MARSNet: Multimodal Adaptive Refined Small Object Detection with Fusion Enhanced by Omni Perception

Feifei Xu[a], Lupeng Sun[a], Dongyang Li[a,*], Guoxiang Wu[a], Chenchuan Lv[a]

[a]*Shanghai University of Electric Power, Shanghai, China*

Abstract

In real world applications such as uncrewed aerial vehicle (UAV) surveillance and autonomous driving, detecting tiny or distant objects is a fundamental operational requirement. To ensure robust, round the clock perception in these complex scenarios, fusing visible and thermal (RGB-T) modalities has emerged as a crucial solution. However, while existing RGB-T detectors improve general robustness by exploiting complementary multimodal cues, reliable RGB-T small object detection remains a formidable practical challenge. Because tiny objects occupy merely a few pixels, their discriminative evidence is highly vulnerable to degradation throughout the detection pipeline; furthermore, rescuing this fragile evidence while satisfying strict real time inference demands is notoriously difficult. In this paper, we identify and explicitly address a Coupled Degradation Chain in RGB-T small object detection, which encompasses early spatial evidence erosion, modality specific noise propagation, and cross scale localization distortion. To overcome this cascading degradation, we propose MARSNet, a novel Multimodal Adaptive Refined Small Object Detection framework empowered by a fusion strategy enhanced by omni perception. MARSNet systematically refines small object evidence through three harmonized components. First, the Dual Scale Object Enhancement (DualSOE) module fortifies spatial detail perception by transferring structure preserving details from early layers directly to stride aligned semantic features. Second, the Adaptive Fusion (AF) unit elevates modality reliability perception by estimating hierarchical modality quality and adaptively regulating RGB-T aggregation. Third, the Aliasing Resistant Feature Pyramid (ARFP) enhances cross scale consistency perception by drastically mitigating resampling induced aliasing and distortion within the detection neck. Comprehensive empirical evaluations across four challenging RGB-T detection benchmarks, including RGBTDronePerson, M^3FD, LLVIP, and KAIST, demonstrate that MARSNet achieves a superior accuracy and efficiency balance. Notably, under the rigorous RGBTDronePerson official protocol, MARSNet achieves an outstanding 49.89% overall mAP_{50}^{all} and 50.90% tiny object mAP_{50}^{tiny} at 35.48 frames per second (FPS).

Keywords: RGB-T object detection, Small-object detection, Reliability-aware multimodal fusion, Cross-scale feature aggregation, UAV visual perception

1. Introduction

Intelligent perception systems deployed in safety critical applications, such as uncrewed aerial vehicle (UAV) surveillance and autonomous driving, strictly demand robust, round the clock operational capabilities. In these expansive real world scenarios, objects frequently appear at extreme distances within wide fields of view, establishing small object detection as an indispensable yet highly demanding prerequisite. To fulfill this all weather requirement, visible and thermal (RGB-T) object detection has witnessed widespread adoption, as it effectively exploits fine grained appearance details from visible imagery alongside illumination invariant contours from thermal radiation. However, although existing RGB-T fusion methods have made significant strides in general object detection, achieving reliable RGB-T small object detection remains a pressing practical dilemma. This challenge is acutely magnified by strict deployment constraints: tiny objects typically occupy merely a few pixels, making their discriminative evidence extremely fragile and easily eroded, while practical systems must simultaneously rescue these fragile features and maintain rapid processing speeds. As shown in Fig. 1, a deployable RGB-T small object detector must therefore rigorously strike an optimal balance between high precision detection, real time inference efficiency, and moderate model complexity.

For RGB-T small object detection, the primary challenge lies in the inherent fragility of small object evidence. When objects occupy only a few pixels, their discriminative information is typically characterized by weak boundaries, sparse textures, and limited thermal responses. Although such small object evidence may initially exist in shallow features, it is frequently progressively attenuated during early downsampling, contaminated by unreliable multimodal fusion, and distorted by repeated cross scale resampling. We conceptualize this process as the Coupled Degradation Chain, which comprises three closely related stages: early spatial evidence erosion, modality specific noise propagation, and cross scale localization distortion. This chain renders RGB-T small object detection inherently more challenging than general RGB-T object detection, compelling the detector to meticulously preserve, calibrate, and propagate small

*Corresponding author.
Email address: dongyangli.ldy@shiep.edu.cn (Dongyang Li)

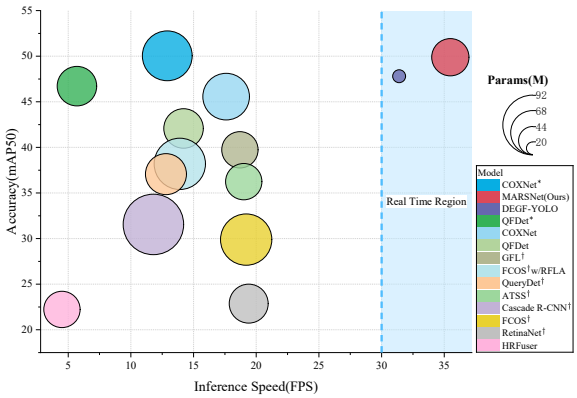

Figure 1: Accuracy and efficiency balance on the RGBTDronePerson benchmark. Each bubble denotes a detector listed in Table 1. The horizontal and vertical axes correspond to inference speed (FPS) and detection accuracy (overall mAP_{50}^{all} under the RGBTDronePerson protocol), respectively, and bubble area is proportional to parameter count (M). The shaded region to the right of the dashed line indicates real time processing (> 30 FPS). Within this regime, MARSNet achieves 49.89% mAP_{50}^{all} at 35.48 FPS.

object evidence across the entire pipeline. To provide visual evidence for the first stage of the Coupled Degradation Chain, Fig. 2 compares the spectral responses produced by standard strided downsampling with those of the structure preserving P_2 to P_3 transfer used in this work. Standard downsampling suppresses a substantial amount of high frequency information, which is crucial for representing small objects with weak boundaries and sparse textures. In contrast, the structure preserving P_2 to P_3 transfer retains stronger peripheral responses in the frequency domain, indicating that finer spatial evidence remains accessible prior to multimodal fusion. This observation elucidates why small objects may be captured in early layers but subsequently lose representational reliability or even disappear after fusion and cross scale propagation. Therefore, the fundamental challenge of RGB-T small object detection is not merely how to combine RGB and T modalities, but rather how to systematically preserve, calibrate, and propagate small object evidence across the entire detection pipeline.

Existing RGB-T object detection and small object detection methods [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17] have achieved significant advancements from different perspectives. RGB-T fusion methods [1, 2, 3, 9, 10, 18] improve multimodal representation by combining complementary RGB and T features. Early feature level fusion methods [1] demonstrated that fusing RGB and T features can improve detection robustness compared with single modality perception. Subsequent methods further introduced cross modal attention [2, 3, 4], dynamic convolution [9], frequency domain modeling [10], and target aware fusion [18] to enhance modality interaction. These methods effectively strengthen RGB-T complementarity and have become the prevailing paradigm for RGB-T object detection. However, most of them perform fusion after backbone feature extraction, where small object evidence may have already been weakened by early downsampling. For RGB-T small

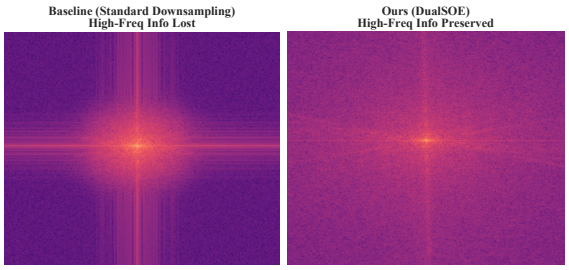

Figure 2: Evidence for early spatial evidence erosion in RGB-T small object detection. Log magnitude spectra of feature maps generated by standard strided downsampling and structure preserving P_2 to P_3 transfer. Compared with standard downsampling, the structure preserving transfer retains stronger high frequency responses, suggesting that more fine spatial evidence remains available for small object localization before multimodal fusion. This visualization provides qualitative supporting evidence rather than standalone quantitative proof.

object detection, such fusion after feature extraction is insufficient because the detector may attempt to fuse features in which weak boundaries and sparse textures are already irreparably lost. RGB-T alignment methods [5, 6, 7] provide another important direction by mitigating spatial inconsistency between RGB and T modalities. Registration based methods [5, 6] improve cross modal spatial correspondence, while RGB-T small object detectors such as COXNet [7] introduce adaptive alignment and scale integration to alleviate feature mismatch. These methods are valuable because spatial misalignment can severely degrade RGB-T detection, especially when objects are small. Nevertheless, alignment alone cannot fully resolve the Coupled Degradation Chain. When small object evidence has already been eroded by downsampling or contaminated by modality specific noise, aligning two degraded feature maps inevitably yields unreliable fused features. Therefore, RGB-T small object detection requires not only cross modal correspondence, but also small object evidence preservation before fusion and modality reliability calibration during fusion. Small object detection methods [8, 12, 13, 14, 19] focus on preserving fine spatial details or improving small object localization. QueryDet [8] accelerates high resolution small object detection through cascaded sparse queries, while SliceSamp [12], Cut Slicing [13], and Space to Depth Convolution (SPDConv) [19] attempt to reduce information loss during downsampling. Semantic aware designs such as SADet [14] further enhance small object representation. These methods underscore the importance of preserving fine spatial evidence. However, most of them are primarily designed for general or single modality detection. In RGB-T scenarios, preserving shallow details is only one part of the problem, because the detector must also handle modality specific noise and varying reliability between RGB and T features. Directly introducing high resolution branches or generic downsampling replacements often inflates computational costs and may not synergize well with multimodal fusion. Feature pyramid and cross scale aggregation methods [15, 16, 17] improve multiscale representation by propagating semantic and localization information across different feature levels. The Feature Pyramid Network (FPN) [15] and its variants, such as bidirectional feature pyramid designs [16],

have become standard architectural components for handling scale variation. Anti aliased convolutional designs [17] further demonstrate that resampling operations can compromise feature stability. However, most detection necks still treat upsampling and downsampling as standard geometric operations, paying inadequate attention to the aliasing and localization distortion caused by repeated cross scale resampling. For RGB-T small object detection, this issue is especially critical because small object evidence is spatially weak and highly sensitive to feature misalignment and resampling artifacts. The preceding analysis demonstrates that existing methods typically address only isolated parts of the RGB-T small object detection challenge. RGB-T fusion methods improve modality interaction, but often operate after early spatial evidence has been weakened. RGB-T alignment methods reduce cross modal mismatch, but do not necessarily recover lost small object evidence. Small object detection methods preserve fine details, but rarely consider modality specific noise in RGB-T scenarios. Feature pyramid methods enhance cross scale aggregation, but often overlook resampling induced localization distortion. These limitations indicate that RGB-T small object detection requires an omni perception of small object evidence, including spatial detail perception before fusion, modality reliability perception during fusion, and cross scale consistency perception after fusion.

Motivated by the preceding analysis, we propose MARSNet[1], a Multimodal Adaptive Refined Small Object Detection framework featuring a fusion strategy enhanced by omni perception. In this work, the fusion strategy enhanced by omni perception refers to a unified fusion pipeline that preserves, calibrates, and propagates small object evidence through three coordinated perception processes. First, we design a Dual Scale Object Enhancement (DualSOE) module to fortify spatial detail perception by transferring structure preserving details from early layers to stride aligned semantic features. Second, we introduce an Adaptive Fusion (AF) unit to enhance modality reliability perception by estimating hierarchical modality quality and adaptively regulating RGB-T aggregation. Third, we develop an Aliasing Resistant Feature Pyramid (ARFP) to bolster cross scale consistency perception by mitigating resampling induced distortion in the detection neck. Through the synergy of these three coordinated components, MARSNet refines small object evidence across the entire detection pipeline, rather than treating multimodal fusion as an isolated operation. The main contributions of this paper are summarized as follows:

- We systematically analyze RGB-T small object detection from the perspective of small object evidence degradation and identify the Coupled Degradation Chain, which encompasses early spatial evidence erosion, modality specific noise propagation, and cross scale localization distortion.
- We propose MARSNet, a Multimodal Adaptive Refined Small Object Detection framework empowered by a fusion strategy enhanced by omni perception, which jointly integrates spatial detail perception, modality reliability perception, and cross scale consistency perception for RGB-T small object detection.
- We design three coordinated components for MARSNet: DualSOE for preserving fine spatial evidence, AF for adaptive RGB-T modality fusion, and ARFP for stabilizing cross scale feature propagation.
- Extensive experiments on four challenging RGB-T detection datasets, including RGBTDronePerson, M^3FD, LLVIP, and KAIST, demonstrate that MARSNet achieves a superior accuracy and efficiency balance, maintaining real time inference efficiency without sacrificing detection accuracy.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 describes the proposed MARSNet architecture in detail. Section 4 presents the experimental results and comprehensive analyses. Finally, Section 5 concludes the paper.

