



ARTICLE

Coarse-to-Fine Alignment and Reliability-Driven Fusion for Multispectral Object Detection

Ye Li, Haotian Ning, Tong Xiao, Xiaoke Su and Weiwei Zhang*

School of Computer Science and Artificial Intelligence, Zhengzhou University of Light Industry, Zhengzhou, China

*Corresponding Author: Weiwei Zhang. Email: anqikeli@126.com

Received: 24 July 2026; Accepted: 25 August 2026; Published: 22 October 2026

ABSTRACT: To address the challenges of cross-modal spatial misalignment, modality reliability discrepancies, and inadequate supervision in visible-infrared (RGB-IR) multispectral object detection, this paper proposes Coarse-to-Fine Alignment and Reliability-Driven Fusion Network (CFR-Net), a robust framework designed for complex real-world environments. First, we introduce the Scale-aware Offset Alignment and Residual (SOAR) module. By employing a coarse-to-fine offset estimation mechanism, SOAR effectively rectifies global displacements and compensates for local residual misalignments, while leveraging high-frequency residual information to enhance fine-grained representations in small-object regions. Based on the aligned features, the Reliability-aware Quality-guided Fusion (RQF) module is further designed, which adaptively adjusts the fusion weights of RGB and IR features according to region-level modality quality and cross-modal conflict intensity, thereby enhancing reliable complementary information and suppressing low-quality or conflicting responses. Furthermore, an Alignment-aware Fusion Consistency (AFC) Loss is formulated to provide explicit process-level supervision by jointly optimizing feature alignment, fusion consistency, and hard-sample detection, thereby ensuring a more principled learning process that extends beyond standard detection objectives. Experimental results on the DroneVehicle and M3FD datasets demonstrate that a mean Average Precision at an intersection over union (IoU) threshold of 0.50 (mAP50) of 82.1% and 85.7% is achieved by CFR-Net, with a compact architecture of only 6.5M parameters and 16.9 giga floating-point operations (GFLOPs). Consequently, CFR-Net provides a robust and computationally efficient solution for multispectral object detection, achieving competitive performance with minimal architectural overhead.

KEYWORDS: Multispectral object detection; RGB-IR feature fusion; cross-modal alignment; reliability-aware fusion

1 Introduction

With the rapid development of intelligent transportation, unmanned aerial vehicle (UAV) inspection, aerial surveillance, and security monitoring, object detection in open environments places higher demands on visual perception systems. Most existing detection systems rely on visible (RGB) images. Under sufficient illumination, RGB images provide detailed information such as color, texture, and edges. However, in nighttime, low-light, rainy, foggy, and occluded scenes, they are easily affected by insufficient illumination, motion blur, and noise, which weakens or even removes target features [1]. Infrared (IR) images are formed from the thermal radiation of objects and are insensitive to changes in ambient illumination, so they can provide relatively stable target responses at night and under adverse weather conditions [2]. However, IR images lack color and fine-grained texture information, and their ability to represent structural details is weaker than that of visible images. Therefore, RGB and IR are complementary in information representation, and effective fusion of the two modalities is a core problem in multispectral object detection [2].

Existing visible-infrared (RGB-IR) multispectral object detection methods mainly include image-level fusion, feature-level fusion, and decision-level fusion. Image-level fusion combines information from the two modalities at the pixel level, but effectively preserving both shared and complementary modality information in the fused representation remains challenging [3]. Decision-level fusion usually performs single-modality detection first and then merges the detection results. This strategy preserves the detection ability of each modality, but it introduces a large computational cost. Feature-level fusion introduces cross-modal interaction during feature extraction and achieves a favorable balance between detection accuracy and computational efficiency, making it a widely used technical route [1,4]. In recent years, methods based on cross-modal attention, feature enhancement, and modality complementarity modeling have further improved fusion-based detection performance [5]. However, cross-modal misalignment and modality imbalance still limit the effective integration of complementary information across modalities [6]. Therefore, RGB-IR detection continues to face several challenges, including cross-modal spatial misalignment, modality reliability differences and local response conflicts, and insufficient fusion supervision [7].

Cross-modal spatial misalignment directly affects the effectiveness of RGB-IR feature fusion. Since RGB and IR sensors differ in imaging mechanisms and optical parameters, spatial offsets may still occur even when the two sensors are rigidly mounted, due to platform motion, asynchronous acquisition, and parallax effects [8]. During feature fusion, if features from the two modalities do not correspond accurately, the model aggregates information at mismatched locations, which introduces structural noise and degrades the quality of target representation [9]. This problem becomes more severe in small-object detection, where slight pixel-level offsets may shift target responses or cause feature aliasing [10,11]. Existing implicit alignment methods usually use deformable attention or sparse sampling to reduce positional bias, but their ability to handle large-range misalignment remains limited [7,12]. Explicit correction methods can improve spatial consistency by predicting spatial correction cues, but maintaining both global alignment stability and fine-grained local calibration accuracy remains challenging, especially when residual misalignment affects small-object regions [9,11,13]. Therefore, a more stable and fine-grained cross-modal spatial alignment mechanism is needed, with special consideration of the alignment requirements in small-object regions.

Modality reliability differences and local response conflicts also weaken the discriminative ability of fused features. RGB images usually provide stronger discriminative information in daytime and texture-rich scenes, while IR images are more stable at night, under low illumination, and in adverse weather conditions [2]. If a fusion module always assigns fixed weights to the two modalities, noise, blur, or false responses from the low-quality modality may enter the fused features, causing the fusion result to be worse than the single-modality result. This phenomenon is known as fusion degradation [14]. Moreover, the relative reliability of RGB and IR features varies across illumination conditions and spatial regions. If the fusion module cannot adaptively balance inter- and intra-modality information, low-quality or redundant responses may propagate into the fused representation and adversely affect the final detection results [15]. Existing methods based on quality-aware weighting, shared-specific feature decomposition, and cross-modal attention can alleviate this problem to some extent, but their modeling of local reliability variations and conflict responses remains insufficient [14].

Insufficient fusion supervision further limits the training effectiveness of multispectral detection models. Standard detection losses mainly act on the final prediction results, making it difficult to directly constrain the cross-modal alignment process or determine whether the fused features are better than single-modality features. When the input modalities suffer from spatial misalignment, quality degradation, or local response conflicts, relying only on classification loss and bounding-box regression loss is often insufficient to identify the source of errors, and the supervision for difficult small-object samples is also limited [7,16]. Existing task-driven fusion losses, cross-modal consistency losses, and modality-balanced optimization

methods improve the training process from different perspectives. However, most of them focus on a single problem and have not yet formed a unified supervision mechanism that simultaneously addresses cross-modal spatial alignment, reliable modality selection, and small-object optimization [17].

To address cross-modal spatial misalignment, modality reliability discrepancy with local response conflicts, and insufficient fusion supervision in RGB-IR multispectral object detection, this paper proposes Coarse-to-Fine Alignment and Reliability-Driven Fusion Network (CFR-Net). Specifically, the proposed framework employs the Scale-aware Offset Alignment and Residual (SOAR) module to achieve scale-aware cross-modal offset alignment, utilizes the Reliability-aware Quality-guided Fusion (RQF) module to perform reliability-aware quality-guided fusion, and introduces the Alignment-aware Fusion Consistency (AFC) Loss to provide alignment consistency and fusion quality constraints during training, thereby improving the robustness of RGB-IR multispectral object detection in complex scenarios. The main contributions of this paper are summarized as follows:

1. The SOAR module addresses cross-modal spatial misalignment through a coarse-to-fine offset estimation mechanism. This mechanism corrects global offsets and compensates for local detail deviations, thereby improving feature alignment accuracy in small-object regions.
2. The RQF module performs reliability-aware fusion to handle modality reliability differences and local response conflicts. It dynamically adjusts the fusion process according to modality quality and conflict intensity, which enhances reliable complementary information and suppresses low-quality or contradictory features.
3. AFC Loss provides process-level supervision when fusion constraints are insufficient. It imposes training constraints on cross-modal alignment, fusion effectiveness, and hard small-object optimization, enabling more direct supervision of the fusion process beyond the standard detection loss.
4. Experiments on two public RGB-IR detection benchmarks demonstrate that our method achieves competitive detection performance with a favorable accuracy–efficiency trade-off.

