GM-DETR: Generalized Muiltispectral DEtection TRansformer with Efficient Fusion Encoder for Visible-Infrared Detection

Yiming Xiao, Fanman Meng*, Qingbo Wu, Linfeng Xu, Mingzhou He, Hongliang Li
University of Electronic Science and Technology of China, China
{ymxiao,hiram}@std.uestc.edu.cn, {fmmeng,qbwu,lfxu,hlli}@uestc.edu.cn

Abstract

Multispectral object detection based on RGB and IR achieves improved accurate and robust performance by integrating complementary information from different modalities. However, existing methods predominantly focus on effectively fusing information from both modalities to enhance detection performance, and rarely study the diversified utilization of RGB and IR data and explore the adaptability of the model to practical application scenarios. We first analyze the utilization of datasets for multispectral object detection, and compare their testing performance. To better leverage datasets and address more generalized model application scenarios, we propose a Generalized Multispectral DEtection TRansformer (GM-DETR) with a two-stage training strategy. Specifically, we design the Modality-Specific Feature Interaction (MSFI) module to extract the high-level information from RGB and IR, and propose the Cross-Modality-Scale feature Fusion (CMSF) module for fusing RGB and IR modalities, which performs multi-scale cross-modalities fusion. Our GM-DETR achieves state-of-the-art performance on FLIR and LLVIP benchmark datasets.

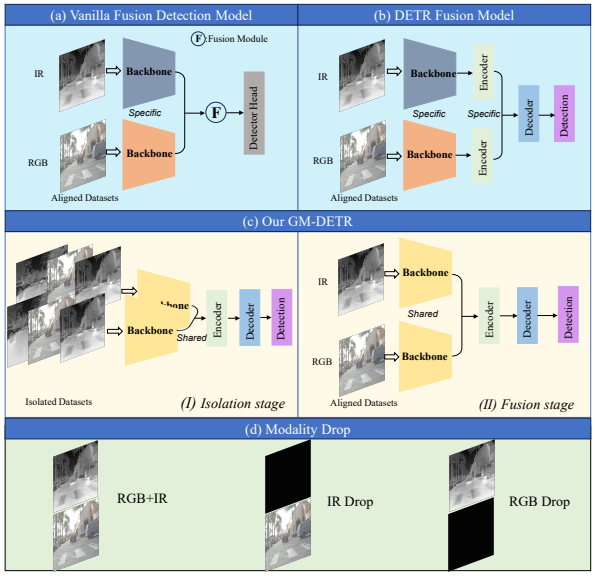

Figure 1. (a) (b) Existing multispectral object detection frameworks. (c) The two-stage GM-DETR, the two backbones in GM-DETR are parameter-shared. (d) The term "Modality Drop" describes a situation in which, within the initial RGB+IR image pair, the model encounters a missing modality image. Such instances are common in practical applications.

1. Introduction

Object detection algorithms are widely used in many applications, such as autonomous driving, and object tracking [17]. By the fact that visible (RGB) sensor is easily affected by highly dynamic and complex environmental factors, such as rain, fog, low light, and glare, infrared (IR) sensors that exhibit insensitivity to these weather changes and illumination variations are also combined to help the object detection. Intuitively, the fusion of complementary RGB and IR images can enhance the accuracy, robustness, and reliability of object detection. However, the discrepancies between modalities present a significant challenge in efficiently extracting and fusing the features of IR and RGB images.

Many multispectral object detection methods have been proposed to address this challenge. Dual-branch of detection networks are introduced to handle these two modalities [1, 7, 20, 23, 26]. The issue lies in the fact that prior methods are constrained by their reliance on modality-specific backbones, limiting models to training solely on pairs of temporally and spatially aligned IR and RGB images. This significantly restricts the available training data for the model.

Meanwhile, as shown in Fig. 1(a) and Fig. 1(b), the existing methods directly fuse IR and RGB data in the training session, this training strategy may limit the capability of model to handle only the IR and RGB two-stream input,

*Corresponding Author: (Email:fmmeng@uestc.edu.cn)

hindering a comprehensive understanding of each modality's features, as show in [23], when multispectral detection models encounter modality drop in input data, as illustrated in Fig. 1 (d), their performance significantly deteriorates compared to detection models trained on single-modality datasets. To address the issue and fully utilize isolated IR and RGB datasets, we propose a two-stage training strategy for the multispectral object detection. The first stage involves training with isolated datasets, enabling enhanced feature extraction for IR and RGB modalities. The second stage introduces fusion training with aligned datasets. A visual representation of the training procedure is presented in Fig. 1(c).

