Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Reliability-Guided RGB-IR Object Detection Network with Complementary Information Decoupling for Autonomous Driving

Miaomiao Yang

yangmiaomiao@chzc.edu.cn

Chuzhou Polytechnic

Fan Guo

Chuzhou Polytechnic

Lin Cheng

Chuzhou Polytechnic

Ping Fang

Chuzhou Polytechnic

Article

Keywords: Autonomous driving, RGB-IR object detection, multispectral image fusion, feature-level fusion, road object detection

Posted Date: July 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10082289/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on October 5th, 2026.
See the published version at https://doi.org/10.1038/s41598-026-71402-8.

A Reliability-Guided RGB-IR Object Detection Network with Complementary Information Decoupling for Autonomous Driving

Miaomiao Yang[a],*, Fan Guo[a], Lin Cheng[a] and Ping Fang[a]
a Chuzhou Polytechnic, Chuzhou, China, 239000;
* Correspondence: yangmiaomiao@chzc.edu.cn

Abstract: Autonomous driving systems require robust object detection under complex open-road conditions, where RGB images are sensitive to illumination changes and background interference, while infrared images may suffer from weak textures and unreliable thermal responses. To improve RGB-IR detection performance, this paper proposes a Feature-level Scale-aware Multispectral Fusion Network, termed FSMF, for autonomous driving scenarios. The network consists of three key components. First, a Reliability-Guided Modulation module estimates RGB and IR feature reliability from channel, spatial, and cross-modal perspectives, and adaptively adjusts their contributions before fusion. Second, a Complementary Information Decoupling module separates shared semantics, complementary differences, and redundant interference, thereby enhancing target-related cross-modal cues while suppressing ineffective responses. Third, a Scale-Aware Fusion Pyramid Network performs differentiated enhancement for shallow, middle, and deep features to improve multi-scale feature representation. Experiments on M3FD and FLIR demonstrate the effectiveness of FSMF. The proposed FSMF-s achieves 85.9% mAP@50 on M3FD, and 72.2% mAP@50 and 38.3% mAP@50:95 on FLIR, outperforming mainstream detectors and existing RGB-IR methods. Ablation results further verify the effectiveness of the proposed modules.

Keywords: Autonomous driving; RGB-IR object detection; multispectral image fusion; feature-level fusion; road object detection

1 Introduction

The safe operation of autonomous driving systems highly depends on accurate perception of the surrounding road environment, in which object detection plays a key role in locating and identifying pedestrians, vehicles, cyclists, traffic facilities, potential obstacles, and other targets. Existing reviews on low-light multimodal object detection have pointed out that achieving stable perception under complex illumination conditions is an important research issue in intelligent transportation scenarios [1]. As advanced driver assistance systems and autonomous vehicles gradually move from structured road testing to open traffic scenarios, visual perception models are required not only to maintain high detection accuracy under sufficient daytime illumination, but also to provide stable and reliable target responses in non-ideal environments, such as nighttime scenes, tunnel entrances and exits, strong backlight, rain and fog, smoke and dust, and local occlusion [2]. Traditional RGB image-based detection methods can fully utilize color, texture, edge, and geometric structure information, and thus show strong target representation capability under normal visual conditions. However, their imaging quality is easily affected by changes in ambient illumination, weather interference, and sensor noise [3]. When the brightness of the target is close to that of the background, the signal-to-noise ratio of visible images is low, or local regions are disturbed by strong light, the model is prone to missed detections, false detections, and localization deviations. Therefore, RGB-IR detection research under complex illumination conditions has begun to emphasize the balance between lightweight structures and robust low-light perception [4]. In recent years, studies on infrared and visible-light cooperative detection have further shown that relying solely on a single visible-light modality makes it difficult to stably cover

complex perception conditions in open-road environments, while introducing infrared or thermal imaging modalities can improve the robustness of target perception under low illumination, occlusion, and complex backgrounds [5].

Infrared (IR) imaging provides an information source complementary to visible-light vision for autonomous driving perception by sensing the thermal radiation differences between targets and the environment. Related studies on thermal infrared object detection have also demonstrated the application value of thermal imaging in low-light and nighttime perception. Compared with RGB images, infrared images are less dependent on ambient illumination and can highlight pedestrians, vehicle engine regions, and other targets with obvious thermal responses under nighttime and low-light conditions. Therefore, they are of great value for all-weather road object detection [6]. Meanwhile, infrared images also suffer from insufficient texture details, unclear edge structures, and similar responses between targets and backgrounds in thermal crossover scenarios. Spatial–frequency fusion studies have also pointed out that infrared and visible images have obvious differences in structural details and frequency responses [7]. Therefore, RGB and IR are not simply redundant sensors, but have differentiated advantages in color and texture, contour structure, low-light adaptability, and target saliency [8]. How to reasonably utilize the information from the two modalities in different scenarios is the key to improving the robustness of autonomous driving object detection. An implicit cross-attention framework has also shown that cross-modal relationship modeling can effectively enhance complementary feature representation in RGB-T object detection [9]. Related studies have verified the effectiveness of RGB-IR fusion from the perspectives of content-guided fusion and self-calibration, while also pointing out that modality differences caused by different imaging mechanisms may limit the effectiveness of direct fusion [10]. The cascade alignment-guided Transformer further indicates that simply enhancing feature interaction is not sufficient in RGB-IR detection; it is also necessary to address spatial misalignment and semantic inconsistency between modalities [11].

For RGB-IR object detection, existing studies have developed relatively diverse fusion paradigms. Infrared–visible object detection methods usually combine the information of the two modalities at the input level, feature level, or decision level [12]. Input-level fusion directly combines the two modalities at the image or channel dimension. This strategy has a simple structure and can preserve relatively complete spatial details at the front end of the network. However, due to the differences between RGB and IR in imaging mechanisms, grayscale distributions, and noise characteristics, such methods often struggle to achieve stable alignment in shallow spatial representations [13]. Decision-level fusion performs confidence weighting or ensemble integration after each modality independently completes prediction. It can preserve the independent discriminative capability of each single-modal branch, but since the two modalities have already generated predictions separately before deep interaction, the fusion process usually has difficulty utilizing fine-grained complementary information in intermediate features, such as edges, textures, thermal responses, and local structures [14]. In contrast, feature-level fusion uses a dual-branch network to encode RGB and IR separately and performs cross-modal interaction at shallow, middle, or deep layers. This enables the model to exploit complementary cues at different semantic levels while preserving modality-specific characteristics, and has therefore gradually become a more commonly used technical route in current RGB-IR detection [15]. Some RGB-IR detection methods based on multi-receptive-field attention and frozen backbones also show that designing more effective interaction structures at the feature level can help improve dual-modal detection performance [16].

Although multimodal fusion detection can effectively improve detection performance in complex road environments, existing methods still have several issues in their fusion mechanisms that deserve

further investigation [17]. First, the reliability of RGB and IR is not fixed. Under sufficient daytime illumination, RGB usually provides richer texture, color, and road structure information, whereas in nighttime or low-light environments, IR often provides more stable target contours and thermal responses [18]. However, in cases of thermal crossover, similar background temperature, or strong infrared noise, the IR modality may also introduce interference. Therefore, simple concatenation, addition, or fixed-weight fusion of the two modalities can easily transfer noise from unreliable modalities into the fused features [19]. Second, RGB-IR fused features contain shared semantic information, complementary differential information, and redundant interference information at the same time, while coarse-grained fusion methods struggle to explicitly distinguish target-related cues from ineffective background responses [20]. Some RGB-IR fusion studies following the strategy of "remove first and then select" have also shown that ineffective information should be suppressed first during fusion, after which valuable cross-modal cues can be further selected [13]. Third, targets in autonomous driving scenarios show obvious scale differences. Distant pedestrians, cyclists, and traffic signs often appear as small targets and rely more on edges, textures, and thermal responses in shallow features, while medium- and large-scale targets require local structures, category semantics, and contextual information to be considered simultaneously [21]. If the same fusion strategy is adopted for features at different scales, it may easily lead to the loss of small-target details or insufficient high-level semantic representation. Studies on multi-scale dual-attention fusion have also shown that cross-layer feature interaction plays an important role in infrared–visible object detection [12]. In addition, RGB-T detection studies in complex scenarios, such as UAV and air–ground cooperative environments, further indicate that modal temporal misalignment, viewpoint changes, and target scale variations can significantly increase the difficulty of multimodal fusion [22]. Therefore, RGB-IR detection models for autonomous driving scenarios need to consider modality reliability, complementary information selection, scale differences, and computational complexity simultaneously, so as to achieve more stable and deployable perception capability in complex environments [6].