2 Related Work

Cross-modal misalignment arises from different sensor optics, platform motion, asynchronous capture, and parallax [8,9]. It makes RGB and IR responses refer to different locations and disproportionately harms small objects with limited spatial support [11]. Existing work therefore addresses registration, in-network correspondence, and residual detail preservation.

Input registration corrects images before detection but adds computation and is difficult to optimize end to end [9]. Implicit methods use deformable attention, sparse sampling, or cross-modal queries, as in CF-Deformable DETR [7], DPDETR [12], and MS-DETR [17]; their correction remains bounded by the sampling range. Attention-based fusion methods can alleviate the sensitivity of cross-modal interaction to spatial misalignment to some extent. ICAFusion [1] employs parameter-shared iterative bidirectional cross-attention to progressively exchange complementary information between RGB and IR features. C²Former [4] combines global cross-modal correspondence modeling with offset-guided feature calibration to align features before attention-based fusion. Explicit methods predict spatial corrections: CoDAF [9] estimates feature offsets, CAGTDet [13] performs cascade region-of-interest (RoI) calibration, while C2RF [18] and CLAFusion [19] strengthen cross-modal correspondence through joint registration-fusion learning or contrastive learning.

Small objects amplify residual alignment error because a few pixels can shift responses into the background. COXNet [11] couples cross-modal alignment with multi-scale feature aggregation. ESM-YOLO+ [20] employs mask-enhanced pixel-level fusion to mitigate cross-modal misalignment and scale heterogeneity, while introducing structural supervision during training to preserve fine-grained

spatial information. DHiF [21] explicitly exploits high-frequency information to retain structural details. IC-Fusion [22] adopts an IR-dominant fusion strategy based on the boundary and contour information contained in high-frequency IR components. RGB-thermal (RGBT)-Tiny [23] highlights the importance of preserving fine-grained target details, while the DVTOD benchmark [24] demonstrates the need for explicit cross-modal alignment under real-world visible-thermal misalignment.

Reliability-aware fusion addresses scene-dependent modality quality: RGB favors well-lit texture, whereas IR remains stable at night [1,2]. Fixed weights can cause fusion degradation [14]. TINet [15], M²D-LIF [14], and GLFDet [25] adaptively calibrate multimodal features through illumination guidance, mono-modality distillation, or global-local feature optimization; however, these strategies may still miss fine-grained regional reliability variations. FDFuse [3] decomposes multimodal features into modality-shared and modality-specific components to model cross-modal correlation and complementarity; RSDet [26], DACFusion [27], and FMPFNet [28] select or calibrate complementary information. Global attention and simple weighting, however, still propagate local conflicts caused by degradation or residual misalignment.

The final detection loss mainly supervises classification and localization outputs and provides only indirect guidance for spatial alignment and modality selection during fusion [7,14]. Task-oriented image fusion methods improve downstream utility through task-specific guidance. TDFuse [29] and ITS [30] introduce task-related supervision, while T-OAR [31] conditions the fusion process on downstream task instructions through dynamic prompts. CF-Deformable DETR [7] and DPDETR [12] impose cross-modal consistency constraints, while DENet [32] dynamically balances RGB and IR contributions according to illumination conditions and cross-modal similarity, and MO R-CNN [33] regulates the relative contributions of modality branches. Although these strategies improve individual components, they do not jointly supervise spatial alignment, reliability-aware modality selection, and challenging small-object instances.

In summary, existing RGB-IR multispectral object detection methods have made progress in cross-modal spatial alignment, reliability-aware fusion, and training supervision, but several limitations remain. For cross-modal spatial misalignment, implicit alignment methods are convenient for end-to-end training, yet they lack an explicit offset modeling mechanism. Explicit correction methods can enhance cross-modal spatial consistency, but maintaining global alignment stability while preserving fine-grained local calibration remains challenging, especially under residual misalignment in small-object regions. For modality reliability differences and local response conflicts, existing methods can adjust the fusion process based on partial quality cues, but they still insufficiently model region-level reliability variations and suppress conflicting responses. For insufficient fusion supervision, existing supervision strategies mainly focus on final detection outputs or a single training objective, making it difficult to jointly constrain cross-modal alignment, reliable modality selection, and difficult small-object samples.

To address the limitations described above, we propose CFR-Net, which comprises three closely integrated components, namely SOAR, RQF, and AFC Loss. Existing alignment methods commonly rely on single-stage spatial deformation or local offset estimation, whereas SOAR adopts a coarse-to-fine alignment strategy that first corrects dominant global spatial shifts and then refines residual local misalignment. This design improves cross-modal alignment stability while better preserving fine-grained information in small-object regions. Compared with fixed fusion rules or simple attention-based gating, RQF jointly models modality quality and local cross-modal response conflicts and dynamically adjusts the fusion process according to region-specific reliability, thereby retaining complementary information while suppressing low-quality or conflicting feature responses. At the optimization level, AFC extends supervision beyond the final detection outputs to the alignment and fusion processes, jointly promoting spatial consistency and effective cross-modal fusion while strengthening the optimization of challenging small-object instances. Together, these modules form a progressive optimization framework that integrates spatial correction,

reliability-aware fusion, and process-level supervision, thereby improving RGB-IR multispectral object detection in complex real-world scenarios.

3 Proposed Method

3.1 Architecture Overview

CFR-Net inserts SOAR followed by RQF at the P3-P5 fusion levels of dual-branch YOLOv8n. The aligned, reliability-weighted features enter the standard neck and detection head, while AFC Loss supervises alignment, fusion, and small objects only during training as shown in Fig. 1.

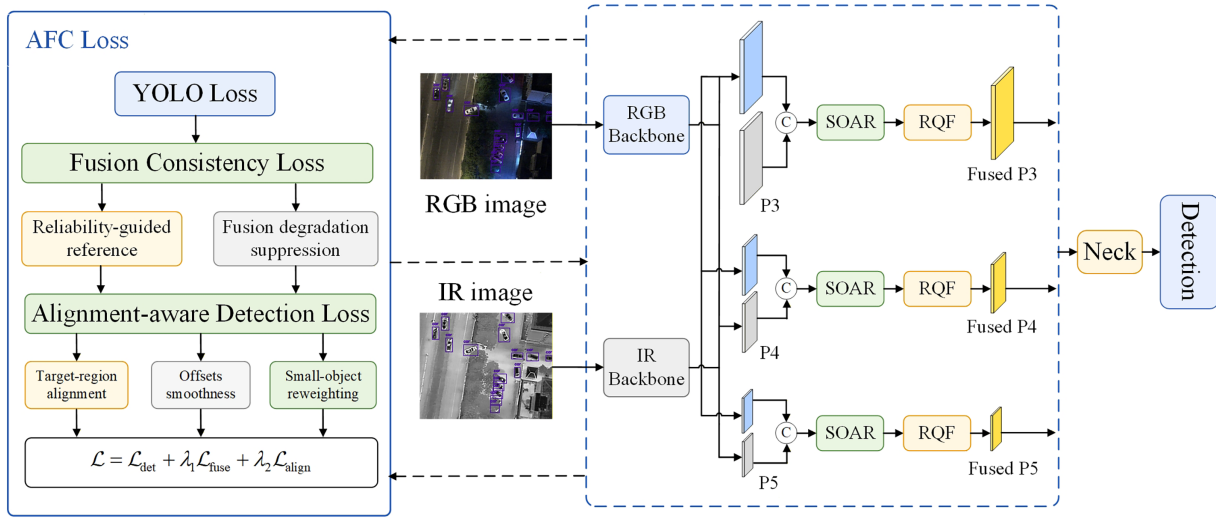


Figure 1: Overall architecture of the proposed CFR-Net framework.

3.2 Scale-Aware Offset Alignment and Residual Module

As shown in Fig. 2, the SOAR module consists of three stages arranged from left to right, including coarse alignment, fine alignment, and residual compensation. The two feature cubes on the left denote the input RGB feature F_v and IR feature F_t , while the feature cube on the right denotes the IR feature F_t^{soar} corrected by SOAR. The module first performs an initial correction for large-scale cross-modal spatial misalignment through global offset estimation. It then further removes local residual offsets through fine-grained residual correction and compensates for high-frequency details in small-object regions that may be weakened during resampling.

