

InfraNet: Quality-Aware RGB Guidance for Efficient Infrared Object Detection

Zichao Feng^{1*}, Haodong Zhu^{1,2*}, Jingying Yang¹, Sheng Xu^{3†},
Yangyang Ren^{1,2}, Yuguang Yang¹, Xuhui Liu⁴, Juan Zhang¹,
Tian Wang¹, Linlin Yang³, and Baochang Zhang¹

¹ Beihang University, Beijing, China

² Zhongguancun Academy, Beijing, China

³ Communication University of China, Beijing, China

⁴ King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia

Abstract. Robust object detection under adverse visual conditions remains a long-standing challenge for multi-modal perception systems. Existing fusion-based methods typically require both RGB and infrared (IR) inputs, and treat them equally during both training and inference, which compromises their robustness when the RGB modality becomes unreliable or unavailable. In this case, we propose **InfraNet**, an IR-centric quality-aware framework that regulates RGB guidance during training and supports flexible RGB-IR or IR-only deployment. InfraNet employs an asymmetric architecture where the primary IR pathway extracts multi-scale infrared features for predictions, while the auxiliary RGB pathway provides reliability-controlled supervisory signals. The core of InfraNet is **QualGate**, a quality-aware fusion module that learns a task-oriented control signal to suppress unreliable RGB guidance and compensate IR features during cross-modal training. Built upon InfraNet, we design two architectural variants: a lightweight IR-only architecture InfraNet-IR and an RGB-IR architecture InfraNet-RGB-IR. Our method is evaluated through extensive experiments on four benchmark datasets (LLVIP, FLIR-Aligned, M³FD, and DroneVehicle), showing strong or competitive accuracy in challenging low-light and adverse weather conditions. Notably, InfraNet maintains high efficiency in IR-only inference, making it both accurate and computationally efficient.

Keywords: Infrared Object Detection · Modality-Asymmetric Learning · Quality-Aware Cross-Modal Fusion

1 Introduction

Object detection is a crucial task in computer vision, with widespread applications in fields such as remote sensing [11], video surveillance [36], and autonomous

* Equal contribution. {fengzichao, HaodongZhu}@buaa.edu.cn

† Corresponding author. xsheng001@gmail.com

driving [59]. While significant advancements have been made in detection algorithms, many current approaches still rely predominantly on RGB images [14, 66]. Nevertheless, they face performance challenges in difficult conditions, such as low-light environments, glare, and adverse weather. This is due to the inherent dependency of RGB sensors on ambient light, which limits their effectiveness in poorly lit settings or highly fluctuating lighting conditions. To address these limitations, infrared (IR) sensing offers a robust alternative by capturing the thermal radiation emitted by objects, enabling consistent detection even under poor visibility or severe lighting variations [19]. While IR imaging provides stable responses under low-light and adverse weather conditions, RGB imaging contributes rich spatial and color details under favorable illumination. Building upon this complementarity, multi-modal object detection has emerged as an effective paradigm that integrates RGB and IR modalities to achieve more reliable perception across diverse environments. Depending on the integration stage, cross-modality fusion networks are generally categorized into pixel-level fusion [37, 70], feature-level fusion [8, 72], and decision-level fusion frameworks [27, 28].

While significant progress has been made in multi-modal object detection, most current methods treat the RGB and IR modalities equally, assuming that both provide complementary information to enhance detection performance. However, these approaches still overlook a crucial issue: the unreliability of the RGB modality under challenging conditions. In environments with poor lighting or adverse weather, RGB images can become severely degraded, often resulting in entirely dark or noisy images with little discernible object structure. When such degraded RGB features are indiscriminately fused with the more reliable IR features, they can introduce irrelevant or even harmful information, ultimately impairing the overall detection accuracy. As illustrated in Fig. 1, we visualize IR and RGB backbone features from a YOLOv8 model trained on the M³FD

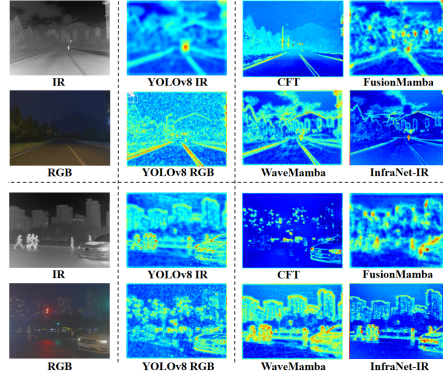


Fig. 1: Visualization of RGB and IR feature quality under nighttime conditions on M³FD dataset. While RGB features exhibit substantial noise and degradation, IR features remain robust and clear. Existing fusion methods (CFT, FusionMamba, WaveMamba) introduce additional noise compared to pure IR features, motivating our IR-centric design that regulates RGB guidance during training.

dataset using channel summation for feature visualization. Under nighttime conditions, the RGB features exhibit substantial noise, while the IR features remain relatively intact. Additionally, when comparing traditional fusion methods such as CFT [9], FusionMamba [8], and WaveMamba [72], we visualize their fused backbone features after RGB-IR training and find that the fusion process increases the noise level compared to the original IR features, further undermining

the reliability of the fusion outcome. Building on these insights, *we aim to prioritize the more reliable IR modality, while selectively incorporating RGB features only when they provide useful and stable information, thereby enhancing detection robustness and accuracy.*

In this paper, we propose **InfraNet**, a quality-aware framework for flexible RGB-IR/IR-only object detection. The core of InfraNet lies in **QualGate**, a quality-aware fusion module designed to regulate cross-modal guidance. Deployed at different scales, QualGate learns a task-oriented reliability control signal from RGB features. It suppresses unreliable RGB contributions and compensates the IR features during cross-modal training, providing more stable supervisory signals for the IR-centered representation. Built upon InfraNet, we design two architectural variants: **(1) InfraNet-IR:** A lightweight single-branch network optimized for computational efficiency, trained with RGB auxiliary supervision but deployed with IR-only inference. **(2) InfraNet-RGB-IR:** A higher-capacity dual-branch structure that processes both RGB and IR inputs during both training and inference, utilizing QualGate modules for reliability-controlled cross-modal fusion. These two architectures are implemented and trained as two independent networks. Extensive experiments conducted on four benchmark datasets demonstrate that InfraNet achieves strong accuracy for both dual-modality and IR-only inference. Notably, it maintains high efficiency in the IR-only setting and remains robust in challenging low-light and adverse weather conditions. In summary, our main contributions are summarized below:

- We propose **InfraNet**, a quality-aware RGB-guided framework for infrared object detection. InfraNet uses RGB as reliability-controlled auxiliary guidance during training to enhance the learning of IR-centered features, while allowing IR-only deployment.
- We design **QualGate**, a quality-aware fusion module that learns a task-oriented control signal to regulate cross-modal contributions by suppressing unreliable RGB guidance and compensating IR features to mitigate the risk of negative transfer.
- Extensive experiments on four RGB-IR benchmarks show that InfraNet achieves competitive or superior performance across different deployment settings, including efficient IR-only inference and higher-capacity RGB-IR inference.

