PRIMARY-COMPLEMENTARY BIDIRECTIONAL ENHANCEMENT AND QUERY-WISE MODALITY PREFERENCE SELECTION FOR VISIBLE-INFRARED OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing visible-infrared object detectors commonly fuse heterogeneous features into a single representation for detection. However, this paradigm may weaken modality-specific discriminative cues and fail to explicitly model object-level modality preferences. To address these challenges, we propose BEPS-Net, a query-based visible-infrared object detection framework. Specifically, we introduce Primary-Complementary Bidirectional Enhancement (PCBE), which treats each modality as the primary representation in one enhancement direction and incorporates complementary information from the other. This design maintains distinct visible-primary and infrared-primary representations, helping preserve modality-specific discriminative cues while exploiting cross-modal complementarity. Furthermore, we introduce Query-wise Modality Preference Selection (QMPS), which integrates query initialization from both representations with adaptive modality routing, allowing each object query to dynamically reassess its modality preference at every decoder layer and select informative cues from the two enhanced representations. Extensive experiments demonstrate that BEPS-Net outperforms existing state-of-the-art visible-infrared detectors on multiple benchmarks, while the proposed designs can be effectively extended to visible-infrared semantic segmentation. The code is available at: `https://anonymous.4open.science/r/BEPSNet`.

1 INTRODUCTION

Object detection is essential for visual perception in applications such as autonomous driving and intelligent surveillance. Despite substantial progress, visible-only detectors remain vulnerable when appearance cues are degraded by insufficient illumination or poor contrast. Infrared imagery provides complementary information by capturing thermal radiation and is less dependent on ambient illumination, whereas visible imagery offers rich texture, color, and structural details. Their complementary sensing characteristics motivate visible-infrared object detection, which aims to exploit both modalities for reliable detection across diverse imaging conditions Zhang & Demiris (2023); Zhou et al. (2025); Li et al. (2025); Zhao et al. (2025); Hu et al. (2025b); Tang et al. (2026).

Feature-level fusion is a widely adopted approach to exploiting this complementarity. Through convolutional aggregation Liu et al. (2016); Zhang et al. (2020), attention-based interaction Fang & Wang (2022); Shen et al. (2024), or cross-modal alignment Zhang et al. (2019); Yuan et al. (2022), many existing methods follow a common "single-representation fusion" paradigm, in which visible and infrared features are integrated into a single fused representation for subsequent detection, as illustrated in Figure 1(a). This prevailing paradigm requires a single representation to accommodate both complementary information and modality-specific discriminative cues. Achieving this goal can be challenging because the two modalities may provide markedly different responses to the same object: a target that is clearly distinguishable in one modality may be difficult to separate from the background in the other. Useful detection evidence may reside primarily in one modality, and integrating such heterogeneous responses into a single representation risks weakening this evidence. *Key Observation 1* in Figure 1 illustrates this concern: some targets that are detected using a single

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

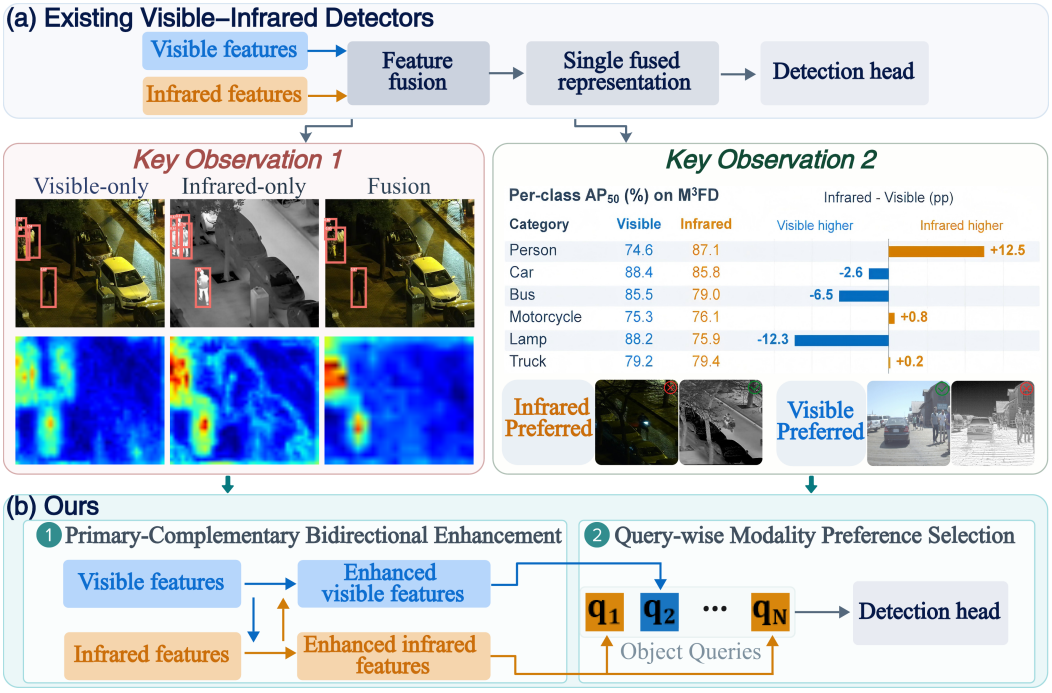

Figure 1: (a) Existing visible-infrared object detectors commonly fuse features from the two modalities into a single representation for subsequent detection. Two observations motivate our design: single-representation fusion may weaken modality-specific cues, as illustrated by the detection results and feature responses (*Key Observation 1*), while modality preferences vary across objects, as evidenced by the class-wise AP_{50} differences on M^3FD (*Key Observation 2*). (b) Our framework addresses these issues with two key designs: Primary-Complementary Bidirectional Enhancement and Query-wise Modality Preference Selection.

modality are less reliably detected after fusion. The corresponding feature visualizations also show that certain target responses evident in the infrared features become less distinct in the fused features. These qualitative observations highlight a potential limitation: integrating heterogeneous features into a single representation can attenuate useful evidence available in an individual modality.

Our second observation is that modality preferences vary across objects under changing environmental conditions (*Key Observation 2* in Figure 1). The results evaluated on visible-only and infrared-only inputs using the same detector (i.e., D-FINE Peng et al. (2025)) reveal clear modality advantages across object categories. For example, infrared improves AP_{50} for *Person* by 12.5 percentage points over visible imagery, whereas visible imagery outperforms infrared by 12.3 percentage points for *Lamp*. In the illustrated examples, the person in low light is more distinguishable in infrared imagery, whereas the daytime car exhibits clearer appearance and structural cues in visible imagery. These results indicate that neither modality should be assumed to provide the most informative evidence for all targets. Although adaptive fusion can modulate modality contributions across different feature locations or channels, it does not provide explicit object-level preference modeling. Therefore, the detector should also be able to exploit these representations in an object-specific manner.

Motivated by these two observations, we propose BEPS-Net, a novel visible-infrared object detection framework built around two corresponding designs: **P**rimary-**C**omplementary **B**idirectional **E**nhancement (PCBE) and **Q**uery-wise **M**odality **P**reference **S**election (QMPS), as shown in Figure 1(b). The former addresses how to enhance each modality while preserving its modality-specific discriminative cues, whereas the latter focuses on how to select useful cues from the resulting representations according to the preference of each object query. Specifically, PCBE organizes cross-modal interaction around an explicit primary-complementary relationship. In each direction, one modality serves as the primary representation and receives complementary information from the other, producing enhanced visible and infrared features, respectively. Maintaining these two outputs allows the detector to access both a visible-primary and an infrared-primary representation after

cross-modal interaction. To exploit these enhanced representations at the object level, we further introduce QMPS. Different objects within the same image may favor different modalities, making a single image-level modality preference insufficient to capture their individual needs. We therefore adopt a DETR-like set prediction framework Carion et al. (2020), in which object queries serve as instance-level prediction units and are associated with individual objects through set-based matching. This formulation provides an explicit unit for modeling modality preference at a finer granularity: each query can express a preference for its corresponding candidate object, allowing different objects within the same image to favor different modalities. Building on this formulation, object queries are initialized from the two modality-enhanced representations. During decoding, each query interacts separately with these representations, and a router then predicts the query's modality preference and selects the response from the preferred branch. This selection is performed at every decoder layer, allowing modality choices to be revised as the queries are progressively refined and mitigating the effects of potentially suboptimal early choices. Extensive experiments demonstrate that BEPS-Net consistently improves performance over existing visible-infrared detection methods and can be effectively extended to visible-infrared semantic segmentation, further validating the effectiveness of the proposed designs.

- We propose BEPS-Net, a novel visible-infrared object detection framework designed to address the potential attenuation of modality-specific discriminative cues under single-representation fusion and the insufficient consideration of object-level modality preferences.
- We design Primary-Complementary Bidirectional Enhancement, which enables each modality to incorporate complementary information from the other while facilitating the preservation of its own modality-specific discriminative cues under reciprocal primary-complementary roles.
- We introduce Query-wise Modality Preference Selection, which explicitly models object-level modality preferences by adaptively routing each object query to select preferred evidence from the two enhanced representations.
- Extensive experiments demonstrate that BEPS-Net outperforms existing state-of-the-art methods.

2 Related Work

2.1 Object Detection

CNN-based object detectors generally follow two-stage or one-stage designs. Two-stage methods Ren et al. (2016) generate region proposals before classification and localization, whereas one-stage methods, exemplified by YOLO Redmon et al. (2016), directly predict categories and bounding boxes from image features. DETR Carion et al. (2020) reformulates object detection as direct set prediction with bipartite matching, where object queries serve as instance-level prediction units and are progressively refined through decoder interactions with image features. Subsequent developments improve multi-scale feature interaction and convergence through deformable attention Zhu et al. (2021), strengthen query selection and iterative refinement Zhang et al. (2023), and further advance detection efficiency and localization accuracy Zhao et al. (2024b); Peng et al. (2025). In visible-infrared detection, however, the potential of object queries for object-level cross-modal information modeling remains relatively underexplored.