In the coarse alignment stage, SOAR estimates the dominant spatial offset between the RGB feature and the IR feature. It first constructs an initial cross-modal difference map D_0 from the two modality features. Since appearance differences alone cannot fully reflect the cross-modal misalignment relationship, SOAR further introduces the spatial gradient cues ∇F_v and ∇F_t , together with the phase-correlation cue. The spatial gradients are extracted by Sobel operators along the horizontal and vertical directions, where ∇F_v and ∇F_t denote the resulting Sobel-gradient maps of the RGB and IR features, respectively, each containing gradient responses in both directions. The phase-correlation cue Φ_0 is used to capture the global translational relationship between the two modalities. Specifically, F_v and F_t are first averaged along the channel dimension. Then, the normalized cross-power spectrum between them is computed and transformed back by the inverse Fourier transform to obtain:

$$\Phi_0 = \mathcal{F}^{-1} \left(\frac{\mathcal{F}(\text{Avg}_c(F_v)) \cdot \overline{\mathcal{F}(\text{Avg}_c(F_t))}}{|\mathcal{F}(\text{Avg}_c(F_v)) \cdot \overline{\mathcal{F}(\text{Avg}_c(F_t))}| + \varepsilon} \right) \quad (1)$$

where $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote the Fourier transform and inverse Fourier transform, respectively, $\overline{\mathcal{F}(\cdot)}$ denotes the complex conjugate, and ε is a numerical stability term. In the first stage of Fig. 2, D_0 , ∇F_v , ∇F_t and Φ_0 are fed into the coarse offset estimator E_c to obtain the coarse offset field:

$$\Delta_c = E_c([\nabla F_v; \nabla F_t; D_0; \Phi_0]) \quad (2)$$

where $[\cdot]$ denotes concatenation along the channel dimension. The first Warp module in Fig. 2 uses Δ_c to perform bilinear resampling on the IR feature, producing the coarsely aligned IR feature:

$$F_t^c = \text{Warp}(F_t, \Delta_c) \quad (3)$$

In the second stage of Fig. 2, the coarsely aligned F_t already alleviates the dominant spatial offset, but local residual misalignment may still remain. Therefore, the fine alignment stage no longer estimates the full offset directly. Instead, it only estimates the residual offset after coarse alignment. Its computation follows the same process as Δ_c . The computed D_1 , ∇F_v , ∇F_t^c and Φ_1 are fed into the fine offset estimator E_f to obtain the fine offset field:

$$\Delta_f = E_f([F_v; F_t^c; D_1; \nabla F_v; \nabla F_t^c; \Phi_1]) \quad (4)$$

The second Warp module in Fig. 2 performs the second correction based on Δ_f and F_t^c producing the finely aligned IR feature:

$$F_t^a = \text{Warp}(F_t^c, \Delta_f) \quad (5)$$

The coarse and fine offset fields jointly form the overall offset field of SOAR:

$$\Delta \approx \Delta_c + \Delta_f \quad (6)$$

With this coarse-to-fine design, SOAR handles large-range misalignment and local residual misalignment separately, thereby reducing the optimization difficulty of single-stage offset estimation. The third stage in Fig. 2 is the residual compensation branch, which restores high-frequency details that may be smoothed during alignment and resampling. For small objects, high-frequency responses such as edges, textures, and local structures often provide critical discriminative cues. Therefore, geometric alignment alone is insufficient, and high-frequency residual compensation is also required. However, high-frequency components may also contain noise or irrelevant texture patterns, especially in low-quality IR images or complex backgrounds. Directly adding high-frequency residuals without discrimination may introduce unreliable information into the aligned feature representation. To alleviate this issue, SOAR employs a learnable compensation gate that adaptively controls the contribution of high-frequency residual features according to the input content. Therefore, informative structural details can be enhanced while potentially noisy high-frequency responses are suppressed. As shown on the right side of Fig. 2, a low-pass filter (LPF) is first applied to extract the low-frequency component.

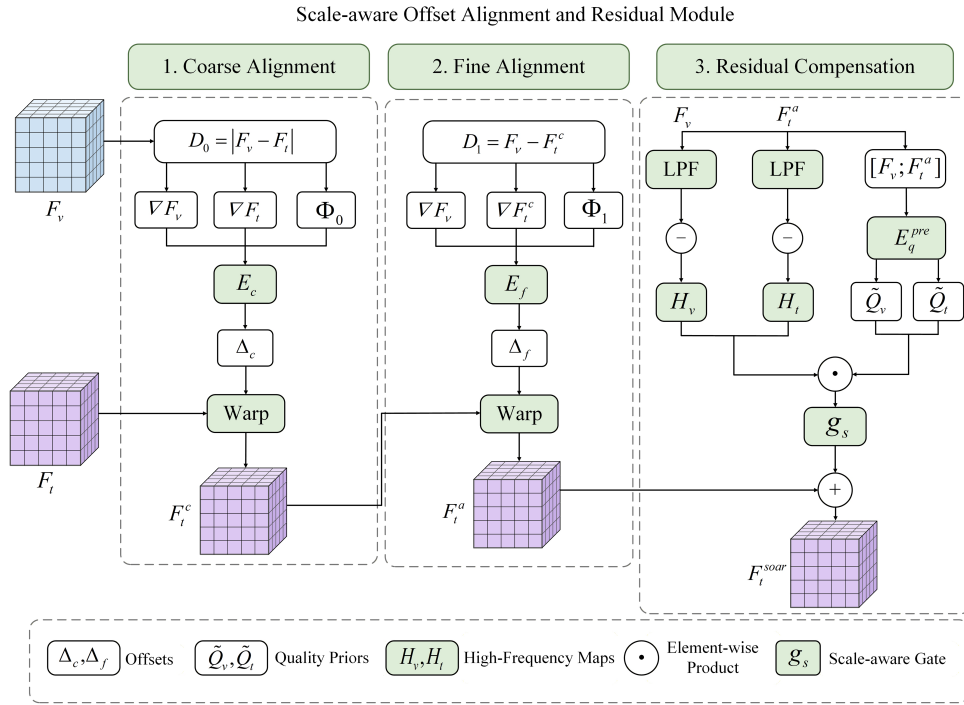


Figure 2: Architecture of the scale-aware offset alignment and residual module.

The low-frequency component is then subtracted from the original feature to obtain the high-frequency residual feature. Meanwhile, the shallow quality prior estimator E_q^{pre} in Fig. 2 generates two initial reliability priors from F_v and F_t^a :

$$(\tilde{Q}_v, \tilde{Q}_t) = E_q^{pre}([F_v; F_t^a]) \quad (7)$$

The high-frequency residual features and the quality priors $\tilde{Q}_v$ and $\tilde{Q}_t$ are then used together to generate the small-object compensation gate g_s :

$$g_s = \sigma(E_g([F_v - \text{LPF}(F_v); F_t^a - \text{LPF}(F_t^a); \tilde{Q}_v; \tilde{Q}_t])) \quad (8)$$

Meanwhile, the high-frequency residual is mapped into residual-detail tokens and added to the finely aligned feature F_t^a , yielding the final SOAR output:

$$F_t^{soar} = F_t^a + g_s \odot E_T([F_v - \text{LPF}(F_v); F_t^a - \text{LPF}(F_t^a)]) \quad (9)$$

where $\odot$ denotes element-wise multiplication. When their channel dimensions are inconsistent, g_s is broadcast along the channel dimension to C channels. Therefore, the SOAR output F_t^{soar} not only completes cross-modal spatial alignment, but also supplements high-frequency details in small-object regions, providing more reliable IR features for the subsequent RQF fusion.