2. Related Work

2.1. RGB-T Object Detection: Fusion and Alignment

Leveraging the complementary information provided by visible red, green, and blue (RGB) and thermal infrared (T) modalities, RGB-T object detection has emerged as an effective paradigm for robust perception under complex sensing conditions. Early RGB-T detection methods predominantly adopted deterministic fusion strategies. For instance, Liu et al. [1] proposed a two stream backbone wherein RGB and T features were fused via element wise addition or channel concatenation, demonstrating the fundamental efficacy of feature level fusion for multimodal detection. Driven by the increasing demand for robust cross modal representations, subsequent methods introduced attention based or gated fusion mechanisms to augment RGB-T interactions. For example, Fang et al. [2] and Fu et al. [3] incorporated channel spatial interactions and long range dependency modeling, whereas Wang et al. [9] and Zhu et al. [10] introduced dynamic cross scale coupling and frequency domain representations to improve semantic interaction. Target aware fusion methods, such as TFDet [18], further exploited object level cues for cross modal feature aggregation. Collectively, these approaches effectively improved RGB-T complementarity and have established a prominent research trajectory for RGB-T object detection.

Beyond feature fusion, cross modal alignment remains equally critical for RGB-T object detection, as spatial inconsistencies between RGB and T images can significantly degrade feature aggregation. Several methods have attempted to alleviate this issue by explicitly modeling modality offsets or establishing more coherent feature spaces. For instance, alignment oriented RGB-T detectors [5, 6, 7] improved cross modal correspondence through feature calibration, adaptive sampling, or scale aware integration. These techniques are highly valuable, given that inaccurate RGB-T correspondence can severely compromise

[1]The source code will be released upon publication.

detection performance, particularly for small objects. However, existing RGB-T fusion and alignment methods predominantly focus on optimizing feature interactions after extraction. In the context of RGB-T small object detection, crucial small object evidence is frequently already attenuated by early downsampling before fusion or alignment mechanisms can even take effect. Furthermore, when a single modality exhibits noise or unreliable responses, fixed confidence fusion strategies risk propagating modality specific noise into the fused representation. Diverging from these conventional approaches, our MARSNet introduces an Adaptive Fusion (AF) unit to fortify modality reliability perception, wherein modality quality estimation and adaptive RGB-T aggregation are jointly optimized to ensure the reliable fusion of small object evidence.

2.2. Small Object Detection and Spatial Detail Preservation

Small object detection remains an enduring challenge because tiny objects occupy merely a few pixels and provide scarce spatial evidence for accurate recognition and localization. During repeated spatial downsampling operations, fine boundaries, sparse textures, and local structural cues are highly susceptible to attenuation or complete loss. To circumvent this issue, existing literature has extensively explored both data level and architecture level strategies. Data level techniques, such as multiscale training and copy paste augmentation, artificially increase the prevalence of small objects during the training phase. Conversely, architecture level methods typically endeavor to enhance high resolution representations, broaden contextual receptive fields, or introduce auxiliary shallow prediction heads to preserve fine spatial details. For example, QueryDet [8] accelerated high resolution small object detection via cascaded sparse queries. Recent detail preserving downsampling mechanisms, such as SliceSamp [12], Cut Slicing [13], and Space to Depth Convolution (SPDConv) [19], have attempted to mitigate information loss during spatial resolution reduction. Furthermore, semantic aware methods such as SADet [14] further improved small object representations by explicitly enhancing semantic discrimination. Collectively, these works validate the paramount importance of preserving fine spatial evidence for successful small object detection. However, the majority of these architectures are exclusively tailored for general or single modality detection, wherein the detector is solely required to maintain spatial details from a singular visual stream. In the realm of RGB-T small object detection, preserving object evidence is significantly more complex, as shallow RGB and T features frequently exhibit disparate noise patterns, boundary qualities, and reliability levels. Moreover, the direct introduction of high resolution prediction heads inevitably incurs substantial computational overhead and prediction redundancy, both of which are highly undesirable for real time RGB-T deployments. Consequently, RGB-T small object detection demands a more targeted spatial detail preservation strategy—one that fortifies small object evidence prior to multimodal fusion without imposing prohibitive network wide overhead. To this end, we design the Dual Scale Object Enhancement (DualSOE) module specifically for spatial detail perception. This module transfers structure preserving details from early layers to stride aligned semantic features and subsequently compensates them through contextual and frequency aware refinements.

2.3. Feature Pyramid Networks and Cross Scale Consistency

Feature Pyramid Networks (FPNs) [15] and their subsequent variants have become the de facto standard architectures for addressing scale variations in object detection. By seamlessly integrating top down semantic propagation with lateral connections, FPNs substantially enrich multiscale feature representations for objects of varying sizes. Subsequent studies have further enhanced cross scale aggregation by deploying bidirectional pathways, learnable fusion weights, and adaptive feature routing mechanisms. For instance, EfficientDet [16] introduced a bidirectional feature pyramid equipped with learnable fusion weights, enabling highly efficient multiscale feature aggregation. Other investigations have also explored channel refinement, lightweight feature aggregation, and anti aliasing operations [17] to concurrently improve feature stability and detection efficiency. Despite the proven ability of these methods to improve multiscale representations, the majority of detection necks continue to treat upsampling and downsampling merely as standard geometric operations. While this conventional design paradigm may suffice for normal sized objects, it proves inherently unreliable for RGB-T small object detection. Given that small object evidence is spatially fragile, repeated cross scale resampling inevitably introduces severe aliasing artifacts, texture degradation, and localization distortions. These distortions compound to profoundly degrade the fused RGB-T representation, particularly when early spatial evidence has already been eroded and modality specific noise has infiltrated the pipeline. Consequently, cross scale aggregation must transcend the mere fusion of multiresolution features to actively enforce cross scale consistency during the propagation of small object evidence. Deviating from conventional FPN based architectures, MARSNet proposes an Aliasing Resistant Feature Pyramid (ARFP) explicitly designed for cross scale consistency perception. This module rigorously stabilizes feature propagation via anti aliasing upsampling, conservative downsampling, and content aware gated routing.

3. Methodology

3.1. Problem Definition

Let $\mathcal{D}$ denote a spatially aligned RGB-T object detection dataset. To maintain notational consistency with existing RGB and infrared literature, the superscript "ir" designates the thermal (T) branch throughout our mathematical formulations. Each training sample comprises an RGB image, a corresponding thermal image, and a set of object annotations:

$$\begin{aligned} &\mathcal{D} = \left\{(\mathbf{I}_i^{\mathrm{rgb}}, \mathbf{I}_i^{\mathrm{ir}}, \mathbf{Y}_i)\right\}_{i=1}^{N}, \\ &\mathbf{I}_i^{\mathrm{rgb}} \in \mathbb{R}^{H\times W\times 3}, \quad \mathbf{I}_i^{\mathrm{ir}} \in \mathbb{R}^{H\times W\times 1}, \\ &\mathbf{Y}_i = \{(\mathbf{b}_{ij}, c_{ij})\}_{j=1}^{n_i}, \quad \widehat{\mathbf{Y}}_i = f_\theta(\mathbf{I}_i^{\mathrm{rgb}}, \mathbf{I}_i^{\mathrm{ir}}). \end{aligned} \tag{1}$$

where $\mathbf{b}_{ij} \in \mathbb{R}^4$ denotes a 2D bounding box and c_{ij} represents the associated class label [20, 21]. The primary objective is to optimize a parameterized detector f_θ capable of predicting $\widehat{\mathbf{Y}}_i$ from

the aligned RGB-T pair. The principal difficulty stems from the acute sensitivity of small objects to early spatial downsampling. Following l consecutive stride 2 reductions, the feature resolution degrades to $H/2^l \times W/2^l$; consequently, an object originally of size $h \times w$ is mapped onto a mere $h/2^l \times w/2^l$ cells. In standard architectures, the P_5 level possesses a stride of 32, dictating that objects smaller than 32×32 pixels inevitably collapse into subpixel footprints. Within this regime, quantization errors and contextual dilution become overwhelmingly dominant [16]. Concurrently, the signal degradations in both the RGB and T branches exhibit distinct spatial nonstationarity, rendering fixed rule fusion strategies fundamentally inadequate for real world scenarios. Consequently, achieving robust RGB-T small object detection strictly demands both meticulous detail preservation and location adaptive reliability calibration. In this framework, the overarching objective transcends mere detection from aligned RGB-T inputs; it fundamentally involves the rigorous preservation of small object evidence against the Coupled Degradation Chain. Specifically, a superior detector must preserve spatial details prior to multimodal fusion, dynamically calibrate modality reliability during fusion, and rigorously maintain cross scale consistency after fusion. These functional imperatives define the core methodological objectives of MARSNet.

3.2. Overall Architecture of MARSNet

As illustrated in Fig. 3, MARSNet constitutes an omni perception enhanced RGB-T small object detection framework. Operating upon a dual stream feature extraction paradigm, the framework systematically refines small object evidence via three harmonized perception stages: spatial detail perception, modality reliability perception, and cross scale consistency perception. Specifically, the DualSOE module reinforces small object evidence prior to multimodal fusion, the AF unit executes reliability aware aggregation during fusion, and the ARFP detection neck stabilizes cross scale feature propagation subsequent to fusion. Through this architectural synergy, MARSNet confronts the Coupled Degradation Chain via a holistic, pipeline level strategy, rather than treating multimodal fusion as a disjointed, isolated operation.

The overall feedforward pipeline is mathematically formalized as follows:

$$\begin{aligned} \mathcal{P}^m &= \mathcal{B}_m(\mathbf{I}^m) = \{P_2^m, P_3^m, P_4^m, P_5^m\}, \\ F_s^m &= \begin{cases} \mathrm{DualSOE}(P_2^m, P_3^m), & s = 3, \\ P_s^m, & s \in \{4,5\}, \end{cases} \\ F_s^{\mathrm{fus}} &= \mathrm{AF}(F_s^{\mathrm{rgb}}, F_s^{\mathrm{ir}}), \quad s \in \{3,4,5\}, \\ \{P_3^{\mathrm{out}}, P_4^{\mathrm{out}}, P_5^{\mathrm{out}}\} &= \mathrm{ARFP}(F_3^{\mathrm{fus}}, F_4^{\mathrm{fus}}, F_5^{\mathrm{fus}}). \end{aligned} \tag{2}$$

Here, $\mathcal{B}_m$ denotes the modality specific backbone, and P_s^m represents the extracted feature map at a downsampling stride of 2^s [21, 16]. Crucially, we design DualSOE to use P_2 exclusively as a high resolution enhancement source, deliberately avoiding the computational overhead of introducing an additional P_2 prediction head. The AF unit subsequently performs reliability aware fusion across pyramid levels P_3 to P_5, yielding the fused features that ARFP refines into the final cross scale representations for the detection heads. The mapping between the Coupled Degradation Chain and the MARSNet architecture is thus entirely explicit: DualSOE counteracts early spatial evidence erosion through spatial detail perception; AF suppresses modality specific noise propagation through modality reliability perception; and ARFP mitigates cross scale localization distortion through cross scale consistency perception.

3.3. Spatial Detail Perception via DualSOE

As evidenced by the spectral analysis in Fig. 2, standard strided downsampling frequently suppresses the high frequency responses that are paramount for accurate small object localization. To counteract this loss, we design DualSOE to perform spatial detail perception prior to RGB-T fusion. Its primary objective is to safeguard the fragile small object evidence that is otherwise heavily attenuated during early spatial downsampling. Rather than appending an explicit high resolution P_2 detection head, DualSOE defines P_2 as an auxiliary high resolution evidence source and transfers structure preserving details into the stride aligned P_3 feature space. This structural design inherently fortifies small object evidence before multimodal aggregation, while successfully circumventing the heavy computational burden and prediction redundancy typically associated with an extra high resolution detection branch. Traditional backbone networks predominantly rely on strided convolutions and pooling layers to achieve downsampling, which inherently perform local aggregation followed by spatial decimation. From a digital signal processing perspective, downsampling without adequate low pass prefiltering inevitably aliases high frequency components into the baseband; moreover, empirical implementations of strided convolutions or max pooling tend to indiscriminately suppress weak high frequency details. This specific form of distortion disproportionately degrades the representation of small objects, which are primarily defined by high frequency edges, corners, and sparse textures [17]. According to multirate signal processing theory, for a decimation factor of D, the spectrum of the output signal $Y(e^{j\omega})$ satisfies the following aliasing superposition equation:

$$Y(e^{j\omega}) = \frac{1}{D} \sum_{k=0}^{D-1} X\left(e^{j(\omega+2\pi k)/D}\right). \tag{3}$$

Consequently, intermediate features that are not strictly bandlimited inevitably suffer from irreversible spectral overlap. To mitigate this catastrophic loss of structural information, we formulate the $P_2 \to P_3$ transition around a Space to Depth Convolution (SPDConv) operation [19]. The Space to Depth (S2D) operator performs a deterministic spatial to channel rearrangement, after which a subsequent convolution reencodes the inflated channel dimension to match the prerequisite depth of P_3. Within our architectural configuration, a decimation factor of $D = 2$ governs the $P_2 \to P_3$ structural transition. For the sake of brevity and conceptual clarity, Eq. (3) formulates multirate aliasing in the 1D domain; however, this principle naturally generalizes to 2D spatial feature maps. Given $P_2^m \in \mathbb{R}^{C \times H/4 \times W/4}$, the $P_2 \to P_3$