Based on the above analysis, this paper proposes a reliability–complementarity–scale-aware RGB-IR object detection network for autonomous driving scenarios. The proposed method takes RGB images and infrared (IR) images as dual-modal inputs, extracts multi-scale features of the two modalities separately through a dual-branch backbone network, and conducts targeted modeling for three issues: dynamic changes in modality reliability, the mixture of complementary information and redundant information, and differences in fusion requirements for multi-scale targets. The main contributions of this paper are summarized as follows:

(1) A Reliability-Guided Modulation (RGM) module is proposed. It estimates the reliability of RGB and IR features from three aspects: channel stability, spatial saliency, and cross-modal consistency. A modality-competitive normalization mechanism is further used to dynamically adjust the contributions of the two modalities, alleviating the problem that fixed fusion strategies tend to introduce noise under complex illumination and unstable thermal response conditions.

(2) A Complementary Information Decoupling (CID) module is proposed. It explicitly models shared semantic information, complementary differential information, and redundant interference information in RGB-IR features, enabling the network to selectively preserve target-related information, enhance cross-modal complementary cues, and suppress ineffective background responses during the fusion process, thereby improving the discriminability and interpretability of the fused features.

(3) A Scale-Aware Fusion Pyramid Network (SAFPN) is proposed. To address the significant scale variations of targets in autonomous driving scenarios, detail preservation, structural balance, and global

context enhancement strategies are designed for shallow, middle, and deep layers, respectively. Adaptive fusion of multi-scale features is achieved through top-down semantic propagation and bottom-up detail feedback, improving the model's detection capability for distant small targets, occluded targets, and complex road scenes.

2 Related work

2.1 Modeling Modality Reliability for Adaptive Fusion

One of the key issues in RGB-IR object detection is that the reliability of different modalities changes with scene conditions. Visible-light images usually contain richer color, texture, and edge information in daytime and structurally clear scenes, while infrared images can provide more stable thermal responses under nighttime, low-light, and severe occlusion conditions. To address this issue, recent studies have gradually shifted from static fusion to dynamic fusion and task-coordinated optimization. Zhang et al. [23] proposed E2E-MFD, which integrates multimodal image fusion and object detection into an end-to-end framework, enabling fused feature learning to directly serve the detection task and alleviating the inconsistency of optimization objectives in the traditional "fusion first, detection later" pipeline. Zeng et al. [24] proposed MMI-Det, which analyzes the feature complementarity in multimodal detection from the perspective of visible and infrared information integration, emphasizing that the fusion module should consider both cross-modal information interaction and detection task requirements. Guo et al. [25] proposed DAMSDet, which enables the detector to dynamically select more effective modality representations according to target regions through modality-competitive query selection and adaptive feature fusion. Hu et al. [26] revisited the self-attention mechanism in multispectral detection, showing that attention structures in cross-modal fusion should not only enhance interaction, but also control computational complexity and redundant responses. Zhang et al. [27] further reconsidered early fusion strategies and pointed out that simple input-level fusion is easily affected by modality differences, but may still retain strong information integration capability when properly designed. Hu et al. [28] proposed EI²Det, which introduces edge-guided and illumination-aware interaction into visible-infrared detection, enabling the model to adjust the contributions of dual-modal features according to illumination conditions. Jang et al. [29] improved detection robustness under different imaging conditions from a condition-adaptive perspective by using the conversion relationship between visible and thermal infrared images. Zhao et al. [30] revisited multimodal detection from the perspective of single-modal feature learning, pointing out that multimodal fusion should not weaken the discriminative capability of each single-modal branch itself.

2.2 Complementary Information Modeling and Cross-Modal Fusion for RGB-IR Features

In addition to changes in modality reliability, RGB-IR object detection also needs to address the mixture of shared semantics, complementary differences, and redundant interference. RGB images can provide texture, color, and spatial structure information, while infrared images are more likely to highlight the contours of thermal targets. However, noise, misalignment, and background interference also exist between the two modalities. Shen et al. [31] proposed ICAFusion, which models the global relationship between RGB and IR through iterative cross-attention to enhance the representation of cross-modal complementary information. Zhang et al. [32] proposed TFDet, which highlights pedestrian-region features from the perspective of target-aware fusion and reduces the interference of background regions during the fusion process. Li et al. [33] proposed M2FNet, which uses a mask-guided multi-level fusion mechanism to enable the network to focus on target-related regions at different feature layers and explore both consistent and differential information.Yuan et al. [34] proposed C²Former, which models inter-modal relationships through a calibration and complementary Transformer, establishing a

connection between cross-modal alignment and complementary enhancement. Zhou et al. [35] proposed COFNet, which introduces object box-level masks and contrastive learning into multispectral detection to strengthen the cross-modal consistency and discriminability of target regions. Dong et al. [36] proposed Fusion-Mamba, which uses a state-space modeling mechanism to associate cross-modal features and alleviates pseudo responses caused by modality differences through a gating strategy. Wu et al. [37] further introduced textual semantics into RGB-IR detection and connected category semantics with dual-modal visual features through consistency and difference modeling, indicating that shared information and complementary information should be treated differently.

2.3 Multi-Scale Target-Aware Efficient Fusion Detection

Target scale variations are very obvious in autonomous driving, UAV, and remote sensing scenarios. Distant pedestrians, cyclists, traffic signs, and small obstacles usually occupy only a few pixel regions, and therefore rely more on shallow edges, textures, and local thermal responses. Medium- and large-scale targets require more structural semantics and contextual information. Therefore, RGB-IR detection needs to address not only modality fusion, but also the differentiated fusion requirements of features at different scales. Zhu et al. [38] proposed WaveMamba, which combines wavelet-based frequency-domain decomposition with the Mamba structure to enhance high-frequency details and long-range dependency modeling in RGB-IR detection. Wang et al. [39] proposed an RGB-T detection method based on multi-receptive-field attention and multimodal supervision, enhancing target representation at different levels through multi-scale receptive fields. Wan et al. [40] constructed the YOLOv11-RGBT framework and systematically analyzed fusion positions, fusion strategies, and modality weight allocation in single-stage multispectral detection, providing a reference for the design of real-time detection structures. Ying et al. [41] constructed the RGBT-Tiny small object detection benchmark and pointed out that existing RGBT datasets are still insufficient for fully evaluating small object detection capability in terms of target scale, category diversity, and registration quality. Peng et al. [42] proposed COXNet, which improves RGBT tiny object detection performance through cross-layer fusion, dynamic alignment, and scale integration, indicating that small object detection needs to consider both cross-modal misalignment and scale adaptation. Lv et al. [43] proposed DGE-YOLO, which designs dual-branch feature extraction and an efficient multi-scale attention mechanism for UAV detection tasks, balancing detection accuracy and computational efficiency under complex backgrounds. Yan et al. [44] constructed MFE-DETR based on the DETR framework and improved object detection capability in aerial RGB-IR scenarios through multimodal feature enhancement. Gu et al. [45] introduced feature enhancement and dynamic gated fusion for UAV multimodal detection, further demonstrating that complex aerial scenarios require adaptive adjustment of fusion ratios according to target scales and modality states.

2.4 Motivation

Overall, the above studies show that RGB-IR object detection, dynamic modality fusion, and multi-scale feature modeling have made considerable progress in complex-environment perception. However, for open-road autonomous driving scenarios, existing methods still have certain limitations in modality reliability estimation, complementary information utilization, and scale-adaptive fusion.

(1) Existing RGB-IR fusion methods often adopt concatenation, addition, attention weighting, or fixed-structure fusion. Although these methods can enhance dual-modal information interaction, they insufficiently consider the dynamic changes in the reliability of RGB and IR under different scenarios. Under strong light, low illumination, thermal crossover, or strong infrared noise, the contributions of different modalities to the detection task are not consistent. Without an explicit reliability estimation

mechanism, the fusion process may easily introduce noise from unreliable modalities into detection features, thereby affecting the stability of the model.

(2) Most existing cross-modal fusion methods mainly focus on enhancing the interaction between RGB and IR, but lack further distinction of the internal information components in fused features. RGB-IR features simultaneously contain target semantics jointly supported by the two modalities, differential information with complementary value, and redundant interference caused by background, misalignment, and modality conflicts. If these types of information are not explicitly decoupled, the model may amplify ineffective responses while enhancing complementary cues, thereby reducing the discriminability and interpretability of the fused features.

(3) Target scales vary significantly in autonomous driving scenarios, and targets at different scales have different dependencies on dual-modal information. Distant pedestrians, cyclists, and traffic signs rely more on shallow details and thermal response variations. Medium-scale targets require both local structures and category semantics, while large-scale targets depend more on global contextual information. Existing methods usually adopt similar fusion strategies for features at different levels, making it difficult to fully adapt to the perception requirements of targets at different scales. This may lead to the loss of small-target details or insufficient high-level semantic representation.

