Redundancy-Guided Hierarchical Fusion for RGBT Object Detection

Yidan Sun[1], Yujie Wu[2], Qiming Yang [3, (✉)]

[1] Institute of Deep-sea Science and Engineering, Chinese Academy of Sciences
[2] School of Mathematical Sciences, School of Mathematical Sciences Harbin Engineering University
[3] Institute of Software, Chinese Academy of Sciences
yangqiming@iscas.ac.cn

Abstract. In dual-branch RGBT object detection, symmetric architectures are widely adopted without examining whether different feature levels exhibit distinct redundancy patterns. In this paper, we first analyze the weight sparsity of a dual-branch model and observe that redundancy increases with depth, and at deeper layers the RGB branch becomes substantially more redundant than the thermal branch. Based on this observation, we propose a redundancy-guided hierarchical fusion strategy (Red-HiFusion)—feature concatenation at shallow low-redundancy levels, bidirectional cross-attention at middle medium-redundancy levels, and unidirectional IR-query attention at deep asymmetric high-redundancy levels to allow the less redundant thermal features guide the more redundant RGB features. Red-HiFusion achieves 85.4% $mAP50$ on M^{3}FD and 97.4% $mAP50$ on LLVIP with only 14.9 M parameters and 35.0 GFLOPs. Extensive ablations show that this hierarchical design consistently outperforms full-bidirectional and full-unidirectional baselines while adding negligible computation. Compared to state-of-the-art approaches, our model attains competitive detection accuracy and has an order of magnitude fewer parameters, offering an interpretable and efficient solution for lightweight RGBT detection.

Keywords: RGBT Object Detection, Feature Fusion, Attention Mechanism.

1 Introduction

Object detection is a fundamental task in computer vision with widespread applications including autonomous driving, intelligent surveillance, robotics, and human–computer interaction [1,2,3]. In recent years, deep-learning-based detectors have achieved remarkable success on standard benchmarks, largely driven by the rich appearance cues (color, texture, edge) available in visible images [4,5,6,7].

However, the performance of visible-based detectors degrades severely under challenging illumination and weather conditions, where visible images lose contrast and become noisy [1,8]. This inherent limitation of the visible spectrum has motivated the use of complementary modalities. Thermal infrared (IR) images, which capture emitted thermal radiation, are insensitive to lighting variations and can provide reliable object

silhouettes even in complete darkness [9]. By combining RGB and IR, RGB-Thermal infrared (RGBT) fusion has become a critical technique for robust perception under adverse conditions such as nighttime, as RGB imaging alone often fails in these scenarios [10].

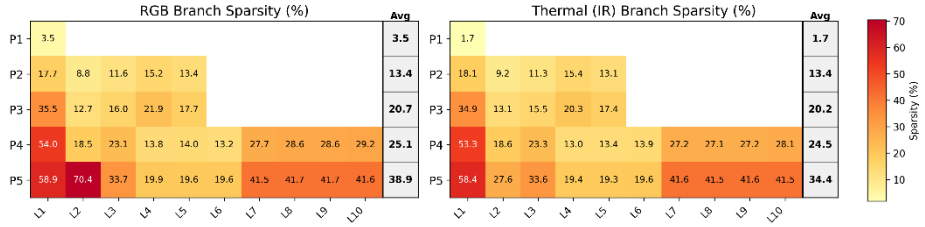

Fig. 1. Sparsity heatmaps of RGB and IR branches across P1–P5 levels. Each row is a feature level; columns are convolutional layers (left) and per-level average sparsity (rightmost gray column).

To achieve effective RGBT object detection, extensive research has been conducted. Early works such as Halfway Fusion [11] demonstrated the effectiveness of middle-level fusion. Since then, numerous dual-branch networks have been proposed, and this architecture has become the dominant paradigm in RGBT detection [12,13,14,15,16]. However, most of these designs implicitly assume that all layers in both branches contribute equally and that no computational waste exists. Model compression studies [17] have revealed that different layers contribute unequally to the final output: layers with low redundancy are more informative and critical to performance, while highly redundant layers contribute little. This raises an important yet under-explored question: *Do the RGB and IR branches share the same contribution pattern across different levels of the network?* And if not, can we exploit this asymmetry to design more efficient and effective fusion?

To answer this question, we conduct a systematic redundancy analysis on a trained baseline dual-branch model using convolution weight sparsity—defined as the proportion of weights whose magnitude falls below a fixed small threshold—as the redundancy indicator. Weight sparsity has been extensively studied in pruning literature as a practical measure of model redundancy [18], and recent comprehensive surveys have further consolidated it as a core concept in sparsity-based model compression [19]. Magnitude-based pruning methods operate under the principle that weights with small magnitude contribute little to network output and can be safely pruned [18]; consequently, a layer with high sparsity contains many such low-magnitude parameters and is considered a candidate for pruning.