3.3 Reliability-Aware Quality-Guided Fusion Module

As shown in Fig. 3, the input of the RQF module is the aligned feature produced by SOAR. The module consists of five stages: feature decomposition, quality estimation, conflict modeling, local-global interaction, and dynamic gated fusion. The two inputs on the left side of Fig. 3 are the RGB feature $X_v = F_v$ and the

SOAR-corrected IR feature $X_t = F_t^{soar}$. In the feature decomposition stage, X_v and X_t are decomposed into shared features and modality-specific features, respectively. The shared features represent structural and semantic information jointly described by RGB and IR, while the modality-specific features preserve information unique to each modality, such as texture and color in RGB and thermal radiation responses in IR. For each modality $m \in \{v, t\}$, the decomposition process is formulated as:

$$X_m^{sh} = P_{sh}(X_m), X_m^{sp} = P_{sp}^m(X_m) \quad (10)$$

where $P_{sh}(\cdot)$ denotes the shared projector, and $P_{sp}^m(\cdot)$ denotes the modality-specific projector. In Fig. 3, the RGB branch is decomposed into X_v^{sh} and X_v^{sp} , while the IR branch is decomposed into X_t^{sh} and X_t^{sp} . These features are then fed into the quality estimator Q to estimate the reliability of the RGB and IR modalities at each spatial location. The quality map is defined as:

$$Q_m = \sigma \left(\sum_{k=1}^7 \alpha_{m,k} \phi_k(X_m) \right) \quad (11)$$

where $\phi_k(\cdot)$ denotes the k -th quality indicator function, and $\alpha_{m,k}$ denotes the corresponding learnable weight. A larger Q_m^{ij} indicates that the feature of modality m is more reliable at spatial location (i, j) . In the conflict modeling stage of Fig. 3, the aligned cross-modal features X_v and X_t are used to compute cosine similarity. For spatial location (i, j) the channel-wise cosine similarity is defined as:

$$S_{\cos}^{ij} = \frac{\langle X_v^{ij}, X_t^{ij} \rangle}{\|X_v^{ij}\|_2 \cdot \|X_t^{ij}\|_2 + \varepsilon} \quad (12)$$

where X_v^{ij} and X_t^{ij} denote the RGB and IR feature vectors at spatial location (i, j) , respectively. Based on this similarity, the conflict map is defined as:

$$M_{\text{conf}}^{ij} = \frac{1 - S_{\cos}^{ij}}{2} \quad (13)$$

When M_{conf}^{ij} approaches 0, the two modalities show high consistency at this location. When M_{conf}^{ij} approaches 1, strong modal conflict exists at this location. This conflict map participates in the gated fusion stage to adjust the fusion strategy. Then, in the local-global cross-modal interaction stage of Fig. 3, this stage consists of local-window cross-attention for shared features and global prototype correction for modality-specific features. For shared features, the LWCA module denotes local-window cross-attention. For modality-specific features, the quality map Q_m is used to compute the quality-weighted prototype of each modality:

$$p_m = \frac{\sum_{i=1}^H \sum_{j=1}^W Q_m^{ij} X_m^{\text{sp},ij}}{\sum_{i=1}^H \sum_{j=1}^W Q_m^{ij} + \varepsilon} \quad (14)$$

where p_m denotes the global quality-weighted prototype of modality m . The prototype of the other modality is used to correct the modality-specific feature of the current modality. The shared enhanced feature and the corrected modality-specific feature are then integrated into a unified modality-enhanced feature:

$$Z_m = U_m [LWCA(X_m^{sh}, X_m^{sh}; r); PAttn(X_m^{sp}, p_m)] \quad (15)$$

where Z_m denotes the enhanced feature of either RGB or IR. In the dynamic quality-conflict gate in Fig. 3, the fusion weights of the two modalities are adaptively determined according to the quality map Q_m , the conflict map M_{conf} , and the enhanced feature Z_m . First, the dynamic gating weights of the two modalities are defined as:

$$G_m = \sigma(E_m([Z_m; Q_m; M_{conf}])) \quad (16)$$

where G_m controls the contribution of the enhanced RGB feature and the enhanced IR feature, respectively. Based on these two gating weights, the initial mixed feature is obtained as:

$$F_{mix} = \sum_{m \in \{v, t\}} G_m \odot Z_m \quad (17)$$

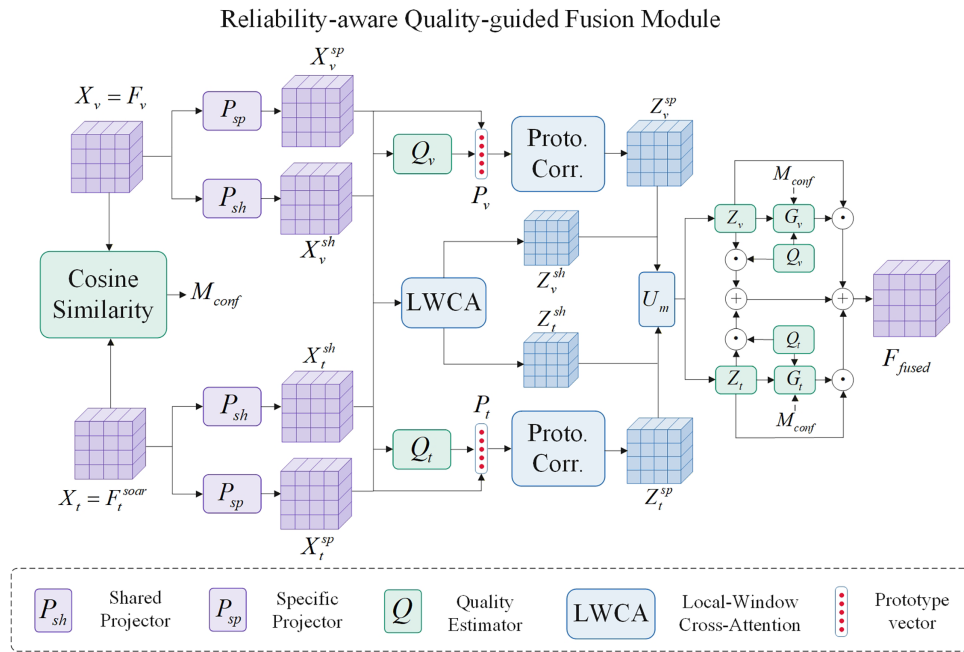


Figure 3: Architecture of the reliability-aware quality-guided fusion module.

To prioritize the more reliable modality in high-conflict regions, normalized reliability weights are further computed from the quality maps, and a fallback feature is constructed as:

$$F_{rel} = \sum_{m \in \{v, t\}} \frac{Q_m}{Q_v + Q_t + \varepsilon} \odot Z_m \quad (18)$$

here, F_{mix} denotes the regular dynamic fusion path, and F_{rel} denotes the reliability fallback path. When the conflict between modalities is weak, the fusion result mainly depends on F_{mix} . When the conflict is strong, it relies more on F_{rel} . During fusion, two modulation factors are computed:

$$\eta = \sigma(E_\eta([G_v; G_t; M_{conf}])) \quad (19)$$

$$\rho = \sigma(E_\rho(M_{conf})) \quad (20)$$

here, η modulates the overall strength of the mixed feature, while ρ controls the proportion of the reliability fallback feature in conflict regions. The final fused feature produced by RQF is:

$$F_{\text{fused}} = (1 - \rho) \odot (\eta \odot F_{\text{mix}}) + \rho \odot F_{\text{rel}} \quad (21)$$

here, F_{fused} serves as the main output, while Q_m and M_{conf} are produced as auxiliary outputs. These auxiliary outputs not only participate in the fusion computation of the current module, but are also used by AFC Loss during training for conflict-region selection, fusion degradation constraint, and small-object weighted supervision.

It is worth noting that while the cosine-similarity-based conflict map M_{conf} effectively captures directional discrepancies between cross-modal feature responses, it inherently focuses on directional consistency in the feature space. In scenarios with severe cross-modal polarity inversion, RGB and IR features corresponding to the same object may exhibit opposite directions while still containing complementary semantic information. In such cases, M_{conf} may interpret these responses as local conflicts. In CFR-Net, the potential impact of this limitation is partially mitigated by the reliability fallback path in Eq. (18), which adaptively emphasizes the modality with higher estimated quality when strong cross-modal conflict is detected. Nevertheless, cosine similarity alone cannot completely distinguish true semantic conflicts from complementary responses with opposite directions. Exploring more advanced conflict measures, such as cross-modal mutual information or higher-order feature correlation, therefore remains a promising direction for future improvement.

