DETR is Secretly a Multispectral Detector: Zero-Parameter Adaptation via Semantic Alignment

Xiangyang Li[1], Chunna Tian[1*], Zhiwei Jiang[2], Wushuai Jin[1], Pengyang Niu[1], and Lingqiao Liu[3]

[1] School of Electronic Engineering, Xidian University, Xi'an 710071, China.
{lxy, wsjin98, 24021110856}@stu.xidian.edu.cn, chnatian@xidian.edu.cn
[2] School of Physics, Xidian University, Xi'an 710071, China.
23201110762@stu.xidian.edu.cn
[3] The University of Adelaide, Adelaide 5005, Australia.
lingqiao.liu@adelaide.edu.au

Abstract. Multispectral object detection enhances perceptual robustness in challenging environments by fusing complementary information from RGB and infrared modalities. However, existing methods predominantly follow a parameter-expanding paradigm, introducing customized fusion modules that inevitably increase computational overhead and disrupt the topological consistency with unimodal pre-trained weights. In this paper, we challenge this convention and propose ZPA-MDETR, a zero-parameter adaptive multispectral framework without additional learnable fusion parameters. By re-examining the DETR architecture, we reveal that its inherent query-memory interaction mechanism provides a natural foundation for cross-modal fusion. Specifically, ZPA-MDETR strictly preserves the architecture of the unimodal DETR and achieves flexible and effective fusion via an asymmetric input organization strategy, where one modality serves as the memory, and the other initializes object queries. To bridge the cross-modal distribution discrepancy, we introduce parameter-free semantic alignment constraints during training. Extensive experiments on four public datasets across different scenes demonstrate that ZPA-MDETR outperforms state-of-the-art methods while maintaining the same parameter scale as the unimodal framework, establishing a favorable trade-off between detection accuracy and deployment efficiency. The source code is available at https://github.com/UserXiangYang/ZPA-MDETR.

Keywords: Multispectral object detection · DETR-based detector · Zero-parameter adaptation · Semantic alignment

1 Introduction

Precise and reliable object detection is a critical prerequisite in fields such as autonomous driving and all-weather surveillance [24], [15]. Visible imaging provides

* Corresponding author.

rich texture and appearance information, but its performance degrades significantly in low-light or adverse weather conditions. In contrast, infrared imaging relies on the thermal radiation emitted by objects and is thus insensitive to illumination variations, yet it typically lacks fine texture details. Consequently, multispectral object detection (MOD), which aims to fuse complementary information from RGB and infrared (RGB-IR) modalities to improve robustness in diverse and complex environments, has attracted increasing attention as an effective solution to the limitations of single-modality perception [8], [45], [36], [34].

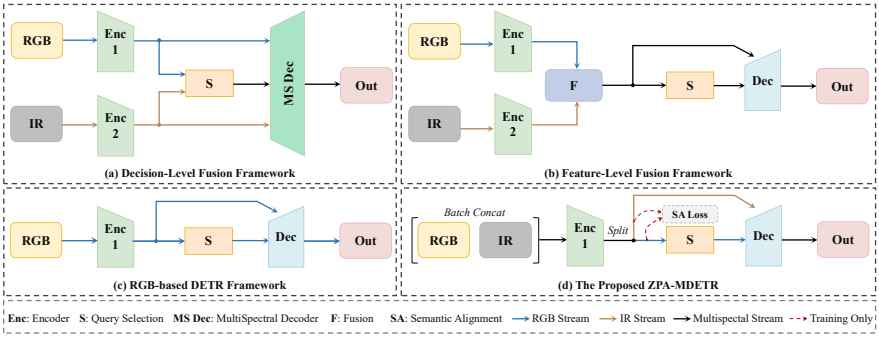

Fig. 1: Structural comparison between existing DETR-based multispectral detection frameworks and the proposed ZPA-MDETR, which preserves the architecture of the single-modal DETR.

Currently, most MOD methods follow two architectural paradigms: CNN-based and DETR-based. CNN-based detectors are known for their engineering practicality [32], whereas DETR-based detectors are favored for their end-to-end optimization and structural simplicity [20]. Although naive input-level fusion provides a parameter-free alternative, the inherent modality gap between RGB and IR inputs can limit the effective exploitation of complementary cross-modal information [21]. Consequently, to achieve robust cross-modal interaction, existing MOD methods [8], [36], [30], [45], [12], [13] predominantly adopt a parameter-expanding design paradigm. Taking DETR-based frameworks as an example, as shown in Fig. 1(a) and Fig. 1(b), both feature-level and decision-level fusion methods introduce additional structural complexity. Although such customized designs can yield performance gains, they also bring two practical limitations. First, their added fusion modules usually lack corresponding unimodal pretrained weights for initialization [8], which limits the full exploitation of pretrained representations. Second, additional fusion networks increase the parameter size and memory footprint, creating a bottleneck for deploying multispectral detectors on resource-constrained edge devices.

This situation raises a fundamental scientific question: *In MOD tasks, does achieving efficient cross-modal fusion necessarily require modifying unimodal network architectures or introducing additional model parameters?*

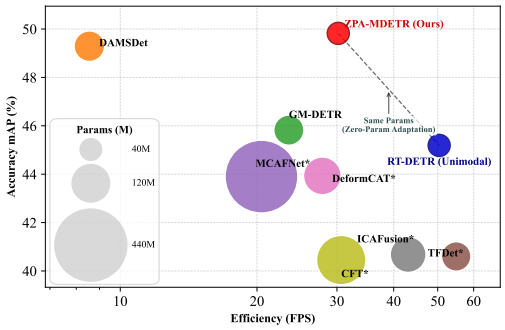

Fig. 2: Efficiency comparison of different MOD methods. Methods marked with * are CNN-based, while the others are DETR-based. RT-DETR is a unimodal detector.

This paper provides a negative answer. We re-examine the DETR architecture and observe that its decoding process fundamentally follows a query-memory information interaction paradigm, in which object queries aggregate object-level representations from the feature memory via cross-attention within a unified semantic space. This indicates that, with effective semantic alignment between modalities, the inherent query-memory mechanism in DETR can implicitly model cross-modal interactions without relying on explicit fusion structures or additional parameter expansion.

Based on the observations above, we propose a zero-parameter adaptive multispectral DETR framework, termed ZPA-MDETR. Here, zero-parameter refers to introducing no additional learnable fusion modules and preserving the same parameter scale and network topology as the unimodal DETR. As shown in Fig. 1(c) and Fig. 1(d), ZPA-MDETR fully preserves the network architecture and parameter scale of the single-modal DETR. The RGB-IR modalities are concatenated along the batch dimension and encoded by a shared encoder, followed by an asymmetric input organization strategy that assigns one modality as memory and the other as object queries. These are then processed by the DETR decoder via cross-attention. This design ensures that multimodal information is collaboratively modeled solely by modifying input organization, without introducing additional network structures or learnable fusion parameters.

