Rethinking Self-Attention for Multispectral Object Detection

Sijie Hu, Fabien Bonardi, Samia Bouchafa, Helmut Prendinger, Désiré Sidibé

Rethinking Self-Attention for Multispectral Object Detection

Sijie Hu, Fabien Bonardi, Samia Bouchafa, Helmut Prendinger, Désiré Sidibé

***Abstract*—Data from different modalities, such as infrared and visible images, can offer complementary information, and integrating such information can significantly enhance the capabilities of a system to perceive and recognize its surroundings. Thus, multi-modal object detection has widespread applications, particularly in challenging weather conditions like low-light scenarios. The core of multi-modal fusion lies in developing a reasonable fusion strategy, which can fully exploit the complementary features of different modalities while preventing a significant increase in model complexity. To this end, this paper proposes a novel lightweight cross-fusion module named Channel-Patch Cross Fusion (CPCF), which leverages Channel-wise Cross-Attention (CCA), Patch-wise Cross-Attention (PCA) and Adaptive Gating (AG) to encourage mutual rectification among different modalities. This process simultaneously explores commonalities across modalities while maintaining the uniqueness of each modality. Furthermore, we design a versatile intermediate fusion framework that can leverage CPCF to enhance the performance of multi-modal object detection. The proposed method is extensively evaluated on multiple public multi-modal datasets, namely FLIR, LLVIP, and DroneVehicle. The experiments indicate that our method yields consistent performance gains across various benchmarks and can be extended to different types of detectors, further demonstrating its robustness and generalizability. Our codes are available at https://github.com/Superjie13/CPCF_Multispectral.**

***Index Terms*—Multispectral, Attention, Intermediate fusion, Object detection, Deep learning.**

I. Introduction

OBJECT detection involves extracting items of interest from input data and locating their positions, which has a wide range of applications in different areas, such as autonomous driving [1], security surveillance [2], and disaster relief [3]. In recent years, numerous advanced object detection methods have emerged [4]–[6], demonstrating outstanding performance on various tasks with color images, i.e., RGB, as inputs [7]. However, real-world scenarios are often dynamically changing, making it highly challenging to collect enough data to detect all objects in a scene using only the color modality. For instance, the image quality captured by RGB cameras at night typically deteriorates significantly, substantially reducing the accuracy and robustness of detection results.

On the other hand, due to the stability in imaging under different lighting conditions, thermal infrared cameras are frequently employed in low-light situations to enhance the system's ability to capture scene information. For instance, thermal images can be used to provide full-time geometric characteristics of objects, e.g., shape and contour, while color images are capable of providing rich texture information when light is sufficient. Therefore, an effective fusion strategy is needed to fully exploit the complementary features among different modalities. In this context, according to the location of fusion occurrence, multi-modal fusion can be categorized into early fusion, late fusion, and intermediate fusion [8]. Specifically, early fusion directly concatenates multi-modal data into a unified multi-channel input, which is then fed into a general object detection framework. Conversely, late fusion independently processes data from different modalities and integrates the outputs at the point of decision-making by an additional fusion operation. Between early and late fusion, intermediate fusion incrementally merges features of different modalities through a flexible structure design, allowing the features to maintain their independence while interacting. Although intermediate fusion presents advantages, devising efficient fusion modules that can accurately combine diverse features while preserving the integrity of the original data remains a substantial challenge. In response to this, various studies [9]–[12] have sought to uncover latent correlations among different modalities using attention mechanisms. Particularly, some recent works [9], [10], [13], [14] have leveraged self-attention [15] to fuse multi-modal features achieving encouraging results. However, the extensive computation required by these fusion modules significantly constrains their potential in multi-modal fusion.

In this paper, we reconsider self-attention within the context of multi-modal visual data fusion, with a particular focus on uncovering the complementary traits among different modalities by simplifying the computation of self-attention across their contexts. To do so, we deduce attention maps across two distinct dimensions: channel and spatial [16]. Furthermore, we believe mutual calibration of modal features serves as an effective way to explore complementarity across multiple modalities. To this end, we propose a lightweight cross-attention fusion module, termed channel-patch cross fusion (CPCF), which is composed of a channel-wise cross-attention (CCA), a patch-wise cross-attention (PCA) and an adaptive gating (GA) unit. Specifically, we employ parametric-free operations such as average pooling and max pooling to model

Manuscript received XX XX, XX; accepted XX XX, XX. Date of publication XX XX, XX; date of current version XX XX, XX. This article was recommended for publication by Associate Editor X. XX and Editor X. XX upon evaluation of the Reviewers' comments. (Corresponding author: Sijie Hu.

Sijie Hu, Fabien Bonardi, Samia Bouchafa and Désiré Sidibé are with Université Paris-Saclay, Univ Evry, IBISC Laboratory, 91020, Evry, France. (e-mail: sijie.hu@universite-paris-saclay.fr)

Helmut Prendinger is with the Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo, 101-8430, Japan.

Digital Object Identifier

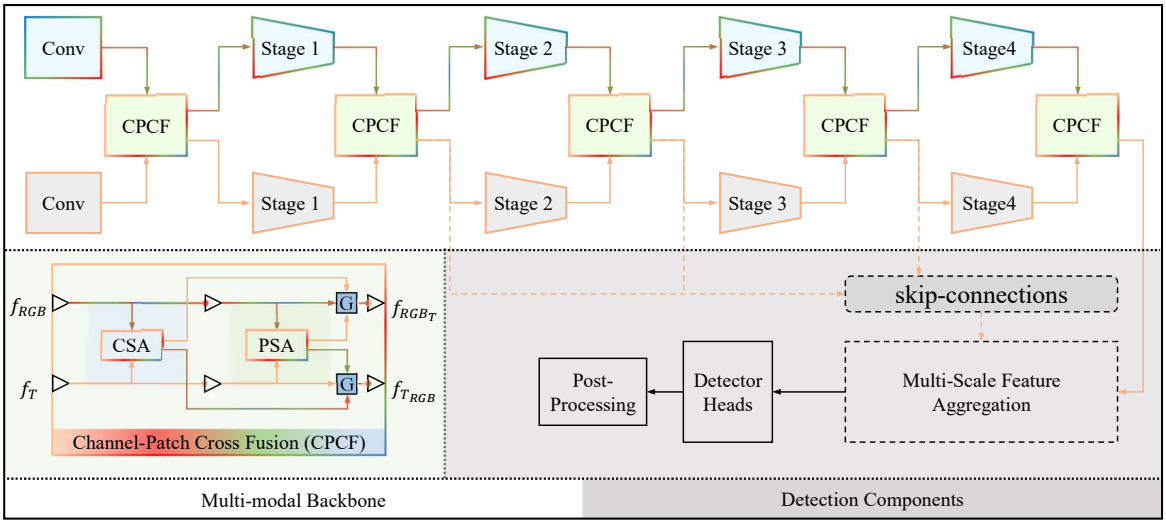

Fig. 1. Overview of CPCF-based Object Detection Framework. The upper part is the general multi-modal backbone, the lower part is the detection related components. CPCF denotes the proposed channel-patch cross fusion module, CSA and PSA denote the proposed channel-wise and patch-wise cross-attention modules, and G denotes the proposed adaptive gating unit.

the characteristics of each modality and incorporate cross-attention to reconstruct complementary awareness across different modalities in terms of channels and spatial dimensions, thus ensuring the complementarity of different modalities while maintaining their independence. Note that due to our efficient design, the extra complexity and parameters introduced by our CPCF are negligible. In addition, we argue that the representation of information in the channel and spatial dimensions changes as forward propagation is performed. Thus, we design an adaptive gating unit to enable the fusion module to adapt to these changes effectively. To showcase the efficacy of CPCF, we create a general intermediate fusion framework, as depicted in Figure 1, which can be extended to various detectors. Then, we conduct extensive experiments on the standard multispectral dataset named FLIR [17] and LLVIP [18] and a more challenging oriented object detection dataset called DroneVehicle [9]. Our results demonstrate that our proposed approach can remarkably improve the performance of object detection without significantly increasing the complexity of the model.

The contributions of this paper are summarized as follows:

- We propose a lightweight channel-patch cross fusion (CPCF) module to construct cross-modal features in both channel and spatial dimensions, during which the CPCF module adaptively leverages the properties specific to one modality to calibrate the features of another, thus effectively modeling the complementary properties between modalities and optimizing the representability of features in the data stream.
- We design an intermediate fusion framework based on CPCF, which can be flexibly integrated into various object detection methods to efficiently leverage multi-modal cues to boost the performance of models.
- We conduct extensive experiments and analyses on different types of multi-modal datasets and obtain optimal results. Simultaneously, we validate the generalization ability of our method on different detectors, which further shows its robustness and versatility.