To enhance the effectiveness and feasibility of the model's two-stage training, inspired by the approach utilized in [30], we employ the backbone with shared parameters for feature extraction. Subsequent two encoders are introduced to extract independent features for each modality. Furthermore, for the detection task, the DETR-based approach seamlessly aligns with this configuration. Specifically, we introduce a novel multispectral object detector based on DETR. The encoder of DETR [2, 18, 29] obtains the CNN features through the backbone and utilizes a Transformer [24] structure to convert the features into high-dimensional representations. Leveraging the distinctive Transformer structure in the encoder of DETR, we first use parameter-shared backbones to extract IR and RGB features. Subsequently, we design a Modality-Specific Feature Interaction (MSFI) module to conduct self-attention operations on the top-level CNN features of different modalities.

The subsequent step addresses the crucial challenge of integrating information from IR and RGB modalities, a pivotal aspect for further improvement based on prior research. We introduce the Cross-Modality-Scale feature Fusion (CMSF) module, designed to facilitate the effective fusion of features from diverse modalities across multiple scales. The fusion process outputs multi-scale fused features, which are then fed into the decoder for further analysis.

The main contributions of our method can be summarized as follows:

- We introduce the two-stage training strategy for the multispectral object detection.
- We propose GM-DETR, a multispectral object detection model that integrates an efficient fusion encoder consisting of two modules. The MSFI module is designed to extract global information from different modalities. The CMSF module is proposed to efficiently fuse information across modalities and scales.
- We conduct experiments on two public datasets, i.e., FLIR [6] and LLVIP [12], and demonstrate that our proposed method outperforms other state-of-the-art methods.

2. Related Work

2.1. Multispectral Object Detection

In multispectral object detection tasks, algorithms leveraging the fusion of infrared (IR) and visible (RGB) images garner significant attention, particularly in applications such as autonomous driving and surveillance. Existing methods mostly rely on pairwise pixel-aligned IR and RGB images. The FLIR [6], KAIST [11], and LLVIP [12] are important benchmarks. The FLIR and KAIST are datasets for autonomous driving scenes, while the LLVIP is a dataset for pedestrian detection in surveillance vision.

The multispectral detection algorithms are mostly based on general object detection algorithms such as Faste-RCNN [21], RetinaNet [15], YoloV5, and DETR [2]. The key to achieving multispectral object detection lies in how to integrate the information of the two modalities, IR and RGB, including early-fusion, mid-fusion, and late-fusion.

The idea of early-fusion is to directly merge single-channel IR images with 3-channel RGB images into a 4-channel image, which is then treated as input to the conventional detection network. This direct method is the most intuitive approach. Recently, alternative algorithms have emerged utilizing image generation techniques to merge IR and RGB into a 3-channel fused image. This fused image is subsequently employed as input for training conventional detection networks.

Mid-fusion operates at intermediate feature layers. The model will have two inputs, obtaining RGB and IR features separately. By designing fusion modules to interact at the feature level, the model can better understand the correlation between the two modalities. CFT [20] extends YOLOv5 into two branches and achieves intra-modal and inter-modal information fusion by introducing Transformer modules. LRAF-Net [7] enhances feature fusion by improving modality complementarity, introducing long-range dependency, and achieving fusion through a long-range feature fusion module. CSAA [1] utilizes channel switching and spatial attention to enhance performance while minimizing computational consumption based on Faster R-CNN[21]. MS-DETR [26] utilizes modality-specific backbones and a multi-modal Transformer decoder to fuse IR and RGB features. ICAFusion [23] introduces an iterative feature fusion module to enhance features, simultaneously reducing model complexity and computational costs.

Late-fusion entails the merging of output results from two detectors that are trained separately using RGB and IR inputs. This is achieved using non-maximum suppression (NMS) to fuse the results of the two modalities. [4] proposed Probabilistic Ensembling to cope with the unaligned images from RGB and IR. However, this method increases the model's parameter count and introduce additional time for post-processing.