Additionally, prior DETR-based studies [26], [23] suggest that effective query-memory interactions benefit from a consistent latent semantic space. To alleviate cross-modal distribution discrepancies, we introduce parameter-free semantic alignment constraints during training. Specifically, spatial alignment enhances semantic consistency across modalities at corresponding spatial locations, while channel alignment reduces cross-modal redundancy through cross-correlation modeling. We further observe that the optimal modality assignment for memory and object query is correlated with unimodal information quality, as discussed in Sec. 4.5. This observation motivates semantic alignment to improve the compatibility between modality-specific memories before decoder interaction, thereby supporting more reliable query-memory interaction.

Owing to the aforementioned design, ZPA-MDETR enables flexible and effective cross-modal information interaction without modifying the unimodal DETR architecture or introducing additional learnable fusion parameters. As shown in Fig. 2, ZPA-MDETR preserves the same parameter scale as the unimodal framework while achieving a favorable trade-off between accuracy and model size. Although dual-stream inputs reduce inference speed due to increased data throughput, this overhead stems solely from the enlarged input scale rather than architectural redundancy. Under the ZPA-MDETR framework, MOD shifts from structural fusion design to semantic alignment within a unified representation space, which can be realized via parameter-free training-stage regularizers.

Overall, our contributions are highlighted as follows:

- We reveal the inherent limitations of the parameter-expanding MOD designs in terms of deployment efficiency and newly introduced fusion parameters, and we advocate a shift toward parameter-preserving cross-modal modeling.
- We propose ZPA-MDETR, a zero-parameter adaptive MOD framework that strictly preserves the unimodal DETR architecture and parameter scale, enabling efficient multimodal collaboration through asymmetric input organization and query-memory interaction.
- We introduce parameter-free spatial and channel semantic alignment constraints during training to mitigate cross-modal distribution discrepancies and support query-memory interaction.
- Extensive experiments demonstrate that ZPA-MDETR achieves superior performance over existing methods while maintaining unimodal model size, establishing a favorable trade-off among accuracy, efficiency, and scale.

2 Related Work

2.1 Multispectral Object Detection

The advancement of MOD has been largely driven by the emergence of large-scale multispectral benchmarks [41], [16], [22], [46]. These MOD frameworks typically extend single-modality detectors into dual-stream architectures, broadly categorized by their fusion paradigms and alignment mechanisms.

To effectively exploit cross-modal complementary information, existing methods are mainly divided into feature-level and decision-level approaches [19]. Feature-level fusion methods [50], [30], [35], [49] introduced dedicated modules to facilitate intermediate feature interaction, whereas decision-level fusion detectors [3], [18] adaptively aggregated predictions through dynamic weighting or ensemble mechanisms. More recently, transformer-based components, such as self-attention [29], cross-attention [42], and positional encoding [11], have been leveraged to improve cross-modal interactions.

Beyond fusion design, auxiliary supervision has been widely explored to improve feature representation. Specifically, illumination-aware classification has been used to adaptively weight RGB-IR fusion [50], [12], while pixel-level segmentation has guided the network to focus on object-relevant regions [45], [19].

Other auxiliary tasks, such as edge detection [12], [39] and super-resolution [43], have also been utilized to guide multispectral information fusion.

Despite these advances, customized fusion strategies usually introduce modules beyond the unimodal detector topology. Although pretrained weights can still be reused for structurally consistent components, the additional fusion modules require new initialization and increase the overall parameter size. These constraints present bottlenecks for deploying multispectral detectors on resource-constrained edge devices. This challenge motivates the development of principled frameworks that can effectively exploit cross-modal complementary information while maintaining architectural simplicity.

2.2 DETR-based Object Detection

DETR-based detectors have attracted considerable attention due to their end-to-end formulation, which eliminates hand-crafted components such as anchor design and non-maximum suppression [2].

Subsequent research primarily focused on improving convergence speed and detection accuracy. Specifically, Deformable DETR [52] introduced deformable attention to sample sparse key points, significantly accelerating training and improving performance on multi-scale objects. Anchor DETR [33] and DAB-DETR [23] refined query initialization strategies by incorporating anchor priors or encoding queries as dynamic box coordinates, respectively, thereby stabilizing training and enhancing localization. DINO [40] improved performance through a denoising training strategy. RT-DETR [48] established the first practical real-time end-to-end detection transformer, balancing accuracy and speed. Building upon RT-DETR, DEIM [14] addressed convergence bottlenecks caused by sparse supervision through an efficient training paradigm. Despite these advances, the aforementioned methods are designed for single-modal object detection.

Recently, DETR-based architectures have been adapted for MOD. Specifically, MS-DETR [37] introduced modality-specific backbones and cross-attention mechanisms to fuse RGB-IR features. GM-DETR [36] proposed modality-specific interaction modules and cross-scale fusion encoders, coupled with a staged training strategy to improve generalization. MiPa [25] adopted a mixed patch training strategy for any-modal detection within the DINO framework. DAMSDet [8] employed a modality-competitive query selection strategy to dynamically choose relevant features for each object and a multispectral deformable cross-attention module to aggregate multi-level semantic features.

However, these approaches introduced dedicated fusion encoders or additional interaction modules, leading to increased architectural complexity. To overcome this limitation, the proposed ZPA-MDETR builds upon the single-modal DETR framework and achieves effective multimodal information fusion through a redesigned input organization and parameter-free semantic alignment constraints, without introducing additional learnable fusion modules.

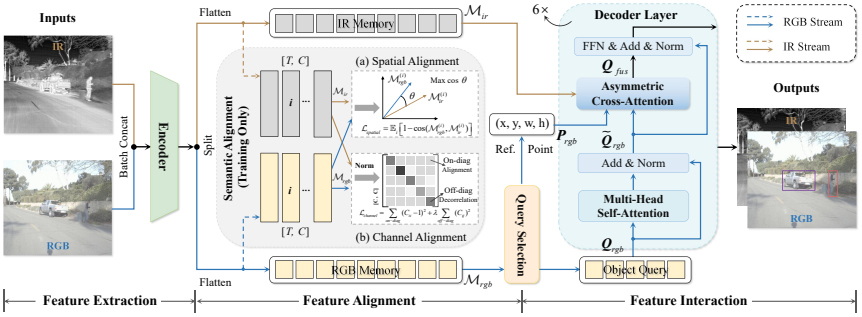

Fig. 3: Overview of the proposed ZPA-MDETR framework, which preserves the network topology of single-modal DETR (*e.g.*, RT-DETR [48]). The framework comprises three main components: a shared encoder for multimodal feature extraction, parameter-free semantic alignment for enhancing cross-modal consistency, and a standard DETR decoder that performs cross-modal interaction via asymmetric query-memory assignment. For clarity, the iterative update of reference points in the decoder is omitted.

3 Methodology

3.1 Overall Framework

The overall architecture of ZPA-MDETR is shown in Fig. 3. ZPA-MDETR preserves the network topology and parameter scale of the single-modal DETR framework. Effective cross-modal modeling is achieved solely through semantic role assignment and parameter-free semantic alignment constraints. Specifically, RGB-IR images are processed by a shared encoder for unified feature extraction, while asymmetric roles are assigned to the two modalities in the decoder. Cross-modal information interaction is then realized via the original deformable cross-attention mechanism in DETR. This design enables ZPA-MDETR to leverage intrinsic multimodal modeling capabilities without architectural modifications or additional learnable fusion parameters.