3.4 Alignment-Aware Fusion Consistency Loss

Standard YOLO loss supervises final detections but not cross-modal alignment, conflict-region fusion, or difficult small objects. AFC Loss therefore adds fusion-consistency and alignment-aware detection terms to the standard objective:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_1 \mathcal{L}_{\text{fuse}} + \lambda_2 \mathcal{L}_{\text{align}} \quad (22)$$

here, $\mathcal{L}_{\text{det}}$ denotes the standard YOLO detection loss, $\mathcal{L}_{\text{fuse}}$ denotes the fusion consistency loss, and $\mathcal{L}_{\text{align}}$ denotes the alignment-aware detection loss. λ_1 and λ_2 are the corresponding weighting coefficients.

The fusion consistency loss constrains the output feature of RQF. According to the previous definition, Z_v and Z_t denote the RGB and IR enhanced features in RQF, respectively; F_{fused} denotes the final fused feature, Q_v and Q_t denote the quality maps of the two modalities, and M_{conf} denotes the conflict map. For high-conflict regions, the fused feature is expected to be close to the reliability-guided fallback feature, which provides a more stable reference by aggregating the enhanced RGB and IR features according to their estimated quality. Meanwhile, the discriminative ability of the fused feature should not be weaker than that of the more informative single-modality representation. Therefore, the fusion consistency loss is defined as:

$$\mathcal{L}_{\text{fuse}} = \text{Avg}_{\Omega_{\text{conf}}} \| F_{\text{fused}} - F_{\text{rel}} \|_2^2 + \beta_{\text{deg}} \text{Avg} [\max(\mathcal{S}(Z_v), \mathcal{S}(Z_t)) - \mathcal{S}(F_{\text{fused}}) + \epsilon]_+ \quad (23)$$

where $\Omega_{\text{conf}} = \{(i, j) \mid M_{\text{conf}}^{ij} > \theta_{M_{\text{conf}}}\}$ denotes the high-conflict region determined by the conflict map M_{conf} . F_{rel} denotes the reliability fallback feature defined in Eq. (18). It is constructed by aggregating the enhanced RGB and IR features in a quality-weighted manner according to the quality maps Q_v and Q_t , thereby providing a more reliable reference representation for the fused feature in high-conflict regions. $\mathcal{S}(\cdot)$ denotes the task-related quality scoring function, ϵ is a safety margin, and β_{deg} is the balancing coefficient for the fusion degradation constraint. This loss constrains the fused feature to approach the reliability-guided reference feature in high-conflict regions while suppressing fusion degradation.

The alignment-aware detection loss constrains the alignment result of SOAR and strengthens small-object detection supervision. According to the previous definition, F_v denotes the RGB feature, F_t^{soar} denotes the aligned IR feature output by SOAR, and the total offset field is $\Delta = \Delta_c + \Delta_f$. The alignment-aware detection loss is defined as:

$$\mathcal{L}_{\text{align}} = \text{Avg}_{\Omega_{\text{obj}}} \| F_v - F_t^{soar} \|_2^2 + \mu \text{Avg} \| \nabla \Delta \|_2^2 + \beta_{\text{small}} \text{Avg}_n \left(w_n \mathcal{L}_{\text{det}}^{(n)} \right) \quad (24)$$

here, Ω_{obj} denotes the target region. The first term constrains the consistency between the RGB feature F_v and the aligned IR feature F_t^{soar} within the target region. The second term constrains the spatial smoothness of the total offset field Δ , where $\nabla \Delta$ denotes the gradient of the offset field and μ is the smoothness regularization coefficient. The third term is the weighted detection loss, and β_{small} is the balancing coefficient for the small-object weighting term. The composite weight of the n -th object is defined as:

$$w_n = \exp \left(-\alpha \frac{a_n}{a_{\text{max}}} \right) (1 - p_n)^\gamma \overline{M}_{\text{conf}} \quad (25)$$

where a_n denotes the area of the n -th object, a_{max} denotes the maximum object area in the current batch, p_n denotes the prediction confidence, and $\overline{M}_{\text{conf}}$ denotes the average conflict intensity computed from the conflict map M_{conf} within the target box. α and γ control the scale weight and the hard-sample weight, respectively. This weighting scheme assigns stronger supervision to small, low-confidence, and high-conflict objects.

4 Experiments

4.1 Experimental Setup

During training, all input images are uniformly resized to 640×640 pixels. All models are trained for 300 epochs using the stochastic gradient descent optimizer with a batch size of 16. The initial learning rate, weight decay, and momentum are set to 0.005, 0.0005, and 0.937, respectively, and a five-epoch learning-rate warm-up strategy is adopted. The weights assigned to the bounding-box regression loss, classification loss, and Distribution Focal Loss are set to 7.5, 0.5, and 1.5, respectively. Moderate data augmentation is applied during training to improve model generalization. Specifically, the hue, saturation, and value gains in the hue, saturation, and value (HSV) color space are set to 0.015, 0.7, and 0.4, respectively. The translation and scaling factors are set to 0.1 and 0.5, while the probability of horizontal flipping is set to 0.5. Rotation, shearing, perspective transformation, vertical flipping, MixUp, and Copy-Paste are disabled to avoid potential interference with target structures and cross-modal spatial correspondence caused by excessive geometric transformations or sample mixing. All experiments are conducted on Ubuntu 22.04 using Python 3.10, PyTorch 2.1.0, and CUDA 12.1. The hardware platform consists of a single NVIDIA RTX 5090 GPU with 32 GB of memory.

Following the Common Objects in Context (COCO) convention [34], we define object scale according to the bounding-box area $A = w \times h$ in the original image coordinates: small objects have $A < 32^2$, medium objects have $32^2 \leq A < 96^2$, and large objects have $A \geq 96^2$ pixels². Counting each registered visible-infrared image pair once, DroneVehicle contains 66,761 small objects among 452,566 instances (14.75%), and 38.17% of its image pairs contain at least one small object. M3FD contains 14,072 small objects among 34,408 instances (40.90%), and 62.45% of its image pairs contain at least one small object.

We evaluate on M3FD and DroneVehicle. M3FD contains paired visible-infrared traffic scenes with illumination, background, and scale variation; DroneVehicle contains aligned UAV RGB-IR imagery dominated by small vehicles, clutter, pose variation, and inter-class similarity. We report mean Average Precision at

an intersection over union (IoU) threshold of 0.50 (mAP50), mean Average Precision at an intersection over union (IoU) threshold of 0.50 to 0.95 (mAP50:95), parameters, and giga floating-point operations (GFLOPs) to measure detection, localization, model size, and computation.

4.2 Comparison and Analysis

CFR-Net is developed on a dual-stream YOLOv8n architecture [35]. YOLOv8n is selected primarily because its compact model size aligns well with the lightweight and efficiency-oriented objectives of this study. More importantly, adopting a lightweight backbone limits the influence of excessive model capacity on performance improvements, thereby allowing the effectiveness of the proposed cross-modal modules to be assessed more objectively. It should be noted that SOAR, RQF, and AFC Loss in CFR-Net are designed around feature representation and detection optimization. SOAR and RQF do not rely on YOLO-specific operations and therefore have the potential to be adapted to other detection architectures with compatible feature representations. Although the implementation of AFC Loss is associated with detector-specific training interfaces, its core idea of introducing process-level supervision for feature alignment, fusion effectiveness, and hard-sample optimization is not restricted to YOLO-based detectors. With appropriate adaptation, this supervision strategy can also be extended to other object detection frameworks. To ensure a fair and controlled comparison, all ablation experiments use identical backbone architectures, data splits, and training configurations, with only the module or loss term under investigation being replaced or removed.

4.2.1 Experiments on the DroneVehicle Dataset

As shown in Table 1, CFR-Net achieves 82.1% mAP50 and 61.7% mAP50:95 on the DroneVehicle dataset, with only 6.5M parameters and 16.9 GFLOPs. Compared with COXNet, CFR-Net improves mAP50 and mAP50:95 by 3.3 and 4.9 percentage points, respectively, while reducing the parameter count by 64.8M and the computational cost by 85.7 GFLOPs. Although ICAFusion obtains slightly higher detection accuracy, exceeding CFR-Net by 2.1 points in mAP50 and 0.7 points in mAP50:95, it requires 120.2M parameters and 191.6 GFLOPs. In contrast, CFR-Net uses approximately 94.6% fewer parameters and 91.2% fewer GFLOPs, while maintaining comparable detection performance. These results demonstrate that CFR-Net provides a more favorable balance between detection accuracy and computational efficiency.