The rest of this paper is organized as follows: Section II reviews the existing works related to our method. The overall framework is presented in Section III with the details of each component. The experimental setup and the results are presented and discussed in Section IV. Finally, Section V ends this paper with a conclusion and discussion.

II. Related Works

A. Unimodal Object Detection

Unimodal object detection typically employs RGB images as input, which can be categorized into two- and single-stage approaches. Two-stage approaches divide object detection into two distinct phases, i.e., the regional proposal phase and the target classification and bounding box regression phases. As a trailblazing effort, RCNN [19] leverages the selective search algorithm [20] to generate numerous potential regions, then employs SVM and a regressor for classification and bounding box prediction tasks. Next, FastRCNN [21] and FasterRCNN [5] upgrade this idea within a deep learning framework, further improving training efficiency and model performance. On the other hand, single-stage object detection frameworks, represented by YOLO [22], directly predict the category and location of objects in a single forward propagation, eliminating the need for a region proposal stage, thereby greatly enhancing the detection speed. Especially, some variants [23]–[25] of YOLO are gradually catching up with the two-stage detector in terms of detection accuracy while maintaining high operating speed. Recently, YOLOX [4] has transformed the YOLO detector into an anchor-free style, further enhancing processing speed.

Moreover, in some special scenarios, such as remote sensing images, traditional axis-aligned bounding boxes cannot accurately describe the state of objects. For this reason, oriented object detectors [26], [27] are designed to align the bounding

boxes with the orientation of the targets. These detectors rely on existing object detection frameworks and predict the direction of bounding boxes through additional modules. For instance, S^2A-Net [28] introduces a feature alignment module and an oriented detection module for mitigating the misalignment between oriented anchors and axis-aligned convolutional features. Then, the PSC [29] utilizes an additional phase shift encoder to achieve an accurate prediction of the orientation.

In this work, we implement our method within different detectors and conduct extensive experiments on different types of datasets to tackle multi-modal object detection tasks under various scenarios.

B. Multispectral Object Detection

Multispectral object detection is a vibrant research field in the computer vision community. It typically blends multi-modal data through early fusion, late fusion, or intermediate fusion strategies. In early fusion, RGB and IR images are concatenated at pixel level to form a 4-channel input, and then features are extracted with a regular object detection framework [30]. However, early fusion forgets modality-specific properties during feature forward propagation, which can lead to suboptimal results [12], [31]. Conversely, late fusion [32] process each modality independently through separate models and the results are merged at the decision level [33]. Yet, assembling multiple detectors results in more false positive cases and slower detection speeds [34]. On the other hand, intermediate fusion lies between the two, in which features from different modalities interact with each other while still preserving their individuality [9], [11], [34]–[36]. For instance, GFD-SSD [35] employs two encoders to handle RGB and thermal images and utilizes a gating unit to merge features from the intermediate layers. UA-CMDet [9] leverages uncertainty-awareness to reduce the detection bias caused by high-uncertainty objects. Moreover, MBNet [37] focuses on the grasp of differential modality and illumination to alleviate both the modality and feature imbalance in the dual-modality network. Inspired by the attention mechanism, GAFF [11] and ECISNet [36] leverage spatial attention to learn the adaptive weighting and fusion of different modalities. Besides, ICAFusion [13] and CAT [14] utilize transformer blocks [15] to learn global feature correlations across modalities, which, although yielding promising results, also introduce a significant computational burden. Recently, CMAFF [38] and CSAA [34] reengineer the fusion module, reducing the computational complexity of the fusion process while emphasizing the complementary nature of modalities.

In this work, inspired by self-attention [15], we calculate cross-attention separately from both channel and spatial dimensions and leverage the learnable gating units to adaptively integrate different attention at different levels rather than treating them equally.

III. Methods

In this section, we propose a lightweight cross-fusion module named CPCF, which can efficiently build long-range dependencies from one modality to another in both channel and spatial axes. Building upon CPCF, we further design a general intermediate fusion object detection framework to effectively exploit multi-modal information. In the following, we will detail the proposed intermediate fusion framework and the associated modules.

A. Framework Overview

As shown in Figure 1, the overall multi-modal object detection framework is composed of two parts. The first part is a general multi-modal backbone, an intermediate fusion-based feature extractor for refining and fusing multi-modal information. The second part is the detection related components, which provide modules, such as skip connections and detection heads, for different types of detectors. Typically, the multi-modal backbone originates from prevalent single-modality backbones, such as ResNet [39] and CSPDarknet [40], which are composed of several convolution stages, enabling a more efficient and comprehensive encoding of information from inputs. As illustrated in the upper half of Figure 1, we employ a symmetrical structure to separately process information from different modalities. Meanwhile, the proposed CPCF module is deployed subsequent to each convolution stage to calculate the awareness across different modalities and recalibrating the multi-modal features. Afterward, the calibrated features are propagated to the components specific to the object detection tasks, as shown in the lower right of Figure 1. Taking YOLOX [4] as an example, the fused features from different CPCF modules are aggregated via a feature pyramid module to multiple object detection heads for multi-scale prediction. In addition, for the two-stage detector, like RCNN [19], a region proposal module is operated to receive the outputs from the last CPCF module.

B. Multi-modal Cross-Attention

While different visual modalities carry complementary information valuable for perception tasks, they also contain a considerable amount of redundant data and noise, factors that can potentially influence the efficiency of data analysis and interpretation. In this context, we propose a multi-modal cross-attention mechanism that calibrates one modality with the features of another. This structure amplifies the complementary characteristics between modalities while diminishing redundant information, thereby fostering a more effective and integrated multi-modal representation. We argue that the feature representation of a modality can be reflected in both channel and spatial dimensions. Thus, we create channel-wise cross-attention (CCA) and patch-wise cross-attention modules (PCA) to establish both channel and spatial relationships among different modalities, enabling cross-modality feature recalibration and ensuring a more coherent and effective multi-modal data integration.

1) Channel-wise Cross-Attention: In a feature map, a channel is usually treated as a feature detector [16], thus channel-wise cross-attention (CCA) is designed to highlight beneficial channels across different modalities and suppress noise-included ones. To this purpose, CCA considers feature channels of two modalities parallelly and associates different

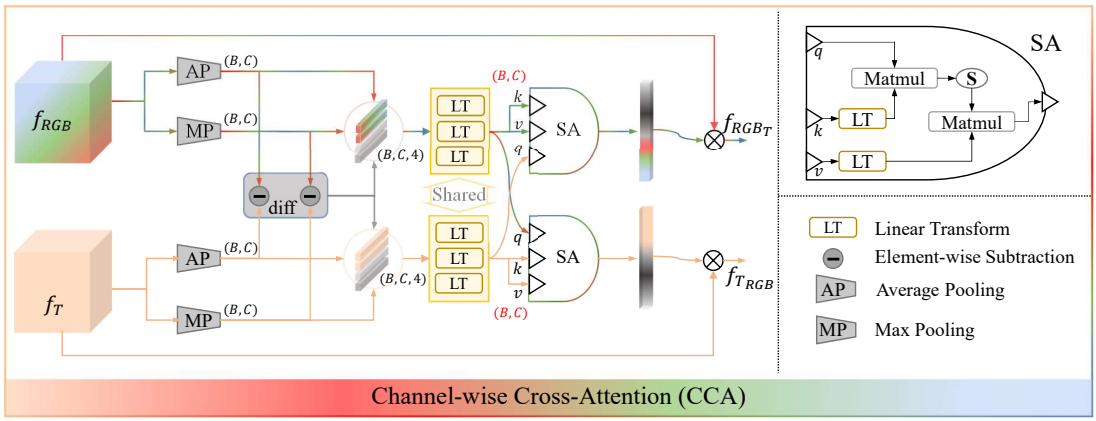

Fig. 2. Architecture of Channel-wise Cross-Attention. SA denotes the variant of the self-attention block used in our CCA module, diff means subtraction operation to compute the differential signals of two modalities, B and C represent the input batch and channel.

attention weights to different channels. The overall architecture of CCA is shwon in Figure 2.