3.2 Asymmetric Input Organization

Given an RGB image $\boldsymbol{I}_{rgb} \in \mathbb{R}^{3\times H\times W}$ and an IR image $\boldsymbol{I}_{ir} \in \mathbb{R}^{1\times H\times W}$, where H and W denote the image height and width, respectively, we first convert $\boldsymbol{I}_{ir}$ into a three-channel input $\boldsymbol{I}'_{ir} \in \mathbb{R}^{3\times H\times W}$ via channel-wise replication. The RGB and IR inputs are then concatenated along the batch dimension to construct a unified input, formulated as:

$$\boldsymbol{I}_{unified} = \text{Concat}(\boldsymbol{I}_{rgb}, \boldsymbol{I}'_{ir}), \tag{1}$$

where $\boldsymbol{I}_{unified} \in \mathbb{R}^{2\times 3\times H\times W}$. This unified input is processed by an encoder to extract multi-scale features for both modalities. The encoded features are split along the batch dimension and flattened into modality-specific memories:

$$\mathcal{M}_{rgb}, \mathcal{M}_{ir} = \text{Flatten}(\text{Split}(\text{Encoder}(\boldsymbol{I}_{unified}))), \tag{2}$$

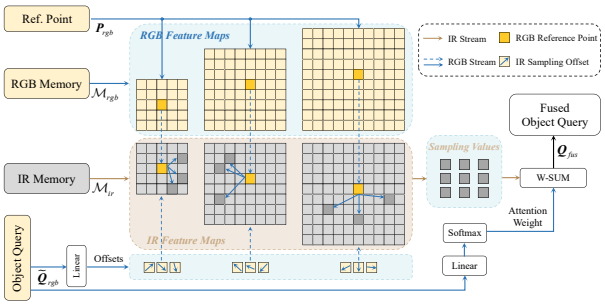

Fig. 4: The pipeline of the asymmetric cross-attention module. The module takes the RGB query $\widetilde{\boldsymbol{Q}}_{rgb}$ as input to predict sampling offsets and attention weights, which are used to sample features from the IR memory $\mathcal{M}_{ir}$ based on the RGB reference points $\boldsymbol{P}_{rgb}$. The sampled IR features are then aggregated to produce the fused query $\boldsymbol{Q}_{fus}$.

where $\mathcal{M}_{rgb}, \mathcal{M}_{ir} \in \mathbb{R}^{T\times C}$ denote the RGB and IR memories, respectively. Here, T represents the total spatial sequence length of each memory, and $C = 256$ denotes the unified channel dimension. Subsequently, the RGB-IR feature interaction strictly inherits the query selection strategy and decoder cross-attention mechanism from the standard single-modal DETR framework [48]. The core distinction lies in the asymmetric data flow governed by our semantic role assignment. Specifically, the RGB memory $\mathcal{M}_{rgb}$ is fed into the query selection scheme to generate the initial object queries $\boldsymbol{Q}_{rgb} \in \mathbb{R}^{K\times C}$ and the corresponding reference points $\boldsymbol{P}_{rgb} \in \mathbb{R}^{K\times 4}$, while the global contextual information is provided by the IR memory $\mathcal{M}_{ir}$ during cross-attention. Here, $K = 500$ denotes the total number of queries, comprising 300 top-ranked selected queries based on confidence scores and 200 denoising queries. The denoising queries are incorporated to stabilize training [17], [40].

Asymmetric Deformable Cross-Attention. Following the DETR architecture [2], we employ a six-layer cascaded decoder with identical layer structures. Taking the first layer as an example, the initial object queries $\boldsymbol{Q}_{rgb} \in \mathbb{R}^{K\times C}$ are first processed by multi-head self-attention to yield the context-enhanced queries $\widetilde{\boldsymbol{Q}}_{rgb} \in \mathbb{R}^{K\times C}$. As shown in Fig. 4, $\widetilde{\boldsymbol{Q}}_{rgb}$ interacts with the multi-scale IR memory $\mathcal{M}_{ir}$ via asymmetric deformable cross-attention, yielding the fused queries $\boldsymbol{Q}_{fus} \in \mathbb{R}^{K\times C}$. The formulation is defined as follows:

$$\boldsymbol{Q}_{fus} = \widetilde{\boldsymbol{Q}}_{rgb} + \sum_{h=1}^{N_h} \boldsymbol{W}_h \left[\sum_{n=3}^{N_l} \sum_{s=1}^{N_s} A_{rgb}^{(hns)} \cdot \boldsymbol{W}'_h \mathcal{M}_{ir}^{(n)} \left(\Phi_n(\hat{\boldsymbol{p}}_{rgb}) + \Delta \boldsymbol{p}_{rgb}^{(hns)} \right) \right], \quad (3)$$

where h, n, and s index the attention head, feature level, and sampling point, respectively. In our implementation, we set $N_h = 8$, $N_l = 5$, and $N_s = 3$. The matrices $\boldsymbol{W}_h$ and $\boldsymbol{W}'_h$ represent learnable linear projections of the h-th attention head. The attention weight $A_{rgb}^{(hns)}$ and the sampling offset $\Delta \boldsymbol{p}_{rgb}^{(hns)}$ correspond to the s-th sampling location in the h-th attention head at the n-th feature level. These two quantities are dynamically predicted from the RGB object query

$\widetilde{\boldsymbol{Q}}_{rgb}$, enabling RGB-guided feature sampling from the IR modality. The function $\Phi_n(\hat{\boldsymbol{p}}_{rgb})$ projects the normalized 2D reference point $\hat{\boldsymbol{p}}_{rgb}$, derived from the center of the predicted RGB reference box $\boldsymbol{P}_{rgb}$, onto the coordinate system of the n-th feature level. To improve scale robustness and stabilize optimization, the sampling offset $\Delta \boldsymbol{p}_{rgb}^{(hns)}$ is obtained by scaling the raw offset prediction using the width and height of $\boldsymbol{P}_{rgb}$ via element-wise multiplication.

3.3 Parameter-Free Semantic Alignment

Through semantic role assignment, RGB-derived queries can leverage the stable structural priors encoded in the IR memory to facilitate effective cross-modal interactions. Since the decoder architecture is inherently modality-agnostic, the native attention mechanism can naturally aggregate complementary information once cross-modal features are aligned within a unified semantic space [26], [23]. To this end, we introduce a set of parameter-free spatial-channel semantic alignment constraints during training that act as semantic regularizers to promote cross-modal consistency.

Spatial Semantic Alignment. To enforce spatial consistency across modalities, we impose a spatial alignment constraint on the RGB and IR memories $\{\mathcal{M}_{rgb}, \mathcal{M}_{ir}\}$ at corresponding spatial locations, formulated as:

$$\mathcal{L}_{spatial} = 1 - \frac{1}{T}\sum_{t=1}^{T} \frac{\mathcal{M}_{rgb}^{(t)} \cdot \mathcal{M}_{ir}^{(t)}}{\|\mathcal{M}_{rgb}^{(t)}\|_2 \|\mathcal{M}_{ir}^{(t)}\|_2}, \tag{4}$$

where t indexes the spatial feature vector, and T denotes the total spatial sequence length of each memory, as described in Sec. 3.2. This constraint encourages the shared encoder to learn spatially consistent semantic representations, enabling queries derived from RGB reference points to reliably attend to the corresponding regions in the IR memory.