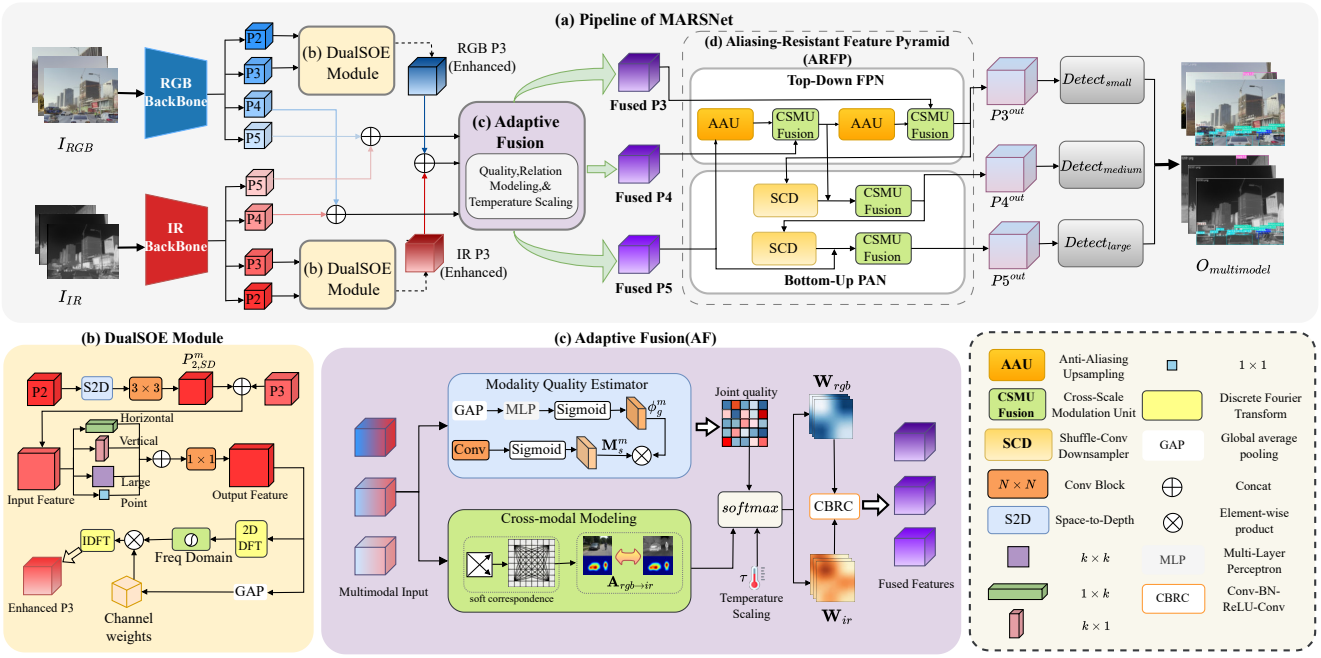

Figure 3: Overview of the MARSNet framework. (a) Dual stream YOLOv8 backbones extract multiscale RGB and T features. (b) DualSOE performs spatial detail perception by transferring structure preserving detail from P2 to P3 through SPDConv, directional context aggregation, and frequency domain channel recalibration. (c) AF performs modality reliability perception on P_3 to P_5 via hierarchical quality estimation, reduced cross modal relation modeling, and adaptive temperature control. (d) ARFP performs cross scale consistency perception by combining neck specific AAU, SCD, and CSMU to preserve localization consistency during feature propagation.

detail transfer stage is formulated as:

$$\begin{aligned} \widetilde{P}_2^m &= \mathcal{T}_{sd}(P_2^m), \\ \widetilde{P}_2^m(:,i,j) &= \begin{bmatrix} P_2^m(:,2i,2j) \\ P_2^m(:,2i+1,2j) \\ P_2^m(:,2i,2j+1) \\ P_2^m(:,2i+1,2j+1) \end{bmatrix}, \\ P_{2,sd}^m &= \mathcal{F}_{\text{conv}}(\widetilde{P}_2^m). \end{aligned} \tag{4}$$

Because $\mathcal{T}_{sd}$ operates deterministically, it explicitly relocates the four subpixel samples within each 2×2 spatial neighborhood into the channel dimension prior to the learnable 3×3 reencoding. While the S2D rearrangement strictly preserves all spatial samples, the ensuing convolution acts as a learnable feature compression and thus should not be conflated with a mathematically lossless propagation guarantee. Although $P_{2,sd}^m$ successfully maps shallow samples into the P_3 stride space, directly injecting these features remains highly susceptible to background clutter. To resolve this, DualSOE selectively reintegrates the transferred details under strict semantic guidance:

$$\begin{aligned} X^m &= \text{Concat}(P_{2,sd}^m, P_3^m), \\ [X_{\text{enh}}^m, X_{\text{id}}^m] &= \text{Split}(X^m), \\ Y_{\text{dir}}^m &= \mathcal{A}\Big(\mathcal{F}_h(X_{\text{enh}}^m), \mathcal{F}_v(X_{\text{enh}}^m), \\ &\qquad \mathcal{F}_l(X_{\text{enh}}^m), \mathcal{F}_p(X_{\text{enh}}^m)\Big), \\ \gamma^m &= \Gamma(\text{GAP}(Y_{\text{dir}}^m)). \end{aligned} \tag{5}$$

Here, $\mathcal{F}_h$, $\mathcal{F}_v$, $\mathcal{F}_l$, and $\mathcal{F}_p$ represent horizontal, vertical, large context, and pointwise convolutional branches, respectively. This compact multi branch structure concurrently extracts elongated structural support and localized channel responses, whereas X_{id}^m maintains an identity pathway to prevent the destructive excessive modification of the base semantic features. The aggregated directional responses are subsequently recalibrated via a lightweight frequency domain channel pathway before being merged back with the identity stream:

$$\begin{aligned} S^m &= \text{DFT}_{2D}(Y_{\text{dir}}^m), \quad \widehat{S}_c^m(u,v) = \gamma_c^m S_c^m(u,v), \\ Y_{\text{freq}}^m &= \Re(\text{IDFT}_{2D}(\widehat{S}^m)), \\ \widehat{X}^m &= \mathcal{F}_{\text{merge}}(\text{Concat}(X_{\text{id}}^m, Y_{\text{freq}}^m)), \\ F_3^m &= \mathcal{F}_{\text{out}}(\widehat{X}^m). \end{aligned} \tag{6}$$

Because the scalar γ^m is applied channel wise rather than to specific frequency bins, this operation is correctly interpreted as a frequency domain channel recalibration, distinguishing it from explicit frequency band filtering. The resulting tensor F_3^m effectively restores the transferred fine structural details under robust semantic guidance, subsequently serving as the high fidelity input for downstream cross modal fusion.

3.4. Modality Reliability Perception via Adaptive Fusion

Following the spatial detail perception stage, the Adaptive Fusion (AF) module is introduced to perform dynamic modality reliability perception during the RGB-T fusion phase. In authentic RGB-T environments, the informational reliability of the RGB and T modalities invariably fluctuates across different spatial locations. For instance, RGB features are highly susceptible to severe degradation under poor illumination, whereas T

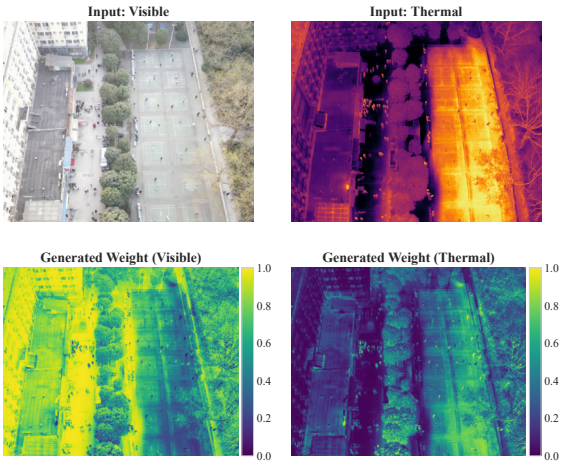

Figure 4: Visualization of modality reliability perception in AF. Top: input RGB and thermal images (T). Bottom: spatial weight maps predicted by AF, showing how the model adjusts modality confidence according to local scene quality and preserves complementary small object evidence.

features often suffer from blurred object boundaries and internal thermal noise. Consequently, traditional fixed confidence fusion mechanisms are fundamentally unsuited for the demands of RGB-T small object detection. The AF module estimates hierarchical modality quality and adaptively regulates RGB-T aggregation, actively suppressing unreliable modality specific noise while faithfully preserving complementary cross modal evidence. In our framework, we execute the fusion of RGB and T features across the feature pyramid levels $s \in \{3, 4, 5\}$. Let $F_s^m \in \mathbb{R}^{C_s \times H_s \times W_s}$ denote the feature map for modality $m \in \{\text{rgb}, \text{ir}\}$ at level s. The AF module initially derives a scene level reliability prior alongside a local spatial clarity map, subsequently utilizing their product to finely modulate the incoming modality evidence:

$$\begin{aligned} \phi_g^m &= \sigma(\text{MLP}(\text{GAP}(F_s^m))), \\ M_s^m &= \sigma(\text{Conv}(F_s^m)), \\ Q_s^m(x,y) &= \phi_g^m M_s^m(x,y), \\ \widetilde{F}_s^m(x,y) &= Q_s^m(x,y) F_s^m(x,y). \end{aligned} \tag{7}$$

Here, ϕ_g^m denotes a scalar scene level reliability score, whereas M_s^m and Q_s^m are single channel spatial maps that are broadcast along the feature channel dimension. This multiplicative formulation inherently ensures that the fusion process remains conservative whenever either global or local evidence is deemed weak, circumventing the need for explicit modality quality annotations.

While quality estimation effectively suppresses unreliable regions, it lacks the capacity to explicitly model complementary cross modal relationships. To address this, the AF module integrates a spatially reduced cross attention mechanism: queries are retained on the full resolution spatial grid, whereas keys and values are efficiently projected from reduced feature tokens. Let $X_m = \widetilde{F}_s^m$ and $X_{m,r} = \mathcal{S}_{r_s}(X_m)$, where the reduction operator $\mathcal{S}_{r_s}$ is implemented via strided average pooling or a lightweight strided convolution. After flattening the spatial dimensions, we obtain $\bar{X}_m \in \mathbb{R}^{C_s \times N_q}$ and $\bar{X}_{m,r} \in \mathbb{R}^{C_s \times N_k}$, with $N_q = H_s W_s$ and $N_k = N_q / r_s^2$. The RGB to T relational mapping is computed as:

$$\begin{aligned} A_{\text{rgb}\to\text{ir}} &= \text{Softmax}\Big(\frac{(W_q \bar{X}_{\text{rgb}})^\top (W_k \bar{X}_{\text{ir},r})}{\sqrt{d_k}}\Big), \\ Y_{\text{rgb}\leftarrow\text{ir}} &= (W_v \bar{X}_{\text{ir},r}) A_{\text{rgb}\to\text{ir}}^\top, \\ \widehat{X}_{\text{rgb}} &= X_{\text{rgb}} + W_o \,\text{reshape}(Y_{\text{rgb}\leftarrow\text{ir}}). \end{aligned} \tag{8}$$

The reverse T to RGB cross relation is computed symmetrically. Compared with standard full attention, whose dominant computational cost scales quadratically as $O(N_q^2)$, the reduced relation formulated in Eq. (8) dramatically lowers the complexity to $O(N_q N_k) = O(N_q^2 / r_s^2)$, thereby rendering nonlocal cross modal interactions strictly compatible with real time inference constraints. Let $\mathcal{M} = \{\text{rgb}, \text{ir}\}$ and $\boldsymbol{\phi}_g = [\phi_g^{\text{rgb}}, \phi_g^{\text{ir}}]$. The extracted quality maps and cross modal relation cues are subsequently mapped into final fusion weights via an adaptive temperature control mechanism:

$$\begin{aligned} \tau &= \text{Softplus}(\mathcal{H}(\boldsymbol{\phi}_g)) + \epsilon, \\ L^m &= \psi(\text{Concat}(\widetilde{F}_s^m, \widehat{X}_m)), \\ W^m(x,y) &= \frac{\exp(L^m(x,y)/\tau)}{\sum_{k\in\mathcal{M}} \exp(L^k(x,y)/\tau)}, \\ F_s^{\text{fus}} &= \sum_{m\in\mathcal{M}} W^m \odot \widehat{X}_m. \end{aligned} \tag{9}$$

Here, the temperature parameter τ serves as a per sample, per scale scalar that is uniformly shared across all spatial locations within a specific pyramid level; conversely, granular spatial adaptivity is purely driven by the preceding quality and evidence maps. A diminutive τ sharpens the softmax distribution, fostering aggressive modality competition when one stream demonstrates overwhelming reliability, whereas a larger τ promotes conservative, uniform spatial averaging under conditions of joint modality uncertainty. Fig. 4 visualizes the resulting spatially adaptive weights.