Fig. 1 reports the sparsity of all convolutional layers in each branch across *P1–P5* levels, visualized through sparsity heatmaps where lighter yellow indicates lower redundancy and darker red indicates higher redundancy. The rightmost column (gray background) shows the average sparsity per level for each branch. From **Fig. 1** we derive two key findings. First, redundancy increases with depth. For both branches, the average sparsity rises monotonically from *P1* to *P5* (e.g., RGB: from 3.5% to 38.9%; IR: from 1.7% to 34.4%). The heatmaps also exhibit a clear color shift from light yellow to deep red at deeper levels. This suggests that deeper layers contain more parameters

that are not fully utilized. Second, at the deepest level (*P5*), the RGB branch exhibits consistently higher redundancy than the IR branch. As shown in the *P5* row of **Fig. 1**, the average sparsity of RGB (38.9%, rightmost column) is higher than that of IR (34.4%). Notably, the highest-sparsity layer in the RGB branch reaches 70.4%, while the highest in the IR branch is only 58.4%. This indicates that thermal information in the deep layers remains relatively more active, while certain RGB parameters become nearly zero. The observed asymmetry in layer-wise redundancy suggests that a one-size-fits-all compression or fusion strategy may be suboptimal.

Based on these observations, we propose a **redundancy-guided hierarchical fusion strategy (Red-HiFusion)** with three principles: low-redundancy shallow layers use only concatenation, without dedicated attention modules; similar-redundancy middle layers adopt bidirectional cross-attention; asymmetric-redundancy deep layers employ unidirectional fusion where the less redundant thermal branch serves as the Query to guide the more redundant RGB branch.

The main contributions of this work are threefold:

- We reveal the hierarchical redundancy pattern in dual-branch RGBT detection networks through weight sparsity analysis, discovering that the RGB branch at deep layers is significantly more redundant than the IR branch across multiple datasets.
- We propose Red-HiFusion, a redundancy-guided design principle that prescribes different fusion strategies (none, bidirectional, or unidirectional with less-redundant as Query) according to layer-wise redundancy characteristics.
- We validate the approach with comprehensive ablations, achieving a +4.9% *mAP50* gain over the baseline on M^3FD with only +0.7 M additional parameters, demonstrating an optimal accuracy-efficiency trade-off.

2 Related Work

2.1 RGBT Object Detection

RGBT object detection leverages visible and thermal infrared images to achieve robust perception under adverse conditions such as nighttime [10]. Early work by Liu *et al.* [11] proposed Halfway Fusion, a dual-stream architecture that performs feature fusion at middle-level convolutional layers and outperforms early or late fusion strategies. Subsequent dual-branch architectures have become dominant, where two parallel backbones extract features followed by cross-modal fusion modules. IRDFusion [20] introduces an iterative differential feedback mechanism but applies uniform fusion across all feature levels. DAMSDet [21] adopts a Transformer with modality competitive query selection, yet still treats shallow and deep layers identically. TFDet [22] employs target-aware fusion, and ProbEn [23] provides a probabilistic late-fusion perspective. A common limitation of these methods is the implicit assumption that all layers contribute equally; none have systematically examined whether the RGB and IR branches exhibit different redundancy patterns at different depths.

2.2 Cross-Attention for RGBT Fusion Strategies

Cross-attention mechanisms have been widely used in multi-modal fusion due to their ability to capture fine-grained complementary cues. RDCRNet [24] leverages cross-attention for feature alignment. IFCAF [25] extends fusion from discrete grids to continuous implicit representations using implicit neural representations. CFRNet [26] refines fused features with thermal information via a cross-attention module. Despite their effectiveness, existing cross-attention approaches deploy attention uniformly across all pyramid levels and fix the query direction arbitrarily, leading to unnecessary parameters and suboptimal fusion.

2.3 Redundancy Analysis

Model compression studies have shown that deep convolutional networks contain substantial redundancy. Magnitude-based pruning [18] and structured pruning methods such as channel pruning [27] and ThiNet [28] have demonstrated that removing low-magnitude or low-importance channels can accelerate networks with minimal accuracy loss. However, most redundancy analysis focuses on pruning single-modality models for efficiency, rather than exploiting redundancy patterns to guide cross-modal fusion design. The question of whether RGB and IR branches exhibit systematically different redundancy characteristics has not been explored in RGBT detection.

3 Method

3.1 Redundancy-Guided Hierarchical Fusion Architecture

Based on the redundancy analysis presented in Sec.1, we propose a hierarchical fusion architecture that allocates fusion capacity according to layer-wise redundancy characteristics. The overall architecture, illustrated in **Fig. 2**, consists of a dual-branch backbone network for feature extraction, followed by hierarchical fusion modules that are selectively applied at different feature pyramid levels.

For the backbone, we adopt a dual-branch YOLO11 architecture [1413]. Both the RGB branch $\mathcal{F}_{RGB}$and the IR branch $\mathcal{F}_{IR}$ extract multi-scale features $\{P3, P4, P5\}$ from their respective inputs, where P_k denotes features at the k th level of the feature pyramid. The detection head then processes the fused features to produce final detection results.

The key novelty of our architecture lies in the hierarchical, redundancy-guided fusion strategy applied to *P3, P4*, and *P5*. According to the diagnostic insights, we configure each level as follows:

- *P3* (low redundancy, both branches): only use concatenation as a simple fusion, without dedicated attention modules. Because shallow layers already have low sparsity and are efficiently utilized; adding fusion modules would introduce unnecessary computation and potentially interfere with already optimized features.

- *P4* (similar moderate redundancy, both branches): adopt bidirectional fusion (RGB↔IR). Since both modalities exhibit comparable sparsity, full cross-modal information exchange is beneficial.
- *P5* (asymmetric redundancy, RGB more redundant): employ IR-Query unidirectional fusion (IR→RGB). Because the IR branch is less redundant and maintains stronger discriminative signal, it should serve as the Query to guide the more redundant RGB branch.

