Pattern Recognition 176 (2026) 113189

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

IRDFusion: Iterative relation-map difference guided feature fusion for multispectral object detection

Jifeng Shen [a,*], Haibo Zhan [a], Xin Zuo [b], Heng Fan [c], Xiaohui Yuan [c], Jun Li [d], Wankou Yang [e]

[a] School of Electrical and Information Engineering, Jiangsu University, 212013, Zhenjiang, China
[b] School of Computer Science and Engineering, Jiangsu University of Science and Technology, 212003, Zhenjiang, China
[c] Department of Computer Science and Engineering, University of North Texas, Denton, TX, 76207, USA
[d] School of Computing, Nanjing Normal University, 210046, Nanjing, Jiangsu, China
[e] School of Automation, Southeast University, 210096, Nanjing, China

ARTICLE INFO

Keywords:
Multispectral object detection
Cross-modal feature fusion
Mutual feature refinement module
Differential feature feedback module

ABSTRACT

Current multispectral object detection methods often retain extraneous background or noise during feature fusion, limiting perceptual performance. To address this, we propose a feature fusion framework based on cross-modal feature contrastive and screening strategy, diverging from conventional approaches. The proposed method adaptively enhances salient structures by fusing object-aware complementary cross-modal features while suppressing shared background interference. Our solution centers on two novel, specially designed modules: the Mutual Feature Refinement Module (MFRM) and the Differential Feature Feedback Module (DFFM). The MFRM enhances intra- and inter-modal feature representations by modeling their relationships, thereby improving cross-modal alignment and discriminative power. Inspired by feedback differential amplifiers, the DFFM dynamically computes inter-modal differential features as guidance signals and feeds them back to the MFRM, enabling adaptive fusion of complementary information while suppressing common-mode noise across modalities. To enable robust feature learning, the MFRM and DFFM are integrated into a unified framework, which is formally formulated as an Iterative Relation-Map Differential Guided Feature Fusion mechanism, termed IRDFusion. IRDFusion enables high-quality cross-modal fusion by progressively amplifying salient relational signals through iterative feedback, while suppressing feature noise, leading to significant performance gains. In extensive experiments on FLIR, LLVIP and M^3FD datasets, IRDFusion achieves state-of-the-art performance and consistently outperforms existing methods across diverse challenging scenarios, demonstrating its robustness and effectiveness. Code will be available at https://github.com/61s61min/IRDFusion.git.

1. Introduction

Multispectral object detection employs data from multiple spectral bands, such as visible and infrared light, for object recognition and localization. It is widely applied in autonomous driving and video surveillance tasks in poor weather conditions (e.g. darkness, fog, rain or snow). Compared to single-spectrum data, multispectral data can more comprehensively reflect the spectral characteristics of the object and its background, thereby significantly improving the robustness and accuracy of detection. It is worth noting that multispectral object detection is different from general multimodal object detection. Multispectral detection focuses on information captured from different spectral bands of the optical sensor system (e.g., RGB and infrared) [1–3], where the modalities are physically correlated and often exhibit strong structural consistency. In contrast, multimodal detection usually involves heterogeneous sources such as images, texts, LiDAR point clouds, or audio [4,5], where the modalities differ not only in physical characteristics but also in semantic representation, requiring more complex alignment and fusion strategies.

Although current multispectral object detection approaches have achieved significant progress by exploring cross-modal fusion strategies, several intrinsic limitations still remain. Modality-specific reconstruction methods (e.g., SCFR [6]) try to preserve unique information, but they often overlook redundant background features that are simultaneously present in both modalities, thereby weakening the discriminability of fused representations. Transformer-based approaches (e.g., DAMSDet [7], ICAFusion [2]) have attempted to capture global complementary information and address misalignment, but their heavy reliance

* Corresponding author.
E-mail address: shenjifeng@ujs.edu.cn (J. Shen).

https://doi.org/10.1016/j.patcog.2026.113189
Received 3 September 2025; Received in revised form 18 January 2026; Accepted 26 January 2026
Available online 29 January 2026

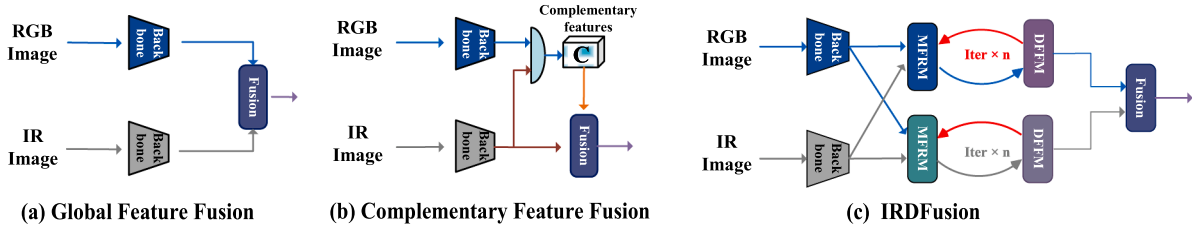

Fig. 1. Comparison of the fusion structures between our method and existing methods.

on stacked attention blocks introduces high computational burden and excessive parameterization, which restricts scalability and real-time applicability. Alignment-driven strategies (e.g., CAGT [8]) mitigate spatial misalignment at the region level, yet they are less effective in filtering modality-shared noise, leading to fused features that remain contaminated by background artifacts.

These limitations highlight a fundamental challenge: How can we preserve complementary, object-aware cross-modal structures while simultaneously suppressing modality-shared redundant background, without introducing unnecessary complexity? To better understand this challenge, we reinterpret multispectral fusion from a modeling perspective. RGB and IR features naturally exhibit both shared structural patterns (e.g., object contours, scene layout) and complementary modality-specific cues (e.g., thermal contrast, color texture). Traditional methods typically attempt to integrate these features through a single-step fusion function, implicitly assuming that complementary information can be extracted in one pass. However, this assumption rarely holds under complex imaging conditions. In practice, salient complementary cues are often weak, noisy, or partially missing in one modality; modality-shared background interference frequently dominates; and the relationship between modalities varies spatially and dynamically. We formulate multispectral fusion as an iterative refinement problem rather than a one-shot operation. Instead of executing a fixed fusion step, we view multispectral fusion as a progressive optimization process. At each iteration, the model jointly considers: the current RGB/IR feature representations, and a differential guidance signal capturing cross-modal discrepancies from the previous iteration.

This differential signal selectively highlights modality-unique, object-aware cues while suppressing common-mode background responses. Feeding this signal back into the next iteration enables the model to progressively amplify informative complementary structures and progressively filter redundant noise. This reinterpretation naturally leads to an iterative fusion framework capable of handling complex cross-modal interactions more effectively than single-pass approaches.

Motivated by this formulation, we propose IRDFusion as illustrated in Fig. 1(c), an Iterative Relation-Map Differential Guided Fusion framework that integrates two mutually reinforcing components: (1) the Mutual Feature Refinement Module (MFRM), which enhances intra- and inter-modal semantic relations, and (2) the Differential Feature Feedback Module (DFFM), which extracts, weights, and reinjects differential cues as guidance to steer the fusion process. Through repeated interaction between these two modules, IRDFusion progressively strengthens salient complementary features while suppressing shared background noise, resulting in highly discriminative and well-aligned fused representations.

We conduct extensive experiments on FLIR, LLVIP, and M3FD datasets, demonstrating that IRDFusion consistently outperforms existing methods across all metrics and challenging scenarios. Ablation studies further validate the effectiveness of both iterative refinement and differential feedback mechanisms. Beyond quantitative results, qualitative visualizations show that IRDFusion noticeably reduces both false detections and missed detections by recovering weak complementary cues that previous methods typically overlook.

In summary, our main contributions are as follows:

- A Mutual Feature Refinement Module (MFRM) is proposed to enhance modal-specific features of object candidates between two modalities, ensuring robust feature alignment.
- Inspired by the feedback differential amplifier circuits, a Differential Feature Feedback Module (DFFM) is proposed to calculate complementary discriminative features between the two modalities and simultaneously filters redundant information.
- The MFRM and DFFM are jointly optimized to effectively integrate discriminative complementary information from different modalities through a dynamic differential relationship map feedback mechanism, which provides a new strategy for progressive multispectral feature fusion.
- The proposed method IRDFusion, building on MFRM and DFFM, achieves state-of-the-art performance on the FLIR, LLVIP and M^3FD datasets.

The rest of this paper is organized as follows: Section 2 reviews related work on multispectral object detection, summarizing existing methods and their advantages and limitations; Section 3 describes the details of our proposed method, including the model architecture and key techniques; Section 4 presents experimental results, comparing the performance of our method with existing approaches; Section 5 concludes the paper and discusses future research directions.