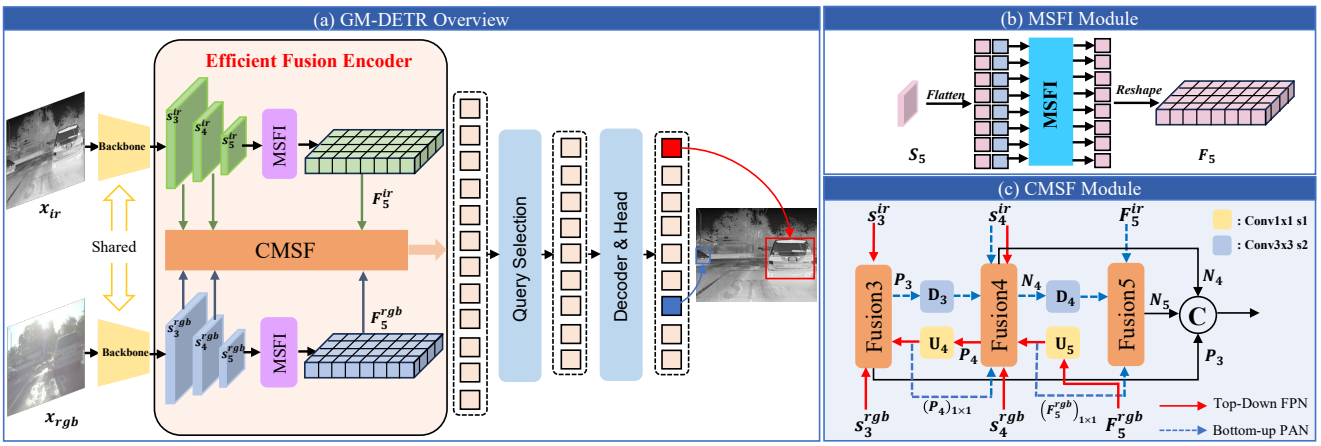

Figure 2. The architecture of our **GM-DETR**. (a) The overview: The RGB and IR images are jointly inputted into a parameter-shared backbone network. This enables the independent operation of the MSFI module for each modality to extract global information. The features from both modalities are then fused within the CMSF module. These fused features are subsequently directed into the downstream detection module. (b) The MSFI module. (c) The CMSF module.

Previous work has mostly focused on exploring how to integrate IR and RGB information without considering the utilization of data. They only consider pixel-aligned IR and RGB datasets, which not only limits the data available for model training but also to some extent fixes the model's input as one stream of RGB and one stream of IR. This may make it difficult to handle the problem of modality drop that may occur in practical application scenarios [23].

2.2. Pre-training method

The RGB pre-training methods [3, 9, 10] get significant advancements on the ImageNet [22]. Leveraging backbone networks pre-trained on ImageNet has proven to be an effective strategy for enhancing model performance in object detection tasks [8, 31]. Concurrently, [27] devise task-oriented staged pre-training methods to guide lightweight networks in learning features from visible and infrared images. For the task of multispectral object detection, besides using pre-trained models in the backbone, we design the GM-DETR network to enable pre-training supervised by detection task labels on IR and RGB datasets.

3. Methodology

As presented in Fig. 2, GM-DETR consists of two parameter-shared backbones, an efficient fusion encoder, and a Transformer decoder with prediction heads. It is noteworthy that GM-DETR is built upon the Real-Time DEtection TRansformer (RT-DETR) [18], the first real-time end-to-end object detector.

For a pair of visible and infrared image $\{x_{ir}, x_{rgb}\}$, two parameter-shared CNN backbones are firstly employed to extract features from the two-stream images. Then, the RGB and IR features of the last three stages of the backbones $\{\mathrm{S}_3^{\mathrm{ir}}, \mathrm{S}_4^{\mathrm{ir}}, \mathrm{S}_5^{\mathrm{ir}}; \mathrm{S}_3^{\mathrm{rgb}}, \mathrm{S}_4^{\mathrm{rgb}}, \mathrm{S}_5^{\mathrm{rgb}}\}$ are used as inputs of the encoder, and the Efficient Fusion Encoder is proposed to combine and transform multi-scale multispectral features into image feature sequences. Subsequently, a fixed number of image features are selected from the encoder output sequence as initial target queries through the Query Selection. These queries are finally fed into the Decoder to iteratively refine the detection results.

The model's two-stage training procedure is divided into the Isolation stage and Fusion stage, utilizing two types of data. Paired IR and RGB datasets $\{x_{ir}, x_{rgb}, y\}$ are used in the Fusion stage, where the labels y is shared between the images x_{ir} and x_{rgb}. Additionally, there are isolated annotated IR datasets $\{x_{ir}, y_{ir}\}$ and RGB datasets $\{x_{rgb}, y_{rgb}\}$, consist of labels y_{ir} and y_{rgb} corresponding to the independent IR image x_{ir} and RGB image x_{rgb}, respectively. Isolated dataset is employed in the Isolation stage. For detailed information about the datasets, please refer to Section.4.1.

The CNN backbone and Transformer decoder in our network are similar to previous works. Interested readers can refer to [18, 29] for more details. In the following sections, we will elaborate on the proposed efficient fusion encoder and two-stage training strategy.

3.1. Efficient Fusion Encoder

The architecture of the efficient fusion encoder is depicted in Fig. 2. The encoder comprises two modules, the Modality-Specific Feature Interaction (MSFI) module and the Cross-Modality-Scale feature Fusion (CMSF) module.

Modality-Specific Feature Interaction module. This module is exclusively applied to high-level feature layers $\{\mathrm{S}_5^{\mathrm{ir}}, \mathrm{S}_5^{\mathrm{rgb}}\}$, which contain abundant semantic information. Utilizing self-attention at these high-level layers improves

the ability to capture relationships between entities in the image which facilitates subsequent modules for object detection and recognition. The process can be formulated as:

$$\begin{aligned} \mathbf{Q} = \mathbf{K} = \mathbf{V} &= \text{Flatten}(\text{S}_5) \\ \text{F}_5 &= \text{Reshape}(\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V})) \end{aligned} \tag{1}$$

where $Attn$ represents multi-head self-attention, and $Reshape$ is inverse operation of $Flatten$, the shape of F_5 is the same as S_5.