This design ensures that fusion resources are spent precisely where they are most beneficial, while avoiding wasteful computation in layers that are already efficiently utilized. Note that *P1* and *P2* features are not fed into the detection head in our YOLO11-based architecture and thus do not require explicit fusion modules.

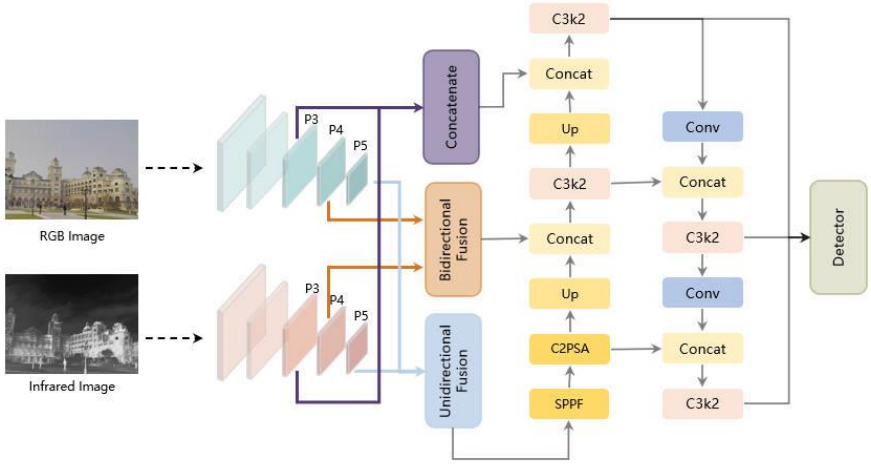

Fig. 2. Overview of the proposed Red-HiFusion architecture.

3.2 Axial-Dynamic Network

We design three types of cross-modal attention modules, whose structures are illustrated in **Fig. 3**. Their placements correspond to the hierarchical strategy defined above. Below we detail each module.

IR-Query Unidirectional Fusion: This module is used at *P5*, where the RGB branch is more redundant. The goal is to let the less redundant thermal features guide the RGB features. **Fig. 3** (a) illustrates its internal structure. The input consists of two feature maps: RGB features $F_{RGB} \in R^{C\times H\times W}$ and IR features $F_{IR} \in R^{C\times H\times W}$.

Step 1: Sequence preparation and normalization. Both feature maps are flattened into sequences of length $N = H \times \mathrm{W}$ and layer-normalized:

$$r = LayerNorm\big(Flatten(F_{RGB})\big), \tag{1}$$

$$i = LayerNorm\big(Flatten(F_{IR})\big), \tag{2}$$

where r, i $\in R^{N\times C}$.

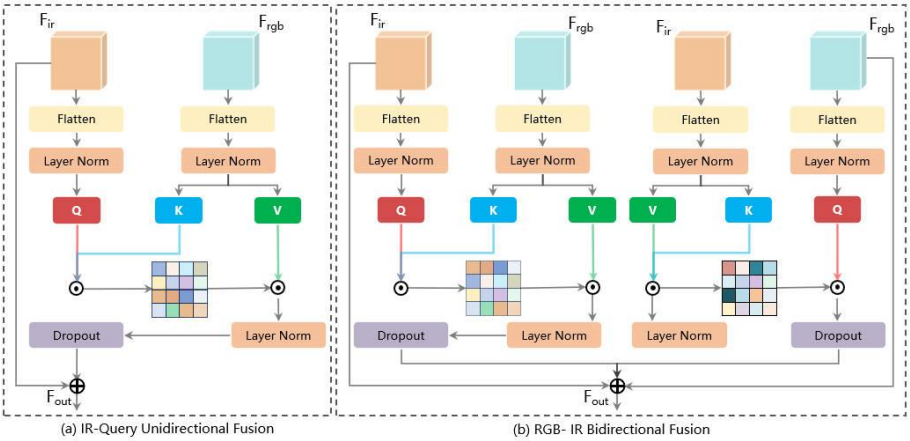

Fig. 3. Detailed architecture of unidirectional and bidirectional fusion module. (a) IR-Query unidirectional fusion module. (b) RGB- IR bidirectional fusion module.

Step 2: Linear projections for cross-attention. We use multi-head attention with h heads. For a single head (head dimension $d = \frac{C}{h}$), the query, key, and value are:

$$Q = iW_Q, K = rW_K, V = rW_V, \tag{3}$$

where $W_Q, W_K, W_V \in R^{C\times C}$ are learnable matrices. The IR branch provides the query, and the RGB branch provides key and value.

Step 3: Scaled dot-product attention. The attention weight matrix is

$$\tilde{F} = Attn \cdot V = Softmax\left(\frac{Q\cdot K^T}{\sqrt{d}}\right)\cdot V. \tag{4}$$

Step 4: Output projection and gated residual connection. The sequence $\tilde{F}$ is reshaped to $\tilde{F}_{reshape} \in R^{C\times H\times W}$, then passed through a linear layer and a dropout layer:

$$\hat{F} = Dropout\left(Linear\left(Reshape(\tilde{F})\right)\right). \tag{5}$$

Finally, a learnable gated residual adds a scaled copy of the original IR features:

$$F_{out} = \hat{F} + \gamma \cdot F_{IR}. \tag{6}$$

where γ is a trainable scalar (initialized to 0.5). This residual preserves the original IR features while the attention output $\hat{F}$ encodes complementary cross-modal information guided by IR queries.