2. Related work

2.1. Object detection

Object detection is a fundamental task in the field of computer vision, primarily categorized into one-stage and two-stage detectors. One-stage detectors, such as YOLO [9], and RetinaNet [10], perform direct regression on feature maps, achieving high detection speeds. Methods like DETR [11] further simplify the detection pipeline by directly regressing object center points or employing Transformers for end-to-end detection. In contrast, two-stage detectors, such as R-CNN [12] and FPN [13], first generate candidate regions and then perform refined classification and bounding box regression, typically achieving higher accuracy. Moreover, detection methods can be divided into anchor-based and anchor-free approaches. Anchor-based methods, such as YOLO and RetinaNet, rely on predefined anchor boxes for object prediction, while anchor-free methods, such as FCOS [14], locate object center points or boundary points directly, reducing reliance on anchor box design and lowering computational complexity. Recent improvements on the DETR framework, such as DINO [15], further enhance performance and training efficiency through contrastive denoising training and improved query selection.

Beyond detection architectures, learning robust feature representations is critical for handling challenges such as severe occlusion and scale variation. Recent advancements in related fine-grained recognition tasks have demonstrated effective strategies for this purpose. For instance, dynamic spatial interaction [16] and attentive multi-granularity perception [17] have proven beneficial for refining object features.

Furthermore, adaptive task decoupling [18] and content-adaptive occlusion handling [19] offer valuable insights for suppressing interference in crowded or occluded scenarios. In our research, we select the DETR framework due to its end-to-end training capability, simplified detection pipeline, and effective global context modeling, which enhances detection performance, especially in complex scenes.

2.2. Multispectral feature fusion for detection

Multispectral Object Detection combines RGB and thermal modalities to improve detection performance in complex scenarios. Early studies such as ConvNet [1] introduce a multispectral pedestrian dataset and significantly reduce detection errors through an ACF-based extension method. IAF R-CNN [20] incorporates an illumination-aware mechanism and a multitask learning framework, enhancing robustness under varying lighting conditions.

In terms of feature alignment and modality fusion, AR-CNN [21] proposes a region feature alignment module and features reweighting method to address weak alignment, improving multimodal fusion. Similarly, MCHE-CF [22] introduces multiscale homogeneity enhancement and confidence fusion, improving modality complementarity through a channel attention mechanism. LG-FAPF [23] improves discrimination and fusion by aggregating features with locality guidance and pixel-level fusion. MMA [24] proposes an explicit features modulation method guided by masks, significantly improving the performance of multi-task learning in object detection and box-level segmentation, especially under scale variations and occlusion. SPA and AFA [25] improve the quality of feature fusion in multispectral pedestrian detection through the scale-aware permutated attention mechanism with adjacent-branch feature aggregation module, effectively reducing the miss rate for small pedestrians.

Despite these advances, many fusion methods still face two core challenges, echoing insights from oriented object detection: effectively suppressing modality-shared background while preserving salient complementary object features, and achieving adaptive feature integration without excessive complexity. Analogous to how DFDet [26] leverages contextual priors and SFRNet [27] refines discriminative features for oriented detection, multispectral fusion can benefit from task-specific module designs that explicitly model both commonality and complementarity between modalities.

In the context of Transformer and attention-based fusion methods, CFT [28] utilizes self-attention mechanisms for inter-modality interaction, while ICAFusion [2] employs iterative cross-attention to reduce model complexity and improve performance. TFDet [29] introduce object awareness and attention mechanisms, significantly improving detection accuracy and background suppression. For redundant information suppression, RISNet [30] minimizes redundancy between RGB and IR images, optimizing complementary fusion and improving detection performance. LGADet [31] improves inference speed by utilizing a lightweight backbone and an anchor-free detection framework and employs a hybrid attention mechanism to enhance accuracy.

Finally, PIAFusion [32] proposes a progressive image fusion network based on illumination awareness, which adaptively optimizes infrared and visible light image fusion under varying lighting conditions while preserving object and texture details. GAFF [33] employs attention mechanisms to dynamically weigh and fuse multispectral features, significantly improving detection accuracy with low computational cost.

In contrast, our proposed IRDFusion model introduces a novel relational differential feedback mechanism for feature fusion. Specifically, IRDFusion first strengthens semantic information across modalities, while simultaneously emphasizing discriminative differential cues. It then extracts and feeds back inter-modal differences as guidance signals, thereby amplifying complementary object features and suppressing redundant background information. Through this iterative feedback process, IRDFusion progressively refines cross-modal alignment, leading to enhanced accuracy and robustness compared to existing fusion approaches.

3. The proposed method

3.1. Problem formulation

Let $F_v, F_t \in \mathbb{R}^{C\times H\times W}$ denote the feature maps extracted from the visible and thermal modalities. Conventional fusion approaches, such as MBNet [30] and ICAFusion [2], typically learn a static, one-shot mapping function as Eq. (1):

$$F_{\text{fused}} = \mathcal{F}(F_v, F_t) \tag{1}$$

While differential-based methods like MBNet [30] leverage differential features for modality weighting, they treat the difference computation as a terminal step, failing to re-utilize this signal to refine the input representation. Distinct from prior open-loop differential and one-shot fusion methods, IRDFusion reformulates multispectral feature fusion as an iterative closed-loop optimization process. It transforms the difference from a static weight into a dynamic guidance signal $G^{(k-1)}$, which optimizes subsequent iterations through a feedback mechanism, as formulated in Eq. (2):

$$F^{(k)}_{\text{fused}} = \mathcal{F}^{(k)}_{\text{IRDFusion}}\left(F_v, F_t, G^{(k-1)}\right), \tag{2}$$

where $G^{(k-1)}$ encodes cross-modal discrepancies. By injecting this signal back into the loop, the framework acts as a differential feedback amplifier, progressively enhancing object-aware structures while attenuating common-mode background noise.

3.2. Architecture

As shown in Fig. 2, the model first employs a dual-branch backbone to extract features from both RGB and thermal modalities, while the proposed IRDFusion module is utilized to progressively fuse cross-modal features. IRDFusion enhances feature representations by amplifying inter-modal differences and leveraging them as guidance signals to steer the fusion process step by step. The fused representations are subsequently processed by a Simple Feature Pyramid (SFP) neck [34], followed by a Transformer Encoder, and finally fed into the multiple parallel detection heads of Co-DETR. The detection head design remains consistent with Double-Co-DETR [34]. This architecture effectively integrates complementary cross-modal cues, leading to substantially improved detection performance under challenging conditions.

3.3. Mutual feature refinement module (MFRM)

The Mutual Feature Refinement Module (MFRM) is designed to enhance the feature representations between two modalities, thereby improving cross-modal consistency and discriminative capability. Its core idea is to leverage the self-attention matrix of a single modality to interact with the weighted Value features of both modalities within a Transformer structure. In this way, MFRM amplifies cross-modal representations and produces more informative fused features. Specifically, as illustrated in Fig. 3, the features of the two modalities are first projected through distinct weight matrices W to generate Query, Key, and Value matrix. These vectors are then processed through self-attention, as described in Eq. (3), to obtain the corresponding attention matrices A_i, $i \in \{v, t\}$ for each modality.

$$\begin{aligned}\left[Q_i, K_i, V_i\right] &= F_i \cdot \left[W_i^q, W_i^k, W_i^v\right] \\ A_i &= Softmax\frac{\left(Q_i \cdot K_i^T\right)}{\sqrt{d}}\end{aligned} \tag{3}$$

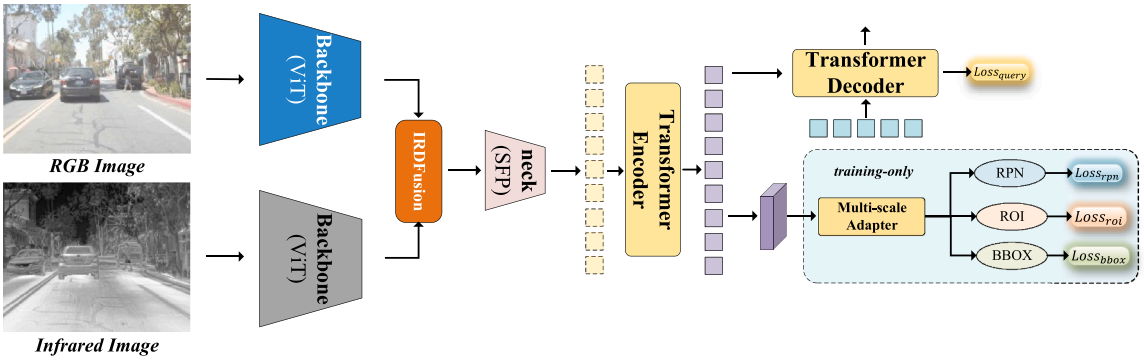

Fig. 2. The architecture of our detection pipeline.