Cross-Modality-Scale feature Fusion module. We propose a module for cross-modality and cross-scale feature fusion, inspired by the characteristics of RGB and IR multispectral images. This module follows a PANet-like structure [16] and consists of two branches: the Top-Down FPN [14] and the Bottom-Up PAN [16]. Firstly, we utilize F_5^{rgb} which has rich feature connection information, to perform top-down upsampling. This process helps to obtain high-resolution features with rich semantic information. Subsequently, we perform bottom-up downsampling using S_4^{ir} and F_5^{ir} to aggregate features with IR features. The module produces a feature pyramid that synthesizes multi-scale information from both the RGB and IR modalities. The FPN process can be formulated as:

$$\begin{aligned} \text{P}_4 &= \mathbf{F_4}(\mathbf{U_5}(\text{F}_5^{\text{rgb}}), \text{S}_4^{\text{rgb}}, \text{S}_4^{\text{ir}}) \\ \text{P}_3 &= \mathbf{F_3}(\mathbf{U_4}(\text{P}_4), \text{S}_3^{\text{rgb}}, \text{S}_3^{\text{ir}}) \end{aligned} \tag{2}$$

where $\mathbf{U_4}(\cdot)$ and $\mathbf{U_5}(\cdot)$ are both composed of a 1×1 convolution layer followed by an upsample operation. $\mathbf{F_3}(\cdot,\cdot,\cdot)$ and $\mathbf{F_4}(\cdot,\cdot,\cdot)$ are FPN fusion blocks that concatenate multi-level features. These blocks then generate two separate paths of output using two 1×1 convolution layers. One branch of features undergoes N *RepBlocks*, and the outputs from the two paths are fused by an element-wise addition operation.

The PAN process can be formulated as:

$$\begin{aligned} \text{N}_4 &= \mathbf{A_4}(\mathbf{D_3}(\text{P}_3), (\text{P}_4)_{1\times1}, \text{S}_4^{\text{ir}}) \\ \text{N}_5 &= \mathbf{A_5}(\mathbf{D_4}(\text{N}_4), (\text{F}_5^{\text{rgb}})_{1\times1}, \text{F}_5^{\text{ir}}) \end{aligned} \tag{3}$$

where $\mathbf{D_3}(\cdot)$ and $\mathbf{D_4}(\cdot)$ perform downsample operations as 3×3 convolution layers with a stride of 2. $(\cdot)_{1\times1}$ is the output after a 1×1 convolution layer. $\mathbf{A_4}(\cdot,\cdot,\cdot)$ and $\mathbf{A_5}(\cdot,\cdot,\cdot)$ are PAN fusion blocks, and their structure resembles the aforementioned FPN fusion blocks.

The final multi-scale fused features output by the Efficient Fusion Encoder are $\{\text{P}_3, \text{N}_4, \text{N}_5\}$, These features are fed into the Transformer decoder with prediction heads in RT-DETR [18] to obtain the detection results.

3.2. Two-stage training

The DETR framework is data-hungry [25]. The datasets available for multispectral object detection are already limited in terms of quantity. Inspired by [13], we introduced the Isolation stage on the basis of previous fusion training to enhance the feature extraction capabilities of the object detection model for both IR and RGB modalities.The whole training strategy is illustrated in Figure. 1(c).

Isolation stage. In the Isolation stage, we combined the IR and RGB datasets into a merged dataset. The two-stream input data for the model consists of two identical, replicated IR $\{x_{ir}, x_{ir}\}$ (or RGB $\{x_{rgb}, x_{rgb}\}$) images. The model is supervised by the respective labels $\{y_{ir}\}$ or $\{y_{rgb}\}$ for object detection training. This stage aims to enhance the model's feature extraction capabilities for both IR and RGB images.

Isolation stage can be conducted using data that is independently separated from the aligned data initially employed for Fusion stage training. Moreover, there is the option to utilize other merged isolated IR or RGB datasets for training.

Fusion stage. In the Fusion stage, the model is fed with temporally and spatially aligned infrared and visible data $\{x_{ir}, x_{rgb}\}$. The model is supervised by their shared label $\{y\}$ during training. The setup in this stage is designed with reference to the training process of previous multispectral object detection algorithms. This stage aims to enable the model to effectively learn the fusion of features from infrared and visible light for object detection.

Training losses. For the Isolation stage and Fusion stage, where the only difference lies in the input of two data streams, both tasks involve object detection. Therefore, we use the same object detection loss function in two stages. $\mathcal{L}_{total}$ is:

$$\mathcal{L}_{total} = \mathcal{L}_{box} + \mathcal{L}_{cls} \tag{4}$$

where L_{box} is the box regression loss, L_{cls} is the classification loss, the overall loss function is consistent with previous work [18].