2 Related Works

2.1 RGB-Infrared Multi-modal Object Detection

RGB-infrared (IR) multi-modal detection leverages complementary characteristics: RGB provides rich texture under favorable illumination, while thermal IR maintains robust signatures in darkness and adverse weather [15, 29, 41]. Existing fusion paradigms can be categorized into pixel-level, feature-level, and decision-level strategies, each with distinct trade-offs. **Pixel-level fusion** operates in the

raw image domain by merging RGB and IR inputs before feature extraction. Representative approaches leverage adversarial learning [29, 70], correlation-driven decomposition [69], or semantic-aware integration [37, 42, 43] to combine complementary information. Despite preserving fine-grained details, this paradigm introduces substantial preprocessing overhead and struggles with high-level semantic alignment, frequently resulting in inferior detection accuracy. **Feature-level fusion** extracts hierarchical representations via dual-stream backbones and integrates cross-modal features at intermediate network layers. Pioneering works employ cross-modal refinement [64] or attention mechanisms [26, 40, 56] to enhance feature complementarity. State-of-the-art methods leverage coarse-to-fine fusion strategies [68], Transformer architectures [20], or state-space models with wavelet transforms [72], achieving remarkable performance gains. YOLO-based extensions [10, 60, 67] adapt single-stage detectors with multi-modal backbones for efficient deployment. A critical limitation, however, is that all these approaches mandate *dual-modal input at inference*, rendering them impractical when RGB sensors are unavailable or heavily corrupted. **Decision-level fusion** processes each modality independently through separate detectors and ensembles predictions via score aggregation or uncertainty weighting [5, 22, 27, 28]. This paradigm offers robustness to single-modal failure and accommodates heterogeneous architectures, yet sacrifices cross-modal representation learning during feature extraction. Consequently, decision-level fusion often underperforms feature-level methods when both modalities are reliable, while incurring doubled computational cost due to parallel detector deployment. Unlike conventional fusion that treats two modalities equally, we propose a quality-aware auxiliary supervision paradigm to enhance the reliability of cross-modal fusion, supporting flexible architecture design. Specifically, built upon InfraNet, we design two architectural variants, one for IR-only inference and the other for RGB-IR inference.

2.2 Auxiliary Learning and Adaptive Fusion

Auxiliary branch learning leverages additional network branches to enhance the main task through complementary supervision during training. Related paradigms include learning using privileged information (LUPI) [46] and knowledge distillation [13], which transfer knowledge from auxiliary signals or pre-trained teachers. However, these approaches assume fixed privileged information or employ fixed teacher models, limiting their flexibility when the reliability or availability of modalities changes. Adaptive fusion mechanisms handle quality variations through illumination-aware weighting [54, 73] or attention-based interactions [3, 21, 61]. Yet most methods perform feature-level weighting without considering the deployment-oriented asymmetry between training-time guidance and test-time inputs.

Recent RGB-T/RGB-X works have also explored dynamic or reliability-aware fusion at the feature level, including scene-specific fusion and adaptive weighting mechanisms for adverse conditions and cross-modal interaction [2, 7, 23, 34, 58, 65]. While effective, most of these methods are designed for dual-modal inference and

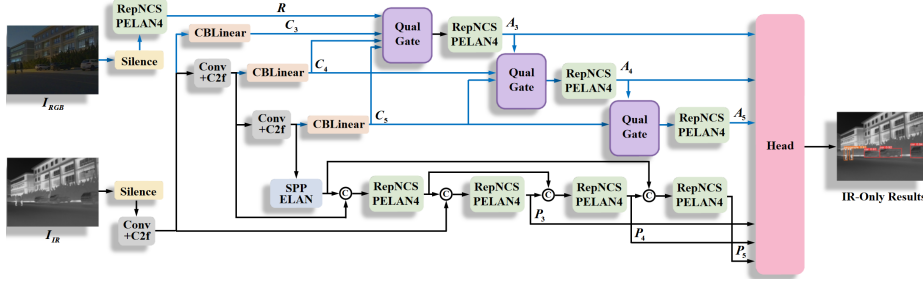


Fig. 2: Overview of InfraNet-IR (YOLOv9-based). During training, a main IR branch (black) and an auxiliary RGB branch (blue) interact via QualGate modules at three fusion sites (A3, A4, A5). Through quality-aware suppression/amplification, QualGate constrains the cross-branch interaction to promote IR-consistent representations while preventing degraded RGB cues from contaminating the IR stream, which justifies removing the RGB branch at test time. At inference, only the main IR branch and detection head are retained, yielding a pure IR detector with zero additional overhead.

primarily focus on reweighting or enhancing fusion itself. In contrast, InfraNet-IR follows a privileged-guidance setting, where RGB is used only during training and can be completely removed at inference. We also relate our design to reliability-learning modules such as CRLM and CAGF in RGB-T tracking/segmentation [31, 65] and LUPI-style methods such as HalluciDet [35]. Compared with these approaches, our key distinction is a lightweight scalar control score q that regulates *RGB suppression* and *IR compensation*, together with a practical IR-only export path for deployment.

In this paper, we propose a quality-aware auxiliary supervision framework that uses RGB as task-oriented guidance for IR representation learning. Instead of treating RGB as a required test-time modality, InfraNet-IR removes the RGB branch and all fusion modules at deployment, while InfraNet-RGB-IR keeps both modalities for higher-capacity dual-modal inference.

3 Method

3.1 Quality-Aware Fusion Module

RGB informativeness varies drastically across training samples: it remains rich and informative in daylight conditions but degrades significantly under low-light, fog, or motion blur scenarios. Naive fusion methods that treat RGB as equally reliable across all conditions inevitably inject noise when RGB quality is poor, contaminating otherwise informative IR features and destabilizing training. To address this, we propose *QualGate*, a quality-aware fusion module that learns a task-oriented reliability control signal and regulates cross-modal guidance for robust representation learning. As illustrated in Fig. 3, QualGate takes K scales IR feature $\{C_k\}_{k=1}^K$, and the auxiliary feature A_{in} which can be RGB feature or fused feature as input, and targets quality-aware fused output A_{out} .

We instantiate QualGate at each fusion site to perform *dual-path reliability modulation*: (i) *RGB suppression* to prevent negative transfer from unreliable auxiliary cues, and (ii) *IR amplification* to compensate the IR stream when RGB guidance becomes weak. Given multi-scale IR features $\{C_k\}_{k=1}^K$ and an auxiliary feature A_{in} (from the RGB branch at the current site), QualGate outputs a quality-aware fused feature A_{out} .

Reliability control score. We predict a scalar reliability-control score $q \in (0, 1)$ from A_{in} :

$$q = \sigma\left(\text{MLP}_{\text{qual}}(\text{GAP}(A_{in}))\right), \quad (1)$$

where MLP_{qual} is a lightweight two-layer bottleneck network. We adopt a scalar design for efficiency and optimization stability, consistent with our IR-centric deployment goal. Here, q is optimized for the detection objective and is used as a control signal for RGB guidance, rather than as a hand-crafted image-quality metric.

RGB suppression gate. The auxiliary contribution is explicitly suppressed by $w_{\text{rgb}}(q)$, implemented as $w_{\text{rgb}}(q) = q$, such that unreliable RGB guidance is down-weighted via $w_{\text{rgb}}(q) \cdot A_{in}$.

IR amplification gate. In parallel, we apply an inverse reliability gain to the IR stream:

$$\alpha(q) = \text{clip}(1.5 - q; 1.0, 1.5), \quad (2)$$

which increases IR contribution when RGB is degraded. This simple clipped form is used as a default non-suppressive IR compensation rule; we do not assume the constant 1.5 to be uniquely optimal.