RGB-Query Unidirectional Fusion: The symmetric module where RGB acts as the query and IR as key/value is defined analogously. It is used only in ablation studies to verify the importance of the query direction. Its description is omitted for brevity.

RGB-IR Bidirectional Fusion: At *P4*, both branches have similar moderate redundancy, making full bidirectional exchange beneficial. The RGB-IR bidirectional fusion module (as shown in **Fig. 3** (b)) takes the same input as RGB features $F_{RGB} \in R^{C\times H\times W}$ and IR features $F_{IR} \in R^{C\times H\times W}$. Its processing consists of five stages.

Step 1: Sequence preparation and normalization. Both feature maps are flattened into sequences of length $N = H \times \mathrm{W}$ and layer-normalized exactly as in IR-Query unidirectional fusion module.

Step 2: Two parallel cross-attention branches. For IR-Query branch, the IR modality is used as the query to attend over the RGB modality, which supplies the key and value.

$$Q_{IR\to RGB} = iW_{Q_{IR\to RGB}}, K_{IR\to RGB} = rW_{K_{IR\to RGB}}, V_{IR\to RGB} = rW_{V_{IR\to RGB}}, \tag{7}$$

$$\tilde{F}_{IR\to RGB} = Softmax\left(\frac{Q_{IR\to RGB}\cdot K_{IR\to RGB}^{T}}{\sqrt{d}}\right)\cdot V_{IR\to RGB}. \tag{8}$$

And for RGB-Query branch, the RGB features act as the query, and the IR features provide both the key and the value.

$$Q_{RGB\to IR} = iW_{Q_{RGB\to IR}}, K_{RGB\to IR} = rW_{K_{RGB\to IR}}, V_{RGB\to IR} = rW_{V_{RGB\to IR}}, \tag{9}$$

$$\tilde{F}_{RGB\to IR} = Softmax\left(\frac{Q_{RGB\to IR}\cdot K_{RGB\to IR}^{T}}{\sqrt{d}}\right)\cdot V_{RGB\to IR}. \tag{10}$$

$\tilde{F}_{IR\to RGB}$ and $\tilde{F}_{RGB\to IR}$ are reshaped to $\tilde{F}_{IR\to RGB,reshape}, \tilde{F}_{RGB\to IR,reshape\in}R^{C\times H\times W}$ respectively, then passed through its own output projection with dropout:

$$F_{IR\to RGB} = Dropout\left(Linear\left(\tilde{F}_{IR\to RGB,reshape}\right)\right), \tag{11}$$

$$F_{RGB\to IR} = Dropout\left(Linear\left(\tilde{F}_{RGB\to IR,reshape\in}\right)\right). \tag{12}$$

Step 3: Learnable weighted fusion. The two directional results are combined using a learnable gate vector (α, β):

$$F_{Bi} = \alpha \cdot F_{IR\to RGB} + \beta \cdot F_{RGB\to IR}, \tag{13}$$

where α and β are trainable scalars (initialized to 0.5 each).

Step 4: Residual connection. Both original RGB and IR features are added to the fused result:

$$F_{res} = F_{Bi} + F_{RGB} + F_{IR}. \tag{14}$$

Step 5: Final projection. A $Conv_{1\times 1}$ produce the final output:

$$F_{out} = Conv_{1\times 1}(F_{res}), \tag{15}$$

with output channel dimension C.

The RGB-IR bidirectional fusion module uses separate parameters for the two attention directions, enabling asymmetric transformations. The learnable fusion gate balances the two directions, and the dual residual preserves unidirectional information.

Empirically, this design outperforms naive concatenation or single-direction fusion at the middle level.

4 Experiments

4.1 Experimental Settings

We evaluate our method on two widely used RGBT object detection benchmarks.

M^{3}FD [29]: A challenging multi-spectral dataset containing daytime, night, and low-light scenes. It contains 4,200 aligned RGBT image pairs captured in campus and road scenes, with six classes (person, vehicle, bus, motorcycle, lamp post, truck) and a total of 33,603 annotated labels. Since there is no official dataset partitioning method, we use train/test splits provided by [30].

LLVIP [31]: This dataset includes 15488 paired RGB and IR images captured in very dark scenes, with pedestrians labeled as the primary category. It is specifically designed for low-light vision applications, making it an ideal benchmark for evaluating robustness under challenging illumination conditions. We follow the official training and testing partition to ensure a fair comparison with prior works.

We report average precision at IoU threshold 0.5 ($mAP50$) and mean average precision (mAP) computed over IoU thresholds from 0.5 to 0.95 at 0.05 intervals. For model efficiency, we report parameter count (Params) and Giga Floating-Point Operations (GFLOPs).

For experimental configuration, all input images are uniformly resized to 640 × 640. We adopt the SGD optimizer with a batch size of 16, and employ the cosine annealing strategy for learning rate adjustment. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU.

4.2 Comparisons with State-of-the-art Detectors

We compare our proposed method with existing RGBT detection methods on the M^{3}FD and LLVIP datasets. **Table 1** presents the quantitative results on M^{3}FD, while **Table 2** presents results on LLVIP.