4. Experiment

4.1. Experimental Setup

Datasets. We evaluate the proposed method on two popular datasets, i.e. aligned FLIR and LLVIP. Furthermore, in the Isolation stage, we separate and combine the IR and RGB data from aligned FLIR and LLVIP for training to substantiate the benefits of the two-stage approach.

FLIR: The FLIR dataset [6] provides RGB and IR image pairs for day and night scenes. However, the original dataset contains many unaligned images and only has annotations for IR images, making training difficult. We use an aligned version from [28] with 5,142 image pairs of which 4,129 are for training, and 1,013 are for testing. It covers 3 classes: "person", "car", and "bicycle". The subsequent mentions of FLIR refer to the aligned version.

LLVIP: The LLVIP dataset [12] contains 15,488 aligned visible and infrared image pairs for pedestrian detection in dim surveillance environments. 12,025 pairs are used for training and 3,463 for testing. The images are strictly aligned with pedestrians as the only object category.

Methods	train	test	mAP50	mAP75	mAP
YOLOv5	RGB	RGB	59.9	21.7	28.1
YOLOv5	merge	RGB	65.4	25.5	31.5
YOLOv5	IR	IR	75.4	34.9	39.5
YOLOv5	merge	IR	77.8	36.8	41.4
RT-DETR[18]	RGB	RGB	69.7	26.5	33.5
RT-DETR[18]	merge	RGB	70.1	28.1	34.4
RT-DETR[18]	IR	IR	80.5	39.6	43.6
RT-DETR[18]	merge	IR	82.2	40.0	44.5

Table 1. Baseline with various compositions on FLIR dataset. The table displays our own training results.

Implementation Details. The proposed detection model is adapted from the RT-DETR model in the PaddleDetection library [19] and trained on an NVIDIA GeForce RTX 4090 GPU. The parameter-shared backbone is the HGNetv2X pre-trained on ImageNet [22] from PaddleClas2 [5]. The MSFI module consists of 1 Transformer layer and the fusion block in the CMSF module consists of 3 *RepBlocks*. In addition, the Transformer decoder consists of six layers. The training strategy, hyperparameters, and data augmentation for the synchronized RGB and IR images largely follow RT-DETR [18]. The model is trained for 60 epochs. For the two-stage training epochs, the Isolation stage training using the original data is conducted for 10 epochs, followed by 50 epochs of Fusion stage training. The batch size is set as 4. The image size is 640×640.

4.2. Baseline With Different Training Datasets

To determine whether training the model by directly merging images from two modalities would cause confusion, we conducted tests on two baseline detection models, we utilized the merged FLIR [6] dataset for training the baseline network. The results are shown in Table.1.

We utilize YOLOv5, the baseline from prior research, along with RT-DETR[18], the base model of our own model. The original baseline involved training the model separately on IR (or RGB) and then testing it on the corresponding modality. The experimental results reveal a significant superiority of RT-DETR[18] over YOLOv5 across diverse data compositions. Moreover, when considering the scenario of merged training data, both test on IR and RGB exhibit a notable performance improvement for YOLOv5 and RT-DETR[18]. This suggests that the straightforward combination of RGB and IR data contributes to enhancing model performance.

4.3. Comparison With State-of-the-Art Methods

Considering the diversity in the training data, we design two training strategies for our method. The first strategy follows the design of previous work, involving only the Fusion stage, denoted as Ours(fus). The second strategy involves two stages, where the IR and RGB data is merged for training in the Isolation stage, and then training in the Fusion stage, denoted as Ours(iso+fus).

FLIR. We compare our proposed method with previous work, including single-modality detection methods: RT-DETR*[18], training on the merged FLIR dataset, and multi-modality detection methods: CFT [20], CSAA [1], ICAFusion[23], and LRAF-Net [7]. The results of these existing methods and our proposed method on the FLIR dataset are shown in Table. 2. It can be observed that the mAP of the single-modality IR baseline of RT-DETR surpasses previous multi-modality fusion detection methods. This indicates that the improvement in the baseline model for multispectral object detection tasks brings about significant enhancements. Our model trained only in the Fusion stage, shows an improvement of 2.5% in mAP compared to the previous state-of-the-art method. It also outperforms the baseline of RT-DETR in the IR modality by 1.3%. This indicates that our model effectively integrates information from both IR and RGB.

When using the two-stage training strategy, there is an additional 0.5% improvement in mAP compared to training only in the Fusion stage. This suggests that changing to a two-stage training strategy, without introducing new training data, enables the model to achieve stronger performance.