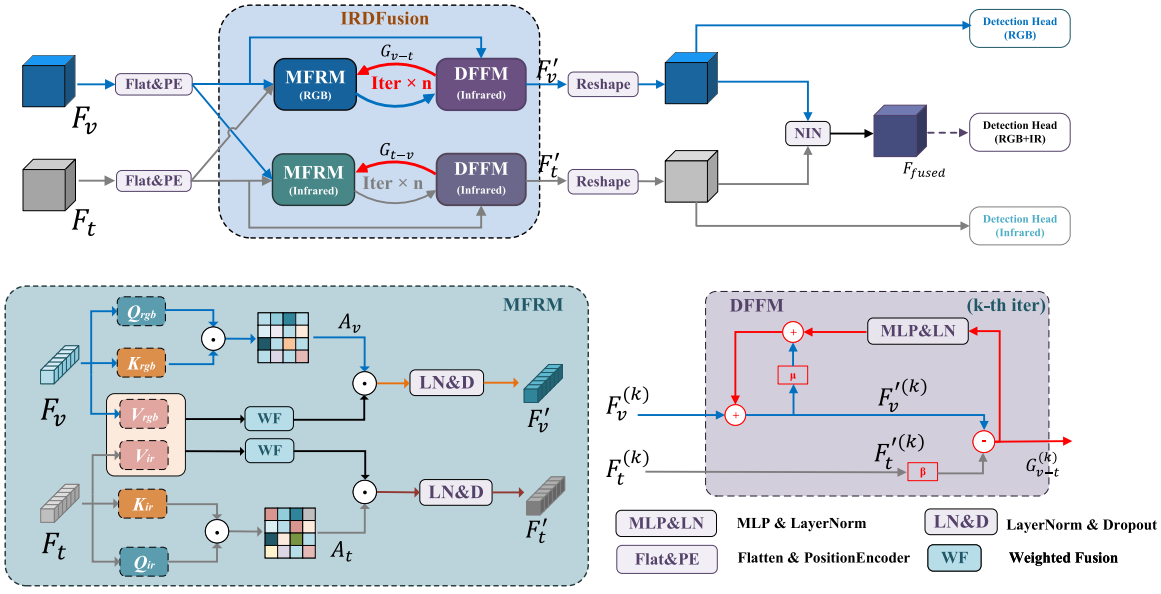

Fig. 3. Illustration of the proposed IRDFusion module. The red data stream is DFFM. In this context, $LN\&D$ refers to Layer Normalization (LN) and Dropout, $MLP\&LN$ refers to the MLP layer followed by Layer Normalization, β and μ are learnable parameters. $Flat\&PE$ refers to organizing feature maps into sequences and adding positional encoding information and *Reshape* refers to converting the sequence back into a feature map. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where F_i represents the input features of the RGB or IR modalities, respectively. Q_i, K_i, V_i denote the query, key, and value matrices, $\cdot$ represents matrix multiplication, while W_i^q, W_i^k, W_i^v are the weight matrices for linear transformation and A_i represents the attention matrix, and d represents the feature dimension.

Unlike standard cross-attention in ICAFusion [2]. ICAFusion performs cross-attention via $(Q_v \cdot K_t^\top)$, which risks structural misalignment when modalities diverge significantly. In contrast, MFRM adopts a structure-preserving strategy: it leverages the self-attention map of the primary modality $(Q_v \cdot K_v^\top)$ to aggregate the Value features of the auxiliary modality. This ensures that complementary cues are integrated without disrupting the intrinsic spatial topology of the primary modality. MFRM preserves the structural topology of the primary modality. To integrate complementary cues, we fuse the Value vectors from both branches using an adaptive weighting mechanism. This mechanism dynamically scales the fusion ratio based on input characteristics. Finally, the fused features are obtained by applying the self-attention matrix to these refined Value vectors. This process is formally expressed in Eq. (3.3):

$$\begin{aligned} F_v' &= A_v(V_v + \lambda_v V_t) \\ F_t' &= A_t(V_t + \lambda_t V_v) \end{aligned} \tag{4}$$

where F_i' is the final fused features, A_i represents the attention matrix and $i \in \{t, v\}$.

During the fusion of the Value vectors, we introduce a learnable parameter in Eq. (5), drawing inspiration from [35]. This parameter allows the model to adaptively adjust the fusion process, improving robustness by enabling the model to scale the feature fusion based on the characteristics of the input data. This adaptive mechanism contributes to improved performance and greater flexibility in feature alignment.

$$\lambda_i = \exp\left(\lambda_{q1} \cdot \lambda_{k1}\right) - \exp\left(\lambda_{q2} \cdot \lambda_{k2}\right) + \lambda_{init}, i \in \{v, t\} \tag{5}$$

where λ_v and λ_t are the fusion weights for the modalities, controlled by learnable vectors $\lambda_{q1}, \lambda_{q2}, \lambda_{k1}, \lambda_{k2}$ and the initial weight λ_{init}.

3.4. Differential feature feedback module (DFFM)

The Differential Feature Feedback Module (DFFM) introduces a closed-loop feedback mechanism to amplify inter-modal differences. Inspired by differential feedback amplifier circuits, DFFM leverages inter-modal differential features as guidance signals for dynamic cross-modal fusion. As illustrated in the lower part of Fig. 3, taking the RGB branch as an example, the differential features between the RGB and IR modalities are first computed, with a learnable parameter β introduced to adaptively control their contribution. The resulting differential features are

then weighted and fed back into the RGB features, amplifying inter-modal difference signals and guiding the MFRM in extracting discriminative cues from the other modality. Through iterative feedback, the DFFM progressively enhances complementary information while filtering redundant noise, leading to more robust and adaptive cross-modal representations. To provide an intuitive explanation of these differential features, we analyze the internal mechanism mathematically. Merely subtracting features may be physically opaque; however, by expanding the subtraction using the fusion process from the MFRM, we reveal that the differential signal is actually derived from the interaction between attention maps and value vectors as Eq. (6):

$$\begin{aligned} G_{v-t} &= F_v^{'} - \beta F_t^{'} \\ &= A_v(V_v + \lambda_v V_t) - \beta A_t(V_t + \lambda_t V_v) \\ &= (A_v - \beta\lambda_t A_t)V_v - (\beta A_t - \lambda_v A_v)V_t \\ &= C_{(v-t)}V_v - C_{(t-v)}V_t \end{aligned} \tag{6}$$

where G_{v-t} represents the difference features of the IR modality relative to the RGB modality. Based on the derivation above, we can identify the term $C_{(v-t)}$ as the Relation Map Difference. This term indicates that IRDFusion does not simply perform pixel-wise subtraction; instead, it implicitly learns a differential attention map. This map highlights specific spatial regions where the structural relationships between the two modalities diverge. Finally, this structural difference signal is processed through a multi-layer perceptron (MLP) and normalization layers as Eq. (7) before being used to refine the input features for the next iteration:

$$\begin{aligned} F_v^{(k+1)} &= \mu \cdot F_v^{(k)} + \alpha \cdot \mathrm{MLP}\left(\mathrm{LN}\left(G_{v-t}^{(k)}\right)\right) \\ F_t^{(k+1)} &= F_t \end{aligned} \tag{7}$$

where α and μ are learnable parameters. MLP and LN denote MLP layer and Layer Normalization, respectively. $F_i^{(k)}$ refers to the output features of the kth iteration of MFRM layer , while $F_i^{(k+1)}$ denotes the input features for the $(k+1)$th iteration of MFRM layer. It is worth mentioning that the thermal image feature F_t is fixed during the feature refinement for the RGB image branch.

DFFM serves as a dynamic closed-loop amplifier. Unlike static difference maps of MBNet, DFFM extracts inter-modal discrepancies and actively feeds them back as guidance signals to iteratively refine the feature representation, progressively suppressing common-mode noise while enhancing salient object details.

3.5. Loss function

In this work, we adopt the CoDetr loss, as in [34], for training. The CoDetr loss function integrates multiple components to optimize classification and localization performance. The main detection head (CoDINOHead [15]) uses Quality Focal Loss for classification, effectively addressing class imbalance issues, and uses L1 Loss and GIoU Loss for bounding box regression and localization accuracy, respectively.

In addition to the main detection head, CoDetr also includes three auxiliary detection heads. The RPN Head [36] applies Cross Entropy Loss for object-background classification and utilizes L1 Loss to refine bounding box proposals. The ROI Head [36] adopts Cross Entropy Loss for category prediction and employs GIoU Loss to improve the precision of bounding box regression. The Bbox Head [36] utilizes Focal Loss for classification, GIoU Loss for regression, and Cross-Entropy Loss for centerness prediction, contributing to improved detection accuracy.