On M^{3}FD, Red-HiFusion achieves 85.4% $mAP50$ and 59.3% mAP. Specifically, Red-HiFusion significantly outperforms SuperFusion (+1.9%), IGNet (+3.9%), DAMSDet (+5.2%), and TFDet (+20.6%), demonstrating the effectiveness of redundancy-guided hierarchical design. Moreover, it achieves performance is comparable to Fusion-Mamba (85.4% vs. 85.0% $mAP50$) and approaches that of IRDFusion (85.4% vs. 89.9%), albeit with a dramatically lower computational cost—requiring only 11.1% of IRDFusion's parameters and 28.5% of its GFLOPs. Additionally, Red-HiFusion attains a higher mAP (59.3%) than Fusion-Mamba (57.5%), indicating superior localization quality at stricter IoU thresholds.

Table 1. The comparison results with state-of-the-art methods on M3FD dataset. The best results are in bold. The second best are double-underlined, and the third best are in underlined.

Method	Params (M)	GFLOPs	mAP50	mAP
SuperFusion	-	-	83.5	56.0
IGNet	-	-	81.5	54.5
DAMSDet	-	-	80.2	52.9
MMFN	176.4	-	<u>86.2</u>	-
TFDet	-	-	64.8	41.0
Fusion-Mamba	244.6	-	85.0	<u>57.5</u>
IRDFusion	134.0	122.8	**89.9**	**59.8**
Red-HiFusion (ours)	14.9	35.0	<u>85.4</u>	<u>59.3</u>

On LLVIP, our method achieves 97.4% $mAP50$ and 66.8% mAP. The mAP metric—which averages precision over IoU thresholds from 0.5 to 0.95—is higher than all compared methods, including IRDFusion (65.5%) and Fusion-Mamba (62.8%). This suggests that our hierarchical fusion not only improves detection recall but also produces more precise bounding box localization.

Table 2. The comparison results with state-of-the-art methods on LLVIP dataset. The best results are in bold. The second best are double-underlined, and the third best are in underlined.

Method	Params (M)	GFLOPs	mAP50	mAP
Halfway fusion	-	-	91.4	55.1
GAFF	-	-	94.0	55.8
ProbEn	-	-	93.4	51.5
CSAA	-	-	94.3	59.2
MMFN	176.4	-	<u>97.2</u>	-
TFDet	-	-	96.0	59.4
RSDet	68.5	136.0	95.8	61.3
Fusion-Mamba	244.6	-	96.8	<u>62.8</u>
IRDFusion	134.0	122.8	**97.9**	<u>65.5</u>
Red-HiFusion (ours)	14.9	35.0	<u>97.4</u>	**66.8**

Quality analysis. Why does our method achieve superior mAP despite slightly lower $mAP50$ than IRDFusion on M^3FD? The mAP metric penalizes inaccurate bounding boxes more heavily than $mAP50$. Our hierarchical design—particularly the shallow-layer concatenation (preserving fine spatial details) and deep-layer IR-query attention (improving object-level feature discrimination)—leads to better localization.

Efficiency analysis. With only 14.9 M parameters and 35.0 GFLOPs, our method is an order of magnitude smaller than IRDFusion (134.0 M and 122.8 GFLOPs), Fusion-Mamba (244.6 M), and RSDet (68.5 M). This efficiency stems from our redundancy

analysis: by avoiding unnecessary fusion at low-redundancy shallow layers and using lightweight unidirectional attention at deep layers, we eliminate the heavy parameters of bidirectional modules where they are least beneficial.

4.3 Qualitative Analysis

To visually demonstrate the advantage of Red-HiFusion, we present detection results on representative challenging samples from M^3FD and LLVIP in **Fig. 4**.

For M^3FD, the first row shows a rainy scene with overlapping pedestrians. Both competing methods produce multiple false-positive boxes on the same person, while Red-HiFusion outputs a single accurate box per person. This false-positive reduction is attributed to the deep-layer IR-query attention, where the less redundant thermal branch suppresses the noisy RGB responses that cause duplicate detections. In the second row (a rainy scene featuring a lamp), both competing methods miss the lamp, whereas Red-HiFusion detects it correctly, benefiting from the thermal modality's robustness under poor illumination. The third row presents a sunny scene where the baseline gives a loose box not covering the feet, and SuperFusion incorrectly detects a single person as multiple targets; Red-HiFusion provides a tight and correct box. The loose box of the baseline reflects the lack of fine spatial detail in uniform fusion; by contrast, our shallow-layer concatenation preserves high-resolution cues, leading to accurate localization. In the fourth row, a sunny scene with partial occlusion by bushes, the two comparison methods generate duplicate boxes on the partially occluded person, while Red-HiFusion produces a clean single detection. This again benefits from the IR-query mechanism, which leverages thermal information to disambiguate occluded structures. The fifth row shows a sunny scene with distant small persons. The baseline has both missed or false detections, SuperFusion misses several targets, and Red-HiFusion misses only two, demonstrating superior recall for tiny objects. The recall improvement is largely due to the bidirectional attention at the middle level, which exchanges complementary features between modalities, enhancing the representation of small-scale targets.