2.2 Visible-Infrared Object Detection

Visible-infrared object detection exploits complementary appearance and thermal cues, commonly through feature-level fusion. Existing approaches perform cross-modal interaction through convolutional aggregation Liu et al. (2016); Zhang et al. (2020), attention-based fusion Fang & Wang (2022); Shen et al. (2024), and state-space modeling Dong et al. (2025); Zhu et al. (2025). To accommodate varying modality reliability, adaptive mechanisms adjust modality contributions across spatial locations and channels Fang & Wang (2022); Cao et al. (2023) or according to illumination conditions Li et al. (2019); Guan et al. (2019). Cross-modal alignment further compensates for spatial discrepancies between the two modalities Zhang et al. (2019); Yuan & Wei (2024). Other studies seek to

mitigate modality interference Wang et al. (2025) or strengthen unimodal representation learning through knowledge distillation Zhao et al. (2025). Despite improving cross-modal interaction and fusion quality, many existing approaches ultimately integrate information from both modalities into a single fused representation for subsequent detection. Query-based visible-infrared detectors Xiao et al. (2024); Xing et al. (2024); Medeiros et al. (2025) have also been adopted to explore cross-modal interaction. For example, GM-DETR Xiao et al. (2024) employs an efficient fusion encoder for visible-infrared feature fusion. Nevertheless, they mainly rely on fused representations without explicitly modeling query-wise modality preferences. Our work focuses on effectively exploiting cross-modal complementarity while preserving modality-specific discriminative cues through primary-complementary bidirectional enhancement, followed by query-wise modality preference selection that allows each object hypothesis to select evidence from its preferred modality-enhanced representation.

3 METHOD

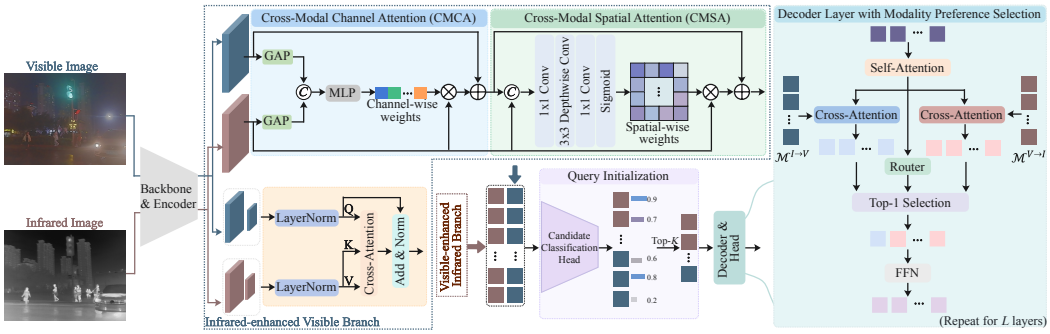

Figure 2: Overview of our BEPS-Net framework.

3.1 OVERVIEW

As illustrated in Figure 2, given an infrared-visible image pair, we first extract multi-scale features using a weight-shared backbone and encoder, obtaining $\mathcal{E}^I = \{\mathcal{E}_l^I\}_{l=3}^5$ and $\mathcal{E}^V = \{\mathcal{E}_l^V\}_{l=3}^5$, where I and V denote the infrared and visible modalities, respectively, and $l \in \{3, 4, 5\}$ indicates different feature scales. Rather than integrating the two modalities into a single fused representation, we perform primary-complementary bidirectional enhancement. In each enhancement direction, one modality serves as the primary representation while complementary information is introduced from the other modality. This process produces two modality-enhanced representations, $\mathcal{M}^{I\to V} = \{\mathcal{M}_l^{I\to V}\}_{l=3}^5$ and $\mathcal{M}^{V\to I} = \{\mathcal{M}_l^{V\to I}\}_{l=3}^5$, where $\mathcal{M}^{I\to V}$ denotes the infrared-enhanced visible representation and $\mathcal{M}^{V\to I}$ denotes the visible-enhanced infrared representation. Based on the two modality-enhanced representations, a query initialization module identifies the Top-K candidates across the two modality-enhanced representations and uses them to initialize the object queries for the decoder. The initialized queries are subsequently refined by the decoder, where each query interacts separately with the two modality-enhanced representations and QMPS selects the preferred response at each decoder layer, ultimately producing the final detection predictions.

3.2 PRIMARY-COMPLEMENTARY BIDIRECTIONAL ENHANCEMENT

To exploit cross-modal complementarity while effectively preserving modality-specific cues, we introduce PCBE, which constructs two reciprocal enhancement branches. In each branch, one modality serves as the primary representation while the other provides complementary information. To accommodate multi-scale features, we employ lightweight channel-spatial interaction for high-resolution features and cross-attention for low-resolution features. The two branches share the same enhancement design but maintain independent parameters for modality-specific optimization.

High-Resolution Enhancement. High-resolution features (i.e., $\mathcal{E}_3$) preserve rich local structures and fine-grained spatial details. However, applying global attention to such features incurs substan-

tial computational overhead. We therefore employ a lightweight enhancement module consisting of Cross-Modal Channel Attention (CMCA) and Cross-Modal Spatial Attention (CMSA). Taking the infrared-enhanced visible branch as an example, CMCA first extracts global channel descriptors from the visible and infrared features:

$$\mathbf{d}^V = \mathrm{GAP}\left(\mathcal{E}_3^V\right), \qquad \mathbf{d}^I = \mathrm{GAP}\left(\mathcal{E}_3^I\right), \tag{1}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling. The two descriptors are then concatenated and fed into an MLP to generate cross-modal channel weights:

$$\mathbf{A}_c^{I\to V} = \sigma\left(\mathrm{MLP}\left([\mathbf{d}^V; \mathbf{d}^I]\right)\right), \tag{2}$$

where $[\cdot;\cdot]$ denotes concatenation and $\sigma(\cdot)$ denotes the sigmoid function. The complementary infrared feature is modulated by the cross-modal channel weights and then added to the visible feature through a residual connection:

$$\hat{\mathcal{E}}_3^{I\to V} = \mathcal{E}_3^V + \mathbf{A}_c^{I\to V} \odot \mathcal{E}_3^I, \tag{3}$$

where $\odot$ denotes element-wise multiplication with spatial broadcasting. To further exploit complementary information in the spatial dimension, CMSA jointly considers the channel-enhanced visible feature $\hat{\mathcal{E}}_3^{I\to V}$ and the infrared feature $\mathcal{E}_3^I$. Specifically, the two features are concatenated along the channel dimension and processed by a convolutional block to generate a cross-modal spatial attention map:

$$\mathbf{A}_s^{I\to V} = \sigma\left(\mathcal{S}\left([\hat{\mathcal{E}}_3^{I\to V}; \mathcal{E}_3^I]\right)\right), \tag{4}$$

where $\mathcal{S}(\cdot)$ consists of a 1×1 convolution, a 3×3 depthwise convolution, and another 1×1 convolution. The resulting spatial attention map is used to spatially modulate the complementary infrared feature, which is then added to the channel-enhanced visible representation through a residual connection:

$$\mathcal{M}_3^{I\to V} = \hat{\mathcal{E}}_3^{I\to V} + \mathbf{A}_s^{I\to V} \odot \mathcal{E}_3^I. \tag{5}$$

Through sequential channel and spatial enhancement, complementary infrared features are incorporated into the visible-primary representation. The residual pathway provides a direct connection from the visible features to the enhanced representation, supporting the retention of modality-specific cues. The visible-enhanced infrared representation $\mathcal{M}_3^{V\to I}$ is obtained symmetrically by exchanging the roles of the two modalities.

Low-Resolution Enhancement. For low-resolution features $\mathcal{E}_4$ and $\mathcal{E}_5$, we employ multi-head cross-attention to capture long-range cross-modal dependencies. Taking the infrared-enhanced visible branch as an example, the visible and infrared features are first normalized, with the visible feature serving as the query and the infrared feature providing the key and value:

$$\mathbf{A}_l^{I\to V} = \mathrm{MHA}\left(\bar{\mathcal{E}}_l^V, \bar{\mathcal{E}}_l^I, \bar{\mathcal{E}}_l^I\right), \quad l \in \{4, 5\}, \tag{6}$$

where $\bar{\mathcal{E}}_l^V$ and $\bar{\mathcal{E}}_l^I$ denote the normalized visible and infrared features, respectively. The infrared response is incorporated into the visible-primary representation through a residual connection followed by normalization:

$$\mathcal{M}_l^{I\to V} = \mathrm{LN}\left(\bar{\mathcal{E}}_l^V + \mathbf{A}_l^{I\to V}\right). \tag{7}$$

$\mathcal{M}_l^{V\to I}$ is obtained symmetrically by exchanging the two modalities for $l = 4, 5$.

3.3 Query-wise Modality Preference Selection

Query Initialization. Based on the two modality-enhanced representations, we flatten and concatenate their multi-scale features to construct a joint candidate pool. Let $\mathbf{M}$ denote the candidate features after projection into the query embedding space. A candidate classification head predicts class logits $\mathbf{z}_j = \mathcal{H}_{\mathrm{cls}}(\mathbf{M}_j)$. The ranking score of candidate j is defined as

$$s_j = \max_{c\in\{1,\ldots,C\}} \sigma(z_{j,c}), \qquad \Omega = \mathrm{TopK}(\{s_j\}_{j=1}^N, K), \tag{8}$$

where σ is the sigmoid function and C is the number of object categories, and N is the total number of candidates. Ω contains the indices of the K highest-scoring candidates. The selected candidate features initialize the decoder queries. This global selection allows candidates from the two modality-enhanced representations to compete for the same query slots, enabling high-confidence object hypotheses from either branch to initialize the decoder queries.