Channel Semantic Alignment. To prevent feature collapse and reduce cross-modal redundancy, we further introduce a channel-level semantic alignment constraint. Let $\mathcal{Z}_{rgb}$ and $\mathcal{Z}_{ir}$ denote the batch-normalized representations of the memories $\{\mathcal{M}_{rgb}, \mathcal{M}_{ir}\}$ along the channel dimension. We compute the cross-correlation matrix $\mathcal{C} \in \mathbb{R}^{C \times C}$ between the two modalities as:

$$\mathcal{C}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{Z}_{rgb}^{(t,i)} \cdot \mathcal{Z}_{ir}^{(t,j)}, \tag{5}$$

where i and j index the channel dimension. The objective of channel semantic alignment is to promote a diagonal correlation structure, formulated as:

$$\mathcal{L}_{channel} = \sum_{i} (1 - \mathcal{C}_{ii})^2 + \lambda \sum_{i} \sum_{j \neq i} \mathcal{C}_{ij}^2, \tag{6}$$

where $\lambda = 0.005$. The diagonal term aligns corresponding channels to ensure cross-modal semantic consistency, while the off-diagonal term penalizes inter-channel correlations to reduce redundancy and promote feature diversity.

Table 1: Performance comparisons on the FLIR dataset. All FPS results are measured on an NVIDIA RTX 3090 GPU. † denotes that query and memory are sourced from IR and RGB, respectively.

Method	Modal	Backbone	Accuracy (%)						Efficiency	
			mAP	mAP_{75}	mAP_{50}	AP_{50}			FPS	Params (M)
						Person	Car	Bicycle		
CNN-based Multispectral Object Detectors										
CFT [29]	RGB-IR	CSPDarknet53	40.45	35.17	78.35	84.09	89.50	61.48	30.66	205.92
MMI-Det [39]	RGB-IR	CSPDarknet53	40.50	35.80	79.80	-	-	-	29.49	205.92
EI^2Det [12]	RGB-IR	CSPDarknet53	-	-	80.20	84.90	89.40	66.30	52.48	116.83
RDGNet [34]	RGB-IR	YOLOv8	-	-	80.60	85.60	91.00	65.30	-	-
ICAFusion [30]	RGB-IR	CSPDarknet53	40.68	33.88	82.84	84.93	89.81	73.78	43.04	108.50
MCAFNet [49]	RGB-IR	CSPDarknet53	43.90	40.20	81.50	-	-	-	20.46	447.78
CDC-YOLOFusion [35]	RGB-IR	CSPDarknet53	44.70	-	83.10	-	-	-	30.81	161.11
$ProbEn_3$-w/GAFF [3]	RGB-IR	ResNet101	45.48	42.86	83.68	86.87	89.16	75.02	-	-
TFDet [45]	RGB-IR	Swin-Tiny	46.60	-	86.60	-	-	-	54.92	73.27
DeformCAT [13]	RGB-IR	CSPDarknet53	46.93	43.65	86.48	89.89	91.86	77.69	27.86	120.11
DETR-based Multispectral Object Detectors										
RT-DETR [48]	RGB	ResNet50	33.88	27.61	69.47	69.17	82.33	56.90	50.37	40.88
DEIM-RT-DETR [14]	RGB	ResNet50	34.52	28.45	69.37	69.56	83.50	55.06	50.37	40.88
DINO [40]	RGB	ConvNeXt	35.77	29.69	72.23	75.64	83.92	57.13	20.02	45.13
MiPa [25]	RGB-IR	ConvNeXt	44.80	41.80	81.30	-	-	-	20.02	45.13
RT-DETR [48]	IR	ResNet50	45.19	41.84	81.71	85.07	90.87	69.17	50.37	40.88
DEIM-RT-DETR [14]	IR	ResNet50	45.75	42.83	81.73	85.51	91.49	68.18	50.37	40.88
GM-DETR [36]	RGB-IR	ResNet50	45.80	42.60	83.90	-	-	-	23.51	76.77
DINO [40]	IR	ConvNeXt	46.48	46.56	82.07	86.37	90.70	69.15	20.02	45.13
DAMSDet [8]	RGB-IR	ResNet50	49.29	48.10	86.65	89.74	92.56	77.64	8.55	78.91
ZPA-MDETR† (Ours)	RGB-IR	ResNet50	48.16	44.89	86.65	88.95	91.70	**79.28**	30.20	40.88
ZPA-MDETR (Ours)	RGB-IR	ResNet50	**49.82**	**48.32**	**87.05**	**90.03**	**92.93**	78.19		

3.4 Total Loss Function

The overall training objective consists of the detection task loss and the proposed semantic alignment regularization terms, defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \mathcal{L}_{spatial} + \mathcal{L}_{channel}, \tag{7}$$

where $\mathcal{L}_{det}$ denotes the standard detection loss of RT-DETR [48].

4 Experiments

4.1 Dataset and Metric

FLIR dataset. We utilize the aligned FLIR dataset [41], which comprises 5,142 RGB-IR image pairs with a resolution of 640×512. The dataset is divided into 4,129 training pairs and 1,013 testing pairs annotated with three categories including person, car, and bicycle.

M3FD dataset. The M3FD dataset [22] consists of 4,200 aligned RGB-IR pairs with a resolution of 1024×768, covering six categories: person (people), car, bus, motor (motorcycle), lamp (traffic light), and truck. Following the "M3FD-zxSplit" protocol [44], 2,905 pairs are used for training and 1,295 for testing.

LLVIP dataset. The LLVIP dataset [16] contains 15,488 strictly aligned RGB-IR image pairs with a resolution of 1280×1024, primarily captured under

Table 2: Performance comparisons on the M3FD dataset. † denotes that query and memory are sourced from IR and RGB, respectively.