Inspired by differential amplifier circuits where differential-mode signals are amplified while suppressing common-mode signals [37], [38], we seek to utilize differential modality features to highlight the characteristics of each modality and conceal redundancies contained in multi-modal features. Specifically, given the intermediate feature maps $f_{RGB} \in \mathbb{R}^{C\times H\times W}$ and $f_T \in \mathbb{R}^{C\times H\times W}$ of two modalities, average-pooling (AP) and max-pooling (MP) are applied to compress spatial information, followed by a series of subtraction operation to obtain the cross-modal differential signals. These are then concatenated into compact expressions $f_{RGB}^C \in \mathbb{R}^{C\times 4}$ and $f_T^C \in \mathbb{R}^{C\times 4}$, as expressed as follows:

$$\begin{aligned} f_{diff}^{AP} &= |\mathbf{AP}(f_{RGB}) - \mathbf{AP}(f_T)|, \\ f_{diff}^{MP} &= |\mathbf{MP}(f_{RGB}) - \mathbf{MP}(f_T)|, \\ f_{RGB}^C &= Concat\left([\mathbf{AP}(f_{RGB}), \mathbf{MP}(f_{RGB}), f_{diff}^{AP}, f_{diff}^{MP}]\right), \\ f_T^C &= Concat\left([\mathbf{AP}(f_T), \mathbf{MP}(f_T), f_{diff}^{AP}, f_{diff}^{MP}]\right). \end{aligned} \tag{1}$$

Self-attention [15] encodes the inputs into a set of vectors, i.e., Query (Q), Key (K), and Value (V), and computes the attention map via a matrix multiplication QK^T. After that, the output of self-attention is obtained by another matrix multiplication between the attention map and V, which can be described as follows:

$$f^{SA} = \mathbf{SA}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{D_k}}\right)V. \tag{2}$$

where $\frac{1}{\sqrt{D_K}}$ is a scaling factor. In this manner, the module can construct global attention across tokens.

On the other hand, the compressed channel features are expressed in vector form, making them inherently compatible with self-attention. Thereby, we leverage self-attention to construct long-range dependencies of each channel. To do so, we regard each channel as a token and project them into vectors designated as $Q \in \mathbb{R}^{C\times 1}$, $K \in \mathbb{R}^{C\times 1}$, and $V \in \mathbb{R}^{C\times 1}$ through a straightforward linear transformation:

$$\begin{aligned} Q &= f_X^C W_Q, \\ K &= f_X^C W_K, \\ V &= f_X^C W_V, \end{aligned} \tag{3}$$

where $W_Q \in \mathbb{R}^{4\times 1}$, $W_K \in \mathbb{R}^{4\times 1}$, and $W_V \in \mathbb{R}^{4\times 1}$ are weight matrices of linear transformation, and the subscript X is either RGB or Thermal. When computing self-attention, we swap the Q vector of the two modalities rather than directly using them for the attention calculation, thus forming cross-attention. As illustrated in the SA module in Figure 2, considering that the computational cost of self-attention is quadratic in the vector's length, two linear transformations are employed to compress vectors K and V to reduce the computational burden. The cross-attention scores $S_{RGB}^{CA} \in \mathbb{R}^{C\times 1}$ and $S_T^{CA} \in \mathbb{R}^{C\times 1}$ can be formulated as follows:

$$\begin{aligned} S_{RGB}^{CA} &= \mathbf{SA}(Q_T, K_{RGB}, V_{RGB}), \\ S_T^{CA} &= \mathbf{SA}(Q_{RGB}, K_T, V_T). \end{aligned} \tag{4}$$

Finally, the attention scores from different modalities are normalized to the range $[0, 1]$ through a sigmoid function, and the channel-wise recalibrated features f_{RGB}^{RC} and f_T^{RC} can be described as:

$$\begin{aligned} f_{RGB}^{RC} &= \sigma\left(S_{RGB}^{CA}\right) \otimes f_{RGB}, \\ f_T^{RC} &= \sigma\left(S_T^{CA}\right) \otimes f_T, \end{aligned} \tag{5}$$

where $\sigma(\cdot)$ indicates the sigmoid function, and $\otimes$ indicates element-wise multiplication.

2) Patch-wise Cross-Attention: Contrary to the aforementioned CCA, which attempts to establish long-range attention across channels, patch-wise cross-attention (PCA) aims to model inter-patch connections of different modalities and leverage this information to calibrate the multi-modal features across spatial dimension. To achieve this goal, given the intermediate feature maps $f_{RGB} \in \mathbb{R}^{C\times H\times W}$ and $f_T \in \mathbb{R}^{C\times H\times W}$ and patch size $h \times w$, we first apply patch average pooling (PAP) and patch max pooling operations (PMP) to condense local information and reduce the spatial resolution of the features. Then, following the same approach as described

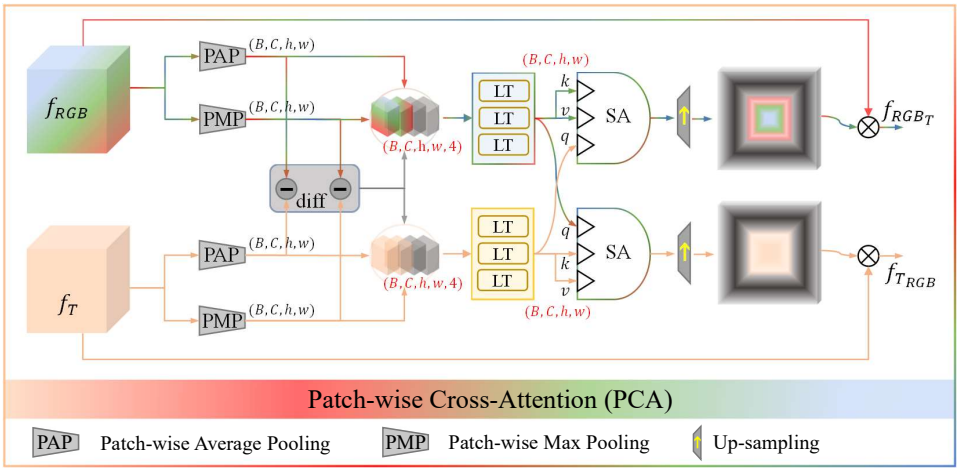

Fig. 3. Architecture of Patch-wise Cross-Attention. SA means the self-attention, and diff denotes the subtraction operation, as in Figure 2.

in Section III-B1, we obtain compact expressions along the spatial dimension. The procedure can be precisely described as:

$$\begin{aligned} f_{diff}^{PAP} &= |\mathbf{PAP}(f_{RGB}) - \mathbf{PAP}(f_T)|, \\ f_{diff}^{PMP} &= |\mathbf{PMP}(f_{RGB}) - \mathbf{PMP}(f_T)|, \\ f_{RGB}^{C} &= Concat\left([\mathbf{PAP}(f_{RGB}), \mathbf{PMP}(f_{RGB}), f_{diff}^{PAP}, f_{diff}^{PMP}]\right), \\ f_{T}^{C} &= Concat\left([\mathbf{PAP}(f_{T}), \mathbf{PMP}(f_{T}), f_{diff}^{PAP}, f_{diff}^{PMP}]\right), \end{aligned} \tag{6}$$

where $f_{RGB}^{C} \in \mathbb{R}^{C\times N\times 4}$ and $f_{T}^{C} \in \mathbb{R}^{C\times N\times 4}$ denote the compact RGB and Thermal features, $N = hw$ denotes patch numbers.

Next, we utilize two separate linear transformation blocks to encode f_{RGB}^{C} and f_{T}^{C} into their corresponding $Q \in \mathbb{R}^{N\times C}$, $K \in \mathbb{R}^{N\times C}$, and $V \in \mathbb{R}^{N\times C}$ vectors, and the cross-attention scores $S_{RGB}^{CA} \in \mathbb{R}^{N\times 1}$ and $S_{T}^{CA} \in \mathbb{R}^{N\times 1}$ can be computed through Equation 4. Finally, the patch-wise recalibrated features f_{RGB}^{RP} and f_{T}^{RP} can be formulated as:

$$\begin{aligned} f_{RGB}^{RP} &= \sigma\left(UPS\left(S_{RGB}^{CA}\right)\right) \otimes f_{RGB}, \\ f_{T}^{RP} &= \sigma\left(UPS\left(S_{T}^{CA}\right)\right) \otimes f_{T}, \end{aligned} \tag{7}$$

where $UPS(\cdot)$ denotes up-sampling the size of attention scores to the input resolution. The details of PCA are depicted in Figure 3.