During the training phase, we introduce a weak cycle consistency regularization term applied exclusively to the reduced token sets to stabilize cross modal correspondence learning, introducing zero computational overhead during inference. Let $\widetilde{A}_{\text{rgb}\to\text{ir}}$ and $\widetilde{A}_{\text{ir}\to\text{rgb}}$ denote the reduced affinity matrices computed symmetrically from $\bar{X}_{\text{rgb},r}$ and $\bar{X}_{\text{ir},r}$. The training objective is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{cyc}} &= \frac{1}{N_k^2} \Big\| \widetilde{A}_{\text{rgb}\to\text{ir}} \widetilde{A}_{\text{ir}\to\text{rgb}} - I \Big\|_F^2, \\ \mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{det}} + \lambda \mathcal{L}_{\text{cyc}}. \end{aligned} \tag{10}$$

Here $\mathcal{L}_{\text{det}}$ is the standard detection loss of the baseline detector and λ controls the auxiliary regularization strength. Because strict one to one pixel correspondence is frequently violated in occluded, crowded, or severely misaligned local regions, $\mathcal{L}_{\text{cyc}}$ functions strictly as a weak auxiliary training regularizer and is entirely decoupled from the inference pipeline.

3.5. Cross Scale Consistency Perception via ARFP

Subsequent to the RGB-T fusion process, the preserved small object evidence must be robustly propagated across dis-

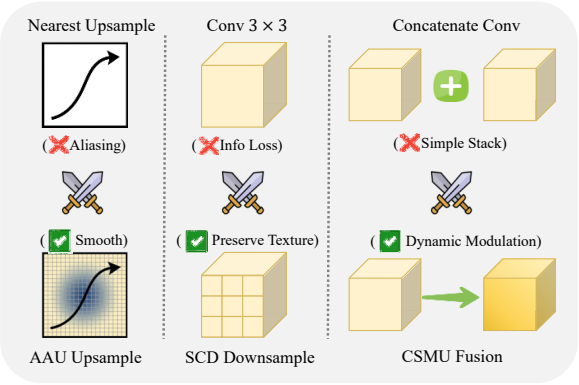

Figure 5: Structural comparison between standard building blocks and the proposed Aliasing Resistant Feature Pyramid (ARFP) components. The rows contrast, from top to bottom: (Top) neck specific Anti Aliasing Upsampling (AAU) against standard nearest neighbor interpolation; (Middle) Signal Fidelity Conservative Downsampling (SCD) against conventional strided convolution; and (Bottom) Cross Scale Modulation Unit (CSMU) against naive concatenation. Rather than generic plug in blocks, these components are redesigned for the RGB-T detection neck to preserve small object evidence and maintain bounded cross scale routing.

parate feature scales. However, the repeated sequence of upsampling and downsampling inherent to standard detection necks frequently introduces severe aliasing artifacts, catastrophic texture loss, and systematic localization distortion. These deleterious effects are particularly detrimental to RGB-T small object detection, primarily because small object representations are spatially fragile and acutely sensitive to geometric resampling errors. To explicitly resolve this structural bottleneck, the ARFP module is meticulously engineered to enforce cross scale consistency perception. It rigorously stabilizes hierarchical feature propagation by integrating Anti Aliasing Upsampling (AAU), Signal Fidelity Conservative Downsampling (SCD), and a bespoke Cross Scale Modulation Unit (CSMU). A standard detection neck is tasked with bidirectionally propagating both localization sensitive low level cues and context rich high level semantics across scales. In the specific context of RGB-T small object detection, this propagation is severely compromised by three coupled architectural failures: top down interpolation spawns high frequency spectral replicas and grid like aliasing; bottom up spatial reduction actively suppresses weak, high frequency textual details; and direct cross level fusion inadvertently amplifies the semantic mismatch between coarse semantic descriptors and fine grained localization cues. The proposed ARFP architecture elegantly dismantles these failure modes through the targeted application of neck specific AAU, SCD, and the content aware gated routing mechanism embedded within the CSMU. Together, these modules guarantee the preservation of cross scale consistency throughout pyramid propagation (Fig. 5 and Fig. 6).

The AAU module is deployed exclusively within the top down propagation path. It seamlessly couples standard nearest neighbor upsampling with deterministic low pass filtering, guaranteeing that spectral replicas are aggressively attenuated prior to lateral fusion, rather than delegating this corrective burden to subsequent convolutional layers. Let Ω_{hf} denote the high frequency region beyond the after upsampling Nyquist limit, and Ω the full frequency plane:

$$\begin{aligned} &Y_{\uparrow} = \text{Up}_{\text{nearest}}(X; s), \\ &\text{AAU}(X) = \delta * Y_{\uparrow}, \quad \sum_{i,j} \delta(i,j) = 1, \\ &\mathcal{E}_{\text{HF}} = \frac{\sum_{(\omega_x,\omega_y)\in\Omega_{\text{hf}}} |Y(e^{j\omega_x}, e^{j\omega_y})|^2}{\sum_{(\omega_x,\omega_y)\in\Omega} |Y(e^{j\omega_x}, e^{j\omega_y})|^2}. \end{aligned} \tag{11}$$

Here, the operator $*$ denotes depthwise spatial convolution, and δ represents a normalized, deterministic smoothing kernel. The high frequency energy ratio, denoted as $\mathcal{E}_{\text{HF}}$, serves as a quantitative diagnostic metric to evaluate the efficacy of spectral replica suppression. This specific mathematical formulation prioritizes the strict attenuation of interpolation artifacts over generic amplitude compensation, thereby immensely stabilizing the top down semantic propagation.

Fig. 7 visually corroborates the aliasing artifacts mathematically quantified by $\mathcal{E}_{\text{HF}}$. As demonstrated in Fig. 7(a), conventional nearest neighbor upsampling generates a highly diffuse and blurred spatial response around densely packed tiny objects following top down propagation and lateral fusion at P_3; conversely, AAU consistently yields a tightly compact activation pattern exhibiting substantially reduced peripheral signal leakage. Correspondingly, the frequency spectra plotted in Fig. 7(b) reveal prominent off center spectral replicas induced by nearest neighbor upsampling, whereas AAU effectively neutralizes these anomalous copies via its integrated deterministic low pass filtering. Because this deterministic filtering operation strictly precedes lateral fusion, the injected top down signal becomes vastly less susceptible to grid like interpolation artifacts, ultimately providing an exceptionally clean semantic map for subsequent multiscale feature aggregation. This specific visualization serves as empirical qualitative support validating the anti aliasing design rationale, rather than functioning as a standalone proof of the comprehensive ARFP system effect.

The SCD module is systematically deployed within the bottom up pathway, directly countering the phenomenon where repeated spatial downsampling progressively erases weak localization evidence. As detailed in Fig. 6(b), SCD intelligently couples a standard convolutional branch—dedicated to semantic feature aggregation—with a parallel depthwise convolutional branch optimized for conservative texture retention; these dual representations are subsequently integrated via a channel shuffle operation:

$$\begin{aligned} &Z = \text{Concat}(\text{Conv}_s(X), \text{DWConv}_s(X)), \\ &\text{SCD}(X) = \text{Shuffle}(Z), \\ &\text{vec}(\text{Shuffle}(Z)) = P_{\text{sh}}\,\text{vec}(Z), \\ &P_{\text{sh}}^{-1} = P_{\text{sh}}^{\top}. \end{aligned} \tag{12}$$

In practice, the parallel branches are configured to each produce exactly half of the total output channels, and an auxiliary pointwise projection is dynamically invoked solely when explicit channel dimension alignment is required. Because the

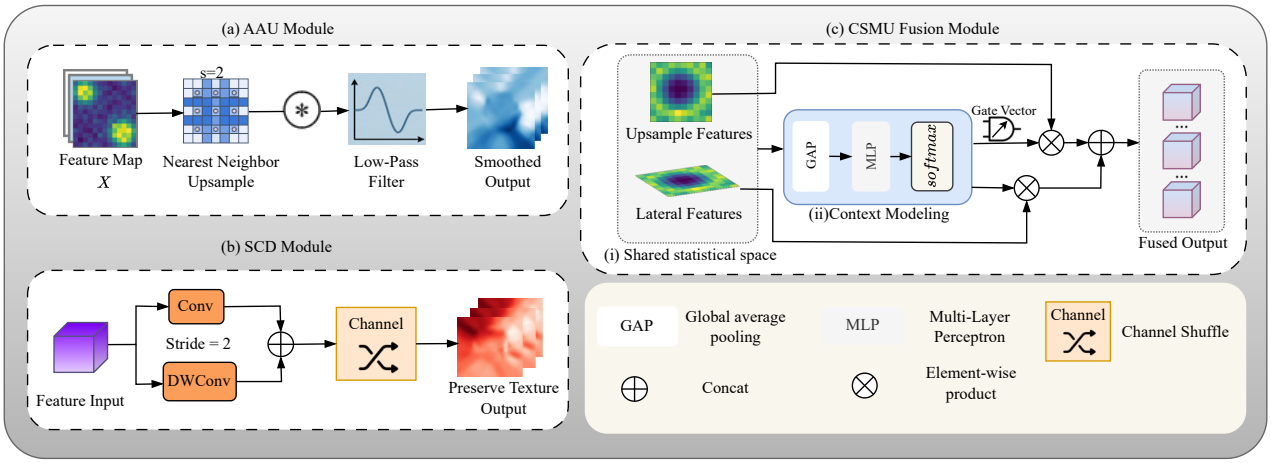

Figure 6: Schematic illustration of the core ARFP components for cross scale consistency perception. (a) **Anti Aliasing Upsampling (AAU)** combines nearest neighbor interpolation with deterministic low pass filtering to suppress spectral replicas before lateral fusion. (b) **Signal Fidelity Conservative Downsampling (SCD)** is designed with dual branch conservative downsampling and channel shuffle to preserve weak texture during spatial reduction. (c) **Cross Scale Modulation Unit (CSMU)** performs content aware gated routing between upsampled semantic flow and lateral localization flow.

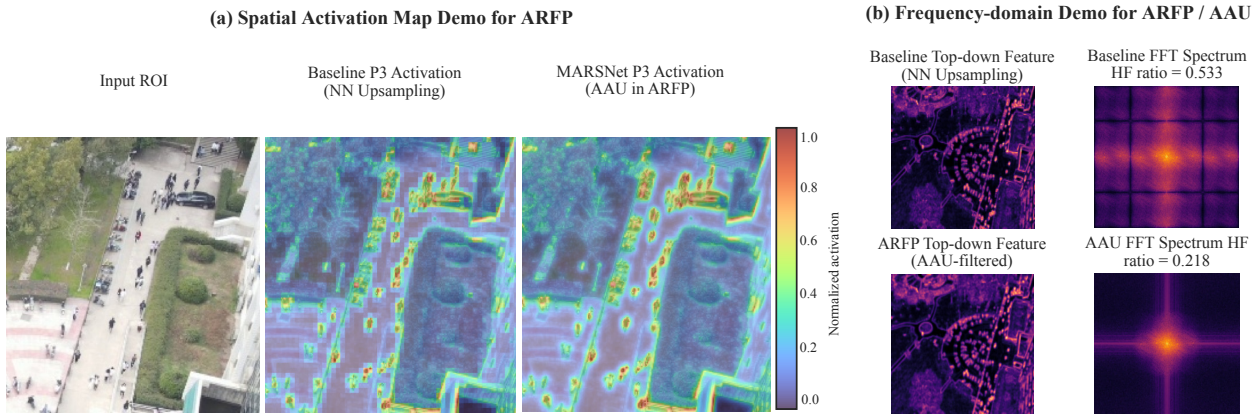

Figure 7: Illustrative feature level visualization of aliasing suppression by AAU. (a) Spatial domain response comparison after top down upsampling and lateral fusion at P_3. Relative to standard nearest neighbor upsampling, AAU yields a more compact response pattern around densely packed tiny objects with less peripheral leakage. (b) Log magnitude frequency spectra of the top down feature before lateral fusion. Nearest neighbor upsampling exhibits stronger off center spectral replicas, whereas AAU attenuates these replicas through deterministic low pass filtering.

channel shuffle operates as a strictly deterministic mathematical permutation, it seamlessly intertwines the two complementary feature branches without discarding any spatial samples. This guarantees that the bottom up pathway successfully retains high frequency textural cues while simultaneously integrating higher order semantic abstraction.

The CSMU is explicitly responsible for orchestrating content aware, dynamic routing between the upsampled semantic flow and the lateral localization flow. Let X_{up} and X_{lat} denote these respective inputs. Functionally, the CSMU first geometrically aligns these disparate inputs into a shared statistical representation space, subsequently predicting a set of strictly simplex constrained routing weights:

$$\begin{aligned} &U = \mathcal{T}(X_{\text{up}}), \quad L = \mathcal{T}(X_{\text{lat}}), \\ &z = [\text{GAP}(U); \text{GAP}(L); \text{GAP}(|U - L|)], \\ &[\alpha_{\text{up}}, \alpha_{\text{lat}}] = \text{Softmax}(\text{MLP}(z)), \quad \alpha_{\text{up}} + \alpha_{\text{lat}} = 1, \\ &\text{CSMU}(X_{\text{up}}, X_{\text{lat}}) = \alpha_{\text{up}} U + \alpha_{\text{lat}} L. \end{aligned} \tag{13}$$

Because the generated routing weights mathematically lie on the probability simplex, the CSMU executes adaptive, proportional soft routing rather than an arbitrary, unconstrained feature mixing operation. This design philosophy fundamentally diverges from the static branch weighting paradigm utilized in architectures like BiFPN [16]: CSMU dynamically predicts routing weights conditioned entirely upon the current input sample, and explicitly models the statistical discrepancy between the semantic and localization feature flows. Furthermore, the imposed simplex constraint rigorously bounds the final output to a

convex combination of the two aligned inputs, completely precluding the risk of unrestrained signal amplification from either independent branch.