Modality Preference Selection. Since the relative usefulness of visible-primary and infrared-primary representations may vary across objects, we allow each object query to adaptively select the modality-enhanced representation that provides more informative evidence for its current object hypothesis. At the t-th decoder layer, the input object queries first perform self-attention to model inter-query dependencies, yielding the updated query representation $\bar{\mathbf{Q}}^t$. The updated queries then independently interact with the two modality-enhanced representations through cross-attention:

$$\mathbf{Z}_t^{I\to V} = \mathrm{CA}\left(\bar{\mathbf{Q}}^t, \mathcal{M}^{I\to V}\right), \qquad \mathbf{Z}_t^{V\to I} = \mathrm{CA}\left(\bar{\mathbf{Q}}^t, \mathcal{M}^{V\to I}\right), \tag{9}$$

where $\mathrm{CA}(\cdot,\cdot)$ denotes the cross-attention operation Hu et al. (2025b). The router estimates branch preferences from the self-attended query embeddings, and the resulting decisions are applied to the two cross-attention responses.

$$\mathbf{P}^t = \mathrm{Softmax}\left(\mathrm{Router}\left(\bar{\mathbf{Q}}^t\right)\right), \tag{10}$$

where $\mathbf{P}^t \in \mathbb{R}^{K\times 2}$. For the i-th query, Top-1 routing selects the branch with the highest preference probability:

$$g_i^t = \underset{m\in\{0,1\}}{\mathrm{argmax}}\, P_{i,m}^t. \tag{11}$$

Accordingly, the routed response is

$$\mathbf{Z}_i^t = \mathbb{1}(g_i^t = 0)\mathbf{Z}_{t,i}^{I\to V} + \mathbb{1}(g_i^t = 1)\mathbf{Z}_{t,i}^{V\to I}, \tag{12}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. During training, a straight-through estimator is employed for the hard Top-1 routing, enabling gradients to propagate through the corresponding soft routing probabilities while retaining discrete branch selection in the forward pass.

After both cross-attention responses have been computed, the router selects one response for each query. The selected response is then added to the current query representation via residual addition (omitted in Figure 2 for clarity), followed by further refinement through the feed-forward network. We perform modality preference selection at every decoder layer, allowing the preferred branch to be updated as the object queries are progressively refined, which mitigates the effect of potentially suboptimal routing decisions made at earlier stages.

4 Experiments

4.1 Main Results

We evaluate our method on LLVIP Jia et al. (2021), FLIR Zhang et al. (2020), M^{3}FD Liu et al. (2022), and VEDAI Razakarivony & Jurie (2016), using standard COCO AP as the evaluation metric. Detailed dataset information and experimental results on VEDAI are provided in Appendices A and C.1, respectively. Comparison methods include single-modality baselines, CNN-based visible-infrared detectors, and DETR-based visible-infrared detectors. Furthermore, we extend the proposed designs to visible-infrared semantic segmentation and validate their effectiveness on the MFNet dataset Ha et al. (2017), as detailed in Appendix C.3.

Comparison with Existing Methods. As shown in Table 1, BEPS-Net achieves excellent performance across all three benchmarks. In terms of the mAP metric, BEPS-Net reaches 70.8%, 50.1%, and 62.2% on LLVIP, FLIR, and M^{3}FD, respectively. Compared with the state-of-the-art visible–infrared detectors, it improves mAP by 1.2 percentage points over DAMSDet on LLVIP, 0.8 percentage points over DAMSDet on FLIR, and a notable 2.8 percentage points over FreDFT on M^{3}FD. These consistent improvements demonstrate the effectiveness of BEPS-Net across datasets with different scene characteristics and category compositions.

The results also show that using both modalities does not always outperform the best-performing single modality. On LLVIP, FLIR, and M^{3}FD, the strongest single-modality detectors achieve 67.9%, 46.5%, and 54.7% mAP, respectively, surpassing some existing visible-infrared detectors. This indicates that directly collapsing heterogeneous modality features into a single representation may make it difficult to preserve modality-specific discriminative cues, potentially degrading detection performance. In contrast, BEPS-Net maintains consistent improvements over strong single-modality

Table 1: Comparison of detection performance on the LLVIP, FLIR, and M^{3}FD datasets. Bold font denotes the best result, underlining indicates the second-best result. † denotes DETR-based multispectral detectors. Per-class results on M^{3}FD are reported in terms of AP_{50}.

	Methods	LLVIP		FLIR		M^{3}FD						
		mAP	mAP_{50}	mAP	mAP_{50}	Person	Car	Bus	Motorcycle	Lamp	Truck	mAP
Visible	YOLOv11 Jocher & Qiu (2024)	51.1	89.1	32.1	65.0	73.0	92.2	92.6	70.1	79.6	85.7	54.7
	DINO Zhang et al. (2023)	54.5	92.0	35.8	72.2	74.9	90.7	90.6	78.4	85.3	82.6	52.6
	RT-DETR Zhao et al. (2024b)	53.7	91.3	33.9	69.4	75.9	88.9	86.3	75.3	89.1	81.9	52.7
	DEIM Huang et al. (2025)	54.0	91.8	34.5	69.4	73.7	89.7	89.4	78.0	84.5	80.0	53.4
	D-FINE Peng et al. (2025)	55.4	92.1	34.9	70.1	74.6	88.4	85.5	75.3	88.2	79.2	52.8
Infrared	YOLOv11 Jocher & Qiu (2024)	66.3	96.2	41.0	75.7	84.3	90.4	89.7	61.6	60.3	81.1	51.6
	DINO Zhang et al. (2023)	67.7	97.3	46.5	82.1	85.4	87.5	86.4	69.2	68.2	77.7	48.0
	RT-DETR Zhao et al. (2024b)	66.0	96.7	45.2	81.7	85.9	85.4	80.1	67.9	77.9	75.4	49.0
	DEIM Huang et al. (2025)	67.9	97.5	45.8	81.7	87.0	85.6	79.8	74.0	77.6	76.8	50.2
	D-FINE Peng et al. (2025)	67.0	96.5	44.7	81.7	87.1	85.8	79.0	76.1	75.9	79.4	50.0
Visible–Infrared	TFDet Zhang et al. (2024b)	59.4	96.0	46.6	86.6	80.1	83.2	85.6	82.3	85.2	83.2	53.6
	ICAFusion Shen et al. (2024)	65.3	97.4	41.4	79.2	83.0	91.0	92.3	76.5	85.0	88.9	53.5
	DAMSDet † Guo et al. (2024)	69.6	97.9	49.3	86.6	74.5	93.2	78.7	80.0	75.0	79.6	52.9
	MiPa † Medeiros et al. (2025)	66.5	**98.2**	44.8	81.3	-	-	-	-	-	-	-
	MS-DETR † Xing et al. (2024)	64.9	97.5	39.9	78.7	-	-	-	-	-	-	-
	EI^{2}Det Hu et al. (2025a)	63.9	98.0	-	-	82.8	91.4	91.6	77.2	84.7	**89.3**	55.5
	FD^2-Net Li et al. (2025)	-	96.2	-	82.9	83.7	**93.6**	82.7	78.1	87.8	73.8	-
	DeformCAT Hu et al. (2025b)	66.1	97.8	43.2	84.7	80.0	89.5	80.3	56.3	68.1	65.6	46.5
	Fusion-Mamba Dong et al. (2025)	64.3	97.0	47.0	84.9	80.3	92.8	91.9	73.0	84.8	87.1	57.5
	COFNet Zhou et al. (2025)	65.9	97.7	44.6	83.6	-	-	-	-	-	-	-
	FENet Liu et al. (2025)	62.6	95.8	-	-	81.4	91.6	92.2	80.0	69.5	85.4	55.8
	DWSF-Net Yang et al. (2026)	-	97.4	41.8	82.3	83.7	**93.6**	92.6	74.3	87.9	**89.3**	-
	FreDFT Wu et al. (2026a)	66.0	**98.2**	42.8	83.5	85.4	93.4	90.3	78.1	**90.0**	88.2	59.4
	Ours	**70.8**	97.8	**50.1**	**87.5**	**92.4**	93.0	**93.3**	**85.9**	89.3	85.9	**62.2**

baselines, demonstrating its effectiveness in mitigating the limitations of single-representation fusion.

The category-wise single-modality results on M^{3}FD show consistent differences in modality performance across the evaluated detectors. Specifically, infrared shows clear advantages for *Person*, whereas visible imagery performs better for *Lamp*, *Car*, and Bus. This cross-detector consistency indicates that modality preference is inherently associated with object categories and imaging conditions. Existing visible-infrared detectors employ a single fused representation, which cannot explicitly model such object-level modality preferences; even when the fusion stage can modulate the contributions of different modalities, it does so only at the input level. These results provide direct motivation for modeling modality preference at the object-query level. By enabling individual object queries to select the preferred modality that better matches their object hypotheses, QMPS effectively improves category-wise performance on M^{3}FD, thereby yielding overall performance gains.

Table 2: Cross-dataset generalization performance from M^{3}FD to LLVIP.

Method	Input	mAP	mAP_{50}
D-FINE Peng et al. (2025)	V	24.7	49.9
DEIM Huang et al. (2025)	V	26.6	55.1
D-FINE Peng et al. (2025)	I	38.3	64.3
DEIM Huang et al. (2025)	I	36.2	58.2
DAMSDet Guo et al. (2024)	V+I	49.7	80.1
DSAFormer Wu et al. (2026b)	V+I	43.4	74.3
FreDFT Wu et al. (2026a)	V+I	47.1	79.4
Ours	V+I	**55.8**	**87.1**

Table 3: Complexity comparison with multimodal detectors. Inference time is reported on FLIR.

Method	Param.	FLOPs	Inference Time (ms)
ICAFusion Shen et al. (2024)	120.2M	192.7G	**42**
DAMSDet Guo et al. (2024)	79.1M	134.7G	109
MMFN Yang et al. (2024)	176.4M	160.1G	81
DeformCAT Hu et al. (2025b)	120.1M	**132.0G**	45
DSAFormer Wu et al. (2026b)	138.8M	359.9G	87
FreDFT Wu et al. (2026a)	152.6M	464.5G	106
Ours	**31.7M**	178.9G	78

Cross-Dataset Generalization. To evaluate cross-dataset generalization, we train BEPS-Net on M³FD and directly test it on LLVIP without target-domain fine-tuning. As shown in Table 2, the single-modality detectors exhibit relatively limited transfer performance, whereas the visible-

infrared methods achieve consistently better results, indicating that exploiting complementary information from both modalities remains beneficial under domain shift. Among the compared methods, our approach achieves the best transfer performance, highlighting the advantage of the proposed paradigm in learning representations that generalize effectively across datasets.