Based on this, this paper proposes a reliability-complementarity-scale-aware RGB-IR object detection network for autonomous driving scenarios. The proposed method conducts modeling from three aspects, namely dynamic modulation of modality reliability, decoupled fusion of complementary information, and multi-scale adaptive feature fusion, so as to improve the detection robustness of the model under complex illumination, modality degradation, occlusion interference, and scale variations.

3 Method

3.1 Overall structure

The overall architecture of the proposed RGB-IR multispectral object detection network is shown in Fig. 1. The network takes registered RGB images and infrared (IR) images as dual-modal inputs, and adopts a dual-branch feature extraction structure to obtain multi-scale features from the two modalities separately. Specifically, the RGB image and the IR image are fed into the RGB Backbone and the IR Backbone for feature encoding, respectively. The two branches maintain the same structure and share feature extraction weights to enhance the consistency of dual-modal features in the semantic space. After passing through the backbone network, RGB features and IR features at three scales, namely P3, P4, and P5, are obtained. Among them, P3 retains more spatial detail information and is suitable for representing small targets and boundary structures, P4 contains both local structural and semantic information, and P5 contains stronger high-level semantic and contextual information.

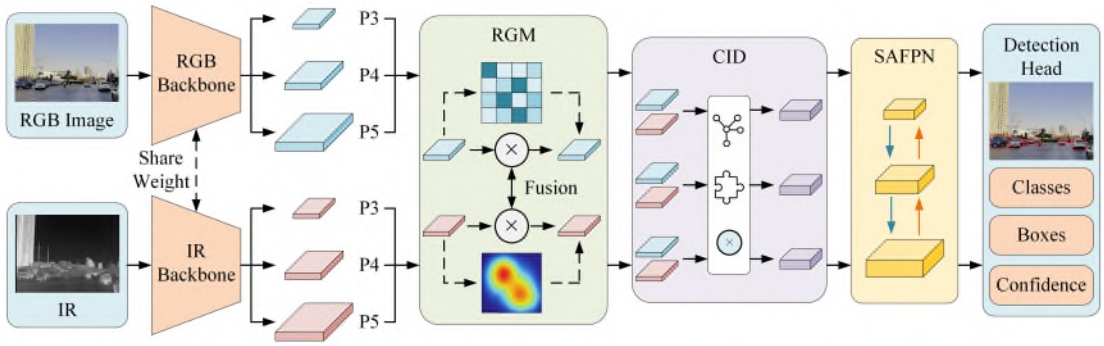


Fig. 1 Overall architecture of the FSMF

After obtaining the dual-modal multi-scale features, the network sequentially introduces the

Reliability-Guided Modulation (RGM) module, the Complementary Information Decoupling (CID) module, and the Scale-Aware Fusion Pyramid Network (SAFPN) module. The RGM module is used to estimate the reliability of the RGB and IR modalities at different channels and spatial positions before feature fusion, and adaptively modulates the features of the two modalities to reduce the influence of low-quality modality responses on the fusion results. Subsequently, the modulated dual-modal features are fed into the CID module to model shared semantic information, complementary differential information, and redundant interference information, thereby obtaining more discriminative fused features. On this basis, SAFPN further performs multi-scale enhancement on the fused features, improving the network's representation capability for targets at different scales through shallow-level detail preservation, middle-level structural balance, and deep-level contextual modeling. Finally, the scale-enhanced features are fed into the detection head to complete target category prediction, bounding box regression, and confidence estimation. Through the above design, the proposed network can achieve more stable RGB-IR object detection under complex illumination, modality degradation, occlusion interference, and scale variations.

3.2 Reliability-Guided Modulation Module

In autonomous driving environments, the effectiveness of RGB and infrared modalities varies significantly with scene conditions. In daytime or well-illuminated conditions, RGB images usually contain richer texture, color, and road structure information. Under nighttime, low-light, or glare conditions, infrared images can provide more stable target contours and thermal responses. However, the infrared modality is not always reliable. When the background temperature is close to the target temperature, thermal crossover occurs, or sensor noise is strong, infrared features may introduce ineffective responses. Based on this observation, this paper designs a Reliability-Guided Modulation (RGM) module, enabling the network to first determine the reliability of different modalities at the current scale and in the current region before fusion. The structure is shown in Fig. 2.

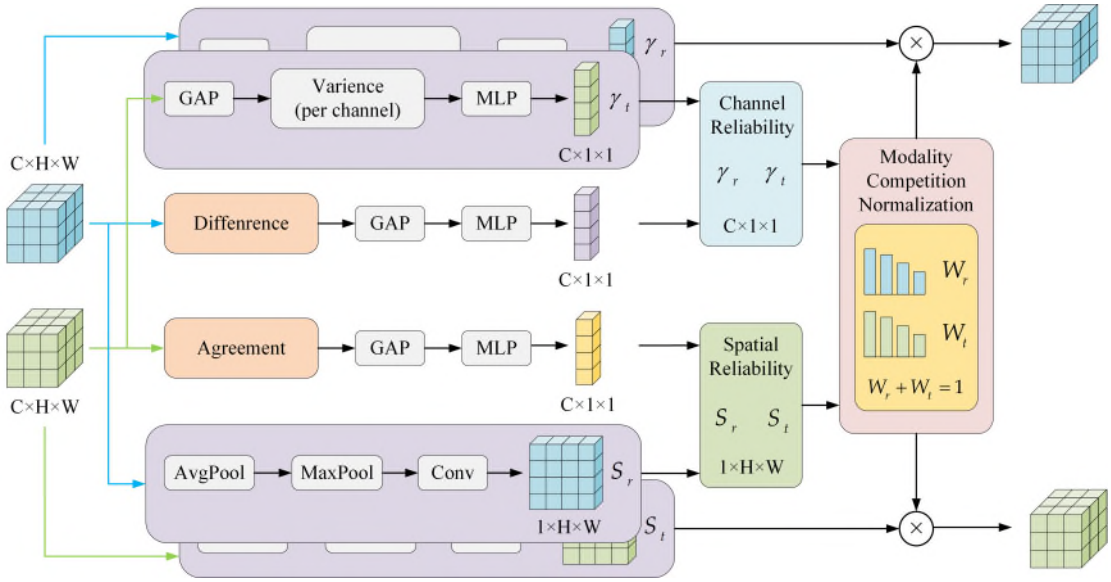


Fig. 2 Architecture of RGM

RGM does not generate a modality weight solely based on global pooling. Instead, it estimates reliability from three aspects: channel stability, spatial saliency, and cross-modal consistency. For the feature F_l^m at the l-th layer, where $m \in \{r, t\}$, we first calculates its channel response descriptor as follows:

$$u_l^m = GAP\left(\left|F_l^m\right|\right) \tag{1}$$

$$v_l^m = GAP\left(\left(F_l^m - \mu\left(F_l^m\right)\right)^2\right) \tag{2}$$

where GAP denotes global average pooling, u_l^m reflects the overall channel response intensity, and v_l^m reflects the stability of the channel response. For the detection task, channels with stable and strong responses usually contain more reliable target information, whereas channels with high fluctuations, weak responses, or severe conflicts with the other modality are more likely to be associated with noise or background interference.

To further characterize the relationship between the two modalities, this paper introduces a cross-modal difference descriptor and a consistency descriptor:

$$D_l = \left|F_l^r - F_l^t\right| \tag{3}$$

$$A_l = F_l^r \square F_l^t \tag{4}$$

where D_l is used to characterize the differential information between RGB and infrared features, and A_l is used to represent the regions jointly activated by the two modalities. In general, jointly activated regions are more likely to correspond to stable targets or important structures, whereas regions with large differences may contain either complementary information or modality conflicts and background noise. Therefore, using only differential features is not sufficient; it is also necessary to make judgments by considering the intrinsic quality of each modality and the spatial foreground response.

At the channel dimension, RGM generates channel reliability weights through a lightweight perception function:

$$\gamma_l^m = Sigmoid\left(MLP\left(Concat\left(u_l^m, v_l^m, GAP\left(D_d\right), GAP\left(A_l\right)\right)\right)\right) \tag{5}$$

where $\gamma_l^m \in R^{C\times 1\times 1}$ denotes the channel reliability of modality m at the l-th layer. At the spatial dimension, RGM further constructs a spatial reliability map based on the average response, maximum response, and cross-modal difference:

$$s_l^m = Sigmoid\left(Conv\left(Concat\left(Avg_c\left(F_l^m\right), Max_c\left(F_l^m\right), Avg_c\left(D_l\right), Avg_c\left(A_l\right)\right)\right)\right) \tag{6}$$

where Avg_c and Max_c denote average pooling and max pooling along the channel dimension, respectively. s_l^m represents the reliability of modality m at different spatial positions. Subsequently, the channel reliability and spatial reliability are jointly used to form the modality reliability tensor:

$$R_l^m = \gamma_l^m \otimes s_l^m \tag{7}$$

To prevent the RGB and infrared branches from being enhanced without constraint at the same time, this paper adopts a modality-competitive normalization strategy, so that the contributions of the two modalities at the same scale remain relatively constrained:

$$W_l^r, W_l^t = Soft\ max\left(\left[R_l^r, R_l^t\right]\right) \tag{8}$$

where $W_l^r + W_l^t = 1$. This design enables the network to adaptively emphasize the more reliable modality under different scene conditions, while avoiding the simultaneous amplification of noise from ineffective modalities.

