April 2026). The IRSTD-1K dataset can be obtained via https://github.com/RuiZhang97/ISNet (accessed on 14 January 2026). The USOD dataset is provided at https://github.com/yemu1138178251/FFCA-YOLO (accessed on 3 May 2026). All datasets were utilized in strict accordance with the terms and licenses stipulated by their respective original providers.

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

References

1. Borja-Jaimes, V.; Valdez-Martínez, J.S.; Beltrán-Escobar, M.; Ramírez-Zúñiga, G.; Reyes-Mayer, A.; Calixto-Rodríguez, M. Robust Backstepping-Sliding Control of a Quadrotor UAV with Disturbance Compensation. *Computation* **2026**, *14*, 51. [CrossRef]
2. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
3. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
4. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162. [CrossRef]
5. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 28 August 2025).
6. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:cs.CV/2502.12524.
7. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. In *Proceedings of the European Conference on Computer Vision (ECCV 2024)*; Springer Nature: Cham, Switzerland, 2025; pp. 1–21.
8. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 107984–108011. [CrossRef]
9. 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 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974. [CrossRef]
10. Lv, W.; Zhao, Y.; Chang, Q.; Huang, K.; Wang, G.; Liu, Y. RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer. *arXiv* **2024**, arXiv:cs.CV/2407.17140.
11. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.; Shum, H.Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
12. Peng, Y.; Li, H.; Wu, P.; Zhang, Y.; Sun, X.; Wu, F. D-FINE: Redefine Regression Task of DETRs as Fine-grained Distribution Refinement. In Proceedings of the Thirteenth International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025.
13. Huang, S.; Lu, Z.; Cun, X.; Yu, Y.; Zhou, X.; Shen, X. DEIM: DETR with Improved Matching for Fast Convergence. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 15162–15171.
14. Zhao, X.; Zhang, W.; Zhang, H.; Zheng, C.; Ma, J.; Zhang, Z. ITD-YOLOv8: An Infrared Target Detection Model Based on YOLOv8 for Unmanned Aerial Vehicles. *Drones* **2024**, *8*, 161. [CrossRef]
15. Han, Y.; Wang, C.; Luo, H.; Wang, H.; Chen, Z.; Xia, Y.; Yun, L. LRDS-YOLO enhances small object detection in UAV aerial images with a lightweight and efficient design. *Sci. Rep.* **2025**, *15*, 22627. [CrossRef] [PubMed]
16. Xiao, N.; Hong, X.; Zheng, Z. BDK-YOLOv8: An Enhanced Algorithm for UAV Infrared Image Object Detection. *IEEE Access* **2024**, *12*, 191129–191139. [CrossRef]
17. Wang, C.; Wang, R.; Wu, Z.; Bian, Z.; Huang, T. YOLO-UIR: A Lightweight and Accurate Infrared Object Detection Network Using UAV Platforms. *Drones* **2025**, *9*, 479. [CrossRef]
18. Wang, S.; Jiang, H.; Li, Z.; Yang, J.; Ma, X.; Chen, J.; Tang, X. PHSI-RTDETR: A Lightweight Infrared Small Target Detection Algorithm Based on UAV Aerial Photography. *Drones* **2024**, *8*, 240. [CrossRef]
19. Liu, Y.; Jin, T.; Hui, B. DISO-DETR: A Detail-Enhanced Infrared Small Object Detection Transformer Based on UAV Imagery. In Proceedings of the 2025 7th International Conference on Software Engineering and Computer Science (CSECS), Taicang, China, 21–23 March 2025; pp. 1–6. [CrossRef]
20. Xu, K.; Song, C.; Xie, Y.; Pan, L.; Gan, X.; Huang, G. RMT-YOLOv9s: An Infrared Small Target Detection Method Based on UAV Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
21. Yao, T.; Pan, Y.; Li, Y.; Ngo, C.W.; Mei, T. Wave-ViT: Unifying Wavelet and Transformers for Visual Representation Learning. In *Proceedings of the European Conference on Computer Vision (ECCV 2022)*; Springer Nature: Cham, Switzerland, 2022; pp. 328–345. [CrossRef]
22. Li, Q.; Yang, Z.; Cheng, J.; Wang, Q. Spatial–Frequency Feature Learning for Infrared Small Target Detection. *IEEE Trans. Aerosp. Electron. Syst.* **2026**, *62*, 4883–4894. [CrossRef]
23. Lin, Z.; Huang, G.; Li, M.; Yuan, X.; Yue, G.; Pun, C.M.; Cheng, L. SFCANet: Channel Attention in Spatial-Frequency Domain for Infrared Small Target Detection. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 13363–13379. [CrossRef]