Complexity Comparison. Table 3 reports the model complexity and inference efficiency of BEPS-Net in comparison with several infrared-visible detectors. All methods are evaluated with paired inputs of 640 × 640 on an NVIDIA V100 GPU, using a batch size of 1. BEPS-Net contains only 31.7M parameters, substantially fewer than the competing methods, whose parameter counts range from 79.1M to 176.4M. This reduction benefits from the parameter-sharing backbone and encoder, together with the lightweight design of PCBE. Despite maintaining two modality-enhanced representations, BEPS-Net requires a moderate computational cost of 178.9G FLOPs and achieves an inference latency of 78 ms, which is comparable to several existing multimodal detectors. Overall, BEPS-Net provides a favorable trade-off between detection performance and computational complexity.

Table 4: Component-wise ablation study of BEPS-Net on LLVIP, FLIR, and M^{3}FD.

Setting	Components		mAP (%)		
	PCBE	QMPS	LLVIP	FLIR	M^{3}FD
Baseline	–	–	66.7	46.2	54.1
+ PCBE	✓	–	69.0	48.4	58.1
+ QMPS	✓	✓	**70.8**	**50.1**	**62.2**

Table 5: Ablation study of different query initialization strategies on LLVIP, FLIR, and M^{3}FD.

Query Initialization	mAP (%)		
	LLVIP	FLIR	M^{3}FD
$I \rightarrow V$	69.7	48.8	61.3
$V \rightarrow I$	70.5	49.0	61.6
Fixed Allocation	70.6	49.3	61.9
Ours	**70.8**	**50.1**	**62.2**

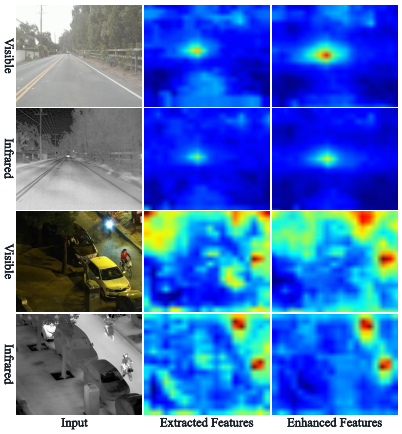

Figure 3: Visualization of feature responses before and after PCBE.

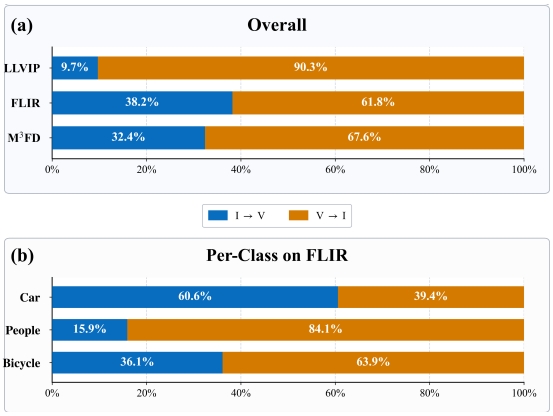

Figure 4: Final-layer routing statistics for queries matched to ground-truth instances. Overall ratios are computed across all matched queries on LLVIP, FLIR, and M^{3}FD; per-category ratios on FLIR are grouped by the matched ground-truth labels.

4.2 Ablation Study

To investigate the effectiveness of the proposed components, we adopt a progressive ablation setting by first introducing PCBE and then incorporating QMPS, with the results reported in Table 4. We also compare different query initialization strategies in Table 5. Additional ablation studies are provided in Appendix D.

Ablation Study of PCBE. The baseline fuses visible and infrared features into a single representation, achieving 66.7% mAP on LLVIP, 46.2% mAP on FLIR, and 54.1% mAP on M^{3}FD. Notably, these results are still lower than the strongest single-modality baselines on the three datasets (67.9%, 46.5%, and 54.7%, respectively), suggesting that collapsing heterogeneous features into a single representation may lead to performance degradation. Introducing PCBE, where queries interact with both modality-enhanced representations and their responses are combined using learnable weights, improves the performance to 69.0% on LLVIP, 48.4% on FLIR, and 58.1% on M^{3}FD. The

consistent improvements demonstrate that PCBE effectively exploits cross-modal complementarity while alleviating the degradation associated with a single fused representation. We further visualize the feature responses before and after enhancement, as shown in Fig. 3. Compared with the extracted features, the enhanced features exhibit more concentrated and discriminative responses around the target regions. These observations indicate that PCBE effectively preserves the intrinsic characteristics of each modality while selectively enhancing complementary cues from the other, thereby providing more discriminative modality-enhanced representations for improved detection performance.

Ablation Study of QMPS. Building upon PCBE, incorporating QMPS yields additional gains of 1.8, 1.7, and 4.1 percentage points on LLVIP, FLIR, and M^3FD, respectively, demonstrating the effectiveness of query-wise modality preference selection across different datasets. We further analyze the routing decisions of the object queries at the final decoder layer, as shown in Fig. 4. At the dataset level, the two branches are selected with clearly different proportions, indicating that QMPS does not rely exclusively on a single modality-enhanced representation but adaptively routes different object queries according to their modality preferences. As shown by the per-class results on FLIR, the preferred modality-enhanced representation varies substantially across object categories. These results further demonstrate that QMPS effectively models object-level modality preferences across diverse scenarios, leading to consistent improvements in detection performance.

Effect of Query Initialization. As shown in Table 5, both single-branch initialization variants yield lower performance, as restricting candidate selection to one representation may overlook informative object hypotheses provided by the other. Fixed Allocation improves mAP to 70.6%, 49.3%, and 61.9% on LLVIP, FLIR, and M^3FD, respectively, demonstrating the benefit of utilizing candidates from both representations. In contrast to the predefined branch-wise quota, Top-K jointly ranks candidates from both representations, dynamically determining how the K query slots are allocated between the two branches and yielding additional gains across all three datasets. Figure 5(a) shows that jointly Top-K converges faster during early training on FLIR than the other initialization strategies. Meanwhile, the spatial distribution in Fig. 5(b) shows that most selected queries are concentrated around target regions. These results suggest that the proposed strategy provides object-relevant initial query hypotheses for subsequent decoding and refinement.



Figure 5: (a) Training loss curves during the first 12 epochs. (b) Spatial distribution of queries initialized by our method.

5 Conclusion

In this work, we revisit visible-infrared object detection based on two key observations: (i) collapsing heterogeneous modalities into a single representation may weaken modality-specific discriminative cues; and (ii) modality preference varies across objects under different scenarios, necessitating object-level modeling of such preferences. Motivated by these observations, we propose BEPS-Net, a novel framework that combines bidirectional modality enhancement with query-wise modality preference selection. Specifically, PCBE constructs visible-primary and infrared-primary representations by introducing complementary information from the other modality while helping to preserve their respective modality characteristics. QMPS further exploits these representations by adaptively selecting the preferred modality-enhanced representation for each object query throughout the decoding process. Extensive experiments demonstrate consistent improvements over existing visible-infrared detectors, while comprehensive ablation studies and cross-task extension experiments verify the effectiveness of the proposed components and design choices.

REFERENCES

Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. Multimodal object detection by channel switching and spatial attention. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 403–411, 2023.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289. IEEE, 2022.

Wenhao Dong, Haodong Zhu, Shaohui Lin, Xiaoyan Luo, Yunhang Shen, Guodong Guo, and Baochang Zhang. Fusion-mamba for cross-modality object detection. *IEEE Transactions on Multimedia*, 2025.

Qingyun Fang and Zhaokui Wang. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognition*, 130:108786, 2022. ISSN 0031-3203. doi: 10.1016/j.patcog.2022.108786.

Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019.

Junjie Guo, Chenqiang Gao, Fangcen Liu, Deyu Meng, and Xinbo Gao. 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.

Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 5108–5115. IEEE, 2017.

Ke Hu, Yudong He, Yuan Li, Jiayu Zhao, Song Chen, and Yi Kang. Ei^{2}det: Edge-guided illumination-aware interactive learning for visible-infrared object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(7):6265–6278, 2025a. doi: 10.1109/TCSVT.2025.3528574.

Yu Hu, Xiaobo Chen, Sheng Wang, Luyang Liu, Hengyang Shi, Lihong Fan, Jing Tian, and Jun Liang. Deformable cross-attention transformer for weakly aligned rgb–t pedestrian detection. *IEEE Transactions on Multimedia*, 2025b.

Shihua Huang, Zhichao Lu, Xiaodong Cun, Yongjun Yu, Xiao Zhou, and Xi Shen. Deim: Detr with improved matching for fast convergence. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15162–15171. IEEE, 2025.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3489–3497. IEEE, 2021.

Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. URL `https://github.com/ultralytics/ultralytics`.

Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019.

Gongyang Li, Yike Wang, Zhi Liu, Xinpeng Zhang, and Dan Zeng. Rgb-t semantic segmentation with location, activation, and sharpening. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1223–1235, 2022.

Ke Li, Di Wang, Zhangyuan Hu, Shaofeng Li, Weiping Ni, Lin Zhao, and Quan Wang. Fd2-net: Frequency-driven feature decomposition network for infrared-visible object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4797–4805, 2025.

Mingjian Liang, Junjie Hu, Chenyu Bao, Hua Feng, Fuqin Deng, and Tin Lun Lam. Explicit attention-enhanced fusion for rgb-thermal perception tasks. *IEEE Robotics and Automation Letters*, 8(7):4060–4067, 2023.

Fangcen Liu, Chenqiang Gao, Fang Chen, Pengcheng Li, Junjie Guo, and Deyu Meng. A fusion-enhanced network for infrared and visible high-level vision tasks. *IEEE Transactions on Multimedia*, 2025.

Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas. Multispectral deep neural networks for pedestrian detection. *arXiv preprint arXiv:1611.02644*, 2016.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan 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*, pp. 5802–5811, 2022.

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 8081–8090. IEEE, 2023.

Yongan Lu, Renbo Luo, Yihan Yao, Jialing Lian, Guohao Liao, and Zhifeng Wu. Mv-moe: A multi-view expert fusion framework for multispectral object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

Heitor R. Medeiros, David Latortue, Eric Granger, and Marco Pedersoli. Mixed patch visible-infrared modality agnostic object detection. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 9005–9014, February 2025.

Kaiyue Men, Chengyou Wang, Xiao Zhou, and Zhiliang Qin. Mamenet: Modal alignment and multi-scale feature enhancement network for rgb-ir small object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

Yansong Peng, Hebei Li, Peixi Wu, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. D-fine: Redefine regression task of detrs as fine-grained distribution refinement. In *International Conference on Learning Representations*, volume 2025, pp. 44015–44031, 2025.

Sebastien Razakarivony and Frederic Jurie. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187–203, 2016.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145: 109913, 2024.

Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82: 28–42, 2022.

Zhanyan Tang, Zhihao Wu, Mu Li, Jie Wen, Bob Zhang, Yong Xu, and Jianqiang Li. Adaptive fine-grained fusion network for multimodal uav object detection. *IEEE Transactions on Image Processing*, 2026.

Yike Wang, Gongyang Li, and Zhi Liu. Sgfnet: Semantic-guided fusion network for rgb-thermal semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33 (12):7737–7748, 2023.

Yu Wang, Shikui Wei, Sen Xu, Ying Qin, and Yao Zhao. Confidence-driven unimodal interference removal for enhanced multimodal object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Wencong Wu, Xiuwei Zhang, Hanlin Yin, Shun Dai, Hongxi Zhang, and Yanning Zhang. Fredft: Frequency domain fusion transformer for visible-infrared object detection. *IEEE Transactions on Multimedia*, 2026a.

Wencong Wu, Xiuwei Zhang, Hanlin Yin, Hongxi Zhang, and Yanning Zhang. Dual sparse aggregation transformer for multispectral object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026b.

Yiming Xiao, Fanman Meng, Qingbo Wu, Linfeng Xu, Mingzhou He, and Hongliang Li. Gm-detr: Generalized muiltispectral detection transformer with efficient fusion encoder for visible-infrared detection. In *2024 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 5541–5549. IEEE, 2024.

Yinghui Xing, Shuo Yang, Song Wang, Shizhou Zhang, Guoqiang Liang, Xiuwei Zhang, and Yanning Zhang. Ms-detr: Multispectral pedestrian detection transformer with loosely coupled fusion and modality-balanced optimization. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):20628–20642, 2024.

Fan Yang, Binbin Liang, Wei Li, and Jianwei Zhang. Multidimensional fusion network for multispectral object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 35 (1):547–560, 2024.

Fan Yang, Wei Li, Lei Li, Min Yang, and Jianwei Zhang. Dwsf-net: A dynamic wavelet-based spatial-frequency fusion network for multispectral object detection. *IEEE Transactions on Multimedia*, 2026.

Maoxun Yuan and Xingxing Wei. C^2former: Calibrated and complementary transformer for rgb-infrared object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

Maoxun Yuan, Yinyan Wang, and Xingxing Wei. Translation, scale and rotation: cross-modal alignment meets rgb-infrared vehicle detection. In *European Conference on Computer Vision*, pp. 509–525. Springer, 2022.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023.

Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 26964–26973. IEEE, 2024a.

Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International Conference on Image Processing*, pp. 276–280. IEEE, 2020.

Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5127–5137, 2019.

Xingchen Zhang and Yiannis Demiris. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10535–10554, 2023.

Xue Zhang, Xiaohan Zhang, Jiangtao Wang, Jiacheng Ying, Zehua Sheng, Heng Yu, Chunguang Li, and Hui-Liang Shen. Tfdet: Target-aware fusion for rgb-t pedestrian detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):13276–13290, 2024b.

Guoqiang Zhao, Junjie Huang, Xiaoyun Yan, Zhaojing Wang, Junwei Tang, Yangjun Ou, Xinrong Hu, and Tao Peng. Open-vocabulary rgb-thermal semantic segmentation. In *European Conference on Computer Vision*, pp. 304–320. Springer, 2024a.

Shenlu Zhao, Yichen Liu, Qiang Jiao, Qiang Zhang, and Jungong Han. Mitigating modality discrepancies for rgb-t semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9380–9394, 2023.

Tianyi Zhao, Boyang Liu, Yanglei Gao, Yiming Sun, Maoxun Yuan, and Xingxing Wei. Rethinking multi-modal object detection from the perspective of mono-modality feature learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6364–6373, 2025.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974, 2024b.

Mingliang Zhou, Yunyao Li, Guangchao Yang, Xuekai Wei, Huayan Pu, Jun Luo, and Weijia Jia. Cofnet: contrastive object-aware fusion using box-level masks for multispectral object detection. *IEEE Transactions on Multimedia*, 2025.

Wujie Zhou, Shaohua Dong, Caie Xu, and Yaguan Qian. Edge-aware guidance fusion network for rgb–thermal scene parsing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 3571–3579, 2022.

Wujie Zhou, Han Zhang, Weiqing Yan, and Weisi Lin. Mmsmcnet: Modal memory sharing and morphological complementary networks for rgb-t urban scene semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7096–7108, 2023.

Haodong Zhu, Wenhao Dong, Linlin Yang, Hong Li, Yuguang Yang, Yangyang Ren, Qingcheng Zhu, Zichao Feng, Changbai Li, Shaohui Lin, et al. Wavemamba: Wavelet-driven mamba fusion for rgb-infrared object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11219–11229, 2025.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and editing of the manuscript, including suggesting rephrasings, correcting grammatical and spelling errors, and improving the clarity, coherence, and readability of the text. We have not used generative AI tools for research ideation, methodology, experimental design, or data collection and analysis; the remaining disclosure items are not applicable. We have reviewed all AI-assisted text to ensure its technical correctness and consistency with the authors' intended meaning. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies visible–infrared object detection using established benchmark datasets and does not involve new data collection or intervention with human participants.

REPRODUCIBILITY STATEMENT

The Method section describes the architecture of BEPS-Net. Training settings and dataset details are provided in the Appendix. Code and configurations for reproducing the main experiments are available in our anonymous repository at `https://anonymous.4open.science/r/BEPSNet`.

A DATASETS

The LLVIP Jia et al. (2021) dataset includes 15,488 infrared-visible image pairs (12,025 for training and 3,463 for testing) and focuses on low-light surveillance scenarios for pedestrian detection.

The FLIR Zhang et al. (2020) dataset contains three categories (person, car, and bicycle) and encompasses a variety of driving scenarios and illumination conditions. Following prior works, we use its updated aligned version. It includes 5,142 infrared-visible image pairs (4,129 for training and 1,013 for testing).

The M^3FD dataset Liu et al. (2022) contains paired infrared-visible images covering six object categories. We adopt the same scene-based partition as DAMSDet Guo et al. (2024), comprising 3,368 training pairs and 831 validation pairs.

The VEDAI dataset Razakarivony & Jurie (2016) contains 1,210 paired RGB and infrared aerial images. We use the 1024×1024 image version. A fixed partition of 1,089 training pairs and 121 testing pairs is used for BEPS-Net and all baselines. Each oriented annotation is converted into the tightest axis-aligned bounding box enclosing its four vertices.

B Implementation Details

We use HGNetv2 and the efficient encoder Zhao et al. (2024b) for feature extraction and multi-scale encoding. We select $K = 300$ candidates to initialize the decoder queries. The decoder consists of six layers, and modality preference selection is performed once at each decoder layer. We train the model for 36 epochs on each dataset and adopt the AdamW optimizer with an initial learning rate of 2.5×10^{-4}, a backbone learning rate of 1.25×10^{-5}, and a weight decay of 1.25×10^{-4}. BEPS-Net is trained on a single NVIDIA V100 GPU with a batch size of 4. We follow the set-prediction training paradigm of DETR-based detectors, using one-to-one Hungarian matching and a weighted combination of Varifocal classification loss, L_1 box regression loss, and generalized IoU loss. The loss weights are set to 1, 5, and 2 for the classification, L_1 regression, and GIoU terms, respectively. Additionally, the selected candidates used for query initialization and their corresponding box predictions are supervised by an auxiliary detection loss.

C Additional Experimental Results

Table 6: Comparison of detection performance on the VEDAI dataset. V and I denote the visible and infrared modalities, respectively. Bold and underlined values indicate the best and second-best results.

Method	**Input**	**mAP**
YOLO11 Jocher & Qiu (2024)	V	44.1
DINO Zhang et al. (2023)	V	44.7
RT-DETR Zhao et al. (2024b)	V	47.7
YOLO11 Jocher & Qiu (2024)	I	43.6
DINO Zhang et al. (2023)	I	41.7
RT-DETR Zhao et al. (2024b)	I	42.6
YOLOFusion Fang & Wang (2022)	V+I	45.8
ICAFusion Shen et al. (2024)	V+I	<u>56.6</u>
DAMSDet Guo et al. (2024)	V+I	55.3
COFNet Zhou et al. (2025)	V+I	56.2
MV-MoE Lu et al. (2026)	V+I	55.9
MAMENet Men et al. (2026)	V+I	53.8
Ours	V+I	**57.0**

C.1 Comparison on the VEDAI Dataset.

As shown in Table 6, BEPS-Net achieves the best result on VEDAI with 57.0% mAP. While most visible–infrared methods benefit from cross-modal complementarity and outperform the single-modality baselines, some earlier fusion approaches, such as YOLOFusion, still underperform the strongest single-modality detector. Although the margin over recent multimodal methods is rela-

tively modest, the leading performance of BEPS-Net indicates that the proposed design remains effective in aerial-view scenarios.

