Dual-Perspective Alignment Learning for Multimodal Remote Sensing Object Detection

Yanfeng Liu, *Student Member, IEEE,* Wei Guo, Chaojun Yao, and Lefei Zhang, *Senior Member, IEEE*

***Abstract*—Recently, anchor-based detectors can achieve decent performance in multimodal remote sensing scenarios, whereas their anchor-free counterparts fail to reach comparable results. To remedy this problem, we first comprehensively investigate the misalignment issues in multimodal features and detection heads, and present a dual-perspective alignment learning (DPAL) framework for multimodal remote sensing object detection. Particularly, we design a cross-modal alignment module (CMAM), which utilizes the multiscale dilation strategy and differentiable alignment function with channel-wise modulation for cross-modal feature integration. Additionally, to cope with the misalignment problem in regression and classification heads, we propose a task-head alignment module (THAM). It presents a novel pseudo-anchor mechanism, introduces a semi-fixed offset generation strategy to capture task-variant sampling coordinates, and ultimately deploys an offset knowledge transfer mechanism with deformable alignment for anchor-free detection heads. Extensive experiments on four multimodal object detection datasets show impressive results of the proposed DPAL framework. The project code is released at *https://github.com/lyf0801/DPAL*.**

***Index Terms*—Object detection, remote sensing, RGB-Infrared imagery, anchor-free, alignment learning.**

I. Introduction

OBJECT detection aims to recognize the category and location information of potential objects, which serves as a fundamental visual task for remote sensing image processing, providing wide applications for land cover classification [1], scene understanding [2], and image caption [3]. With the development of deep learning [4], object detection has made remarkable achievements in optical aerial imagery [5], [6]. However, in harsh environments, such as low-light conditions, there is still a huge challenge, as the full picture of objects cannot be captured by optical remote sensing sensors solely.

Subsequently, some researchers propose a vision task known as multispectral/multimodal object detection [7], which aims to identify and localize objects from optical and infrared image pairs. Early research efforts focus on pedestrian and vehicle detection by incorporating visible and thermal images in traffic scenarios. Moreover, some researchers present object detection in optical and infrared remote sensing images [8]. Recently, benefiting from the fast inference and straightforward design of continuously updated YOLO detectors, numerous anchor-based YOLO algorithms have been proposed for multimodal remote sensing images [9], [10], [11]. For example, Zhang et al. [12] present a super-resolution auxiliary decoder to facilitate object detection. To our best knowledge, anchor-based approaches have been extensively investigated in multimodal remote sensing images [11], [12], [13], which show excellent results on public datasets. However, a notable drawback of anchor-based detectors is the necessity of manual predefinition of massive multiscale anchors, which is time-consuming and might not be optimally adaptable to complicated scenarios. In contrast, anchor-free detectors eliminate predefined anchors, they enable end-to-end one-stage regression, which simplifies the detection process and potentially reduces computational overhead. Particularly, Huang et al. [14] observe a universal performance degradation of anchor-free methods relative to their anchor-based counterparts. The inherent limitations of anchor-free detectors [15], typically the absence of predefined anchors and two-stage regression, become more pronounced in multimodal remote sensing applications. For instance, Su et al. [16] highlight that FCOS exhibits a significant performance gap compared to anchor-based methods in both single- and dual-modal scenarios on VEDAI and DroneVehicle datasets.

To bridge the above deficiency, we pay attention to investi-

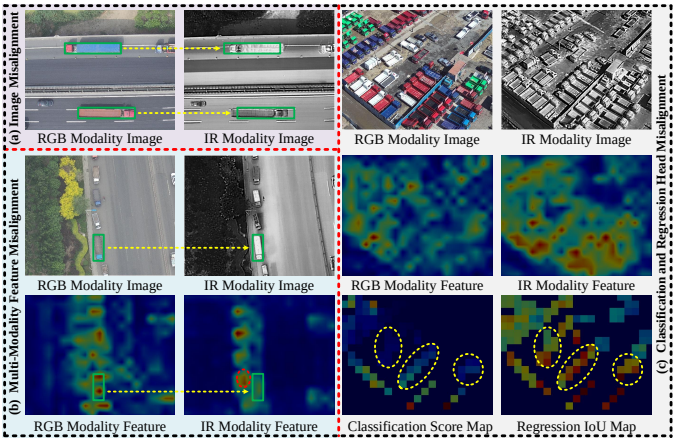

Fig. 1. Typical misalignment issues that exist in multimodal object detection as follows: (a) multi-modality image misalignment; (b) multi-modality feature misalignment; (c) classification and regression head feature misalignment.

Manuscript received 28 November 2024; revised 15 March 2025 and 28 May 2025; accepted 3 June 2025. Date of publication xx xxx 2025; date of current version xx xxx 2025. This work was supported by the National Natural Science Foundation of China under Grant 62431020, the Fundamental Research Funds for the Central Universities under Grant 2042025kf0030, and Seed Funding Project of Multisensor Intelligent Detection and Recognition Technologies R&D Center of CASC. The numerical calculations in this work had been supported by the supercomputing system in the Supercomputing Center of Wuhan University. *(Corresponding author: Lefei Zhang.)*

Yanfeng Liu and Lefei Zhang are with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan, 430072, P. R. China (email: liuyanfeng99@whu.edu.cn; zhanglefei@whu.edu.cn).

Wei Guo and Chaojun Yao are with the Multisensor Intelligent Detection and Recognition Technologies R&D Center of CASC, Chengdu, P. R. China (email: guowei3158@126.com; ycjhit@163.com).

Digital Object Identifier 10.1109/TGRS.2025.xxxxxxx

gating a novel anchor-free detection framework for multimodal remote sensing imagery. Generally, existing anchor-free detectors cannot reach comparable results to anchor-based methods [17]. However, the reasons behind the aforementioned problem and potential solutions are not yet clear. In this article, we first argue that the significant concerns are several misalignment issues in multimodal anchor-free detectors shown in Fig. 1.

Specifically, we find three misalignment problems as follows. **First**, some object instances in infrared and optical images are not aligned due to diverse differences in multispectral imaging sensors. Although it is possible to manually align infrared and optical images, serious misalignment problems still remain. As illustrated in Fig. 1(a), the locations of two trucks are shifted between optical and infrared images, and their offsets are distinct from each other. Conversely, the trucks marked with green boxes in Fig. 1(b) are almost strictly aligned. These non-regular multispectral imaging discrepancies pose a severe challenge for multimodal feature fusion in deep neural networks [18]. **Second**, as revealed in Fig. 1(b), although the identical objects are seemingly aligned in multimodal images, suffering from inconsistent downsampling and imprecise fusion of convolutional networks, the activation positions of the same object in cross-modal features appear to be misaligned. This is also an essential factor that we believe constrains the performance of anchor-free detectors. **Third**, we observe a significant difference between IoU scores from the regression head and category predictions from the classification head in Fig. 1(c). We attribute two main reasons as follows: firstly, the restricted point features of 3×3 convolutions cannot depict complete and accurate object regions, and secondly, the lack of RoI alignment and two-stage regression further increases the misalignment of regression and classification features. The aforementioned misalignment problems inevitably exacerbate the challenge of anchor-free multimodal object detection.

To address the above-mentioned problems, we propose a **D**ual-**P**erspective **A**lignment **L**earning (DPAL) approach for multimodal remote sensing object detection. The uncertain misalignment of optical and infrared images is an inherent attribute of the model inputs [18], and thus we cannot elegantly cope with this problem in an end-to-end detection framework. Hence, we propose to incorporate the adaptive and generalized spatial alignment learning strategy in the multimodal feature fusion phase to unify the multispectral image misalignment and multimodal feature misalignment issues at the deep feature level. Besides, we also present a pseudo-anchor generation and a deformable alignment mechanism into anchor-free detection heads to alleviate the misalignment issues between the regression head and the classification head, ultimately unleashing the potential of multimodal anchor-free detectors [19]. Different from the existing multispectral detection approaches [20], the proposed two unsupervised alignment-based learning modules are equipped with an anchor-free baseline model for adaptive multimodal feature integration, head feature coordinate resampling, and offset knowledge transfer. As shown in Fig. 2(c), the presented methodology is distinguished from the existing single-modal and multimodal anchor-free detection paradigms [15], and our main contribution and novelty is to propose two plug-and-play modules to handle the misalignment problems

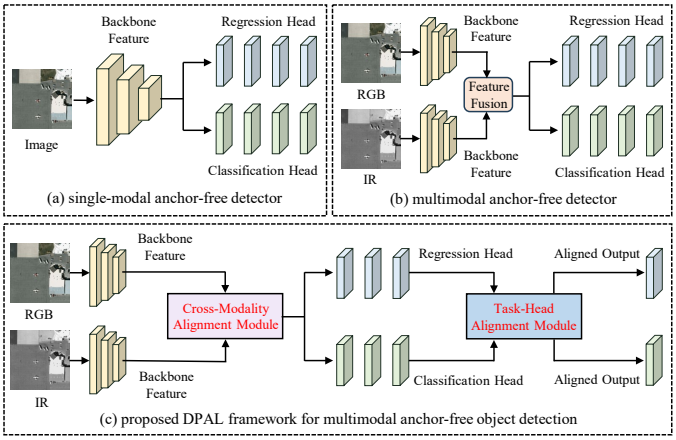

Fig. 2. Comparison of the proposed dual-perspective alignment framework with existing single-modality and multimodal anchor-free object detectors.

in anchor-free object detectors, i.e., cross-modal alignment module (CMAM) and task-head alignment module (THAM). With the aforementioned efforts, we aim for the presented approach to reach comparable performance with anchor-based methods, e.g., RetinaNet [21], paving a way to compensate for the shortcomings of anchor-free detectors and bringing a new perspective and insight to the remote sensing community. To reveal the effectiveness of the proposed DPAL, we conduct experiments on four widely-used multispectral datasets, including two datasets for remote sensing and two ones for traffic scenarios. Extensive quantitative, qualitative, and visual studies show the advantages of DPAL, as well as some merits of both alignment learning modules via ablation study.

The main contributions of this article are listed as follows:

1) We summarize several misalignment issues in multimodal object detection, and argue they are critical factors in limiting the performance of anchor-free detectors.
2) To tackle the uncertain misalignment issue between multimodal features, we propose a cross-modal alignment module (CMAM) with adaptive offset learning, differential alignment, and channel modulation mechanisms.
3) To handle the misalignment problem between regression and classification sub-task heads, we present a task-head alignment module (THAM) with pseudo-anchor strategy, semi-fixed offset learning, and offset knowledge transfer.
4) Based on the above research efforts, we design DPAL framework, providing new research insights for multispectral or multimodal remote sensing object detection.

The remaining article is organized as follows. Section II provides the related studies. Section III presents the methodology of the proposed DPAL framework. We conduct experiments in Section IV and draw a conclusion in Section V.

II. Related Work

Here, we summarize related studies on remote sensing multimodal detection and alignment learning for object detection.