Multi-scale IR injection and fusion. For each scale, we align IR features by a channel adapter and apply a content gate:

$$F_k = \text{CA}(C_k), \quad g_k = \sigma\left(\text{MLP}_{\text{gate}}(\text{GAP}(F_k))\right). \quad (3)$$

Finally, QualGate fuses the suppressed auxiliary term with the amplified IR injection:

$$A_{out} = w_{\text{rgb}}(q) \cdot A_{in} + \sum_{k=1}^K \alpha(q) \cdot g_k \cdot F_k. \quad (4)$$

A larger q preserves more auxiliary guidance and keeps the IR gain close to nominal scale; a smaller q reduces the auxiliary contribution and increases the IR compensation, yielding more stable IR representation shaping for IR-only inference.

Discussion. This dual-path design is empirically necessary: removing either the suppression or amplification branch leads to consistent drops (Table 5), and the learned q exhibits task-oriented behavior under controlled RGB degradations (Table 6).

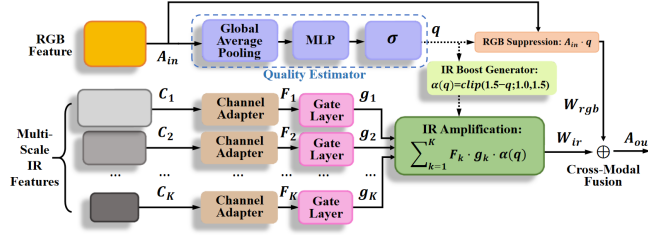


Fig. 3: Architecture of QualGate. It takes as input K scales IR features $\{C_k\}_{k=1}^K$ and the auxiliary feature A_{in} , and targets a quality-aware fused output A_{out} . For fusion, the control branch predicts a scalar score q for auxiliary guidance and compensates IR features accordingly based on $\alpha(q)$.

3.2 InfraNet

To avoid degraded RGB contaminating informative IR features and inducing negative transfer, we propose **InfraNet**, a quality-aware RGB guided framework based on QualGate. InfraNet is a plug-and-play design for common multi-scale detectors such as ResNet [12] and YOLOv5/v8/v12 [17], by preserving the dual-branch structure and strategically placing QualGate modules at corresponding multi-scale fusion sites. Based on InfraNet, we design two deployment configurations: (1) InfraNet-IR, a lightweight IR-only architecture for efficient single-modal inference, and (2) InfraNet-RGB-IR, a higher-capacity dual-branch architecture for RGB-IR fusion. Both configurations utilize QualGate modules for quality-aware feature processing.

Fig. 2 illustrates InfraNet-IR, which built on YOLOv9 components (*i.e.*, Conv, C2f, SPPELAN, RepNCSPPELAN4) [48]. Given paired RGB images I_{RGB} and infrared images I_{IR} , our InfraNet adopts two asymmetric pathways: a main IR branch (black arrows) and an auxiliary RGB branch (blue arrows). The IR pathway encodes multi-scale features $\{P_3, P_4, P_5\}$ for detection via successive Conv+C2f and SPPELAN stages, producing hierarchical representations. Meanwhile, the RGB pathway extracts base features through a lightweight stem and a RepNCSPPELAN4 block. At three designated fusion sites, the QualGate modules adaptively integrate RGB cues with the corresponding IR features, yielding $\{A_3, A_4, A_5\}$. The RGB guidance strength is regulated by the learned control score. The fused features are then processed by RepNCSPPELAN4 blocks and passed to the detection head. For InfraNet-IR, the detection head processes both feature sets ($\{A_3, A_4, A_5\}$ and $\{P_3, P_4, P_5\}$) during training, a lightweight network processes only IR features ($\{P_3, P_4, P_5\}$) during inference, without requiring RGB input or fusion components.

For the design of InfraNet-RGB-IR, we retain the full dual-stream backbone-neck with RGB-IR inputs and multi-scale QualGate fusion during inference, where RGB and IR flows jointly produce and fuse the feature maps, resulting in a higher-capacity RGB-IR detector than the pruned InfraNet-IR variant. More details can be found in the supplementary.

4 Training and Inference

4.1 Training

Given paired inputs $(I_{\text{IR}}, I_{\text{RGB}})$, the InfraNet extracts a multi-scale hierarchy $\{P_3, P_4, P_5\}$ from the main branch and $\{A_3, A_4, A_5\}$ through the quality-aware fusion mechanism. Based on the detection head, standard detection losses (*i.e.*, box regression, classification, and distribution-based localization) are employed under task-aligned assignment at each scale. Let $\mathcal{L}_{\text{det}}^{(\text{main})}$ and $\mathcal{L}_{\text{det}}^{(\text{aux})}$ denote the main and auxiliary branch detection losses, respectively. For both architectural variants (*i.e.*, InfraNet-IR and InfraNet-RGB-IR), we employ the same asymmetric weights,

$$\mathcal{L} = \mathcal{L}_{\text{det}}^{(\text{main})} + w_{\text{aux}} \mathcal{L}_{\text{det}}^{(\text{aux})}, \quad (5)$$

with w_{aux} set to 0.25 by default. This choice ensures that optimisation is dominated by the infrared pathway, while the auxiliary branch provides reliability-adaptive guidance that enhances infrared representations.

4.2 Inference

The two architectural variants, InfraNet-IR and InfraNet-RGB-IR, follow different inference settings.

InfraNet-IR. During training, InfraNet-IR takes paired RGB and IR inputs, where RGB serves as privileged auxiliary guidance for shaping the IR representation. During inference, only IR is used for prediction: the RGB auxiliary branch and all fusion operators are removed, and only the infrared backbone-neck hierarchy $\{P_s\}$ and the detection head are retained.

InfraNet-RGB-IR. InfraNet-RGB-IR adopts a dual-stream backbone that processes both RGB and IR inputs during training and inference. QualGate continues to regulate RGB-IR interactions at multiple scales.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on four established RGB-IR benchmarks: **LLVIP** [15] (15,488 aligned image pairs; pedestrian detection), **FLIR-Aligned** [44] (5,142 aligned pairs; automotive detection with three classes), **M³FD** [29] (4,200 pairs spanning six classes under diverse conditions), and **DroneVehicle** [41] (28,439 UAV-captured pairs; five-class vehicle detection).

Implementation. All experiments are conducted on a single NVIDIA RTX 4090 GPU with PyTorch. Models are trained for 300 epochs using SGD (momentum 0.937, weight decay 5×10^{-4}) with cosine learning rate decay ($10^{-2} \rightarrow 10^{-4}$) and batch size 32 under the single-GPU setting. Dual-branch training employs $w_{\text{aux}} = 0.25$. Standard YOLO augmentations (Mosaic, MixUp, HSV, horizontal

flip, scaling) are applied, with geometric transforms synchronized across modalities to preserve spatial alignment. We evaluate the models based on InfraNet-IR and InfraNet-RGB-IR, to demonstrate the flexibility of our framework, separately. For simplicity, we use Ours (IR) to represent InfraNet-IR, and Ours (RGB-IR) to represent InfraNet-RGB-IR. This ensures consistent experimental conditions across all main comparisons and ablation studies.