LLVIP. In our LLVIP dataset experiments, the baseline data presented below still follows the settings and results of previous work, where the baseline is trained and tested on their respective modalities (IR or RGB). We additionally compared our method with MS-DETR [26], a Deformable-DETR-based multispectral pedestrian detection method. The specific results are shown in Table. 2. Compared with the single-modality RT-DETR, Our model trained by two-stage improves the AP by 2.3%. Compared with the state-of-the-art multi-modality object detection method, the AP is improved by 3.9%. Although slightly lower in terms of AP50 metric compared to the state-of-the-art method, our approach exhibits significant improvements in high IoU AP, indicating higher precision in human bounding box localization by our model. This indicates that the fusion of IR and RGB information in our model is also effective on the LLVIP dataset.

4.4. Ablation Experiments

Visualization: Fig. 3 shows the advantages of the GM-DETR model. Networks trained on a single modality alone struggle to accurately predict boxes on inputs from both

Methods	Data	FLIR			LLVIP		
		mAP50	mAP75	mAP	AP50	AP75	AP
YOLOv5	RGB	67.8	25.9	31.8	91.8	50.9	50.8
RT-DETR[18]	RGB	69.7	26.5	33.5	91.5	59.5	54.2
YOLOv5	IR	73.9	35.7	39.5	95.6	72.2	61.9
RT-DETR[18]	IR	80.5	39.6	43.6	97.3	78.4	67.9
CFT[20]	RGB+IR	78.7	35.5	40.2	97.5	72.9	63.6
CSAA[1]	RGB+IR	79.2	37.4	41.3	94.3	66.6	59.2
ICAFusion[23]	RGB+IR	79.2	36.9	41.4	-	-	-
LRAF-Net[7]	RGB+IR	80.5	-	42.8	**97.9**	-	66.3
MS-DETR[26]	RGB+IR	-	-	-	**97.9**	76.3	66.1
Ours(fus)	RGB+IR	83.6	42	45.3	97.5	80.6	69.6
Ours(iso+fus)	RGB+IR	**83.9**	**42.6**	**45.8**	97.4	**81.4**	**70.2**

Table 2. Comparison of performances on The FLIR dataset and The LLVIP dataset. The best accuracy is indicate in bold.