A. Remote Sensing Multimodal Object Detection

Since Razakarivony et al. [8] release the VEDAI dataset incorporating multimodal image pairs of optical and infrared

modalities, remote sensing multimodal object detection has continued to attract the research interest in recent years [10], [11], [12]. As an early deep learning-based research work, Sharma et al. [11] propose a mid-level multimodal fusion detector based on YOLO, and Fang et al. [10] introduce modal-invariant and modal-specific attention mechanisms for parallel multimodal feature fusion for joint detection.

Overall, existing related studies almost focus on employing single-modal detectors, e.g., YOLO, as a baseline to design modal-adaptive, spatial-wise, channel-wise, global-wise, local-wise, or self-attention mechanisms to refine multimodal features [9], [13], [16]. For example, Zhang et al. [13] present a feature enhancement module with multiscale convolutions and feature fusion blocks via learnable spatial weights. Su et al. [16] propose a low-rank enhancement approach and a dynamic illumination-aware mask module to unbiasedly and compatibly extract multimodal features from the frequency domain.

Besides, several works investigate DETR-like models for multimodal remote sensing object detection. For instance, Zhu et al. [22] introduce a multispectral DETR framework via deformable attention mechanism. Guo et al. [23] present modality competitive query selection mechanism and multispectral deformable cross-attention module with DETR framework.

Furthermore, some researchers pay attention to detecting small or tiny objects from multimodal remote sensing imagery, such as persons and vehicles. For example, a RGB-Thermal tiny person detection model [24] is proposed based on the quality-aware learning strategy and cross-modal enhancement module. Xu et al. [25] leverage spatial and channel attention mechanisms to combine RGB and infrared features for airborne small object detection. Based on physical simulation images, Cao et al. [26] introduce a template matching approach with a dynamic template library for small object detection.

Recently, there are some interesting ideas about state space model [27] and cross-modal distillation [28] for multimodal remote sensing object detection, which have contributed significantly to this field with novel research insights.

B. Alignment Learning for Object Detection

Some existing studies focus on alignment learning for object detection tasks, and most of them are proposed for unimodal object detection. We divide these efforts into object proposal perspective, feature pyramid perspective, and detection head level. For object proposal alignment, Han et al. [29] introduce the align convolution for sampling points of oriented object proposals in remote sensing imagery. Furthermore, Xie et al. [30] extend align convolution to adaptive aligned convolution by constraining deformable convolution and align convolution. With respect to feature pyramid perspective, Xu et al. [31] propose pyramidal representative feature alignment for adaptive object detection. Huang et al. [32] introduce deformable convolution to integrate adjacent-stage spatial features in feature pyramid networks. Song et al. [33] present an accurate feature alignment approach via graph matching for image and point cloud pairs. Wang et al. [34] also design an adjacent alignment module to dynamically integrate multiscale spatial features for salient object detection. Regarding detection head level, Feng et al. [35] propose task-aligned head for one-stage detectors. Xie et al. [36] present a deformable alignment approach to explore the correlation between regression and classification heads. Zhao and Wang [37] present several task-specific loss functions to align the disagreement problem of both subtasks.

Furthermore, researchers extend alignment to domain adaptive, few-shot, and 3D object detection. For instance, He et al. [38] propose a partial alignment-based asymmetric detector for domain adaptive object detection. Wang et al. [39] combine intermediate domain image generation and domain-adversarial training via an augmented feature alignment framework. Chu et al. [40] align source-similar and source-dissimilar samples in the unified feature space by adversarial learning for source-free object detection. Han et al. [41] present an attention-based feature alignment mechanism for few-shot object detection.

In addition, some research works concentrate on misaligned multispectral object detection [42], [43], [44]. Among them, Fu et al. [42] propose to adapt a single-stage detector trained on aligned multimodal imagery to non-aligned visible-infrared image pairs. Chen et al. [43] present the attentive positional alignment to match pedestrian regions between complementary modalities. Furthermore, a deformable convolution-based alignment approach for multimodal feature fusion is proposed in [44]. However, none of the existing works have adequately considered the multi-level misalignment challenges in multispectral aerial imagery. To bridge this gap, we first investigate a dual-perspective alignment framework in this article.

III. Methodology

Here, we describe the methodology of the proposed DPAL framework. We firstly present the overview of DPAL, then introduce two alignment learning modules, i.e., CMAM and THAM, and finally the anchor-free loss function is discussed.

A. Overview of the Proposed DPAL Framework

As illustrated in Fig. 3, the proposed DPAL framework consists of three parts, i.e., a modal-specific encoder, five CMAMs, and a THAM. For an optical (RGB) and infrared (IR) image pair I_{rgb} and I_{ir}, DPAL first feeds them into the modal-specific encoder, which is composed of dual independent parallel branches for individual modality images. Specifically, each branch includes a ResNet50/PVTv2 backbone and a feature pyramid network (FPN) with lateral connection [45], and yields five multiscale modal-specific features for both I_{rgb} and I_{ir}, which denote as $F_{rgb}^1 \sim F_{rgb}^5$ and $F_{ir}^1 \sim F_{ir}^5$.

As shown in Fig. 3(b), for a multimodal feature pair F_{rgb}^i and F_{ir}^i at the same level, CMAM utilizes them as input information, and employs a multiscale dilated learning method to generate several spatial offset groups for both RGB and IR features. Then, we deploy a parameter-free differential alignment strategy to perform multi-group alignment for individual modality features, and obtain the aligned RGB and IR features. Finally, we introduce a channel-wise modulation and fusion solution to integrate aligned RGB and IR features. Thus, a combined cross-modal alignment representation is produced, which indeed reduces background noise and produces more accurate and fine-grained spatial activation for remote sensing

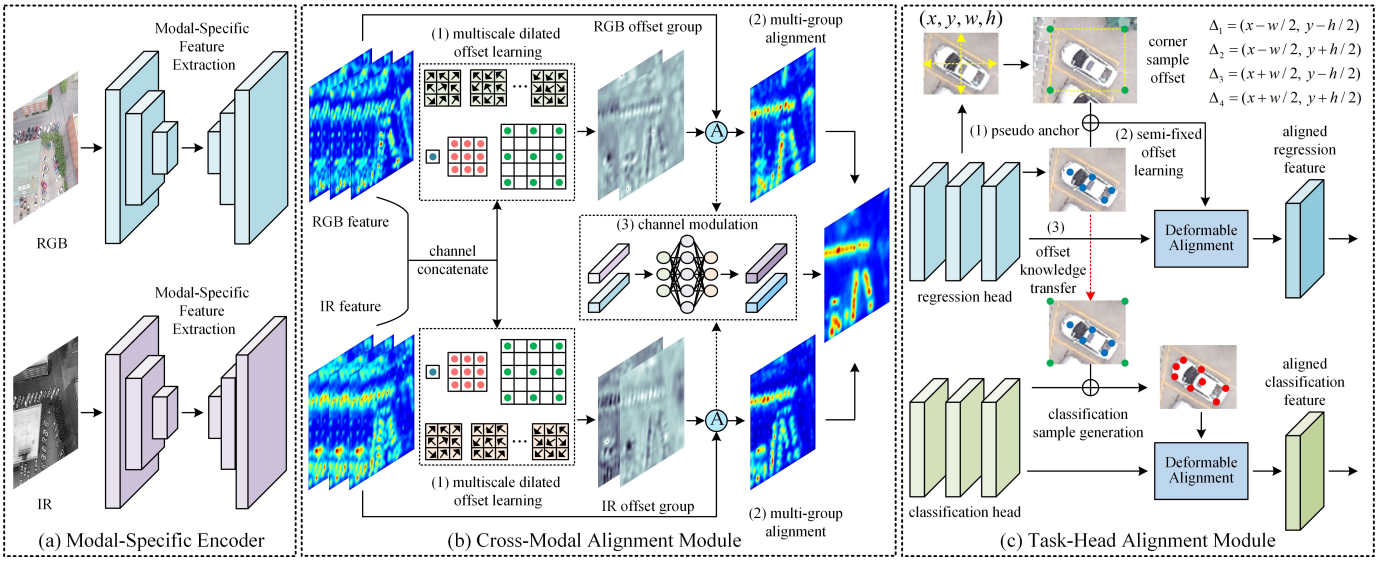

Fig. 3. Illustration of the proposed DPAL framework. (a) Modal-specific encoder. (b) Cross-modal alignment module. (c) Task-head alignment module.

objects. Considering that FPN adopts a five-level architecture, DPAL also contains five independent CMAMs to acquire five cross-modal aligned features in a multiscale pyramidal manner.

With respect to THAM, we illustrate its methodology in Fig. 3(c). It first feeds the cross-modal aligned feature as input, and utilizes two parallel groups of three 3×3 convolutions for initial feature generation of regression and classification heads. Then, a pseudo-anchor mechanism and a semi-fixed offset learning strategy are presented to project the regression offsets, thereby the aligned regression features for anchor-free regression prediction could be calculated via the deformable alignment method. Motivated by the fact that classification and regression differ in their preferences for spatial features [46], we argue the regression offsets are not an optimal shape for the classification task. Therefore, we perform classification sample generation from initial classification features, and then design an offset knowledge transfer function to refine classification offsets via residual learning. In the same way, the deformable alignment approach is also utilized to calculate aligned classification features for anchor-free object classification prediction.

B. Cross-Modal Alignment Module (CMAM)

Existing remote sensing multimodal object detection algorithms widely propose elaborate multimodal feature fusion blocks, but ignore the feature misalignment problem, which is an intrinsic issue in deep convolutional networks [34], and the principal factors include discrete downsampling and imprecise fusion. As shown in Fig. 1(a)-(b), there simultaneously exists indeterminate misalignment at the image and feature levels for multispectral image pairs. Obviously, this co-existing uncertain misalignment problem exacerbates the confusion between RGB and IR features, further induces multimodal feature misalignment, and seriously hinders the fusion generalization. To address this issue, we design a novel CMAM module that makes efforts to align the cross-modal features and generate accurate multimodal integrated features adaptively.

As illustrated in Fig. 3(b), the core process of CMAM consists of three steps, i.e., 1) multiscale dilated offset learning, 2) group-wise alignment mapping, and 3) cross-modal channel-wise modulation and fusion. Overall, the purpose of CMAM is to adaptively leverage two coarse and unaligned RGB and IR features, F_{rgb} and F_{ir}, then employ the above techniques and ultimately obtain the aligned cross-modal representation.