Metrics. We report standard metrics mAP_{50} (IoU=0.5) and mAP (IoU=0.5:0.95). Besides, for FLIR, we report additional Precision, Recall, and F1-score; for M³FD and DroneVehicle, we report per-class AP_{50} .

Training overhead and convergence. QualGate is lightweight, implemented as two-layer MLPs. For example, during training, InfraNet-IR (YOLOv8) introduces an auxiliary RGB branch (26.1M parameters) for privileged guidance, while at inference it keeps an IR-only model (28.2M) with *no extra deployment cost*. In practice, the added training overhead is acceptable—the model can be trained on a single consumer-grade GPU. Moreover, under the standard 300-epoch protocol, our method converges stably and typically reaches a stable performance regime around 130 epochs.

5.2 Comparison with State-of-the-Art Methods

LLVIP. Table 1 presents pedestrian detection results on the LLVIP dataset. Our method achieves strong performance across different backbone architectures. Specifically, our RGB-IR variants attain **70.5% mAP** (YOLOv8) and **70.6% mAP** (YOLOv12) improving over the strong recent method WaveMamba (66.0%). Notably, our IR-only deployment also demonstrates strong performance, with ResNet50-IR achieving 68.4% mAP while requiring no additional computational cost at inference, thereby validating the effectiveness of our auxiliary supervision approach.

FLIR-Aligned. Table 2 presents automotive thermal detection results on the FLIR-Aligned dataset. Our method achieves strong performance across multiple metrics: YOLOv5-RGB-IR attains the highest recall (81.8%), F1-score (83.3%), and mAP₅₀ (89.5%), while YOLOv12-RGB-IR

Table 1: Comparison results with SOTA methods on LLVIP dataset. The best results are highlighted in **red**. The second and third best results are highlighted in **green** and **blue**, respectively.

Methods	Mode	Backbone	mAP_{50}	mAP
(ACM MM'23) TIRDet [53]	IR	CSPD53	96.3	64.2
(WACV'24) HalluciDet [35]	IR	Faster RCNN	88.3	-
(CVPRW'23) CSAA [3]	RGB+IR	ResNet50	94.3	54.2
(2024) RSDet [68]	RGB+IR	ResNet50	95.8	61.3
(TIV'24) MMSANet [21]	RGB+IR	ResNet50	96.4	62.5
(ICCV'25) WaveMamba [72]	RGB+IR	ResNet50	97.3	65.0
Ours (RGB-IR)	RGB+IR	ResNet50	98.1	69.2
Ours (IR)	IR	ResNet50	97.7	68.4
(CVPR'24) Text-IF [61]	RGB+IR	Transformer	94.1	60.2
(IF'23) DIVFusion [43]	RGB+IR	YOLOv5	89.8	52.0
(PR'24) ICAFusion [37]	RGB+IR	YOLOv5	95.2	60.1
(ICCV'25) WaveMamba [72]	RGB+IR	YOLOv5	97.6	65.2
Ours (RGB-IR)	RGB+IR	YOLOv5	98.3	70.2
Ours (IR)	IR	YOLOv5	97.3	68.1
(ICHMS'24) MDMDet [60]	RGB+IR	YOLOv7	96.5	61.5
YOLOv8l-IR [47]	IR	YOLOv8	95.2	62.1
YOLOv8l-RGB [47]	RGB	YOLOv8	91.9	54.0
(RS'24) ACDF-YOLO [10]	RGB+IR	YOLOv8	96.5	61.3
(ICHMS'24) MSFF [60]	RGB+IR	YOLOv8	96.8	61.9
(Sensors'24) FAWDet [67]	RGB+IR	YOLOv8	97.1	62.1
(ICCV'25) WaveMamba [72]	RGB+IR	YOLOv8	98.3	66.0
Ours (RGB-IR)	RGB+IR	YOLOv8	98.1	70.5
Ours (IR)	IR	YOLOv8	97.3	68.8
Ours (RGB-IR)	RGB+IR	YOLOv12	98.0	70.6
Ours (IR)	IR	YOLOv12	97.2	67.6

Table 2: Comparison results with SOTA methods on FLIR-Aligned dataset. The best results are highlighted in **red**, the second and third best results are highlighted in **green** and **blue**, respectively. For Precision/Recall/F1/ mAP_{50} / mAP , larger is better; for Parameters and Inference time, smaller is better.

Methods	Mode	Backbone	Precision	Recall	F1	mAP_{50}	mAP	Parameters	Inference time (ms)
(ACM MM'23) TIRDet [53]	IR	CSPD53	-	-	-	81.4	44.3	-	-
(TITS'23) MFPT [73]	RGB+IR	ResNet50	79.3	72.9	76.0	80.0	41.9	200.0M	80.0
(ICCV'25) WaveMamba [72]	RGB+IR	ResNet50	84.2	77.9	80.9	86.5	47.9	193.2M	53.2
Ours (RGB-IR)	RGB+IR	ResNet50	83.5	80.4	81.9	88.6	54.3	73.8M	14.4
Ours (IR)	IR	ResNet50	84.9	78.8	81.7	87.3	53	42.1M	9.4
(PRL'24) CrossFormer [20]	RGB+IR	YOLOv5	78.1	72.8	75.4	79.3	42.1	340.0M	80.0
(ICCV'25) WaveMamba [72]	RGB+IR	YOLOv5	83.9	79.7	81.7	85.8	47.5	45.6M	44.1
Ours (RGB-IR)	RGB+IR	YOLOv5	85.1	81.8	83.3	89.5	55.1	52.2M	13.2
Ours (IR)	IR	YOLOv5	84.9	80.4	82.6	88.1	53.8	26.0M	11.6
YOLOv8-IR [47]	IR	YOLOv8	75.1	65.3	69.9	72.9	38.3	43.7M	22.0
YOLOv8-RGB [47]	RGB	YOLOv8	71.6	62.2	66.6	66.3	28.2	43.7M	22.0
(GRSL'24) ESSFN [54]	RGB+IR	YOLOv8	81.4	73.5	77.2	80.8	42.3	80.2M	47.0
(ICCV'25) WaveMamba [72]	RGB+IR	YOLOv8	84.2	80.9	82.5	88.4	48.1	69.1M	40.0
Ours (RGB-IR)	RGB+IR	YOLOv8	85.8	79.9	82.8	89.4	55.2	54.3M	13.4
Ours (IR)	IR	YOLOv8	84.6	80.9	82.7	88.5	54.0	28.2M	12.5
Ours (RGB-IR)	RGB+IR	YOLOv12	84.3	81.7	83.0	89.3	55.8	56.8M	13.2
Ours (IR)	IR	YOLOv12	85.0	79.5	82.2	87.7	53.2	24.7M	12.0

Table 3: Comparison results with SOTA methods on DroneVehicle dataset. The best results are highlighted in **red**. The second and third best results are highlighted in **green** and **blue**, respectively.