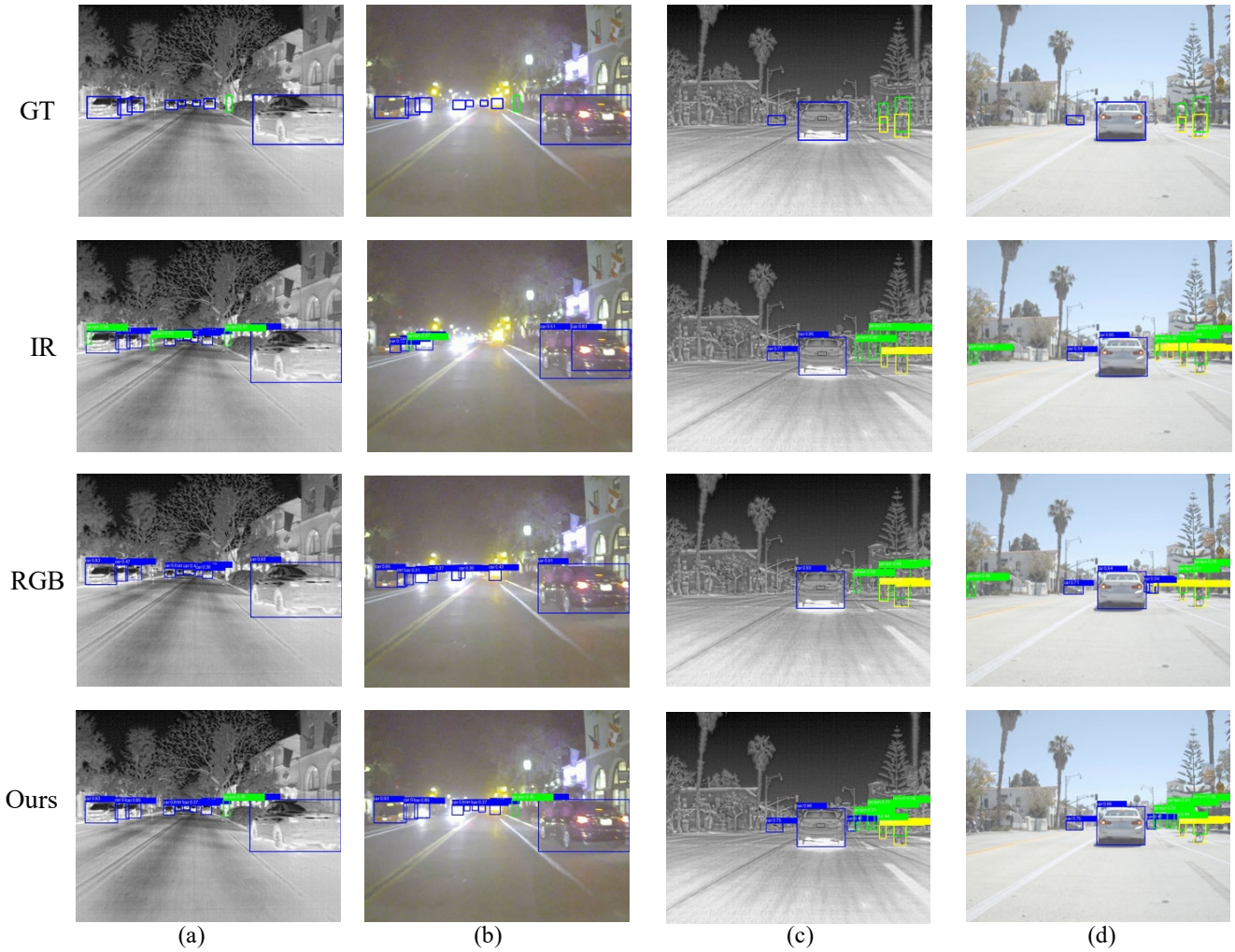

Figure 3. The visualization of the GTs and the detection results from the single-modality baseline and ours method. People, cars, and bicycles are respectively indicated by green, blue, and yellow bounding boxes.

modalities simultaneously. For instance, when the image in Fig. 3 (b) is fed into the baseline network trained on IR images, it fails to predict distant vehicles with headlights in the dark accurately. Simultaneously, for the corresponding infrared image in Fig. 3 (a) fed into the baseline network trained on IR images, there are excessive false detections of the human category. It can be observed that the results of our method are closest to the ground truths (GTs). Moreover, for Fig. 3 (c) and (d), it can be seen that the GTs in the dataset do not annotate distant objects. However, our method not only detects the objects included in the GTs but also detects distant vehicles and people not included in the GTs.

Input Ablation study: For multispectral object detection tasks, in real-world scenarios, modalities may be missing, meaning that data from one modality cannot be input. This leads to input data from only a single modality. We compare the performance of different models in this scenario. For our model, when modal drop occurs, the data input is transformed into two identical copies of a single-modality input. Table. 3 shows the results of the ablation studies.

Methods	-	IR	RGB
YOLOv5	-	73.9	67.8
RT-DETR[18]	-	80.5	69.7
Methods	RGB+IR	IR+IR	RGB+RGB
ICAFusion[23]	79.2	66(-7.9)	57.8(-10)
Ours(fus)	83.6	78.7(-1.8)	57.6(-12.1)
Ours(iso+fus)	83.9	80.6(+0.1)	64.3(-5.4)

Table 3. Comparison with different input on FLIR dataset

ICAFusion [23] similarly conducted experiments on this matter. The metric being mAP50.

The experimental results indicate that baseline networks trained on single-modality data achieve optimal results on their respective modalities. When dealing with modality drop in multi-spectral detection models, it can be seen that the ICAFusion gets significant performance degradation compared to its baseline network Yolov5. For our model, after employing a two-stage training strategy, when only IR input is available, the model's performance is comparable to the baseline RT-DETR. Meanwhile, with only RGB input, the model gets a decrease of only 5.4%, and fused training strategy is 12.1%. This not only demonstrates the effectiveness of our two-stage training strategy but also indicate that our model is adept at handling situations involving the drop of the modality drop, demonstrating its generalization properties.

Module Ablation study: The ablation experiments of GM-DETR modules are shown in Table. 4, evaluating the model's mAP50 performance on the FLIR dataset with RGB+IR, IR+IR, RGB+RGB date input. For the RGB+IR data input, Starting from a baseline of 82.4%, adding the MSFI module to the baseline resulted in performance declines of 0.6%. This suggests that directly performing weighted sum on features extracted solely by the backbone is intuitive and yields satisfactory performance. Adding the CMSF module to the baseline, the mAP50 on the IR+IR gets a improvement of 1.2%, but decreases 2% on the RGB+RGB data input. Moreover, since the design of MSFI and CMSF is integrated, the model utilizes global information extracted by MSFI to guide the CMSF module in fusing information from both modalities. Upon the concurrent integration of MSFI and CMSF, the model exhibited a 1.2% performance enhancement compared to the baseline. However, this concurrent integration also led to overfitting of the RGB+IR input data type. With the implementation of the two-stage training strategy, the model demonstrated a significant improvement across IR+IR and RGB+RGB data inputs, achieving optimal performance across all three input types.

MSFI	CMSF	Two-stage	RGB+IR	IR+IR	RGB+RGB
			82.4	77.9	58.6
✓			81.8	78.3	56.0
	✓		82.5	79.1	56.6
✓	✓		83.6	78.7	57.6
✓	✓	✓	**83.9**	**80.6**	**64.3**

Table 4. Ablation studies of GM-DETR on FLIR dataset. MSFI, CMSF, Two-stage refer to Modality-Specific Feature Interaction module, Cross-Modality-Scale feature Fusion module, Two-stage trainin strategy. As for CMSF ablation studies, we replace CMSF with an additive operation on each feature layer. The metrics is mAP50.

methods	Params(M)	FLOPs(G)	FPS
RT-DETR[18]	66	117	336
GM-DETR	70	176	218

Table 5. Comparison of Parameters, FLOPS, and FPS

4.5. Runtime Analysis

We present the total number of learnable parameters, floating-point operations per second (FLOPs), and frames per second (FPS) on a RTX 4090 (TensorRT fp16) in Table. 5.