After modulation by the reliability weights, the features of the two modalities are updated as follows:

$$\tilde{F}_l^t = F_l^t \square \left(1 + W_l^t\right) \tag{9}$$

$$\tilde{F}_l^r = F_l^r \square \left(1 + W_l^r\right) \tag{10}$$

3.3 Complementary Information Decoupling Module

After RGM modulation, the RGB and infrared features already contain preliminary reliability differences. However, if they are directly added or concatenated, the fused features will still mix shared semantics, complementary cues, and redundant interference. In fact, not all information in the dual-modal features is beneficial for detection. The components jointly responded to by RGB and infrared features usually contain relatively stable target semantics. The components that differ significantly between the two modalities but are related to the target may contain complementary information, such as RGB textures and infrared contours. In contrast, the components that differ significantly but are irrelevant to the target may be redundant interference caused by illumination noise, thermal backgrounds, sensor bias, or registration errors. The structure is shown in Fig. 3.

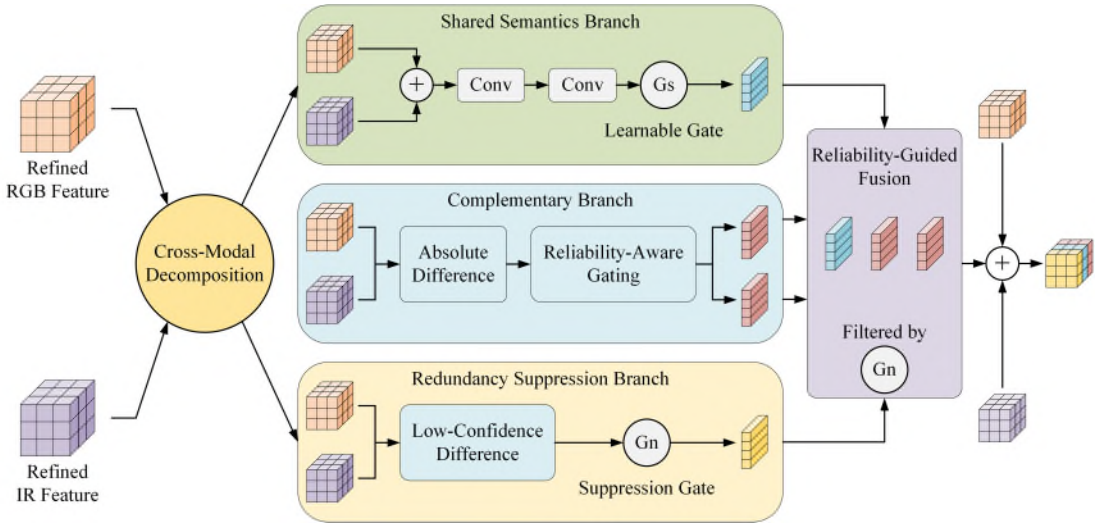


Fig. 3 Architecture of CID

To this end, this paper proposes a Complementary Information Decoupling (CID) module, which explicitly constructs a shared semantic branch, a complementary difference branch, and a redundancy suppression branch during the fusion stage, enabling the network to selectively fuse effective information.

Given the RGM-modulated features $\tilde{F}_l^t$ and $\tilde{F}_l^r$, the shared semantic candidate feature is first constructed as follows:

$$S_l^0 = Conv_s\left(\tilde{F}_l^t + \tilde{F}_l^r\right) \tag{11}$$

At the same time, the complementary difference candidate feature is constructed as follows:

$$C_l^0 = Conv_c\left(\left|\tilde{F}_l^t - \tilde{F}_l^r\right|\right) \tag{12}$$

where S_l^0 emphasizes the information jointly responded to by the two modalities, and C_l^0 emphasizes the differences between the two modalities. To avoid misidentifying ineffective differences as complementary information, CID further constructs a shared gate and a complementary gate:

$$G_l^s = Sigmoid\left(Conv\left(Concat\left(S_l^0, \tilde{F}_l^t \square \tilde{F}_l^r\right)\right)\right) \tag{13}$$

$$G_l^c = Sigmoid\left(Conv\left(Concat\left(C_l^0, S_l^0, W_l^r, W_l^t\right)\right)\right) \tag{14}$$

where G_l^s denotes the confidence of shared semantics, and G_l^c denotes the confidence of complementary differences. The shared semantics are determined by the joint activation degree of the two modalities, while the complementary information depends not only on modality differences but also on the constraint of reliability weights. This can reduce the possibility of enhancing meaningless differences in background regions. The shared semantic feature is represented as follows:

$$S_l = G_l^s \square S_l^0 \tag{15}$$

The complementary information is further divided into an RGB-dominated complementary component and an infrared-dominated complementary component:

$$C_l^r = G_l^c \square W_l^r \square Conv_r\left(\tilde{F}_l^r - \tilde{F}_l^t\right) \tag{16}$$

$$C_l^t = G_l^c \square W_l^t \square Conv_r\left(\tilde{F}_l^t - \tilde{F}_l^r\right) \tag{17}$$

where C_l^r mainly preserves the texture, color, and local structural information in the RGB modality that is more valuable than that in the infrared modality, while C_l^t mainly preserves the thermal responses, contours, and low-light target information in the infrared modality that is more valuable than that in the RGB modality. On this basis, CID constructs a redundant interference estimation branch:

$$N_l = Conv_n\left(Concat\left(C_l^0, 1 - G_l^s, 1 - G_l^c\right)\right) \tag{18}$$

where N_l is used to estimate responses that belong neither to stable shared semantics nor to reliable complementary differences. This part usually corresponds to background thermal noise, illumination disturbances, modality conflicts, or ineffective edges. To reduce the influence of redundant information on the fusion results, this paper designs a redundancy suppression gate:

$$G_l^n = 1 - Sig\,mo\,id\left(Conv\left(N_l\right)\right) \tag{19}$$

The final fused feature is represented as follows:

$$H_l = Conv_f\left(Concat\left(S_l, C_l^r, C_l^t\right) \square G_l^n\right) + Conv_{res}\left(\tilde{F}_l^r + \tilde{F}_l^t\right) \tag{20}$$

where the first term denotes the effective fused feature after shared–complementary modeling, and the second term is a residual term used to maintain the stable transmission of the original dual-modal information. To further enhance the decoupling effect, a lightweight structural constraint can be introduced during training to keep the shared semantic component and the complementary difference component relatively independent in the feature space:

$$L_{dec} = \frac{1}{L}\sum_{l=3}^{5} Frobenius\left(Norm\left(S_l\right)^T Norm\left(C_l^r + C_l^t\right)\right)_F^2 \tag{21}$$

where $Norm$ denotes feature normalization. This constraint encourages the shared semantics and complementary differences to represent different types of information, thereby avoiding their mutual entanglement during the fusion process.

3.4 Scale-Aware Fusion Pyramid Network

Autonomous driving object detection involves obvious scale variations. Distant pedestrians, cyclists, traffic signs, and small vehicles usually occupy only a few pixels, and therefore rely more on edges, contours, and local details in shallow features. Medium-scale targets require both local structures and category semantics. Large vehicles or close-range targets depend more on deep global contextual information. If the same fusion strategy is adopted for different scales, the details of small targets may be easily covered by deep semantic information, or large targets may lack sufficient global modeling. The structure is shown in Fig. 4.