24. 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*, 1–17. [CrossRef]
25. Liu, Y.; Lin, Z.; Li, B.; Liu, T.; An, W. FM-Net: Frequency-Aware Masked-Attention Network for Infrared Small Target Detection. *Remote Sens.* **2025**, *17*, 2264. [CrossRef]
26. Pan, J.; Chen, X.; Dong, Z.; Zhang, M.; Guo, H. Edge-Prior Guided Dual-Branch Enhancement Network for Infrared Small Target Detection. *Appl. Sci.* **2026**, *16*, 2929. [CrossRef]
27. Li, R.; Xiao, C.; Yin, Q.; An, W.; Chen, N.; Ying, X.; Li, M.; Wang, Y. Dynamic High-Frequency Convolution for Infrared Small Target Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2026**, *36*, 7676–7680. [CrossRef]
28. Zhu, Y.; Ma, Y.; Fan, F.; Huang, J.; Yao, Y.; Zhou, X.; Huang, R. Toward Robust Infrared Small Target Detection via Frequency and Spatial Feature Fusion. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15. [CrossRef]
29. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
30. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787. [CrossRef]
31. Yang, G.; Lei, J.; Tian, H.; Feng, Z.; Liang, R. Asymptotic Feature Pyramid Network for Labeling Pixels and Regions. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 7820–7829. [CrossRef]
32. Wang, C.; He, W.; Nie, Y.; Guo, J.; Liu, C.; Wang, Y.; Han, K. Gold-YOLO: Efficient Object Detector via Gather-and-Distribute Mechanism. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 51094–51112.
33. 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, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 6896–6904. [CrossRef]
34. 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:cs.CV/2506.17733.
35. Qin, M.; Song, Y.; Zhao, Q.; Yang, X.; Che, Y.; Yang, X. A3-FPN: Asymptotic content-aware pyramid attention network for dense visual prediction. *Pattern Recognit.* **2026**, *179*, 113793. [CrossRef]
36. Du, Z.; Hu, Z.; Zhao, G.; Jin, Y.; Ma, H. Cross-Layer Feature Pyramid Transformer for Small Object Detection in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–14. [CrossRef]
37. Zhou, S.; Pan, J.; Shi, J.; Chen, D.; Qu, L.; Yang, J. Seeing the Unseen: A Frequency Prompt Guided Transformer for Image Restoration. In *Proceedings of the European Conference on Computer Vision (ECCV 2024)*; Springer Nature: Cham, Switzerland, 2025; pp. 246–264.
38. Sun, H.; Wang, R.; Li, Y.; Yang, L.; Lin, S.; Cao, X.; Zhang, B. SET: Spectral Enhancement for Tiny Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 4713–4723.
39. Yuan, M.; Meng, D.; Xi, Z.; Zhao, T.; Zhao, S.; Dai, Y.; Wei, X. Seeing Through the Noise: Improving Infrared Small Target Detection and Segmentation from Noise Suppression Perspective. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Denver, CO, USA, 3–7 June 2026; pp. 27783–27792.
40. 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]
41. Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; Guo, J. ISNet: Shape Matters for Infrared Small Target Detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 867–876. [CrossRef]
42. Zhang, Y.; Xu, C.; Yang, W.; He, G.; Yu, H.; Yu, L.; Xia, G.S. Drone-based RGBT tiny person detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *204*, 61–76. [CrossRef]
43. Zhang, Y.; Ye, M.; Zhu, G.; Liu, Y.; Guo, P.; Yan, J. FFCA-YOLO for Small Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15. [CrossRef]
44. Zhao, X.; Xia, Y.; Zhang, W.; Zheng, C.; Zhang, Z. YOLO-ViT-Based Method for Unmanned Aerial Vehicle Infrared Vehicle Target Detection. *Remote Sens.* **2023**, *15*, 3778. [CrossRef]
45. Xiao, Y.; Xu, T.; Xin, Y.; Li, J. FBRT-YOLO: Faster and Better for Real-Time Aerial Image Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 8673–8681. [CrossRef]
46. Liu, B.; Jiang, Q.; Wang, P.; Yao, S.; Zhou, W.; Jin, X. IRMSD-YOLO: Multiscale Dilated Network With Inverted Residuals for Infrared Small Target Detection. *IEEE Sens. J.* **2025**, *25*, 16006–16019. [CrossRef]

47. Vasu, P.K.A.; Gabriel, J.; Zhu, J.; Tuzel, O.; Ranjan, A. FastViT: A Fast Hybrid Vision Transformer using Structural Reparameterization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 5762–5772. [CrossRef]
48. Lu, W.; Chen, S.B.; Li, H.D.; Shu, Q.L.; Ding, C.H.Q.; Tang, J.; Luo, B. LEGNet: A Lightweight Edge-Gaussian Network for Low-Quality Remote Sensing Image Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Honolulu, HI, USA, 19–20 October 2025; pp. 2865–2874.
49. Lu, W.; Chen, S.B.; Tang, J.; Ding, C.H.Q.; Luo, B. A Robust Feature Downsampling Module for Remote-Sensing Visual Tasks. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 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.