This comprehensive loss design strikes a balance between robust classification and precise localization. The auxiliary detection heads complement the main detection head, further improving overall detection performance.

4. Experiments

4.1. Implementation details

We use the double-co-detr framework from [34] for our experiments. All experiments are carried out using PyTorch on a system equipped with an Intel i7-9700 CPU, 64 GB of RAM, and a Nvidia RTX 3090 GPU (24 GB of memory). The image input size is set to 640×640, and the data augmentation is followed by the v1 version from [34], with all other settings remaining consistent with those in the original paper. In the final experiments, the FLIR and LLVIP datasets are trained for 12 epochs, while the M^3FD dataset is trained for 36 epochs. In the ablation studies, NiNfusion is used as the baseline, and all experiments are trained for 12 epochs with a batch size of 1 with the same settings. Our code and model will be released for reproducing our results.

4.2. Dataset and evaluation metric

FLIR [37]: FLIR is a high-quality dataset for multispectral object detection, consisting of paired IR and RGB images. It is primarily used in autonomous driving and surveillance applications. FLIR includes object categories such as “person”, “car”, and “bicycle”. With detailed annotations, it is particularly suitable for research on cross-modal fusion between infrared and visible light modalities.

LLVIP [38]: LLVIP focuses on pedestrian detection in low-light conditions. LLVIP contains paired RGB and IR images and is specifically designed to address the challenges of multispectral pedestrian detection in low-light scenarios. It is widely used for research on multimodal data fusion methods in low-illumination environments.

M3FD [39]: It includes 4200 infrared and visible light aligned image pairs collected under various environments such as different lighting, seasons, and weather scenarios. It encompasses six typical categories in automated driving and road surveillance. M^3FD is partitioned according to an 8:2 training/testing split as provided in [40].

Mean Average Precision (mAP): The COCO mAP (mean average precision) evaluation metric is a standard for assessing object detection models. The mean AP is calculated across all classes and IoU thresholds (from 0.5 to 0.95, in increments of 0.05), offering a comprehensive evaluation of detection accuracy and localization precision. mAP50 measures AP at IoU = 0.5, focusing on sufficient overlap, while mAP75 evaluates AP at IoU = 0.75, requiring stricter localization.

4.3. State-of-the-art comparison

4.3.1. Comparison on the FLIR dataset

As shown in Table 1, our method achieves the best mAP50 while maintaining competitive mAP75, improving the baseline by 3.5% and surpassing DAMSDet by 1.7%, demonstrating its superior detection capability at lower IoU thresholds. For mAP75 and overall mAP, although slightly below the best-performing method, our approach still improves the baseline by 4.0% and 3.8%, respectively, indicating robust performance across stricter evaluation criteria. These gains stem from the core design of IRDFusion: the MFRM reinforces semantic features while suppressing redundant background, and the DFFM dynamically extracts and feeds back inter-modal differences to guide fusion. This iterative feedback progressively amplifies complementary object information and filters common-mode noise, enhancing both cross-modal alignment and discriminative power.

4.3.2. Comparison on the LLVIP dataset

As shown in Table 2, our method also achieves the best performance across all three metrics, with improvements of 0.4%, 2.4%, 1.4% in mAP50, mAP75 and overall mAP compared to the baseline, and gains of 0.5%, 4.0%, 1.3% over DAMSDet [7]. These improvements demonstrate the effectiveness of IRDFusion in pedestrian detection: by

Table 1
Comparison on the FLIR dataset. The bold text represents the best result, and the underlined text represents the second best result. '-' is used to indicate papers that have been published on arXiv but have not been officially published yet.

Methods	Year	Source	mAP50	mAP75	mAP
GAFF [33]	2021	CVPR	72.9	32.9	36.6
CFT [28]	2021	–	78.7	35.5	40.2
MFPT [41]	2023	TITS	80.0	–	–
LRAF-Net [42]	2023	TNNLS	80.5	–	42.8
ICAFusion [2]	2024	PR	79.2	36.9	41.4
MMFN [43]	2024	TCSVT	80.8	–	–
RSDet [44]	2024	–	81.1	–	41.4
UniRGB-IR [45]	2024	–	81.4	40.2	44.1
YOLOXCPCF [46]	2024	TITS	82.1	41.2	44.6
SCFR [6]	2024	TITS	82.3	35.7	–
GM_DETR [47]	2024	CVPR	83.9	42.6	45.8
MMPedestron [48]	2024	ECCV	86.4	–	–
DAMSDet [7]	2024	ECCV	<u>86.6</u>	**48.1**	<u>49.3</u>
Fusion-Mamba [49]	2025	TMM	84.9	45.9	47.0
Baseline	2025	–	84.8	44.0	46.9
Ours	2025	–	**88.3**	<u>48.0</u>	**50.7**

Table 2
Comparison on the LLVIP dataset. The bold text represents the best result, and the underlined text represents the second best result.

Methods	Year	Source	mAP50	mAP75	mAP
CFT [28]	2021	–	97.5	72.9	63.6
CSAA [50]	2023	CVPR	94.3	66.6	59.2
LRAF-Net [42]	2023	TNNLS	97.9	–	66.3
UniRGB-IR [45]	2024	–	96.1	72.2	63.2
YOLOXCPCF [46]	2024	TITS	96.4	75.4	65.2
MMFN [43]	2024	TCSVT	97.2	–	–
GM_DETR [47]	2024	CVPR	97.4	<u>81.4</u>	<u>70.2</u>
SCFR [6]	2024	TITS	97.5	–	–
MS_DETR [51]	2024	TITS	97.9	76.3	66.1
DAMSDet [7]	2024	ECCV	97.9	79.1	69.6
ICAFusion [2]	2024	PR	**98.4**	76.2	64.5
Fusion-Mamba [49]	2025	TMM	97.0	–	64.3
Baseline	2025	–	<u>98.0</u>	80.7	69.5
Ours	2025	–	**98.4**	**83.1**	**70.9**

Table 3
Comparison on the M^3FD dataset. The bold text represents the best result, and the underlined text represents the second best result.

Methods	Year	Source	mAP50	mAP
TarDAL [39]	2022	CVPR	80.5	54.1
CDDFusion [52]	2023	CVPR	81.1	54.3
IGNet [53]	2023	MM	81.5	54.5
DAMSDet [7]	2024	ECCV	80.2	52.9
MMFN [43]	2024	TCSVT	86.2	–
ICAFusion [2]	2024	PR	**90.8**	<u>60.9</u>
Fusion-Mamba [49]	2025	TMM	<u>88.0</u>	**61.9**
Baseline	2025	–	87.1	58.2
Ours	2025	–	**90.8**	**61.9**

reinforcing semantic structures and leveraging differential cues, the framework enhances feature alignment and preserves discriminative information, particularly under challenging conditions such as occlusion or crowding, leading to more accurate detection.

4.3.3. Comparison on the M^3FD dataset

As presented in Table 3, our method attains the highest performance across all metrics, matching the best model in overall mAP while achieving a 2.8% improvement in mAP50. Compared to the baseline, both metrics increase by 3.7%. This improvement reflects the capability of IRDFusion to progressively refine cross-modal features: the MFRM consolidates cross-modal information, and the DFFM emphasizes complementary differences, enabling the model to effectively suppress noise and enhance discriminative cues, which is particularly beneficial in complex multi-spectral scenarios.

4.4. Ablation studies

4.4.1. Different modules

We conducted ablation studies using CoDetr with NiNfusion as the baseline (first row in Table 4) to evaluate the contributions of the proposed modules.

MFRM Module: Introducing the MFRM module led to substantial improvements in all metrics for the "bicycle" and "person" categories, with the most pronounced gains observed for "bicycle". This suggests that features corresponding to "bicycle" are not prominent in either modality individually; by amplifying and fusing complementary features from both modalities, MFRM significantly enhances their representation, thereby improving detection performance. Notably, the growth rate for the "bicycle" category is the largest among all classes.

While the MFRM significantly improves overall detection, the "car" category exhibits a specific decline in high-IoU (e.g. MAP75) metrics. Detailed analysis reveals a trade-off inherent to the MFRM design. The module aggregates global contextual information for reinforcing semantic consistency. While this effectively highlights the semantic regions of objects, it inadvertently induces a "feature over-smoothing" effect. As visualized in Fig. 4, this aggregation process effectively acts as a low-pass filter. For rigid objects like cars that require sharp boundary definitions, high-frequency edge details are blurred by fusing them with surrounding background textures, such as shadows or metallic reflections in RGB images. This results in spatial ambiguity at object boundaries, where the feature gradient becomes indistinct compared to the raw input. Consequently, while the regression head can correctly classify the object, it struggles to achieve the pixel-perfect localization required for mAP75. Despite this localized drawback, the overall metrics (mAP50, mAP75, and mAP) increase by 1.5%, 2.0%, and 1.7% respectively, validating the overall efficacy of the object detection performance under the DFFM-guided MFRM collaborative optimization framework.