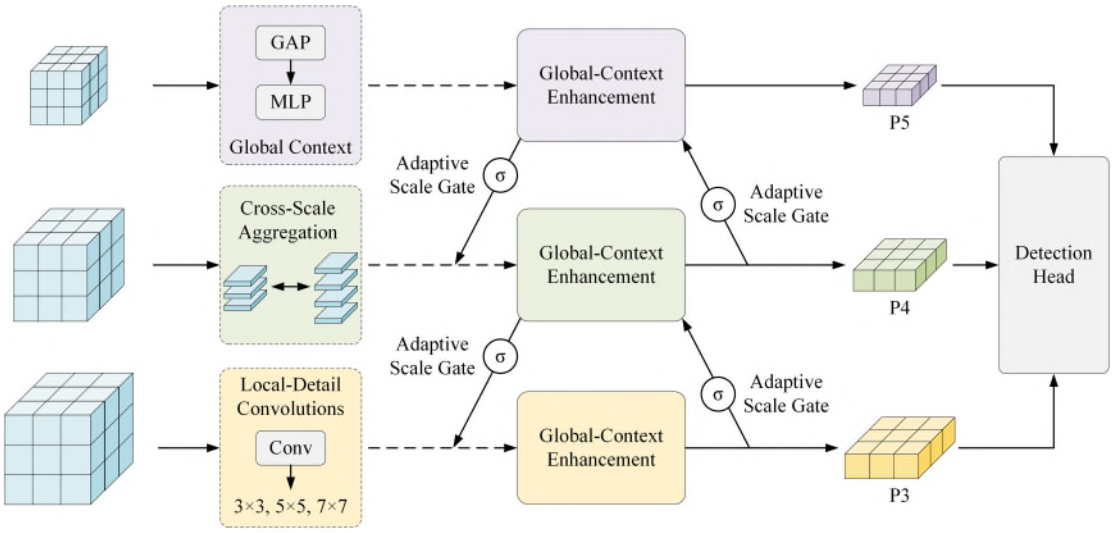


Fig. 4 Architecture of SAFPN

To this end, this paper proposes a Scale-Aware Fusion Pyramid Network (SAFPN) module. This module does not simply follow the traditional FPN/PAN structure. Instead, it designs differentiated enhancement strategies according to the feature properties of the three scales, P3, P4, and P5, so that features at different levels undertake different fusion tasks. For the shallow feature H3, which has a higher spatial resolution and is suitable for representing small targets and target edges, this paper designs a detail-preserving enhancement branch:

$$E_3 = DWConv_{3\times3}\left(H_3\right) + DWConv_{5\times5}\left(H_3\right) \tag{22}$$

$$G_3^d = Sigmoid\left(Conv\left(Concat\left(H_3, E_3\right)\right)\right) \tag{23}$$

$$D_3 = H_3 + G_3^d \square E_3 \tag{24}$$

where $DWConv$ denotes depthwise separable convolution, E_3 is used to extract local edge and texture structures, and G_3^d is used to control the intensity of detail enhancement. This branch can strengthen small-target boundaries, infrared contours, and local RGB textures, preventing shallow fine-grained information from being weakened during the subsequent fusion process. For the middle-level feature H_4, it contains both certain spatial structures and relatively strong semantic information. This paper designs a structure-balanced enhancement branch, which jointly models information from the shallow and deep layers together with the current layer:

$$G_4^s = Sigmoid\left(Conv\left(Concat\left(Down\left(D_3\right), H_4, Up\left(H_5\right)\right)\right)\right) \tag{25}$$

$$D_4 = H_4 + G_4^s \odot Conv\left(Concat\left(Down\left(D_3\right), H_4, Up\left(H_5\right)\right)\right) \tag{26}$$

where Up and $Down$ denote the upsampling and downsampling operations, respectively. The role of this branch is to introduce shallow-level details and deep-level semantics into the middle layer, enabling medium-scale targets to obtain both structural and semantic constraints. For the deep-level feature H_5, it has a lower spatial resolution but stronger semantic representation and contextual perception capabilities. We design a global context enhancement branch:

$$G_5 = Sigmoid\left(MLP\left(GAP\left(H_5\right)\right)\right) \tag{27}$$

$$D_5 = H_5 + G_5 \odot Conv\left(Concat\left(H_5\right), Down\left(H_4\right)\right) \tag{28}$$

This branch enhances deep-level features through global context weights, enabling the network to better understand the relationship between targets and backgrounds in complex traffic scenarios. It is especially suitable for scenes with dense vehicles, occlusion, and significant illumination variations.

After completing differentiated enhancement at the three scales, SAFPN further introduces a bidirectional scale interaction path. The top-down path transfers deep semantic information to shallow features, allowing small-target features to obtain stronger category semantics. The bottom-up path feeds shallow details back to high-level features, enabling deep features to retain more boundary and local structural information. To avoid indiscriminate information transmission across different scales, this paper introduces scale gating during cross-layer interaction, so that the network can selectively receive information from adjacent scales according to the requirements of the current level. The bidirectional interaction process can be simplified as follows:

$$T_i = D_i + \alpha_i \odot Up\left(T_{i+1}\right), i = 4,3 \tag{29}$$

$$P_i = T_i + \beta_i \odot Down\left(P_{i-1}\right), i = 4,5 \tag{30}$$

where T_i denotes the intermediate feature after top-down semantic propagation, P_i denotes the final scale-enhanced output feature, and α_i and β_i denote the scale gating weights for the top-down and bottom-up paths, respectively. Finally, SAFPN outputs $\left\{P_3, P_4, P_5\right\}$, which is then fed into the

detection head to complete classification and bounding box regression.

Compared with traditional FPN or PAN structures, SAFPN is characterized by the fact that it does not adopt the same fusion logic for all scales. Instead, it is specifically designed according to the functional differences of features at different levels. The shallow layer focuses on detail preservation, the middle layer focuses on balancing structure and semantics, and the deep layer focuses on global context modeling. Meanwhile, the bidirectional gated interaction makes cross-scale information transmission more controllable, which can reduce the coverage of shallow details by high-level semantics and also avoid the interference of low-level noise with deep semantic information. Therefore, SAFPN can improve the model's adaptability to distant small targets, occluded targets, and complex road backgrounds while maintaining multi-scale representation capability.

4 Experiments

4.1 Datasets

M3FD: The M3FD dataset was proposed by Liu et al. in 2022 for RGB-IR image fusion and multi-modal object detection [46]. It contains 4200 aligned visible-infrared image pairs, totaling 8400 images, most of which have a resolution of 1024×768. The dataset provides 34,407 bounding box annotations for six categories, including People, Car, Bus, Motorcycle, Lamp, and Truck. Its scenes cover various imaging conditions, such as day, night, overcast, foggy, tunnel, coastal, woodland, grassland, and other challenging environments. Since M3FD does not provide an official split, we divide it into training, validation, and test sets at a ratio of 8:1:1.

FLIR: The Aligned FLIR dataset is derived from the original FLIR ADAS dataset by Zhang et al [47]. To address the misalignment between visible and thermal images in the original dataset, the authors manually aligned the RGB-IR image pairs. The processed dataset includes 4128 training pairs and 1013 validation pairs, and the commonly used detection categories are person, bicycle, and car. Owing to its improved cross-modal alignment, this dataset is widely used for evaluating RGB-IR object detection and feature fusion methods.

4.2 Evaluation Metrics

The main evaluation metrics used in this study are mean average precision (mAP), mAP@50, and mAP@75. Specifically, mAP@50 measures the average accuracy when the IoU threshold is set to 0.5, while mAP@75 uses a stricter threshold to better reflect the accuracy of bounding box localization. mAP calculates the average AP value for IoU thresholds ranging from 0.5 to 0.95 (in 0.05 increments), thereby providing a more rigorous and comprehensive evaluation of the model. Together, these metrics reflect the detection accuracy and localization quality of the proposed method. Their respective formulas are as follows:

$$Precision = \frac{TP}{(TP + FP)} \tag{31}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{32}$$

$$AP = \int_0^1 P(R)dR \tag{33}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{34}$$

where TP denotes true positives, indicating correctly detected objects; FP denotes false positives, indicating background regions or incorrect categories that are falsely predicted as objects; and FN denotes false negatives, indicating ground-truth objects that are missed by the detector.

4.3 Implementation Details

All experiments are conducted under a unified training configuration. The proposed method is trained on a workstation equipped with an NVIDIA GeForce RTX 2080Ti GPU. During training, the input images are uniformly resized to 640×640, the batch size is set to 8, and the model is trained for 300 epochs. The number of data loading workers is set to 2. Adam is adopted as the optimizer, with an initial learning rate of 0.001, and the random seed is fixed at 2026 to ensure experimental reproducibility. In addition, GFLOPs are calculated with an input size of 640×640. FPS is measured on the same NVIDIA GeForce RTX 2080Ti GPU with a batch size of 1 after model warm-up.