Methods	Mode	Backbone	mAP_{50}	mAP	Car	Bus	Truck	Freight-car	Van
(ECCV'22) TSFADet [62]	RGB+IR	ResNet50	73.1	44.1	89.9	89.8	67.9	63.7	54.0
(GRSL'24) GLFNet [18]	RGB+IR	ResNet50	71.4	54.8	90.3	88.0	72.7	53.6	52.6
(TGRS'24) LF-MDet [39]	RGB+IR	ResNet50	71.8	51.3	-	-	-	-	-
(TGRS'24) C^2 Former- S^2 ANet [63]	RGB+IR	ResNet50	74.2	47.3	90.2	89.8	68.3	64.4	58.5
(2024) DMM [71]	RGB+IR	ResNet50	77.2	55.8	90.4	88.7	77.8	63.0	66.0
(JST'24) FFODNet [52]	RGB+IR	ResNet50	76.3	56.9	-	-	-	-	-
(ICCV'25) WaveMamba [72]	RGB+IR	ResNet50	79.3	59.9	94.6	90.2	79.6	68.0	64.1
Ours (RGB-IR)	RGB+IR	ResNet50	83.9	64.1	98.0	97.3	80.5	78.0	65.7
Ours (IR)	IR	ResNet50	83.5	63.0	97.8	97.1	79.3	75.0	68.3
(YAC'24) CAFN-IA [57]	RGB+IR	YOLOv5	69.3	56.1	89.1	90.8	62.0	57.3	47.1
(RSL'24) multimodal DINO [38]	RGB+IR	YOLOv5	72.5	50.3	89.5	88.8	75.4	54.3	54.3
(MTA'23) SLBAF-Net [6]	RGB+IR	YOLOv5	76.8	49.5	90.2	89.9	76.0	68.2	59.9
(CVPR'24) CSOM-ODAF [4]	RGB+IR	YOLOv5	77.1	50.1	90.1	89.8	75.6	68.2	61.8
(ICCV'25) WaveMamba [72]	RGB+IR	YOLOv5	79.5	60.2	94.8	90.1	80.1	68.2	64.3
Ours (RGB-IR)	RGB+IR	YOLOv5	84.1	63.8	98.1	97.2	82.1	76.2	67.1
Ours (IR)	IR	YOLOv5	83.6	63.1	98.0	96.9	79.6	75.5	68.2
(Sensors'23) Dual-YOLO [1]	RGB+IR	YOLOv7	71.9	55.2	95.9	91.6	69.7	55.9	46.6
YOLOv8l-IR [47]	IR	YOLOv8	71.9	49.1	93.4	91.9	69.3	53.7	51.1
YOLOv8l-RGB [47]	RGB	YOLOv8	70.2	48.6	92.5	91.7	68.8	47.8	50.2
(AEORS'24) YOLOFIV [49]	RGB+IR	YOLOv8	64.7	53.1	95.9	91.6	64.2	34.6	37.3
(TGRS'24) CRSIOD [50]	RGB+IR	YOLOv8	73.2	50.8	95.6	92.2	71.7	50.5	55.8
(IJAE0G'24) M2FNet [16]	RGB+IR	YOLOv8	76.8	50.4	95.8	93.1	75.9	59.8	59.4
(Sensors'24) IV-YOLO [45]	RGB+IR	YOLOv8	74.6	56.8	97.2	94.3	65.4	63.1	53.0
(RS'24) DAAB-FFPN [51]	RGB+IR	YOLOv8	75.2	56.3	-	-	-	-	-
(ICCV'25) WaveMamba [72]	RGB+IR	YOLOv8	79.8	60.5	95.0	90.6	80.4	68.5	64.5
Ours (RGB-IR)	RGB+IR	YOLOv8	85.0	64.9	98.1	97.3	82.6	79.9	67.3
Ours (IR)	IR	YOLOv8	84.0	63.4	98.0	97.0	80.0	76.9	68.4
Ours (RGB-IR)	RGB+IR	YOLOv12	84.8	64.7	98.0	97.3	82.6	78.3	67.6
Ours (IR)	IR	YOLOv12	84.7	63.9	98.1	97.3	81.3	77.7	69.3

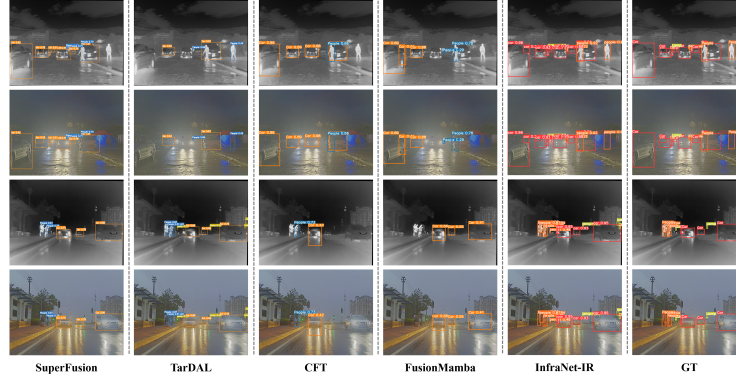


Fig. 4: Visualization of detection results of several cross-modality object detection methods on M³FD dataset. InfraNet-IR produces tighter bounding boxes, fewer false positives, and more consistent detections across diverse illumination conditions. Unlike prior methods (SuperFusion, TarDAL, CFT, FusionMamba), which often suffer from RGB degradation and background leakage, InfraNet-IR maintains robust localization by relying on IR-dominant features enhanced through quality-aware supervision. GT annotations are provided in the last column.

achieves the best mAP (55.8%). Notably, our IR-only variants demonstrate superior efficiency: YOLOv12-IR requires only 24.7M parameters and ResNet50-IR achieves 9.4ms inference time, substantially outperforming competing fusion methods while maintaining competitive accuracy. This validates our auxiliary supervision paradigm that achieves strong performance without deployment overhead.

DroneVehicle. Table 3 presents UAV-based 5-class vehicle detection results on the DroneVehicle dataset. Our YOLOv12-IR achieves the best results among the compared methods with **84.7% mAP_{50}** and **63.9% mAP**, surpassing WaveMamba (79.8%/60.5%) by substantial margins of **+4.9%/+3.4%** (+6.1%/+5.6% relative)—the largest absolute gains across all datasets. Per-class analysis shows particularly strong performance on challenging small-object categories: **Car** 98.1% AP_{50} (rank 1st), **Bus** 97.3% (1st), **Freight-car** 77.7% (+9.2% over WaveMamba), **Van** 69.3% (+4.8%).

These substantial small-object improvements stem from our hierarchical multi-scale IR tip aggregation: at the finest detection scale A3, fusing P3/P4/P5 features provides robust representations resilient to RGB degradation from atmo-

Table 4: Comparison results on M³FD subsets in AP50. The best results are highlighted in **red**, the second best results are highlighted in **blue**, respectively.