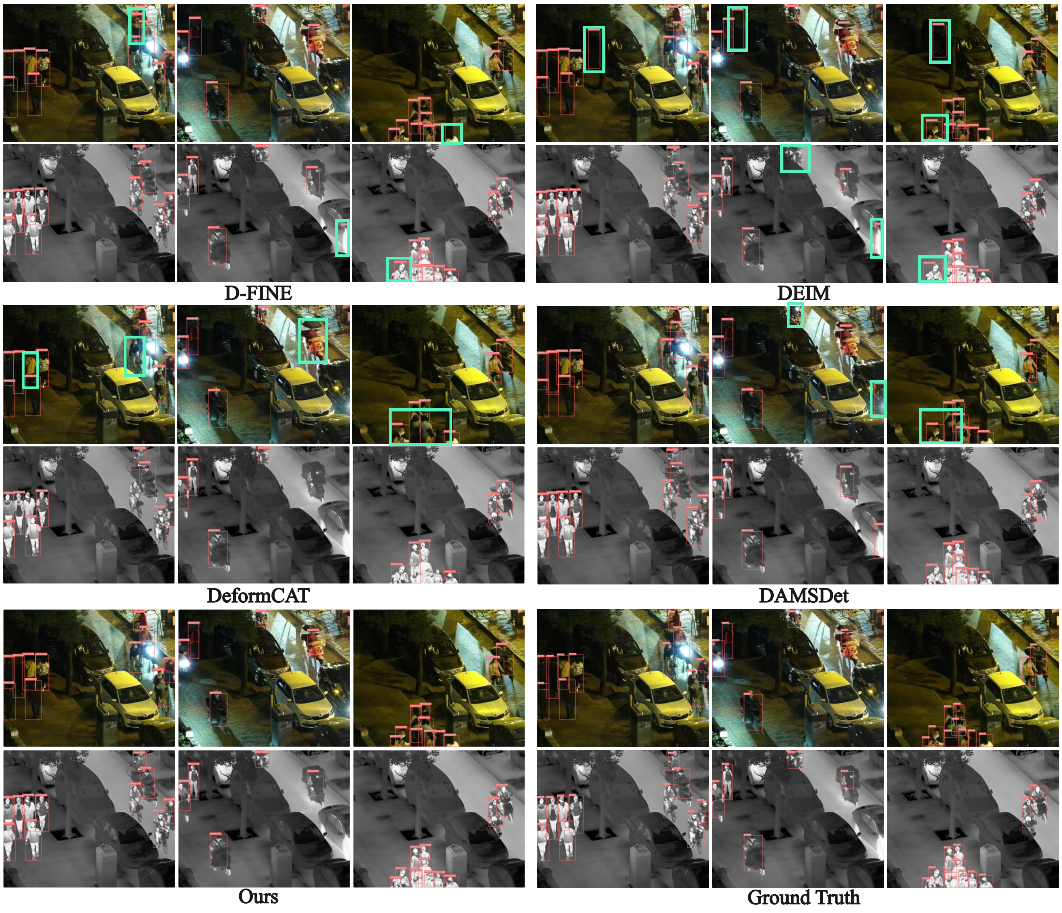

Figure 6: Qualitative comparison on the LLVIP dataset. Cyan boxes indicate missed or inaccurate detections. For multimodal methods, the highlighted results are shown on the visible inputs, whereas for unimodal methods, both visible and infrared inputs are presented for comparison.

C.2 Qualitative Comparison

Figures 6–8 present qualitative comparisons between BEPS-Net and single-modality detectors (D-FINE Peng et al. (2025) and DEIM Huang et al. (2025)), evaluated separately on visible and infrared inputs, as well as visible-infrared detectors (DeformCAT Hu et al. (2025b) and DAMSDet Guo et al. (2024)).

Figure 6, on the LLVIP dataset, shows nighttime pedestrian scenes with uneven illumination, crowded groups, and partial occlusion. The compared methods exhibit missed detections in closely spaced pedestrian groups and false detections in low-light regions. Notably, several pedestrians detected in the infrared-only results remain undetected by certain visible-infrared methods, illustrating that using both modalities does not necessarily improve detection for every object. This observation is consistent with the concern motivating our design. BEPS-Net effectively alleviates these issues by detecting more pedestrians in closely spaced groups and reducing false detections under low-light conditions.

Figure 7, on the FLIR dataset, presents daytime and nighttime road scenes containing objects at different apparent scales. The highlighted errors include missed small vehicles and undetected pedestrians or bicycles along the roadside. Visible-only and infrared-only detectors exhibit different failure patterns, while the visible-infrared baselines still miss some of these targets. BEPS-Net detects more

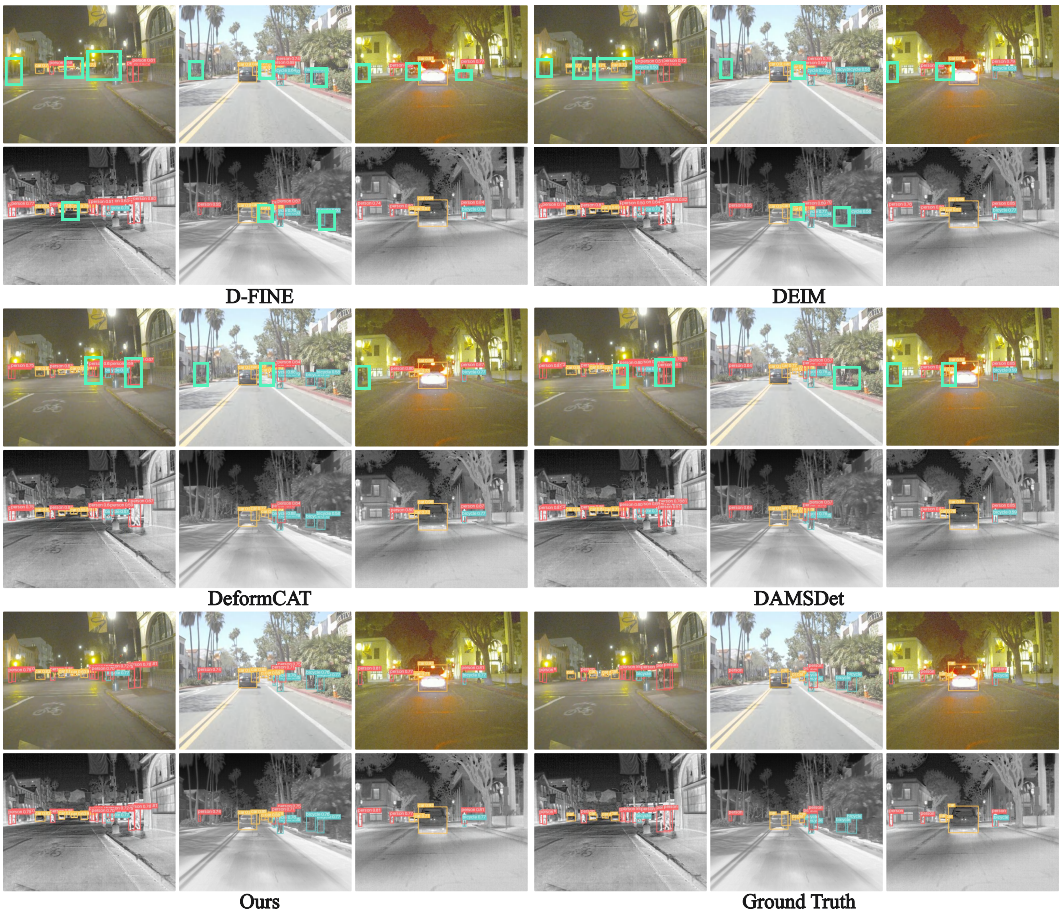

Figure 7: Qualitative comparison on the FLIR dataset. Cyan boxes indicate missed or inaccurate detections. For multimodal methods, the highlighted results are shown on the visible inputs, whereas for unimodal methods, both visible and infrared inputs are presented for comparison.

small and distant objects while maintaining detections of larger nearby vehicles. These examples illustrate improved object coverage across scale and illumination variations in the selected traffic scenes.

Figure 8, on the M^3FD dataset, shows urban traffic scenes containing vehicles, pedestrians, and lamps under different lighting conditions. Large foreground vehicles coexist with small roadside targets, while headlight glare and low background contrast further complicate the nighttime example. The compared methods exhibit missed detections and inaccurate predictions around partially occluded objects and small targets. BEPS-Net recovers several of the highlighted instances and produces predictions that more closely match the ground-truth annotations. The different visibility of objects within the same scene also motivates modeling object-level modality preferences.

Overall, these qualitative results demonstrate that BEPS-Net achieves more reliable detections in challenging scenes involving low illumination, scale variation, and occlusion.