C. Channel-Patch Cross Fusion

The architecture of channel-patch cross fusion (CPCF) is shown in the lower left of Figure 1. In CPCF, we integrate the proposed CCA and PCA into the fusion process, thus allowing for the effective utilization of multi-modal cues and enhancing the representative capability of the fused features. However, treating channel and spatial attention equally during this process may lead to suboptimal results. The feature extraction process is characterized by the continuous compression of spatial resolution and expansion of channel dimensions. Throughout this process, the quantity of information across different dimensions does not remain constant. In response to this situation, we design an adaptive gating (AG) unit that dynamically allocates weights to different attention mechanisms, which allows a more responsive and adaptive fusion. More specifically, two learnable scaling factors, denoted as α_1 and α_2, are defined to dynamically adjust the weights of CCA and PCA during training. Then, the corresponding weights s_1 and s_2 can be computed as:

$$\begin{aligned} s_1 &= \frac{\sigma(\alpha_1/T)}{\sigma(\alpha_1/T) + \sigma(\alpha_2/T)}, \\ s_2 &= 1 - s_1, \end{aligned} \tag{8}$$

where the $\sigma(\cdot)$ denotes the sigmoid function, T is a temperature coefficient used to smooth the scaling weights.

In summary, given the input feature maps f_{RGB} and f_T and recalibrated feature maps f_{RGB}^{CR}, f_{T}^{CR}, f_{RGB}^{PR}, and f_{T}^{PR}, the fused features are obtained as:

$$\begin{aligned} f_{RGB}^{Fuse} &= f_{RGB} + s_1 \cdot f_{T}^{CR} + s_2 \cdot f_{T}^{PR}, \\ f_{T}^{Fuse} &= f_{T} + s_1 \cdot f_{RGB}^{CR} + s_2 \cdot f_{RGB}^{PR}. \end{aligned} \tag{9}$$

TABLE I
DATASET SETUP. HBB/OBB REFER TO HORIZONTAL/ORIENTED BOUNDING BOX, RESPECTIVELY.

Setup	FLIR	LLVIP	DroneVehicle
Class Num.	3	1	5
Modality	RGB&Thermal	RGB&IR	RGB&IR
Box Type	HBB	HBB	OBB
Img Size (original)	640×512	1280×1024	640×512
Img Size (train)	640×512	640×512	640×512
Epoches	13	13	36
Leraning Rate	2e-3	2e-3	2.5e-3
Batch Size	8	8	2
Train/Val/Test (pairs)	4139/1013/-	12025/-/3463	17990/1469/8980

IV. EXPERIMENTS

In this section, we initially perform experiments on two general-purpose object detection benchmarks, specifically FLIR [17] and LLVIP [18], to assess the efficacy of our proposed methods. Subsequently, we extend our testing to a more

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART MULTISPECTRAL METHODS AND OUR BASELINES ON FLIR VALIDATION SET BY MAP IN PERCENTAGE.

Method	Backbone	Fusion	Modality	$mAP_{50}\uparrow$	$mAP_{75}\uparrow$	mAP ↑	Param. (M)↓
Fcos [6]	ResNet50	-	RGB	59.3	20.2	26.7	32.12
Fcos [6]	ResNet50	-	Thermal	69.4	28.3	33.7	32.12
YOLOv5 [41]	Darknet53	-	RGB	65.2	21.9	29.3	7.03
YOLOv5 [41]	Darknet53	-	Thermal	78.9	32.9	39.2	7.03
YOLOX [4]	Darknet53	-	RGB	62.8	22.2	28.9	8.94
YOLOX [4]	Darknet53	-	Thermal	76.4	36.3	40.2	8.94
Multi-modal methods							
GAFF [11]	ResNet18	GAFF	RGB-T	72.9	32.9	37.5	23.75
CFT [42]	CFB	CFT	RGB-T	78.7	35.5	40.2	206.03
YOLOFusion [38]	Darknet53	CMAFF	RGB-T	76.6	-	39.8	12.52
UA-CMDet [9]	Darknet53	UA-CM	RGB-T	78.6	-	-	-
CSAA [34]	ResNet50	CSAA	RGB-T	79.2	37.4	41.3	-
ICAFusion [13]	Darknet53	DMFF	RGB-T	79.2	36.9	41.4	120.21
Our baselines							
FcosCAT	ResNet50	Concatenate	RGB-T	68.0	25.5	32.1	32.13
FcosSUM	ResNet50	MLSum	RGB-T	70.4	28.9	34.5	55.63
YOLOv5CAT	Darknet53	Concatenate	RGB-T	77.0	31.5	38.1	7.03
YOLOv5SUM	Darknet53	MLSum	RGB-T	79.2	34.6	40.2	11.2
YOLOXCAT	Darknet53	Concatenate	RGB-T	77.4	36.9	41.0	8.94
YOLOXSUM	Darknet53	MLSum	RGB-T	76.7	37.7	41.2	13.15
Our implementation with CPCF							
FcosCPCF	ResNet50	CPCF	RGB-T	73.4	32.0	37.0	61.28
YOLOV5CPCF	Darknet53	CPCF	RGB-T	81.6	37.0	41.8	12.67
YOLOXCPCF	Darknet53	CPCF	RGB-T	**82.1**	**41.2**	**44.6**	14.61

challenging DroneVehicle [9] dataset, which targets oriented object detection. Finally, we illustrate a series of studies to ablate different components and analyze the effectiveness of our designs.

A. Datasets

FLIR: The FLIR dataset is a benchmark extensively used for evaluating multispectral object detection, comprising a substantial number of paired RGB and thermal images. In our experiments, we utilize the aligned-FLIR dataset [17], wherein RGB-Thermal image pairs are correctly aligned. This dataset features 5142 image pairs, spanning three object categories: 'person', 'car', and 'bicycle', gathered from daytime to nighttime. Among these, 4139 pairs are for training, while the remaining 1013 pairs are allocated for testing.

LLVIP: The LLVIP [18] is a recently introduced, large-scale dataset explicitly designed for pedestrian detection in visible-infrared contexts. It contains 15488 image pairs, with 12,025 pairs for training and 3,463 pairs for testing. A notable characteristic of this dataset is that a majority of the images are captured under extremely low light conditions. Furthermore, all images within the dataset are stringently aligned in terms of time and space.

DroneVehicle: The DroneVehicle dataset [9] is a newly released multi-modal benchmark specifically designed for oriented vehicle detection from a drone's perspective. It encompasses five distinct vehicle categories, namely 'car', 'truck', 'bus', 'van', and 'freight car'. This dataset comprises 28,439 RGB-Infrared image pairs that capture a variety of settings, including urban roads, residential areas, and parking lots, from day to night with a resolution of 640×512. The dataset is composed of 17,990 image pairs for training, 1,469 for validation, and 8980 pairs reserved for testing.

B. Implementation Details

Utilizing the proposed CPCF, we design an intermediate fusion architecture that can be seamlessly integrated into a range of object detection frameworks. For the practical implementation, we build our model based on a popular object detection codebase MMDetection [43], and train our models on a single NVIDIA RTX3090 GPU. In all experiments, we initialize the backbone networks using the weights pre-trained on COCO [7] for general-purpose object detection. For oriented object detection tasks, the backbone networks are initialized with weights pre-trained on the ImageNet [44]. To train the models, we employ the SGD optimizer with an initial learning rate of 2e-3 and a momentum of 0.9. For data augmentation, we apply random flipping and scale the images to a resolution of 640×512. In the case of the FLIR dataset, we additionally leverage the Mosaic data augmentation technique [4] to further enrich the data for methods within the YOLO family. Subsequently, all models are trained in 13 epochs with a batch size of 8. For the DroneVehicle dataset, we set the batch size to 2 and train the model for 36 epochs. The setups of different datasets are shown in Table I. For all experiments, the hyper-parameter T mentioned in Equation 8 is set to 1.0, and the patch size $h \times w$ in PCA is set to 8×10.

Baselines: To comprehensively evaluate our method, we first implement two fusion strategies, namely Concatenate and Multi-level Sum (MLSum), for multi-modal data fusion. Specifically, Concatenate means that we concatenate RGB and thermal images along the channel dimension, exemplifying an early fusion method. While MLSum represents an intermediate fusion method, maintaining the same structure as depicted in Figure 1, we substitute the CPCF with a summation operation at each stage. Furthermore, we take into account

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART MULTISPECTRAL METHODS AND OUR BASELINES ON LLVIP TESTING SET BY MAP IN PERCENTAGE.