The comprehensive ARFP hierarchical propagation process can thus be formally written as a sequence of top down semantic refinements followed immediately by bottom up spatial reinforcements:

$$\begin{aligned} \widetilde{P}_5 &= F_5^{\text{fus}}, \\ \widetilde{P}_l &= \text{CSMU}(\text{AAU}(\widetilde{P}_{l+1}), F_l^{\text{fus}}), \quad l = 4, 3, \\ P_3^{\text{out}} &= \widetilde{P}_3, \\ P_l^{\text{out}} &= \text{CSMU}(\text{SCD}(P_{l-1}^{\text{out}}), \widetilde{P}_l), \quad l = 4, 5. \end{aligned} \tag{14}$$

This formalized propagation rule explicitly delineates the specific role of each constituent subblock: AAU cleanses top down semantic flows prior to lateral integration; SCD rigorously preserves weak textural features during bottom up spatial reduction; and CSMU dynamically calibrates the balance between abstract semantic context and fine grained localization evidence at every single cross scale merging junction.

4. Experiments

Our empirical evaluation is systematically designed to address the following four core research questions: (1) whether MARSNet achieves an optimal accuracy and efficiency balance for RGB-T small object detection; (2) whether it generalizes robustly across diverse RGB-T scenarios; (3) whether the qualitative visual results substantiate the claim of small object evidence preservation; and (4) whether the three omni perception components are individually necessary and mutually complementary.

4.1. Datasets, Metrics, and Implementation Details

To rigorously evaluate MARSNet under the specific challenges emphasized in this work, we selected four RGB-T benchmarks exhibiting complementary characteristics. RGBTDronePerson serves as the primary testbed for evaluating dense UAV small object detection, whereas M^{3}FD assesses multiclass RGB-T robustness under diverse environmental conditions. Furthermore, LLVIP focuses on low light pedestrian detection and modality reliability, and KAIST further examines multispectral pedestrian detection under varying day/night and occlusion settings.

- RGBTDronePerson [11] contains 6,125 image pairs at a 640×512 resolution, featuring UAV scenes characterized by severe scale variation and high object density. It provides annotations for the Person, Rider, and Crowd categories. This benchmark is particularly challenging because UAV platform motion induces dynamic RGB-T misalignment, while crowded regions cause severe interobject occlusion. Notably, the original benchmark reports that approximately 98% of the annotated objects are smaller than 20 pixels.

- M^{3}FD [21] consists of 4,200 aligned RGB-T image pairs at mostly 1024×768 resolution and contains six object categories. Incorporating diverse scenes and complex environmental conditions, such as daytime, overcast, nighttime, fog, and smoke, it serves as a robust testbed for evaluating model resilience against environmental noise and signal degradation.

- LLVIP [20] comprises 15,488 released visible and infrared image pairs and serves as a standard benchmark for pedestrian detection in low light environments. The original visible and thermal sensors possessed resolutions of 1920×1080 and 1280×720, respectively, while the released pairs are registered and cropped to ensure matched fields of view. Dominated by relatively large objects, this dataset proves highly informative for studying cross modal reliability imbalances; specifically, it reveals whether the fusion module can effectively exploit thermal structures when visible evidence is severely compromised by extreme low light conditions.

- KAIST [22] is a widely adopted multispectral pedestrian detection benchmark for autonomous driving scenarios, containing color and thermal image pairs at 640×480 resolution across day and night traffic scenes. In addition to the standard processed test split and its All, Day, and Night partitions, we evaluated two diagnostic subsets from the KAIST annotations: Tiny/Far, corresponding to the far scale subset with pedestrian bounding box height below 45 pixels, and Occluded, defined as the union of partially and heavily occluded pedestrians.

Implementation details. All experiments evaluating MARSNet and its ablation variants were executed on a single NVIDIA RTX 3090 GPU. MARSNet was implemented in PyTorch, with the network architecture instantiated upon a YOLOv8 backbone. The network was optimized using stochastic gradient descent (SGD) with a momentum of 0.937 and a weight decay of 5×10^{-4}. We initialized the learning rate at 1×10^{-2} and regulated its decay via a cosine annealing schedule. Training spanned 300 epochs with a batch size of 16, utilizing automatic mixed precision (AMP) training. During training, geometric augmentations, including random horizontal flipping, scaling, and cropping, were applied synchronously to both the RGB and thermal images, as well as to their corresponding bounding box annotations. Photometric augmentation, such as color jittering, was applied exclusively to the visible branch. To achieve an optimal balance between detection accuracy and computational cost, images from LLVIP and M^{3}FD were resized to 640×640 pixels prior to network ingestion. Conversely, we preserved the original resolution for RGBTDronePerson to retain the fine grained cues of densely distributed tiny objects, while all KAIST evaluations strictly adhered to the standard multispectral protocol. Inference latency was measured using a batch size of 1 under FP16 precision after 50 warm up iterations, and averaged over 300 forward passes. Unless otherwise specified, NMS time was included and data loading time was excluded. Unavailable entries in the comparison tables are denoted by "–". Regarding the internal hyperparameters of the proposed modules, the token reduction stride r_s in the AF module was set to 2 to balance

global interaction and computational efficiency. For the adaptive temperature control, the numerical stability constant ϵ was set to 1×10^{-6}. During training, the auxiliary cycle consistency loss $\mathcal{L}_{\text{cyc}}$ was weighted by a coefficient of 0.1.

Evaluation metrics. For RGBTDronePerson, M^{3}FD, and LLVIP, we report the Mean Average Precision (mAP) and mAP_{50} metrics, adhering strictly to standard object detection evaluation protocols. For comparisons with state of the art (SOTA) methods on RGBTDronePerson, we followed the official tiny evaluation protocol [11] and reported mAP_{25}, $\text{mAP}_{50}^{\text{tiny}}$, scale partitioned mAP_{50}, and $\text{mAP}_{50}^{\text{all}}$. The scale partitioned columns in Table 1 correspond to the official evaluator defined tiny1, tiny2, tiny3, and small subsets, and are therefore reported as AP_{50}-style metrics rather than COCO style $\text{AP}_{50:95}$. For internal ablation studies on RGBTDronePerson, we additionally utilized a unified COCO style evaluator and reported $\text{mAP}_{50:95}$ and all object $\text{mAP}_{50}^{\text{all}}$. For the KAIST benchmark, we followed the standard multispectral pedestrian detection protocol and reported the log average miss rate (MR^{-2}) on the processed test split. Day and Night results were evaluated on the standard processed KAIST test split with 2,252 image pairs, including 1,455 daytime and 797 nighttime pairs. The Tiny/Far subset corresponds to pedestrians whose bounding box heights are below 45 pixels, and the Occluded subset contains pedestrians annotated as either partially occluded or heavily occluded. We evaluated only the person category as positive instances, while people, person, cyclist, and invalid ignored regions were handled following the KAIST/Caltech style evaluation protocol.

4.2. Comparison with State of the Art Methods

1) On RGBTDronePerson: RGBTDronePerson constitutes the most direct benchmark for the proposed RGB-T small object detection problem, given that the vast majority of its instances are tiny and densely distributed. Consequently, the empirical results on this dataset intrinsically reflect the capacity of MARSNet to preserve and refine fragile small object evidence. Table 1 presents the comparative results under the official tiny object protocol. MARSNet achieves an impressive 50.90% $\text{mAP}_{50}^{\text{tiny}}$ and 49.89% $\text{mAP}_{50}^{\text{all}}$, while efficiently operating at 35.48 FPS with a computational overhead of merely 98.20 GFLOPs. Although COXNet* reports a marginally higher $\text{mAP}_{50}^{\text{all}}$, it incurs a massive computational burden of 123.59 GFLOPs and runs below the real time threshold at 12.9 FPS; MARSNet therefore provides a substantially more favorable real time accuracy and efficiency balance. Notably, HRFuser [23] is included as a representative high resolution RGB-T fusion baseline. While HRFuser maintains high resolution representations through a multiresolution fusion design, which undeniably aids in preserving spatial details, this architecture inherently leads to a severe inference speed bottleneck, as reflected by its 4.5 FPS. In contrast, MARSNet preserves fine grained spatial structures via its proposed DualSOE and ARFP modules, without abandoning the highly efficient multiscale paradigm of YOLO family detectors, thereby demonstrating immense practical applicability for real time UAV deployments.

2) On M^{3}FD and LLVIP: To verify that the efficacy of the proposed fusion strategy enhanced by omni perception is not narrowly limited to a single UAV dataset, we further extend our evaluation to the M^{3}FD and LLVIP benchmarks. As shown in Tables 2 and 3, MARSNet achieves 92.1% mAP_{50} and the highest overall mAP of 65.1% on M^{3}FD, matching WaveMamba in AP_{50} while improving the stricter COCO style mAP by 0.7 points. Additionally, it achieves the best AP_{50} on scale sensitive categories, including People, Car, and Motorcycle. These results compellingly demonstrate MARSNet's capability to learn robust multiclass RGB-T representations across highly diverse environmental conditions. On the LLVIP benchmark, our method achieves the highest mAP of 70.4% and matches the best mAP_{50} of 98.3% among all compared methods, strongly corroborating the effectiveness of the modality reliability perception mechanism under extreme low light conditions.

3) On KAIST: To further validate MARSNet on multispectral pedestrian detection, we provide a representative comparison on KAIST using the standard KAIST/Caltech style log average miss rate. As shown in Table 4, MARSNet achieves an MR^{-2} of 11.45 on the All subset, outperforming CDC-YOLOFusion by 0.40. More importantly, the performance gains are markedly more pronounced on the most challenging subsets: MARSNet reduces the miss rate from 56.30 to 51.20 on Tiny/Far pedestrians and from 44.10 to 39.80 on Occluded pedestrians. These results confirm that MARSNet remains highly effective for tiny, distant, and occluded pedestrian cases, where small object evidence is inherently weak and cross scale localization consistency is of paramount importance.

4.3. Qualitative Analysis of Small Object Evidence

To qualitatively validate that the quantitative gains stem directly from superior small object evidence preservation, we conduct targeted visual evaluations spanning three distinct aspects: dense and occluded UAV scenes, comparisons with advanced RGB-T detectors, and category specific robustness under complex environments.

1) Visualization of Dense and Occluded Scenarios: To demonstrate the robustness of MARSNet in crowded UAV scenes, we first visualize dense region detection results on the RGBTDronePerson dataset in Fig. 8. Compared with the basic dual stream baseline, MARSNet successfully recuperates multiple closely proximate tiny objects that are otherwise prone to omission or spatial merging. This confirms that spatial detail perception and cross scale consistency perception synergistically improve instance recall in crowded aerial scenes.

Fig. 9 further compares MARSNet with DEGF-YOLO under low illumination nighttime conditions and high density daytime scenes. In the illustrated cases, although DEGF-YOLO can detect many salient pedestrians, it remains susceptible to missed detections and unstable bounding box localization when tiny objects are densely distributed, weakly contrasted, or partially occluded. In contrast, MARSNet recovers more weakly visible objects and maintains highly consistent bounding box localizations across both the visible and thermal inputs. This corroborates that proactively preserving small object evidence prior to

Table 1: Performance comparison on the RGBTDronePerson dataset. We follow the official tiny object evaluation protocol and report mAP_{25}, mAP_{50}^{tiny}, scale partitioned mAP_{50}, and mAP_{50}^{all}. Higher is better for all AP metrics and FPS, while lower is better for GFLOPs and Params. Bold indicates the best result in each column, underlined denotes the second best result, and "–" denotes unavailable entries. Here, "*" marks the higher computation variant of the corresponding method.