4.4 Comparison Experiment

4.4.1. Results on M3FD

Table 1 summarizes the quantitative experimental results of different methods on the M3FD dataset. In terms of overall detection accuracy, FSMF-s achieves the best results in Recall and mAP@50, reaching 79.9% and 85.9%, respectively. Compared with YOLOv11s, the Recall of FSMF-s increases from 78.7% to 79.9%, and mAP@50 increases from 83.9% to 85.9%. Compared with YOLOv8s, its Recall increases from 77.5% to 79.9%, and mAP@50 increases from 83.1% to 85.9%. For the Precision metric, FSMF-s achieves 86.4%, which is slightly lower than the 86.8% of YOLOv8s, but higher than those of YOLOv11s, EMOD, GEM-YOLO, and FSMF-n. For the mAP@50:95 metric, FSMF-s achieves 55.3%, which is higher than the 55.0% of YOLOv11s, 53.9% of EMOD, 50.6% of GEM-YOLO, and 53.8% of FSMF-n. This indicates that the proposed method still has good overall detection performance under stricter localization evaluation conditions.

Table 1 Comparison of experimental results in terms of detection accuracy on the M3FD.

Models	Precision	Recall	mAP@50	mAP@50:95	Parameters (M)
Yolov8n	83.8	70.7	77.6	51.2	3.0
Yolov8s	86.8	77.5	83.1	53.3	11.1
Yolov11n	84.2	66.6	76.1	49.2	2.6
Yolov11s	85.6	78.7	83.9	55.0	9.4
EMOD	84.9	77.9	83.0	53.9	16.6
GEM-YOLO	82.3	70.4	77.6	50.6	3.3
FSMF-n	84.7	69.9	79.7	53.8	3.0
FSMF-s	86.4	79.9	85.9	55.3	13.6

In terms of model scale, FSMF-s has 13.6M parameters, which is higher than YOLOv8s with 11.1M and YOLOv11s with 9.4M, but lower than EMOD with 16.6M. Meanwhile, compared with EMOD, FSMF-s improves mAP@50 from 83.0% to 85.9%, achieving better detection results with fewer parameters. The lightweight version FSMF-n has 3.0M parameters, the same as YOLOv8n and lower than GEM-YOLO with 3.3M. Its mAP@50 reaches 79.7%, which is higher than 77.6% of YOLOv8n, 76.1% of YOLOv11n, and 77.6% of GEM-YOLO. Its mAP@50:95 also reaches 53.8%, outperforming the above lightweight comparison methods.

Overall, FSMF-s achieves the highest Recall and mAP@50 on the M3FD dataset, while FSMF-n shows certain lightweight advantages with a smaller parameter scale. In terms of Precision and Recall, FSMF-s achieves a Precision of 86.4%, which is close to 86.8% of YOLOv8s, while its Recall reaches 79.9%, the highest among all methods. The higher Recall indicates that the proposed method can detect more real targets and reduce missed detections. This result is consistent with the characteristics of the M3FD dataset, which contains complex scenes such as nighttime, foggy weather, tunnels, grasslands, and woodlands. It shows that the reliability-guided modulation module can dynamically adjust the

contributions of RGB and IR modalities according to scene differences, enabling the model to maintain more stable target responses under different illumination and background conditions. The qualitative detection results of multiple methods on M3FD are shown in Fig. 5.

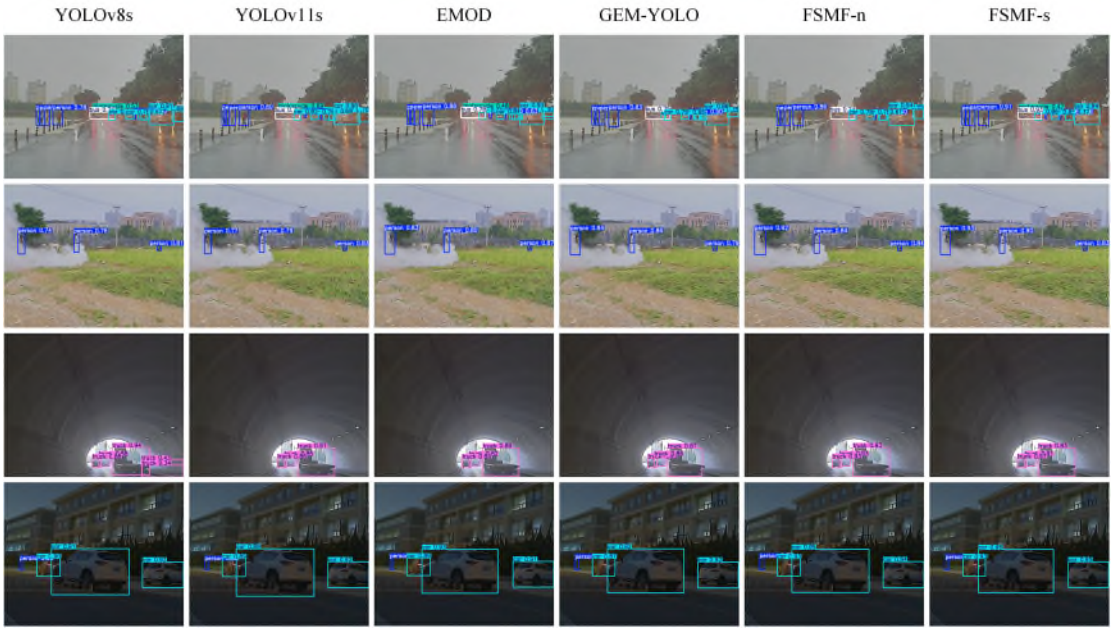


Fig. 5 Qualitative comparison of detection results of different methods on the M3FD dataset.

4.4.2 Results on FLIR

Table 2 summarizes the quantitative experimental results of different methods on the FLIR dataset. In terms of overall detection accuracy, FSMF-s achieves the best results in Precision, Recall, mAP@50, and mAP@50:95, reaching 77.3%, 65.1%, 72.2%, and 38.3%, respectively. Compared with YOLOv11s, the Precision of FSMF-s increases from 73.0% to 77.3%, Recall increases from 55.3% to 65.1%, mAP@50 increases from 61.5% to 72.2%, and mAP@50:95 increases from 30.0% to 38.3%. Compared with EMOD, the Precision of FSMF-s increases from 74.9% to 77.3%, Recall increases from 61.5% to 65.1%, mAP@50 increases from 69.5% to 72.2%, and mAP@50:95 increases from 36.5% to 38.3%. These results indicate that FSMF-s has good overall detection performance on the FLIR dataset.

Table 2 Comparison of experimental results in terms of detection accuracy on the FLIR.

Models	Precision	Recall	mAP@50	mAP@50:95	Parameters (M)
Yolov8n	68.9	51.2	57.8	27.8	3.0
Yolov8s	69.5	57.3	61.8	29.5	11.1
Yolov11n	69.0	54.2	60.0	29.1	2.6
Yolov11s	73.0	55.3	61.5	30.0	9.4
EMOD	74.9	61.5	69.5	36.5	11.8
GEM-YOLO	73.6	60.7	67.3	35.1	3.3
FSMF-n	68.5	61.9	67.3	35.2	3.0
FSMF-s	77.3	65.1	72.2	38.3	13.6

In terms of model scale, FSMF-s has 13.6M parameters, which is higher than YOLOv8s with 11.1M, YOLOv11s with 9.4M, and EMOD with 11.8M, but it achieves better detection accuracy across all metrics. The lightweight version FSMF-n has 3.0M parameters, the same as YOLOv8n and lower than GEM-YOLO with 3.3M. In terms of detection performance, FSMF-n achieves a Precision of 68.5%, a Recall of 61.9%, an mAP@50 of 67.3%, and an mAP@50:95 of 35.2%, among which Recall, mAP@50,

and mAP@50:95 are significantly higher than those of YOLOv8n and YOLOv11n. Compared with GEM-YOLO, FSMF-n maintains the same mAP@50 of 67.3%, while mAP@50:95 increases from 35.1% to 35.2%, indicating that FSMF-n still has certain competitiveness with a smaller parameter scale. Overall, FSMF-s achieves the best overall detection results on the FLIR dataset, while FSMF-n maintains good detection performance under lightweight conditions. The qualitative detection results of multiple methods on FLIR are shown in Fig. 6.