Method	Backbone	Fusion	Modality	mAP_{50}↑	mAP_{75}↑	mAP ↑
Fcos [6]	ResNet50	-	RGB	86.8	45.2	46.5
Fcos [6]	ResNet50	-	Thermal	94.2	62.1	57.4
YOLOv5 [41]	Darknet53	-	RGB	88.0	47.8	48.0
YOLOv5 [41]	Darknet53	-	Thermal	94.7	62.4	58.2
YOLOX [4]	Darknet53	-	RGB	89.3	48.3	48.6
YOLOX [4]	Darknet53	-	Thermal	94.4	67.3	60.6
Multi-modal methods						
ECISNet [36]	ResNet50	ECIS	RGB-T	95.7	-	-
UA-CMDet [9]	Darknet53	UA-CM	RGB-T	96.3	-	-
CSAA [34]	ResNet50	CSAA	RGB-T	94.3	66.6	59.2
CFT [42]	CFB	CFT	RGB-T	**97.5**	72.9	63.6
Our baselines						
FcosCAT	ResNet50	Concatenate	RGB-T	94.5	61.6	57.9
FcosSUM	ResNet50	MLSum	RGB-T	95.1	64.8	58.5
YOLOv5CAT	Darknet53	Concatenate	RGB-T	95.1	62.7	58.2
YOLOv5SUM	Darknet53	MLSum	RGB-T	95.6	65.8	59.4
YOLOXCAT	Darknet53	Concatenate	RGB-T	93.4	65.8	58.1
YOLOXSUM	Darknet53	MLSum	RGB-T	93.4	69.0	61.0
Our implementation with CPCF						
FcosCPCF	ResNet50	CPCF	RGB-T	96.0	69.5	60.6
YOLOv5CPCF	Darknet53	CPCF	RGB-T	96.1	70.1	62.0
YOLOXCPCF	Darknet53	CPCF	RGB-T	96.4	**75.4**	**65.0**

detectors that utilize either RGB or thermal inputs, serving as unimodal baselines for comparison. Note that our design pertains only to the encoding part of the model, which allows us to evaluate our method across various detectors such as Fcos [6], YOLOX [4], and S^2A-Net [28]. For each detector, we conduct experiments based on the aforementioned baselines to assess the generalization capability of our proposals.

C. Evaluation Metrics

In our evaluation, we adopt three standard COCO metrics [7], namely mean Average Precision (mAP), mAP_{50}, and mAP_{75}, to quantify the effectiveness of the proposed method. During this process, the Intersection over Union (IoU) is employed as a criterion to classify positive and negative samples. More concretely, a detected instance is deemed a positive sample only when the IoU between the predicted bounding box and the ground truth bounding box surpasses a designated threshold, denoted as τ. Consequently, for mAP_{50} and mAP_{75}, the threshold τ is set at 0.5 and 0.75, respectively. The mAP, on the other hand, is computed with the threshold τ ranging from 0.5 to 0.95 in increments of 0.05.

D. Comparative Studies

1) Quantitative Results: We compare the proposed fusion methods with our baselines and other state-of-the-art methods on FLIR, LLVIP, and DroneVehicle datasets. The experimental results on the FLIR dataset are illustrated in Table II. Note that Fcos [6], YOLOv5 [41], and YOLOX [41] are initially designed for RGB-based object detection, while GAFF [11], CFT [42], YOLOFusion [38], and UA-CMDet [9] are multi-modal-based object detection methods. Then, we extend the unimodal methods to multi-modal based on Concatenate and MLSum and present them as our multi-modal baselines, detailed in Section IV-B. The results show that multi-modal-based methods significantly outperform unimodal-based methods, illustrating that the model can obtain more task-relevant cues from the multi-modal inputs. In addition, our proposed CPCF achieves remarkable performance gains on different detectors, and our methods surpass our baselines and other state-of-the-art methods by a large margin. For example, our YOLOXCPCF outperforms RGB and thermal-based YOLOX by 19.3% and 5.7% on mAP_{50}. Also, compared to our multi-modal baselines, the method exceeds the Concatenate and MLSum-based fusion methods by 4.7% and 5.4% on mAP_{50}, 3.6% and 3.4% on mAP, which shows the advancement and efficiency of our CPCF. Notably, in our multi-modal baselines, the methods based on MLSum outperform those based on concatenation on nearly all metrics. This further illustrates that compared to directly concatenating inputs from different modalities, using an intermediate fusion strategy is more effective in extracting multi-modal information, thereby enhancing the performance of the model. We also observe the consistent performance boosts across various detector types, demonstrating not only the efficacy of our method but also its robust capacity for generalization. In addition, our YOLOv5CPCF and YOLOXCPCF also outperform existing multi-modal methods.

Table III presents the results of our methods and the competing methods on the LLVIP dataset. As can be seen, our methods achieve superior performance than our baselines and other existing methods. Moreover, compared to unimodal methods, it is evident that multi-modal methods significantly improve the regression accuracy of bounding boxes. For instance, our YOLOXCPCF shows a marked increase on the mAP_{75} metric, improving by 26.6% over YOLOX (RGB) and 7.6% over YOLOX (Thermal). Similarly, we obtain a consistent performance improvement even with other types of detectors.

Different from FLIR and LLVIP datasets, DroneVehicle is a more challenging large-scale dataset targeting oriented object detection in low-light conditions. We compare our methods with the state-of-the-art oriented object detectors on the DroneVehicle dataset and report the experimental results in Table IV. Specifically, we modify the detection heads to support oriented detection on standard object detectors, such as FasterRCNN [5] and RetinaNet [45]. Moreover, for state-of-the-art oriented object detection methods, such as S^2A-Net [28], and PSC [29], we adapt their feature extraction structures to accommodate multi-modal inputs. It can be seen that the multi-modal-based methods are considerably better than the unimodal-based methods. For instance, our S^2A-NetCPCF demonstrates a significant improvement over methods based on RGB or thermal images, with the mAP_{50} increases of 12.3% and 3.9%, respectively. In addition, our multi-modal baselines achieve competitive results compared to existing multi-modal-based state-of-the-art methods, and the model results are further enhanced with the benefit of our proposed CPCF strategy. All the experiments conducted on these datasets validate the versatility of our approach across different types of detectors and its generalizability in various

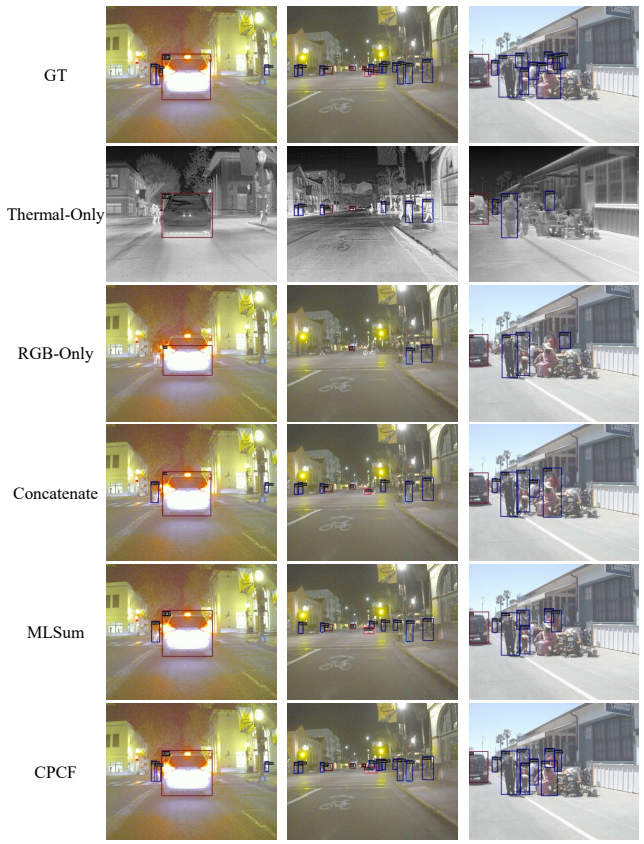

Fig. 4. Qualitative comparison of four baselines and our proposed method on FLIR validation set. Only bounding boxes with a confidence greater than 0.7 are displayed.