For LLVIP, the sixth row is a nighttime scene with a pedestrian and a car. The baseline misses the pedestrian, and ProbEn falsely detects the car as a person, while Red-HiFusion correctly identifies both. The thermal-branch-guided attention at deep levels reduces false positives on non-pedestrian objects, while the shallow concatenation retains enough detail to capture the pedestrian. In the seventh row, a nighttime scene with overlapping pedestrians, the baseline misses one person, ProbEn produces three bounding boxes for two people, and Red-HiFusion outputs exactly two accurate boxes, demonstrating its ability to resolve adjacent persons and avoid false positives. The bidirectional fusion at the middle level enables cross-modal information exchange that clarifies spatial boundaries between nearby individuals, while the deep IR-query attention prevents the model from generating redundant boxes.

Overall, Red-HiFusion consistently reduces missed or false detections across diverse adverse conditions. The hierarchical design—shallow concatenation preserving details, middle bidirectional attention enabling feature exchange, and deep unidirectional

IR-query suppressing noise—directly explains the observed qualitative improvements and aligns with the quantitative results.

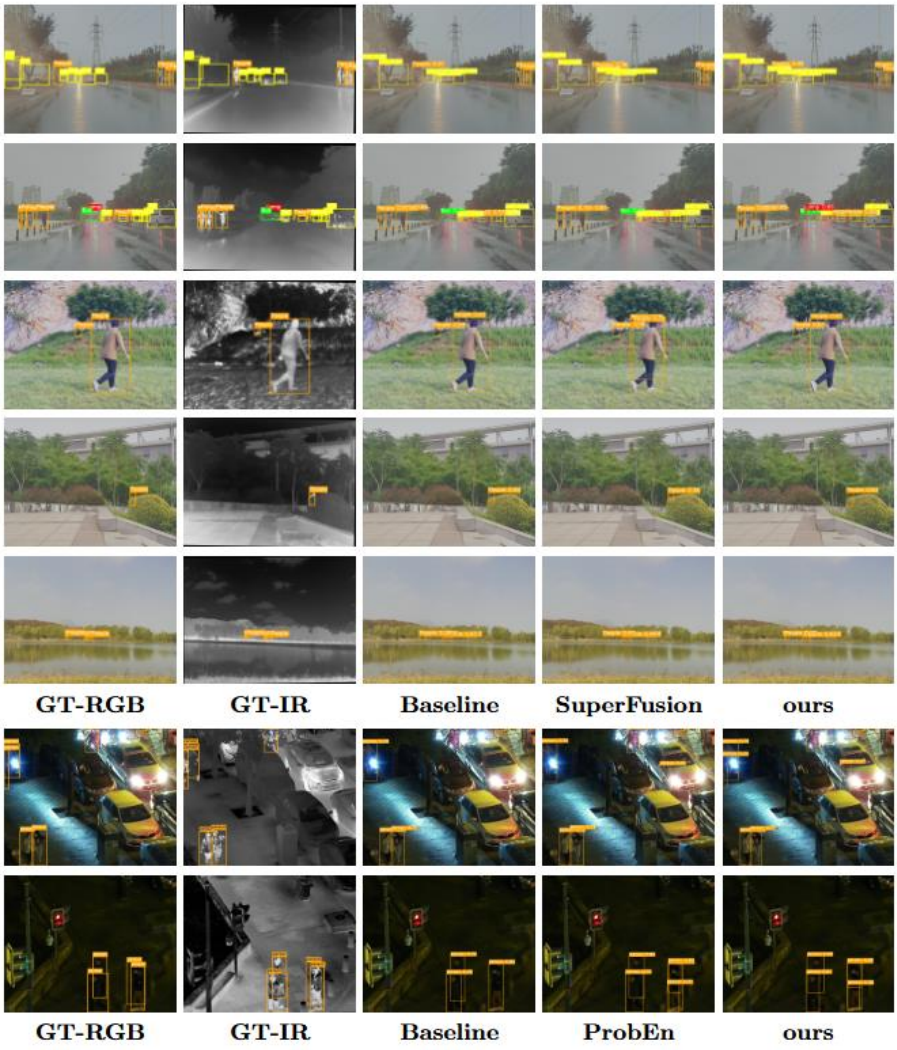

Fig. 4. Qualitative comparison on challenging scenes.

4.4 Ablation Study

To validate our design choices, we conduct comprehensive ablation experiments on both datasets. **Table 3** presents the results on M^3FD and LLVIP with nine configurations covering different fusion assignments at $P3$, $P4$, and $P5$.

Table 3. Ablation study on M3FD and LLVIP. "–" denotes concatenation only. "RGB / IR" indicate unidirectional attention with the specified branch as query. "Bi" denotes bidirectional attention. Best results in bold.

No.	*P3*	*P4*	*P5*	LLVIP	M^3FD	Params(M)
				mAP50	mAP50	
A1	-	-	-	95.0	80.5	14.2
A2	-	RGB	IR	96.7	83.8	14.6
A3	-	IR	IR	96.0	84.6	14.6
A4	IR	IR	-	96.8	80.6	14.7
A5	IR	IR	IR	94.0	79.2	14.8
A6	RGB	RGB	RGB	96.1	82.1	14.8
A7	-	Bi	RGB	97.0	84.0	14.9
A8	Bi	Bi	-	96.5	85.2	15.3
A9	Bi	Bi	IR	97.1	84.3	15.4
A10	Bi	Bi	RGB	96.6	83.8	15.4
A11	-	RGB	Bi	97.0	84.5	15.9
A12	-	IR	Bi	96.1	83.7	15.9
A13	-	Bi	Bi	96.9	83.6	16.2
A14	Bi	Bi	Bi	97.0	84.0	16.8
Red-HiFusion	-	Bi	IR	**97.4**	**85.4**	**14.9**