Table 1: Comparison of results on the DroneVehicle dataset.

Methods	mAP50	mAP50:95	Params. (M)	GFLOPs
CFT [36]	65.4	45.7	206.0	224.4
TAF-YOLO [37]	65.8	46.7	4.52	8.4
UA-CMDet [8]	64.0	40.1	139.7	-
C ² Former [4]	72.8	42.8	100.8	89.9
DEYOLO [5]	81.7	59.5	6.0	-
ICAFusion [1]	84.2	62.4	120.2	191.6
COXNet [11]	78.8	56.8	71.3	102.6
Ours	82.1	61.7	6.5	16.9

Note: The results obtained by our method are highlighted in bold.

Fig. 4 qualitatively compares the detection results of the dual-stream baseline, ICAFusion, COXNet, and CFR-Net against the ground-truth annotations. Yellow and green arrows indicate missed detections and false positives, respectively. In scenes with dense small targets and complex backgrounds, CFR-Net reduces both

errors and produces predictions that more closely match the ground truth, demonstrating the effectiveness of its cross-modal alignment and reliability-aware fusion.

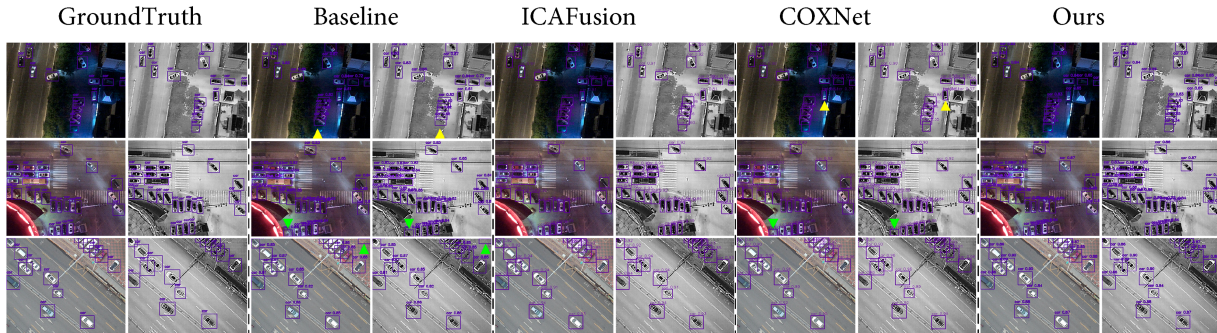


Figure 4: Visual comparison of detection results.

4.2.2 Experiments on the M3FD Dataset

As reported in Table 2, CFR-Net achieves 85.7% mAP50 and 56.3% mAP50:95 on the M3FD dataset. Compared with the lightweight DEYOLO, it improves mAP50 and mAP50:95 by 0.6 and 2.8 percentage points, respectively, with only 0.5M additional parameters. Although MDM remains ahead by 0.9 and 1.6 points, CFR-Net reduces the parameter count and computational cost by 95.7% and 92.9%, respectively, requiring only 6.5M parameters and 16.9 GFLOPs, compared with 149.7M parameters and 237.8 GFLOPs for MDM. These results demonstrate that CFR-Net achieves a more favorable balance between detection accuracy and computational efficiency.

Table 2: Comparison of results on the M3FD dataset with different algorithms.

Methods	mAP50	mAP50:95	Param. (M)	GFLOPs
CFT [36]	71.4	40.8	206.0	224.4
MMI-Det [38]	83.5	51.9	207.6	229.2
Fusion-Mamba [39]	85.0	57.5	244.6	-
DEYOLO [5]	85.1	53.5	6.0	-
MDM [40]	86.6	57.9	149.7	237.8
Ours	85.7	56.3	6.5	16.9

Note: The results obtained by our method are highlighted in bold.

4.3 Ablation Experiments

4.3.1 Effectiveness Analysis of SOAR and RQF Modules

Table 3 evaluates SOAR and RQF on DroneVehicle using the stronger IR-only YOLOv8n model, which achieves 75.2% mAP50 and 53.1% mAP50:95, as the baseline. Introducing SOAR alone increases the two metrics to 80.6% and 59.8%, respectively, confirming the effectiveness of spatial offset correction and detail compensation. RQF alone further achieves 81.4% mAP50 and 60.6% mAP50:95, demonstrating the benefit of reliability-aware fusion. When both modules are integrated, CFR-Net reaches 82.1% mAP50 and 61.7% mAP50:95, outperforming the baseline by 6.9 and 8.6 percentage points, respectively. Under the stricter mAP50:95 metric, the combined model exceeds SOAR and RQF by 1.9 and 1.1 points, indicating that spatial alignment and reliability modeling provide complementary benefits.

Table 3: Ablation study of SOAR and RQF modules.

SOAR	RQF	Data	mAP50	mAP50:95
×	×	RGB	72.3	50.7
×	×	IR	75.2	53.1
✓	×	RGB + IR	80.6	59.8
×	✓	RGB + IR	81.4	60.6
✓	✓	RGB + IR	82.1	61.7

Note: ✓ and × indicate that the corresponding module is enabled and disabled, respectively. The results obtained by our method are highlighted in bold.

4.3.2 Effectiveness Analysis of AFC Loss Components

Table 4 presents the ablation results for AFC. Here, $\mathcal{L}_{\text{det}}$ denotes the standard detection loss, while $\mathcal{L}_{\text{fuse}}$ and $\mathcal{L}_{\text{align}}$ enforce fusion consistency and alignment consistency, respectively. The RGB-only baseline achieves 72.3% mAP50 and 50.7% mAP50:95, whereas the IR-only baseline achieves 75.2% mAP50 and 53.1% mAP50:95. With SOAR and RQF enabled, combining $\mathcal{L}_{\text{align}}$ with $\mathcal{L}_{\text{det}}$ yields 81.5% mAP50 and 60.8% mAP50:95, while combining $\mathcal{L}_{\text{fuse}}$ with $\mathcal{L}_{\text{det}}$ yields 81.8% mAP50 and 61.1% mAP50:95. Jointly applying both auxiliary losses further improves the performance to 82.1% mAP50 and 61.7% mAP50:95, indicating that fusion supervision and alignment supervision provide complementary benefits.

Table 4: Ablation study of AFC loss components.

SOAR	RQF	$\mathcal{L}_{\text{det}}$	$\mathcal{L}_{\text{fuse}}$	$\mathcal{L}_{\text{align}}$	Data	mAP50	mAP50:95
×	×	✓	×	×	RGB	72.3	50.7
×	×	✓	×	×	IR	75.2	53.1
✓	✓	✓	×	✓	RGB + IR	81.5	60.8
✓	✓	✓	✓	×	RGB + IR	81.8	61.1
✓	✓	✓	✓	✓	RGB + IR	82.1	61.7

Note: ✓ and × indicate that the corresponding component is enabled and disabled, respectively. The results obtained by our method are highlighted in bold.

4.3.3 Effectiveness Analysis of the Proposed Fusion Strategy

As shown in Table 5, fixed Add and Concat fusion achieve 79.5% and 80.3% mAP50, together with 58.8% and 59.4% mAP50:95, respectively. In comparison, the proposed fusion method reaches 82.1% mAP50 and 61.7% mAP50:95, outperforming Add by 2.6 and 2.9 percentage points and Concat by 1.8 and 2.3 percentage points. These improvements demonstrate that region-adaptive quality assessment and conflict-aware regulation are more effective than fixed fusion rules for integrating complementary RGB and IR information.