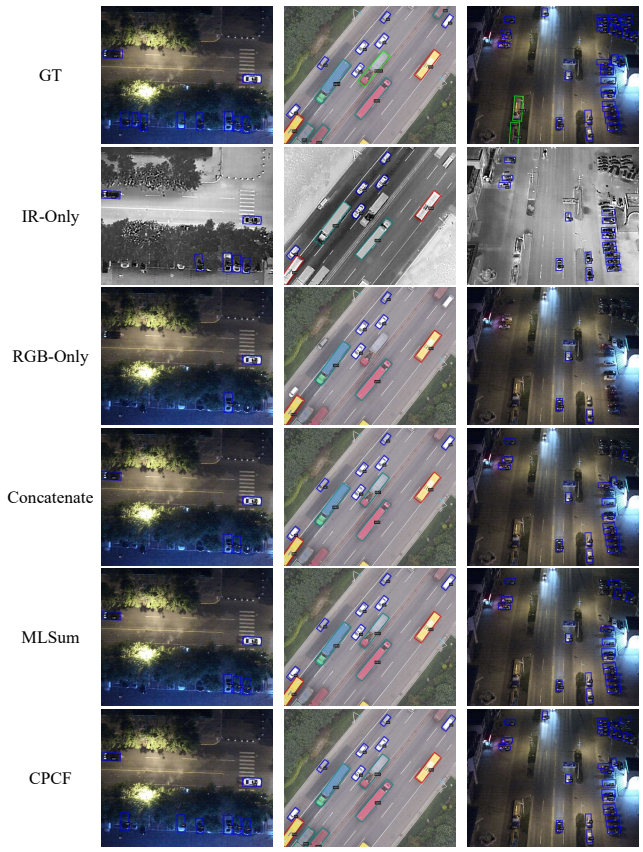

Fig. 5. Qualitative comparison of four baselines and our proposed method on DroneVehicle validation set. Only bounding boxes with a confidence greater than 0.7 are displayed.

scenarios.

2) Qualitative Results: In Figure 4, we compare the detection results of our proposed CPCF with different baselines on the FLIR validation set. In the experiment, we use YOLOX [4] as the base detector to generate single-modality detection results, i.e., RGB-Only and Thermal-Only. Then, for multi-modal fusion, we generated detection results based on Concatenate and MLSum, refer to Section IV-B. As shown in the second and third rows of the figure, the RGB images provide rich texture information under clear weather conditions, while thermal images offer more object clues under low-light conditions. It can be seen that the results generated utilizing RGB images are superior to those generated by thermal images under clear weather conditions, which could be attributed to the lack of texture information in thermal images making it difficult to distinguish different individuals in dense objects. This phenomenon is reversed under low-light conditions, illustrating the complementarity between RGB and thermal images. On the other hand, multi-modal methods attempt to leverage this complementarity. As can be seen from the fourth and fifth rows, multi-modal methods clearly outperform unimodal ones. More specifically, a simple concatenation of RGB and thermal images at the input stage can combine information from different modalities to a certain extent, but it falls short when detecting targets that are unclear in appearance or partially obscured. The use of intermediate fusion strategies can alleviate this issue, but still struggles to handle complex scenarios. CPCF, by employing channel and spatially correlated attention during the intermediate fusion process, effectively utilizes clues from different modalities, achieving the best detection results, as shown in the last row.

Figure 5 illustrates the detection results on the DroneVehicle validation set. We use S^2A-Net [28] as our foundational oriented object detection framework. As can be seen, although our multi-modal baseline improves detection results, it still falls short in detecting obscured or densely clustered objects. For instance, the baseline methods lose the obscured vehicle in the first column scenario and fail to identify the densely packed objects in the upper right corner of the scene in the last column. In contrast, our proposed method demonstrates stable results under these scenarios, further attesting to the effectiveness of our approach.

3) Ablation Study: In this section, we conduct ablation experiments on the FLIR dataset for a detailed analysis of our designs. The CPCF consists of three modules: channel-wise cross-attention (CCA), patch-wise cross-attention (PCA), and adaptive gating (AG) unit. As presented in Table V, we use YOLOX as a case study and progressively incorporate these modules into the model to investigate their individual contributions to the overall performance. Specifically, we employ YOLOXSUM as our multi-modal baseline for a fair

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART MULTISPECTRAL METHODS AND OUR BASELINES ON DRONEVEHICLE TESTING SET BY MAP IN PERCENTAGE.

Method	Modality	Fusion	mAP_{50}↑	mAP_{75}↑	mAP ↑
FasterRCNN [5]	RGB	-	63.0	28.6	31.4
FasterRCNN [5]	Thermal	-	71.9	49.6	43.6
RetinaNet [45]	RGB	-	58.0	26.9	29.5
RetinaNet [45]	Thermal	-	66.6	48.2	41.4
S^2A-Net [28]	RGB	-	64.1	29.4	32.3
S^2A-Net [28]	Thermal	-	74.4	52.5	45.9
PSC [29]	RGB	-	66.9	32.0	33.8
PSC [29]	Thermal	-	75.3	54.8	46.9
Multi-modal methods					
UA-CMDet [9]	RGB-T	UA-CM	63.3	-	-
ECISNet [36]	RGB-T	ECIS	76.0	-	-
Our baselines					
FasterRCNNCAT	RGB-T	Concatenate	74.1	49.5	44.4
FasterRCNNSUM	RGB-T	MLSum	74.7	52.0	45.7
RetinaNetCAT	RGB-T	Concatenate	69.7	49.3	43.1
RetinaNetSUM	RGB-T	MLSum	70.1	51.0	43.8
S^2A-NetCAT	RGB-T	Concatenate	75.7	53.2	46.6
S^2A-NetSUM	RGB-T	MLSum	76.1	55.8	47.8
PSCCAT	RGB-T	Concatenate	75.6	55.4	47.2
PSCSUM	RGB-T	MLSum	77.3	58.0	48.8
Our implementation with CPCF					
FasterRCNNCPCF	RGB-T	CPCF	76.1	52.8	46.6
RetinaNetCPCF	RGB-T	CPCF	72.9	53.01	45.7
PSCCPCF	RGB-T	CPCF	77.8	**58.1**	49.4
S^2A-NetCPCF	RGB-T	CPCF	**79.2**	57.9	**49.7**

TABLE V
ABLATION STUDY OF THE COMPONENTS OF OUR CPCF ON FLIR DATASET. ● AND ○ INDICATE ACTIVATED AND INACTIVATED COMPONENTS, RESPECTIVELY.

Method	CCA	PCA	AG	mAP_{50}↑	mAP_{75}↑	mAP ↑
YOLOXSUM	○	○	○	76.7	37.7	41.2
YOLOXCCA	●	○	○	80.7 (+4.0)	38.6 (+0.9)	43.0 (+1.8)
YOLOXPCA	○	●	○	80.8 (+4.1)	39.8 (+2.1)	43.1 (+1.9)
YOLOXCPCF	●	●	○	81.1 (+4.4)	39.9 (+2.2)	43.4 (+2.2)
YOLOXCPCF	●	●	●	**82.1 (+5.4)**	**41.2 (+3.5)**	**44.6 (+3.4)**

comparison, as shown in the first row of the table. We then replace the summation operation in YOLOXSUM with CCA and PCA respectively. The results from the second and third rows show that the model's performance in terms of mAP improved by 1.8% and 1.9% with the application of CCA and PCA, respectively. Finally, to illustrate the role of AG, we conduct experiments using fixed weights of 0.5 and dynamic weights produced by AG and obtain performance boosts of 2.2% and 3.4%, respectively, as shown in the last two rows of the table. The results reveal that compared to manually setting fixed weights, employing AG can greatly enhance the model's performance. This further suggests that different weights should be assigned to different attention mechanisms at various stages of the model to adapt to the changes in information volume in the channel and spatial dimensions. Therefore, we conclude that the introduction of CCA and PCA can provide more efficient feature extraction capabilities for intermediate fusion from both channel and spatial dimensions, thereby enhancing model performance. Moreover, the dynamic weight allocation mechanism of AG can further optimize fusion efficiency according to changes of information in channel and spatial dimensions, thereby dealing with complex multi-modal data more effectively.

TABLE VI
COMPARISON OF MLP-BASED AND OUR SELF-ATTETNION-BASED CROSS-ATTENTION (CA) ON FILR, LLVIP, AND DRONEVEHICLE DATASETS.

Method	Dataset	CA	mAP_{50}↑	mAP_{75}↑	mAP ↑	Param. (M)↓
YOLOX CPCF	FLIR	MLP-Based	79.7	38.9	42.3	6.50
		Ours	**82.1**	**41.2**	**44.6**	**1.03**
YOLOX CPCF	LLVIP	MLP-Based	94.8	71.7	62.8	6.50
		Ours	**96.4**	**75.4**	**65.0**	**1.03**
S^2A-Net CPCF	Drone Vehicle	MLP-Based	78.0	57.1	48.8	6.50
		Ours	**79.2**	**57.9**	**49.7**	**1.03**

E. Attention Analysis

To further illustrate the effectiveness of the proposed cross-attention mechanism, we employ MLP-based channel and spatial attention to replace our CCA and PCA modules. Notably, the MLP, which squeezes concatenated feature maps into attention maps [16], is widely used in various attention mechanisms, such as [36], [38]. As shown in Table VI, our proposed self-attention-based CCA and PCA significantly outperform the MLP-based attention mechanisms on different datasets. Moreover, we also quantify the parameters of a single standalone cross-attention module, which takes a feature map of size 128×168 with 512 channels as input. As shown in the last column of the table, compared to the MLP-based cross-attention, our method saves approximately 85% of the parameters, proving its higher efficiency.