Method	Modal	Backbone	mAP	mAP_{75}	mAP_{50}	AP_{50}					
						Person	Car	Bus	Motor	Lamp	Truck
CNN-based Multispectral Object Detectors											
ShaPE [44]	RGB-IR	ResNet50	35.40	36.80	55.70	65.80	79.10	63.00	41.90	30.20	54.20
EME [44]	RGB-IR	ResNet50	37.10	39.60	57.60	68.50	81.30	63.30	42.70	35.90	53.90
CFT [29]	RGB-IR	CSPDarknet53	40.39	42.63	66.93	76.26	87.04	65.28	48.61	65.00	59.40
TFDet [45]	RGB-IR	CSPDarknet53	41.00	-	64.80	-	-	-	-	-	-
EI^2Det [12]	RGB-IR	CSPDarknet53	41.26	42.13	69.58	77.52	87.03	66.57	57.72	64.07	64.58
ICAFusion [30]	RGB-IR	CSPDarknet53	41.43	42.02	69.82	77.56	87.51	69.36	64.37	57.45	62.68
DeformCAT [13]	RGB-IR	CSPDarknet53	46.49	49.78	73.28	79.96	89.53	80.22	**68.11**	56.25	65.64
DETR-based Multispectral Object Detectors											
DEIM-RT-DETR [14]	IR	ResNet50	37.52	39.42	60.35	78.94	81.68	61.09	51.72	31.44	57.25
MiPa [25]	RGB-IR	ConvNeXt	38.95	40.25	63.67	83.32	84.42	60.79	58.78	32.45	62.27
RT-DETR [48]	IR	ResNet50	38.33	38.86	62.51	79.42	80.43	66.67	57.12	33.02	58.38
RT-DETR [48]	RGB	ResNet50	40.36	40.60	66.01	57.59	86.63	69.52	57.53	60.44	64.34
DEIM-RT-DETR [14]	RGB	ResNet50	40.90	41.70	65.65	58.41	87.83	66.67	54.55	66.17	60.26
GM-DETR [36]	RGB-IR	ResNet50	45.72	47.15	71.25	79.76	87.62	72.82	57.40	65.40	64.50
MM-DETR [9]	RGB-IR	ResNet50	-	-	73.39	82.53	89.05	74.30	62.39	66.35	65.63
ZPA-MDETR (Ours)	RGB-IR	ResNet50	48.39	50.68	74.36	**84.32**	90.25	76.03	64.59	64.09	66.89
ZPA-MDETR† (Ours)	RGB-IR	ResNet50	**50.58**	**53.18**	**77.26**	82.99	**91.44**	**83.16**	66.58	**69.29**	**70.10**

low-light conditions. The dataset is split into 12,025 training pairs and 3,463 testing pairs, with annotations limited to the person category.

RGBTDronePerson dataset. The RGBTDronePerson dataset [46] comprises 6,125 RGB-IR image pairs with a resolution of 640×512. It is characterized by small object scales and dense scenes, and is split into 4,900 training pairs and 1,225 testing pairs. Following [46], experiments are conducted on the person, rider, and crowd categories, while the uncertain class is excluded.

Evaluation Metrics. Following standard protocols, we adopt mean average precision (mAP) [27] as our evaluation metric. We report mAP, mAP_{50}, and mAP_{75} to evaluate overall performance, along with class-wise AP_{50}.

4.2 Implementation Details

Our ZPA-MDETR is built upon the single-modal RT-DETR framework [48], a representative DETR-based detector, following its original architecture and data augmentation settings. Unless otherwise specified, ZPA-MDETR uses RGB features to initialize object queries and IR features as memory. A ResNet50 backbone [10] pretrained on the COCO dataset is used in the main experiments. To examine backbone robustness, we further evaluate ZPA-MDETR using different backbones, and provide the detailed results in the Supplementary Material. To ensure a fair comparison, input images are resized to 640×640 for the FLIR [41], M3FD [22], and RGBTDronePerson [46] datasets, and to 1024×1024 for the LLVIP [16] dataset. The number of training epochs is set to 40, 50, 15, and 40 for FLIR, M3FD, LLVIP, and RGBTDronePerson, respectively, with corresponding batch sizes of 4, 6, 4, and 6. All models are trained end-to-end on a single NVIDIA GeForce RTX 3090 GPU.

Table 3: Performance comparisons on the LLVIP dataset. † denotes that query and memory are sourced from IR and RGB, respectively.

Method	Modal	Backbone	mAP	AP_{75}	AP_{50}
CNN-based Multispectral Object Detectors					
CSSA [1]	RGB-IR	ResNet101	59.20	66.60	94.30
$ProbEn_3$-w/GAFF [3]	RGB-IR	ResNet101	59.71	68.99	93.61
UniRGB-IR [38]	RGB-IR	ViT-Base	63.20	72.20	96.10
CMA-Det [31]	RGB-IR	CSPDarknet53	-	-	97.10
CFT [29]	RGB-IR	CSPDarknet53	63.60	72.90	97.50
Fusion-Mamba [5]	RGB-IR	CSPDarkNet53	64.30	-	97.00
M-SpecGene [51]	RGB-IR	ViT-Base	65.30	75.40	97.40
DeformCAT [13]	RGB-IR	CSPDarknet53	66.13	77.07	97.82
SIA [4]	RGB-IR	CSPDarkNet53	67.30	-	97.50
DETR-based Multispectral Object Detectors					
RT-DETR [48]	RGB	ResNet50	53.70	58.14	91.28
DEIM-RT-DETR [14]	RGB	ResNet50	53.99	58.40	91.79
DINO [40]	RGB	ConvNeXt	54.46	59.68	91.99
RT-DETR [48]	IR	ResNet50	66.01	77.24	96.74
MS-DETR [37]	RGB-IR	ResNet50	66.10	76.30	97.90
DINO [40]	IR	ConvNeXt	67.71	79.12	97.34
DEIM-RT-DETR [14]	IR	ResNet50	67.90	**80.50**	97.54
ZPA-MDETR† (Ours)	RGB-IR	ResNet50	64.88	76.26	97.82
ZPA-MDETR (Ours)	RGB-IR	ResNet50	**69.14**	78.79	**98.11**

4.3 Comparison with State-of-the-Art Methods

Comparisons on FLIR. As shown in Table 1, ZPA-MDETR achieves the best performance on the FLIR dataset in terms of the main detection metrics. Compared to the strong CNN-based DeformCAT [13], our method improves mAP by 2.89%, highlighting the advantage of transformer-based cross-modal modeling. Furthermore, compared with the second-best DETR-based DAMSDet [8], ZPA-MDETR not only achieves superior detection accuracy but also substantially improves inference efficiency. Specifically, our model runs at 30.20 FPS, which is approximately 3.5× faster than DAMSDet, while reducing the parameter count by nearly 50%. It should also be noted that parameter preservation does not imply zero computational overhead. Compared with the single-modal RT-DETR, ZPA-MDETR maintains the same parameter scale, while its FLOPs increase from 67.84G to 129.03G due to dual-modal input processing rather than additional learnable fusion modules. Overall, ZPA-MDETR achieves a favorable accuracy-efficiency trade-off by effectively leveraging complementary RGB-IR information while maintaining the parameter scale of RT-DETR [48].

Comparisons on M3FD. As shown in Table 2, ZPA-MDETR achieves 50.58% mAP, outperforming all existing MOD methods. Compared to DeformCAT [13], our method improves mAP by 4.09%. It also substantially surpasses the second-best DETR-based GM-DETR [36] by 4.86% in mAP. Notably, RT-DETR performs significantly worse under these challenging conditions, achieving only 40.36% and 38.33% mAP on RGB and IR inputs, respectively. In contrast, ZPA-MDETR yields improvements exceeding 10%, highlighting the effectiveness of the proposed method in mitigating environmental interference.

Comparisons on LLVIP. The RGB modality is severely degraded in the LLVIP dataset, resulting in a significant modality imbalance and challenging cross-modal fusion. As shown in Table 3, ZPA-MDETR achieves 69.14% mAP,

Table 4: Performance comparisons on the RGBTDronePerson dataset. † denotes that query and memory are sourced from IR and RGB, respectively.