The baseline (A1) is a simple dual-stream architecture. It achieves 95.0% $mAP50$ on LLVIP and 80.5% on M^3FD with 14.2 M parameters. Adding a single unidirectional attention module at $P4$ or $P5$ while keeping other levels as concatenation improves performance consistently. For instance, configuration A2 ($P4$ with RGB-query, $P5$ with IR-query) raises M^3FD $mAP50$ to 83.8%, a gain of 3.3%. Comparing the two $P5$-only variants, IR-query (A3) gives 84.6% $mAP50$ on M^3FD, outperforming RGB-query (A2, 83.8%). This confirms the earlier observation that at deep layers the less redundant thermal branch serves as a better query. However, extending attention to the shallow $P3$ level (A4 and A5) degrades performance: M^3FD $mAP50$ drops to 80.6% and 79.2%, even slightly below the baseline. This validates that low-redundancy shallow layers do not benefit from complex attention mechanisms.

Next, the use of bidirectional fusion (Bi) at various levels is examined. Using Bi at $P4$ while keeping $P5$ unidirectional (A7 and ours) generally yields strong results. Applying Bi at both $P4$ and $P5$ (A13) increases parameters to 16.2M but gives an M^3FD $mAP50$ of 83.6%, which is lower than the optimal. Adding Bi also at $P3$ (A8–10, A14) further raises parameter counts without consistent accuracy gains, supporting the hierarchical principle.

The optimal configuration is ours: $P3$ concatenation, $P4$ bidirectional fusion, and $P5$ IR-query unidirectional fusion. It achieves the highest $mAP50$ on both datasets — 97.4% on LLVIP and 85.4% on M^3FD — with only 14.9 M parameters, just 0.7 M more than the baseline. This design exactly follows the redundancy-guided rule:

shallow layers with low redundancy use simple concatenation; middle layers with similar redundancy adopt bidirectional exchange; deep layers with asymmetric redundancy, where RGB is more redundant, employ unidirectional attention with IR as the query. Finally, a comparison between ours and the fully bidirectional variant (A14) shows that ours saves 1.9 M parameters while delivering +1.4% $mAP50$ on M^3FD and +0.4% $mAP50$ on LLVIP. This demonstrates that matching the fusion strategy to layer-wise redundancy is both more parameter-efficient and more accurate than applying a uniform bidirectional fusion across all levels.

5 Conclusion

In this paper, we have presented a redundancy-guided hierarchical fusion architecture for dual-branch RGBT object detection. Our starting point was a simple but previously unexplored question: Do symmetric dual-branch designs contain systematic redundancy that could be exploited for more efficient and effective fusion? Through systematic weight sparsity analysis, we discovered that redundancy increases monotonically with depth, and crucially, at deep layers the RGB branch becomes substantially more redundant than the IR branch. Based on this diagnosis, we proposed a design principle: low-redundancy shallow layers use only concatenation; similar-redundancy middle layers adopt bidirectional fusion; asymmetric-redundancy deep layers employ unidirectional fusion with the less redundant modality as the Query.

Red-HiFusion achieves competitive detection accuracy on M3FD and LLVIP benchmarks with an order of magnitude fewer parameters than most existing methods. The hierarchical design consistently outperforms full-bidirectional and full-unidirectional baselines while adding negligible computation, offering an interpretable and efficient solution for lightweight RGBT detection. Future directions include extending the analysis to other backbones, developing dynamic redundancy estimation, and applying the principle to other multi-modal tasks.

References

1. Rana, C., et al.: Artificial intelligence based object detection and traffic prediction by autonomous vehicles–a review. Expert Systems with Applications 255, 124664 (2024)
2. Islam, M.M., Chowdhury, I.J., Mahboob, T.Z., Mazumder, M.S.J., Hossain, M.J., Biswas, M.S., Rone, P.D.: A comprehensive review on object detection in the context of autonomous driving. In: 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS). pp. 1860–1864 (2024). https://doi.org/10.1109/ICUIS64676.2024.10867258
3. Contreras, M., Jain, A., Bhatt, N.P., Banerjee, A., Hashemi, E.: A survey on 3d object detection in real time for autonomous driving. Frontiers in Robotics and AI 11, 1212070 (2024)
4. Jocher, G., Qiu, J.: Ultralytics yolo11 (2024), https://github.com/ultralytics/ultralytics
5. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
6. Jocher, G.: Yolov5. GitHub repository (2020), https://github.com/ultralytics/yolov5

7. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7464–7475 (2023)
8. Shu-Yin, C., Ting-Yu, L.: Low-brightness object recognition based on deep learning. Computers, Materials, & Continua 79(2), 1757 (2024)
9. Ma, J., Ma, Y., Li, C.: Infrared and visible image fusion methods and applications: A survey. Information fusion 45, 153–178 (2019)
10. Song, K., Zhao, Y., Huang, L., Yan, Y., Meng, Q.: Rgb-t image analysis technology and application: A survey. Engineering Applications of Artificial Intelligence 120, 105919 (2023)
11. Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. arXiv preprint arXiv:1611.02644 (2016)
12. Wang, L., Bao, Z., Lu, D.: Gem-yolo: A lightweight and real-time rgbt object detector with gated multimodal fusion. Sensors 26(7), 2035 (2026)
13. Wan, D., Lu, R., Fang, Y., Lang, X., Shu, S., Chen, J., Shen, S., Xu, T., Ye, Z.: Yolov11-rgbt: Towards a comprehensive single-stage multispectral object detection framework. arXiv preprint arXiv:2506.14696 (2025)
14. Qingyun, F., Zhaokui, W.: Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. Pattern Recognition 130, 108786 (2022)
15. Zhao, F., Lou, W., Feng, H., Ding, N., Li, C.: Mfmg-net: Multispectral feature mutual guidance network for visible–infrared object detection. Drones 8(3), 112 (2024)
16. Wang, J., Tian, X., Dai, S., Zhuo, T., Zeng, H., Liu, H., Liu, J., Zhang, X., Zhang, Y.: Rgb-t object detection via group shuffled multi-receptive attention and multi-modal supervision. In: International Conference on Pattern Recognition. pp. 284–298. Springer (2024)
17. Liu, L., Zhang, S., Kuang, Z., Zhou, A., Xue, J.H., Wang, X., Chen, Y., Yang, W., Liao, Q., Zhang, W.: Group fisher pruning for practical network compression. In: International Conference on Machine Learning. pp. 7021–7032. PMLR (2021)
18. Han, S., Pool, J., Tran, J., Dally, W.: Learning both weights and connections for efficient neural network. Advances in neural information processing systems 28 (2015)
19. Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., Peste, A.: Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. Journal of Machine Learning Research 22(241), 1–124 (2021)
20. Shen, J., Zhan, H., Zuo, X., Fan, H., Yuan, X., Li, J., Yang, W.: Irdfusion: Iterative relation-map difference guided feature fusion for multispectral object detection. Pattern Recognition p. 113189 (2026)
21. Guo, J., Gao, C., Liu, F., Meng, D., Gao, X.: Damsdet: Dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion. In: European Conference on Computer Vision. pp. 464–481. Springer (2024)
22. Zhang, X., Zhang, X., Wang, J., Ying, J., Sheng, Z., Yu, H., Li, C., Shen, H.L.: Tfdet: Target-aware fusion for rgb-t pedestrian detection. IEEE Transactions on Neural Networks and Learning Systems 36(7), 13276–13290 (2024)
23. Chen, Y.T., Shi, J., Ye, Z., Mertz, C., Ramanan, D., Kong, S.: Multimodal object detection via probabilistic ensembling. In: European Conference on Computer Vision. pp. 139–158. Springer (2022)
24. Li, Y., Zhan, W., Jiang, Y., Guo, J.: Rdcrnet: Rgb-t object detection network based on cross-modal representation model. Entropy 27(4), 442 (2025)
25. Liu, Z., Zhu, C., Li, Y., Ye, P.: A novel implicit cross-attention framework for rgb-t object detection. Expert Systems with Applications p. 131570 (2026)

26. Deng, B., Liu, D., Cao, Y., Liu, H., Yan, Z., Chen, H.: Cfrnet: Cross-attention-based fusion and refinement network for enhanced rgb-t salient object detection. Sensors 24(22), 7146 (2024)
27. He, Y., Zhang, X., Sun, J.: Channel pruning for accelerating very deep neural networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1389–1397 (2017)
28. Luo, J.H., Wu, J., Lin, W.: Thinet: A filter level pruning method for deep neural network compression. In: Proceedings of the IEEE international conference on computer vision. pp. 5058–5066 (2017)
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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5802–5811 (2022)
30. Liang, M., Hu, J., Bao, C., Feng, H., Deng, F., Lam, T.L.: Explicit attention-enhanced fusion for rgb-thermal perception tasks. IEEE Robotics and Automation Letters 8(7), 4060–4067 (2023)
31. Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: Llvip: A visible-infrared paired dataset for low-light vision. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3496–3504 (2021)
32. Tang, L., Deng, Y., Ma, Y., Huang, J., Ma, J.: Superfusion: A versatile image registration and fusion network with semantic awareness. IEEE/CAA Journal of Automatica Sinica 9(12), 2121–2137 (2022)
33. Li, J., Chen, J., Liu, J., Ma, H.: Learning a graph neural network with cross modality interaction for image fusion. In: Proceedings of the 31st ACM international conference on multimedia. pp. 4471–4479 (2023)
34. Yang, F., Liang, B., Li, W., Zhang, J.: Multidimensional fusion network for multispectral object detection. IEEE Transactions on Circuits and Systems for Video Technology 35(1), 547–560 (2024)
35. Dong, W., Zhu, H., Lin, S., Luo, X., Shen, Y., Guo, G., Zhang, B.: Fusion-mamba for cross-modality object detection. IEEE Transactions on Multimedia (2025)
36. Zhang, H., Fromont, E., Lefèvre, S., Avignon, B.: Guided attentive feature fusion for multispectral pedestrian detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 72–80 (2021)
37. Cao, Y., Bin, J., Hamari, J., Blasch, E., Liu, Z.: Multimodal object detection by channel switching and spatial attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 403–411 (2023)
38. Zhao, T., Yuan, M., Jiang, F., Wang, N., Wei, X.: Removal then selection: A coarse-to-fine fusion perspective for rgb-infrared object detection. IEEE Transactions on Intelligent Transportation Systems 27(2), 2504–2519 (2026). https://doi.org/10.1109/TITS.2025.3638627