Methods	GFLOPs	Params(M)	FPS	mAP_{25}	mAP_{50}^{tiny}	mAP_{50}^{tiny1}	mAP_{50}^{tiny2}	mAP_{50}^{tiny3}	mAP_{50}^{small}	mAP_{50}^{all}
RetinaNet (2017) [24]	49.48	59.56	19.4	38.92	23.34	4.69	13.57	32.66	15.77	22.87
Cascade R-CNN (2018) [25]	76.23	92.35	11.8	42.47	31.99	0.00	29.43	37.77	33.61	31.55
FCOS (2019) [26]	**47.21**	78.32	19.2	45.21	30.71	9.40	22.73	34.87	26.00	29.89
ATSS (2020) [27]	48.73	55.31	19.0	53.14	37.47	16.92	24.16	43.59	24.62	36.24
GFL (2020) [28]	49.22	55.45	18.7	56.91	41.67	12.47	30.23	47.68	26.72	39.74
QueryDet (2022) [8]	121.67	62.58	12.8	55.16	37.75	17.89	24.90	44.05	25.52	37.07
FCOS w/ RFLA (2022) [29]	110.62	78.32	13.9	51.41	39.45	25.87	30.93	44.31	25.32	38.20
CFT (2021) [2]	112.02	206.26	10.8	37.32	22.83	16.72	18.14	27.52	8.14	22.69
HRFuser (2023) [23]	54.17	55.47	4.5	33.24	22.50	0.00	26.73	26.26	23.85	22.23
TINet (2023) [30]	92.80	100.86	13.0	40.34	28.60	0.00	24.12	34.99	**34.97**	28.30
QFDet (2023) [11]	81.43	60.19	14.2	57.34	44.04	20.27	30.09	50.36	26.78	42.08
QFDet* (2023) [11]	242.82	60.25	5.7	61.62	48.75	22.15	37.91	53.71	28.41	46.72
COXNet (2026) [7]	51.27	71.11	17.6	59.01	47.18	27.37	35.55	52.56	29.74	45.57
COXNet* (2026) [7]	123.59	75.86	12.9	**62.76**	**51.82**	23.08	40.10	**56.76**	30.89	**50.04**
DEGF-YOLO (2026) [31]	–	**19.60**	31.4	–	–	–	–	–	–	47.80
MARSNet (Ours)	98.20	56.80	**35.48**	62.10	50.90	**28.50**	**41.20**	55.90	30.60	49.89

Table 2: Quantitative comparison with SOTA methods on the M^3FD dataset. Best and second best results are highlighted in **bold** and underlined, respectively.

Methods	Backbone	Overall Metrics		Class wise AP_{50} (%)					
		mAP_{50}	mAP	People	Car	Bus	Motorcycle	Lamp	Truck
RFNet (2022) [5]	YOLOv5	79.4	53.2	79.4	91.1	78.2	72.8	85.0	69.9
TarDAL (2022) [21]	YOLOv5	80.5	54.1	81.5	94.8	81.3	69.3	87.1	68.7
DIDFuse (2020) [32]	YOLOv5	77.2	51.4	78.5	91.0	78.3	66.3	83.2	66.3
SDNet (2021) [33]	YOLOv5	79.3	52.9	79.4	91.1	78.2	72.8	85.0	69.0
DetFusion (2022) [34]	YOLOv5	80.8	53.8	80.8	92.5	83.0	69.4	87.8	71.4
CDDFuse (2023) [35]	YOLOv5	80.5	53.6	81.1	91.6	82.4	71.1	86.0	71.2
IGNet (2023) [36]	YOLOv5	81.5	54.5	81.6	92.8	82.4	73.0	86.9	72.1
SuperFusion (2022) [6]	YOLOv7	83.5	56.0	83.7	91.0	93.2	77.4	70.0	85.8
EMMA (2024) [37]	YOLOv5	82.9	55.4	82.0	93.5	83.2	77.7	87.7	73.5
KCDNet (2024) [38]	YOLOv5	83.2	56.3	83.3	90.9	88.4	84.1	80.3	72.1
CRSIOD (2024) [39]	YOLOv5	84.0	57.2	78.4	92.2	93.3	73.9	80.5	85.5
TSJNet (2024) [40]	YOLOv7	86.0	58.9	81.8	91.8	**95.7**	86.8	70.4	**89.3**
RI-YOLO (2024) [41]	YOLOv8	83.6	56.6	84.1	91.6	88.5	73.5	75.7	87.8
MRD-YOLO (2024) [42]	YOLOv8	86.6	59.3	–	–	–	–	–	–
MMFN (2025) [43]	YOLOv5	86.2	57.4	83.0	93.2	92.1	73.7	87.6	87.4
WaveMamba (2025) [10]	YOLOv8	**92.1**	64.4	92.6	95.9	95.6	86.1	**93.8**	88.6
MARSNet (Ours)	YOLOv8	**92.1**	**65.1**	**93.8**	**97.0**	92.9	**90.5**	91.8	86.5

fusion and maintaining cross scale consistency after fusion are critically advantageous for challenging aerial RGB-T scenes.

2) Qualitative Comparison with Advanced RGB-T Detectors: Fig. 10 provides additional qualitative comparisons against COXNet, COXNet*, and QFDet. While these methods exhibit strong detection capabilities across various visible or thermal scenes, their predictions frequently become fragmented in crowded regions where tiny objects are spatially adjacent or generate weak sensor responses. By comparison, MARSNet consistently yields more complete detections and clearer instance separations within the highlighted regions. This observation perfectly aligns with the quantitative results on RGBTDronePerson, especially the superior tiny object performance of MARSNet, and further substantiates the effectiveness of jointly preserving small object evidence, calibrating modality reliability, and stabilizing cross scale feature propagation.

3) Visualization of Category Specific Performance: To comprehensively profile the performance envelope across diverse object classes, we present a radar chart visualizing the class wise AP_{50} on the M^3FD dataset in Fig. 11. Contrasting MARSNet against six strong competing methods from Table 2, the visualization reveals that our framework establishes a pronounced advantage on categories inherently highly sensitive to scale variations and visual degradation (e.g., People, Car, and Motorcycle). Mean-

Table 3: Quantitative comparison on the LLVIP dataset. Multimodal methods are uniformly denoted as RGB-T for consistency. Best and second best results are highlighted in **bold** and underlined, respectively.

Methods	Input	mAP (%)	mAP_{50} (%)
RetinaNet (2017) [24]	RGB	42.8	88.0
Cascade R-CNN (2018) [25]	RGB	51.7	90.3
RetinaNet (2017) [24]	IR	55.1	94.8
Cascade R-CNN (2018) [25]	IR	60.4	94.9
RSDet (2026) [44]	RGB-T	61.3	95.8
FAWDet (2024) [4]	RGB-T	62.1	97.1
LRAF-Net (2024) [3]	RGB-T	66.3	97.9
UniRGB-IR (2025) [45]	RGB-T	63.2	96.1
WaveMamba (2025) [10]	RGB-T	66.0	**98.3**
CDC-YOLOFusion (2025) [9]	RGB-T	69.3	97.9
MARSNet (Ours)	RGB-T	**70.4**	**98.3**

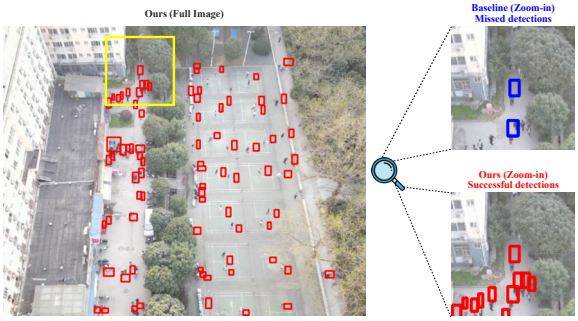

Figure 8: Qualitative comparison on the RGBTDronePerson dataset. The primary panel displays full scene detections generated by MARSNet. The zoomed insets highlight a dense region marked by the yellow box: the upper right inset presents baseline predictions (blue), and the lower right inset presents MARSNet predictions (red). MARSNet successfully recovers several closely spaced tiny objects overlooked by the baseline.

while, it maintains highly competitive performance on inherently large and static objects (e.g., Bus, Lamp, and Truck). These visual results validate that the performance gains of our framework are heavily concentrated on challenging categories where small object evidence and modality reliability are most difficult to preserve.

4) Discussion on Modality Limitations for Tiny Objects: As observed during our exploratory analyses on tiny objects, relying solely on the thermal infrared branch often yields suboptimal detection precision. This limitation is fundamentally rooted in the underlying sensing physics: thermal radiation footprints usually lack fine grained textures and sharp spatial gradients at subpixel scales; consequently, tiny objects tend to degenerate into amorphous hotspots. This observation directly motivates our Adaptive Fusion (AF) module. Instead of blindly trusting the thermal stream, the AF module designs temperature scaling to dynamically downweight thermal confidence when spatial evidence is severely degraded, permitting the DualSOE enhanced visible structures to dominate the final bounding box decision whenever they present demonstrably more reliable cues.

4.4. Ablation Study on Fusion Enhanced by Omni Perception

To rigorously quantify the individual and synergistic contributions of the three perception components within MARSNet, we conducted stepwise ablation studies on the RGBTDronePerson benchmark. The baseline architecture was a dual stream YOLOv8 detector equipped with element wise RGB-T fusion and a standard PANet neck. DualSOE, AF, and ARFP were then progressively integrated to independently evaluate the impacts of spatial detail perception, modality reliability perception, and cross scale consistency perception, respectively. The ablation results are presented in Table 5, and the corresponding accuracy and efficiency trajectory is plotted in Fig. 12.

1) Effect of Spatial Detail Perception: We first evaluated the contribution of DualSOE by integrating the proposed P_2 to P_3 structure preserving enhancement bridge into the baseline, strictly avoiding the introduction of an explicit P_2 detection head or the wholesale replacement of all stride 2 downsampling operators. As detailed in Table 5, the integration of DualSOE elevated the baseline mAP^{all}_{50} from 43.15% to 45.80% (+2.65%). This demonstrates that spatial detail perception effectively mitigates early spatial evidence erosion, thereby establishing a robust foundation for small object evidence preservation.

2) Effect of Modality Reliability Perception: Building upon the DualSOE configuration, we designed the Adaptive Fusion (AF) module to replace the naive element wise addition fusion. Table 5 shows that this integration further improved the mAP^{all}_{50} to 47.95% (+2.15%). This confirms the critical necessity of modality reliability perception over deterministic fusion rules, especially since modality specific noise risks being propagated by fixed confidence aggregation.

3) Effect of Cross Scale Consistency Perception: Finally, we integrated the complete ARFP detection neck into the architecture. The holistic combination of all three modules achieved peak performance, elevating the mAP^{all}_{50} to 49.89% (+1.94% over the DualSOE+AF variant). This stepwise gain definitively validates that cross scale consistency perception successfully maintains localization consistency and suppresses resampling induced distortions during hierarchical feature propagation.

4) Component Level Analysis: We further conducted component level ablations to rigorously validate the specific internal architectural design choices within DualSOE, AF, and ARFP. As evidenced in Table 6, substituting our proposed mechanisms with generic alternatives universally yielded suboptimal performance. For example, substituting the DualSOE downsampling path with SliceSamp reduced the dense crowd metric within the DualSOE block from 33.80% to 29.10%, indicating that simple information preserving sampling is insufficient without directional context aggregation and channel recalibration. Likewise, replacing AF with a generic attention block (e.g., CBAM) led to overconfident, deterministic fusion. The optimal results were consistently achieved by our complete configurations, confirming that the internal mechanisms within each module are intricately coordinated for the demanding RGB-T small object detection task.

Table 4: Representative quantitative comparison on the KAIST benchmark. Following the standard multispectral pedestrian detection protocol, we report the KAIST/Caltech style log average miss rate (MR^{-2}), where lower values indicate better performance. All, Day, and Night are evaluated on the processed KAIST test split. Tiny/Far denotes pedestrians with bounding box height below 45 pixels, and Occluded denotes the union of partially and heavily occluded pedestrians. Best and second best results are highlighted in **bold** and underlined, respectively.

Method	Backbone	All ($MR^{-2}\downarrow$)	Day ($MR^{-2}\downarrow$)	Night ($MR^{-2}\downarrow$)	Tiny/Far ($MR^{-2}\downarrow$)	Occluded ($MR^{-2}\downarrow$)
Halfway Fusion [1]	VGG-16	25.75	24.88	26.59	75.20	58.40
MBNet [46]	ResNet-50	14.50	14.30	14.80	65.10	48.30
TFDet [18]	ResNet-50	12.35	11.90	12.80	58.40	45.20
CDC-YOLOFusion [9]	YOLOv8	11.85	11.20	12.30	56.30	44.10
MARSNet (Ours)	**YOLOv8**	**11.45**	**11.05**	**11.95**	**51.20**	**39.80**

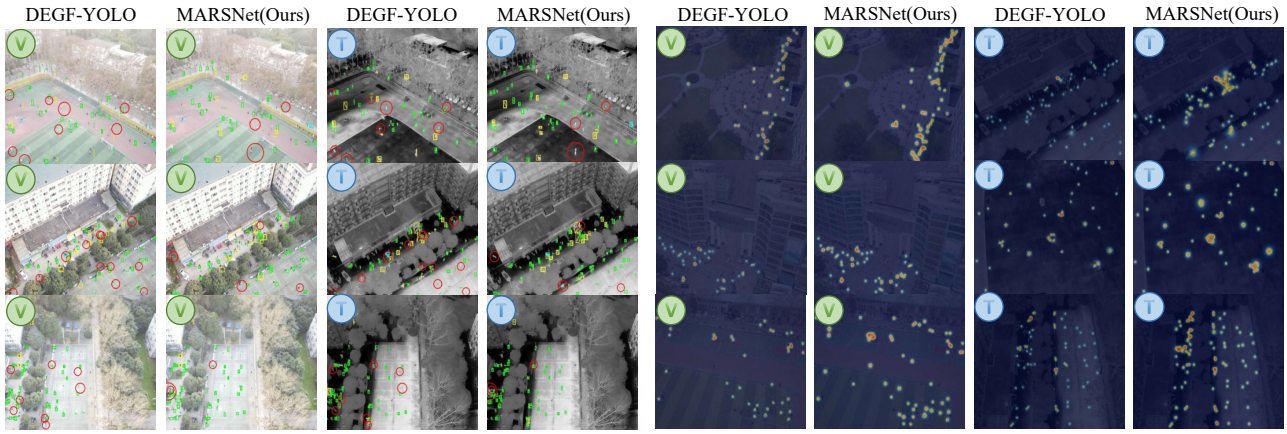

Figure 9: Extended qualitative comparison with current advanced methods on the RGBTDronePerson benchmark. Examples illustrate performance across low illumination nighttime conditions and high density daytime scenes. Compared with current advanced RGB-T detectors, MARSNet maintains more stable localization and recall for densely distributed tiny objects across complex aerial environments.