1) Multiscale Dilated Offset Learning: First of all, considering the limited receptive field of standard convolution, we propose to utilize multiscale dilated convolutions to project more generalized spatial offsets for both RGB and IR features. Specifically, our presented multiscale dilated convolutions include 1×1 point-wise convolution, standard 3×3 convolution, and 3×3 convolution with dilation rate of 2. With respect to two coarse RGB and IR features F_{rgb} and F_{ir}, the multiscale dilated offset approach first concatenates them as a unified tensor and deploys various convolution operators in a parallel manner, to generate RGB and IR offsets as follows:

$$\Delta_{rgb}, \Delta_{ir} = \mathcal{C}_{3\times3}(\mathcal{C}_{1\times1}(\mathcal{C}_{3\times3}([F_{rgb}, F_{ir}]), \mathcal{C}^2_{3\times3}([F_{rgb}, F_{ir}]))), \quad (1)$$

where $\mathcal{C}_{i\times i}$ denotes the function of standard $i \times i$ convolution, $\mathcal{C}^2_{3\times3}$ represents 3×3 convolution with dilation rate of 2, and $[\cdot,\cdot]$ is channel-wise concatenation. In particular, we present multi-group alignment mapping in CMAM, so that Δ_{rgb} and Δ_{ir} both contain several offset groups with a wider view.

2) Group-Wise Alignment Mapping: Motivated by group convolution, we first divide the initial F_{rgb} and F_{ir} into n groups, i.e., $F^1_{rgb}, ..., F^n_{rgb}$ and $F^1_{ir}, ..., F^n_{ir}$, to bring more adaptive diversity, then split multi-group offsets Δ_{rgb} and Δ_{ir} into $\Delta^1_{rgb}, ..., \Delta^n_{rgb}$ and $\Delta^1_{ir}, ..., \Delta^n_{ir}$ as prior knowledge, and finally leverage a non-parametric alignment function to project aligned RGB and IR features in Fig. 3(b). Mathematically,

$$\begin{aligned} F'_{rgb} &= \mathcal{C}_{3\times3}(\left[\mathcal{F}_A(F^1_{rgb}, \Delta^1_{rgb}), ..., \mathcal{F}_A(F^n_{rgb}, \Delta^n_{rgb})\right]), \\ F'_{ir} &= \mathcal{C}_{3\times3}(\left[\mathcal{F}_A(F^1_{ir}, \Delta^1_{ir}), ..., \mathcal{F}_A(F^n_{ir}, \Delta^n_{ir})\right]). \end{aligned} \quad (2)$$

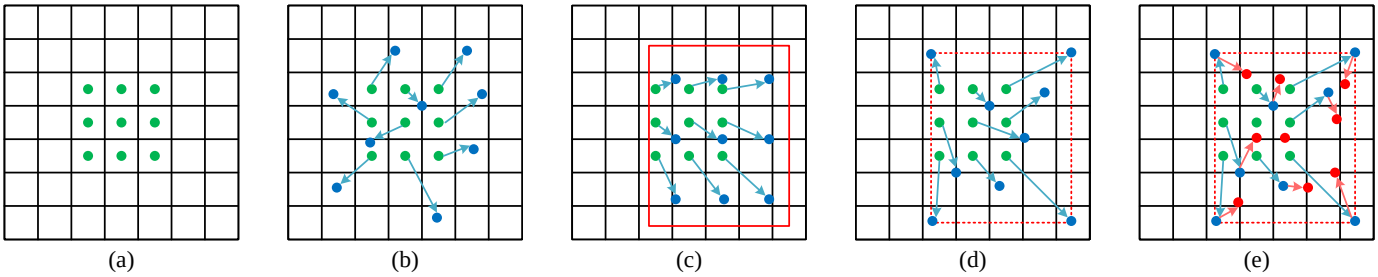

Fig. 4. Illustration of various sampling point methods with 3×3 kernels. (a) standard convolution with fixed sampling points. (b) deformable convolution [47]. (c) AlignConv [29]. (d) the proposed anchor-free regression head alignment mechanism. (e) the proposed anchor-free classification alignment mechanism. Note that the red box represents proposal anchor in anchor-based detectors, and red dashed boxes denotes our presented pseudo anchors in anchor-free detectors.

Here, F'_{rgb} and F'_{ir} denote the aligned RGB and IR features, and $\mathcal{F}_A(\cdot,\cdot)$ indicates the alignment function as follows:

$$\mathcal{F}_A(F,\Delta) = \sum_{w'}^{W}\sum_{h'}^{H} F_{w',h'} \cdot \max(0, 1-|w+\Delta^x_{w,h}-w'|) \cdot \max(0, 1-|h+\Delta^y_{w,h}-h'|), \tag{3}$$

where w' and h' are the horizontal and vertical index of a subgroup feature F, and $\Delta = \{\Delta^x, \Delta^y\}$ denote the learnable transformation offset elements for pixels in F. Typically, this alignment approach employs a bi-linear interpolation kernel on the spatial location ($h+\Delta^x_{w,h}$, $w+\Delta^y_{w,h}$) to resample and align F, hence refining cross-modal features. More implementation details for how to find partial derivatives are discussed in [48].

3) Channel-Wise Modulation and Fusion: As presented in Fig. 3(b), before constructing the final cross-modal alignment representations, a simple yet effective channel-wise attention mechanism is introduced to suppress noise and remove background false activation of the aligned RGB and IR features. First, we employ the global average pooling (GAP) operations to capture the overall abstract semantic knowledge and generate two global features, i.e., $v_{rgb} = \text{GAP}(F'_{rgb})$ and $v_{ir} = \text{GAP}(F'_{ir})$. Then, two fully connected layers with activation functions are utilized to perform channel-wise nonlinear mapping and global adaptive knowledge reconstruction. Formally, the above-mentioned process is defined as follows:

$$\left[v'_{rgb}, v'_{ir}\right] = \text{Tanh}(W_2 \cdot \sigma(W_1 \cdot [v_{rgb}, v_{ir}] + b_1) + b_2), \tag{4}$$

where $\text{Tanh}(\cdot)$ and $\sigma(\cdot)$ represent Tanh and Sigmoid activation functions, respectively. W_1 and W_2 denote the weight matrices of fully connected layers, b_1 and b_2 indicate the bias matrices.

Eventually, we deploy a self-contained integration method to compute the distinctive cross-modal alignment representation by simultaneously leveraging the aligned multimodal features F'_{rgb}, F'_{ir} and refined global semantic vectors v'_{rgb}, v'_{ir}, i.e.,

$$F_{align} = \mathcal{C}_{3\times3}((1+v'_{rgb}) \cdot F'_{rgb} \oplus (1+v'_{ir}) \cdot F'_{ir}), \tag{5}$$

where $\oplus$ denotes element-wise summation function. As shown in Fig. 3(b), the ultimate cross-modal alignment feature provides the finer-grained spatial response, which maximizes activation of regions of interest in multimodal image pairs.

In summary, CMAM adopts a multi-group offset learning strategy, a parameter-free differentiable alignment function, and a channel-wise modulation approach, which fully utilizes various learnable operators to eliminate the misalignment issue of multimodal images in feature perspective as much as possible and ultimately generates adaptive cross-modal aligned representations, thus offering to anchor-free detection heads.

C. Task-Head Alignment Module (THAM)

Since there are no pre-defined anchors to ensure as much recall as possible, existing anchor-free object detectors [15] always underperform their anchor-based counterparts. In addition, due to the overhead view and wide imaging scopes of aerial sensors, extensive ground objects possess multiscale characteristics. As illustrated in Fig. 4(a), the fixed point-based convolutional features cannot capture sufficient receptive fields, especially for large-scale and irregular remote sensing objects. Furthermore, classification and regression do not share the similar preferences of spatial features, e.g., classification prefers the topology, edge, and texture of objects, while regression additionally relies more on the contextual information around objects [46]. Recently, several research works propose to utilize deformable convolution [47] or aligned convolution [29] (as shown in Fig. 4(b) and 4(c)) to roughly enlarge the receptive fields [36]. However, they neglect the intrinsic correlation between regression and classification, and struggle to generalize to anchor-free detectors. Based on the above deficiency, we present THAM to tackle these issues.

As illustrated in Fig. 3(c), THAM mainly consists of three novel technical procedures as follows: 1) pseudo-anchor mechanism, 2) semi-fixed offset learning, and 3) offset knowledge transfer. Here, we describe the detailed phases as follows. First of all, we employ dual separate branches of three 3×3 convolutions to obtain the initial features of classification and regression heads from the output of CMAM (i.e., F_{align}), and term them as F_{reg} and F_{cls}. Motivated by the predefined anchors of anchor-based detectors, THAM introduces a learnable pseudo anchor (x, y, w, h) for each spatial point in the aligned cross-modal feature F_{align} to provide the region scope information for a potential object. As for the pseudo anchor, four corner coordinates of a region, i.e., x_{min}, x_{max}, y_{min}, y_{max}, are calculated as follows:

$$\begin{aligned} x_{min} &= x - w/2, \quad x_{max} = x + w/2, \\ y_{min} &= y - h/2, \quad y_{max} = y + h/2. \end{aligned} \tag{6}$$

To capture contextual information around the object, we consider these four corner coordinates of pseudo-anchor as

four fixed resampling points. Thus, these four corner offsets $\Delta_{corner1}, ..., \Delta_{corner4}$ can be denoted as (x_{min}, y_{min}), (x_{min}, y_{max}), (x_{max}, y_{min}), and (x_{max}, y_{max}), respectively. We regard the above four corner points as spatial constraint conditions, and project five adaptive offsets within the range of pseudo anchors from F_{reg} by two convolutional layers, i.e., $\{\Delta^1_{reg}, ..., \Delta^5_{reg}\}$, where the offset generation function $\mathcal{F}_{offset}(\cdot)$ is defined as follows:

$$\mathcal{F}_{offset}(x) = \sigma(\mathcal{C}_{3\times3}(\text{ReLU}(\mathcal{C}_{3\times3}(x)))). \tag{7}$$

Essentially, four fixed corner offsets can serve as contextual knowledge for an object, while the other five adaptive sampling points could characterize the local information. Based on this insight, we combine four fixed corner offsets and five conditional offsets as the kernel resampling knowledge of a 3×3 convolution for regression subtask, defined as Δ_{reg}, and named this process as semi-fixed offset learning as illustrated in Fig. 4(d). In contrast to fully adaptive offset generation of deformable convolution version 2 (DCNv2) [49] and internal well-distributed sampling within anchors of AlignConv [29] in Fig. 4, the presented semi-fixed offset learning proposes a pseudo-anchor mechanism and incorporates the advantages of the fully adaptive learning within feature maps in Fig. 4(b) and the uniform sampling within anchors in Fig. 4(c), which greatly compensates for the misalignment of regression head.

On account of distinct spatial feature preferences between classification and regression sub-tasks, we argue that regression offsets can not provide an effective representation for classification. Based on this deficiency, we present two additional efforts on top of the fundamental classification sample generation from initial classification features. First, we assume the central coordinate of the pseudo anchor (x, y) as the geometric center of potential objects, and consider it as a fixed offset (defined as Δ_{center}) that benefits object classification. Therefore, we only need to generate eight resampling offsets for the classification offset learning via Eq. 7, termed as $\{\Delta^1_{cls}, ..., \Delta^8_{cls}\}$. Second, to model the intrinsic correlation between regression and classification, and considering the prior knowledge of regression offset, an offset residual transfer strategy from regression offset Δ_{reg} to classification offset Δ_{cls} is designed to refine classification offset. The above steps for classification offset generation could be defined as

$$\Delta_{cls} = \Delta_{reg} \oplus \{\Delta_{center}, \Delta^1_{cls}, \Delta^2_{cls}, ..., \Delta^8_{cls}\}. \tag{8}$$

Fig. 4(e) illustrates the proposed classification offset method.