C.3 Extension to Visible-Infrared Semantic Segmentation

Architectural adaptation. To investigate the extensibility of our approach beyond object detection, we adapt PCBE and QMPS to the query-based framework Mask2Former Cheng et al. (2022) for visible-infrared semantic segmentation. Following the query-based segmentation paradigm illustrated in Figure 9, visible and infrared images are processed separately by a backbone and a pixel decoder whose weights are shared across modalities. Each pixel-decoding branch produces multi-scale features ($1/32$, $1/16$, and $1/8$) for query refinement and high-resolution pixel embeddings

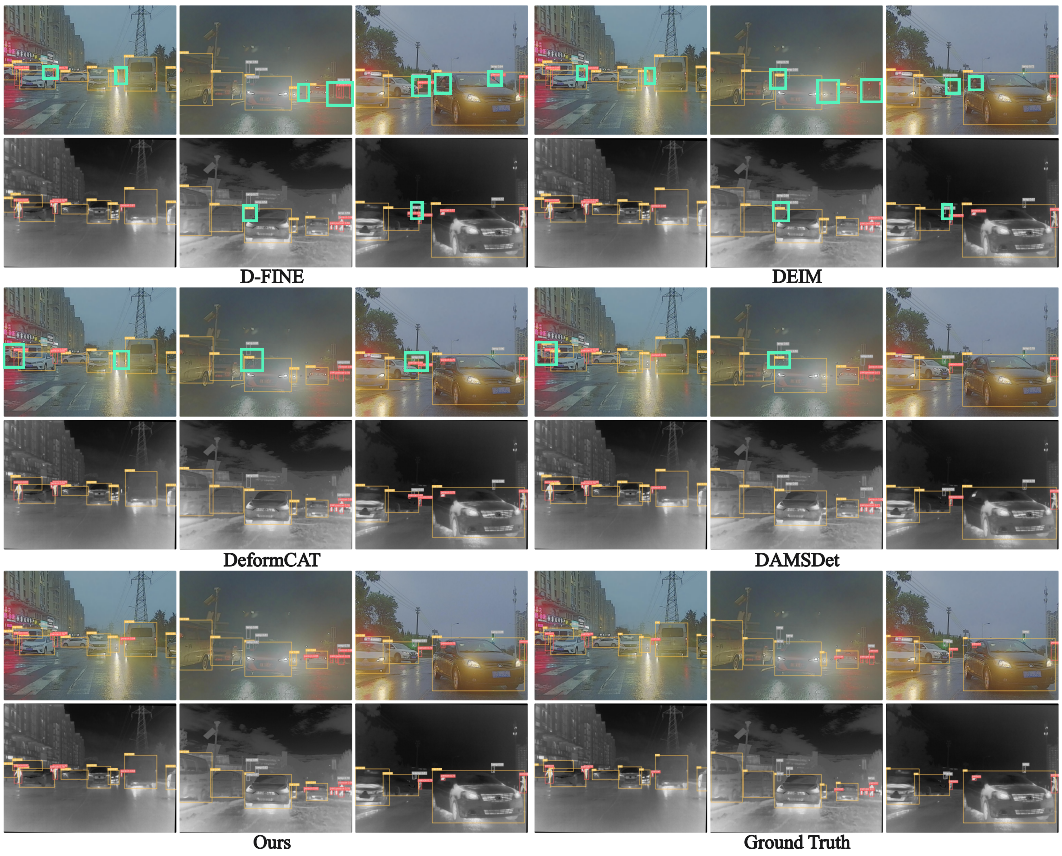

Figure 8: Qualitative comparison on the M^{3}FD dataset. Cyan boxes indicate missed or inaccurate detections. For multimodal methods, the highlighted results are shown on the visible inputs, whereas for unimodal methods, both visible and infrared inputs are presented for comparison.

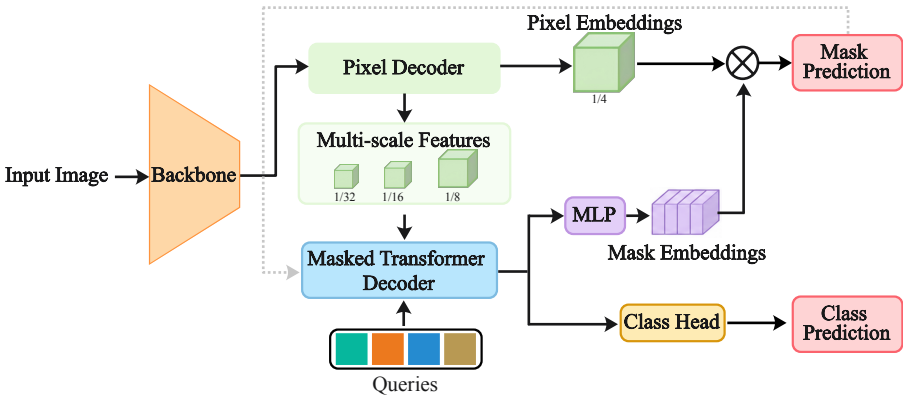

Figure 9: Overview of the query-based Mask2Former architecture for image segmentation.

(1/4) for mask prediction. PCBE enhances the corresponding outputs of the two branches, maintaining distinct visible-primary and infrared-primary representations. Specifically, channel-spatial enhancement is applied to the 1/8 features and extended to the 1/4 pixel embeddings, while cross-attention is used at 1/16 and 1/32.

We retain Mask2Former's original learnable object queries. The enhanced multi-scale features are supplied to successive decoder layers in the original 1/32–1/16–1/8 cycle. Each layer preserves the sequence of masked cross-attention, self-attention, and FFN. Within masked cross-attention, the

same query interacts separately with the two modality-enhanced features under a same mask derived from the preceding stage. A query-wise router predicts a Top-1 routing decision from the incoming query and selects one attention response.

The refined queries are mapped to class predictions and mask embeddings through a classification head and an MLP, respectively. Each mask embedding is combined with both modality-enhanced pixel embeddings through a channel-wise inner product, and the same routing decision selects the resulting mask logits. For query i at layer t, the selected attention response and mask logits are

$$Z_{t,i} = \sum_{m\in\{V,I\}} g^t_{i,m} Z^m_{t,i}, \qquad S^t_i(x) = \sum_{m\in\{V,I\}} g^t_{i,m} \left\langle e^t_i, F^m(x) \right\rangle, \tag{13}$$

where m indexes the visible-primary or infrared-primary branch, $Z^m_{t,i}$ is its masked-attention response, F^m denotes its enhanced $1/4$ pixel embeddings, $e^t_i = \mathrm{MLP}(q^t_i)$ is the mask embedding of the refined query, and g^t_i is a one-hot routing gate trained with a straight-through estimator. This design maintains a consistent modality preference between query refinement and mask prediction.

Table 7: Quantitative comparison results on the MFNet dataset. Best and second-best results are highlighted in **bold** and underline (mIoU).

Method	Modality	Car	Person	Bike	Curve	Car stop	Cone	Bump	Guardrail	mIoU
Mask2Former Cheng et al. (2022)	V	87.8	62.4	63.7	40.5	38.3	50.2	50.4	7.5	55.4
Mask2Former Cheng et al. (2022)	I	85.6	68.6	57.2	39.7	33.8	32.9	48.9	8.6	52.9
EGFNet Zhou et al. (2022)	V+I	87.6	69.8	58.8	42.8	33.8	48.3	47.1	7.2	54.8
SeAFusion Tang et al. (2022)	V+I	84.2	71.1	58.7	33.1	20.1	40.4	33.9	0.1	48.8
LASNet Li et al. (2022)	V+I	84.2	67.1	56.9	41.1	39.6	48.8	40.1	**18.9**	54.9
EAEFNet Liang et al. (2023)	V+I	87.6	**77.6**	63.8	48.6	35.3	52.4	58.3	14.2	58.9
MDRNet Zhao et al. (2023)	V+I	87.1	69.8	60.9	47.8	34.2	50.2	55.1	8.2	56.8
SGFNet Wang et al. (2023)	V+I	88.4	76.6	64.3	45.8	31.2	**57.1**	55.4	6.1	57.6
MMSMC Zhou et al. (2023)	V+I	89.2	69.1	63.5	46.4	41.9	48.8	57.6	8.8	58.1
SegMiF Liu et al. (2023)	V+I	87.8	71.4	63.2	47.5	31.1	48.9	50.3	0.1	56.1
MRFS Zhang et al. (2024a)	V+I	**89.4**	75.4	**64.9**	**49.2**	37.2	53.1	**58.8**	5.4	59.1
OpenRS Zhao et al. (2024a)	V+I	87.4	68.9	63.2	46.8	43.2	49.5	56.4	7.4	58.5
Ours	V+I	89.1	68.8	63.7	48.6	**53.5**	56.3	52.2	9.6	**60.0**

Dataset and evaluation protocol. We evaluate our model on the MFNet Ha et al. (2017) dataset, which contains 1,569 RGB-T image pairs captured by the InfRec R500 sensor in both visible and thermal infrared spectra, covering diverse day and night urban driving scenes. The dataset provides eight semantic categories, including car, person, bike, curve, car stop, guardrail, color cone, and bump. Both visible and thermal images have a spatial resolution of 480×640 pixels. Following previous visible-infrared semantic segmentation methods Zhao et al. (2023); Wang et al. (2023); Zhou et al. (2023); Zhang et al. (2024a), we adopt the standard training/testing split of MFNet for evaluation. We use mean intersection over union (mIoU) as the evaluation metric.

Comparison with existing methods. Table 7 compares our Mask2Former-based extension with unimodal baselines and existing visible-infrared semantic segmentation methods. Our method achieves a mIoU of 60.0%, improving upon the visible-only and infrared-only Mask2Former baselines by 4.6 and 7.1 percentage points, respectively. These improvements demonstrate the effectiveness of exploiting cross-modal complementary information within the query-based segmentation framework. The category-wise results further reveal the complementary characteristics of different modalities: infrared imagery achieves a higher IoU than visible imagery on the *Person* category by 6.2 points, while visible imagery provides a substantially better representation for the *Cone* category with a 17.3 point improvement. Compared with the stronger unimodal Mask2Former baseline for each category, our method improves seven out of eight foreground categories, indicating the benefit of jointly utilizing visible and infrared information.