DFFM Module: Upon incorporating the DFFM module, all metrics substantial gains overall, despite slight declines in specific categories, confirming that differential feature feedback effectively guides the fusion of relevant cross-modal features. The "bicycle" category again shows the largest gains, with mAP50, mAP75 and mAP increasing by 9.9%, 6.2%, 8.8%, respectively, while the "person" category improves by 0.3%, 8.6%, 4.1%. These results indicate that DFFM is particularly effective for objects that are visually blurred or weakly represented in both modalities, as it selectively reinforces critical features while suppressing redundant information.

MFRM + DFFM module: In summary, the ablation study clearly demonstrates that both MFRM and DFFM modules play complementary and crucial roles in enhancing cross-modal feature fusion. MFRM primarily amplifies and consolidates weak but informative features, while DFFM guides the fusion process to emphasize relevant features and suppress interference, together achieving significant improvements in detection performance across challenging categories. The combination of two modules leads to overall improvements of 3.5%, 4.0%, 3.8% across mAP50, mAP75 and mAP respectively.

4.4.2. Iteration number

We have conducted ablation experiments to investigate the effect of the number of iterations in the DFFM module, with results summarized in Table 5. Overall, the model achieves optimal performance at the fourth iteration in terms of mAP50. Although certain categories reach their peak performance at different iteration counts, these variations are minor, supporting the selection of four iterations as the standard setting for all experiments.

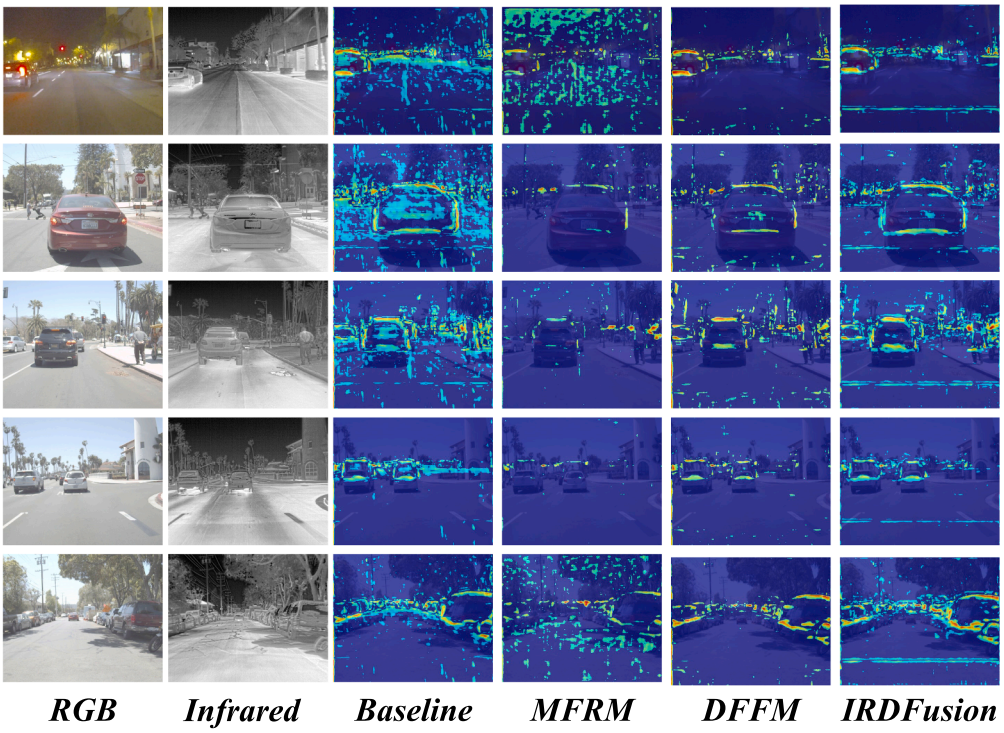

Fig. 4. Visual comparison of feature representations generated by different modules. The RGB and Infrared columns show the source modalities. The MFRM column exhibits strong internal activation for semantic consistency but results in blurred object boundaries due to feature over-smoothing from context aggregation. In contrast, the DFFM column captures sharp high-frequency differential cues, effectively preserving object contours. By synergizing the semantic strength of MFRM with the boundary precision of DFFM, the IRDFusion framework (last column) compensates for spatial ambiguity, achieving both precise object localization and robust background suppression.

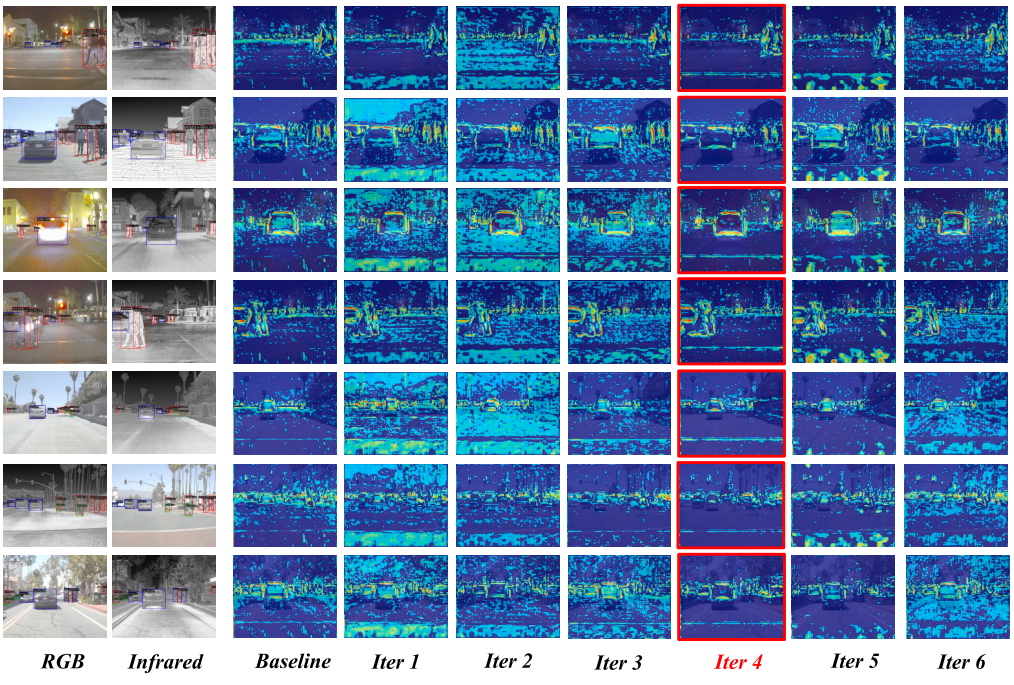

Fig. 5. Visualization of different iterations of IRDFusion. The red markers indicate the optimal iteration number. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Effects of different modules.

MFRM	DFFM	mAP50				mAP75				mAP			
		person	car	bicycle	all	person	car	bicycle	all	person	car	bicycle	all
		89.1	92.4	72.9	84.8	46.2	70.4	15.5	44.0	48.5	62.5	28.7	46.9
✓		89.4 (+0.3)	92.2 (−0.2)	77.3 (+4.4)	86.3 (+1.5)	48.1 (+1.9)	66.5 (−3.9)	23.2 (+7.7)	46.0 (+2.0)	49.3 (+0.8)	61.2 (−1.3)	35.4 (+6.7)	48.6 (+1.7)
	✓	89.7 (+0.6)	93.2 (+1.0)	79.5 (+6.6)	87.5 (+2.7)	48.1 (+1.9)	70.7 (+0.3)	19.2 (+3.7)	46.0 (+2.0)	49.5 (+1.0)	64.2 (+1.7)	33.9 (+5.2)	49.2 (+2.3)
✓	✓	89.4 (+0.3)	92.6 (+0.2)	82.8 (+9.9)	88.3 (+3.5)	54.8 (+8.6)	67.6 (−2.8)	21.7 (+6.2)	48.0 (+4.0)	52.6 (+4.1)	62.2 (-0.3)	37.5 (+8.8)	50.7 (+3.8)

Table 5
Effects of different iteration numbers.