Based on two offset representations Δ_{reg} and Δ_{cls} and two task-specific features F_{reg} and F_{cls}, we can utilize the deformable alignment approach [49] to calculate aligned features for anchor-free regression and classification prediction, i.e.,

$$F'_{reg} = \mathcal{F}_D(F_{reg}, \Delta_{reg}) \oplus F_{reg}, \ F'_{cls} = \mathcal{F}_D(F_{cls}, \Delta_{cls}) \oplus F_{cls}, \tag{9}$$

where $\mathcal{F}_D(\cdot,\cdot)$ indicates the function of deformable alignment learning for a spatial point x_p in a feature map as follows:

$$\mathcal{F}_D(x_p, \Delta_p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta_{p_k}) \cdot \Delta m_k, \tag{10}$$

where p_k denotes 9 sampling points $\{(0,0), (-1,0), ..., (1,1)\}$ of a standard 3×3 convolution, k is the sampling index, Δ_{p_k} is our proposed spatial offset index, w represents convolution weights, and m is the modulator scalar in DCNv2 [49]. The operation procedure of THAM is summarized in Algorithm 1.

Algorithm 1 : Task-Head Alignment Learning Mechanism

Input: regression feature F_{reg}, classification feature F_{cls}
Output: regression output F'_{reg}, classification output F'_{cls}
THAM (F_{reg}, F_{cls}):
  Generate pseudo anchor $(x, y, w, h) \leftarrow F_{reg}$
  Calculate corner coordinates $\{x_{min}, y_{min}, x_{max}, y_{max}\}$
  Project four corner offsets $\Delta_{corner1}, ..., \Delta_{corner4}$
  Obtain regression offsets $\Delta_1, ..., \Delta_5 \leftarrow F_{reg}$ via Eq. (7)
  for $i = 1, ..., 5$ **do**
    $\Delta_{i,x} \leftarrow \min(\max(\Delta_{i,x},\ x_{min}),\ x_{max})$
    $\Delta_{i,y} \leftarrow \min(\max(\Delta_{i,y},\ x_{min}),\ x_{max})$
  end for
  $\Delta_{reg} \leftarrow \{\Delta_{corner1}, ..., \Delta_{corner4},\ \Delta_1, ..., \Delta_5\}$
  Define object center sample offset $\Delta_{center} \leftarrow (x, y)$
  Obtain regression offsets $\Delta_1, ..., \Delta_8 \leftarrow F_{cls}$ via Eq. (7)
  for $i = 1, ..., 8$ **do**
    $\Delta_{i,x} \leftarrow \min(\max(\Delta_{i,x},\ x_{min}),\ x_{max})$
    $\Delta_{i,y} \leftarrow \min(\max(\Delta_{i,y},\ x_{min}),\ x_{max})$
  end for
  $\Delta_{cls} \leftarrow \Delta_{reg}$, Δ_{center}, and $\Delta_1, ..., \Delta_8$ via Eq. (8)
  $F'_{reg} \leftarrow F_{reg}$ and Δ_{reg} via Eqs. (9) and (10)
  $F'_{cls} \leftarrow F_{cls}$ and Δ_{cls} via Eqs. (9) and (10)
Return F'_{reg}, F'_{cls}

In conclusion, the presented THAM enables anchor-free detectors to better adaptively resample object regions by developing feasible alignment learning mechanisms for classification and regression subtasks, and thus facilitates the collaborative object detection in multimodal remote sensing images.

D. Loss Functions for Anchor-Free DPAL Detector

The presented DPAL framework adopts the anchor-free paradigm of FCOS [19] with a total loss function containing a classification loss, a regression loss, and a centerness loss, respectively. Because of no predefined anchors, the regression objective is different from anchor-based detectors. Formally, suppose a bounding box is $B = \{x_1, y_1, x_2, y_2, c^*\}$, where (x_1, y_1) and (x_2, y_2) are the coordinates of the upper-left and lower-right corners, and c^* denotes the category label. If a spatial point (x, y) lies within B in the feature map, then the regression target (l, t, r, b) for point (x, y) is represented as

$$\begin{aligned} l &= (x - x_1)/s, \quad t = (y - y_1)/s, \\ r &= (x_2 - x)/s, \quad b = (y_2 - y)/s, \end{aligned} \tag{11}$$

where s is the stride factor for the corresponding feature stage.

FCOS defines a center score based on the above regression target (l, t, r, b) for each predicted bounding box, i.e.,

$$s_c = \sqrt{\frac{\min(l, r)}{\max(l, r)} \times \frac{\min(t, b)}{\max(t, b)}} \in [0, 1], \tag{12}$$

Obviously, if the center of the predicted bounding box is close to the center of ground truth, the center score will be

TABLE I
TYPICAL DETAILED INFORMATION OF FOUR PUBLIC MULTISPECTRAL OBJECT DETECTION DATASETS FOR OUR EXPERIMENTS

Datasets	Image Type	Image Size	Train Set	Test Set	Classes	Category Names	Instances
VEDAI [8]	Aerial+Infrared	1024×1024	1,089	121	8	car, pickup, camping car, truck, other, tractor, boat, van	3,644
DroneVehicle [50]	Drone+Infrared	712×840	17,990	8,980	5	car, truck, bus, van, freight car	953,087
FLIR [51]	RGB+Thermal	512×640	4,129	1,013	3	bicycle, car, person	40,752
M3FD [52]	RGB+Thermal	640×640	2,905	1,295	6	person, car, bus, motorcycle, lamp, truck	34,408

approaching 1. Conversely, the center score will be close to 0. Thus, the total loss function of DPAL can be formulated as

$$L_{total} = \frac{1}{N_{pos}} \sum \left(L_{cls} + \lambda_2 L_{reg} + \lambda_3 L_{center} \right), \quad (13)$$

where N_{pos} denotes the total number of positive samples, and L_{cls} indicates the focal loss function [21] as follows:

$$L_{cls} = -\alpha_t (1 - c_t)^{\gamma} log(c_t), \quad (14)$$

where α_t and γ are hyperparameters for focal loss, and we set α_t and γ to 0.25 and 2, respectively. c_t is defined as follows:

$$c_t = \begin{cases} c & \text{if } y = 1, \\ 1 - c & \text{otherwise}, \end{cases} \quad (15)$$

where $c \in [0, 1]$ indicates the predicted category classification probability for the class with label $y = 1$.

L_{center} is defined as a binary cross-entropy loss for positive samples. Thus, for a predicted centerness score $s_c \in [0, 1]$ and a ground truth score $s_c^* \in [0, 1]$, L_{center} is denoted as

$$\mathcal{L}_{center}(s_c, s_c^*) = -s_c^* \log(s_c) - (1 - s_c^*) \log(1 - s_c), \quad (16)$$

In addition, $L_{reg} = 1 - \text{GIoU}$, which indicates the GIoU loss [19] for bounding box regression optimization.

IV. EXPERIMENTS

A. Experimental Protocol

1) Datasets: We perform experiments on four public multispectral object detection datasets, including two remote sensing optical and infrared object detection datasets (VEDAI [8] and DroneVehicle [50]), and two traffic visible and thermal object detection datasets (FLIR [51] and M3FD [52]). Their major information is briefly summarized as follows:

VEDAI: is released by Razakarivony et al. [8] in 2016, which consists of 1,246 pairs of RGB and IR aerial images with more than 3,700 object instances. This dataset provides two spatial versions, i.e., 512×512 and 1024×1024, where 1024×1024 version is adopted for our experiments. Following Sharma et al. [11], eight object categories with 3,644 instances are chosen for detection, i.e., car, pickup, camping car, truck, other, tractor, boat, and van. Referring to [12], we split 1,089 pairs for training and 121 pairs for test with cross-validation.

DroneVehicle: is a large-scale RGB-Infrared vehicle detection dataset [50] captured by drones. It contains 28,439 image pairs with 953,087 instances under various lighting conditions, including five categories: car, truck, bus, van, and freight car. It contains 17,990 pairs for training, 1,469 for validation, and 8,980 for testing. The image size is uniformly 712×840. Following [53], [54], [16], we transfer the bounding boxes for horizontal object detection. In this work, we train the model in the training dataset and make inferences in the test dataset.

FLIR: is a complicated RGB-Thermal object detection dataset in traffic scenes proposed by [51]. It includes three object categories: bicycle, car, and person, and consists of 5,142 image pairs with the size of 512×640. Referring to the official split [51], we adopt 4,129 image pairs for the training set, and the other 1,013 pairs for the test set. The number of total object instances is 40,752 within the FLIR dataset.

M3FD: provides object detection annotations with six categories, i.e., person, car, bus, motorcycle, lamp, and truck [52]. It contains 4,200 image pairs with various image sizes, but without the official dataset split. Since M3FD is segmented from multiple video sequences, random splits will result in similar samples and information leakage between training and test sets. Thus, following [55], we obtain 2,905 training pairs and 1,295 test pairs. Additionally, we unify the image size to 640×640 pixels for both training and inference phases.

The detailed dataset comparison of is shown in Table I.

2) Evaluation Metrics: As the basic and widely utilized metrics for object detection, we choose AP and mAP50 for performance indicators. The definitions of them are as follows:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i, \ \text{AP} = \int_0^1 P(R) dR, \quad (17)$$

where AP is the average precision of each class, P is precision, R is recall, and mAP is mean average precision of N classes. mAP50 means the IoU threshold of truth positive is 0.5.

3) Implementation Details: We conduct experiments on 4 NVIDIA RTX 4090 GPUs with Ubuntu 20.04 system and PyTorch2.1 toolbox. We deploy the anchor-free FCOS with a ResNet50 backbone as baseline. The training batch sizes for VEDAI, DroneVehicle, FLIR, and M3FD are set to 2, 6, 8, and 6 for ResNet-based DPAL, and 2, 3, 8, and 6 for PVT-based DPAL, respectively. We employ AdamW optimizer with an initial learning rate of 1e-4 and a weight decay of 1e-4 for training. The training process includes 500 warm-up iterations, with total epochs set to 40 for DroneVehicle and FLIR, and 60 for VEDAI and M3FD. The learning rate will drop to 1e-5 after 30 epochs for all datasets. The total training time of PVT-based DPAL for DroneVehicle is nearly 10.6 hours on 4 NVIDIA 4090 GPUs, and for VEDAI, FLIR, and M3FD datasets are approximately 5.4 hours, 6.1 hours, and 7.8 hours on a single NVIDIA RTX 4090 GPU, respectively. The group number of CMAM is set to 32, and the dilation rate of dilated convolution is set to 2. Empirically, we follow FCOS and set both λ_2 and λ_3 to 1 which works well in experiments.