Compared with existing visible-infrared semantic segmentation methods, our method achieves the highest mIoU of 60.0%, surpassing MRFS, EAEFNet, and OpenRS by 0.9, 1.1, and 1.5 percentage points, respectively. The category-wise comparisons further show that our method achieves competitive performance across different semantic categories. These results indicate that the proposed

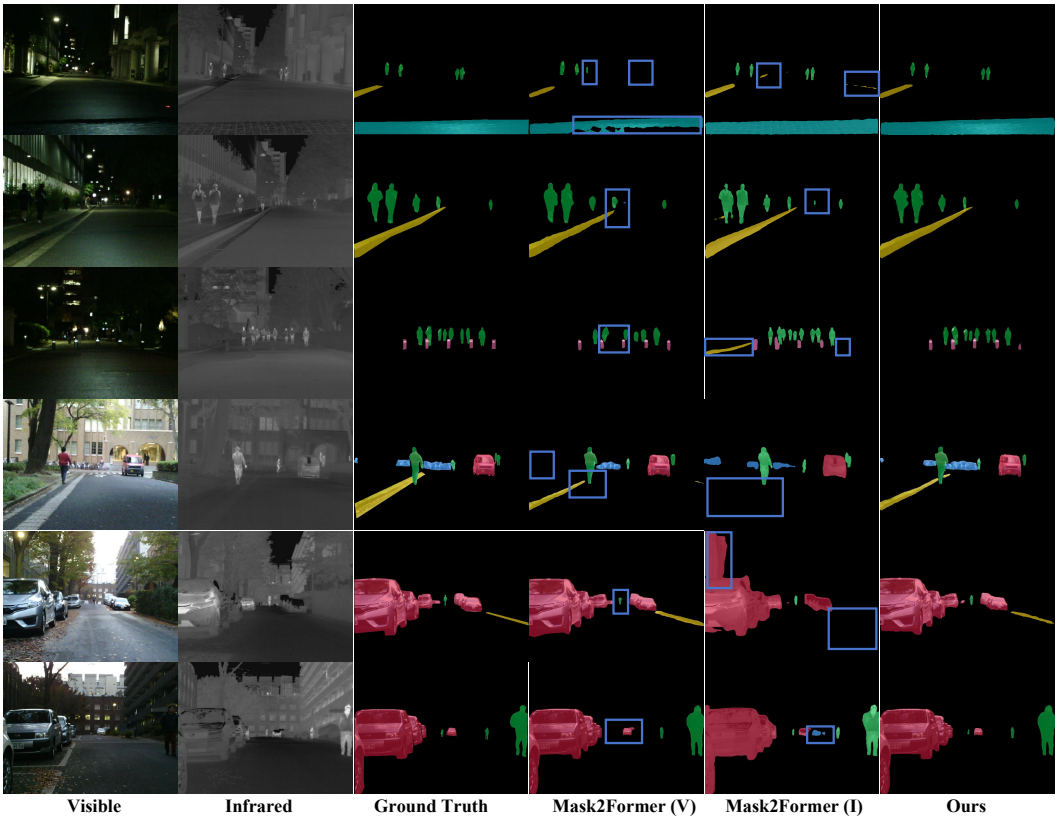

Figure 10: Qualitative comparison on the MFNet dataset. The results of Mask2Former with visible-only input, infrared-only input, and our visible-infrared extension are compared with the ground truth. Our method produces more accurate and complete segmentation results under diverse illumination conditions. Blue boxes indicate inaccurate segmentation regions.

approach can effectively exploit complementary modality information while adapting to category-dependent visual characteristics.

The comparisons also reveal that integrating visible and infrared modalities does not always guarantee improvements over strong unimodal models. Specifically, EGFNet, SeAFusion, and LASNet achieve lower mIoU values than the visible-only Mask2Former baseline by 0.6, 6.6, and 0.5 percentage points, respectively. Moreover, category-wise results show that even methods with higher overall mIoU can still underperform the unimodal baseline on certain categories, indicating that the benefits of multimodal fusion are not consistently distributed across different object categories. These observations further support the motivation that modeling modality preferences is beneficial for reliable visible-infrared semantic segmentation.

Figure 10 presents qualitative comparisons on the MFNet dataset among the visible-only Mask2Former, infrared-only Mask2Former, and our visible-infrared extension. The selected scenes cover diverse urban environments with different illumination conditions, including daytime and nighttime scenarios. Compared with single-modal baselines, our method generates segmentation results that are generally closer to the ground truth, demonstrating the advantage of incorporating complementary information from visible and infrared modalities. In daytime scenes, visible images provide rich appearance and boundary information, while infrared inputs may suffer from limited semantic details. However, relying solely on visible information can still lead to incomplete segmentation of small or ambiguous objects. In nighttime scenes, infrared imagery provides stronger responses for objects under low illumination, but infrared-only segmentation may lack sufficient spatial and structural details, resulting in inaccurate boundaries or confused predictions. The infrared-only Mask2Former shows degraded segmentation quality in some regions, especially for objects requiring precise shape information. By jointly utilizing visible and infrared representa-

tions, our method better preserves object structures while maintaining robustness to illumination variations. Overall, the proposed visible-infrared extension effectively combines the complementary characteristics of visible and infrared modalities, leading to more reliable segmentation results across different scene conditions.

D Additional Ablation Studies

Table 8: Ablation study of unidirectional and bidirectional cross-modal enhancement strategies on the LLVIP and FLIR datasets.

Variant	LLVIP		FLIR	
	mAP	**mAP$_{50}$**	**mAP**	**mAP$_{50}$**
Baseline	66.7	96.5	46.2	82.7
$I \rightarrow V$ only	67.6	96.8	47.7	85.8
$V \rightarrow I$ only	68.1	97.0	47.9	86.0
Bidirectional	**69.0**	**97.3**	**48.4**	**86.3**

Effect of the Bidirectional Enhancement Strategy. We investigate the effectiveness of bidirectional enhancement by comparing a single-representation fusion baseline, two unidirectional variants, and the bidirectional design. As shown in Table 8, both $I \rightarrow V$ and $V \rightarrow I$ consistently improve over the baseline, achieving $67.6\%/47.7\%$ and $68.1\%/47.9\%$ mAP on LLVIP/FLIR, respectively. This demonstrates the benefit of preserving one modality as the primary representation while incorporating complementary cues from the other. However, each unidirectional variant provides only a single modality-enhanced representation, leaving no alternative representation for subsequent modality preference selection. In contrast, bidirectional enhancement maintains both visible-primary and infrared-primary representations, thereby providing a selection space for different object queries. Even when queries simply interact with both modality-enhanced representations and their responses are combined using learnable weights, the bidirectional design further improves the performance to 69.0% and 48.4% mAP on LLVIP and FLIR, respectively. These results support the benefit of maintaining both enhancement directions and provide a stronger basis for subsequent query-wise modality preference selection.

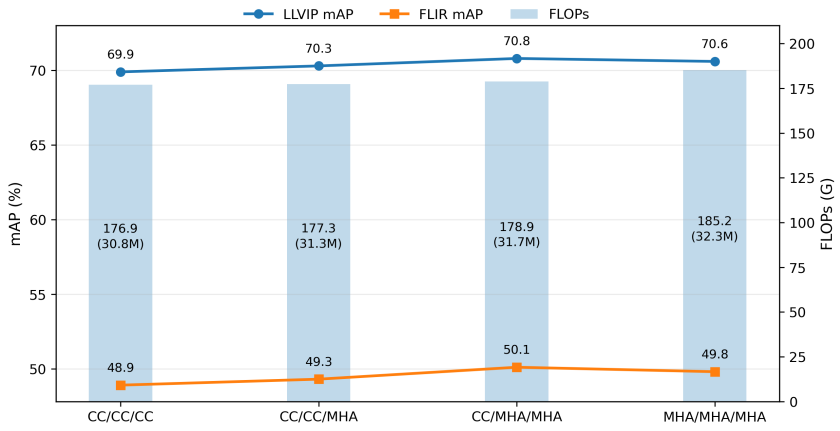

Figure 11: Scale-specific enhancement analysis. CC denotes CMCA+CMSA, and each x-axis configuration follows the order $\mathcal{E}_3/\mathcal{E}_4/\mathcal{E}_5$.

Scale-Specific Enhancement Design. We further investigate the effectiveness of adopting different enhancement mechanisms across feature scales by progressively replacing channel-spatial enhancement with cross-modal multi-head attention (MHA) from lower- to higher-resolution features, starting from $\mathcal{E}_5$, followed by $\mathcal{E}_4$, and finally $\mathcal{E}_3$. As shown in Fig. 11, introducing MHA at $\mathcal{E}_5$ and

subsequently at $\mathcal{E}_4$ progressively improves detection performance with only a moderate increase in computational cost. However, further replacing channel-spatial enhancement at the high-resolution $\mathcal{E}_3$ with MHA brings no additional performance gain while noticeably increasing FLOPs. Therefore, employing channel-spatial enhancement at $\mathcal{E}_3$ and reserving MHA for $\mathcal{E}_4$ and $\mathcal{E}_5$ provides a more favorable accuracy–complexity trade-off.

Table 9: Ablation study of layer-wise modality preference selection on LLVIP and FLIR.

Strategy	**LLVIP**		**FLIR**	
	mAP	**mAP$_{50}$**	**mAP**	**mAP$_{50}$**
First Layer Only	70.4	97.7	49.5	87.0
Layer-wise Selection	**70.8**	**97.8**	**50.1**	**87.5**

Effect of Layer-wise Preference Selection. We compare the proposed layer-wise strategy with a variant that performs modality preference selection only at the first decoder layer and keeps the selected branch fixed throughout subsequent decoding. As shown in Table 9, the First Layer Only variant achieves 70.4% mAP on LLVIP and 49.5% on FLIR. However, modality preferences inferred from early-stage queries may become suboptimal as the object queries are progressively refined across decoder layers. By re-estimating the modality preference at each layer, our layer-wise strategy allows individual queries to update their preferred modality-enhanced representation throughout the decoding process and revise potentially suboptimal decisions made at earlier stages. This further improves the performance to 70.8% mAP on LLVIP and 50.1% on FLIR, demonstrating the benefit of progressively adapting modality preferences during query refinement.

Table 10: Ablation study of different modality preference selection granularities on LLVIP and FLIR.

Selection Granularity	**LLVIP**		**FLIR**	
	mAP	**mAP$_{50}$**	**mAP**	**mAP$_{50}$**
Input-wise Selection	69.5	97.7	48.4	86.3
Object-wise Selection	**70.8**	**97.8**	**50.1**	**87.5**

Effect of Preference Selection Granularity. We further investigate the effect of modality preference selection granularity by comparing input-wise and object-wise selection. For each input pair, the input-wise variant makes a single preference selection and applies the same selection to all object queries within that input. As shown in Table 10, this strategy achieves 69.5% mAP on LLVIP and 48.4% on FLIR. While such input-level selection can capture an overall modality preference for the entire image, sharing the same preference across all queries limits its ability to accommodate the potentially different modality requirements of individual objects. Allowing each query to independently select its preferred modality-enhanced representation improves the performance to 70.8% mAP on LLVIP and 50.1% on FLIR.