Table 5: Ablation study on the RGBTDronePerson dataset. Configurations are evaluated by toggling the **DualSOE**, **AF**, and **ARFP** modules, corresponding to spatial detail perception, modality reliability perception, and cross scale consistency perception, respectively. The first row establishes the baseline. The best results are highlighted in **bold**.

Modules			Accuracy (%)		FPS↑
DualSOE	AF	ARFP	$mAP_{50:95}$ ↑	mAP^{all}_{50} ↑	
			27.55	43.15	**40.50**
✓			28.90	45.80	38.20
	✓		28.05	44.35	39.10
		✓	27.85	43.90	39.70
✓	✓		30.15	47.95	36.80
✓		✓	29.95	47.35	36.90
	✓	✓	29.20	46.20	37.40
✓	✓	✓	31.55	**49.89**	35.48

4.5. Efficiency Analysis and Limitation Discussion

To quantitatively evaluate the accuracy and efficiency balance conceptualized in Fig. 1, we present a detailed computational cost breakdown in Table 7. The baseline model operated at 40.50 FPS. The stepwise integration of DualSOE, AF, and ARFP incurred strictly marginal computational overheads of +1.6G, +1.7G, and +2.4G FLOPs, respectively, resulting in a highly practical final inference speed of 35.48 FPS. Furthermore, when compared to directly appending an explicit high resolution P_2 prediction head, the complete MARSNet achieved higher detection accuracy (+4.69% mAP^{all}_{50}) while maintaining vastly superior operational efficiency (+4.28 FPS). This result unequivocally confirms that MARSNet improves small object evidence refinement through a fusion strategy enhanced by omni perception, rather than relying on computationally burdensome high resolution prediction heads.

Despite its highly favorable accuracy and efficiency balance, MARSNet remains a task specific architecture explicitly optimized for RGB-T small object detection, rather than a universal replacement for all generic RGB-T fusion pipelines. It may occasionally struggle in extreme edge cases where both the RGB and T modalities are simultaneously severely degraded by environmental interference, when thermal responses become completely saturated, or when cross modal geometric misalignment drastically exceeds the effective spatial receptive range of the fusion module. These specific failure modes indicate that explicit geometric alignment mechanisms and deployment oriented model compression remain critical directions for future improvement.

5. Conclusion

In this paper, we propose MARSNet, a novel Multimodal Adaptive Refined Small Object Detection framework empowered by a fusion strategy enhanced by omni perception for RGB-T small object detection. We identified and explicitly addressed

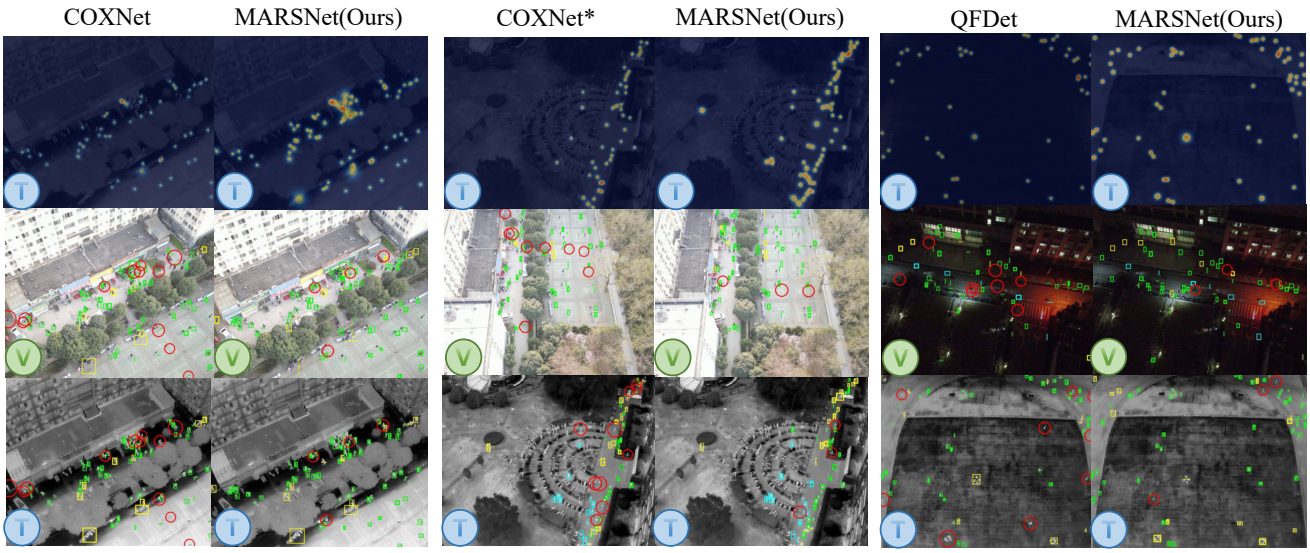

Figure 10: Qualitative comparison with current advanced methods on the RGBTDronePerson benchmark. The examples cover challenging aerial scenes with dense tiny objects, weak object contrast, and modality dependent degradation. The visual results indicate that MARSNet provides more complete tiny object recall and more stable localization in crowded regions.

Table 6: Component level ablations on the RGBTDronePerson benchmark. Each block tests alternative design choices for DualSOE, AF, and ARFP under the stepwise construction of MARSNet. We additionally report the dense crowd metric (mAP_{50}^{Crowd}) alongside the overall mAP_{50}^{all} to highlight how the three omni perception components improve heavily occluded tiny object detection.

Component	Variant / Configuration	mAP_{50}^{Crowd}	mAP_{50}^{all}	Note / Rationality
DualSOE	Baseline (No high resolution enhancement)	28.45	43.15	Severe early detail loss
	+ SliceSamp [12] replacement	29.10	43.82	Lacks structural integrity
	+ SPDConv only transfer	31.50	44.05	No contextual reintegration
	+ Full DualSOE	**33.80**	**45.80**	**Detail transfer + directional context and channel recalibration**
AF	Concat/Add Fusion	33.80	45.80	Ignores modality quality
	Generic Attention (e.g., CBAM)	34.25	46.52	Deterministic and overconfident
	Quality aware only	34.80	47.10	Lacks entropy control
	Temperature only (Global)	34.50	46.85	Lacks spatial granularity
	+ Full AF	**35.55**	**47.95**	**Uncertainty calibrated**
ARFP	Standard FPN/PAN Neck	35.55	47.95	Introduces aliasing artifacts
	+ AAU only	36.60	48.60	Suppresses top down aliasing
	+ LW Downsample Baseline	36.15	48.45	Discards bottom up texture
	+ Naive Add/Concat Routing	37.10	48.80	Unbounded semantic mismatch
	+ Full ARFP (AAU+SCD+CSMU)	**37.96**	**49.89**	**Improves localization consistency**

the Coupled Degradation Chain in RGB-T small object detection, demonstrating how fragile small object evidence is systematically degraded by early spatial evidence erosion, modality specific noise propagation, and cross scale localization distortion. To comprehensively overcome these cascading issues, MARSNet systematically refines small object evidence through three harmonized perception components. Specifically, DualSOE fortifies spatial detail perception by transferring structure preserving details from early network layers directly to stride aligned semantic features. The AF module elevates modality reliability perception by estimating hierarchical modality quality and adaptively regulating RGB-T aggregation. Finally, the ARFP module enhances cross scale consistency perception by drastically mitigating resampling induced aliasing and distortion within the detection neck.

Comprehensive empirical evaluations across the RGBTDronePerson, M^3FD, LLVIP, and KAIST benchmarks conclusively substantiate the superior effectiveness of MARSNet under highly diverse RGB-T detection scenarios. In particular, MARSNet achieves an outstanding 49.89% mAP_{50}^{all} and 50.90% mAP_{50}^{tiny} at 35.48 FPS under the rigorous RGBTDronePerson official protocol, establishing a highly favorable accuracy and efficiency balance for real time dense UAV small object detection. The results on M^3FD, LLVIP, and KAIST further confirm that MARSNet successfully generalizes to multiclass RGB-T detection, low light pedestrian detection, and multispectral pedestrian detection tasks. In depth ablation studies validate that DualSOE, AF, and ARFP provide vital, synergistic, and complementary improvements corresponding directly to spatial detail, modality reliability, and cross scale consistency perception, respectively.

Table 7: Module level cost breakdown on the RGBTDronePerson benchmark. The table quantifies the incremental computational overhead versus the accuracy gains as DualSOE, AF, and ARFP are integrated sequentially.

Configuration	FLOPs (G)	Params (M)	FPS	Δ FLOPs	Δ Params	Δ FPS	mAP^{all}_{50}
Baseline	92.5	54.2	40.50	-	-	-	43.15
Baseline + Explicit P_2 Head	115.4	60.5	31.20	+22.9	+6.3	-9.30	45.20
MARSNet w/o AF & ARFP (DualSOE)	94.1	54.8	38.20	+1.6	+0.6	-2.30	45.80
MARSNet w/o ARFP (+ AF)	95.8	55.5	36.80	+3.3	+1.3	-3.70	47.95
Full MARSNet	**98.20**	**56.8**	**35.48**	**+5.7**	**+2.6**	**-5.02**	**49.89**

Note: Δ columns are cumulative against the Baseline row. The incremental (row to row) overhead is +1.6G FLOPs / +0.6M parameters for DualSOE, +1.7G / +0.7M for AF, and +2.4G / +1.3M for ARFP. The AF module's cycle consistency loss is dropped during inference. FPS values match those reported in Table 5.

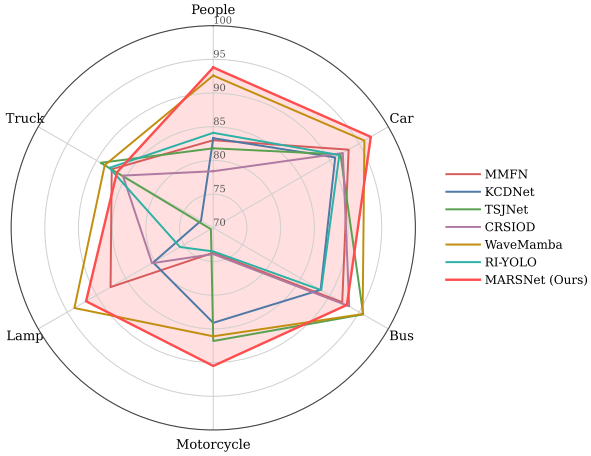

Figure 11: Class wise performance comparison on the M^{3}FD dataset. The radar chart contrasts MARSNet with six strong competing methods from Table 2 using class wise AP_{50} metrics.

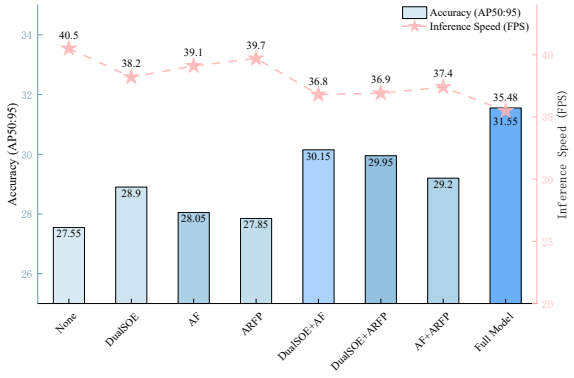

Figure 12: Visualization of ablation study outcomes on RGBT-DronePerson. The bars show the overall $AP_{50:95}$ of different module combinations, while the dashed line reports the corresponding inference speed. The full model achieves the best accuracy while maintaining real time inference.

Despite these proven advantages, MARSNet is not without practical limitations; it may struggle under extreme scenarios of simultaneous dual modality degradation, thermal response saturation, or severe spatial misalignment that exceeds the effective receptive range of the fusion module. Future work will focus on exploring explicit geometric alignment mechanisms, more deployment friendly model compression techniques, and broader multisensor extensions to further facilitate robust, open world RGB-T small object detection.

References

[1] J. Liu, S. Zhang, S. Wang, D. Metaxas, Multispectral deep neural networks for pedestrian detection, in: Proceedings of the British Machine Vision Conference (BMVC), BMVA Press, 2016, pp. 73.1–73.13. doi:10.5244/C.30.73.
URL https://doi.org/10.5244/C.30.73

[2] F. Qingyun, H. Dapeng, W. Zhaokui, Cross-modality fusion transformer for multispectral object detection, arXiv preprint arXiv:2111.00273 (2021).