Iter Num	mAP50				mAP75				mAP			
	person	car	bicycle	all	person	car	bicycle	all	person	car	bicycle	all
Baseline	89.1	92.4	72.9	84.8	46.2	70.4	15.5	44	48.5	62.5	28.7	46.9
1	89.0	92.8	74.5	85.4	45.5	69.5	12.9	42.6	48.1	63.2	28.4	46.6
2	89.0	**93.0**	75.3	85.8	48.2	**70.7**	15.7	44.9	49.3	63.8	30.9	48.0
3	88.8	**93.0**	77.2	86.3	46.3	70.6	18.0	45	48.2	**63.9**	32.1	48.1
4	**89.4**	92.6	**82.8**	**88.3**	**54.8**	67.6	**21.7**	**48.0**	**52.6**	62.2	**37.5**	**50.7**
5	86.9	91.3	80.4	86.2	46.4	66.6	19.9	44.3	47.4	61.2	35.2	47.9
6	88.8	91.7	76.5	85.7	48.8	67.9	17.2	44.7	49.6	62.3	31.1	47.7

We also provide the visualization of fused feature maps of different iterations in Fig. 5. It is clear to observe that, when the number of iterations is too low, some redundant features are not fully suppressed. Limited feedback on differential features at lower iteration prevents effective elimination of redundant or noisy information within both modalities, reducing the model's ability to integrate meaningful cross-modal features. By the fourth iteration, the DFFM module sufficiently removes irrelevant information while maximizing the fusion of critical features, resulting in optimal detection performance.

Conversely, we also find that further increasing the number of iterations leads to performance degradation. With excessive iterations, the differential features become progressively weakened, and interactions between modalities can introduce adverse effects. In particular, the amplification of background noise increases, which interferes with accurate feature integration. We attribute this decline to the model overemphasizing differential features at the expense of effectively consolidating features, thereby undermining overall fusion quality.

In summary, this ablation study demonstrates that an appropriate number of iterations is crucial for balancing the suppression of irrelevant features and the integration of meaningful cross-modal information. The fourth iteration achieves this balance, ensuring robust and effective performance across categories.

4.4.3. Ablation study of single-branch detection

We also conducted ablation experiments to evaluate the contributions of each branch after integrating the IRDFusion module, as shown in Table 6. In the RGB branch, incorporating infrared features through IRDFusion leads to improvements across all metrics compared to the baseline that relies solely on RGB features. Similarly, in the IR branch, fusing RGB features via IRDFusion yields notable gains in most metrics. These results demonstrate that both single-modality branches benefit significantly from cross-modal feature integration, with the independent IR branch generally outperforms the RGB branch.

For specific categories such as "bicycle," the improvements are particularly pronounced. This is because the features of "bicycle" are relatively indistinct in either RGB or IR modalities alone, and only through cross-modal fusion can salient object cues be sufficiently enhanced for reliable detection. Conversely, for certain "person" and "bicycle" instances, slight decreases in mAP75 and overall mAP are observed, likely due to interference from background clutter in the RGB modality, which may introduce minor noise into the fused features.

Finally, integrating IRDFusion across both branches achieves the best overall performance, confirming that joint fusion of RGB and IR modalities produces the most robust results. However, for the "car" category, a minor decline in mAP75 and overall mAP is observed, even compared to the IR branch alone. This is attributed to relatively clear car contours in the IR modality but blurred RGB features with nearby irrelevant background, causing the fusion to introduce slight misalignment and reduce IoU. Despite these localized declines, their overall impact on the model's performance is negligible, underscoring the effectiveness of IRDFusion in enhancing cross-modal feature representation through iterative differential guidance.

4.5. Different detection frameworks

We further evaluated the adaptability of the proposed IRDFusion module by integrating it into different detection frameworks. First, in the anchor-based YOLOv5 framework, with NiNfusion as the baseline, experiments were conducted on the FLIR, LLVIP, and M^3FD datasets, as shown in Table 7. Our method achieves consistent improvements across all datasets and metrics. Specifically, on FLIR, all three metrics increase by over 3%; on LLVIP, the improvement in mAP75 is particularly notable, indicating enhanced precision in pedestrian localization; and on M^3FD, while the gains are smaller, there is still measurable improvement. Although YOLOv5 is designed for real-time applications and emphasizes speed, the integration of IRDFusion demonstrates that its feature fusion mechanism can still provide significant accuracy gains.

Similarly, in the CoDETR framework, a DETR variant, IRDFusion achieves substantial improvements, particularly on FLIR. On LLVIP, the most pronounced gain is observed in mAP75, suggesting that IRDFusion effectively enhances high-precision localization of pedestrians. On M^3FD, mAP50 shows a notable increase, with slight improvements in mAP75 and overall mAP. These results indicate that the

Table 6
The individual detection results of different branches.

Method	Output	mAP50				mAP75				mAP			
		person	car	bicycle	all	person	car	bicycle	all	person	car	bicycle	all
Baseline	RGB	76.8	86.3	61.0	74.7	25.4	51.9	10.7	29.3	34.3	50.9	22.7	36.0
	IR	88.9	93.0	70.4	84.1	53.0	68.6	15.1	45.6	51.8	62.9	28.2	47.6
	RGB + IR	89.1	92.4	72.9	84.8	46.2	70.4	15.5	44.0	48.5	62.5	28.7	46.9
IRDFusion	RGB	78.5 (+1.7)	87.1 (+0.8)	66.2 (+5.2)	77.2 (+2.5)	29.4 (+4.0)	55.5 (+3.6)	14.4 (+3.7)	33.1 (+3.8)	37 (+2.7)	53.3 (+2.4)	25.9 (+3.2)	38.7 (+2.7)
	IR	89.5 (+0.6)	93.0 (+0.0)	78.2 (+7.8)	86.9 (+2.8)	52.4 (−0.6)	70.3 (+1.7)	14.7 −(0.4)	45.8 (+0.2)	51.7 (−0.1)	63.5 (+0.6)	29.9 (+1.7)	48.4 (+0.8)
	RGB + IR	89.4 (+0.3)	92.6 (+0.2)	82.8 (+9.9)	88.3 (+3.5)	54.8 (+8.6)	67.6 (−2.8)	21.7 (+6.2)	48.0 (+4.0)	52.6 (+4.1)	62.2 (−0.3)	37.5 (+8.8)	50.7 (+3.8)

Table 7
Comparison on different detection frameworks.

Detector	Backbone	Method	FLIR			LLVIP			M3FD		
			mAP50	mAP75	mAP	mAP50	mAP75	mAP	mAP50	mAP75	mAP
YOLOv5	CSPDarknet	Baseline	79.9	34.6	40.2	96.8	71.2	62.7	89.1	65.8	59.8
		Ours	84.8	38.1	43.4	97.9	75.7	65.5	89.8	65.8	60.5
CoDetr	ViT	Baseline	84.8	44.0	46.9	98.0	80.7	69.5	87.1	61.1	58.2
		Ours	88.3	48.0	50.7	98.4	83.1	70.9	90.8	65.4	61.9

Table 8
Comparison with different attention methods.

	mAP50				mAP75				mAP			
	person	car	bicycle	all	person	car	bicycle	all	person	car	bicycle	all
Self-Attention	88.6	91.9	79.4	86.6	49.9	67.2	31.4	49.5	50.2	62.0	38.5	50.2
Cross-Attention	90.4	93.4	76.5	86.8	51.6	71.9	22.4	48.6	51.3	65.0	34.7	50.3
IRDFusion	89.4	92.6	82.8	88.3	54.8	67.6	21.7	48.0	52.6	62.2	37.5	50.7

iterative feedback mechanism of IRDFusion, through MFRM and DFFM, consistently strengthens cross-modal feature alignment and discriminability, regardless of the underlying detection architecture.

Collectively, these experiments demonstrate that IRDFusion is highly generalizable and robust across different frameworks and datasets. As a plug-and-play module, it can be seamlessly integrated into both speed-oriented frameworks like YOLOv5 and performance-oriented frameworks like CoDETR, consistently enhancing detection performance through adaptive cross-modal fusion.

4.6. Comparison of alternative methods of replacement

To further validate the effectiveness of the proposed fusion design, we conducted comparative experiments by replacing the MFRM modules with alternative mechanisms, including Cross-Attention in ICAFusion [2] and standard self-attention [54] in Transformer. The results in Table 8 indicate that, although these alternatives offer certain advantages, they still underperform compared to IRDFusion. Specifically, When fed back into the MFRM, differential information requires both intra-modal consistency and cross-modal alignment. Using only self-attention or cross-attention fails to simultaneously satisfy these two requirements, leading to suboptimal performance. In contrast, IRDFusion combines MFRM and DFFM to simultaneously reinforce cross-modal features and dynamically extract differential cues. The iterative feedback mechanism amplifies complementary object information while suppressing redundant noise, leading to more stable, discriminative, and well-aligned feature representations. These results demonstrate that the unique design of IRDFusion is crucial for achieving superior cross-modal fusion performance compared to conventional attention-based alternatives.