Method	Backbone	Day	Overcast	Night	Challenge
Infrared [29]	YOLOv5	58.9	80.3	79.5	70.9
Visible [29]	YOLOv5	59.1	82.4	78.7	75.9
DenseFuse [24]	Dense-CNN	60.8	75.9	80.6	83.7
FusionGAN [32]	CNN	60.1	81.6	79.8	66.7
RFN [56]	RFN	59.2	79.6	80.3	82.7
GANMcC [33]	CNN	60.3	79.6	81.1	82.7
DDcGAN [25]	CNN	59.4	78.0	77.1	68.9
MFEIF [30]	CNN	60.7	77.0	81.2	68.3
U2Fusion [55]	VGG-16	60.4	79.3	78.3	83.6
TarDAL [29]	YOLOv5	61.3	82.3	81.6	84.6
Ours (RGB-IR)	YOLOv5	87.6	86.8	96.6	99.5
Ours (IR)	YOLOv5	82.7	81.2	92.3	99.5
Ours (RGB-IR)	YOLOv8	87.7	87.4	96.8	99.5
Ours (IR)	YOLOv8	83.5	81.4	92.1	99.5
Ours (RGB-IR)	YOLOv12	87.6	88.7	96.7	99.5
Ours (IR)	YOLOv12	81.2	80.8	92.3	99.4

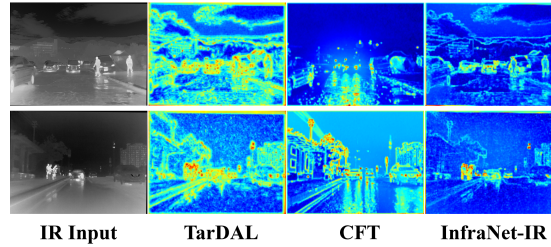


Fig. 5: Feature-level visualization on M³FD nighttime examples. InfraNet-IR produces more target-aligned responses with less background activation. This visualization is used to diagnose the effect of RGB-guided training under degraded RGB conditions; the deployed InfraNet-IR model still performs IR-only inference.

spheric scattering at high altitudes—a critical advantage for aerial imaging where RGB visibility is severely compromised by weather, haze, and altitude-induced degradation.

M³FD. We first evaluate our model on the M³FD benchmark, and complete quantitative results are provided in the Appendix. To further characterize performance under different illumination conditions, we also report results on the commonly used Day, Overcast, Night, and Challenge subsets of M³FD. As shown in Table 4, InfraNet obtains strong results across these subsets, especially in night scenes where RGB visibility is severely compromised. Since many recent RGB-IR detection methods do not report results under exactly the same subset protocol, Table 4 is used as a diagnostic comparison rather than an exhaustive current-SOTA ranking. The IR-only variants also maintain strong night performance, indicating that RGB-guided training can improve the IR representation without requiring RGB at deployment. Fig. 4 shows that our model achieves tighter localization and fewer false positives than representative fusion baselines. To further inspect the feature behavior under degraded RGB conditions, we visualize intermediate features in Fig. 5. These visualizations are used as training-time feature-level diagnostics; the deployed InfraNet-IR model still uses only IR input.

5.3 Ablation Studies

We conduct comprehensive ablation studies to validate our design choices and dissect the contribution of each component. Unless otherwise specified, all ablations use YOLOv8 with consistent training protocols (300 epochs, SGD optimizer, standard augmentation).

QualGate Design Dissection. Table 5 isolates the contribution of QualGate from both the *module* and *mechanism* perspectives. Replacing QualGate with HallucDet-style pseudo-RGB guidance or existing adaptive weighting modules (CAGE/CRLM) leads to inferior results, indicating that generic reliability-aware fusion is insufficient for our IR-centric setting. More importantly, removing either branch of the proposed modulation degrades performance: disabling IR

amplification ($\alpha(q) \equiv 1$) causes a clear drop, and disabling RGB suppression ($w_{\text{rgb}} \equiv 1$) also hurts accuracy. These results support our central design choice that a learned scalar reliability score q should jointly *suppress unreliable RGB guidance* and *compensate the IR stream*, which is critical for robust IR representation learning and IR-only deployment.

Table 5: QualGate design dissection on LLVIP using YOLOv8 (InfraNet-IR setting). We compare QualGate with representative alternatives (HallucDet-style pseudo-RGB guidance, CAGF, and CRLM), and further ablate its two key components: RGB suppression and IR amplification. The full QualGate consistently achieves the best performance, validating that the gain comes from the proposed *suppress+amplify* reliability modulation rather than generic feature reweighting.

Setting/Backbone	mAP_{50}	mAP_{50-95}
A. QualGate ablation (LLVIP, YOLOv8-IR)		
HallucDet (Faster RCNN) [35]	88.3	-
Ours training with pseudo-RGB by HallucDet	95.2	64.9
Ours replacing QualGate with CAGF [65]	95.7	66.8
Ours replacing QualGate with CRLM [31]	96.9	66.7
Ours with Full QualGate	97.3	68.8
w/o IR amp ($\alpha(q) \equiv 1$)	97.1	67.2
w/o RGB supp ($w_{\text{rgb}} \equiv 1$)	97.1	68.4
Naive ($q \equiv 1$)	96.9	67.7

Controlled RGB Degradation and q Validation. Table 6 analyzes the behavior of the learned scalar control score q under controlled RGB perturbations, without using any extra quality annotations. Under fog corruption, q decreases mildly as degradation becomes stronger, while our method remains more stable than the naive variant, especially at severe degradation levels. For brightness perturbation, q varies more mildly and is not intended to serve as a handcrafted image-quality metric. These results support our interpretation of q as a task-oriented reliability control signal for limiting harmful RGB guidance, rather than as an explicit estimator of perceptual RGB quality. Additional analyses on reliability granularity, fusion-site placement, and IR amplification range are provided in the supplementary material.

Deployment Flexibility Analysis. Table 7 (a) compares three deployment configurations on M3FD: a conventional IR baseline and our two InfraNet-based detectors (InfraNet-IR and InfraNet-RGB-IR). InfraNet-IR improves the baseline by +1.4 AP50 and +1.8 AP with the same parameters and inference time, showing that auxiliary RGB supervision strengthens IR feature learning without extra deployment cost. When RGB is available, InfraNet-RGB-IR yields larger gains (+5.1 AP50, +6.8 AP), confirming the benefit of quality-aware RGB-IR fusion.

Auxiliary Branch Loss Weight. Table 7 (b) evaluates the auxiliary branch loss weight w_{aux} on LLVIP. $w_{\text{aux}} = 0.25$ achieves optimal performance (68.8% mAP),

Table 6: Controlled RGB degradation analysis on LLVIP with fixed IR input. We progressively degrade only the RGB modality (Fog / Brightness) and report the predicted quality score q together with detection performance (AP50/AP). “Naive” disables reliability control by forcing $q \equiv 1$. The results show that learned reliability control improves robustness under degraded RGB conditions, without requiring q to be interpreted as a perceptual image-quality score.

Level	<i>Fog</i>			<i>Brightness</i>		
	q	Ours	Naive	q	Ours	Naive
0.2	0.441	97.0/67.6	95.8/66.9	0.459	97.5/67.8	96.7/66.7
0.4	0.439	97.0/67.7	95.6/66.1	0.465	97.5/67.6	96.8/67.0
0.6	0.437	97.1/67.7	95.3/65.9	0.473	97.3/67.1	96.2/65.3
0.8	0.435	97.1/67.7	94.4/64.5	0.481	96.9/66.9	95.2/65.4
1.0	0.434	97.1/67.8	92.6/61.8	0.484	96.5/66.4	93.6/63.9

Table 7: (a) Deployment flexibility analysis on M3FD using YOLOv8. InfraNet-IR achieves zero-overhead IR-only deployment with superior performance, while InfraNet-RGB-IR offers optional dual-modal fusion. (b) Ablation study on w_{aux} using YOLOv8 on LLVIP.