TABLE II
QUANTITATIVE COMPARISON OF OBJECT DETECTION WITH SEVERAL STATE-OF-THE-ART METHODS ON THE VEDAI DATASET [8]

Methods	Publication	Year	Modality	Car	Pickup	Camping	Truck	Other	Tractor	Boat	Van	mAP
YOLOrs [11]	JSTARS	2021	I	82.03	73.92	63.80	54.21	43.99	54.39	21.97	43.38	54.71
			R	85.25	72.93	70.31	50.65	42.67	76.77	18.65	38.92	57.00
			R+I	84.15	78.27	68.81	52.60	46.75	67.88	21.47	57.91	59.73
RetinaNet† [21]	TPAMI	2020	I	88.68	84.50	87.05	81.30	40.14	51.79	71.39	78.78	72.95
			R	91.35	82.48	79.76	78.16	60.81	75.35	69.58	69.02	75.82
			R+I	93.36	87.86	**90.61**	83.39	51.91	72.94	74.22	68.24	77.82
FCOS† [19]	TPAMI	2022	I	82.19	75.52	89.01	92.73	56.13	59.25	73.70	74.15	75.34
			R	88.96	82.42	87.56	86.35	58.65	68.44	78.05	71.48	77.74
			R+I	85.86	79.22	86.08	90.93	57.27	75.36	70.53	80.20	78.20
YOLOFusion [10]	PR	2022	I	86.70	75.90	66.60	77.10	43.00	62.30	70.70	84.30	70.80
			R	91.10	82.30	75.10	78.30	33.30	81.20	71.80	62.20	71.90
			R+I	91.70	85.90	78.90	78.10	54.70	71.90	71.70	75.20	75.90
SuperYOLO [12]	TGRS	2023	I	87.90	81.39	76.90	61.56	39.39	60.56	46.08	71.00	65.60
			R	90.30	82.66	76.69	68.55	53.86	79.48	58.08	70.30	72.49
			R+I	91.13	85.66	79.30	70.18	57.33	80.41	60.24	76.50	75.09
GH-YOLO [56]	TGRS	2023	R+I	89.15	83.57	76.19	59.55	53.05	78.70	59.58	70.71	71.31
FFCA [13]	TGRS	2024	R+I	89.60	85.70	78.70	85.70	48.60	81.80	61.50	67.00	74.80
L-FFCA [13]	TGRS	2024	R+I	91.30	85.50	72.80	79.70	47.30	79.00	56.10	73.90	73.30
C^2Former [44]	TGRS	2024	R+I	87.20	80.70	82.70	77.40	58.40	72.90	71.40	75.20	75.70
CrossYOLO [9]	GRSL	2024	R+I	91.60	**90.60**	78.60	66.60	74.00	77.50	81.50	74.60	79.40
YOLOFIV [57]	JSTARS	2024	R+I	**93.89**	87.42	82.10	80.13	60.84	**82.15**	75.47	79.28	80.16
DPAL-R (Ours)	–	2024	R+I	85.39	80.41	87.09	85.58	**74.94**	76.24	80.31	77.35	80.91
DPAL-P (Ours)	–	2024	R+I	88.10	83.88	86.79	**95.53**	67.98	71.07	**87.06**	**87.05**	**83.43**

FCOS is our baseline, † is our re-implementation, R means RGB modality, I indicates infrared modality.

B. Comparison with State-of-the-Art Methods

Here, we describe the comparison study on four datasets. Particularly, we report Tables II-VI and illustrate Figs. 5-6.

1) Comparison on VEDAI and DroneVehicle: As reported in Table II, we provide five single-modal algorithms (YOLOrs [11], RetinaNet [21], FCOS [19], YOLOFusion [10], and SuperYOLO [12]) for RGB, IR, and RGB+IR object detection on the VEDAI dataset, respectively. Besides, six state-of-the-art methods for multimodal object detection are also introduced, i.e., GH-YOLO [56], FFCA [13], L-FFCA [13], C^2Former [44], CrossYOLO [9], and YOLOFIV [57]. Firstly, we find that the models trained on RGB images show higher performance than that of IR modality. In addition, combining the multimodal images can yield consistent performance improvements, revealing the complementary properties of optical and infrared images. As for our presented framework, although DPAL-R and DPAL-P cannot achieve the most excellent AP on all categories among competitors, they reach the best mAP on the whole dataset, i.e., 80.91% and 83.43%, illustrating its cross-category trade-offs and backbone-agnostic capability.

Compared to state-of-the-art multimodal detectors, most of which are developed based on anchor-based YOLOs. Especially for CrossYOLO [9], it yields better results than anchor-free FCOS [19]. In contrast, our DPAL-P breaks through the limitation of anchor-free paradigm and outperforms the above YOLO-based methods, which demonstrates the validity of the proposed dual-perspective alignment modules, highlighting its robustness and potential for broader applications.

To further analyze the performance of the proposed DPAL under different IoU thresholds and object scales, we report Table III. It can be observed that our DPAL-R outperforms

TABLE III
COMPARISON OF AP UNDER VARIOUS IOU THRESHOLDS ON VEDAI [8]

Methods	Year	AP_{50}	AP_{75}	$AP_{50:95}$	AP_S	AP_M	AP_L
RetinaNet† [21]	2020	77.82	47.37	46.80	39.02	51.24	37.82
FCOS† [19]	2022	78.20	51.69	46.94	43.67	50.29	54.92
SuperYOLO† [12]	2023	74.99	50.58	45.01	39.42	47.98	41.63
FFCA [13]	2024	74.80	–	44.80	44.60	–	–
ICAFusion [20]	2024	76.62	–	44.93	–	–	–
C^2Former [44]	2024	75.70	–	48.30	–	–	–
DPAL-R (Ours)	2024	80.91	54.38	49.52	43.65	51.98	**55.51**
DPAL-P (Ours)	2024	**83.43**	**54.56**	**49.98**	**47.53**	**52.56**	45.15

the baseline FCOS across nearly all metrics, demonstrating the effectiveness of dual-perspective alignment mechanisms. Furthermore, DPAL-P achieves comprehensive superiority over RetinaNet and SuperYOLO in all indicators. The above observation indicates constructive contributions of DPAL.

With respect to the DroneVehicle dataset, we select four single-modality methods, namely SSD512 [58], RFB [59], RetinaNet [21], and FCOS [19] for comparison. Additionally, we introduce six state-of-the-art multimodal object detection approaches, including CGRP [18], FCOS [19], LAIIFusion [60], ICAFusion [20], C^2Former [44], and LFMDet [16], and report their performance on the DroneVehicle dataset in Table IV to validate the strengths and weaknesses of these algorithms. Overall, the proposed DPAL-R and DPAL-P achieve the best average performance among all competitors. Especially, DPAL-P attains 74.82% in the truck category and 94.16% AP in the bus category, which demonstrates its superiority. Unlike the VEDAI dataset, the IR modality in the DroneVehicle dataset provides richer and more valuable

Fig. 5. Zoom in for a better view. Visual detection comparison on the VEDAI test dataset. (a) RGB FCOS baseline. (b) IR FCOS baseline. (c) RGB+IR dual FCOS. (d) DPAL. (e) Ground Truth. Note that the red dashed circles represent false detection, and green dashed circles indicate missing detection.

TABLE IV
QUANTITATIVE COMPARISON ON THE DRONEVEHICLE DATASET [50]

Methods	Year	Modal	Car	Truck	Freight	Bus	Van	mAP
SSD512 [58]	2016	R	82.33	56.01	39.84	85.97	44.47	61.72
RFB [59]	2018	R	77.84	53.60	40.75	82.33	41.71	59.24
RetinaNet† [21]	2020	R	81.58	50.36	40.63	86.18	40.33	59.82
FCOS-R† [19]	2022	R	86.22	54.35	40.34	87.82	36.96	61.14
FCOS-P† [19]	2022	R	87.01	59.73	45.45	88.71	42.70	64.72
SSD512 [58]	2016	I	89.92	65.32	56.97	89.11	54.40	71.14
RFB [59]	2018	I	88.65	62.60	**62.11**	89.22	54.45	71.41
RetinaNet† [21]	2020	I	94.34	59.33	56.56	91.38	51.28	70.58
FCOS-R† [19]	2022	I	91.89	61.16	47.20	89.73	38.70	65.73
FCOS-P† [19]	2022	I	95.09	68.71	56.13	92.76	47.83	72.10
CGRP [18]	2022	R+I	89.90	66.40	60.80	88.90	51.30	71.40
FCOS† [19]	2022	R+I	94.32	70.30	53.18	92.26	42.72	70.55
LAIIFusion [60]	2023	R+I	94.50	54.40	57.90	90.50	33.90	66.20
ICAFusion∗ [20]	2024	R+I	81.60	56.00	33.30	85.70	31.80	57.70
C^2Former∗ [44]	2024	R+I	83.10	69.60	60.50	88.90	55.70	71.60
LFMDet [16]	2024	R+I	82.20	73.60	59.60	86.60	**57.00**	71.80
DPAL-R (Ours)	2024	R+I	**95.34**	72.12	56.07	93.86	45.37	72.55
DPAL-P (Ours)	2024	R+I	95.25	**74.82**	58.69	**94.16**	51.60	**74.91**

† denotes our re-implementation, ∗ indicates the re-implementation in [16] , R denotes RGB modality, I indicates infrared modality.

information, enabling unimodal methods to achieve better results from the infrared images. This is because most scenes in the DroneVehicle dataset are captured under extremely dark conditions, while the IR spectrum reflects more object information than the RGB modality. As for IR baselines, RetinaNet [21] outperforms FCOS-R [19], primarily due to the advantages of anchor-based mechanisms. To address this deficiency, we focus on misalignment issues of multimodal images and introduce a dual-perspective alignment approach at cross-modal feature representation and task-head feature levels. Based on these contributions, the proposed DPAL-R framework achieves 72.55% mAP50 on DroneVehicle, while DPAL-P further improves this to 74.91%, significantly narrowing the performance gap between anchor-based and anchor-free methods for multimodal remote sensing object detection.