Due to the efficient fusion encoder, GM-DETR's parameter (70M) has not significantly increased, while still maintaining the capability for real-time operations (FPS = 218), the additional 59G FLOPs represent the computa-

tional overhead incurred in computing the interaction between dual inputs.

5. Conclusion and Future Work

In this work, we propose GM-DETR, a novel multispectral object detection algorithm. The utilization of the MSFI and CMSF modules facilitates the effective extraction and fusion of features from each modality, enabling the implementation of a two-stage training strategy. Alongside enhancing detection performance, proposing solutions to address the issue of modality drop.Experimental results on two datasets show that our proposed method outperforms existing multispectral object detection models.

In the future, the generalization capability of GM-DETR will see significant improvements with the introduction of additional isolated datasets in the Isolated stage. Furthermore, integrating a semi-supervised learning paradigm into GM-DETR is highly suitable for such models with dual inputs. This will pave the way for new approaches to cross-domain research, enabling adaptation from RGB to IR.

Acknowledgements. This work was supported in part by National Natural Science Foundation of China (No.62271119, U23A20286, and 62071086), Natural Science Foundation of Sichuan Province (2023NSFSC1972), and Independent Research Project of Civil Aviation Flight Technology and Flight Safety Key Laboratory (FZ2022ZZ06).

References

[1] Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. Multimodal object detection by channel switching and spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–411, 2023. 1, 2, 5, 6

[2] 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*, pages 213–229. Springer, 2020. 2

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[4] Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling. In *Computer Vision – ECCV 2022*, pages 139–158, Cham, 2022. Springer Nature Switzerland. 2

[5] Cheng Cui, Ruoyu Guo, Yuning Du, Dongliang He, Fu Li, Zewu Wu, Qiwen Liu, Shilei Wen, Jizhou Huang, Xiaoguang Hu, et al. Beyond self-supervision: A simple yet effective network distillation alternative to improve backbones. *arXiv preprint arXiv:2103.05959*, 2021. 5

[6] FLIR. Flir santa barbara regional thermal dataset. https://flir.app.box.com/s/suwst0b3k9rko35homhr3rny 3102d, 2018. Date of visit: Aug 30 2023. 2, 4, 5

[7] Haolong Fu, Shixun Wang, Puhong Duan, Changyan Xiao, Renwei Dian, Shutao Li, and Zhiyong Li. Lraf-net: Long-range attention fusion network for visible–infrared object detection. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2023. 1, 2, 5, 6

[8] Kaiming He, Ross Girshick, and Piotr Dollar. Rethinking imagenet pre-training. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4917–4926, 2019. 3

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 3

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022. 3

[11] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[12] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504, 2021. 2, 5

[13] Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:72–86, 2021. 4

[14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. 2

[16] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018. 4

[17] Andong Lu, Cun Qian, Chenglong Li, Jin Tang, and Liang Wang. Duality-gated mutual condition network for rgbt tracking. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2022. 1

[18] Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui, Yuning Du, Qingqing Dang, and Yi Liu. Detrs beat yolos on real-time object detection, 2023. 2, 3, 4, 5, 6, 7

[19] Yanjun Ma, Dianhai Yu, Tian Wu, and Haifeng Wang. Paddlepaddle: An open-source deep learning platform from in-

dustrial practice. *Frontiers of Data and Domputing*, 1(1): 105–115, 2019. 5

[20] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021. 1, 2, 5, 6

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 3, 5

[23] 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*, page 109913, 2023. 1, 2, 3, 5, 6, 7

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[25] Wen Wang, Jing Zhang, Yang Cao, Yongliang Shen, and Dacheng Tao. Towards data-efficient detection transformers. In *European conference on computer vision*, pages 88–105. Springer, 2022. 4

[26] Yinghui Xing, Song Wang, Guoqiang Liang, Qingyi Li, Xiuwei Zhang, Shizhou Zhang, and Yanning Zhang. Multispectral pedestrian detection via reference box constrained cross attention and modality balanced optimization. *arXiv preprint arXiv:2302.00290*, 2023. 1, 2, 5, 6

[27] H. Yu, X. Cheng, and W. Peng. Toplight: Lightweight neural networks with task-oriented pretraining for visible-infrared recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3541–3550, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 3

[28] 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 (ICIP)*, pages 276–280. IEEE, 2020. 4

[29] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 2, 3

[30] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5906–5916, 2023. 2

[31] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33:3833–3845, 2020. 3