4.3.4 Robustness Analysis of High-Frequency Residual Compensation

To investigate whether the high-frequency (HF) residual compensation in SOAR remains robust when high-frequency components are dominated by noise rather than useful structural information, we conduct a controlled perturbation experiment. Specifically, random phase perturbations are introduced into the high-frequency residual features to simulate unreliable high-frequency responses while preserving the overall feature distribution. Four variants are compared: (1) Original HF Residual+Gate, which represents the default

SOAR compensation strategy; (2) Noise-Perturbed HF Residual+Gate, where the high-frequency residual is disturbed before being adaptively weighted by the compensation gate; (3) HF Residual Disabled, where the high-frequency compensation branch is removed; (4) Noise-Perturbed HF Residual without Gate, where the perturbed residual is directly added without adaptive modulation. The experimental results are reported in Table 6.

Table 5: Ablation study of different fusion methods.

Method	Data	mAP50	mAP50:95
×	RGB	72.3	50.7
×	IR	75.2	53.1
Add	RGB + IR	79.5	58.8
Concat	RGB + IR	80.3	59.4
Ours	RGB + IR	82.1	61.7

Note: × indicates that no fusion method is applied. The results obtained by our method are highlighted in bold.

Table 6: Results of high-frequency residual noise perturbation and learned gating.

Method	mAP50	mAP50:95
Original HF Residual+Gate	82.1	61.7
Noise-Perturbed HF Residual+Gate	79.1	58.5
HF Residual Compensation Disabled	79.1	58.7
Noise-Perturbed HF Residual without Gating	77.2	56.1

Note: Following common frequency perturbation strategies, the noise perturbation is generated via phase-randomization in the frequency domain, preserving spectral magnitude while scrambling spatial structures.

As shown in Table 6, perturbing the high-frequency residual leads to a performance degradation compared with the original compensation strategy, indicating that unreliable high-frequency responses may negatively affect feature representation. However, the Noise-Perturbed HF Residual+Gate variant maintains comparable performance to the version without high-frequency compensation, while the removal of the gate causes a more significant decline. These results demonstrate that the learnable compensation gate can adaptively regulate residual contributions and suppress unreliable high-frequency responses, enabling SOAR to effectively exploit informative details while maintaining robustness against noisy components.

4.3.5 Efficiency of the Proposed SOAR Module

As shown in Table 7, this ablation study progressively introduces the components of SOAR into the dual-stream YOLOv8n-Concat baseline without including the subsequent RQF module, thereby allowing the relative computational influence of each component to be evaluated under a consistent architecture and fusion setting. The parameter count and computational cost increase gradually as coarse alignment, fine alignment, and residual compensation are introduced. Compared with the baseline, coarse alignment adds only 0.29M parameters and 0.54 GFLOPs, indicating that the initial correction of dominant spatial offsets incurs limited computational overhead. Building on this stage, fine alignment introduces an additional 0.46M parameters and 1.07 GFLOPs, while residual compensation further adds 0.74M parameters and 1.70 GFLOPs. Overall, the additional complexity of SOAR is mainly associated with local residual

offset refinement and high-frequency detail compensation, whereas coarse alignment performs the initial correction of large-scale cross-modal misalignment at a relatively low computational cost.

Table 7: Computational overhead of different SOAR configurations.

Coarse Alignment	Fine Alignment	Residual Compensation	Param. (M)	GFLOPs
×	×	×	4.37	11.52
✓	×	×	4.66	12.06
✓	✓	×	5.12	13.13
✓	✓	✓	5.86	14.83

Note: ✓ and × indicate that the corresponding component is enabled and disabled, respectively.

5 Conclusion

This paper presents CFR-Net, a lightweight framework for RGB-IR multispectral object detection that addresses cross-modal spatial misalignment, differences in modality reliability, and insufficient supervision of the fusion process. SOAR employs coarse-to-fine cross-modal offset estimation to correct global shifts and residual local misalignment sequentially, while residual compensation strengthens detailed responses in small-object regions. RQF evaluates modality reliability from regional quality and local response conflicts, preserving effective cross-modal complementary information while suppressing low-quality or conflicting responses. AFC Loss jointly supervises feature alignment, fusion effectiveness, and difficult small-object detection, thereby providing more direct optimization constraints for cross-modal alignment and fusion beyond the standard detection loss. Experiments on the DroneVehicle and M3FD datasets show that CFR-Net achieves competitive performance in small-object detection, localization accuracy, and robustness under complex backgrounds. Ablation studies further verify the effectiveness of SOAR, RQF, and AFC Loss, demonstrating the complementary roles of cross-modal alignment, reliability-aware fusion, and process-level supervision. CFR-Net contains only 6.5M parameters and requires 16.9 GFLOPs, achieving a favorable balance between detection accuracy and model efficiency.

The current study still has several limitations. The evaluation is restricted to the two public datasets, DroneVehicle and M3FD, and therefore does not fully establish the scalability of CFR-Net to larger datasets and more complex application scenarios. Differences among sensors in spatial resolution, field of view, imaging noise, and spectral response may also affect the accuracy of offset estimation and modality reliability modeling. Future work will evaluate CFR-Net on larger and more diverse multispectral datasets and systematically assess its generalization through cross-dataset and cross-sensor validation. Sensor-adaptive learning and domain generalization will also be investigated to improve robustness under severe misalignment, sensor variation, and incomplete modalities. Model compression, knowledge distillation, and adaptive module activation will be explored to further reduce computational cost. Overall, CFR-Net balances detection accuracy, model efficiency, and adaptability to complex scenarios, providing an effective lightweight solution for RGB-IR multispectral object detection.

Acknowledgement: We sincerely thank our collaborators for their valuable contributions, insightful discussions, and constructive feedback.

Funding Statement: This work was supported by the Henan Provincial University Scientific and Technological Innovation Talent Program (Grant 25HASTIT030), the Humanities and Social Science Fund of the Ministry of Education (Grant 24YJAZH229), and the Science and Technology Research Project of Henan Province (Grants 262102211063, 262102211047, and 262102211032).

Author Contributions: Conceptualization, Ye Li and Weiwei Zhang; methodology, Ye Li, Haotian Ning and Weiwei Zhang; software, Haotian Ning and Weiwei Zhang; validation, Xiaoke Su, Tong Xiao and Weiwei Zhang; formal analysis, Ye Li and Weiwei Zhang; investigation, Ye Li, Haotian Ning and Weiwei Zhang; resources, Ye Li, Haotian Ning and Weiwei Zhang; data curation, Xiaoke Su and Tong Xiao; writing—original draft preparation, Ye Li, Haotian Ning and Weiwei Zhang; writing—review and editing, Ye Li, Haotian Ning and Weiwei Zhang; visualization, Haotian Ning, Xiaoke Su and Tong Xiao; supervision, Weiwei Zhang; project administration, Ye Li and Weiwei Zhang; funding acquisition, Ye Li, Xiaoke Su and Weiwei Zhang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data and the code of this study are available from the corresponding author upon request.

Ethics Approval: Not applicable.