2) Comprehensive Analysis of Practical Value: As shown in Fig. 5, the unimodal baseline exhibits frequent misidentification of similar categories (e.g., classifying vans as cars), while DPAL demonstrates excellent performance for small and tiny objects. Fig. 6 presents visual comparisons on the complex DroneVehicle dataset, where baseline methods struggle with substantial inter-class similarity among vehicle types, resulting in persistent false detections. The visualizations in Fig. 6(d) show that DPAL achieves the most accurate prediction for three vehicle categories (cars, trucks, and buses), overcoming both aerial-view feature limitations and inter-class similarity challenges without any omitted instances or false detections. Notably, DPAL excels in three critical scenarios: 1) low-light/nighttime conditions, 2) naturally occluded objects, and 3) misalignment-dominated environments, exhibiting nearly zero false positives or missed detections. Practically, its anchor-free design eliminates computational overhead from

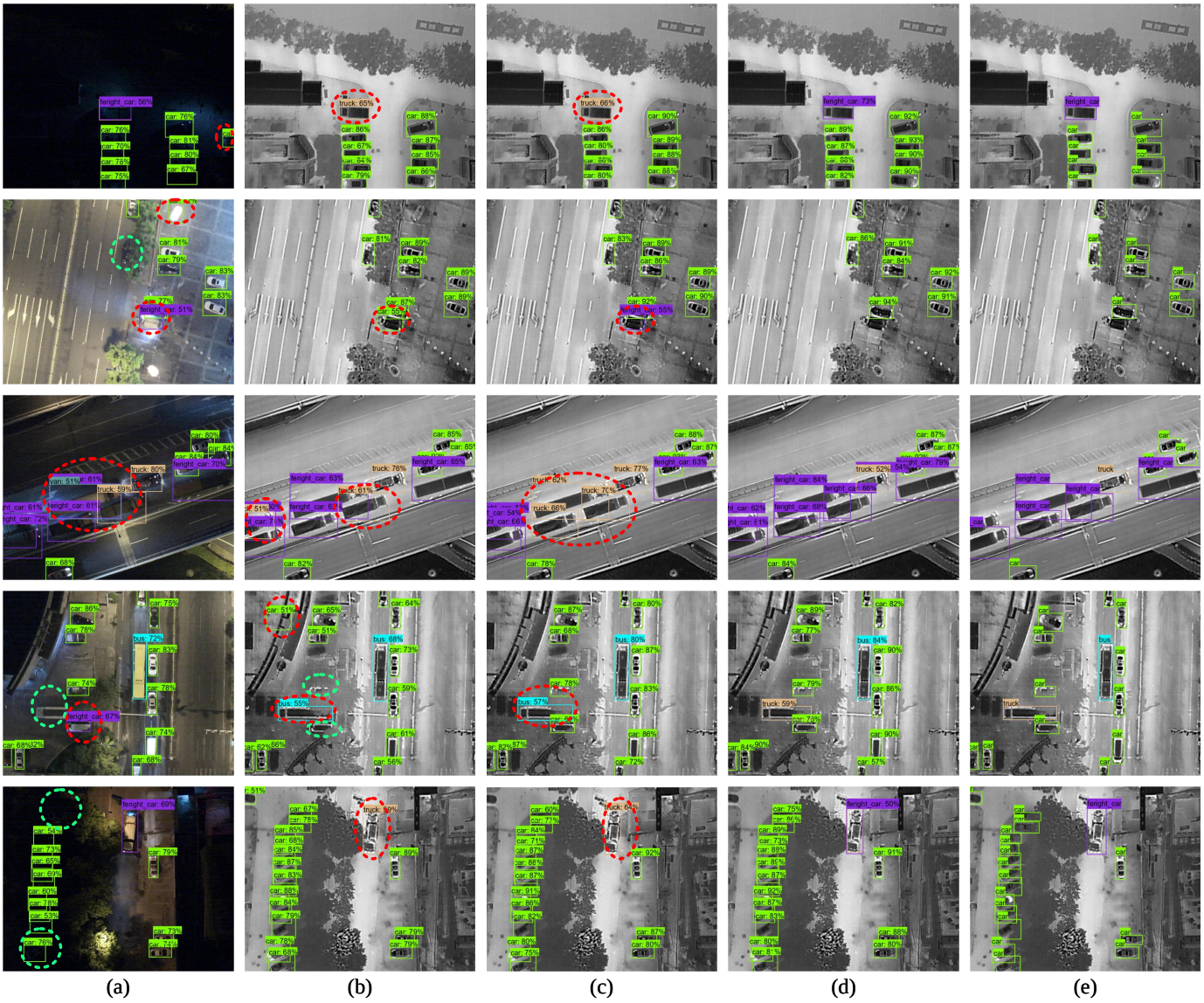

Fig. 6. Zoom in for a better view. Visual detection comparison on the DroneVehicle test dataset. (a) RGB FCOS baseline. (b) IR FCOS baseline. (c) RGB+IR dual FCOS. (d) DPAL. (e) Ground Truth. Note that the red dashed circles represent false detection, and green dashed circles indicate missing detection.

anchor-based methods, significantly improving real-time inference efficiency while maintaining practical applicability.

3) Comparison on FLIR and M3FD: To further verify the generalization of DPAL, we consider two RGB-Thermal object detection datasets in traffic scenes, i.e., FLIR [51] and M3FD [52], and present the quantitative analysis as follows. As for the FLIR dataset, it contains three object categories: bicycle, car, and person. Several state-of-the-art methods, e.g., CMPD [61] and EME [55], are included for a comparison study of the FLIR dataset in Table V. In terms of mAP, our DPAL-P reaches 75.95%, which exceeds all comparison algorithms. Although DPAL-P fails to achieve the best performance on all categories, the overall mAP demonstrates its balanced trade-off among various scenarios and multiple object categories. Similarly to DroneVehicle, the IR images from FLIR have more abundant object information and provide more discriminative features for various unimodal detectors. In the case of the recent EME [55] framework, which achieves the best performance on bicycles but struggles to recognize persons. In contrast, DPAL-P shows predominance in cars and persons, thus achieving a superior mAP. We attribute that these comparison methods fail to take into account the inherent misalignment issue between multimodal features, leading to the performance discrepancy on FLIR. Obviously, benefiting from our research efforts, the misalignment problem is possibly mitigated, thereby a decent performance is delivered.

As for the M3FD dataset, we refer to Zhang et al. [55] and compare the proposed DPAL with recent GFL [63], ShaPE [55] and EME [55] methods in Table VI. In general, the RGB images provide more beneficial knowledge than their IR counterparts. By leveraging RGB and IR images, the unimodal detectors (i.e., FCOS, RetinaNet, and GFL) offer better overall performance than their original version. In terms of mAP, DPAL-R exceeds the state-of-the-art EME by 1.77%, while DPAL-P further improves performance by 2.27% over DPAL-R, reaching the best accuracy in car, bus, lamp, and truck

TABLE V
QUANTITATIVE DETECTION COMPARISON ON THE FLIR DATASET [51]

Methods	Year	Modal	Bicycle	Car	Person	mAP
RetinaNet† [21]	2020	R	55.70	77.80	44.93	59.47
FCOS† [19]	2022	R	48.88	75.59	57.33	60.60
RetinaNet† [21]	2020	I	66.37	84.27	62.17	70.93
FCOS† [19]	2022	I	50.33	78.47	71.81	66.87
CFR [51]	2020	R+I	57.77	84.91	74.49	72.39
RetinaNet† [21]	2020	R+I	67.60	85.17	61.93	71.57
FCOS† [19]	2022	R+I	56.89	83.47	76.37	72.24
CMPD [61]	2023	R+I	59.87	78.11	69.64	69.35
HalluciDet [62]	2024	R+I	–	–	–	70.90
ICA-FCOS [20]	2024	R+I	–	–	–	71.70
EME [55]	2024	R+I	**69.23**	85.10	62.27	72.23
DPAL-R (Ours)	2024	R+I	61.60	84.72	77.57	74.63
DPAL-P (Ours)	2024	R+I	63.14	**85.72**	**78.99**	**75.95**

TABLE VI
QUANTITATIVE DETECTION COMPARISON ON THE M3FD DATASET [52]

Methods	M	Person	Car	Bus	Motor	Lamp	Truck	mAP	SPF
RetinaNet†	R	44.57	74.87	57.80	44.30	36.63	49.70	51.30	.106
	I	59.17	71.17	54.17	35.90	10.43	50.83	46.97	.106
	R+I	60.10	77.27	61.63	45.42	25.67	51.80	53.63	.170
FCOS-R†	R	48.25	78.24	54.38	41.66	43.18	49.36	52.51	.121
	I	67.51	74.06	54.14	40.24	21.40	45.58	50.49	.121
	R+I	69.83	81.70	57.44	42.72	43.02	52.08	57.80	.139
FCOS-P†	R	51.04	79.79	60.33	38.58	40.05	51.39	53.53	.092
	I	67.42	74.25	49.84	34.73	20.39	47.59	49.04	.092
	R+I	69.99	81.92	63.25	41.58	42.72	50.50	58.33	.097
GFL	R	48.67	77.43	60.27	43.50	39.07	49.63	53.10	.110
	I	64.27	73.73	52.50	36.50	15.10	47.37	48.27	.110
	R+I	65.37	79.83	61.20	37.00	34.80	48.73	54.47	.172
ShaPE	R+I	60.20	77.10	58.83	39.77	24.57	51.33	51.97	.149
	R+I	65.80	79.10	62.33	41.33	30.80	53.67	55.50	.151
EME	R+I	61.47	76.40	59.20	43.10	25.97	53.17	53.23	.149
	R+I	68.43	81.23	63.37	43.90	35.77	53.53	57.70	.151
DPAL-R	R+I	**72.12**	82.53	59.91	**45.42**	43.43	53.41	59.47	.168
DPAL-P	R+I	71.54	**83.31**	**68.96**	44.79	**44.45**	**57.38**	**61.74**	.126

categories. As a result, our DPAL framework shows excellent generalizability over multiple datasets, revealing its superiority for universal multimodal object detection.

4) Inference Efficiency Comparison: To analyze the inference efficiency of the proposed method, we introduce a speed indicator on the M3FD dataset, i.e., second per frame (SPF), to compare these algorithms in Table VI. In particular, our baseline FCOS needs average 0.121s to process a single image for RGB or IR modality. As we can see, for multispectral inputs, the presented DPAL models could reach a superior inference speed to RetinaNet [21] and GFL [63]. This is because RetinaNet and GFL, as anchor-based methods, inherently exhibit slower inference speeds compared to the anchor-free FCOS. Their RGB+Infrared multi-modal variants employ early fusion [55], directly integrating high-dimensional backbone features. In contrast, FCOS and our DPAL perform feature fusion after FPN processing, where the channel dimension is efficiently reduced, significantly decreasing the computational overhead. As a result, RetinaNet and GFL demonstrate a substantial inference speed gap between their single-modal and multi-modal versions, whereas FCOS shows minimal differences.

C. Ablation Study

Here, we provide the ablation analysis on VEDAI [8] for ResNet-based DPAL and on DroneVehicle [50] for PVT-based DPAL, with the quantitative results reported in Table VII. We first describe the single-modal anchor-free baseline and then discuss the effects of CMAM and THAM for dual modality.

1) Baseline Setup: Compared to No. 1, 2, and No. 7, 8 in Table VII, it is observed that the baseline performs distinctly in different modalities of VEDAI and DroneVehicle datasets, i.e., the RGB images of VEDAI outperforms the IR images and vice versa for DroneVehicle. Furthermore, the dual-modal FCOS-R baseline achieves 78.20% mAP on VEDAI, while the FCOS-P baseline attains 73.03% mAP on DroneVehicle.