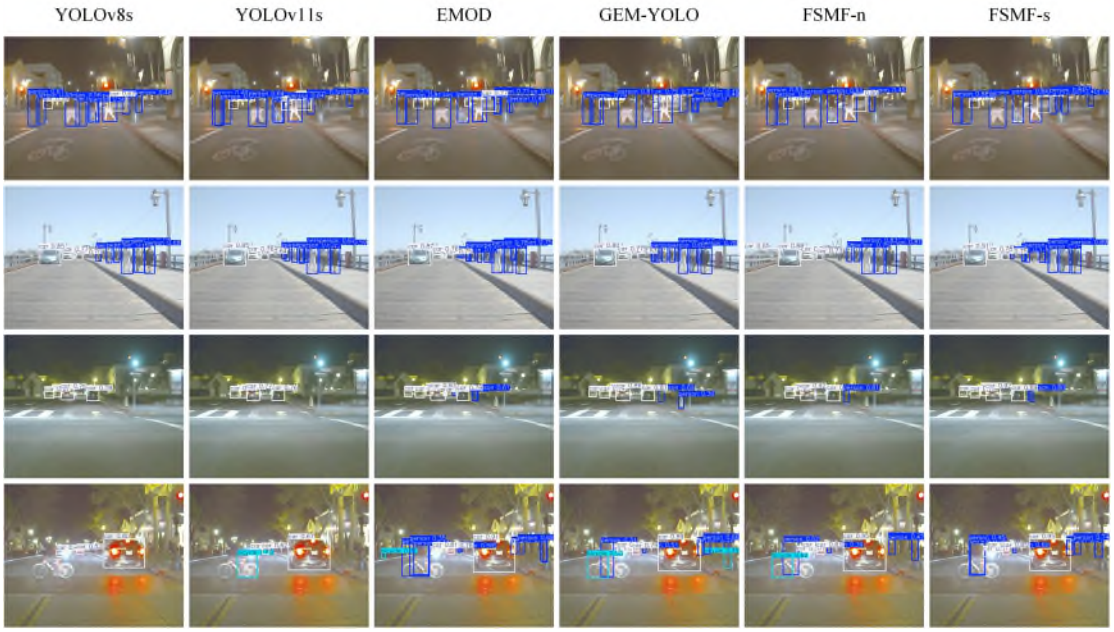


Fig. 6 Qualitative comparison of detection results of different methods on the FLIR dataset.

4.5 Ablation Experiment

Table 3 Comparison of experimental results in terms of detection accuracy on the FLIR.

RGM	CID	SAFPN	mAP@50	mAP@50:95	Parameters (M)
			62.4	30.7	6.0
√			64.8	32.1	6.8
	√		67.5	36.0	7.5
		√	67.5	34.8	11.3
√	√		68.9	35.6	8.8
√		√	70.8	37.4	11.6
	√	√	69.2	36.4	11.0
√	√	√	72.2	38.3	13.6

Table 3 presents the ablation experimental results of the RGM, CID, and SAFPN modules on the FLIR dataset. The baseline model without any of these modules is used as the reference, with mAP@50 and mAP@50:95 of 62.4% and 30.7%, respectively, and 6.0M parameters. First, after introducing RGM alone, mAP@50 increases from 62.4% to 64.8%, mAP@50:95 increases from 30.7% to 32.1%, and the number of parameters increases from 6.0M to 6.8M, indicating that RGM can bring certain detection performance gains with a small model overhead. After introducing CID alone, mAP@50 further increases to 67.5%, and mAP@50:95 reaches 36.0%, showing good overall performance under the single-module setting. When SAFPN is introduced alone, mAP@50 also reaches 67.5%, and mAP@50:95 is 34.8%, but the number of parameters increases to 11.3M, indicating that SAFPN can effectively improve detection accuracy, while also introducing relatively higher model complexity.

In the dual-module combination experiments, the combination of RGM and CID achieves an mAP@50 of 68.9% and an mAP@50:95 of 35.6%. After combining RGM with SAFPN, mAP@50 further increases to 70.8%, and mAP@50:95 reaches 37.4%, achieving the best performance among all dual-module combinations. In contrast, the combination of CID and SAFPN achieves an mAP@50 of 69.2%, which is higher than using either module alone but lower than the RGM and SAFPN combination. This indicates that different modules do not simply have an additive effect, and that a reasonable order of feature modulation and fusion plays an important role in performance improvement. Finally, after introducing RGM, CID, and SAFPN simultaneously, the model achieves the best results, with mAP@50 increasing from 62.4% to 72.2% and mAP@50:95 increasing from 30.7% to 38.3%, while the number of parameters increases to 13.6M. Overall, the three modules jointly and continuously improve the detection performance of the model on the FLIR dataset. The complete model achieves the best detection accuracy, verifying the effectiveness of the combined module design.

5 Conclusion

This paper proposes a reliability-guided and complementary information decoupling RGB-IR multispectral object detection network for autonomous driving scenarios. To address the problems of dynamically changing RGB and IR modality reliability, the mixture of complementary information and redundant interference, and the obvious differences in fusion requirements for targets at different scales in complex road environments, this paper designs a Reliability-Guided Modulation module, a Complementary Information Decoupling fusion module, and a Scale-Aware Fusion Pyramid Network module, respectively. By performing reliability estimation, complementary information selection, and multi-scale adaptive fusion on dual-modal features, the proposed method can obtain more stable object detection results under complex illumination, weak texture, occlusion interference, and scale variations.

Experimental results on the M3FD and FLIR datasets show that the proposed method achieves good detection performance across multiple evaluation metrics. Specifically, FSMF-s achieves an mAP@50 of 85.9% on the M3FD dataset, and mAP@50 and mAP@50:95 of 72.2% and 38.3% on the FLIR dataset, respectively, outperforming various mainstream detection models and existing RGB-IR detection methods. Ablation experiments further verify the effectiveness of the RGM, CID, and SAFPN modules. Compared with the baseline model, the complete model improves mAP@50 from 62.4% to 72.2% and mAP@50:95 from 30.7% to 38.3% on the FLIR dataset. Overall, the proposed method effectively improves the overall performance of RGB-IR object detection and achieves a good balance between detection accuracy and model complexity. Future work will further optimize the model structure, reduce the computational overhead introduced by dual-modal fusion, and verify the generalization ability and deployment performance of the method in more complex traffic scenarios and on edge computing platforms.

Funding: This research was funded by the Key Natural Science Research Project of Universities in Anhui Province, grant number 2024AH051438.

Acknowledgments: This research was funded by the Key Natural Science Research Project of Universities in Anhui Province, grant number 2024AH051438. We also thank the anonymous reviewers for providing critical comments and suggestions that improved the manuscript.

Data availability statement: The datasets used and analyzed in this study are publicly available third-party datasets. The M3FD dataset is available from the TarDAL project repository at

https://github.com/JinyuanLiu-CV/TarDAL. The aligned FLIR dataset used in this study is available from the CFR project repository at https://github.com/zhanghengdev/CFR. No new dataset was generated during the current study.

References

1. Li G, Wang Y, He B, et al (2025) Low-light multimodal object detection: A survey. Computer Science Review 58:100804. https://doi.org/10.1016/j.cosrev.2025.100804
2. Mudavath T, Niranjan V (2026) Seeing beyond the visible: A comprehensive review of thermal image enhancement and object detection techniques. Infrared Physics & Technology 156:106573. https://doi.org/10.1016/j.infrared.2026.106573
3. Xiong Z, Yao Z, Liu X, et al (2025) Efficient Multispectral Object Detection with attentive feature aggregation leveraging zero-shot implicit illumination guidance. Inf Fusion 118:. https://doi.org/10.1016/j.inffus.2025.102939
4. Weng T, Niu X (2026) LMDENet: A Lightweight RGB-IR Object Detection Network for Low-Light Remote Sensing Images. Sensors 26:. https://doi.org/10.3390/s26041130
5. Li Y, Zhan W, Jiang Y, Guo J (2025) RDCRNet: RGB-T Object Detection Network Based on Cross-Modal Representation Model. Entropy 27:. https://doi.org/10.3390/e27040442
6. Xue B, Yu Q, Ding K, et al (2026) WS-R-IR Adapter: A Multimodal RGB–Infrared Remote Sensing Framework for Water Surface Object Detection. Remote Sensing 18:. https://doi.org/10.3390/rs18081220
7. Liu J, Gao J, Liu X, et al (2025) Asymmetric Spatial–Frequency Fusion Network for Infrared and Visible Object Detection. Symmetry 17:. https://doi.org/10.3390/sym17122174
8. Guo X, Yang F, Ji L (2026) DEF-Net: A dual-modal feature enhancement and fusion network for infrared and visible object detection. PLOS ONE 21:e0345815. https://doi.org/10.1371/journal.pone.0345815
9. Liu Z, Zhu C, Li Y, Ye P (2026) A novel implicit cross-attention framework for RGB-T object detection. Expert Systems with Applications 314:131570. https://doi.org/10.1016/j.eswa.2026.131570
10. Ning L, Liu X, Zhou L, Zou X (2025) Cross-Modal Object Detection Based on Content-Guided Feature Fusion and Self-Calibration. Sensors 25:. https://doi.org/10.3390/s25113392
11. Yuan M, Shi X, Wang N, et al (2024) Improving RGB-infrared object detection with cascade alignment-guided transformer. Information Fusion 105:102246. https://doi.org/10.1016/j.inffus.2024.102246
12. Wang W, Mu X, Dai L, Khadka A (2026) Infrared and visible image object detection based on dual-attention multiscale fusion network. In: International Conference on Optoelectronics, Computer Science, and Algorithms (OCSA 2025). SPIE, pp 17–22
13. Zhao T, Yuan M, Jiang F, et al (2026) Removal Then Selection: A Coarse-to-Fine Fusion Perspective for RGB-Infrared Object Detection. IEEE Transactions on Intelligent Transportation Systems 27:2504–2519. https://doi.org/10.1109/TITS.2025.3638627
14. Xiang X, Zhou G, Niu B, et al (2025) Infrared-Visible Image Fusion Meets Object Detection: Towards Unified Optimization for Multimodal Perception. Remote Sensing 17:. https://doi.org/10.3390/rs17213637
15. Zhou L, Xie Y, Yang J, et al (2026) DFDNet: Differential fusion detection network for RGB-IR objects. Expert Systems with Applications 324:132494. https://doi.org/10.1016/j.eswa.2026.132494