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

References

1. Shen J, Chen Y, Liu Y, Zuo X, Fan H, Yang W. ICAFusion: iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognit.* 2024;145(1):109913. doi:10.1016/j.patcog.2023.109913.
2. Liu J, Fan X, Huang Z, Wu G, Liu R, Zhong W, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 5792–801. doi:10.1109/CVPR52688.2022.00571.
3. Cheng M, Huang H, Liu X, Mo H, Wu S, Zhao X. FDFuse: infrared and visible image fusion based on feature decomposition. *IEEE Trans Instrum Meas.* 2025;74:5021413. doi:10.1109/TIM.2025.3551460.
4. Yuan M, Wei X. C²Former: calibrated and complementary transformer for RGB-infrared object detection. *IEEE Trans Geosci Remote Sens.* 2024;62:5403712. doi:10.1109/TGRS.2024.3376819.
5. Chen Y, Wang B, Guo X, Zhu W, He J, Liu X, et al. DEYOLO: dual-feature-enhancement YOLO for cross-modality object detection. In: *Pattern recognition*. Cham, Switzerland: Springer Nature; 2025. p. 236–52. doi:10.1007/978-3-031-78447-7_16.
6. Xiong Z, Yao Z, Liu X, Zhao W, Cao J, Wu X. Efficient multispectral object detection with attentive feature aggregation leveraging zero-shot implicit illumination guidance. *Inf Fusion.* 2025;118(4):102939. doi:10.1016/j.inffus.2025.102939.
7. Fu H, Yuan J, Zhong G, He X, Lin J, Li Z. CF-deformable DETR: an end-to-end alignment-free model for weakly aligned visible-infrared object detection. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*; 2024 Aug 3–9; Jeju, Republic of Korea.
8. Sun Y, Cao B, Zhu P, Hu Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Trans Circuits Syst Video Technol.* 2022;32(10):6700–13. doi:10.1109/TCSVT.2022.3168279.
9. Liu Z, Luo H, Wang Z, Peng X, Zuo H, Zhang J. Cross-modal offset-guided dynamic alignment and fusion for weakly aligned UAV object detection. *Appl Soft Comput.* 2026;203(11):116028. doi:10.1016/j.asoc.2026.116028.
10. Zhou C, Cheng P, Fang J, Zhang Y, Yan Y, Jia X, et al. Optimizing multispectral object detection: a bag of tricks and comprehensive benchmarks. *arXiv:2411.18288*. 2024.
11. Peng P, Xu T, Song L, Zhu M, Fang Y, Li J. COXNet: cross-layer fusion with adaptive alignment and scale integration for RGBT tiny object detection. *IEEE Trans Circuits Syst Video Technol.* 2026;36(1):596–608. doi:10.1109/TCSVT.2025.3595147.
12. Guo J, Gao C, Liu F, Meng D. DPDETR: decoupled position detection transformer for infrared-visible object detection. *arXiv:2408.06123*. 2024.
13. Yuan M, Shi X, Wang N, Wang Y, Wei X. Improving RGB-infrared object detection with cascade alignment-guided transformer. *Inf Fusion.* 2024;105(5):102246. doi:10.1016/j.inffus.2024.102246.

14. Zhao T, Liu B, Gao Y, Sun Y, Yuan M, Wei X. Rethinking multi-modal object detection from the perspective of mono-modality feature learning. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV); 2025 Oct 19–25; Honolulu, HI, USA. p. 6364–73. doi:10.1109/ICCV51701.2025.00600.
15. Zhang Y, Yu H, He Y, Wang X, Yang W. Illumination-guided RGBT object detection with inter- and intra-modality fusion. *IEEE Trans Instrum Meas.* 2023;72:2508013. doi:10.1109/TIM.2023.3251414.
16. Sun T, Xiang D, Ding T, Fang X, Qi Y, Zhao Z. Modality-aware infrared and visible image fusion with target-aware supervision. *arXiv:2509.11476.* 2025.
17. Xing Y, Yang S, Wang S, Zhang S, Liang G, Zhang X, et al. MS-DETR: multispectral pedestrian detection transformer with loosely coupled fusion and modality-balanced optimization. *IEEE Trans Intell Transp Syst.* 2024;25(12):20628–42. doi:10.1109/TITS.2024.3450584.
18. Tang L, Yan Q, Xiang X, Fang L, Ma J. C2RF: bridging multi-modal image registration and fusion via commonality mining and contrastive learning. *Int J Comput Vis.* 2025;133(8):5262–80. doi:10.1007/s11263-025-02427-1.
19. Ma L, Lin S, Zeng J, Wang Y, Jin Z. CLAFusion: misaligned infrared and visible image fusion based on contrastive learning and collaborative attention. *Comput Vis Image Underst.* 2026;263(8):104574. doi:10.1016/j.cviu.2025.104574.
20. Zhang Q, Jia X, Abdelmoniem AM, Zhou L, An J. Small target detection based on mask-enhanced attention fusion of visible and infrared remote sensing images. *arXiv:2603.06925.* 2026.
21. Li R, Xiao C, Yin Q, An W, Chen N, Ying X, et al. Dynamic high-frequency convolution for infrared small target detection. *arXiv:2602.02969.* 2026.
22. Hwang S, Han D, Jeon M. Multispectral detection transformer with infrared-centric sensor fusion. *arXiv:2505.15137.* 2025.
23. Ying X, Xiao C, An W, Li R, He X, Li B, et al. Visible-thermal tiny object detection: a benchmark dataset and baselines. *IEEE Trans Pattern Anal Mach Intell.* 2025;47(7):6088–96. doi:10.1109/TPAMI.2025.3544621.
24. Song K, Xue X, Wen H, Ji Y, Yan Y, Meng Q. Misaligned visible-thermal object detection: a drone-based benchmark and baseline. *IEEE Trans Intell Veh.* 2024;9(11):7449–60. doi:10.1109/TIV.2024.3398429.
25. Chen Z, Ji H, Zhang Y. Global-local feature optimization based RGB-IR *fusi* on object detection on drone view. *Chin J Aeronaut.* 2026;39(1):103781. doi:10.1016/j.cja.2025.103781.
26. Zhao T, Yuan M, Jiang F, Wang N, Wei X. Removal then selection: a coarse-to-fine fusion perspective for RGB-infrared object detection. *IEEE Trans Intell Transp Syst.* 2026;27(2):2504–19. doi:10.1109/TITS.2025.3638627.
27. Qian J, Qiao B, Zhang Y, Liu T, Wang S, Wu G, et al. DACFusion: dual asymmetric cross-attention guided feature fusion for multispectral object detection. *Neurocomputing.* 2025;635(4):129913. doi:10.1016/j.neucom.2025.129913.
28. Wang Z, Zhang Q. Real-time aerial multispectral object detection with dynamic modality-balanced pixel-level fusion. *Sensors.* 2025;25(10):3039. doi:10.3390/s25103039.
29. Zhang K, Zhang P, Sun L, Zhang F, Wan W, Sun J. Task-driven infrared and visible image fusion via detail and semantic dual injection. *Neural Netw.* 2026;202(9):109003. doi:10.1016/j.neunet.2026.109003.
30. Chen B, Luo S, Wu H, Chen M, He C. Infrared and visible image fusion and detection based on interactive training strategy and feature filter extraction module. *Opt Laser Technol.* 2024;179:111383. doi:10.1016/j.optlastec.2024.111383.
31. Yang Z, Zhang Y, Li H, Liu Y. Instruction-driven fusion of infrared-visible images: tailoring for diverse downstream tasks. *Inf Fusion.* 2025;121(5):103148. doi:10.1016/j.inffus.2025.103148.
32. Wang Q, Sun Y, Chi Y, Shen T. RGB-T object detection with failure scenarios. *IEEE J Sel Top Appl Earth Obs Remote Sens.* 2025;18:3000–10. doi:10.1109/JSTARS.2024.3523408.
33. Wang L, Jin B, Huang F, Chen L, Wang Z, He X, et al. MO R-CNN: multispectral oriented R-CNN for object detection in remote sensing image. *arXiv:2509.16957.* 2025.
34. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: *Computer Vision—ECCV 2014.* Cham, Switzerland: Springer International Publishing; 2014. p. 740–55. doi:10.1007/978-3-319-10602-1_48.
35. Jocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8 [software]. Version 8.0.0. 2023 [cited 2026 Jan 1]. Available from: <https://github.com/ultralytics/ultralytics>.

36. Fang Q, Han D, Wang Z. Cross-modality fusion transformer for multispectral object detection. arXiv:2111.00273. 2021. doi:10.48550/arXiv.2111.00273.
37. Zhuo Z, Lu R, Yao Y, Wang S, Zheng Z, Zhang J, et al. TAF-YOLO: a small-object detection network for UAV aerial imagery via visible and infrared adaptive fusion. Remote Sens. 2025;17(24):3936. doi:10.3390/rs17243936.
38. Zeng Y, Liang T, Jin Y, Li Y. MMI-det: exploring multi-modal integration for visible and infrared object detection. IEEE Trans Circuits Syst Video Technol. 2024;34(11):11198–213. doi:10.1109/TCSVT.2024.3418965.
39. Dong W, Zhu H, Lin S, Luo X, Shen Y, Guo G, et al. Fusion-mamba for cross-modality object detection. IEEE Trans Multimed. 2025;27:7392–406. doi:10.1109/TMM.2025.3599020.
40. Zhang Y, Chai L. MDM: modality decoupling for visible and infrared mamba-based object detection. Neural Netw. 2026;202(10):108988. doi:10.1016/j.neunet.2026.108988.