2) Effects of CMAM: We propose CMAM to address the spatial feature misalignment between RGB and IR modalities, which designs multiscale dilated offset learning, group-wise alignment mapping, and channel-wise modulation and fusion to generate cross-modal alignment representation knowledge. To validate the effectiveness of the proposed CMAM module, we equip five CMAMs into the baseline model as an ablation variant. By comparing No. 3, 4 and No. 9, 10 in Table VII, the adoption of CMAM delivers some performance boosts, i.e., 1.63% mAP on VEDAI and 1.23% on DroneVehicle, which indeed reveals the effectiveness of CMAM on both datasets.

3) Effects of THAM: Furthermore, we build an ablation variant in No. 5 and No. 11 in Table VII to demonstrate the effects of THAM. Overall, the introduction of THAM results in a performance increase of 1.78% and 0.81% in terms of mAP on VEDAI and DroneVehicle, respectively. We attribute this performance gain to the novel techniques of THAM, i.e., pseudo-anchor mechanism, semi-fixed offset generation, and offset knowledge transfer. In other words, it is precisely because THAM enables classification and regression subtasks to exploit more conducive task-specific knowledge.

4) Overall Effects: If we simultaneously integrate CMAM and THAM, the completely DPAL-R can reach 80.91% mAP on VEDAI, while DPAL-P achieves 74.91% on DroneVehicle, which indicates their combined and non-conflicted validity.

D. Visual Analysis for Dual-Perspective Alignment

To illustrate why the presented DPAL could bring consistent performance boosts across four datasets, we provide some direct evidences from an intuitive perspective. Here, we show and discuss the visual significance of CMAM and THAM.

1) Visualization of CMAM: As revealed in Fig. 7, we present several typical samples on the challenging DroneVehicle dataset, and the interesting findings are summarized as following three points. **First**, compared to the RGB modality, our baseline detectors can extract more object spatial knowledge in dark conditions from the IR modality; however, there still exists a lot of background noise in IR features. **Second**, as shown in Fig. 7(e), although the RGB+IR baseline features can capture richer object information, the activations are scattered,

TABLE VII
ABLATION STUDY ON THE DRONEVEHICLE DATASET

Ablation study of DPAL-R on the VEDAI [8] dataset				
No.	Modality	CMAM	THAM	mAP
1	I	–	–	75.34
2	R	–	–	77.74
3	R+I	–	–	78.20
4	R+I	✓	–	79.83
5	R+I	–	✓	79.98
6	R+I	✓	✓	**80.91**
Ablation study of DPAL-P on the DroneVehicle [50] dataset				
No.	Modality	CMAM	THAM	mAP
7	I	–	–	71.46
8	R	–	–	64.72
9	R+I	–	–	73.03
10	R+I	✓	–	74.26
11	R+I	–	✓	73.84
12	R+I	✓	✓	**74.91**

lacking precise object localization with high response. Moreover, for dense foreground objects, the dual-modal baseline features exhibit erroneous activation responses and misaligned foreground activations. In contrast, our proposed CMAM in Fig. 7(f) eliminates the misaligned and erroneous activations in the multimodal baseline features, generating more accurate and higher-response clues for remote sensing objects. **Third**, the horizontal and vertical offset maps actually perceive the offset of foreground object locations and accurately depict this offset information of objects. The above discoveries serve as direct evidence of the effectiveness of the proposed CMAM.

2) Visualization of THAM: To further demonstrate the validity of the proposed THAM, we visualize the predicted regression IoU scores and classification scores from detection heads with and without (w/o) THAM, as displayed in Fig. 8. By comparison, there are two significant observations as follows. **First**, leveraging THAM enables the model to predict both a larger number of correct classification scores and higher IoU scores. **Second**, by contrast to the IoU and classification scores of individual models, THAM greatly reduces the misalignment problem of prediction between classification and regression heads, ultimately boosting the detection results. The above findings adequately justify the effectiveness of THAM.

V. CONCLUSION

To cope with the misalignment issues in multimodal remote sensing images, we propose DPAL framework for anchor-free detectors. First, we design a cross-modal alignment module to address the misalignment problem in the multimodal feature fusion phase. Besides, we present a task-head alignment module to handle the misalignment issue between classification and regression heads and address the limitation of point features in fully convolutional detectors. Experiments on four multimodal object detection datasets in both remote sensing and traffic scenes reveal the effectiveness of DPAL. While DPAL effectively addresses the cross-modal misalignment problem, its primary contribution lies in alignment optimization rather than universally boosting detection accuracy. Performance boosts are most pronounced in scenarios with significant modality discrepancies (e.g., uneven illumination, low-light conditions, and occluded scenarios). In the future, we will explore the object proposal level alignment mechanism within anchor-based detectors for multimodal remote sensing imagery.

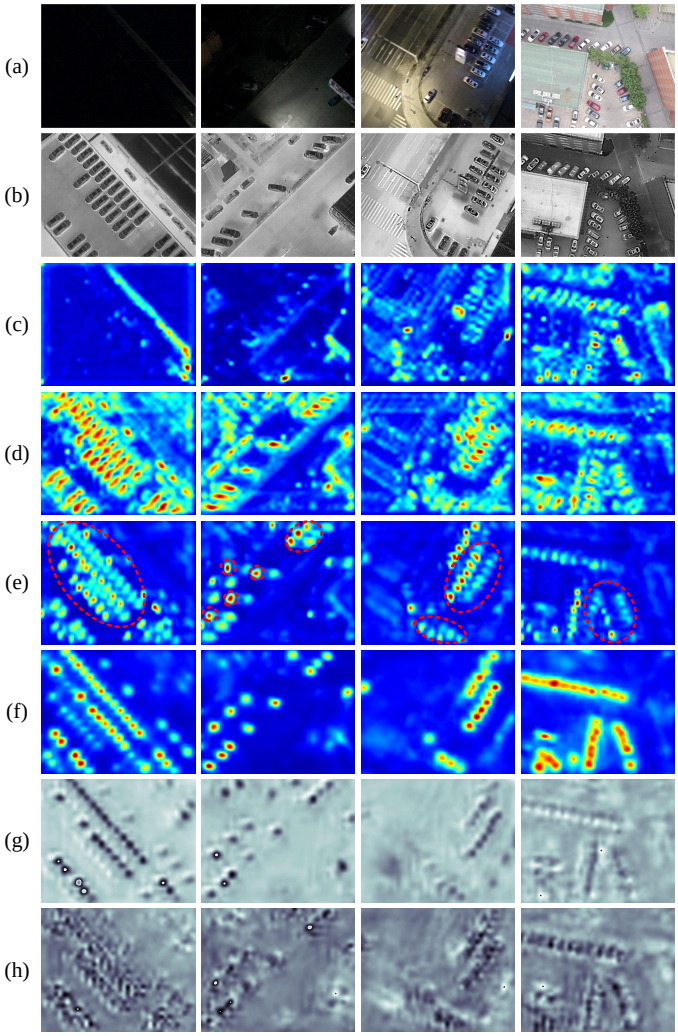

Fig. 7. Typical feature and offset visualization on the DroneVehicle dataset. (a) RGB images. (b) IR images. (c) baseline RGB features. (d) baseline IR features. (e) dual-modal baseline RGB+IR features. (f) RGB+IR features of CMAM. (g) horizontal offset maps. (h) vertical offset maps.

REFERENCES

[1] C. He, Y. Xu, Z. Wu, and Z. Wei, "Connecting Low-Level and High-Level Visions: A Joint Optimization for Hyperspectral Image Super-Resolution and Target Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.

[2] Y. Liu, Z. Xiong, Y. Yuan, and Q. Wang, "Transcending Pixels: Boosting Saliency Detection via Scene Understanding from Aerial Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

[3] Z. Yang, Q. Li, Y. Yuan, and Q. Wang, "HCNet: Hierarchical Feature Aggregation and Cross-Modal Feature Alignment for Remote Sensing Image Captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–11, 2024.

[4] L. Zhang, L. Song, B. Du, and Y. Zhang, "Nonlocal Low-Rank Tensor Completion for Visual Data," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 673–685, 2021.

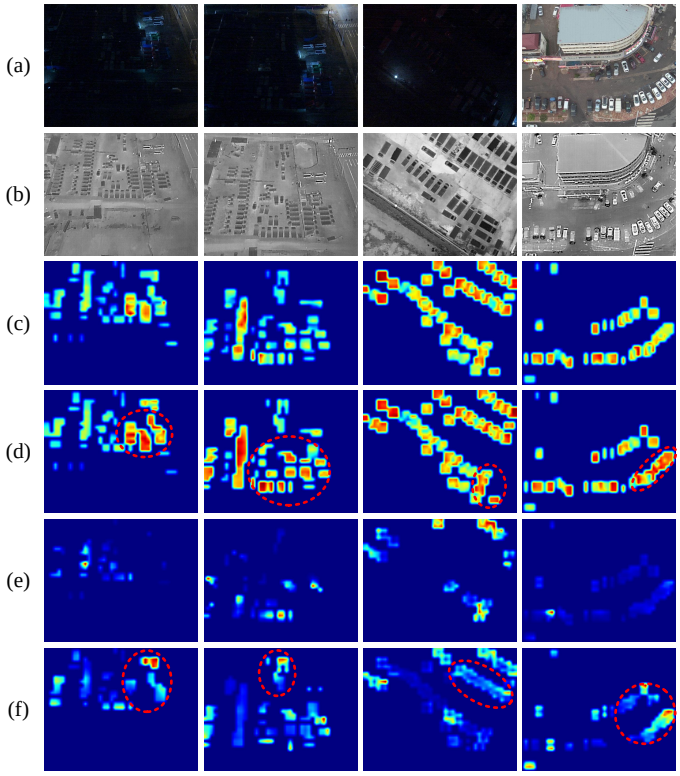

Fig. 8. Comparison of head prediction on DroneVehicle dataset. (a) RGB images. (b) IR images. (c) IoU scores w/o THAM. (d) IoU scores with THAM. (e) Classification scores w/o THAM. (f) Classification scores with THAM.

[5] Y. Liu, Q. Li, Y. Yuan, Q. Du, and Q. Wang, "ABNet: Adaptive Balanced Network for Multiscale Object Detection in Remote Sensing Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[6] Y. Liu, Q. Li, Y. Yuan, and Q. Wang, "Single-Shot Balanced Detector for Geospatial Object Detection," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2022, pp. 2529–2533.

[7] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral Pedestrian Detection: Benchmark Dataset and Baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1037–1045.

[8] S. Razakarivony and F. Jurie, "Vehicle Detection in Aerial imagery : A Small Target Detection Benchmark," *J. Visual Commun. Image Represent.*, vol. 34, pp. 187–203, 2016.

[9] J. Nie, H. Sun, X. Sun, L. Ni, and L. Gao, "Cross-Modal Feature Fusion and Interaction Strategy for CNN-Transformer-Based Object Detection in Visual and Infrared Remote Sensing Imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[10] Q. Fang and Z. Wang, "Cross-Modality Attentive Feature Fusion for Object Detection in Multispectral Remote Sensing Imagery," *Pattern Recognit.*, vol. 130, p. 108786, 2022.