4.7. Visualization

As shown in Fig. 6, the baseline method employs a global feature fusion strategy, which exhibits notable limitations in detection tasks. Specifically, false detections (red triangles) appear in columns 1, 2, 4, 5, 6, and 7, while missed detections (pink triangles) are observed in columns 3 and 8. These errors arise because the baseline fails to differentiate between cross-modal and differential features across modalities, relying instead on coarse global fusion. Such a strategy inadequately integrates complementary cues, leading to the omission or degradation of critical object information and, consequently, reduced overall detection performance.

In contrast, our proposed IRDFusion method explicitly separates cross-modal and differential features, and leverages both types in the fusion process. This fine-grained approach effectively suppresses redundant background information while amplifying complementary cues, significantly reducing both false and missed detections. The visualization results thus illustrate the capability of IRDFusion to produce more precise and discriminative cross-modal representations, further validating the importance of iterative feature fusion in enhancing multispectral object detection performance.

4.8. Failure case

As illustrated in Fig. 7, IRDFusion exhibits limitations when detecting small or heavily occluded objects. Mechanism-level analysis reveals that these failures stem from specific breakdowns in the relation modeling process.

For small objects, the limitation is primarily due to “Relation Map Degeneration.” In deep feature maps with low spatial resolution, small targets occupy very few pixels. This causes the attention mechanism

Fig. 6. Sample detections of our method on FLIR, LLVIP, and M^3FD datasets. The first and second rows are the input RGB and IR images, the third row is the detection result of the baseline, and the fourth row is the detection result of our method. The red triangle markers indicate false detections of the baseline, while the pink triangle markers indicate missed detections of the baseline. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

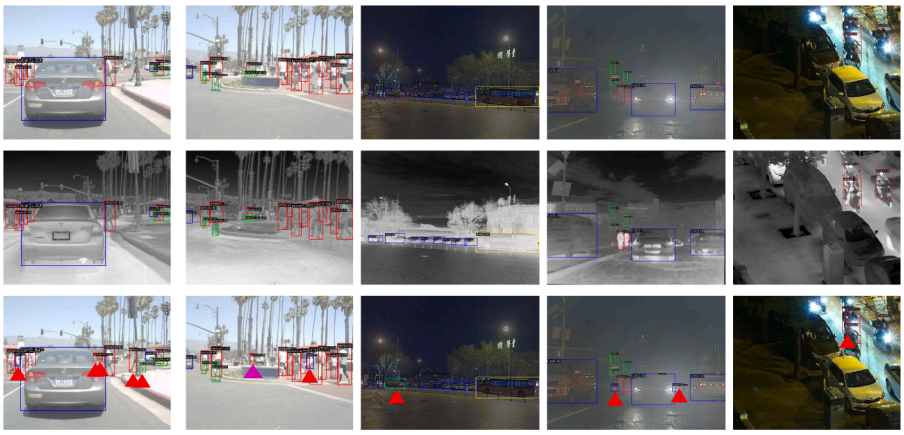

Fig. 7. False detections of our method on FLIR, LLVIP, and M^3FD datasets. The first and second rows are the input RGB and IR images, and the third row is the detection result of our method. The red triangle markers indicate false detections of the IRDFusion, while the pink triangle markers indicate missed detections of the IRDFusion. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 9
Speed comparison.

Detector	Method	FPS	params(M)	GFLOP
YOLOv5	Baseline	45.3	75.4	91.0
	Ours	17.0	134.0	122.8
CoDetr	Baseline	3.6	485.3	944.9
	Ours	3.1	510.5	1213.5

to suffer from over-smoothing, where the relation map loses structural distinctiveness. Consequently, the differential signal extracted by DFFM degrades into ineffective noise rather than meaningful guidance.

For heavily occluded objects, the challenge arises from "Structural Topology Mismatch." Since MFRM utilizes the self-attention map of the primary modality to aggregate features, if the primary view is dominated by an occluder (e.g., a tree in RGB), it incorrectly forces the visible target features in the auxiliary modality (e.g., a person in IR) to align with the occluder's structure. This misalignment leads to the suppression of valid target features. Future work will address these issues by exploring multi-scale and deformable attention mechanisms to preserve details and decouple target features from occlusion.

4.9. Limitations

According to the results in Table 9, it is evident that the introduction of IRDFusion leads to a trade-off between accuracy and efficiency. Specifically, the parameter count increases by approximately 60M in the YOLO framework and 25M in the CoDetr framework, accompanied by a notable reduction in inference speed (FPS). This increase in computational complexity is primarily attributed to the iterative refinement processes within the MFRM and DFFM.

While these metrics indicate a relatively high deployment risk for resource-constrained edge devices, the potential for practical application remains significant. First, in safety-critical scenarios such as autonomous driving at night or surveillance under adverse weather, the substantial gain in detection robustness often outweighs the latency penalty. Second, from a deployment perspective, the current IRDFusion model can serve as a high-performance "teacher" network. Its robust cross-modal representations can be transferred to lightweight "student" models via knowledge distillation, or optimized using network pruning and TensorRT acceleration. These strategies would effectively mitigate the computational burden, making the proposed method feasible for real-time applications in future iterations.

5. Conclusion

In this paper, we presented IRDFusion, a novel multispectral object detection framework that progressively integrates cross-modal features via an iterative feedback mechanism. By synergizing the Mutual Feature Refinement Module (MFRM) for structure-preserving alignment and the Differential Feature Feedback Module (DFFM) for dynamic difference guidance, our approach effectively amplifies salient object signals while suppressing common-mode background noise. Extensive experiments on

FLIR, LLVIP, and M^3FD datasets demonstrate that IRDFusion achieves state-of-the-art performance, showing exceptional robustness in challenging scenarios such as low illumination and complex backgrounds.

Despite these significant performance gains, we acknowledge certain limitations. First, the detection of extremely small or heavily occluded objects remains a bottleneck due to potential feature over-smoothing in attention mechanisms. Second, the iterative nature of the framework introduces higher computational complexity and parameter overhead compared to one-shot fusion methods, which currently poses challenges for real-time deployment on resource-constrained edge devices.

Future work will focus on addressing these challenges to bridge the gap between academic research and industrial application. To tackle the deployment issue, we plan to explore model compression techniques, such as knowledge distillation and network pruning, to significantly reduce computational costs while retaining the benefits of the iterative fusion strategy. Additionally, we aim to incorporate multi-scale feature enhancement and dynamic attention mechanisms to further improve detection performance for small targets and under severe occlusion.

CRediT authorship contribution statement

Jifeng Shen: Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Investigation, Funding acquisition, Conceptualization; **Haibo Zhan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Xin Zuo:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization; **Heng Fan:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis, Conceptualization; **Xiaohui Yuan:** Writing – review & editing, Supervision, Methodology, Formal analysis; **Jun Li:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis, Conceptualization; **Wankou Yang:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization.

Data availability

Data will be made available on request.

Declaration of competing interest

Please check the following as appropriate: All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version. This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue. The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61903164, 62173186 and in part by Natural Science Foundation of Jiangsu Province in China under Grants BK20191427 and the Key R&D Program of Zhejiang Province (2024C04056(CSJ)). Heng Fan receives no financial support for the research, authorship, and/or publication of this article.

References

[1] S.W. Jingjing Liu, Shaoting Zhang, D. Metaxas, Multispectral deep neural networks for pedestrian detection, In: Proceedings of the British Machine Vision Conference (BMVC), 2016, pp. 73.1–73.13.

[2] J. Shen, Y. Chen, Y. Liu, X. Zuo, H. Fan, W. Yang, ICAFusion: iterative cross-attention guided feature fusion for multispectral object detection, Pattern Recognit. 145 (2024) 109913. https://doi.org/10.1016/j.patcog.2023.109913

[3] J. Shen, H. Zhan, S. Dong, X. Zuo, W. Yang, H. Ling, Multispectral state-space feature fusion: bridging shared and cross-parametric interactions for object detection, Inform. Fusion (2025) 103895. https://doi.org/10.1016/j.inffus.2024.103895

[4] C. Hu, H. Zheng, K. Li, J. Xu, W. Mao, M. Luo, L. Wang, M. Chen, Q. Peng, K. Liu, et al., FusionFormer: a multi-sensory fusion in bird's-eye-view and temporal consistent transformer for 3D object detection, arXiv:2309.05257 (2023).

[5] Y. Li, A.W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q.V. Le, et al., Deepfusion: LiDAR-camera deep fusion for multi-modal 3D object detection, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17182–17191.

[6] X. Sun, Y. Zhu, H. Huang, Specificity-guided cross-modal feature reconstruction for RGB-infrared object detection, IEEE Trans. Intell. Transp. Syst. 26 (2024) 950–961.