In an ideal scenario, an effective channel attention mechanism should allocate different weights to various channels based on the amount of information within each channel. Our method concentrates valuable information into a subset of channels and better recognizes these channels with the CCA module, thereby suppressing redundant information while allocating more attention to channels with more information. In Figure 6, we compare the information entropy of different feature channels at different stages depicted in Figure 1. Specifically, we first rank the channels in the feature map according to the channel attention scores, and then calculate the information entropy of the top 16 feature channels and the bottom 16 feature channels, termed top_k and bottom_k, respectively. In the figure, the green distribution describes the information entropy of top_k, while the gray corresponds to bottom_k. We observe that, in the MLP-based channel attention, the information entropy distribution of top_k and bottom_k is strikingly similar. This suggests that the amount of information in top_k and bottom_k is consistent. Therefore, when the attention score of bottom_k is very low, many channels containing valuable information might be suppressed. On the other hand, it is evident that in our method, top_k always contains more information. Particularly in the final stage, as illustrated in the column of L5, the discrepancy between the information distribution of top_k and bottom_k in high-level semantic feature maps is further amplified, demonstrating the effectiveness of our CCA in channel awareness.

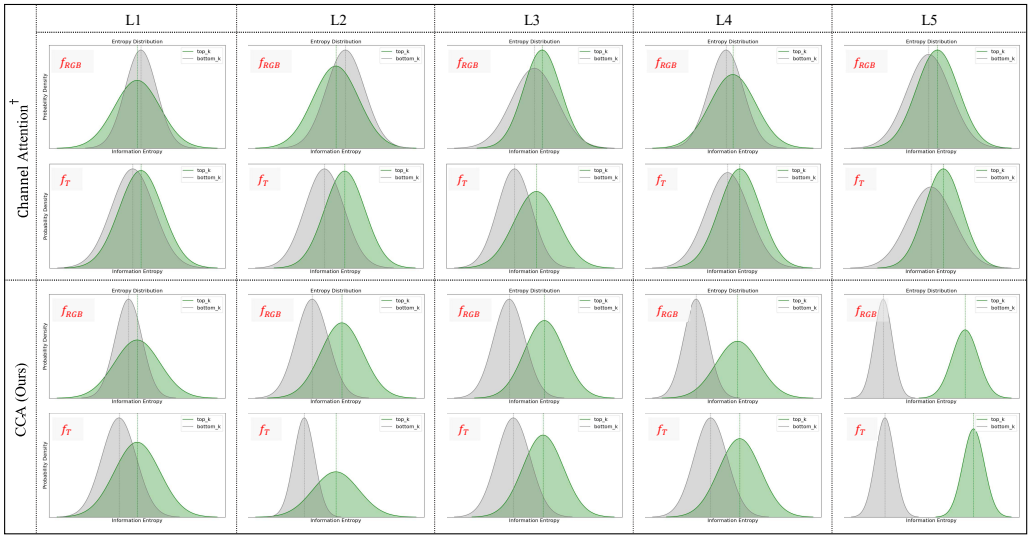

Fig. 6. Comparison of information entropy distributions of top and bottom 16 channels of RGB and Thermal feature maps at different levels. † denotes MLP-based cross-attention. L1-5 represent 5 CPCFs in Figure 1. Green and gray distributions correspond to top_k and bottom_k feature channels, respectively.

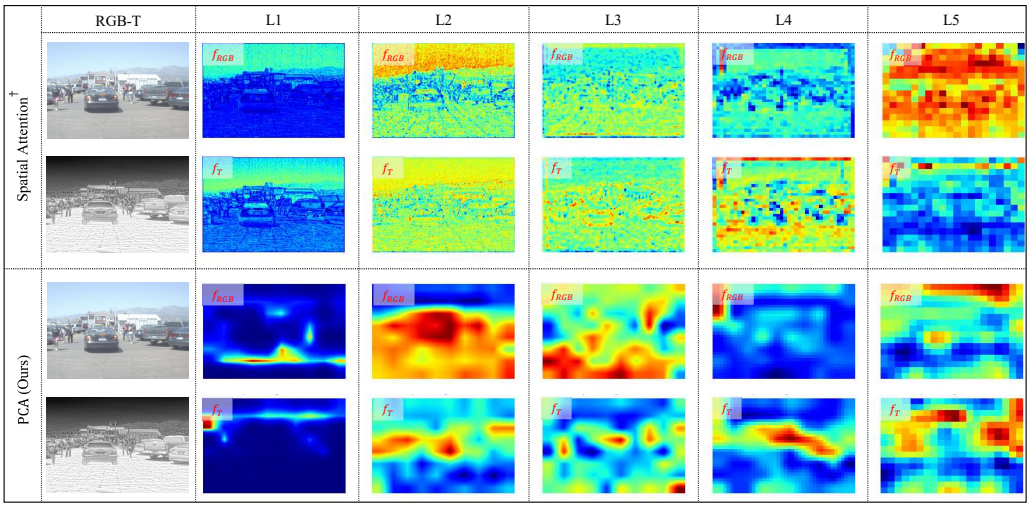

Fig. 7. Comparison of spatial attentions of RGB and Thermal feature maps at different levels. † denotes MLP-based cross-attention. L1-5 represent 5 CPCFs in Figure 1.

Additionally, Figure 7 demonstrates the spatial attention maps in different attention blocks. It is noticeable that, compared to MLP-based spatial attention, the attention maps of different modalities produced by our PCA complement each other to a certain extent., which indicates that our method is able to utilize the complementarity between modalities more efficiently while forming spatial awareness.

F. Speed and Parameter Analysis

To further assess the practicality of our proposed fusion method, we choose the widely used single-stage detectors YOLOv5 and YOLOX as benchmarks and conduct tests to measure the execution speed of our method. In Table VII, we report the total number of learnable parameters, the number of floating-point operations (FLOPs), and the runtime. All models in the experiments are implemented based on MMDetection [43], and running on a laptop equipped with an RTX2080 GPU. It can be observed that multi-modal methods show a decrease in speed compared to unimodal methods. For instance, the runtime of YOLOv5SUM and YOLOXSUM increased by approximately 4ms compared to the single-modality counterparts. This is due to the fact that intermediate fusion introduces additional feature extraction branches, leading to an increase in computational complexity. Additionally, the use of our lightweight fusion module results in a minor increase in runtime. Specifically, CCA adds approximately 3ms, while PCA contributes an extra 5ms. When combining both CCA and PCA, i.e., our CPCF, the runtime increases by around 10ms. Furthermore, compared to our multi-modal baseline, our fusion strategy adds virtually no extra parameters or floating-point operations. In addition, in the last column of Table II, we list the number of parameters for different models. It is evident that our method manages to achieve state-of-the-art performance while ensuring the model remains lightweight.

TABLE VII
COMPARISION OF MODEL PARAMETERS AND FLOPS AND RUNTIME.

Detector	Modality	Param. (M)↓	FLOPs (G)↓	Runtime (ms)↓
YOLOv5	RGB/T	7.03 (-4.17)	6.35 (-4.17)	20.3 (-3.2)
YOLOv5SUM	RGB-T	11.20 (±0.0)	10.52 (±0.0)	23.5 (±0.0)
YOLOv5CCA	RGB-T	11.26 (+0.06)	10.53 (+0.01)	27.1 (+3.6)
YOLOv5PCA	RGB-T	12.60 (+1.40)	10.60 (+0.08)	29.1 (+5.6)
YOLOv5CPCF	RGB-T	12.66 (+1.46)	10.61 (+0.09)	35.1 (+11.6)
YOLOX	RGB/T	8.94 (-4.21)	10.66 (-4.38)	11.6 (-4.3)
YOLOXSUM	RGB-T	13.15 (±0.0)	15.04 (±0.0)	15.9 (±0.0)
YOLOXCCA	RGB-T	13.22 (+0.07)	15.05 (+0.01)	18.9 (+3.0)
YOLOXPCA	RGB-T	14.55 (+1.40)	15.12 (+0.08)	21.5 (+5.6)
YOLOXCPCF	RGB-T	14.61 (+1.46)	15.13 (+0.09)	26.7 (+10.8)

V. CONCLUSION

In this work, we present a lightweight multi-modal cross-fusion method termed CPCF for visible-infrared object detection, which consists of channel-wise cross-attention (CCA), patch-wise cross-attention (PCA), and an adaptive gating (GA) unit. The CCA and PCA are designed to refine valuable cues from the channel and spatial dimensions, respectively, and operate the features of one modality to calibrate another, thereby better integrating the information of different modalities. Moreover, we argue that the useful multi-modal information contained within channel and spatial dimensions can vary during the forward propagation process. To account for this, we design the AG unit to adaptively adjust the attention weights in the channel and spatial dimensions. Subsequently, based on the CPCF, we design a universal intermediate fusion architecture that allows for extension to various types of detectors, facilitating the harnessing of multi-modal information to enhance the model's performance. Finally, we conduct extensive experiments with various object detection frameworks on standard and oriented object detection datasets. The results demonstrate that our method is able to effectively capture information from different modalities and consistently outperform other advanced multi-modal methods. Additionally, thanks to its lightweight design, our method can be incorporated into lightweight object detection models, enabling real-time object detection.