[11] M. Sharma, M. Dhanaraj, S. Karnam, D. G. Chachlakis, R. Ptucha, P. P. Markopoulos *et al.*, "YOLOrs: Object Detection in Multimodal Remote Sensing Imagery," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 1497–1508, 2021.

[12] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "SuperYOLO: Super Resolution Assisted Object Detection in Multimodal Remote Sensing Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[13] Y. Zhang, M. Ye, G. Zhu, Y. Liu, P. Guo, and J. Yan, "FFCA-YOLO for Small Object Detection in Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

[14] Z. Huang, W. Li, X.-G. Xia, H. Wang, and R. Tao, "Task-Wise Sampling Convolutions for Arbitrary-Oriented Object Detection in Aerial Images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 3, pp. 5204–5218, 2025.

[15] J. Shen, W. Zhou, N. Liu, H. Sun, D. Li, and Y. Zhang, "An Anchor-Free Lightweight Deep Convolutional Network for Vehicle Detection in Aerial Images," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24 330–24 342, 2022.

[16] X. Sun, Y. Yu, and Q. Cheng, "Low-Rank Multimodal Remote Sensing Object Detection With Frequency Filtering Experts," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.

[17] F. Chu, J. Cao, Z. Song, Z. Shao, Y. Pang, and X. Li, "Toward Generalizable Multispectral Pedestrian Detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 5, pp. 3739–3750, 2024.

[18] D. Wang, J. Liu, X. Fan, and R. Liu, "Unsupervised Misaligned Infrared and Visible Image Fusion via Cross-Modality Image Generation and Registration," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Nov. 2022, pp. 3508–3515.

[19] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A Simple and Strong Anchor-Free Object Detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1922–1933, 2022.

[20] 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 Recognit.*, vol. 145, p. 109913, 2024.

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020.

[22] J. Zhu, X. Chen, H. Zhang, Z. Tan, S. Wang, and H. Ma, "Transformer Based Remote Sensing Object Detection With Enhanced Multispectral Feature Extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[23] J. Guo, C. Gao, F. Liu, D. Meng, and X. Gao, "DAMSDet: Dynamic Adaptive Multispectral Detection Transformer with Competitive Query Selection and Adaptive Feature Fusion," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2024, pp. 464–481.

[24] Y. Zhang, C. Xu, W. Yang, G. He, H. Yu, L. Yu, and G.-S. Xia, "Drone-Based RGBT Tiny Person Detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 204, pp. 61–76, 2023.

[25] S. Xu, X. Chen, H. Li, T. Liu, Z. Chen, H. Gao *et al.*, "Airborne Small Target Detection Method Based on Multimodal and Adaptive Feature Fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

[26] Y. Cao, L. Guo, F. Xiong, L. Kuang, and X. Han, "Physical-Simulation-Based Dynamic Template Matching Method for Remote Sensing Small Object Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.

[27] S. Wang, C. Wang, C. Shi, Y. Liu, and M. Lu, "Mask-Guided Mamba Fusion for Drone-Based Visible-Infrared Vehicle Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–12, 2024.

[28] Y. Zhang, X. Lei, Q. Hu, C. Xu, W. Yang, and G.-S. Xia, "Learning Cross-Modality High-Resolution Representation for Thermal Small-Object Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

[29] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align Deep Features for Oriented Object Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[30] X. Xie, Z.-H. You, S.-B. Chen, L.-L. Huang, J. Tang, and B. Luo, "Feature Enhancement and Alignment for Oriented Object Detection," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 778–787, 2024.

[31] S. Xu, H. Zhang, X. Xu, X. Hu, Y. Xu, L. Dai *et al.*, "Representative Feature Alignment for Adaptive Object Detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 689–700, 2023.

[32] S. Huang, Z. Lu, R. Cheng, and C. He, "FaPN: Feature-Aligned Pyramid Network for Dense Image Prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 864–873.

[33] Z. Song, C. Jia, L. Yang, H. Wei, and L. Liu, "GraphAlign++: An Accurate Feature Alignment by Graph Matching for Multi-Modal 3D Object Detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2619–2632, 2024.

[34] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid Feature Aligned Network for Salient Object Detection in Optical Remote Sensing Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[35] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-Aligned One-Stage Object Detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3510–3519.

[36] X. Xie, C. Lang, S. Miao, G. Cheng, K. Li, and J. Han, "Mutual-Assistance Learning for Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15 171–15 184, 2023.

[37] L. Zhao and L. Wang, "Task-Specific Inconsistency Alignment for Domain Adaptive Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2022, pp. 14 217–14 226.

[38] Z. He, L. Zhang, Y. Yang, and X. Gao, "Partial Alignment for Object Detection in the Wild," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5238–5251, 2022.

[39] H. Wang, S. Liao, and L. Shao, "AFAN: Augmented Feature Alignment Network for Cross-Domain Object Detection," *IEEE Trans. Image Process.*, vol. 30, pp. 4046–4056, 2021.
[40] Q. Chu, S. Li, G. Chen, K. Li, and X. Li, "Adversarial Alignment for Source Free Object Detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Jun. 2023, pp. 452–460.
[41] G. Han, S. Huang, J. Ma, Y. He, and S.-F. Chang, "Meta Faster R-CNN: Towards Accurate Few-Shot Object Detection with Attentive Feature Alignment," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Jun. 2022, pp. 780–789.
[42] H. Fu, H. Liu, J. Yuan, X. He, J. Lin, and Z. Li, "YOLO-Adaptor: A Fast Adaptive One-Stage Detector for Non-Aligned Visible-Infrared Object Detection," *IEEE Trans. Intell. Veh.*, pp. 1–14, 2024.
[43] N. Chen, J. Xie, J. Nie, J. Cao, Z. Shao, and Y. Pang, "Attentive Alignment Network for Multispectral Pedestrian Detection," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, Nov. 2023, pp. 3787–3795.
[44] M. Yuan and X. Wei, "C²Former: Calibrated and Complementary Transformer for RGB-Infrared Object Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–12, 2024.
[45] Y. Liu, Z. Xiong, Y. Yuan, and Q. Wang, "Distilling Knowledge from Super-Resolution for Efficient Remote Sensing Salient Object Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.
[46] Y. Wu, Y. Chen, L. Yuan, Z. Liu, L. Wang, H. Li *et al.*, "Rethinking Classification and Localization for Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10 186–10 195.
[47] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu *et al.*, "Deformable Convolutional Networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.
[48] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial Transformer Networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 28, Dec. 2015, pp. 1–9.
[49] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More Deformable, Better Results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9308–9316.
[50] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-Based RGB-Infrared Cross-Modality Vehicle Detection Via Uncertainty-Aware Learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6700–6713, 2022.
[51] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Apr. 2020, pp. 276–280.
[52] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong *et al.*, "Target-Aware Dual Adversarial Learning and a Multi-Scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5802–5811.
[53] Y. Wu, X. Guan, B. Zhao, L. Ni, and M. Huang, "Vehicle Detection Based on Adaptive Multimodal Feature Fusion and Cross-Modal Vehicle Index Using RGB-T Images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 8166–8177, 2023.
[54] C. Jiang, H. Ren, H. Yang, H. Huo, P. Zhu, Z. Yao *et al.*, "M2FNet: Multi-Modal Fusion Network for Object Detection From Visible and Thermal Infrared Images," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 130, p. 103918, 2024.
[55] X. Zhang, S.-Y. Cao, F. Wang, R. Zhang, Z. Wu, X. Zhang *et al.*, "Rethinking Early-Fusion Strategies for Improved Multispectral Object Detection," *IEEE Trans. Intell. Veh.*, pp. 1–15, 2024.
[56] J. Zhang, J. Lei, W. Xie, Y. Li, G. Yang, and X. Jia, "Guided Hybrid Quantization for Object Detection in Remote Sensing Imagery via One-to-One Self-Teaching," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.
[57] H. Wang, C. Wang, Q. Fu, B. Si, D. Zhang, R. Kou *et al.*, "YOLOFIV: Object Detection Algorithm for Around-the-Clock Aerial Remote Sensing Images by Fusing Infrared and Visible Features," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 15 269–15 287, 2024.
[58] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu *et al.*, "SSD: Single Shot MultiBox Detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 21–37.
[59] S. Liu, D. Huang, and Y. Wang, "Receptive Field Block Net for Accurate and Fast Object Detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sept. 2018, pp. 385–400.
[60] J. Wu, T. Shen, Q. Wang, Z. Tao, K. Zeng, and J. Song, "Local Adaptive Illumination-Driven Input-Level Fusion for Infrared and Visible Object Detection," *Remote Sens.*, vol. 15, no. 3, p. 660, 2023.
[61] Q. Li, C. Zhang, Q. Hu, H. Fu, and P. Zhu, "Confidence-Aware Fusion Using Dempster-Shafer Theory for Multispectral Pedestrian Detection," *IEEE Trans. Multimedia*, vol. 25, pp. 3420–3431, 2023.
[62] H. R. Medeiros, F. A. G. Peña, M. Aminbeidokhti, T. Dubail, E. Granger, and M. Pedersoli, "HalluciDet: Hallucinating RGB Modality for Person Detection Through Privileged Information," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 1444–1453.
[63] X. Li, C. Lv, W. Wang, G. Li, L. Yang, and J. Yang, "Generalized Focal Loss: Towards Efficient Representation Learning for Dense Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3139–3153, 2023.

Yanfeng Liu (Student Member, IEEE) received the M.S. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2024. He is currently pursuing the Ph.D. degree with the School of Computer Science, Wuhan University, Wuhan, China. He received the Outstanding Master's Thesis Nomination Award from the China Society of Image and Graphics in 2024 and was recognized as the TOP 50 reviewer by IEEE GRSL in 2025.

His research interests include computer vision, remote sensing, and multimedia signal processing.

Wei Guo received the Ph.D. degree from Chongqing University, Chongqing, China, in 2021. She is currently an engineer in Multisensor Intelligent Detection and Recognition Technologies R&D Center of China Aerospace Science and Technology Corporation (CASC), Chengdu, China.

Her research interests include radar target recognition, multimodal target recognition, and machine learning.

Chaojun Yao received the B.E. and M.S. degrees in electronic engineering from Harbin Institute of Technology, Harbin, China, in 2005 and 2007, respectively. He is currently an engineer in Multisensor Intelligent Detection and Recognition Technologies R&D Center of China Aerospace Science and Technology Corporation (CASC), Chengdu, China.

His research interests include cognitive radar and SAR imaging information processing.

Lefei Zhang (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Wuhan University, Wuhan, China, in 2008 and 2013, respectively. He was a Big Data Institute Visitor with the Department of Statistical Science, University College London, U.K., and a Hong Kong Scholar with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China. He is a professor with the School of Computer Science, Wuhan University, Wuhan, China, and also with the Hubei Luojia Laboratory, Wuhan, China. His research interests include pattern recognition, image processing, and remote sensing.

Dr. Zhang serves as an associate editor of IEEE Transactions on Geoscience and Remote Sensing and IEEE Geoscience and Remote Sensing Letters.