Method	Modal	Backbone	mAP	mAP_{75}	mAP_{50}	AP_{50}		
						Person	Rider	Crowd
CNN-based Multispectral Object Detectors								
QFDet [46]	RGB-IR	ResNet50	-	-	42.08	-	-	-
DeformCAT [13]	RGB-IR	CSPDarknet53	12.73	2.87	43.41	47.44	49.15	33.64
EI^2Det [12]	RGB-IR	CSPDarknet53	12.88	2.45	44.48	48.26	46.24	38.93
CBC [47]	RGB-IR	Swin-Small	-	-	43.55	49.76	52.62	28.27
TFDet [45]	RGB-IR	CSPDarknet53	13.59	4.04	42.66	49.35	41.25	37.37
AODet [7]	RGB-IR	CSPDarknet53	-	-	47.66	-	-	-
CDC-YOLOFusion [35]	RGB-IR	CSPDarkNet53	-	-	47.70	52.20	48.00	33.80
DEGF-YOLO [6]	RGB-IR	YOLOv8	-	-	47.80	54.00	49.50	40.00
COXNet [28]	RGB-IR	ResNet50	-	-	50.04	-	-	-
DETR-based Multispectral Object Detectors								
DEIM-RT-DETR [14]	RGB	ResNet50	0.14	0.01	0.41	0.48	0.13	0.62
RT-DETR [48]	RGB	ResNet50	0.87	0.19	3.59	3.72	4.64	2.41
DEIM-RT-DETR [14]	IR	ResNet50	13.32	3.44	42.52	49.74	40.54	37.27
MiPa [25]	RGB-IR	ConvNeXt	16.77	5.96	47.13	59.98	43.36	38.05
RT-DETR [48]	IR	ResNet50	17.35	5.73	51.63	57.10	52.72	45.06
GM-DETR [36]	RGB-IR	ResNet50	20.63	9.37	54.97	59.92	59.66	45.32
ZPA-MDETR† (Ours)	RGB-IR	ResNet50	20.08	7.90	51.41	61.20	51.90	41.14
ZPA-MDETR (Ours)	RGB-IR	ResNet50	**21.01**	**10.75**	**57.13**	**64.15**	**61.10**	**46.15**

Table 5: Ablation study on the FLIR dataset. Baseline, AIO, and SA denote single-modal RT-DETR, asymmetric input organization, and semantic alignment, respectively. Spa. and Cha. indicate spatial and channel semantic alignment.

Baseline	AIO	SA		mAP	mAP_{75}	mAP_{50}	AP_{50}			Params
		Spa.	Cha.				Person	Car	Bicycle	
Query-Memory: IR-RGB										
✓				33.88	27.61	69.47	69.17	82.33	56.90	40.88
✓	✓			47.62	44.70	85.82	**89.38**	**92.15**	75.92	
✓	✓	✓		47.95	44.86	86.51	89.34	92.13	78.07	
✓	✓		✓	47.80	44.77	86.53	88.82	91.81	**79.97**	
✓	✓	✓	✓	**48.16**	**44.89**	**86.65**	88.95	91.70	79.28	
Query-Memory: RGB-IR										
✓				45.19	41.84	81.71	85.07	90.87	69.17	40.88
✓	✓			49.21	47.53	86.58	89.84	93.07	76.83	
✓	✓	✓		49.48	47.77	86.66	89.53	**93.08**	77.36	
✓	✓		✓	49.58	47.94	86.52	89.47	92.77	77.31	
✓	✓	✓	✓	**49.82**	**48.32**	**87.05**	**90.03**	92.93	**78.19**	

outperforming the CNN-based SIA [4] by 1.84% and the RGB-IR DETR-based MS-DETR [37] by 3.04%. It also surpasses the strong single-modal RT-DETR-IR [48] by 3.13%. These results demonstrate that our method effectively mitigates negative transfer from degraded RGB features while leveraging complementary spatial cues to enhance IR representations.

Comparisons on RGBTDronePerson. We evaluate ZPA-MDETR on the RGBTDronePerson dataset to assess its effectiveness in detecting tiny objects. As shown in Table 4, this dataset is challenging due to severe degradation of the RGB modality, where RT-DETR-RGB achieves only 0.87% mAP, while the IR modality reaches 17.35% mAP. Such extreme modality imbalance typically hinders effective cross-modal fusion. In contrast, ZPA-MDETR achieves 57.13% mAP_{50}, outperforming the CNN-based COXNet [28] by 7.09% in mAP_{50}. Compared to the DETR-based GM-DETR [36], our method also yields consistent

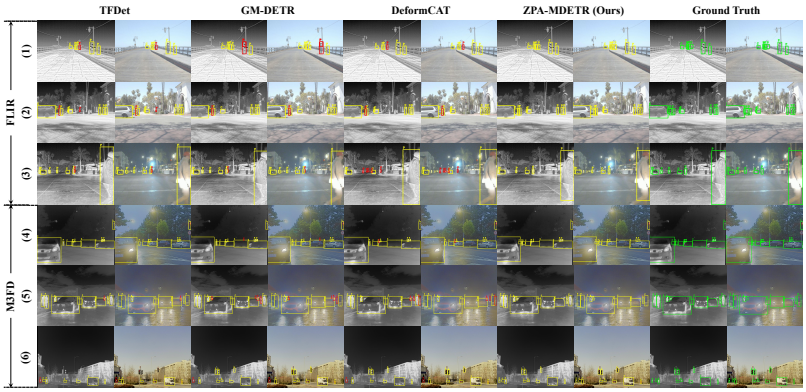

Fig. 5: Qualitative comparison on the FLIR and M3FD datasets. All detections are displayed using a unified confidence threshold of 0.5. Ground-truth annotations are shown in green, while predicted bounding boxes are highlighted in yellow for true positives and red for false positives. Class labels are indicated in the top-left corner of each box: "P" (Person), "C" (Car), and "L" (Lamp), "T" (Truck).

improvements, including a gain of 2.16% in mAP_{50}. These results demonstrate that our method effectively preserves the fine-grained spatial details of tiny objects even when one modality is severely degraded.

Robustness on unaligned DVTOD. We further evaluate ZPA-MDETR on the unaligned DVTOD dataset [31], where RGB and IR images exhibit significant spatial offsets and resolution discrepancies. ZPA-MDETR achieves the best mAP of 56.66%, demonstrating that the proposed asymmetric input organization and semantic alignment remain effective under spatially unaligned conditions. Details are provided in the Supplementary Material.

Qualitative Detection Results. As shown in Fig. 5, ZPA-MDETR consistently produces more accurate and stable predictions compared to recent state-of-the-art methods. In daytime scenes (Rows 1-2), our method effectively localizes distant, small-scale targets that are often missed by competing approaches. Under low-illumination conditions (Rows 3-4), it mitigates ambiguities caused by poor visibility and thermal crossover, resulting in fewer false detections. In adverse environments (Rows 5-6), including rain reflections and background clutter, ZPA-MDETR suppresses environmental noise while preserving object details. These qualitative results indicate that ZPA-MDETR enhances cross-modal feature consistency and enables robust detection across diverse scenarios.