[7] J. Guo, C. Gao, F. Liu, D. Meng, X. Gao, DAMSDet: dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion, In: Proceedings of the European Conference on Computer Vision, 2024, pp. 464–481.

[8] M. Yuan, X. Shi, N. Wang, Y. Wang, X. Wei, Improving RGB-infrared object detection with cascade alignment-guided transformer, Inform. Fusion 105 (2024) 102246. https://doi.org/10.1016/j.inffus.2023.102246

[9] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, In: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, In: Proceedings of the European Conference on Computer Vision, 2020, pp. 213–229.

[12] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[14] Z. Tian, C. Shen, H. Chen, T. He, FCOS: a simple and strong anchor-free object detector, IEEE Trans. Pattern Anal. Mach. Intell. 44 (4) (2020) 1922–1933.

[15] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, H.-Y. Shum, DINO: DETR with improved denoising anchor boxes for end-to-end object detection, In: The Eleventh International Conference on Learning Representations, 2023, pp. 1–23.

[16] Q. Zhang, D. Miao, Q. Zhang, C. Zhao, H. Zhang, Y. Sun, R. Wang, Dynamic frequency selection and spatial interaction fusion for robust person search, Inform. Fusion (2025) 103314.

[17] Q. Zhang, J. Wu, D. Miao, C. Zhao, Q. Zhang, Attentive multi-granularity perception network for person search, Inf. Sci. 681 (2024) 121191.

[18] Q. Zhang, D. Miao, Q. Zhang, C. Wang, Y. Li, H. Zhang, C. Zhao, Learning adaptive shift and task decoupling for discriminative one-step person search, Knowl. Based Syst. 304 (2024) 112483.

[19] C. Zhao, Z. Qu, X. Jiang, Y. Tu, X. Bai, Content-adaptive auto-occlusion network for occluded person re-identification, IEEE Trans. Image Process. 32 (2023) 4223–4236.

[20] C. Li, D. Song, R. Tong, M. Tang, Illumination-aware faster R-CNN for robust multispectral pedestrian detection, Pattern Recognit. 85 (2019) 161–171.

[21] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, Z. Liu, Weakly aligned cross-modal learning for multispectral pedestrian detection, In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5127–5137.

[22] R. Li, J. Xiang, F. Sun, Y. Yuan, L. Yuan, S. Gou, Multiscale cross-modal homogeneity enhancement and confidence-aware fusion for multispectral pedestrian detection, IEEE Trans. Multimed. 26 (2023) 852–863.

[23] Y. Cao, X. Luo, J. Yang, Y. Cao, M. Y. Yang, Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection, Inform. Fusion 88 (2022) 1–11.

[24] J. Shen, Y. Liu, Y. Chen, X. Zuo, J. Li, W. Yang, Mask-guided explicit feature modulation for multispectral pedestrian detection, Comput. Electr. Eng. 103 (2022) 108385.

[25] X. Zuo, Z. Wang, J. Shen, W. Yang, Improving multispectral pedestrian detection with scale-aware permutation attention and adjacent feature aggregation, IET Comput. Vis. 17 (7) (2023) 726–738.

[26] X. Xie, G. Cheng, C. Rao, C. Lang, J. Han, Oriented object detection via contextual dependence mining and penalty-incentive allocation, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–10.

[27] G. Cheng, Q. Li, G. Wang, X. Xie, L. Min, J. Han, SFRNet: fine-grained oriented object recognition via separate feature refinement, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–10.

[28] Q. Fang, D. Han, Z. Wang, Cross-Modality Fusion Transformer for Multispectral Object Detection, arXiv:2111.00273 (2021).

[29] X. Zhang, X. Zhang, J. Wang, J. Ying, Z. Sheng, H. Yu, C. Li, H.-L. Shen, TFDET: target-aware fusion for RGB-T pedestrian detection, IEEE Trans. Neural Netw. Learn. Syst. 36 (2024) 13276–13290.

[30] K. Zhou, L. Chen, X. Cao, Improving multispectral pedestrian detection by addressing modality imbalance problems, In: Proceedings of the European Conference on Computer Vision, 2020, pp. 787–803.

[31] X. Zuo, Z. Wang, Y. Liu, J. Shen, H. Wang, LGADet: light-weight anchor-free multispectral pedestrian detection with mixed local and global attention, Neural Process. Lett. 55 (3) (2023) 2935–2952.

[32] L. Tang, J. Yuan, H. Zhang, X. Jiang, J. Ma, PIAFusion: a progressive infrared and visible image fusion network based on illumination aware, Inform. Fusion 83 (2022) 79–92.

[33] H. Zhang, E. Fromont, S. Lefèvre, B. Avignon, Guided attentive feature fusion for multispectral pedestrian detection, In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 72–80.
[34] C. Zhou, P. Cheng, J. Fang, Y. Zhang, Y. Yan, X. Jia, Y. Xu, K. Wang, X. Cao, Optimizing multispectral object detection: a bag of tricks and comprehensive benchmarks, arXiv:2411.18288 (2024).
[35] T. Ye, L. Dong, Y. Xia, Y. Sun, Y. Zhu, G. Huang, F. Wei, Differential transformer, In: International Conference on Representation Learning, Vol. 2025, 2025, pp. 144–164.
[36] R. Girshick, Fast R-CNN, In: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[37] Free teledyne flir thermal dataset for algorithm training, https://www.flir.com/oem/adas/adas-dataset-form/, (2021).
[38] X. Jia, C. Zhu, M. Li, W. Tang, W. Zhou, LLVIP: a visible-infrared paired dataset for low-light vision, In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3496–3504.
[39] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, Z. Luo, Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5802–5811.
[40] M. Liang, J. Hu, C. Bao, H. Feng, D. Fuqin, T.L. Lam, Explicit attention-enhanced fusion for RGB-thermal perception tasks, IEEE Rob. Autom. Lett. 8 (2023) 1–8.
[41] Y. Zhu, X. Sun, M. Wang, H. Huang, Multi-modal feature pyramid transformer for RGB-infrared object detection, IEEE Trans. Intell. Transp. Syst. 24 (9) (2023) 9984–9995.
[42] H. Fu, S. Wang, P. Duan, C. Xiao, R. Dian, S. Li, Z. Li, LRAF-Net: long-range attention fusion network for visible–infrared object detection, IEEE Trans. Neural Netw. Learn. Syst. 35 (10) (2023) 13232–13245.
[43] F. Yang, B. Liang, W. Li, J. Zhang, Multidimensional fusion network for multispectral object detection, IEEE Trans. Circuits Syst. Video Technol. 35 (2024) 547–560.
[44] T. Zhao, M. Yuan, F. Jiang, N. Wang, X. Wei, Removal and Selection: Improving RGB-Infrared Object Detection via Coarse-to-Fine Fusion, arXiv:2401.10731 (2024).
[45] M. Yuan, B. Cui, T. Zhao, J. Wang, S. Fu, X. Yang, X. Wei, UNIRGB-IR: a unified framework for visible-infrared semantic tasks via adapter tuning, In: Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 2409–2418.
[46] S. Hu, F. Bonardi, S. Bouchafa, H. Prendinger, D. Sidibé, Rethinking self-attention for multispectral object detection, IEEE Trans. Intell. Transp. Syst. 25 (11) (2024) 16300–16311.
[47] Y. Xiao, F. Meng, Q. Wu, L. Xu, M. He, H. Li, GM-DETR: generalized muiltispectral detection transformer with efficient fusion encoder for visible-infrared detection, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 5541–5549.
[48] Y. Zhang, W. Zeng, S. Jin, C. Qian, P. Luo, W. Liu, When pedestrian detection meets multi-modal learning: generalist model and benchmark dataset, In: Proceedings of the European Conference on Computer Vision, Springer, 2024, pp. 430–448.
[49] W. Dong, H. Zhu, S. Lin, X. Luo, Y. Shen, G. Guo, B. Zhang, Fusion-Mamba for cross-modality object detection, IEEE Trans. Multimed. 27 (2025) 1–15.
[50] Y. Cao, J. Bin, J. Hamari, E. Blasch, Z. Liu, Multimodal object detection by channel switching and spatial attention, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 403–411.
[51] Y. Xing, S. Yang, S. Wang, S. Zhang, G. Liang, X. Zhang, Y. Zhang, MS-DETR: multispectral pedestrian detection transformer with loosely coupled fusion and modality-balanced optimization, IEEE Trans. Intell. Transp. Syst. (2024) 20628–20642.
[52] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, L. Van Gool, CDDFuse: correlation-driven dual-branch feature decomposition for multi-modality image fusion, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5906–5916.
[53] J. Li, J. Chen, J. Liu, H. Ma, Learning a graph neural network with cross modality interaction for image fusion, In: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 4471–4479.
[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 1–11.