(a)					(b)			
Method	M ³ FD Performance		Computational Efficiency		Backbone	w_{aux}	AP ₅₀	AP
	AP ₅₀	AP	Params (M)	Speed (ms)				
IR baseline	84.8	57.2	28.2	12.5	YOLOv8	0.0	97.0	68.0
					YOLOv8	0.1	97.2	68.2
					YOLOv8	0.25	97.3	68.8
Ours (IR)	86.2	59.0	28.2	12.5	YOLOv8	0.5	97.1	68.3
Ours (RGB-IR)	89.9	64.0	54.3	13.4	YOLOv8	1.0	97.1	68.1

indicating a balanced trade-off between the use of RGB guidance and the prevention of interference. When w_{aux} is too small (0.0–0.1) or too large (0.5–1.0), performance degrades (68.0–68.3% mAP), validating our asymmetric loss weighting strategy.

6 Conclusion

In this paper, we introduce InfraNet, a modality-asymmetric quality-aware framework for RGB-IR/IR-only object detection. Based on InfraNet, we design two separate architectures, the IR-only detector (InfraNet-IR) and the dual-branch RGB-IR detector (InfraNet-RGB-IR). In both architectures, we keep IR as the primary pathway and use RGB as reliability-controlled guidance when available. This cross-modal guidance is regulated by **QualGate**, which learns a task-oriented control signal to suppress unreliable RGB guidance and compensate IR features during training. Through extensive experiments on four benchmark datasets, we show that InfraNet achieves strong performance under different deployment settings. Notably, InfraNet-IR maintains high inference efficiency by removing the RGB branch and fusion modules at deployment, while InfraNet-RGB-IR provides a higher-capacity option when both modalities are available.

Acknowledgements

The work was supported by the National Key Research and Development Program of China (No.2023YFC3306401). This research was also supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020007, Beijing Natural Science Foundation L244043.

References

1. Bao, C., Cao, J., Hao, Q., Cheng, Y., Ning, Y., Zhao, T.: Dual-yolo architecture from infrared and visible images for object detection. *Sensors* (2023)
2. Bijelic, M., Gruber, T., Mannan, F., Kraus, F., Ritter, W., Dietmayer, K., Heide, F.: Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In: *CVPR*. pp. 11682–11692 (2020)
3. Cao, Y., Bin, J., Hamari, J., Blasch, E., Liu, Z.: Multimodal object detection by channel switching and spatial attention. In: *CVPRW* (2023)
4. Chen, C., Qi, J., Liu, X., Bin, K., Fu, R., Hu, X., Zhong, P.: Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection. In: *CVPR* (2024)
5. Chen, Y.T., Shi, J., Ye, Z., Mertz, C., Ramanan, D., Kong, S.: Multimodal object detection via probabilistic ensembling. In: *ECCV* (2022)
6. Cheng, X., Geng, K., Wang, Z., Wang, J., Sun, Y., Ding, P.: Slbaf-net: Super-lightweight bimodal adaptive fusion network for uav detection in low recognition environment. *Multimedia Tools Appl.* (2023)
7. Deevi, S.A., Lee, C., Gan, L., Nagesh, S., Pandey, G., Chung, S.J.: Rgb-x object detection via scene-specific fusion modules. In: *WACV*. pp. 7366–7375 (2024)
8. Dong, W., Zhu, H., Lin, S., Luo, X., Shen, Y., Guo, G., Zhang, B.: Fusion-mamba for cross-modality object detection. *IEEE Trans. Multim.* **27**, 7392–7406 (2025)
9. Fang, Q., Han, D., Wang, Z.: Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273* (2021)
10. Fei, X., Guo, M., Li, Y., Yu, R., Sun, L.: Acdf-yolo: Attentive and cross-differential fusion network for multimodal remote sensing object detection. *Remote. Sens.* (2024)
11. Gui, S., Song, S., Qin, R., Tang, Y.: Remote sensing object detection in the deep learning era - A review. *Remote. Sens.* **16**(2), 327 (2024)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), <https://arxiv.org/abs/1512.03385>, accessed: June 29, 2026
13. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015)
14. Hussain, M.: Yolov1 to v8: Unveiling each variant-a comprehensive review of YOLO. *IEEE Access* **12**, 42816–42833 (2024)
15. Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: Llvip: A visible-infrared paired dataset for low-light vision. In: *ICCVW* (2021)
16. Jiang, C., Ren, H., Yang, H., Huo, H., Zhu, P., Yao, Z., Li, J., Sun, M., Yang, S.: M2fnet: Multi-modal fusion network for object detection from visible and thermal infrared images. *Int. J. Appl. Earth Obs. Geoinf.* (2024)
17. Jocher, G., chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K., Fang, J., Wong, c., Yifu, Z., Montes, D., et al.: ultralytics/yolov5: v6. 2-yolov5 classification models, apple ml, reproducibility, clearml and deci. ai integrations. *Zenodo* (2022)

18. Kang, X., Yin, H., Duan, P.: Global-local feature fusion network for visible-infrared vehicle detection. *IEEE Geosci. Remote Sens. Lett.* (2024)
19. Kristo, M., Ivasic-Kos, M., Pobar, M.: Thermal object detection in difficult weather conditions using YOLO. *IEEE Access* **8**, 125459–125476 (2020)
20. Lee, S., Park, J., Park, J.: Crossformer: Cross-guided attention for multi-modal object detection. *Pattern Recognit. Lett.* (2024)
21. Li, A., Wang, Z., Wang, F., Liu, Z., Yin, G., Fang, R., Geng, K.: A novel semantic information perception architecture for extreme targets detection in complex traffic scenarios. *IEEE Trans. Intell. Veh.* (2024)
22. Li, C., Song, D., Tong, R., Tang, M.: Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *PR* (2019)
23. Li, H., et al.: Crossmodalnet: A dual-modal object detection network based on cross-modal fusion and channel interaction. *Expert Systems with Applications* p. 129677 (2025)
24. Li, H., Wu, X.J., Kittler, J.: Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing* **28**(5), 2614–2623 (2019)
25. Li, H., Wu, X.J., Kittler, J.: Infrared and visible image fusion using a dual discriminator conditional generative adversarial network. *IEEE Access* **7**, 103964–103977 (2019)
26. Li, J., Chen, J., Liu, J., Ma, H.: Learning a graph neural network with cross modality interaction for image fusion. In: *ACMMM* (2023)
27. Li, Q., Zhang, C., Hu, Q., Zhu, P., Fu, H., Chen, L.: Stabilizing multispectral pedestrian detection with evidential hybrid fusion. *TCSVT* (2024)
28. Li, X., Chen, S., Tian, C., Zhou, H., Zhang, Z.: M²fnet: Mask-guided multi-level fusion for rgb-t pedestrian detection. *IEEE Trans. Multimed.* (2024)
29. Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: *CVPR* (2022)
30. Liu, J., Fan, X., Jiang, J., Liu, R., Luo, Z.: Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology* **32**(1), 105–119 (2022)
31. Liu, L., Li, C., Xiao, Y., Tang, J.: Quality-aware rgbt tracking via supervised reliability learning and weighted residual guidance. In: *ACMMM*. pp. 3129–3137 (2023)
32. Ma, J., Yu, W., Liang, P., Jiang, J., Ling, H.: FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information Fusion* **48**, 11–26 (2019)
33. Ma, J., Zhang, H., Shao, Z., Liang, P., Xu, H.: Ganmcc: A generative adversarial network with multiclassification constraints for ivif. *IEEE Transactions on Instrumentation and Measurement* **70**, 1–14 (2020)
34. Medeiros, H.R., et al.: Mixed patch visible-infrared modality agnostic object detection. In: *WACV* (2025)
35. Medeiros, H.R., Peña, F.A.G., Aminbeidokhti, M., Dubail, T., Granger, E., Pedersoli, M.: Hallucidet: Hallucinating rgb modality for person detection through privileged information. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. pp. 1444–1453 (2024)
36. Payghode, V., Goyal, A., Bhan, A., Iyer, S.S., Dubey, A.K.: Object detection and activity recognition in video surveillance using neural networks. *Int. J. Web Inf. Syst.* **19**(3/4), 123–138 (2023)
37. Shen, J., Chen, Y., Liu, Y., Zuo, X., Fan, H., Yang, W.: Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *PR* (2024)