In the future, it will be worthwhile to further optimize fusion algorithms to enhance the efficiency of model integration and explore the application of CPCF on different types of modality data, such as Depth maps. Moreover, applying CPCF to different computer vision tasks, such as semantic segmentation, to further investigate its generalizability across more architectures is another promising avenue of research.

REFERENCES

[1] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2021.

[2] A. Raghunandan, Mohana, P. Raghav, and H. V. R. Aradhya, "Object detection algorithms for video surveillance applications," in *2018 International Conference on Communication and Signal Processing (ICCSP)*, 2018, pp. 0563–0568.

[3] Y. Pi, N. D. Nath, and A. H. Behzadan, "Convolutional neural networks for object detection in aerial imagery for disaster response and recovery," *Adv. Eng. Informatics*, vol. 43, p. 101009, 2020.

[4] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *ArXiv*, vol. abs/2107.08430, 2021.

[5] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[6] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9626–9635, 2019.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[8] T. Baltruaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 423–443, 2017.

[9] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6700–6713, 2022.

[10] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 12 878–12 895, 2022.

[11] H. ZHANG, É. Fromont, S. Lefèvre, B. Avignon, and U. de Rennes, "Guided attentive feature fusion for multispectral pedestrian detection," *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 72–80, 2021.

[12] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *ArXiv*, vol. abs/1611.02644, 2016.

[13] J. Shen, Y. Chen, Y. Liu, X. Zuo, H. Fan, and W. Yang, "Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection," *Pattern Recognition*, vol. 145, p. 109913, 2024.

[14] W.-Y. Lee, L. Jovanov, and W. Philips, "Cross-modality attention and multimodal fusion transformer for pedestrian detection," in *Computer Vision – ECCV 2022 Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds. Cham: Springer Nature Switzerland, 2023, pp. 608–623.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[17] H. Zhang, É. Fromont, S. Lefèvre, and B. Avignon, "Multispectral fusion for object detection with cyclic fuse-and-refine blocks," *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 276–280, 2020.

[18] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "Llvip: A visible-infrared paired dataset for low-light vision," *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3489–3497, 2021.

[19] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2013.

[20] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, pp. 154–171, 2013.

[21] R. B. Girshick, "Fast r-cnn," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, 2015.

[22] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, Jun 2015.

[23] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *ArXiv*, vol. abs/1804.02767, 2018.

[24] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *ArXiv*, vol. abs/2004.10934, 2020.

[25] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *ArXiv*, vol. abs/2207.02696, 2022.

[26] L. Liu, Z. Pan, and B. Lei, "Learning a rotation invariant detector with rotatable bounding box," *ArXiv*, vol. abs/1711.09405, 2017.

[27] X. Pan, Y. Ren, K. Sheng, W. Dong, H. Yuan, X.-W. Guo, C. Ma, and C. Xu, "Dynamic refinement network for oriented and densely packed object detection," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 204–11 213, 2020.
[28] J. Han, J. Ding, J. Li, and G. Xia, "Align deep features for oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2020.
[29] Y. Yu and F. peng Da, "Phase-shifting coder: Predicting accurate orientation in oriented object detection," *ArXiv*, vol. abs/2211.06368, 2022.
[30] S. Speth, A. Gonçalves, B. Rigault, S. Suzuki, M. Bouazizi, Y. Matsuo, and H. Prendinger, "Deep learning with rgb and thermal images onboard a drone for monitoring operations," *Journal of Field Robotics*, vol. 39, pp. 840 – 868, 2022.
[31] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *The European Symposium on Artificial Neural Networks*, 2016.
[32] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong, "Multimodal object detection via probabilistic ensembling," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 139–158.
[33] D. König, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 243–250, 2017.
[34] Y. Cao, J. Bin, J. Hamari, E. Blasch, and Z. Liu, "Multimodal object detection by channel switching and spatial attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023, pp. 403–411.
[35] Y. Zheng, I. H. Izzat, and S. Ziaee, "Gfd-ssd: Gated fusion double ssd for multispectral pedestrian detection," *ArXiv*, vol. abs/1903.06999, 2019.
[36] Z. An, C. Liu, and Y. Han, "Effectiveness guided cross-modal information sharing for aligned rgb-t object detection," *IEEE Signal Processing Letters*, vol. 29, pp. 2562–2566, 2022.
[37] K. Zhou, L. Chen, and X. Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in *Computer Vision – ECCV 2020*. Cham: Springer International Publishing, 2020, pp. 787–803.
[38] F. Qingyun and W. Zhaokui, "Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery," *Pattern Recognition*, vol. 130, p. 108786, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320322002679
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[40] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "Cspnet: A new backbone that can enhance learning capability of cnn," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1571–1580, 2019.
[41] G. Jocher, "Ultralytics yolov5," Oct. 2020. [Online]. Available: https://github.com/ultralytics/yolov5
[42] F. Qingyun, H. Dapeng, and W. Zhaokui, "Cross-modality fusion transformer for multispectral object detection," *arXiv preprint arXiv:2111.00273*, 2021.
[43] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.
[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, Jun 2009, pp. 248–255.
[45] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

VI. Biography Section

Sijie Hu received his master's degrees in 2017 in the field of Mechanical Engineering and Automation from Hefei University of Technology. He is currently working towards the Ph.D. Degree at the IBISC lab, Science and Technologies of Information and Communication doctoral school, university of Paris-Saclay. His research interests include multi-modal learning, domain adaptation, object detection and tracking, and semantic segmentation.

Fabien Bonardi received in 2017 the Ph.D. degree in electrical and computer engineering from Normandie University, UNIROUEN (LITIS Lab). He is assistant professor at Université d'Évry/Université Paris-Saclay since 2018. His research concerns computer vision, visual perception in robotics and sensors fusion for non-conventional cameras.

Samia Bouchafa received a Ph.D. degree in 1998 in the field of electrical and computer engineering from University Paris VI (now Sorbonne University) and the French Research Institute in Transportation. In 1999, she was an assistant professor at University Paris XI, then an associate professor in 2011. Since 2012, she has had a full professor position at Univ. Evry/Université Paris-Saclay. Since 2019, she has been the director of the IBISC laboratory. Her research concerns computer vision, visual perception, multi-modal vision for autonomous systems, motion analysis, stereovision, visual odometry, and localization.

Helmut Prendinger (Member, IEEE) received the master's and Ph.D. degrees in logic and artificial intelligence from the University of Salzburg, Austria. He held positions as a research associate and a JSPS post-doctoral fellow with The University of Tokyo. In 1996, he was a Junior Specialist with the University of California at Irvine, Irvine. He is currently a Full Professor with the National Institute of Informatics, Tokyo. His current research interests include unmanned aircraft systems traffic management (UTM) and machine learning (ML), especially deep learning for drone use cases. His team contributes to developing the entire UTM system as part of large-scale Japanese government projects. He has published more than 250 refereed papers in international journals and conferences. His H-index is 47.

Désiré Sidibé (Senior Member, IEEE) graduated from Ecole Centrale de Nantes (MSc Eng., 2004) and from University of Montpellier (PhD, 2007). He is currently a Full Professor at Université Evry - Paris Saclay, and a member of the IBISC laboratory. His research interests include computer vision for autonomous vehicles and medical image analysis. He is an IEEE senior member and has authored about 100 papers in international journals and conferences. He serves as an Associate Editor for IEEE Robotics and Automation Letters.