16. Lu B, Liu H, Watanabe H (2026) Enhancing RGB-IR object detection: a frozen backbone approach with multi-receptive field attention. Vis Comput 42:164. https://doi.org/10.1007/s00371-026-04378-1
17. Chen Z, Ji H, Zhang Y (2026) Global–local feature optimization based RGB-IR fusion object detection on drone view. Chinese Journal of Aeronautics 39:103781. https://doi.org/10.1016/j.cja.2025.103781
18. Wang Z, Zhang Q (2025) Real-Time Aerial Multispectral Object Detection with Dynamic Modality-Balanced Pixel-Level Fusion. Sensors 25:. https://doi.org/10.3390/s25103039
19. Tan W, Geng B, Bai X (2026) A study on infrared-visible fusion multimodal object detection algorithm based on cross-modal information bottleneck and minimum redundancy transformation. Sci Rep 16:12991. https://doi.org/10.1038/s41598-026-35339-2
20. Xu G, Sun Y, Chen K, et al (2026) SCFusion: Spatial-Channel Cross-Frequency Guided Fusion Network for Infrared–Visible Image Object Detection. Mathematics 14:969. https://doi.org/10.3390/math14060969
21. Qi H, Qin H, Xiang X, et al (2026) LF-DETR: A Laplacian Frequency Enhanced DETR for Aerial RGB-Infrared Pedestrian Detection. Remote Sensing 18:. https://doi.org/10.3390/rs18030531
22. Hao H, Peng Y, Ye Z, et al (2025) TMRGBT-D2D: A Temporal Misaligned RGB-Thermal Dataset for Drone-to-Drone Target Detection. Drones 9:. https://doi.org/10.3390/drones9100694
23. Zhang J, Cao M, Xie W, et al (2024) E2E-MFD: towards end-to-end synchronous multimodal fusion detection. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, pp 52296–52322
24. Zeng Y, Liang T, Jin Y, Li Y (2024) MMI-Det: Exploring Multi-Modal Integration for Visible and Infrared Object Detection. IEEE Transactions on Circuits and Systems for Video Technology 34:11198–11213. https://doi.org/10.1109/TCSVT.2024.3418965
25. Guo J, Gao C, Liu F, et al (2025) DAMSDet: Dynamic Adaptive Multispectral Detection Transformer with Competitive Query Selection and Adaptive Feature Fusion. In: Leonardis A, Ricci E, Roth S, et al (eds) Computer Vision – ECCV 2024. Springer Nature Switzerland, Cham, pp 464–481
26. Hu S, Bonardi F, Bouchafa S, et al (2024) Rethinking Self-Attention for Multispectral Object Detection. IEEE Transactions on Intelligent Transportation Systems 25:16300–16311. https://doi.org/10.1109/TITS.2024.3412417
27. Zhang X, Cao S-Y, Wang F, et al (2025) Rethinking Early-Fusion Strategies for Improved Multispectral Object Detection. IEEE Transactions on Intelligent Vehicles 10:3728–3742. https://doi.org/10.1109/TIV.2024.3462488
28. Hu K, He Y, Li Y, et al (2025) EI^{2}Det: Edge-Guided Illumination-Aware Interactive Learning for Visible-Infrared Object Detection. IEEE Transactions on Circuits and Systems for Video Technology 35:7101–7115. https://doi.org/10.1109/TCSVT.2025.3539625
29. Jang J, Lee J, Paik J (2025) CAMDet: Condition-Adaptive Multispectral Object Detection Using a Visible-Thermal Translation Model. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp 1–5
30. Zhao T, Liu B, Gao Y, et al (2025) Rethinking Multi-Modal Object Detection From the Perspective of Mono-Modality Feature Learning. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV). pp 6364–6373
31. Shen J, Chen Y, Liu Y, et al (2024) ICAFusion: Iterative cross-attention guided feature fusion for

multispectral object detection. Pattern Recognition 145:109913. https://doi.org/10.1016/j.patcog.2023.109913

32. Zhang X, Zhang X, Wang J, et al (2025) TFDet: Target-Aware Fusion for RGB-T Pedestrian Detection. IEEE Transactions on Neural Networks and Learning Systems 36:13276–13290. https://doi.org/10.1109/TNNLS.2024.3443455
33. Li X, Chen S, Tian C, et al (2024) M2FNet: Mask-Guided Multi-Level Fusion for RGB-T Pedestrian Detection. IEEE Transactions on Multimedia 26:8678–8690. https://doi.org/10.1109/TMM.2024.3381377
34. Yuan M, Wei X (2024) C^2Former: Calibrated and Complementary Transformer for RGB-Infrared Object Detection. IEEE Transactions on Geoscience and Remote Sensing 62:1–12. https://doi.org/10.1109/TGRS.2024.3376819
35. Zhou M, Li Y, Yang G, et al (2025) COFNet: Contrastive Object-Aware Fusion Using Box-Level Masks for Multispectral Object Detection. IEEE Transactions on Multimedia 27:7444–7458. https://doi.org/10.1109/TMM.2025.3599097
36. Dong W, Zhu H, Lin S, et al (2025) Fusion-Mamba for Cross-Modality Object Detection. IEEE Transactions on Multimedia 27:7392–7406. https://doi.org/10.1109/TMM.2025.3599020
37. Wu J, Wang Z, Huang E, et al (2026) Bridging the RGB-IR Gap: Consensus and Discrepancy Modeling for Text-Guided Multispectral Detection
38. Zhu H, Dong W, Yang L, et al (2025) WaveMamba: Wavelet-Driven Mamba Fusion for RGB-Infrared Object Detection. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV). pp 11219–11229
39. Wang J, Tian X, Dai S, et al (2025) RGB-T Object Detection via Group Shuffled Multi-receptive Attention and Multi-modal Supervision. In: Antonacopoulos A, Chaudhuri S, Chellappa R, et al (eds) Pattern Recognition. Springer Nature Switzerland, Cham, pp 284–298
40. Wan D, Lu R, Fang Y, et al (2025) YOLOv11-RGBT: Towards a Comprehensive Single-Stage Multispectral Object Detection Framework
41. Ying X, Xiao C, An W, et al (2025) Visible-Thermal Tiny Object Detection: A Benchmark Dataset and Baselines. IEEE Transactions on Pattern Analysis and Machine Intelligence 47:6088–6096. https://doi.org/10.1109/TPAMI.2025.3544621
42. Peng P, Xu T, Song L, et al (2026) COXNet: Cross-Layer Fusion With Adaptive Alignment and Scale Integration for RGBT Tiny Object Detection. IEEE Transactions on Circuits and Systems for Video Technology 36:596–608. https://doi.org/10.1109/TCSVT.2025.3595147
43. Lv K, Xiao Z, Ren H, et al (2026) DGE-YOLO: Dual-Branch Gathering and Attention for Efficient Accurate UAV Object Detection. Applied Sciences 16:. https://doi.org/10.3390/app16084004
44. Yan Z, Wang M-J-S (2026) MFE-DETR: Multimodal Feature-Enhanced Detection Transformer for RGB–Infrared Object Detection in Aerial Imagery. Symmetry 18:. https://doi.org/10.3390/sym18030417
45. Gu Y, Chen W, Peng D (2026) UAV-based multimodal object detection via feature enhancement and dynamic gated fusion. Pattern Recognition 172:112722. https://doi.org/10.1016/j.patcog.2025.112722
46. Liu J, Fan X, Huang Z, et al (2022) 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). pp 5792–5801
47. Zhang H, Fromont E, Lefevre S, Avignon B (2020) Multispectral Fusion for Object Detection with

Cyclic Fuse-and-Refine Blocks. In: 2020 IEEE International Conference on Image Processing (ICIP). pp 276–280