4.4 Ablation Study and Further Analysis

Effect of Asymmetric Input Organization (AIO). As shown in Table 5, our AIO strategy consistently improves detection performance across both query-memory configurations. In the IR-RGB setting, AIO increases mAP by 13.74%, while in the RGB-IR setting, mAP improves by 4.02%. Additionally, consistent

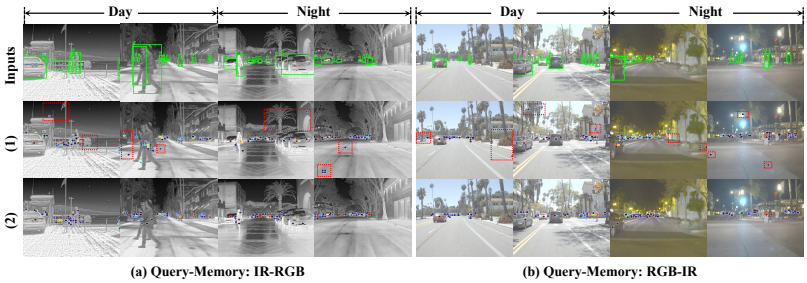

Fig. 6: Visualization of object query distributions. The top-50 (1) initial object queries and (2) refined queries from the final decoder layer. Red boxes highlight queries falling into background. The brighter points indicate higher confidence.

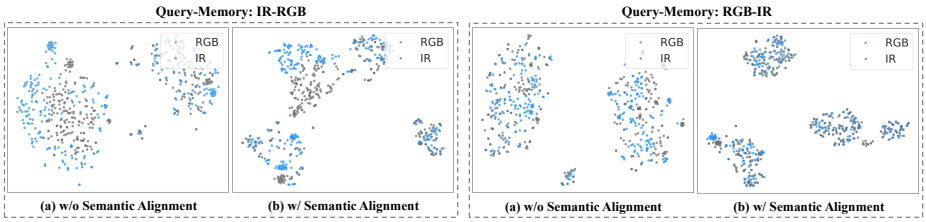

Fig. 7: Visualization of RGB and IR feature distributions using t-SNE on the FLIR (a) without and (b) with semantic alignment constraints.

gains in mAP_{75} and mAP_{50} further indicate enhanced localization precision and overall detection quality. To provide further insight, we visualize the query distributions before and after cross-modal interaction, as shown in Fig. 6. Compared to the initial query distribution at the first decoder layer, the final-layer queries after asymmetric cross-modal attention are more concentrated on object regions. These results demonstrate that AIO enhances query-memory interaction, improves spatial focus, and reduces background interference, providing a more robust foundation for subsequent detection refinement.

Effect of Semantic Alignment (SA). As shown in Table 5, semantic alignment consistently improves detection performance in both query-memory settings. In the IR-RGB setting, spatial SA and channel SA increase mAP by 0.33% and 0.18%, respectively, and their combination brings a larger gain of 0.54%. In the RGB-IR setting, they bring gains of 0.27% and 0.37%, respectively, while their combination improves mAP by 0.61%. These results confirm the complementary effects of spatial and channel SA. By jointly enhancing location-wise consistency and cross-modal channel correlation, they support more reliable query-memory interaction. To intuitively verify alignment quality, inspired by [51], we visualize the feature distributions in Fig. 7. As shown in Fig. 7(a), without SA, there is a distinct distribution gap between the RGB and IR features. In contrast, Fig. 7(b) demonstrates that, with SA, the features of both modalities are compactly clustered and well overlapped. This confirms that the

SA strategy effectively bridges the modality gap, mapping heterogeneous features into a unified semantic space.

4.5 Further Analysis of Query-Memory Modality Selection

To better understand the role of modality selection in cross-modal attention, we analyze the relationship between unimodal performance and the optimal query-memory configuration across different datasets. We do not formulate modality-role assignment as a universal theoretical rule, but view it as a reliability-based practical criterion. As shown in Tables 1, 2, 3, and 4, the modality with stronger unimodal performance usually serves as a more effective memory source, while the other modality is better suited for object queries. Consistent with the experimental results, IR is more reliable on the FLIR [41], LLVIP [16], and RGBT-DronePerson [46] datasets, which contain many low-light or modality-imbalanced scenes, whereas RGB provides stronger unimodal evidence on the relatively balanced M3FD dataset [22]. This trend is consistent with the role of memory as the key-value source in decoder cross-attention, where a more reliable modality provides cleaner evidence for object queries. Meanwhile, the other modality, serving as the query source, acts as a complementary probe to retrieve informative regions from the memory.

5 Conclusion

In this paper, we propose ZPA-MDETR, a zero-parameter adaptive multispectral DETR framework, to address the bottlenecks of deployment efficiency and pre-training compatibility in current MOD methods. Unlike existing parameter-expanding paradigms that rely on customized fusion modules, we re-examine the inherent query-memory interaction mechanism of DETR and exploit it for RGB-IR information interaction without introducing additional learnable fusion parameters. Through an asymmetric input organization strategy and training-only semantic alignment constraints, ZPA-MDETR achieves flexible and effective multimodal collaboration by assigning complementary roles to different modalities while preserving the unimodal detector architecture. Extensive experiments demonstrate that our method outperforms the existing MOD methods while maintaining the same parameter scale as the unimodal DETR framework. This work offers a novel perspective on efficient multimodal perception, advocating a shift from architectural modifications toward parameter-preserving cross-modal interaction and semantic alignment.

Acknowledgements

This research was supported in part by the National Natural Science Foundation of China under Grants 62173265, 62372355; in part by the Natural Science Basic Research Program of Shaanxi under Grant 2023JC-ZD-39.

References

1. 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 (CVPR). pp. 403–411 (2023)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (ECCV). pp. 213–229 (2020)
3. Chen, Y.T., Shi, J., Ye, Z., Mertz, C., Kong, S., Ramanan, D.: Multimodal object detection via probabilistic ensembling. In: European Conference on Computer Vision (ECCV). pp. 139–158 (2022)
4. Cui, M., Nie, J., Sun, H., Xie, J., Cao, J., Pang, Y., Li, X.: Multispectral remote sensing object detection via selective cross-modal interaction and aggregation. Neural Networks **197**, 108533 (2025)
5. 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 **27**, 7392–7406 (2025)
6. Gu, Y., Chen, W., Peng, D.: UAV-based multimodal object detection via feature enhancement and dynamic gated fusion. Pattern Recognition **172**, 112722 (2026)
7. Gui, Z., Zhang, Y., Lei, X., Zhang, R., Yang, W.: AODet: Anti-occlusion for enhanced small object detection in drone-based RGBT imagery. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS). pp. 954–958 (2024)
8. 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 (ECCV). pp. 464–481 (2024)
9. Han, J., Wang, Y., Zhang, Y., Chen, L.: MM-DETR: An efficient multimodal detection transformer with mamba-driven dual-granularity fusion and frequency-aware modality adapters. arXiv preprint arXiv: 2512.00363 (2025)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
11. Hou, S., Yang, M., Zheng, W.S., Gao, S.: Multispectral transformer fusion via exploiting similarity and complementarity for robust pedestrian detection. Pattern Recognition **162**, 111383 (2025)
12. Hu, K., He, Y., Li, Y., Zhao, J., Chen, S., Kang, Y.: 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), 7101–7115 (2025)
13. Hu, Y., Chen, X., Wang, S., Liu, L., Shi, H., Fan, L., Tian, J., Liang, J.: Deformable cross-attention transformer for weakly aligned RGB-T pedestrian detection. IEEE Transactions on Multimedia **27**, 4400–4411 (2025)
14. Huang, S., Lu, Z., Cun, X., Yu, Y., Zhou, X., Shen, X.: DEIM: DETR with improved matching for fast convergence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15162–15171 (2025)
15. Jain, D.K., Zhao, X., Gan, C., Shukla, P.K., Jain, A., Sharma, S.: Fusion-driven deep feature network for enhanced object detection and tracking in video surveillance systems. Information Fusion **109**, 102429 (2024)