38. Sun, X., Yu, Y., Cheng, Q.: Adaptive multimodal feature fusion with frequency domain gate for remote sensing object detection. *Remote Sens. Lett.* (2024)
39. Sun, X., Yu, Y., Cheng, Q.: Low-rank multimodal remote sensing object detection with frequency filtering experts. *IEEE Trans. Geosci. Remote Sens.* (2024)
40. Sun, Y., Cao, B., Zhu, P., Hu, Q.: Detfusion: A detection-driven infrared and visible image fusion network. In: *ACMMM* (2022)
41. Sun, Y., Cao, B., Zhu, P., Hu, Q.: Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *TCSVT* (2022)
42. Tang, L., Deng, Y., Ma, Y., Huang, J., Ma, J.: Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA J. Autom. Sin.* (2022)
43. Tang, L., Xiang, X., Zhang, H., Gong, M., Ma, J.: Divfusion: Darkness-free infrared and visible image fusion. *Inf. Fusion* (2023)
44. Teledyne FLIR: Free flir thermal dataset for algorithm training. <https://www.flir.com/oem/adas/adas-dataset-form/> (2018), accessed: June 29, 2026
45. Tian, D., Yan, X., Zhou, D., Wang, C., Zhang, W.: Iv-yolo: A lightweight dual-branch object detection network. *Sensors* (2024)
46. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. *Neural Networks* **22**(5-6), 544–557 (2009)
47. Varghese, R., M., S.: Yolov8: A novel object detection algorithm with enhanced performance and robustness. In: *ADICS* (2024)
48. Wang, C.Y., Yeh, I.H., Liao, H.Y.M.: Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616* (2024)
49. Wang, H., Wang, C., Fu, Q., Si, B., Zhang, D., Kou, R., Yu, Y., Feng, C.: Yolofiv: Object detection algorithm for around-the-clock aerial remote sensing images by fusing infrared and visible features. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* (2024)
50. Wang, H., Wang, C., Fu, Q., Zhang, D., Kou, R., Yu, Y., Song, J.: Cross-modal oriented object detection of uav aerial images based on image feature. *IEEE Trans. Geosci. Remote Sens.* (2024)
51. Wang, J., Su, N., Zhao, C., Yan, Y., Feng, S.: Multi-modal object detection method based on dual-branch asymmetric attention backbone and feature fusion pyramid network. *Remote Sens.* (2024)
52. Wang, J., Xu, C., Zhao, C., Gao, L., Wu, J., Yan, Y., Feng, S., Su, N.: Multimodal object detection of uav remote sensing based on joint representation optimization and specific information enhancement. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* (2024)
53. Wang, Z., Colonnier, F., Zheng, J., Acharya, J., Jiang, W., Huang, K.: Tirdet: Mono-modality thermal infrared object detection based on prior thermal-to-visible translation. In: *Proceedings of the 31st ACM International Conference on Multimedia*. pp. 2663–2672 (2023)
54. Xu, F., Xu, T., Hong, L., Peng, P., Guo, J., Li, J.: Enhanced spectral-spatial fusion network for multispectral object detection in ground-aerial images. *IEEE Geoscience and Remote Sensing Letters* (2024)
55. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(1), 502–518 (2022)
56. Xu, H., Ma, J., Yuan, J., Le, Z., Liu, W.: Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In: *CVPR* (2022)
57. Xu, J., Mo, B., Zhao, J., Zhao, C., Tao, Y., Han, S.: Cross-modal adaptive fusion object detection based on illumination-awareness. In: *YAC* (2024)

58. Xu, Y., et al.: Mambatrans: Multimodal fusion image translation via large language model priors for downstream visual tasks. arXiv preprint arXiv:2508.07803 (2025)
59. Yang, B., Li, J., Zeng, T.: A review of environmental perception technology based on multi-sensor information fusion in autonomous driving. *World Electric Vehicle Journal* **16**(1), 20 (2025)
60. Yang, H., Bian, W., Wang, J., Bo, Y., Mi, Y.: A dual-modality pedestrian detection method based on multi-scale feature fusion. In: ICHMS (2024)
61. Yi, X., Xu, H., Zhang, H., Tang, L., Ma, J.: Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In: CVPR (2024)
62. Yuan, M., Wang, Y., Wei, X.: Translation, scale and rotation: Cross-modal alignment meets rgb-infrared vehicle detection. In: ECCV (2022)
63. Yuan, M., Wei, X.: C²former: Calibrated and complementary transformer for rgb-infrared object detection. *IEEE Trans. Geosci. Remote Sens.* (2024)
64. Zhang, H., Ma, J.: Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *IJCV* (2021)
65. Zhang, Q., Zhao, S., Luo, Y., Zhang, D., Huang, N., Han, J.: Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In: CVPR. pp. 2633–2642 (2021)
66. Zhao, H., Chu, K., Zhang, J., Feng, C.: Yolo-fsd: An improved target detection algorithm on remote-sensing images. *IEEE Sensors Journal* **23**(24), 30751–30764 (2023)
67. Zhao, P., Ye, X., Du, Z.: Object detection in multispectral remote sensing images based on cross-modal cross-attention. *Sensors* (2024)
68. Zhao, T., Yuan, M., Jiang, F., Wang, N., Wei, X.: Removal then selection: A coarse-to-fine fusion perspective for rgb-infrared object detection. arXiv preprint arXiv:2401.10731 (2024)
69. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., Van Gool, L.: Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: CVPR (2023)
70. Zhao, Z., Xu, S., Zhang, C., Liu, J., Zhang, J., Li, P.: Didfuse: Deep image decomposition for infrared and visible image fusion. In: IJCAI (2020)
71. Zhou, M., Li, T., Qiao, C., Xie, D., Wang, G., Ruan, N., Mei, L., Yang, Y.: Dmm: Disparity-guided multispectral mamba for oriented object detection in remote sensing. arXiv:2407.08132 (2024)
72. Zhu, H., Dong, W., Yang, L., Li, H., Yang, Y., Ren, Y., Zhu, Q., Feng, Z., Li, C., Lin, S., Wang, R., Luo, X., Zhang, B.: Wavemamba: Wavelet-driven mamba fusion for rgb-infrared object detection. *ICCV* (2025)
73. Zhu, Y., Sun, X., Wang, M., Huang, H.: Multi-modal feature pyramid transformer for rgb-infrared object detection. *IEEE Trans. Intell. Transport. Syst.* (2023)