[3] H. Fu, S. Wang, P. Duan, C. Xiao, R. Dian, S. Li, Z. Li, Lraf-net: Long-range attention fusion network for visible-infrared object detection, IEEE Trans. Neural Networks Learn. Syst. 35 (10) (2024) 13232–13245. doi:10.1109/TNNLS.2023.3266452.
URL https://doi.org/10.1109/TNNLS.2023.3266452

[4] P. Zhao, X. Ye, Z. Du, Object detection in multispectral remote sensing images based on cross-modal cross-attention, Sensors 24 (13) (2024) 4098. doi:10.3390/S24134098.
URL https://doi.org/10.3390/s24134098

[5] H. Xu, J. Ma, J. Yuan, Z. Le, W. Liu, RFNet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19647–19656. doi:10.1109/CVPR52688.2022.01906.

[6] L. Tang, Y. Deng, Y. Ma, J. Huang, J. Ma, Superfusion: A versatile image registration and fusion network with semantic awareness, IEEE/CAA J. Autom. Sinica 9 (12) (2022) 2121–2137. doi:10.1109/JAS.2022.106082.

[7] P. Peng, T. Xu, L. Song, M. Zhu, Y. Fang, J. Li, COXNet: Cross-layer fusion with adaptive alignment and scale integration for RGBT tiny object detection, IEEE Trans. Circuits Syst. Video Technol. 36 (1) (2026) 596–608. doi:10.1109/TCSVT.2025.3595147.
URL https://doi.org/10.1109/TCSVT.2025.3595147

[8] C. Yang, Z. Huang, N. Wang, QueryDet: Cascaded sparse query for accelerating high-resolution small object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 13658–13667. doi:10.1109/CVPR52688.2022.01330.
URL https://doi.org/10.1109/CVPR52688.2022.01330

[9] Z. Wang, X. Liao, J. Yuan, Y. Yao, Z. Li, CDC-YOLOFusion: Leveraging cross-scale dynamic convolution fusion for visible-infrared object detection, IEEE Trans. Intell. Veh. 10 (3) (2025) 2080–2093. doi:10.1109/TIV.2024.3443264.
URL https://doi.org/10.1109/TIV.2024.3443264

[10] H. Zhu, W. Dong, L. Yang, H. Li, Y. Yang, Y. Ren, Q. Zhu, Z. Feng, C. Li, S. Lin, et al., WaveMamba: Wavelet-driven mamba fusion for RGB-infrared object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2025, pp. 11219–11229.
URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhu_WaveMamba_Wavelet-Driven_Mamba_Fusion_for_RGB-Infrared_Object_Detection_ICCV_2025_paper.html

[11] Y. Zhang, C. Xu, W. Yang, G. He, H. Yu, L. Yu, G.-S. Xia, Drone-based RGBT tiny person detection, ISPRS J. Photogramm. Remote Sens. 204

(2023) 61–76. doi:10.1016/j.isprsjprs.2023.08.016.
URL https://doi.org/10.1016/j.isprsjprs.2023.08.016

[12] L. He, M. Wang, SliceSamp: A promising downsampling alternative for retaining information in a neural network, Appl. Sci. 13 (21) (2023) 11657. doi:10.3390/app132111657.
URL https://doi.org/10.3390/app132111657

[13] W. Lu, S.-B. Chen, J. Tang, C. H. Q. Ding, B. Luo, A robust feature downsampling module for remote-sensing visual tasks, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–12. doi:10.1109/TGRS.2023.3282048.
URL https://doi.org/10.1109/TGRS.2023.3282048

[14] X. Zhang, D. Yuan, Y. Hu, Z. Wu, X. Zhang, B. Yu, X. Bai, S.-Y. Cao, Y. Jin, B. Yang, H.-L. Shen, SADet: A semantic-aware tiny object detection network against missed detection, Pattern Recognit. 172 (2026) 112624. doi:10.1016/j.patcog.2025.112624.
URL https://doi.org/10.1016/j.patcog.2025.112624

[15] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, S. J. Belongie, Feature pyramid networks for object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 936–944. doi:10.1109/CVPR.2017.106.
URL https://doi.org/10.1109/CVPR.2017.106

[16] M. Tan, R. Pang, Q. V. Le, Efficientdet: Scalable and efficient object detection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 10778–10787. doi:10.1109/CVPR42600.2020.01079.
URL https://openaccess.thecvf.com/content_CVPR_2020/html/Tan_EfficientDet_Scalable_and_Efficient_Object_Detection_CVPR_2020_paper.html

[17] R. Zhang, Making convolutional networks shift-invariant again, in: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, PMLR, 2019, pp. 7324–7334.
URL http://proceedings.mlr.press/v97/zhang19a.html

[18] X. Zhang, X. Zhang, J. Wang, J. Ying, Z. Sheng, H. Yu, C. Li, H.-L. Shen, TFDet: target-aware fusion for RGB-T pedestrian detection, IEEE Trans. Neural Networks Learn. Syst. 36 (7) (2025) 13276–13290. doi:10.1109/TNNLS.2024.3443455.
URL https://doi.org/10.1109/TNNLS.2024.3443455

[19] R. Sunkara, T. Luo, No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects, in: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2022, Grenoble, France, September 19-23, 2022, Proceedings, Part III, Springer, 2022, pp. 443–459. doi:10.1007/978-3-031-26409-2_27.
URL https://doi.org/10.1007/978-3-031-26409-2_27

[20] X. Jia, C. Zhu, M. Li, W. Tang, W. Zhou, LLVIP: A visible-infrared paired dataset for low-light vision, in: IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, QC, Canada, October 11-17, 2021, IEEE, 2021, pp. 3489–3497. doi:10.1109/ICCVW54120.2021.00389.
URL https://doi.org/10.1109/ICCVW54120.2021.00389

[21] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, Z. Luo, Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 5792–5801. doi:10.1109/CVPR52688.2022.00571.
URL https://doi.org/10.1109/CVPR52688.2022.00571

[22] S. Hwang, J. Park, N. Kim, Y. Choi, I. S. Kweon, Multispectral pedestrian detection: Benchmark dataset and baseline, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1037–1045. doi:10.1109/CVPR.2015.7298706.
URL https://openaccess.thecvf.com/content_cvpr_2015/html/Hwang_Multispectral_Pedestrian_Detection_2015_CVPR_paper.html

[23] T. Brödermann, C. Sakaridis, D. Dai, L. V. Gool, HRFuser: A multi-resolution sensor fusion architecture for 2d object detection, in: 26th IEEE International Conference on Intelligent Transportation Systems, ITSC 2023, Bilbao, Spain, September 24-28, 2023, IEEE, 2023, pp. 4159–4166. doi:10.1109/ITSC57777.2023.10422432.
URL https://doi.org/10.1109/ITSC57777.2023.10422432

[24] T. Lin, P. Goyal, R. B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 2999–3007. doi:10.1109/ICCV.2017.324.
URL https://doi.org/10.1109/ICCV.2017.324

[25] Z. Cai, N. Vasconcelos, Cascade R-CNN: delving into high quality object detection, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 6154–6162. doi:10.1109/CVPR.2018.00644.
URL http://openaccess.thecvf.com/content_cvpr_2018/html/Cai_Cascade_R-CNN_Delving_CVPR_2018_paper.html

[26] Z. Tian, C. Shen, H. Chen, T. He, FCOS: fully convolutional one-stage object detection, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019, pp. 9626–9635. doi:10.1109/ICCV.2019.00972.
URL https://doi.org/10.1109/ICCV.2019.00972

[27] S. Zhang, C. Chi, Y. Yao, Z. Lei, S. Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 9756–9765. doi:10.1109/CVPR42600.2020.00978.
URL https://openaccess.thecvf.com/content_CVPR_2020/html/Zhang_Bridging_the_Gap_Between_Anchor-Based_and_Anchor-Free_Detection_via_Adaptive_CVPR_2020_paper.html

[28] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, J. Yang, Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.
URL https://proceedings.neurips.cc/paper/2020/hash/f0bda020d2470f2e74990a07a607ebd9-Abstract.html

[29] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, G. Xia, RFLA: gaussian receptive field based label assignment for tiny object detection, in: Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX, Springer, 2022, pp. 526–543. doi:10.1007/978-3-031-20077-9_31.
URL https://doi.org/10.1007/978-3-031-20077-9_31

[30] Y. Zhang, H. Yu, Y. He, X. Wang, W. Yang, Illumination-guided RGBT object detection with inter- and intra-modality fusion, IEEE Trans. Instrum. Meas. 72 (2023) 1–13. doi:10.1109/TIM.2023.3251414.
URL https://doi.org/10.1109/TIM.2023.3251414

[31] Y. Gu, W. Chen, D. Peng, UAV-based multimodal object detection via feature enhancement and dynamic gated fusion, Pattern Recognit. 172 (2026) 112722. doi:10.1016/j.patcog.2025.112722.
URL https://doi.org/10.1016/j.patcog.2025.112722

[32] Z. Zhao, S. Xu, C. Zhang, J. Liu, J. Zhang, P. Li, DIDFuse: Deep image decomposition for infrared and visible image fusion, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 2020, pp. 970–976. doi:10.24963/ijcai.2020/135.

[33] H. Zhang, J. Ma, SDNet: A versatile squeeze-and-decomposition network for real-time image fusion, Int. J. Comput. Vis. 129 (10) (2021) 2761–2785. doi:10.1007/s11263-021-01501-8.

[34] Y. Sun, B. Cao, P. Zhu, Q. Hu, DetFusion: A detection-driven infrared and visible image fusion network, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 4003–4011. doi:10.1145/3503161.3547902.

[35] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, L. V. Gool, Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5906–5916. doi:10.1109/CVPR52729.2023.00572.
URL https://doi.org/10.1109/CVPR52729.2023.00572

[36] J. Li, J. Chen, J. Liu, H. Ma, Learning a graph neural network with cross modality interaction for image fusion, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 4471–4479. doi:10.1145/3581783.3612135.
URL https://doi.org/10.1145/3581783.3612135

[37] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, K. Zhang, S. Xu, D. Chen, R. Timofte, L. Van Gool, Equivariant multi-modality image fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 25912–25921. doi:10.1109/CVPR52733.2024.02448.
URL https://doi.org/10.1109/CVPR52733.2024.02448

[38] H. Wang, S. Qu, Z. Qiao, X. Liu, KCDNet: Multimodal object detection in modal information imbalance scenes, IEEE Trans. Instrum. Meas. 73 (2024) 1–13. doi:10.1109/TIM.2024.3441019.
URL https://doi.org/10.1109/TIM.2024.3441019

[39] H. Wang, C. Wang, Q. Fu, D. Zhang, R. Kou, Y. Yu, J. Song, Cross-modal oriented object detection of UAV aerial images based on image feature, IEEE Trans. Geosci. Remote. Sens. 62 (2024) 1–21. doi:10.1109/TGRS.2024.3367934.
URL https://doi.org/10.1109/TGRS.2024.3367934

[40] Y. Jie, Y. Xu, X. Li, H. Li, H. Tan, F. Nie, TSJNet: A multi-modality target and semantic awareness joint-driven image fusion network, CoRR abs/2402.01212 (2024). arXiv:2402.01212, doi:10.48550/ARXIV.2402.01212.
URL https://doi.org/10.48550/arXiv.2402.01212

[41] Y. Chen, B. Wang, W. Zhu, J. Yuan, RGB-IR YOLO combining modality-specific reconstruction and information integration, in: 2024 39th Youth Academic Annual Conference of Chinese Association of Automation (YAC), IEEE, 2024, pp. 2045–2050. doi:10.1109/YAC63405.2024.10598725.
URL https://doi.org/10.1109/YAC63405.2024.10598725

[42] C. Sun, Y. Chen, X. Qiu, R. Li, L. You, MRD-YOLO: A multispectral object detection algorithm for complex road scenes, Sensors 24 (10) (2024) 3222. doi:10.3390/S24103222.
URL https://doi.org/10.3390/s24103222

[43] F. Yang, B. Liang, W. Li, J. Zhang, Multidimensional fusion network for multispectral object detection, IEEE Trans. Circuits Syst. Video Technol. 35 (1) (2025) 547–560. doi:10.1109/TCSVT.2024.3454631.
URL https://doi.org/10.1109/TCSVT.2024.3454631

[44] T. Zhao, M. Yuan, F. Jiang, N. Wang, X. Wei, Removal then selection: A coarse-to-fine fusion perspective for RGB-infrared object detection, IEEE Trans. Intell. Transp. Syst. 27 (2) (2026) 2504–2519. doi:10.1109/TITS.2025.3638627.
URL https://doi.org/10.1109/TITS.2025.3638627

[45] M. Yuan, B. Cui, T. Zhao, J. Wang, S. Fu, X. Yang, X. Wei, UniRGB-IR: A unified framework for visible-infrared semantic tasks via adapter tuning, in: Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 2409–2418. doi:10.1145/3746027.3754806.
URL https://doi.org/10.1145/3746027.3754806

[46] K. Zhou, L. Chen, X. Cao, Improving multispectral pedestrian detection by addressing modality imbalance problems, in: Computer Vision – ECCV 2020, Vol. 12363 of Lecture Notes in Computer Science, Springer, 2020, pp. 787–803. doi:10.1007/978-3-030-58523-5_46.
URL https://doi.org/10.1007/978-3-030-58523-5_46