16. 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 Workshops (ICCVW). pp. 3496–3504 (2021)
17. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: DN-DETR: Accelerate DETR training by introducing query denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13619–13627 (2022)
18. Li, Q., Zhang, C., Hu, Q., Zhu, P., Fu, H., Chen, L.: Stabilizing multispectral pedestrian detection with evidential hybrid fusion. IEEE Transactions on Circuits and Systems for Video Technology **34**(4), 3017–3029 (2024)
19. Li, X., Chen, S., Tian, C., Zhou, H., Zhang, Z.: M2FNet: Mask-guided multi-level fusion for RGB-T pedestrian detection. IEEE Transactions on Multimedia **26**, 8678–8690 (2024)
20. Li, Y., Miao, N., Ma, L., Shuang, F., Huang, X.: Transformer for object detection: Review and benchmark. Engineering Applications of Artificial Intelligence **126**, 107021 (2023)
21. Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. arXiv preprint arXiv: 1611.02644 (2016)
22. 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 (CVPR). pp. 5802–5811 (2022)
23. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: DAB-DETR: Dynamic anchor boxes are better queries for DETR. In: International Conference on Learning Representations (ICLR) (2021)
24. Maiettini, E., Pasquale, G., Rosasco, L., Natale, L.: On-line object detection: A robotics challenge. Autonomous Robots **44**(5), 739–757 (2020)
25. Medeiros, H.R., Latortue, D., Granger, E., Pedersoli, M.: MiPa: Mixed patch infrared-visible modality agnostic object detection. arXiv preprint arXiv: 2404.18849 (2024)
26. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional DETR for fast training convergence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3651–3660 (2021)
27. Padilla, R., Netto, S.L., Da Silva, E.A.: A survey on performance metrics for object-detection algorithms. In: International Conference on Systems, Signals and Image Processing (IWSSIP). pp. 237–242 (2020)
28. Peng, P., Xu, T., Zhu, M., Fang, Y., Li, J.: COXNet: Cross-layer fusion with adaptive alignment and scale integration for RGBT tiny object detection. IEEE Transactions on Circuits and Systems for Video Technology **36**(1), 596–608 (2026)
29. Qingyun, F., Dapeng, H., Zhaokui, W.: Cross-modality fusion transformer for multispectral object detection. arXiv preprint arXiv: 2111.00273 (2021)
30. Shen, J., Chen, Y., Liu, Y., Zuo, X., Fan, H., Yang, W.: ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. Pattern Recognition **145**, 109913 (2024)
31. Song, K., Xue, X., Wen, H., Ji, Y., Yan, Y., Meng, Q.: Misaligned visible-thermal object detection: A drone-based benchmark and baseline. IEEE Transactions on Intelligent Vehicles **9**(11), 7449–7460 (2024)
32. Wang, C.Y., Liao, H.Y.M.: YOLOv1 to YOLOv10: The fastest and most accurate real-time object detection systems. arXiv preprint arXiv:2408.09332 (2024)

33. Wang, Y., Zhang, X., Yang, T., Sun, J.: Anchor DETR: Query design for transformer-based object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 2567–2575 (2022)
34. Wang, Z., Tian, T.: Regional defeats global: An efficient regional feature fusion via convolutional architecture for multispectral object detection. Information Fusion **130**, 104110 (2026)
35. Wang, Z., Liao, X., Yuan, J., Yao, Y., Li, Z.: CDC-YOLOFusion: Leveraging cross-scale dynamic convolution fusion for visible-infrared object detection. IEEE Transactions on Intelligent Vehicles **10**(3), 2080–2093 (2025)
36. Xiao, Y., Meng, F., Wu, Q., Xu, L., He, M., Li, H.: GM-DETR: Generalized multispectral detection transformer with efficient fusion encoder for visible-infrared detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 5541–5549 (2024)
37. Xing, Y., Yang, S., Wang, S., Zhang, S., Liang, G., Zhang, X., Zhang, Y.: 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)
38. Yuan, M., Cui, B., Zhao, T., Wang, J., Fu, S., Yang, X., Wei, X.: UniRGB-IR: A unified framework for visible-infrared semantic tasks via adapter tuning. In: Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM). pp. 2409–2418 (2025)
39. Zeng, Y., Liang, T., Jin, Y., Li, Y.: MMI-Det: Exploring multi-modal integration for visible and infrared object detection. IEEE Transactions on Circuits and Systems for Video Technology **34**(11), 11198–11213 (2024)
40. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: International Conference on Learning Representations (ICLR) (2023)
41. Zhang, H., Fromont, E., Lefevre, S., Avignon, B.: Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: IEEE International Conference on Image Processing (ICIP). pp. 276–280 (2020)
42. Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., Stiefelhagen, R.: CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. IEEE Transactions on intelligent transportation systems **24**(12), 14679–14694 (2023)
43. Zhang, J., Lei, J., Xie, W., Fang, Z., Li, Y., Du, Q.: SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2023)
44. Zhang, X., Cao, S.Y., Wang, F., Zhang, R., Wu, Z., Zhang, X., Bai, X., Shen, H.L.: Rethinking early-fusion strategies for improved multispectral object detection. IEEE Transactions on Intelligent Vehicles **10**(6), 3728–3742 (2025)
45. 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 (2025)
46. Zhang, Y., Xu, C., Yang, W., He, G., Yu, H., Yu, L., Xia, G.S.: Drone-based RGBT tiny person detection. ISPRS Journal of Photogrammetry and Remote Sensing **204**, 61–76 (2023)
47. Zhang, Y., Yang, W., Xu, C., Hu, Q., Xu, F., Xia, G.S.: Mitigating the impact of prominent position shift in drone-based RGBT object detection. arXiv preprint arXiv: 2502.09311 (2025)
48. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: DETRs beat YOLOs on real-time object detection. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16965–16974 (2024)
49. Zheng, S., Junfeng, L., Zeng, J.: MCAFNet: Multiscale cross-modality adaptive fusion network for multispectral object detection. Digital Signal Processing **159**, 104996 (2025)
50. Zhou, K., Chen, L., Cao, X.: Improving multispectral pedestrian detection by addressing modality imbalance problems. In: European Conference on Computer Vision (ECCV). pp. 787–803 (2020)
51. Zhou, K., Yang, F., Wang, S., Wen, B., Zi, C., Chen, L., Shen, Q., Cao, X.: M-specgene: Generalized foundation model for RGBT multispectral vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7861–7872 (2025)
52. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (ICLR) (2021)